Cas4 Nucleases Can Effect Specific Integration of CRISPR Spacers

Zhufeng Zhang,a Saifu Pan,a Tao Liu,a Yingjun Li,a Nan Penga

a State Key Laboratory of Agricultural Microbiology, College of Life Science and Technology, Huazhong Agricultural University, Wuhan, Hubei, People’s Republic of China

ABSTRACT Clustered regularly interspaced short palindromic repeat (CRISPR)-Cas systems incorporate short DNA fragments from invasive genetic elements into host CRISPR arrays in order to generate host immunity. Recently, we demonstrated that the Csa3a regulator protein triggers CCN protospacer-adjacent motif (PAM)-dependent CRISPR spacer acquisition in the subtype I-A CRISPR-Cas system of Sulfolobus islandicus. However, the mechanisms underlying specific protospacer selection and spacer insertion remained unclear. Here, we demonstrate that two Cas4 family proteins (Cas4 and Csa1) have essential roles (i) in recognizing the 5’ PAM and 3’ nucleotide motif of protospacers and (ii) in determining both the spacer length and its orientation. Furthermore, we identify amino acid residues of the Cas4 proteins that facilitate these functions. Overexpression of the Cas4 and Csa1 proteins, and also that of an archaeal virus-encoded Cas4 protein, resulted in strongly reduced adaptation efficiency, and the former proteins yielded a high incidence of PAM-dependent atypical spacer integration or of PAM-independent spacer integration. We further demonstrated that in plasmid challenge experiments, overexpressed Cas4-mediated defective spacer acquisition in turn potentially enabled targeted DNA to escape subtype I-A CRISPR-Cas interference. In summary, these results define the specific involvement of diverse Cas4 proteins in in vivo CRISPR spacer acquisition. Furthermore, we provide support for an anti-CRISPR role for virus-encoded Cas4 proteins that involves compromising CRISPR-Cas interference activity by hindering spacer acquisition.

IMPORTANCE The Cas4 family endonuclease is an essential component of the adaptation module in many variants of CRISPR-Cas adaptive immunity systems. The Crenarchaeota Sulfolobus islandicus REY15A carries two cas4 genes (cas4 and csa1) linked to the CRISPR arrays. Here, we demonstrate that Cas4 and Csa1 are essential to CRISPR spacer acquisition in this organism. Both proteins specify the upstream and downstream conserved nucleotide motifs of the protospacers and define the spacer length and orientation in the acquisition process. Conserved amino acid residues, in addition to those recently reported, were identified to be important for these functions. More importantly, overexpression of the Sulfolobus viral Cas4 abolished spacer acquisition, providing support for an anti-CRISPR role for virus-encoded Cas4 proteins that inhibit spacer acquisition.

KEYWORDS CRISPR-Cas, Cas4, spacer acquisition, Sulfolobus, virus

Clustered regularly interspaced short palindromic repeats (CRISPRs) and CRISPR-associated (Cas) proteins generate a diversity of immune systems in most archaea and many bacteria that target invasive viruses and plasmids (1, 2). These diverse systems have been classified into two major classes and at least 6 basic types (types I to VI), which are further divided into multiple subtypes (3). Spacer acquisition into CRISPR arrays constitutes the first stage of the immune reaction (4); however, the molecular mechanisms involved in this process are still only partially understood. Short
DNA fragments of similar length are excised from the invasive genetic element and integrated into the CRISPR loci, facilitated by the conserved core proteins Cas1 and Cas2 (1). The first successful demonstration of spacer acquisition under laboratory conditions was obtained for the *Streptococcus thermophilus* subtype II-A system (5), and more recent studies have focused mainly on type I systems (6), including the subtype I-A systems of *Streptococcus solfataricus* and *Sulfolobus islandicus* (7–9), the subtype I-B system of *Halocarcula hispanica* (10), the subtype I-E system of *Escherichia coli* (11–13), and the subtype I-F systems of *Pseudomonas aeruginosa* (14) and *Pectobacterium atrosepticum* (15).

Spacer acquisition occurs by two related but differing mechanisms, the *de novo* acquisition and primed acquisition pathways. The former requires Cas1 and Cas2 in the subtype I-E system of *E. coli* (12, 16–18), Cas1, Cas2, Csa1, and Cas4 in the subtype I-A system of *S. islandicus* (19), and Cas1, Cas2, Cas9, and Csn2 in *S. thermophilus* subtype II-A systems (20,21). In contrast, primed acquisition involves activation by a preexisting spacer in a CRISPR locus that matches the targeted genetic element, and this then triggers spacer acquisition for different type I CRISPR-Cas systems (15–17, 22, 23).

Protospacer-adjacent motifs (PAM) in the DNA of invasive genetic elements determine the acquisition efficiency and integration specificity during the spacer acquisition process. These motifs have been shown to direct spacer acquisition for several archaea and bacteria (7, 9, 10, 12, 13). Moreover, the protospacer 3′-terminal motif is important for acquisition efficiency in the subtype I-E system of *E. coli* (13). Different subtypes adapt spacers from invasive genetic elements with a range of lengths. For example, most spacers (95%) in subtype I-E CRISPR arrays of *E. coli* are 32 bp (24), and that is likely to reflect the structure of the substrate-binding Cas1-Cas2 complex (25). However, for other systems, including those of subtypes I-A and II-A (26, 27), spacer lengths vary, and moreover, they require proteins in addition to Cas1 and Cas2 for specific adaptation (19–21). Most of these additional proteins belong to the Cas4 family and carry nuclease activities. Recently, we showed that two Cas4 family proteins, Cas4 and Csa1, are essential for CRISPR adaptation in *S. islandicus* (19), and more recent studies have shown that the Cas4 nucleases can facilitate PAM recognition and influence spacer lengths and adaptation efficiency (28–30). In particular, the two Cas4 proteins (Cas4-1 and Cas4-2) in the *Pyrococcus furiosus* subtype I-A system have been observed to have distinct activities: Cas4-1 specifies the upstream PAM, while Cas4-2 specifies the conserved downstream motif (30).

*S. islandicus* REY15A has proven to be a successful model organism for studying crenarchaeal CRISPR-Cas systems. It carries one subtype I-A and two functionally different subtype III-B interference modules that share an adaptation module encoded adjacent to the subtype I-A interference module (31), and, importantly, a comprehensive array of genetic tools has been developed for this strain (26, 32). Recently, we demonstrated that Csa3a, the activator protein of the subtype I-A CRISPR-Cas immune response in *S. islandicus*, couples transcriptional activation of spacer acquisition and DNA repair genes and, moreover, enhances CRISPR-Cas interference of invading genetic elements by activating transcription of CRISPR arrays (9, 19). We also demonstrated that the Cas4 family protein genes cas4 and csa1 were essential for spacer acquisition (19). However, the mechanisms involved require further study, including determining the essential amino acid residues of the different Cas4 proteins. Moreover, the functions of virus-encoded Cas4 family proteins remain to be determined.

In the present study, we focused on the functions of the Cas4 family proteins Cas4 and Csa1 in CRISPR spacer acquisition in *S. islandicus*. We have found new in vivo characteristics of *Sulfolobus* Cas4 compared with previous in vitro findings on *Sulfolobus* Cas4 and with Cas4 proteins of other organisms. Moreover, we provide evidence that *Sulfolobus* virus-encoded Cas4 proteins may play an important role in reducing the efficiency of host CRISPR-Cas interference.
RESULTS

Cas4 family proteins are divided into different subtypes. Many CRISPR-Cas adaptation systems require proteins in addition to Cas1 and Cas2 and, in particular, Cas4 family proteins that are encoded widely in archaeal and bacterial genomes and in their viruses. A phylogenetic tree derived from alignments of 28 Cas4 family proteins, including 21 genome-encoded Cas4 proteins from six archaeal genomes and seven archaeal viral Cas4-like proteins, is presented in Fig. 1. Here, Cas4_Sso1392 does not carry the four conserved cysteine residues (4-C motif) which form an Fe-S cluster. The sequences group into 4 subtypes, i.e., the CRISPR-related Csa1, CRISPR-associated Cas4, solo Cas4, and viral Cas4-like proteins, where the genes of solo Cas4 and viral Cas4 proteins are not linked to CRISPR loci (Fig. 1).

Sulfolobus species generally encode multiple Cas4 proteins (Fig. 1). For example, in the crenarchaeon S. islandicus REY15A, the cas4 and csa1 genes neighbor CRISPR arrays, while two other cas4 genes are not linked to CRISPR arrays (Fig. 1). The euryarchaeon

FIG 1 Unrooted bootstrapped phylogenetic tree showing a subset of archaea and their virus-encoded Cas4 family proteins. Each protein is represented by a species code followed by the gene number (Sso, S. solfataricus; SiRe, S. islandicus REY15A; Pfu, P. furiosus; Sto, Sulfolobus tokodaii; Ttx, Trichomonas tenax; Pcal, Pyrobaculum calidifontis; SIRV, Sulfolobus islandicus rod-shaped virus; SSV, Sulfolobus spindle-shaped virus). This neighbor-joining tree was generated from a T-Coffee alignment of the proteins using Mega7, with pairwise distances between sequences uncorrected. The bootstrap values shown at each node represent the percentage of all trees (10,000 total) agreeing with this topology. The subgroups of CRISPR-associated Csa1, CRISPR-associated Cas4, solo Cas4, and viral Cas4-like proteins are indicated.
**Pyrococcus furiosus** JFW02 carries one CRISPR-linked *cas4-1* gene and a second solo *cas4-2* gene (Fig. 1). However, although the *Cas4-1* (*Pfu 1119*)-encoding gene adjoins a CRISPR array, it diverges strongly from the *Sulfolobus* CRISPR-linked *Csa1* or *Cas4* (Fig. 1). Similarly, *P. furiosus* Cas4-2 (*Pfu 1994*) falls between solo Cas4 and viral Cas4-like proteins (Fig. 1). Several *Sulfolobus* rudiviruses and fuselloviruses encode Cas4 proteins, and their genes cluster in the phylogenetic tree (Fig. 1). This raises two questions: (i) what are the functions of CRISPR-linked Cas4 and Csa1 and the solo Cas4 proteins in the *Sulfolobus* genome, and (ii) what effects, if any, do the viral Cas4 proteins have on the CRISPR-Cas response of the infected host?

**Cas4 proteins define the spacer origin and recognize the PAM and the 3′ conserved DNA motif.** *S. islandicus* REY15A encodes one spacer acquisition module, constituting *Cas1*, *Cas2*, *Cas4*, and *Csa1* (Fig. 2A). The downstream gene, *csa3a*, encodes...
a transcriptional regulator that activates expression of the adaptation module (9). This, in turn, triggers CCN PAM-dependent de novo spacer acquisition (9, 19). Earlier, we demonstrated that deletion of the cas4 and csa1 genes eliminated the capacity for CRISPR spacer acquisition as judged by the inability of the deletion strain to produce larger PCR products in the CRISPR leader-proximal region (19).

In this study, we examined the effects of deletion and overexpression of Cas4 and Csa1 on spacer acquisition by high-throughput sequencing and analysis of newly acquired spacers. The high-throughput data are summarized in Table S1 in the supplemental material. Consistent with the earlier work (9, 19), the high-throughput sequencing data revealed that the control strain, carrying only the empty vector pSeSD, yielded relatively few new spacers, while many were observed in wild-type (wt) cells with csa3a overexpression (Fig. 2B). Moreover, in both strains, most new spacers were derived from the expression vector and few were from host genome DNA (Fig. 2B).

Next, we examined the effects of Cas4 and Csa1 on spacer acquisition by testing cas4 and csa1 deletion strains (Fig. 2B). Very few new spacers were detected for either deletion strain, which indicated that Cas4 and Csa1 were essential for specific spacer acquisition. In addition, we demonstrated that enhancing expression of Cas4, and also Csa1, strongly reduced spacer acquisition (Fig. 2B), possibly due to excess Cas4 or Csa1 largeprotein complex forms, such as a decameric toroid as reported previously (33), disordering the adaptation complex. Furthermore, when we examined the origins of the spacers, we found that the relatively few spacers that were acquired in the cas4 and csa1 deletion or overexpression mutants did show a stronger bias to genomic DNA than to the plasmid (Fig. 2B). This observation also reinforced that the Cas4 proteins were crucial for specific selection of spacers.

Since CRISPR-Cas interference requires PAM recognition, we analyzed the PAM and 3′-nucleotide motif of the cognate protospacers. New spacers in the wt strain carrying the empty vector were derived predominantly from protospacers carrying conserved CCN PAM and 3′-A/G motifs (Fig. 2B and C). Moreover, protospacers in the csa3a overexpression strain also predominantly carried CCN PAM and the 3′-A/G motifs (Fig. 2B and C), consistent with our earlier results (9). In contrast, the CCN PAM was absent from protospacers of both the deletion and overexpression strains of cas4 and csa1 (Fig. 2B and C); the occurrence of the 3′-A/G motif at the +2 site was also reduced in these strains (Fig. 2B and C). These results indicated that proteins Cas4 and Csa1 play a role in locating the PAM and 3′-motif during spacer selection. Thus, the absence of the PAM in these strains will prevent CRISPR-Cas interference from the defective spacers. Potentially, new spacers can accumulate from any DNA site in the four strains, and this will produce a strong bias to host genomic DNA because of the much higher DNA yields relative to those of the low-copy-number csa3a overexpression plasmid (Fig. 2B). It should be noted that few new spacer reads (~100 reads) were identified in the csa1 or cas4 overexpression strains (Fig. 2B). PAM sequences from these few reads showed less conservation. However, new spacers from csa1 or cas4 overexpression strains were matched to different locations on the genomic and plasmid DNAs, and the ratios between unique new spacer reads and total new spacer reads were 0.60 and 0.23, respectively, suggesting that identification of non-PAMs in these strains may be significant.

**Novel amino acid residues essential for Cas4 protein functions.** Cas4 family proteins carry a few highly conserved amino acid residues (34) (Fig. 3A), and we examined whether they were involved in specific protospacer selection. First, we analyzed the amino acid residues in the RecB motifs of the 28 archaeal genome-encoded or archaeal virus-encoded Cas4 proteins (Fig. 3B). We identified a conserved Asp residue in S. islandicus Cas4 (D60 in SiRe_0763) and the CRISPR-linked Csa1 proteins and two conserved Trp/Leu residues in the CRISPR-linked Cas4 proteins (YL100/101 in SiRe_0763) (Fig. 3B). Therefore, we mutated these amino acids separately with a view to establish whether they were essential for specific spacer acquisition.
Cas4 protein (SiRe_0763). (B) Alignment of the RecB domains of Cas4 family proteins.
transformation of p(csa3a+cas4) reactivated highly efficient spacer acquisition, albeit
with a strong bias to genomic DNA (Fig. 3C), transformation with p(csa3a+cas4D60A)
generated a low level of specific spacer acquisition, again with a bias to genomic DNA,
while transformation with the two other p(csa3a+cas4 mutant) plasmids failed to
restore specific spacer acquisition (Fig. 3D). We infer from these results that Cas4
positions E79 and Y100/L101 are critical for selection of the CCN PAM in protospacers.
Cas4 facilitates specific spacer integration. As described above, the CCN PAM was
not detected in most protospacers from the cas4 and csa1 deletion or overexpression
strains or in those deriving from genomic DNA, as indicated in Fig. 2B and 3C and D).
These results demonstrate that at suboptimal levels of Cas4, unspecific spacer integra-
tion occurred. This result provided an explanation for the majority of spacers being
derived from the much more prevalent host genomic DNA. Complementing cas4 in the
cas4 deletion mutant strain that overexpressed csa3a restored the selection of proto-
spacers with conserved 5'-CCN PAMs (Fig. 4). Complementing the cas4 deletion mutant
with cas4D60A, cas4E79A, and cas4Y100AL101A completely or partially restored selec-
tion of plasmid DNA protospacers carrying CCN PAMs (Fig. 4A) but not those derived
from genomic DNA (Fig. 4B). However, CCN PAM-dependent inverted integration and
slipped integration (atypical CCN) accounted for a large proportion in the cas4D60A
complementing strain (Fig. 4B), suggesting that cells with these genomic protospacers
survived CRISPR interference guided by the flipped and slipped integrated spacers. These
results indicate that amino acid residues D60A, E79, and YL100/101 are important for
selection of specific protospacers in Sulfolobus. Particularly, complementing cas4E79A
caused half of the genomic protospacers to have no CCN PAM at either end (Fig. 4B),
suggesting that it was extremely important for CCN PAM selection.
Cas4 nuclease defines spacer length. Spacer lengths vary in Sulfolobus species,
although they fall mainly in the range of 39 to 41 bp (26). Most spacers acquired in
wild-type S. islandicus strains carrying pSeSD or csa3a overexpression vectors were 40
to 41 bp (Fig. 5) similar to the lengths of spacers present in the host CRISPR arrays.
However, deletion of cas4 or csa1 not only strongly reduced spacer acquisition effi-

FIG 4 Percentages of protospacers on the plasmid (A) and on genomic DNA (B) that carry conserved CCN, atypical CCN, and
no CCN motifs. CCN, CCN motif at the 5'-adjacent region of protospacers; Atypical CCN, spacers with flipped, slipped, or
flipped-slipped integration or with a CDN or DCD motif (a single conserved C at the −2 or −3 site of the 5'-adjacent region);
No CCN, the cognate protospacers with no CCN motif at either end. All cognate protospacers were detected in high-
throughput sequencing data.
ciency, as measured by the $n$ value (Fig. 2B and 3C), but also yielded, on average, slightly shorter spacer lengths (Fig. 5). Overexpression of $csa1$ and $cas4$ also produced altered length distributions (Fig. 5). Complementing $cas4$ in the $cas4$ deletion strain did restore spacer lengths to the normal host size distribution, but whereas complementing $cas4Y100AL101A$ did not alter average spacer lengths, complementing $cas4D60A$ and $cas4E79A$ produced increased length distributions (Fig. 5). These results reinforce further that Cas4 is important for producing specific spacer acquisition.

Flipped and slipped spacers inactivate DNA interference of the subtype I-a system. In order to demonstrate that flipped spacers abolish subtype I-A CRISPR-Cas interference activity, we designed a plasmid challenge experiment (illustrated in Fig. 6A). The DNA sequence of spacer 10 from CRISPR locus 2 of *S. islandicus* REY15A was cloned into the *Sulfolobus*-E. coli shuttle vector pSeSD under the control of the araS promoter (35) with different 5′ or 3′ motifs to either activate or escape from subtype I-A and III-B DNA interference activities (Fig. 6A). These challenging plasmids and the control expression vector pSeSD were transformed into *S. islandicus* E233S, and transformation efficiencies were estimated. If the plasmids were targeted by the subtype I-A and III-B interference complexes, guided by CRISPR RNA (crRNA) from spacer 10, the transformation efficiency would be strongly reduced.

The plasmid carrying the protospacer matching spacer 10 with a 5′-CCA PAM showed a very low transformation efficiency compared with that of the control plasmid, consistent with strong subtype I-A DNA interference having occurred (Fig. 6B). Another control plasmid carrying the same protospacer with a 5′-AAG motif showed a relatively low transformation efficiency compared with that of the pSeSD plasmid but a much higher transformation efficiency than that of the plasmid carrying the CCA motif (Fig. 6B). This indicated that the 5′-AAG motif inhibited subtype I-A interference. The third plasmid carrying the spacer with protospacer 5′-AGG and 3′-TGG sequences mimics the protospacer of the flipped spacers inserted into the CRISPR array (Fig. 6A). High “flip” plasmid transformation efficiency, similar to that of the AAG plasmid (Fig. 6B), would indicate that flipped spacers had integrated into CRISPR arrays and were inactive in subtype I-A.

Slipped spacer integration can also impede subtype I-A DNA interference because the crRNAs will fail to recognize the CCN PAM sequence. To test this experimentally, we selected a protospacer from the template strand of the *S. islandicus* REY15A cmr2α gene and generated different mutants carrying single-site mutations (Fig. 6C). The altered
spacers were cloned into the *Sulfolobus* expression vector pSeSD (35) to generate plasmid-borne mini-CRISPRs under the control of the arabinose-inducible promoter (Fig. 6C) (32, 36). These challenging plasmids were then transformed into *S. islandicus* E233S, and transformation efficiencies were estimated. If the plasmid-borne spacer crRNA causes interference at *cmr2α/H9251*, then low transformation efficiency will be observed, and vice versa. Since the designed spacer is from the *cmr2α/H9251* template strand, the crRNA will not base pair with the mRNA and, therefore, will not trigger the subtype III-B DNA interference activity. The transformation efficiency of pSeSD carrying the wild-type protospacer was very low compared with that of the empty vector, consistent with strong subtype I-A interference of *cmr2α* (Fig. 6D). The plasmid carrying the Slip-1 spacer (corresponding to the protospacer with the NCC motif) showed a transformation efficiency that was lower than that of the empty vector control but higher than that of the wild-type spacer plasmid, indicating that moderate plasmid interference had occurred (Fig. 6D). This reflects that a 5′-ACC motif of the protospacer can mediate subtype I-A interference of the target DNA with reduced efficiency (Fig. 6D). No interference by the other Slip spacers was observed at *cmr2α*, except for very weak interference by the Slip +2 spacer (Fig. 6D). We conclude that the slipped spacers strongly reduced or eliminated subtype I-A CRISPR-Cas interference.

A virus-encoded Cas4 protein hinders CRISPR acquisition. The *cas4* genes are present in most known viruses of the *Sulfolobales* (37). *Sulfolobus* spindle-shaped viruses (SSV) encode Cas4 proteins that cluster with a group of Cas4 proteins associated with type I CRISPR-Cas systems of the *Thermococcales*, and a few *Cas4* proteins encoded by Icelandic rudiviruses (e.g., *Sulfolobus islandicus* rod-shaped virus [SIRV] in Fig. 1) also
form a potential branch in this family (37). However, even given the widespread occurrence of cas4 genes in viral genomes, the functions of viral Cas4 in CRISPR spacer acquisition or in virus-host interaction are still unknown. In this study, the cas4 gene of SSV Ragged Hills (SSVRH) was cloned into the csa3a overexpression plasmid and transformed into S. islandicus in order to study its effect on CRISPR spacer acquisition. The positive-control strain overexpressing csa3a (wtOEcsa3a) acquired new spacers, as evidenced by the expanded PCR bands deriving from the leader-proximal CRISPR region (Fig. 7A). In contrast, two strains from single colonies overexpressing both csa3a and SSVRH cas4 failed to yield detectable new spacers (Fig. 7A). PCR products from the leader-proximal region of CRISPR locus 1 in S. islandicus transformants. OEcsa3a, wt cells overexpressing the csa3a activator gene; OE(cs3a + Vcas4)1 and OE(cs3a + Vcas4)2, cells of single colonies 1 and 2 of wt cells overexpressing both csa3a and SSVRH cas4. (B) Analysis of the conserved motifs upstream and downstream of the protospacers detected in different Sulfolobus transformants by high-throughput sequencing data analyses using WebLogo. (C) New spacer lengths are shown at the top, and the strains used are indicated at the left. Black circles show the percentage of protospacers, and red circles indicate the spacer length of the majority. Reads of the new spacers (n) and the percentage of plasmid- or genome-derived spacers are indicated.

**FIG 7** Effect of a viral Cas4 protein on CRISPR spacer acquisition in Sulfolobus. (A) PCR analysis of the leader-proximal region of CRISPR locus 1 in S. islandicus transformants. OEcsa3a, wt cells overexpressing the csa3a activator gene; OE(cs3a + Vcas4)1 and OE(cs3a + Vcas4)2, cells of single colonies 1 and 2 of wt cells overexpressing both csa3a and SSVRH cas4. (B) Analysis of the conserved motifs upstream and downstream of the protospacers detected in different Sulfolobus transformants by high-throughput sequencing data analyses using WebLogo. (C) New spacer lengths are shown at the top, and the strains used are indicated at the left. Black circles show the percentage of protospacers, and red circles indicate the spacer length of the majority. Reads of the new spacers (n) and the percentage of plasmid- or genome-derived spacers are indicated.
the protospacer 3’-A/G motif (Fig. 7B). The length distribution of the adapted spacers from both viral cas4 expression strains was similar to that for the wt cells (Fig. 7C), indicating that viral Cas4 had no effect on spacer length. In summary, expression of SSVRH Cas4 protein strongly reduced spacer acquisition efficiency and reduced the conservation of the 3’-A/G motif (Fig. 7A and B) but did not lead to changes in the S’-CCN PAM or the spacer length (Fig. 7B and C).

**DISCUSSION**

**Conserved amino acid residues define Cas4 functions.** Cas4 family proteins occur widely within CRISPR-Cas adaptation modules of subtype I-A, I-B, I-C, I-D, I-U, and II-B and type V systems (2), and they include Csa1 of *Sulfolobus* subtype I-A systems and Csn2 of the subtype II-A systems of *S. thermophilus* (20) and *Streptococcus pyogenes* (21). High-resolution crystal structures of Cas4 proteins from *S. solfataricus* (33) and *Pyrobaculum caldifontis* (38) indicate that they carry two domains, an N-terminal RecB-like nuclease domain and a C-terminal domain containing an Fe-S cluster coordinated by four conserved cysteine residues. Moreover, biochemical studies have shown that Cas4 can exhibit S’-to-3’ and 3’-to-S’ DNA exonuclease activity, as well as ATP-independent DNA unwinding activity (33, 34, 38), which suggests that they produce single-strand DNA overhangs that are potential intermediates for insertion of new CRISPR spacers. Previously, we have demonstrated that both Cas4 and Csa1 proteins are essential for de novo spacer acquisition (19). Recently, an *in vitro* study has demonstrated that Cas4 of the *Bacillus halodurans* subtype I-C system interacts tightly with Cas1 integrase and processes double-stranded substrates with long 3’ overhangs through site-specific cleavage (29). Cas4 recognizes PAM sequences within the prespacers and prevents integration of unprocessed prespacers, ensuring correct integration (29). An *in vivo* study revealed that Cas4 of the subtype I-D system from the *Synechocystis* sp. strain 6803 pSYSA megaplasmid, in addition to Cas1 and Cas2, facilitates recognition of PAMs and defines the spacer length in a heterogenous host, *E. coli* (28). Although one conserved amino acid residue of Cas4 protein from the *Synechocystis* subtype I-D system has been identified to define the PAM and spacer length (28), more work is required to study the functions of Cas4 proteins in their native hosts.

Most recently, the Cas4 functions of the *P. furiosus* subtype I-A system have been studied *in vivo* (30). The *P. furiosus* JFW02 genome carries two Cas4 genes: cas4-1 is linked with the CRISPR array, and cas4-2 is solo. Shiimori et al. observed distinct activities in the two Cas4 proteins, especially in recognition of the upstream PAM and downstream conserved DNA motif, respectively, in *P. furiosus* (30), while in our study, we found that both Csa1 and Cas4 were crucial for specifying the upstream PAM sequence in *Sulfolobus* (Fig. 2B). Shiimori et al. have also identified one conserved amino acid residue of the RecB motif involved in nuclease activity in both Cas4 proteins for defining the spacer length and orientation in *P. furiosus* (30). We experimentally identified three conserved amino acid residue sites of the RecB motif that were important for Cas4 function, in addition to that found by Shiimori and coworkers: residue D60, conserved in CRISPR-linked Csa1 proteins; Y100/L101, conserved in CRISPR-linked Cas4 proteins; and E79, conserved in all Cas4 family proteins (Fig. 3B). In Fig. 8 we summarize the experimentally studied conserved amino acid residues in the RecB motif of Cas4 family proteins involved in different functions in CRISPR spacer acquisition. D60, identified in *S. islandicus* Cas4, is essential for spacer acquisition and defines spacer length, but it is not responsible for defining the S’-CCN PAM and 3’ nucleotide motif (Fig. 8B). D70, corresponding to *S. solfataricus* Sso0001, is essential to its *in vitro* nuclease activity (34), and D70 corresponds to the residue in Cas4-1 or Cas4-2 of *P. furiosus* and determines acquisition efficiency, spacer length, and integration orientation and recognizes the 5’ or 3’ protospacer motifs *in vivo* (30) (Fig. 8B). E79, identified in *Sulfolobus* Cas4, influences acquisition efficiency and is important for determining spacer length, integration orientation, and the 5’ and 3’ motifs of protospacers (Fig. 8B). However, H91, corresponding to a residue in Cas4-1 or Cas4-2 of *P. furiosus*, is unimportant for Cas4 functions (30) (Fig. 8B). Y1L00/L01, identified in this study, strongly
effects Cas4 functions (Fig. 3D), probably because these amino acid residues affect the nuclease activity or because the conserved hydrophobic residue L101 is important for Cas4 structural integrity. Similarly, the 4-C cluster of both Cas4-1 and Cas4-2 of *P. furiosus* is essential for Cas4 functions (30), consistent with published findings that the Fe-S cluster is required for the structural integrity of Cas4 proteins (33).

*Csa1* belongs to the Cas4 family proteins, which are specific for subtype I-A CRISPR-Cas systems (2). In this study, we report its functions in defining PAM, orientation and spacer length (Fig. 2, 4, and 5). Even though Csa1 proteins carry conserved amino acid residues as Cas4 protein do, a long insert is found in the Csa1 amino acid sequence, suggesting that some different roles may be played by Csa1 proteins.

A putative anti-CRISPR-Cas mechanism evolved by viruses. In this study, we show that overexpression of Cas4 or Csa1 can trigger a strong reduction in CRISPR-Cas spacer acquisition efficiency (Fig. 2B) and induce a larger fraction of new spacers to be integrated in a flipped or slipped manner or to be derived from protospacers with no conserved 5′-CNNPAM motif (Fig. 4). Overexpression of the host-encoded Cas4 reduces spacer acquisition efficiency and causes defective acquisition. We quantified the transcription levels of the *cas4*, *csa1*, and *cas4* mutant genes in the studied strains. As shown in Table S2 in the supplemental material, the transcription level of the *cas4* gene from the wild-type cells overexpressing the *csa3a* and *cas4* genes [wtOE(*csa3a*h11001*cas4*)] was much higher (22.7-fold) than that from the wild-type cells overexpressing only the *csa3a* gene (wtOE(*csa3a*)), and much lower spacer acquisition efficiency was observed in the wtOE(*csa3a*/*H11001*cas4*) strain. However, the transcription level of the *cas4* gene in the Δ*cas4* OE(*csa3a*/*H11001*cas4*) strain was higher than that in the wtOE(*csa3a*/*H11001*cas4*) strain (137.8-fold) and also higher than that in the wtOE(*csa3a*/*H11001*cas4*) strain (6.1-fold). The higher expression level of *cas4* in the Δ*cas4*OE(*csa3a*/*cas4*) strain than in the wtOE(*csa3a*/*cas4*) strain is probably because deletion of the chromosomally encoded *cas4* gene interrupted the *cas4* antisense RNA (39). This antisense RNA against the chromosomally encoded *cas4* transcript might mediate inhibition of *cas4* translation and degradation the *cas4* transcripts. However, this raises a question about how overexpression of *cas4* in the wt strain almost abolished spacer acquisition but overexpression of *cas4* in the Δ*cas4* strain restored adaptation efficiency (Fig. 2B and 3C). In our previous work, the *csa1* transcript level was 4-fold higher than the *cas4* transcript level as seen from the transcriptome data for the wtOEcsa3a strain (19). Therefore, in the wtOEcsa3a strain, Cas4 specifies the PAM on target DNA and probably binds one Cas1 subunit of the Cas1-Cas2 complex for

---

**FIG 8** Summary of experimentally studied conserved amino acid residues of Cas4 family proteins involved in the functions of CRISPR spacer acquisition. (A) Experimentally characterized amino acid residues in Cas4 family proteins, using the Cas4 (*SiRe_0763*) protein from *S. islandicus* REY15A. Conserved residues represented by black lines and letters were identified by Shimori et al. (30), and conserved residues represented by green lines and letters were examined in this study. (B) Functions of the experimentally characterized amino acid residues in the Rec8 nuclease motif of Cas4 proteins. Weak to strong influences are indicated by + to ++++, and no influence is indicated by −. D70 and H91 have functions in Cas4-1 and Cas4-2 from *P. furiosus* (30).
specific integration (in the correct direction) at the leader-proximal site. This is confirmed by the new spacer data from high-throughput sequencing of the leader-proximal regions in the wtOEcsa3a strain showing that most of the protospacers have a 5′-end-adjacent CCN PAM sequence (Fig. 2B and C). However, in the wtOE(csaa3a + cas4) strain, Cas4 proteins at a higher concentration possibly bind all four Cas1 subunits of the Cas1-Cas2 complex. Therefore, the saturated binding of Cas4 with the Cas1-Cas2 complex possibly affects the conformation of the adaptation complex, resulting in much lower acquisition efficiency (Fig. 2). Saturated binding of Cas4 with the Cas1-Cas2 complex also probably fails to specify the PAM and correct direction of spacer acquisition (Fig. 4). In the Δcas4OE(csaa3a + cas4) strain, a much higher expression level of Cas4 protein might make Cas4 form a toroidal structure with 6 subunits or self-assemble (33, 34). Thus, under this condition, less Cas4 binds with the Cas1-Cas2 complex, and the acquisition efficiency restored (Fig. 3). However, to confirm this hypothesis will require much biochemical and structural study of the Cas4-Cas1/Cas2 complex in combination with genetic studies.

Cas4-like nucleases are encoded by archaeal viruses (40, 41), phages (42), and transposable elements (43), and these genetic elements may use their Cas4 proteins to interact with the host-encoded adaptation complex. Moreover, a Cas4-like protein from the rudivirus SIRV2 has been shown to possess both 5′-to-3′ exonuclease and endonuclease activities in vitro (40, 44), suggesting that the viruses encode functional Cas4 nucleases that could participate in CRISPR spacer acquisition. Importantly, the expression of rudivirus SIRV2-encoded Cas4-like protein continuously increased after infection into Sulfolobus cells (45). In this study, expression of SSVRH Cas4 strongly reduced spacer acquisition compared to that for the cells lacking SSVRH Cas4 (Fig. 7A and B). High-throughput sequencing data also support that viral Cas4 hinders de novo spacer acquisition in Sulfolobus (Fig. 7B). Analyzing the new spacers detected from the viral Cas4 expression strain, we found that SSVRH Cas4 has no effects on defining the 5′-CCN PAM or spacer length (Fig. 7B and C) but does affect the 3′-A/G motif (Fig. 7B). Together, these results suggest that some invasive genetic elements may have accrued Cas4 proteins to corrupt CRISPR-Cas spacer acquisition by strongly reducing the spacer acquisition efficiency. In turn, the mobile genetic elements may have evolved to escape from CRISPR-Cas interference by the Cas4-disordered adaptation module (the possible mechanism discussed above), as also inferred from bioinformatic analyses (37).

Although expression of SSVRH Cas4 only reduced spacer acquisition efficiency and did not affect integration orientation or spacer length (Fig. 7), this cannot exclude the effects of viral Cas4 on atypical spacer acquisition. For example, it has been shown recently that a Campylobacter bacteriophage-encoded Cas4 homolog induces de novo spacer acquisition exclusively from host DNA (42). We extracted all the new spacers with fewer than 3 mismatches to genomic DNA from that study (42) and analyzed the 5′ and 3′ motifs of the cognate protospacers. However, no conserved PAM or 3′ motifs were detected on either end of the protospacers, indicating two possibilities: (i) the presence of a DNA interference module counterselects the functional genome-derived spacers with conserved PAMs and (ii) the viral Cas4 produces defective spacer acquisition that would prevent CRISPR-Cas targeting of the genomic DNA (data not shown). Given that only host DNA, and no viral DNA, is sampled into the CRISPR array, both possibilities could be reasonable.

During viral infections, archaeal and bacterial cells have also evolved counteracting defense mechanisms, including the production of antisense RNAs. Thus, a high level of antisense transcripts encompassing cas2 and the upstream region of cas4 is detected upon STSV2 viral infection in S. islandicus. The antisense RNA yields increases significantly at around 7 days postinfection and then gradually decreases (39). This raises the possibility that host cells employ antisense RNAs to control the Cas4 expression level in order to minimize its influence on acquisition efficiency and the consequent inactivation of CRISPR-Cas interference against invasive genetic elements.
MATERIALS AND METHODS

Strains, growth, and transformation of *Sulfolobus*. The *S. islandicus* strains employed, including the genetic host E233S (ΔpaprEΔΔac3) (46), the E233S derivative cas deletion mutants, and the overexpression strains, were cultured in SCVy medium (0.2% sucrose, 0.2% [wt/vol] vitamin-free Casamino Acids (BD Difco vitamin assay), 0.005% yeast extracts, and a mixed vitamin solution) at 78°C. *S. islandicus* cells were transformed by electroporation, and transformants were selected on two-layer phytopel gel plates as described earlier (46). Interference and control plasmid transformation efficiencies were calculated as CFU per microgram of DNA of constructed plasmids. For experiments measuring new spacer uptake at the leader-repeat region of CRISPR loci, cells were grown in SCVy medium.

Plasmid construction. The *csa1*, *cas4*, *cas4* mutant, and viral *cas4* overexpression plasmids were constructed by cloning *cas4* and its mutant genes into the pCsa3a plasmid (9) to form an operon with the *csa3a* gene. The *csa1* gene (SiRe_0760), *cas4* gene (SiRe_0763), and viral *cas4* (p22) gene were amplified from *S. islandicus* REY15A or the fusellovirus SSVRH by PCR using FastPfu DNA polymerase (TransGene, Beijing, China) with the primer sets listed in Table S1 in the supplemental material and cloned into the *csa3a* overexpression plasmid pCsa3a (9), yielding the pCsa3a-Csa1, pCsa3a-Cas4, and pCsa3a-vCas4 plasmids, on which additional Shine-Dalgarno sequences were generated for translational initiation of Csa1 and host and viral Cas4 proteins. For mutations in the *cas4* gene, linear plasmids which were amplified from pCsa3a-Cas4 by PCR using FastPfu DNA polymerase contained the designed *cas4* gene mutations. After purification, these linear plasmids were cyclized by Gibson assembly before transformation into *E. coli* BL21 cells.

To construct interference plasmids for testing the flipped spacers, the S10 plasmor CRISPR locus 2 with different 5' end PAM and 3' end sequences (Fig. 5A) was cloned into the pSeSD vector. For testing the slpped spacers, a protospacer sequence from the *cmr2*a gene template strand and its “slipped” derivatives were cloned into the site between two repeat sequences to form a mini-CRISPR under the control of the araS promoter in pSeSD. Interference plasmids were electroporated into *S. islandicus* strain E233S, and transformation efficiencies were calculated. Primers used for constructing overexpression and interference plasmids are listed in Table S3 in the supplemental material.

PCR amplification of the integrated spacers. Transformants were cultured in 10 ml SCVy medium at 78°C until the optical density at 600 nm (OD600) reached 0.3. Samples were taken from each culture (0.1 ml), and DNA was extracted from cells and employed as a PCR template. Leader-proximal regions of CRISPR locus 1 were amplified using *Taq* polymerase with forward and reverse primers CRISPR-F2 and CRISPR-FS-R (or CRISPR-FS-R), PCR products were separated by 1.5% agarose gel electrophoresis and visualized by ethidium bromide staining to identify the expanded bands. The PCR products were purified using a PCR cleanup kit and sent for high-throughput sequencing. Primers used to amplify the leader-proximal regions are listed in Table S1.

High-throughput sequencing and bioinformatics analysis. Six single colonies on each transformant plate were picked up and mixed in a single tube containing SCVy medium for growth. Genomic DNA was extracted from each transformant and used for amplification of the leader-proximal region. The leader-proximal regions were also amplified from two samples of single colonies carrying a *csa3a* and viral *cas4* overexpressing plasmid. Equal amounts of the purified leader-proximal PCR products were subjected to HiSeq3000 sequencing (National Key Laboratory of Crop Genetic Improvement). After paired-end data assembly and low-quality data filtration, reads containing multiple (≥2) repeats were selected. For reads containing two repeats, the intervening sequence was considered the original first spacer (5'→1'). For reads with three or four repeats, the leader-proximal spacers (5'→2 and 5'→1) were considered the new spacers. Using the BLASTN program against the pSeSD plasmid (35) or the *S. islandicus* REY15A genome sequence (31), the protospacer sequence was identified for each spacer. The 3-bp sequence at the 5' end of each protospacer was considered the PAM region. Perl scripts were run to analyze the protospacers. The conserved motifs of the 5' end- and 3' end-adjacent sequences of the protospacers were analyzed using WebLogo (47). The neighbor-joining tree was generated from a phylogenetic alignment of the proteins using Mega7 (48, 49), with pairwise distances between sequences uncorrected.

Accession number(s). All high-throughput sequencing reads have been deposited at the SRA database with the accession number PRJNA498890.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/JB.00747-18.

SUPPLEMENTAL FILE 1, PDF file, 0.2 MB.

ACKNOWLEDGMENTS

We thank Roger A. Garrett and Qunxin She of the Danish Archaea Centre in the University of Copenhagen, Denmark, for critical reading of the manuscript.

This work was supported by the National Natural Science Foundation of China (no. 31671291 and 91751104 to N.P.), the National Postdoctoral Program for Innovative Talents (no. BX20180112 to T.L.), and the Fundamental Research Funds for the Central Universities (no. 2662015PY038 and 2662015PX199 to N.P.). Funding for open access charges is from the National Natural Science Foundation of China (no. 31671291).
REFERENCES


