Fidelity of prespacer capture and processing is governed by the PAM mediated interactions of Cas1-2 adaptation complex in CRISPR-Cas type I-E system

Kakimani Nagarajan Yoganand, Manasasri Muralidharan, Siddharth Nimkar, and Baskaran Anand*

Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Guwahati 781039, INDIA

Running title: Mechanism of prespacer capture and processing in E. coli

* To whom correspondence should be addressed.
Tel: +91-361-2582223; Fax: +91-361-2582249; Email: banand@iitg.ac.in

Keywords: Prespacer processing, CRISPR-Cas, CRISPR adaptation, Protospacer adjacent motif (PAM), Cas1, Cas2, Prespacer integration, Bacterial immunity, Exonuclease

Abstract
Prokaryotes deploy CRISPR-Cas based RNA guided adaptive immunity to fend off mobile genetic elements such as phages and plasmids. During CRISPR adaptation, which is the first stage of CRISPR immunity, the Cas1-2 integrase complex captures invader-derived prespacer DNA and specifically integrates it at the leader-repeat junction as spacers. For this integration, several variants of CRISPR-Cas systems use Cas4 as an indispensable nuclease for selectively processing the protospacer adjacent motif (PAM) containing prespacers to a defined length. Surprisingly, however, a few CRISPR-Cas systems such as type I-E are bereft of Cas4. Despite the absence of Cas4, how the prespacers show impeccable conservation for length and PAM selection in type I-E remains intriguing. Here, using in vivo and in vitro integration assays, deep sequencing and exonuclease footprinting, we show that Cas1-2/I-E– via the type I-E specific extended C-terminal tail of Cas1 –displays intrinsic affinity for PAM containing prespacers of variable length in Escherichia coli. Although Cas1-2/I-E does not prune the prespacers, its binding protects the prespacer boundaries from exonuclease action. This ensures the pruning of exposed ends by exonucleases to aptly sized substrates for integration into the CRISPR locus. In summary, our work reveals that in few CRISPR-Cas variants such as type I-E, the specificity of PAM selection resides with Cas1-2, whereas the prespacer processing is co-opted by cellular non-Cas exonucleases, thereby offsetting the need for Cas4.

Introduction

Prokaryotes utilize an adaptive immune response mediated by clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR associated proteins (Cas) in order to respond against infections by mobile genetic elements (MGE), viz., phages and plasmids (1-4). CRISPR encompasses a typical architecture that comprises of an array of direct repeats (~30-40 bp) (5), which are partitioned by short spacer sequences of viral origin (6,7). The repository of spacers acts as a vaccination card and this genetic memory acquired during pathogenic invasions guides the adaptive immune response by CRISPR-Cas
Mechanism of prespacer capture and processing in E. coli

The molecular choreography of CRISPR-Cas defence is trichotomized into (I) adaptation, (II) maturation and (III) interference stages (3). Upon phage attack, the CRISPR adaptation machinery derives short stretches of nucleic acids (prespacers) from the invaders and incorporates them into the CRISPR array. This process generates an infection memory and immunizes the host (2,4,8-10). An A-T rich leader region present upstream to the repeat-spacer array regulates the transcription of the CRISPR locus and the subsequent Cas nuclease mediated processing of this inert transcript generates short regulatory mature CRISPR RNA (crRNA) (11-14). In association with other Cas proteins, the crRNA forms a surveillance complex that detects the recurring infections by MGE via base complementarity. This event triggers the Cas nucleases to annihilate the MGE thereby interfering with the spread of infection (1-3,13).

The CRISPR adaptation machinery expands the spacer archive, thus bestowing the host with a compendium of immunological memory to counter the ever-evolving phages. Highly conserved Cas1 and Cas2 spearhead the spacer acquisition (9,10,15), however, previous investigations have also demonstrated the indispensable role of additional Cas proteins and host factors in the uptake of spacers that are derived from phage DNA and RNA (16-28). Though the spacer sequences are extremely variable, the majority of the organisms (CRISPR types I, II and V) display conservation of a 2-5 nt protospacer adjacent motif (PAM) at its origin site in MGE (8,10,29-31). In addition to specifying the prespacer region for integration, PAM also guides the differentiation between self and non-self during the interference step (32-34). Mutations in the PAM region on MGE lead to impaired target recognition, thus evading CRISPR-Cas immune response. In such circumstances, the imperfectly paired surveillance complex at the target region directs the interference machinery and Cas1-2 to display an inflammatory immune response by rapidly acquiring more spacers by a process termed ‘primed adaptation’ (24,35-38).

The CRISPR adaptation pathway proceeds via two sequential events: the capture of the prespacer fragments from invading MGE and the site-specific integration of these captured fragments at the leader-repeat junction. The CRISPR adaptation machinery of Escherichia coli (type I-E) derives its spacers predominantly from the DNA debris generated during the action of multi-subunit RecBCD DNA repair complex (18). Modulation in helicase, exonuclease and endonuclease activities of RecBCD results in the production of single-stranded DNA fragments ranging from tens to thousands of nucleotides in length (39,40). However, the legitimate spacers in E. coli are strictly 33 bp in length and abut a 5’-AAG-3’ PAM (where ‘G’ is destined to be the first residue of the spacer) (10,30,33,35,41,42). Additionally, structural studies have demonstrated that Cas1-2/I-E efficiently binds to partial duplex prespacers that are 33 bp in length (42,43). The existence of such spacer sized DNA fragments is infinitesimal among the RecBCD products. Moreover, a previous study also demonstrated the incorporation of 33 bp spacers that are directly acquired from electroporated longer DNA duplexes (63 bp) (44). These findings augment the involvement of an additional DNA trimming step to generate befitting substrates for CRISPR adaptation.

Recent studies in Sulfolobus solfataricus (type I-A), Sulfolobus islandicus (type I-A), Bacillus halodurans (type I-C), Synechocystis sp.6803 (type I-D), Pyrococcus furiosus (type I-G) and Geobacter sulfurreducens (type I-U) (25,45-51) highlighted the indispensable role of Cas4 nuclease in PAM selection and prespacer processing. The occurrence of Cas4 is prevalent in type I CRISPR-Cas systems except in subtypes I-E and I-F (15). In one case, it was observed that an extended variant of Cas2 with an unorthodox C-terminus DnaQ exonuclease domain assists Streptococcus thermophilus DGCC7710 (type I-E) in prespacer trimming (52). Cas2-DnaQ domain fusion is non-ubiquitous, and even model organisms for spacer acquisition studies like E. coli (type I-E) do not harbour it. Though recent studies envisage the involvement of exonucleases during spacer acquisition in E. coli (53), the
molecular events guiding PAM selectivity and prespacer processing remain obscure.

Upon production of legitimate prespacers, Cas1-2 integrase complex catalyses the prespacer incorporation at the leader adjoining repeat (10,23,24,41,54-56). This polarized mechanism records the chronology of infections by positioning the new spacer towards a promoter encompassing leader. During recurring infections, swift expression of the latest spacers ensures a productive fight by a rapid and robust immune response (54). In conjunction with Cas1-2, various conserved DNA motifs present in the leader and repeat regions mandate the fidelity of prespacer integration (16,17,54,57-65). In E. coli (type I-E), upon interaction with these conserved regions, a sequence-specific genome architectural protein termed integration host factor (IHF) restructures the leader and generates a docking site for Cas1-2 integrase at the leader-repeat junction (16,17,66). In contrast, the presence of Cas1-2 binding site abutting the leader proximal region obviates the involvement of host factors during spacer acquisition in CRISPR type II-A system (54,67,68).

Unlike many of the type I CRISPR-Cas encompassing prokaryotes, E. coli (type I-E) lacks Cas nucleases like Cas4 and Cas2-DnaQ to generate processed prespacers for efficient homing at CRISPR locus (15). Nevertheless, this inadequacy does not appear to hinder PAM selection or spacer size preference (10,30,33,41,44). Intrigued by these observations, we sought to understand how prespacers are selected and tailored to the appropriate size for CRISPR adaptation. Here, we demonstrate that the PAM directed interactions with longer DNA fragments signals Cas1-2 to demarcate potential prespacer boundaries. Upon supplementing the reaction with exonucleases to mimic the cellular environment, we found that Cas1-2-DNA nucleoprotein complex could protect DNA fragments of ~33 bp length. Further, we showed that these protected fragments could be efficiently integrated into the CRISPR locus. These findings demystify the mechanism by which E. coli efficiently scales the fragments of the foreign DNA to generate viable prespacers of desired length, in contrast to other CRISPR-Cas subtypes that possess dedicated prespacer processing nucleases such as Cas4.

Results

Length of the prespacers dictates their integration at the CRISPR array

We utilized previously established in vitro spacer integration assay (17) to understand the parameters that necessitate the uptake of prespacers during CRISPR adaptation. In this assay, we employed Cas1-2 integrase, IHF, prespacers and linear CRISPR DNA (69 bp leader followed by two repeat-spacer units) (Figure 1A and S1A). Spacers in E. coli are routinely derived from the remnant ssDNA fragments generated by the action of RecBCD mediated double-stranded break repair (18). Upon reannealing to their complementary sequences, variably sized DNA fragments that encompass blunt ends, 3’- or 5’-overhangs could be generated. To simulate these conditions in vitro, we employed various types of DNA fragments such as P33 (33 bp duplex), P33[ss] (33 nt ssDNA), P23[3’-5] (23 bp duplex with 5 nt 3’-overhangs), P23[5’-5] (23 bp duplex with 5 nt 5’-overhangs), P23[3’-10] (23 bp duplex with 10 nt 3’-overhangs) and P63 (63 bp blunt duplex) as prespacers in the integration assays (Figure 1B).

Upon incubation with Cas1-2 and IHF, prespacer integration proceeds via a trans-esterification reaction, wherein 3’-OH of the Cas1-2 bound prespacer makes two nucleophilic attacks at the target sites to get itself ligated into the CRISPR array (L-R1 junction in top strand and R1-S1 junction in bottom strand) (Figure 1A) (55,56). To precisely identify the site of nucleophilic attack we employed the CRISPR DNA substrates that were 5’-end labelled with fluorescein (FAM) either on the top strand (CD-T* in Figure 1A (i)) or on the bottom strand (CD-B* in Figure 1A (ii)). Likewise, to monitor the prespacer ligation we used an unlabelled CRISPR DNA (CD-U in Figure 1A (iii) and (iv)) and
Mechanism of prespacer capture and processing in E. coli

prespacers with FAM at their 5'-ends. Here, we observed that P23[3'-5] or P23[5'-5] or P33 could alone make successful nucleophilic attack at the integration sites and result in the generation of top strand (L) and bottom strand (R') cleavage products from CD-T* and CD-B*, respectively (Lanes 3, 4 and 5 in Figure 1C and 1D; Figure S1B and S1C). Further, utilizing 5'-FAM labelled prespacers, we observed the ligation of P23[3'-5], P23[5'-5] and P33 at these nicked sites on the top strand (P+R) and bottom strand (L'+P) (Lanes 3, 5 and 7 in Figure 1E; Figure S1B and S1C).

Interestingly, we did not observe bands corresponding to the integration products when we substituted the reaction mixtures with P33[ss], P23[3'-10] and P63 (Lanes 6, 7 and 8 in Figure 1C and 1D; Lanes 9, 11 and 13 in Figure 1E; Figure S1B and S1C). These findings suggest that either duplex (P33) or partial duplex (comprising of 3'-overhang (P23[3'-5]) or 5'-overhang (P23[5'-5])) prespacers with an effective length of 33 nt are strictly required during CRISPR adaptation. This bias in prespacer size preference could possibly arise due to the weakening of Cas1-2 interaction with long substrate precursors (such as P23[3'-10] and P63 with effective length of 43 nt and 63 nt, respectively) and/or inefficient integration of such DNA fragments at the target site in CRISPR locus. To test whether longer prespacers (>33 nt) have weak interaction with Cas1-2 thereby leading to inefficient integration, we performed EMSA to assess the interaction between Cas1-2 and various prespacers (Figure S2 and 1F). This indicated that the affinity of Cas1-2 towards P23[3'-10] (K_D = 748.5±163.7 nM) is comparable to its affinity to P23[3'-5] (K_D = 648.5±136.2 nM). Similarly, the affinity of P63 (K_D = 2.842±0.372 µM) for Cas1-2 is on par with that of P33 (K_D = 2.885±0.613 µM). These experiments suggest that Cas1-2 can interact with DNA fragments of varying lengths; however, the integration at the target site could be achieved only in the presence of DNA fragments with an effective length of 33 nt (Figure 1G).

Cas1-2 foothold protects potential prespacer regions during exonuclease action

CRISPR adaptation in E. coli mandates precisely sized prespacers (Figure 1). To cater to this need, long DNA fragments generated during RecBCD activity have to be trimmed further by nuclease action. Of the multitude of proteins encoded by the Cas operon, only Cas1 and Cas2 contribute to naïve spacer acquisition in E. coli. We could not notice processing of longer prespacers (P23[3'-10] and P63) even at higher Cas1-2 concentrations (Figure S3A).

The type I-E system is devoid of prespacer processing exonuclease Cas4 and its deficit cannot be complemented by Cas1-2 alone (Figure S3A). Hence, we predicted the involvement of cytoplasmic exonucleases in trimming the longer prespacer to a suitable length. Having established that Cas1-2 indeed binds DNA fragments of variable length (Figure 1F and S2), we sought to test the fate of Cas1-2 bound DNA fragments against exonucleases. Here, Cas1-2-P63 nucleoprotein complex was treated with a mixture containing 5'→3' acting T5 exonuclease (T5exo) and 3'→5' acting Exonuclease III (ExoIII) (Figure 2A). To our surprise, we identified a smear of protected DNA fragments (P63exo+) that ranged from 30-40 nt in the sample containing both P63 and Cas1-2 (Lane 8 in Figure 2A and Lanes 4-13 in Figure S3B), whereas, such protection was not observed when we treated P63 in the absence of Cas1-2 or in the presence of either Cas1 or Cas2 (Lanes 5, 6 and 7 in Figure 2A). Coincidentally, the length of the protected fragments corresponded to legitimate spacer size in E. coli (~33 nt). Additionally, it was noted that this protection was absent when Cas1-2-P63 was treated with ExoIII alone (Lane 11 in Figure 2A). The nuclease generates 5'-overhangs with which Cas1-2 binds weakly (compare P23[3'-5] and P23[5'-5] in Figure 1F and S2). Therefore, it appears that ExoIII seems to dislodge the weakly bound Cas1-2 from its position on the prespacer. In contrast, the treatment of Cas1-2-P63 with 5'→3' acting T5exo resulted in incompletely digested fragments that were predominantly of higher
Mechanism of prespacer capture and processing in E. coli

As the Cas1-2 protected P63exo+ fragments were approximately of E. coli spacer size, we wondered whether they could act as potential prespacers for integration. To test this hypothesis, we purified and utilised P63exo+ DNA fragments as prespacers in spacer integration assay. In line with the previous experiment (Figure 1), we could not observe any integration events when we employed longer prespacer P63 (Lanes 3 and 7 in Figure 2B). To our surprise integration was observed when P63exo+ was employed (Lane 4 and 8 in Figure 2B; Lane 7 in Figure S3C). In this case, though we monitored efficient nucleophilic attack at the top strand (L in Lane 4 of Figure 2B), the integration at the bottom strand seemed to be sparse (R’ in Lane 8 of Figure 2B). By recapitulating these observations, we could suggest that Cas1-2 mediated binding of large DNA fragments secure the boundaries of suitable prespacers from the exonucleolytic action of cellular nucleases in E. coli (Figure 2C).

PAM directed binding of Cas1-2 defines the boundary for prespacers

Since Cas1-2 binding was shown to mark the spacer boundaries (Figure 2), we sought to identify and map these regions. As no protected DNA fragments appeared upon treatment of Cas1-2-P63 complex with ExoIII alone (Lane 11 in Figure 2A), we sought to utilize T5exo digestion (Lane 13 in Figure 2A) to demarcate the prespacer boundaries defined by Cas1-2. To accomplish this, we utilised variants of 63 bp blunt-ended prespacers that encompass fluorescein labelled 3’-end (6-FAM) either on the top (P63T*) or on the bottom (P63B*) strand (Figure 3A). Further, in order to identify the footprints of Cas1-2 on these 3’-end labelled prespacers, we incubated Cas1-2 bound prespacer complex with T5exo. Here, the Cas1-2 binding on the prespacer acts as a roadblock that stalls the 5’→3’ progression of T5exo. The length of the resultant labelled fragments specifies the stalling points of the exonuclease, which in turn, indicates the binding position of Cas1-2 on the prespacer. Utilizing this approach, we mapped the cleavage termination points on the top and bottom strands of the prespacer. After T5exo treatment of P63T*, we could observe ~28 nt labelled fragment. This is indicative of an inherent nuclease stalling point (in the absence of roadblocks such as Cas1-2) around the 28th nt position from the labelled end in P63T* (Lane 2 in Figure 3B; Lane 12 in Figure 2A). However, owing to complete exonucleolytic cleavage of the bottom strand we did not observe such stalling on P63B* (Lane 10 in Figure 3B). Upon T5exo treatment of Cas1-2 bound P63T* complex, we noticed a shift in the nuclease stalling point to ~45 nt position from the labelled end (Lane 4 in Figure 3B). This maps the Cas1-2 binding position to be around 45 nt from the labelled end (P63T* in Figure 3A). Coincidentally, this binding position of Cas1-2 on P63T* is localized around a cognate PAM sequence (5’-AAG-3’ ranging from 47 nt to 49 nt upstream of labelled position) (P63T* in Figure 3A).

Prompted by this finding, we were interested in identifying the extent of protection that Cas1-2 could confer on the bottom strand upon its binding at PAM region. To accomplish this, we treated the Cas1-2 bound P63B* complex with T5exo. Here, the resulting length of the protected fragments upon exonuclease treatment indicated that Cas1-2 complex interaction can guard a region spanning ~45 nt from the labelled end in P63B* (Lane 12 in Figure 3B and P63B* in Figure 3A). As the PAM residues are positioned at 14 nt from the labelled end of the P63B*, the effective length of the protected prespacer from the PAM is ~30 nt. Overall, these results suggest a PAM dependent mechanism by which Cas1-2 could selectively acquire 33 nt prespacers.
Mechanism of prespacer capture and processing in E. coli

with T5exo. Here, we found extended smears and multiple bands upon employing P63mPAMT* (Figure 3A and Lane 8 in Figure 3B) and P63mPAMB* (Figure 3A and Lane 16 in Figure 3B). The varied length of resultant labelled fragments is indicative of numerous stalling points on these P63 mutants (Figure 3A) that occurred due to Cas1-2 binding. These results highlight that Cas1-2 gets specifically recruited towards PAM containing region that plays a crucial role in defining prespacer boundaries. Whereas, such specificity is lost when the DNA fragments lack PAM. Here the promiscuous interaction of Cas1-2 results in the generation of illicit prespacers that defies the productive length of the prespacers for integration (Lanes 8 and 16 in Figure 3B; P63mPAMT* and P63mPAMB* in Figure 3A).

Intrinsic specificity of Cas1-2 circumvents the requirement of Cas4 during PAM selection in E. coli

Having found the role of PAM mediated interaction in selecting the prespacers for uptake, we attempted to understand the intrinsic molecular principles that confer precision to Cas1-2 in PAM selectivity and prespacer scaling. Previous structural studies of CRISPR adaptation complex in E. coli suggested that the extended Cas1 C-terminal tail of apoCas1-2 complex (69) gets organized around the PAM residues upon binding to prespacer DNA (Figure 4) (42). In particular, Q287 and I291 residues of this proline-rich C-terminal tail make direct contacts with the nucleotides of PAM (Figure 4B), thus possibly imparting the PAM specificity. In another striking feature, a pair of Y22 residues that are derived from two different Cas1 protomers scales a 23 bp duplex region of prespacer via stacking interactions at either end (Figure 4B) (42,43). This gating mechanism at the Cas1-2 platform seems to scale the spacer length by facilitating the positioning of 3' overhang at the catalytic groove for integration (Figure 4B). To validate these observations, we employed Cas1-2 variants that encompass either deletion of Cas1 C-terminal tail (ΔC - ΔP279-S305) or a Cas1 point mutant Y22A (Figure S4A) and analysed their footprints upon binding to P63B* and P63T* (Figure 5A and S5D). As a control, we also used a Cas1 variant (5M - Q24H, P202Q, G241D, E276D and L297Q) (Figure 4B) that was previously shown to abrogate PAM selectivity (44).

Similar to Wild-type Cas1-2 (Wt), T5exo treatment of Y22A-P63B* generated a nuclease stalling point at ~45 nt albeit with reduced efficiency (Compare lanes 4 and 8 in Figure 5A). These findings indicate that Cas1 Y22A does not result in the altered PAM specificity (Wt and Y22A in Figure 5B). Despite having a stronger affinity for P63 (Y22A K_D = 1.781±0.202 µM vs Wt K_D = 2.842±0.372 µM) (Figure S5C and S2E), absence of Y22 mediated stacking interaction seems to reduce the prespacer protection ability of Y22A against nuclease action. Additionally, a shift in the nuclease stalling point to 60 nt from the labelled ends of P63B* was observed for ΔC and 5M variants (Lanes 8 and 10 in Figure 5A). Due to its low affinity, 5M seems to display reduced protection of prespacers from nuclease action (ΔC K_D = 2.675±0.282 µM vs 5M K_D = 17.08±3.428 µM) (Figure S5A and S5B). The shift in this nuclease stalling point indeed indicates that the ΔC and 5M variants display an impaired PAM specificity and were randomly interacting at the ends of P63B* (ΔC and 5M in Figure 5B). To fortify our observations, we sought to understand the impact of these mutations on the sequence composition of acquired spacers by expressing Cas1-2 variants in E. coli IYB5101. We observed that the mutations in ΔC and 5M have partly reduced the spacer incorporation efficacy of Cas1-2 (Compare Lanes 2, 4 and 6 in Figure 5C). Surprisingly, Y22A displayed a drastic reduction of spacer uptake in vivo (Compare Lanes 2 and 8 in Figure 5C), despite showing spacer integration in vitro (Figure S4B).

Expanded CRISPR arrays corresponding to the expression of each mutant were purified, and the sequences of newly incorporated spacers were derived from high-throughput sequencing. In line with previous studies, we
observed that the spacers originated from both genome and plasmid (44). Irrespective of the mutations in Cas1-2, the length of the incorporated spacers is strictly conserved (i.e., 33 nt) (Figure 5D). This finding suggests that Y22 mediated stacking interaction with prespacer or the Cas1 C-terminal restructuring is dispensable for the scaling of prespacers. These spacer sequences were mapped onto the plasmid and genome to identify the PAM. Despite the display of precise prespacer scaling by Cas1-2 variants, the specificity towards PAM region appears to be profoundly altered (Figure 5E). In concurrence with previous studies (30,44), we observed that most of the spacers acquired by Wt Cas1-2 encompass a conserved PAM region (5’-AAG-3’, where ‘G’ indicates +1 position of 33 nt spacer) (Figure 5E). In line with the nuclease protection assay (Figure 5A), we did not observe any preference towards the PAM region when we employed 5M or ΔC (Figure 5E). This finding bolsters the involvement of Cas1 C-terminal tail in PAM selectivity. Despite the reduced efficiency in spacer acquisition in vivo (Figure 5C), surprisingly, Y22A has displayed a remarkable precision for PAM selectivity, suggesting that this mutation bestowed high fidelity with respect to PAM recognition (Figure 5E).

Discussion

CRISPR system in E. coli (type I-E) displays precise scaling of prespacer length and stringent selectivity of PAM (30,41,44). Here, we attempted to uncover the elusive molecular events that drive the generation of competent substrates for homing at the leader-repeat junction by CRISPR adaptation complex. During naïve adaptation, prespacers are predominantly foraged from the DNA fragments generated by RecBCD during DNA repair (18) or by Cas3 during primed adaptation (35,36,70). Being helicase-nuclease enzymes, RecBCD and Cas3 action generates varied length of ssDNA fragments (39,71). Despite the association of ssDNA fragments with Cas1 in vivo (72), their integration at the CRISPR locus has been highly inefficient in comparison to duplex and partial-duplex DNA forms (Figure 1) (56). This points to a certitude that duplex formation by annealing of complementary ssDNA strands upon helicase-nuclease action could potentially generate prespacers that elicits CRISPR adaptation.

We demonstrate that irrespective of being a duplex or partial duplex the overall length of the prespacer must be ~33 nt (i.e., E. coli spacer length) for successful integration (Figure 1). A previous study has shown that an in vitro reconstituted Cas1-2 can solely process the prespacers with 10 nt 3’-overhangs to legitimate spacer size (42). In contrast, our experiments show the absence of such prespacer processing even with increasing Cas1-2 concentrations, despite the presence of PAM (P23[3'-10] in Figure S3A). We posit that these variations could be due to differences in the purification strategies of Cas1-2 complex in both studies. Here, we employed Cas1-2 that was generated as a complex in vivo and was further purified extensively, whereas, the previous study (42) utilized in vitro reconstituted Cas1-2 that potentially contained unassembled, free Cas1 protomers. The utilization of high concentrations of the in vitro reconstituted Cas1-2 and thus the increased presence of free Cas1, a known endonuclease (73), could have resulted in the cleavage of the prespacer overhangs (42). Alternatively, because the nuclease activity was only seen at µM concentration of in vitro reconstituted Cas1-2 (42), despite the fact that the integrase activity requires just nM concentration (17,56), it points to the possibility of trace cellular nuclease contamination.

Previous structural studies of E. coli Cas integrase complex reveals that 33 bp length of DNA can be exactly accommodated in between two active site regions of Cas1-2 (42,43). This hints at the fact that the Cas1-2 foothold can only mask 33 bp region and the rest is exposed to potential nuclease action. Mimicking such conditions, we incubated Cas1-2 bound longer DNA fragments (P63) with 3’→5’ (ExoIII) and 5’→3’ (T5exo) acting exonucleases. These reactions resulted in the generation of fragments that are in the range of cognate E. coli spacer size (Figure 2A). Furthermore,
these processed DNA fragments were readily utilized as substrates by Cas1-2 integrase (Figure 2B). Likewise, in B. halodurans (type I-C) and S. thermophilus DGCC7710 (type I-E), the nuclease action of Cas4 and DnaQ (an auxiliary domain of Cas2), respectively, on Cas1-2 bound DNA, generate integration competent prespacers (45,52). Intriguingly, as observed in E. coli, when Cas1-2 of these systems was presented with legitimately sized prespacers, the need for nuclease processing was obviated (45,52). This implies that Cas1-2 can solely catalyze spacer integration, but the generation of productive prespacers involves the action of an additional nuclease. The lack of a known prespacer processing enzyme such as Cas4 in E. coli led us to hypothesize that the trimming action could be complemented by other cellular nucleases. Unlike in other CRISPR variants (25,45-50), the productive pruning of prespacers by non-Cas nucleases is not sequence-specific (Figure 2). Moreover, a parallel study on prespacer generation in E. coli also independently demonstrated that the generic nucleases (such as DNA polymerase III or Exonuclease T) are sufficient for trimming the prespacers upon PAM recognition by Cas1-2 (74). These explain why the involvement of specific nucleases such as Cas4 is precluded for prespacer processing in E. coli (see below).

In addition to spacer length conservation, most prokaryotes display selective uptake of phage-origin prespacers that are bordered by a PAM (8). Recent in vivo studies in various type I organisms (I-A, I-D, I-G and I-U) have underscored the indispensable requirement of Cas4 in PAM selection as well as in prespacer processing (46-50). Moreover, in vitro studies performed with adaptation complex of B. halodurans (type I-C) and S. solfataricus (type I-A) revealed that Cas4 nuclease avoids the processing of free DNA ends that are devoid of PAM sequence (25,45). This preferential activity of Cas4 seems to act as a critical checkpoint in ensuring the productive uptake of infection memory by Cas1-2 in the hosts.

A previous study in E. coli showed that upon expression of Cas1-2, 33 nt prespacer bordered by PAM originated from the longer electroporated DNA (P63) (44). Interestingly, we observed the protection of the same region by Cas1-2 when we performed a nuclease footprinting assay on P63 (Figure 3). These experiments highlight that the Cas1-2 complex alone is sufficient to recognize PAM in E. coli. The footprinting experiments also demonstrate binding of substrates at multiple points when PAM residues of P63 were mutated (Figure 3). These non-specific interactions of Cas1-2 could generate a heterogeneous population of protected prespacers. This explains how the adaptation complex of various type I organisms infrequently uptake prespacers with erroneous PAM (Wt in Figure 5E) (35,75-80).

Structural analysis of Cas1-2-prespacer complex highlights the features that could lead to precise scaling and PAM selection of prespacers. A platform formed by the interaction of a Cas2 dimer with two Cas1 dimers on either side houses the 23 bp duplex region of prespacer (42,43). Stationed at either end of this duplex is the aromatic ring of Y22 residue that stacks the prespacer at the border of the Cas1 catalytic groove and directs the 3’-overhang to position its 5th nt at the catalytic site. Thus, Y22 guided meticulous placement of DNA substrate seems to dictate the length of prespacer (42,43). Furthermore, the flexible C-terminal tail of Cas1 is moulded around the PAM region. The absence of such molecular architecture upon mutating the PAM hints at the role of C-terminal tail in PAM recognition (42). Deployment of Cas1-2 variants that encompass either deletion of Cas1 C-terminal tail (ΔC) or Y22A in spacer integration assays helped to unveil the role of these structural entities in determining the PAM selection (Figure 5). As shown here, the deletion of C-terminal tail resulted in impaired PAM recognition (Figure 5A) and led to an uptake of prespacers that were lacking PAM (ΔC in Figure 5E).

A comparison of the structures of Cas1 from various type I organisms (Figure 6A-D) revealed a striking contrast between Cas1 C-terminal tail of type I-E and other subtypes. The C-terminal tail of E. coli Cas1 is...
**Mechanism of prespacer capture and processing in E. coli**

noticeably longer with 31 amino acid (aa) than the shorter 12 aa tails of *Archeoglobus fulgidus*, *Pyrococcus horikoshii* and *B. halodurans* (Figure 6) (42,81,82). Additional comparative analysis reinforced these observations that type I-E encompasses the longest C-terminal tail with an average length of 29 aa (Figure 6E, S6-9). Coincidentally, CRISPR-Cas subtypes with shorter Cas1 C-terminal tails such as type I-A, I-B and I-C encompass Cas4 (15) and previous studies suggest the indispensability of Cas4 in promoting PAM specificity (25,45-47). In contrast to these systems, it appears that the extended C-terminal tail of Cas1 in *E. coli* (type I-E) compensates for the lack of Cas4 by guiding the PAM selection.

In case of Y22A variant, though the prespacer integration is unaffected *in vitro* (Figure S4B), in line with the previous observations (43), a rampant decrease in spacer integration efficiency was observed *in vivo* (Figure 5C). Y22A conferred reduced prespacer protection against the nucleases in comparison to Wt (Figure 5A). This observation highlights the critical role of Y22 residue in providing the Wt with a better grip on bound DNA (Figure 5A). As Y22A could lack such interactions with its substrates, nucleases might seamlessly dislodge it from the bound prespacer (Figure 5A). This action appears to limit the substrate availability and impede spacer integration *in vivo* (Figure 5C). Despite the reduction in spacer acquisition potential, Y22A showed high fidelity towards “AAG” PAM selection than Wt (Figure 5E). As discussed above, Y22A displays a reduced grip on the prespacers during nuclease mediated processing (Figure 5A). This weak prespacer binding could be further disrupted in the absence of PAM due to the loss of interactions with Cas1 C-terminal tail. Therefore, only the presence of cognate PAM (AAG) is likely to allow Y22A to retain the hold on DNA during prespacer processing leading to selective enrichment (Figure 5E). The strategic position of Cas1 Y22 in adaptation complex appears to have a key role in defining the prespacer length (Figure 4B). Interestingly, our experiments with Y22A resulted in the uptake of prespacers that were predominantly 33 bp in length (Figure 5D). These findings negate the involvement of Y22 stacking interactions in deciding the prespacer boundary. Recent studies in type V-C demonstrated that a mini integrase complex constituted by Cas1 tetramer prefers short (18 bp) spacers (83), likewise in *E. coli*, Cas1-2 structural framework alone appears to be a critical parameter in gauging the length of spacers (42,43).

Our work in conjunction with the previous studies allows us to propose an updated model for prespacer capture in type I systems (Figure 7). During CRISPR adaptation, the dispensability of sequence-specific auxiliary nucleases such as Cas4 seems to be contingent on the type of molecular players that are involved in PAM selection. Though Cas1-2 integrase catalyzes the prespacer homing, in majority of type I systems, PAM selection and prespacer processing require Cas4 (25,45-47). In contrast, in type I-E system, the intrinsic affinity of Cas1-2 integrase alone is sufficient to recognize cognate PAM. This lineage specific remarkable adaptation of Cas1-2/I-E offsets the requirement of PAM specifying Cas4 nuclease. Instead, generic cellular non-Cas nucleases are co-opted to trim the exposed DNA ends of Cas1-2-prespacer complex for generating the legitimate prespacers for integration.

**Experimental procedures**

**Construction of plasmids**

Lists of plasmids, strains and oligonucleotides used in this study are detailed in the Supporting tables S1-S3.

Genes encoding IHFα, IHFβ, Cas1 and Cas2 were amplified using *E. coli* K-12 MG1655 genomic DNA as template. To generate p1R-IHFαβ, a bicistronic cassette encoding IHFα and IHFβ was amplified and inserted at the SspI site of p1R. p13SR-Cas1 was generated by inserting an amplicon encoding Cas1 at SspI site of p13SR and pMS-Cas2 was created by introducing an amplicon encoding Cas2 between BamHI/HindIII sites of pMS (84), respectively. Bicistronic cassettes

---

*Note:* The page number 9 is indicated at the bottom of the page.
Mechanism of prespacer capture and processing in E. coli

(cas1-cas2) expressing 6X Histidine tagged Wt and 5M Cas1 were amplified using pCas1-2[K] (85) and pMut89 (44) as templates, respectively. Amplified fragments encoding Wt and 5M Cas1-2 were inserted between NcoI/NotI sites of pCas1-2[K] to generate pCas1-2H and p5M, respectively. PCR based mutagenesis was used to generate pY22A and pΔC that express Y22A and ΔC variants of Cas1, respectively. All constructs were verified by Sanger sequencing.

Expression and purification of proteins

IHF and Cas1 were purified as described before (17). In order to purify Cas2, E. coli BL21(DE3) harbouring pMS-Cas2 was grown in auto-induction LB media supplemented with 100 µg/ml kanamycin at 37°C, 180 rpm. Upon reaching 0.6 OD_{600}, temperature was shifted to 16°C and thereafter the growth and induction was continued for 16 hours. Subsequently, cells were harvested and washed 2X times with Buffer 1A (20 mM HEPES–NaOH pH 7.4, 500 mM KCl and 10% glycerol). Bacterial pellet was resuspended in Buffer 1A containing 1 mM PMSF and the cells were lysed by sonication. Here, Cas2 encompasses 6X Histidine tagged MBP-SUMO as an N-terminal fusion and a Strep-II tag on the C-terminal end. The clarified fraction of the lysate was applied to a 5 ml MBPTrap HP column (GE healthcare) and was followed by a washing step with Buffer 1A. Thereafter, the bound proteins were eluted with Buffer 1A containing 10 mM maltose. Eluted fractions were mixed with SUMO protease (Ulp1_{403-621}) (in 400:1 ratio of His-MBP-SUMO-Cas2-strep: Ulp1_{403-621}) (84) and the incubation was continued for 60 min at 25°C. Following this, the mixture was loaded onto 5 ml HiTrap IMAC HP column (GE Healthcare) 5X times to facilitate binding of Histidine tagged MBP-SUMO-Cas2-strep, MBP-SUMO and Ulp1_{403-621}. Column flow-through containing Cas2-strep was concentrated using a centrifugal membrane filter (Sartorius). To remove any trace protein contaminants, the concentrated sample was loaded onto HiLoad Superdex 200 pg gel filtration column (GE healthcare), that is pre-equilibrated with Buffer 1B (20 mM HEPES–NaOH pH 7.4, 150 mM KCl and 10% glycerol). Eluted fractions containing Cas2-strep were pooled, concentrated, snap frozen in liquid nitrogen and stored at −80°C until required.

Integrate complex comprising of untagged Cas1 and C-terminal 6X Histidine tagged Cas2 was expressed and purified as described before (61) with minor modifications. Here, E. coli BL21(DE3) transformed using pCas1-2H was grown in 2XYT broth supplemented with 100 µg/ml spectinomycin at 37°C, 180 rpm till 0.6 OD_{600}. Thereafter, the protein expression was induced by addition of 0.7 mM IPTG and the growth was continued at 25°C for 24 hours. Simultaneously, cells were harvested and washed 2X times with Buffer 2A (20 mM HEPES–NaOH pH 7.4, 150 mM KCl, 10% glycerol and 30 mM imidazole). The pellet was resuspended in Buffer 2A containing 1 mM PMSF and cells were lysed by sonication. Thereafter, the lysate was clarified and loaded onto a 5 ml HiTrap IMAC HP column (GE Healthcare) and was followed by a washing step with Buffer 2A. A linear gradient of imidazole (0.03–0.5 M) in Buffer 2A was applied to elute the proteins that were bound to the column resin. The purified fractions that contain the complex of Cas1-2 were pooled and concentrated using a centrifugal membrane filter (Sartorius). To remove trace protein contaminants and un-complexed Cas2, the concentrate was further purified using HiLoad Superdex 200 pg gel filtration column (GE healthcare) that is pre-equilibrated with Buffer 2B (20 mM HEPES–NaOH pH 7.4, 150 mM KCl and 10% glycerol). Eluted fractions containing Cas1-2 integrase were pooled, concentrated, snap frozen in liquid nitrogen and stored at −80°C until required. Similar procedure was implemented to purify 5M, ΔC and Y22A Cas1 variants of Cas1-2 from the IPTG induced E. coli BL21(DE3) cells that harbours p5M, pΔC and pY22A, respectively.
**Mechanism of prespacer capture and processing in E. coli**

**In vitro integration assay**

177 bp CRISPR DNA substrate that encompasses 69 bp leader and two repeat-spacer units of CRISPR 2.1 locus of *E. coli* was amplified using pCSIR-T (85) as a template. CRISPR DNA substrates labelled with 5’-FAM at top strand of the leader end (CD-T*) or at bottom strand of second spacer end (CD-B*) were prepared using PCR. To generate various prespacers (P33, P23[3’-5], P23[5’-5], P23[3’-10], P63, P63mPAM and their 5’-FAM labelled variants), respective oligonucleotides (Table S3) were mixed in a buffer containing 10 mM Tris-Cl pH 8.5. These mixtures were heated to 95°C and gradually allowed to cool to room temperature in order to facilitate the formation of duplex and partial-duplex prespacers. In the case of P33ss, a 33nt long single stranded oligonucleotide was used as a prespacer.

The *in vitro* integration assays were performed as previously described (17) with minor modifications. Briefly, 210 nM of Cas1 or Cas2 or Cas1-2 (Wt, ΔC, Y22A and 5M) was mixed with 550 nM of desired prespacer and incubated at room temperature for 5 minutes. To this mixture, 0.5 µM of IHF and 21 nM of CRISPR DNA substrate were supplemented and incubation was continued at 37°C for 60 min in integrase buffer (20 mM HEPES–NaOH pH 7.4, 25 mM KCl, 10 mM MgCl₂ and 1 mM DTT). Subsequently, the reaction mixtures were supplemented with equal volume of stopping solution (95% Formamide, 5 mM EDTA and 0.025% SDS) followed by heating at 95°C for 20 min. These samples were loaded onto pre-heated 12% denaturing acrylamide gels that were maintained at 50°C and electrophoresed in 1X TBE. Subsequently, gels were stained with EtBr and visualized using a gel documentation system (Bio-Rad). Whereas, in the assays that involve FAM labelled CRISPR DNA or prespacers, gels were imaged without any post-staining step. All the integration experiments were independently repeated at least twice and the representative gel pictures were shown.

**Electrophoretic mobility shift assays**

The binding of Cas1-2 with various prespacers was monitored using electrophoretic mobility shift assays. Here, 100 nM of desired 5’-FAM labelled prespacers (P23(3’-5), P23(5’-5), P33 and P23(3’-10)) were incubated with increasing concentration of Wt Cas1-2 (0.1, 0.15, 0.2, 0.25, 0.45, 0.6, 0.8, 1, 1.5, 2 and 3 µM) in prespacer binding buffer (20 mM HEPES–NaOH pH 7.4, 125 mM KCl, 10 mM MgCl₂ and 1 mM DTT) for 30 min at 37°C. Subsequently, all the samples were directly loaded on 0.8% agarose gel and electrophoresed in 1X TAE at 4°C. Bound fraction for each sample in the gel was estimated by quantifying the amount of DNA at each band using densitometric analysis (Bound fraction of prespacer (%) at X µM Cas1-2 = [(Amount of DNA in absence of Cas1-2 – Amount of unbound DNA at X µM Cas1-2) / (Amount of DNA in absence of Cas1-2)]*100. To estimate dissociation constants (K_D), the resulting plots of bound fraction (%) against Cas1-2 concentration was fitted to a non-linear equation: y = Bmax*x/(KD + x) (where x, y, Bmax and K_D represents Cas1-2 concentration (µM), bound fraction (%), maximum concentration of Cas1-2 bound to prespacer and dissociation constant, respectively). In EMSA that involves 5’-FAM labelled P63 or P63mPAM prespacers, 100 nM of DNA was incubated with 0.2 to 5 µM of Wt / ΔC / Y22A or 0.2 to 20 µM of 5M Cas1-2 variants (Figure S2E, S2F, S5A, S5B and S5C). All the binding experiments were independently repeated thrice and the representative gel pictures were displayed.

To further verify the formation of Cas1-2 nucleoprotein complex, the release of prespacers was monitored upon Proteinase-K treatment of Cas1-2 in each assay. To achieve this, an aliquot of sample containing prespacer and 3 µM Cas1-2 was mixed with 1 mg/ml Proteinase-K and incubated at 37°C for 15 min.

**Exonuclease treatment of Cas1-2 bound DNA fragments**

Exonuclease treatment was performed to identify the extent of protection conferred by
Mechanism of prespacer capture and processing in E. coli

binding of Cas1-2 on to a long DNA fragment. 40 µl of 0.5 µM P63 and 6 µM of either Cas1 or Cas2 or Cas1-2 in prespacer binding buffer were incubated at 37°C for 45 min. Subsequently, 20 µl aliquots of these samples were supplemented with 3 units of either T5 exonuclease (NEB) or Exonuclease III (NEB), or 3 units of the mixture containing both exonucleases and incubation was continued for 60 min at 37°C. Thereafter, all the samples were mixed with equal volume of denaturation buffer that contains 200 mM Tris–Cl pH 8.3, 200 mM Boric acid, 20 mM EDTA, 0.05 % SDS and 8 M Urea followed by heating at 95°C for 15 min. These samples were loaded onto pre-heated 20% denaturing acrylamide gels that were maintained at 50°C and electrophoresed in 1X TBE. Subsequently, gels were stained with EtBr and visualized using a gel documentation system. This experiment was independently repeated thrice and the representative gel picture was shown.

Exonuclease footprinting

T5 exonuclease mediated footprinting was performed to identify the interaction boundaries of Cas1-2 on longer prespacer DNA fragments. Here, 40 µl of 0.5 µM of desired fluorescein labelled P63 variant (P63T*, P63B*, P63mPAMT* and P63mPAMB*) was mixed with 6 µM of Wt or one of the mutant variants of Cas1-2 (Y22A, ΔC and 5M) in prespacer binding buffer and incubated for 45 min at 37°C. Subsequently, 20 µl aliquots of these samples were supplemented with 3 units of T5 exonuclease and incubation was continued for 60 min at 37°C. Thereafter, all the samples were mixed with equal volume of denaturation buffer that contains 200 mM Tris–Cl pH 8.3, 200 mM Boric acid, 20 mM EDTA, 0.05 % SDS and 8 M Urea followed by heating at 95°C for 15 min. These samples were loaded onto pre-heated 20% denaturing acrylamide gels that were maintained at 50°C and electrophoresed in 1X TBE. Subsequently, gels were directly visualized using a gel documentation system. All the footprinting experiments were repeated at least twice and the representative gel pictures were shown.

Spacer acquisition assays

The in vivo spacer acquisition assays were performed as described previously (10,17) with minor modifications. After transformation using plasmids (pCas1-2H, p5M, pY22A and pΔC), E. coli IYB5101 that expresses Wt or a mutant of Cas1-2 was subjected to three cycles of growth and induction in LB media supplemented with 100 µg/ml spectinomycin, 0.2% L-arabinose and 0.1 mM IPTG for 16 hours at 37°C. After each cycle, cultures were diluted to 1:300 times with fresh LB media containing aforementioned supplements and the growth was continued for 16 hours. Thereafter, genomic DNA was isolated according to manufacturer’s protocol (HiPurA bacterial genomic DNA purification kit, Himedia) and this was used as a template for PCR to monitor the spacer integration at CRISPR 2.1. All the PCR amplified samples were resolved on 1.5% agarose gels to identify the DNA bands corresponding to parental and expanded arrays (parental array + n x 61 bp), where n is a positive integer. DNA quantities corresponding to parental and expanded array were quantified by densitometric analysis. Utilizing these values, percentage of spacer integration for each Cas1-2 variant was estimated (% integration = [(Amount of expanded array) / (Amount of parental array + Amount of expanded array)]*100.

High-throughput sequencing and analysis

In order to understand the effect of Cas1-2 mutants on prespacer scaling and PAM selectivity, high-throughput sequencing was performed to derive the sequences of newly incorporated prespacers. Expanded CRISPR arrays corresponding to the expression of each Cas1-2 variant were extracted from the agarose gels (QIAquick Gel Extraction Kit, Qiagen). Approximately 200 ng of each PCR product was further purified using HighPrep magnetic beads (MAGBIO). These purified samples were subjected to DNA end repair and adaptor ligation using Illumina-compatible NEXTflex Rapid DNA sequencing kit (BIOO Scientific,
Mechanism of prespacer capture and processing in E. coli

13

Austin, Texas, U.S.A.). Subsequently, the ligated DNA products were purified with HighPrep magnetic beads and further enrichment was achieved by 8 cycles of PCR with Illumina-compatible primers (NEXTFlex DNA sequencing kit). These amplicons were subjected to an additional step of purification with HighPrep magnetic beads and were sequenced on a Miseq 300 paired end platform.

The paired-end reads were subjected to several pre-processing steps as described below. Firstly, both F and R reads with Phred score less than 20 were removed by utilizing fastq_quality_trimmer from the FASTX-toolkit-version-0.0.13. The remaining F and R reads were trimmed in paired-end mode to remove F [5' - AGATCGGAAGAGCACACGTCTGAACTC CAGTCA-3'] and R [5' - AGATCGGAAGAGCGTCGTGTAGGGAA AGAGTGT-3'] adapter sequences using Cutadapt-1.18 (86). Following these, the leader proximal spacer sequence (S0) was selectively retrieved in FASTA format. These S0 sequences that were derived from E. coli expressing WT and mutants of Cas1-2 were searched against plasmid (pCas1-2[K]) (85) and E. coli K-12 MG1655 genome (GenBank assembly accession: GCA_000005845.2), respectively, using BLASTN (87). From the BLAST hits, we identified the location of spacer sequences on plasmid and E. coli K-12 MG1655 genome, respectively and extracted the triplet sequence corresponding to PAM. The conservation of PAM was analysed using WEB-LOGO (88). For all sequence manipulations including the extraction of S0 and PAM sequences, we employed custom written python codes utilising the Biopython library (89).

Analysis of Cas1 C-terminal tail across CRISPR-Cas type I systems

Locus ID corresponding to Cas1 genes of type I-A, I-B, I-C and I-E were derived from the previous study (90). Utilizing these identifiers, we extracted Cas1 protein sequences of 36 type I-A, 150 type I-B, 129 type I-C and 116 type I-E organisms from NCBI protein database. Subsequent to this, we performed multiple sequence alignments in T-COFFEE web server (91) for Cas1 proteins of each subtype separately (Supporting information files F1, F2, F3 and F4). Utilizing Cas1 crystal structures of A. fulgidus (type I-A; PDB ID: 4N06), P. horikoshii (type I-B; PDB ID: 4WJ0) and E. coli (type I-E; PDB ID: 5DQZ) as a reference, C-terminal tail residues of each Cas1 was extracted from their multiple sequence alignments. Owing to the absence of type I-C Cas1 crystal structure in PDB, Cas1 (BH0341) structure of B. halodurans was predicted by I-TASSER web server (82) using structures corresponding to PDB ID: 3LFX, 4N06 and 2YZS as threading templates. I-TASSER predicted five different models for B. halodurans Cas1. Among these, the model with highest confidence score (C-score = 1.25) was used as a structural reference for predicting the C-terminal tail residues from multiple sequence alignment of 129 type I-C Cas1 proteins.

Utilizing ESPript server (92), various secondary structural element positions were mapped on to the multiple sequence alignments of type I-A, I-B, I-C and I-E Cas1 proteins (Figure S6-S9). All the protein structural representations were generated using ChimeraX (93).
Acknowledgements

This work was supported by grants from Department of Biotechnology (DBT) [BT/PR15925/NER/95/141/2015 and BT/08/IYBA/2014/05]; and Science and Engineering Research Board (SERB) [YSS/2014/000286]. *E. coli* IYB5101 strain was a kind gift from Dr. Udi Qimron, Tel Aviv University, Israel; Vectors pET StreptII TEV LIC cloning vector (p1R) (Addgene #29664) and pET StreptII TEV co-transformation cloning vector (p13SR) (Addgene #48328) were a kind gift from Dr. Scott Gradia, QB3 MacroLab, University of California, USA; pLJSRSF7 (pMS) (Addgene #64693) and pFGET19_Ulp1(Addgene #6497) were a kind gift from Dr. Hideo Iwai, University of Helsinki, Finland; pWUR 1+2 tetO mut89 (pMut89) (Addgene #80102) was a kind gift from Dr. George M. Church, Harvard University, USA; pCSIR-T and pCas1-2[K] were a kind gift from Dr. F.J.M. Mojica, University of Alicante, Spain. We sincerely acknowledge the gracious gesture of the aforementioned scientists for sharing their bacterial strains and plasmids. We also thank all the members of MAB lab for their suggestions on the experiments and critical comments on the manuscript.

Conflict of interest

The authors declare no competing interests.

References

1. Marraffini, L. A., and Sontheimer, E. J. (2008) CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* **322**, 1843-1845
2. Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A., and Horvath, P. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709-1712
3. Hille, F., Richter, H., Wong, S. P., Bratovič, M., Ressel, S., and Charpentier, E. (2018) The Biology of CRISPR-Cas: Backward and Forward. *Cell* **172**, 1239-1259
4. Deveau, H., Barrangou, R., Garneau, J. E., Labonte, J., Fremaux, C., Boyaval, P., Romero, D. A., Horvath, P., and Moineau, S. (2008) Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *J. Bacteriol.* **190**, 1390-1400
5. Kunin, V., Sorek, R., and Hugenholtz, P. (2007) Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol.* **8**, R61
6. Bolotin, A., Quinquis, B., Sorokin, A., and Ehrlich, S. D. (2005) Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551-2561
7. Mojica, F. J., Diez-Villasenor, C., Garcia-Martinez, J., and Soria, E. (2005) Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol.* **60**, 174-182
8. McGinn, J., and Marraffini, L. A. (2019) Molecular mechanisms of CRISPR-Cas spacer acquisition. *Nat. Rev. Microbiol.* **17**, 7-12
9. Jackson, S. A., McKenzie, R. E., Fagerlund, R. D., Kieper, S. N., Fineran, P. C., and Brouns, S. J. (2017) CRISPR-Cas: Adapting to change. *Science* **356**, eaal5056
10. Yosef, I., Goren, M. G., and Qimron, U. (2012) Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic Acids Res.* **40**, 5569-5576
11. Hochstrasser, M. L., and Doudna, J. A. (2015) Cutting it close: CRISPR-associated endoribonuclease structure and function. *Trends Biochem. Sci.* **40**, 58-66
Mechanism of prespacer capture and processing in E. coli

12. Punetha, A., Yoganand, K. N. R., Nimkar, S., and Anand, B. (2018) Cutting it right: Plasticity and strategy of CRISPR RNA specific nucleases. *Proc. Indian Natl. Sci. Acad. B. Biol. Sci.* **84**, 455-477

13. Brouns, S. J., Jore, M. M., Lundgren, M., Westra, E. R., Slijkhuis, R. J., Snijders, A. P., Dickman, M. J., Makarova, K. S., Koonin, E. V., and van der Oost, J. (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960-964

14. Carte, J., Wang, R., Li, H., Terns, R. M., and Terns, M. P. (2008) Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes Dev.* **22**, 3489-3496

15. Makarova, K. S., Wolf, Y. I., and Koonin, E. V. (2018) Classification and Nomenclature of CRISPR-Cas Systems: Where from Here? *CRISPR J.* **1**, 325-336

16. Nunez, J. K., Bai, L., Harrington, L. B., Hinder, T. L., and Doudna, J. A. (2016) CRISPR Immunological Memory Requires a Host Factor for Specificity. *Mol. Cell* **62**, 824-833

17. Yoganand, K. N. R., Sivathanu, R., Nimkar, S., and Anand, B. (2017) Asymmetric positioning of Cas1–2 complex and Integration Host Factor induced DNA bending guide the unidirectional homing of protospacer in CRISPR-Cas type I-E system. *Nucleic Acids Res.* **45**, 367-381

18. Levy, A., Goren, M. G., Yosef, I., Auster, O., Manor, M., Amitai, G., Edgar, R., Qimron, U., and Sorek, R. (2015) CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* **520**, 505-510

19. Pul, U., Wurm, R., Arslan, Z., Geissen, R., Hofmann, N., and Wagner, R. (2010) Identification and characterization of E. coli CRISPR-cas promoters and their silencing by H-NS. *Mol. Microbiol.* **75**, 1495-1512

20. Mohr, G., Silas, S., Stamos, J. L., Makarova, K. S., Markham, L. M., Yao, J., Lucas-Elio, P., Sanchez-Amat, A., Fire, A. Z., Koonin, E. V., and Lambowitz, A. M. (2018) A Reverse Transcriptase-Cas1 Fusion Protein Contains a Cas6 Domain Required for Both CRISPR RNA Biogenesis and RNA Spacer Acquisition. *Mol. Cell* **72**, 700-714.e8

21. Silas, S., Mohr, G., Sidote, D. J., Markham, L. M., Sanchez-Amat, A., Bhaya, D., Lambowitz, A. M., and Fire, A. Z. (2016) Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. *Science* **351**, aad4234

22. Ivancic-Bace, I., Cass, S. D., Wearne, S. J., and Bolt, E. L. (2015) Different genome stability proteins underpin primed and naive adaptation in E. coli CRISPR-Cas immunity. *Nucleic Acids Res.* **43**, 10821-10830

23. Wei, Y., Terns, R. M., and Terns, M. P. (2015) Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. *Genes Dev.* **29**, 356-361

24. Li, M., Wang, R., Zhao, D., and Xiang, H. (2014) Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process. *Nucleic Acids Res.* **42**, 2483-2492

25. Rollie, C., Graham, S., Rouillon, C., and White, M. F. (2018) Prespacer processing and specific integration in a Type I-A CRISPR system. *Nucleic Acids Res.* **46**, 1007-1020

26. Fagerlund, R. D., Wilkinson, M. E., Klykov, O., Barendrecht, A., Pearce, F. G., Kieper, S. N., Maxwell, H. W. R., Capolupo, A., Heck, A. J. R., Krause, K. L., Bostina, M., Scheltema, R. A., Staals, R. H. J., and Fineran, P. C. (2017) Spacer capture and integration by a type I-F Cas1-Cas2-3 CRISPR adaptation complex. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E5122-E5128

27. Westra, E. R., Pul, U., Heidrich, N., Jore, M. M., Lundgren, M., Stratmann, T., Wurm, R., Raine, A., Mescher, M., Van Heereveld, L., Mastop, M., Wagner, E. G., Schnetz, K., Van Der Oost, J., Wagner, R., and Brouns, S. J. (2010) H-NS-mediated repression of CRISPR-based immunity in Escherichia coli K12 can be relieved by the transcription activator LeuO. *Mol. Microbiol.* **77**, 1380-1393

28. Ka, D., Jang, D. M., Han, B. W., and Bae, E. (2018) Molecular organization of the type II-A CRISPR adaptation module and its interaction with Cas9 via Csn2. *Nucleic Acids Res.* **46**, 9805-9815

29. Heler, R., Samai, P., Modell, J. W., Weiner, C., Goldberg, G. W., Bikard, D., and Marraffini, L. A. (2015) Cas9 specifies functional viral targets during CRISPR-Cas adaptation. *Nature* **519**, 199-202
Mechanism of prespacer capture and processing in E. coli

30. Yosef, I., Shitrit, D., Goren, M. G., Burstein, D., Pupko, T., and Qimron, U. (2013) DNA motifs determining the efficiency of adaptation into the Escherichia coli CRISPR array. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 14396-14401

31. Zetsche, B., Gootenberg, J. S., Abudayyeh, O. O., Slaymaker, I. M., Makarova, K. S., Essletzbichler, P., Volz, S. E., Joung, J., van der Oost, J., Regev, A., Koonin, E. V., and Zhang, F. (2015) Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* **163**, 759-771

32. Gleditzsch, D., Pausch, P., Müller-Esparza, H., Özcan, A., Guo, X., Bange, G., and Randau, L. (2019) PAM identification by CRISPR-Cas effector complexes: diversified mechanisms and structures. *RNA Biol.* **16**, 504-517

33. Mojica, F. J., Diez-Villasenor, C., Garcia-Martinez, J., and Almendros, C. (2009) Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733-740

34. Sashital, D. G., Wiedenheft, B., and Doudna, J. A. (2012) Mechanism of Foreign DNA Selection in a Bacterial Adaptive Immune System. *Mol. Cell* **46**, 606-615

35. Datsenko, K. A., Pougach, K., Tikhonov, A., Wanner, B. L., Severinov, K., and Semenova, E. (2012) Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat. Commun.* **3**, 945

36. Semenova, E., Savitskaya, E., Musharova, O., Strotskaya, A., Vorontsova, D., Datsenko, K. A., Logacheva, M. D., and Severinov, K. (2016) Highly efficient primed spacer acquisition from targets destroyed by the Escherichia coli type I-E CRISPR-Cas interfering complex. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 7626-7631

37. Li, M., Wang, R., and Xiang, H. (2014) Haloarcula hispanica CRISPR authenticates PAM of a target sequence to prime discriminative adaptation. *Nucleic Acids Res.* **42**, 7226-7235

38. Musharova, O., Sitnik, V., Vlot, M., Savitskaya, E., Datsenko, K. A., Krivoy, A., Fedorov, I., Semenova, E., Brouns, S. J. J., and Severinov, K. (2019) Systematic analysis of Type I-E Escherichia coli CRISPR-Cas PAM sequences ability to promote interference and primed adaptation. *Mol. Microbiol.* **111**, 1558-1570

39. Dillingham, M. S., and Kowalczykowski, S. C. (2008) RecBCD enzyme and the repair of double-stranded DNA breaks. *Microbiol. Mol. Biol. Rev.* **72**, 642-671

40. Muskavitch, K. M., and Linn, S. (1982) A unified mechanism for the nuclease and unwinding activities of the recBC enzyme of Escherichia coli. *J. Biol. Chem.* **257**, 2641-2648

41. Swarts, D. C., Mosterd, C., van Passel, M. W., and Brouns, S. J. J. (2012) CRISPR interference directs strand specific spacer acquisition. *PLoS One* **7**, e35888

42. Wang, J., Li, J., Zhao, H., Sheng, G., Wang, M., Yin, M., and Wang, Y. (2015) Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems. *Cell* **163**, 840-853

43. Nunez, J. K., Harrington, L. B., Kranzusch, P. J., Engelman, A. N., and Doudna, J. A. (2015) Foreign DNA capture during CRISPR-Cas adaptive immunity. *Nature* **527**, 535-538

44. Shipman, S. L., Nivala, J., Macklis, J. D., and Church, G. M. (2016) Molecular recordings by directed CRISPR spacer acquisition. *Science* **353**, aaf1175

45. Lee, H., Zhou, Y., Taylor, D. W., and Sashital, D. G. (2018) Cas4-Dependent Prespacer Processing Ensures High-Fidelity Programming of CRISPR Arrays. *Mol. Cell* **70**, 48-59.e5

46. Kieper, S. N., Almendros, C., Behler, J., McKenzie, R. E., Nobrega, F. L., Haagsma, A. C., Vink, J. N. A., Hess, W. R., and Brouns, S. J. J. (2018) Cas4 Facilitates PAM-Compatible Spacer Selection during CRISPR Adaptation. *Cell Rep.* **22**, 3377-3384

47. Shiimori, M., Garrett, S. C., Graveley, B. R., and Terns, M. P. (2018) Cas4 Nucleases Define the PAM, Length, and Orientation of DNA Fragments Integrated at CRISPR Loci. *Mol. Cell* **70**, 814-824.e6
Mechanism of prespacer capture and processing in E. coli

48. Liu, T., Liu, Z., Ye, Q., Pan, S., Wang, X., Li, Y., Peng, W., Liang, Y., She, Q., and Peng, N. (2017) Coupling transcriptional activation of CRISPR-Cas system and DNA repair genes by Csa3a in Sulfolobus islandicus. *Nucleic Acids Res.* **45**, 8978-8992.

49. Zhang, Z., Pan, S., Liu, T., Li, Y., and Peng, N. (2019) Cas4 nuclease can effect specific integration of CRISPR spacers. *J. Bacteriol.* **201**, e00747-18.

50. Almendros, C., Nobrega, F. L., McKenzie, R. E., and Brouns, S. J. J. (2019) Cas4–Cas1 fusions drive efficient PAM selection and control CRISPR adaptation. *Nucleic Acids Res.* **47**, 5223–5230.

51. Lee, H., Dhingra, Y., and Sashital, D. G. (2019) The Cas4-Cas1-Cas2 complex mediates precise prespacer processing during CRISPR adaptation. *Elife* **8**, e44248.

52. Drabavicius, G., Sinkunas, T., Silanskas, A., Gasiunas, G., Venclovas, Č., and Siksnys, V. (2018) DnaQ exonuclease-like domain of Cas2 promotes spacer integration in a type I-E CRISPR-Cas system. *EMBO Rep.* **19**, e45543.

53. Radovcic, M., Killelea, T., Savitskaya, E., Wettstein, L., Bolt, E. L., and Ivancic-Bace, I. (2018) CRISPR-Cas adaptation in Escherichia coli requires RecBCD helicase but not nuclease activity, is independent of homologous recombination, and is antagonized by 5′ ssDNA exonucleases. *Nucleic Acids Res.* **46**, 10173-10183.

54. McGinn, J., and Marraffini, L. A. (2016) CRISPR-Cas Systems Optimize Their Immune Response by Specifying the Site of Spacer Integration. *Mol. Cell* **64**, 616-623.

55. Rollie, C., Schneider, S., Brinkmann, A. S., Bolt, E. L., and White, M. F. (2015) Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition. *Elife* **4**, e08716.

56. Nunez, J. K., Lee, A. S., Engelman, A., and Doudna, J. A. (2015) Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature* **519**, 193-198.

57. Wei, Y., Chesne, M. T., Terns, R. M., and Terns, M. P. (2015) Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in Streptococcus thermophilus. *Nucleic Acids Res.* **43**, 1749-1758.

58. Wang, R., Li, M., Gong, L., Hu, S., and Xiang, H. (2016) DNA motifs determining the accuracy of repeat duplication during CRISPR adaptation in Haloarcula hispanica. *Nucleic Acids Res.* **44**, 4266-4277.

59. Goren, M. G., Doron, S., Globus, R., Amitai, G., Sorek, R., and Qimron, U. (2016) Repeat Size Determination by Two Molecular Rulers in the Type I-E CRISPR Array. *Cell Rep.* **16**, 2811-2818.

60. Arslan, Z., Hermanns, V., Wurm, R., Wagner, R., and Pul, U. (2014) Detection and characterization of spacer integration intermediates in type I-E CRISPR-Cas system. *Nucleic Acids Res.* **42**, 7884-7893.

61. Moch, C., Fromant, M., Blanquet, S., and Plateau, P. (2017) DNA binding specificities of Escherichia coli Cas1-Cas2 integrate drive its recruitment at the CRISPR locus. *Nucleic Acids Res.* **45**, 2714-2723.

62. Kieper, S. N., Almendros, C., and Brouns, S. J. J. (2019) Conserved motifs in the CRISPR leader sequence control spacer acquisition levels in Type I-D CRISPR-Cas systems. *FEMS Microbiol. Lett.* **366**, fnz129.

63. Grainy, J., Garrett, S., Graveley, B. R., and Terns, M. P. (2019) CRISPR DNA elements controlling site-specific spacer integration and proper repeat length by a Type II CRISPR-Cas system. *Nucleic Acids Res.* **47**, 8632–8648.

64. Plateau, P., Moch, C., and Blanquet, S. (2019) Spermidine strongly increases the fidelity of Escherichia coli CRISPR Cas1-Cas2 integrase. *J. Biol. Chem.* **294**, 11311-11322.

65. Wright, A. V., Liu, J. J., Knott, G. J., Doxzen, K. W., Nogales, E., and Doudna, J. A. (2017) Structures of the CRISPR genome integration complex. *Science* **357**, 1113-1118.

66. Xiao, Y., Ng, S., Nam, K. H., and Ke, A. (2017) How type II CRISPR-Cas establish immunity through Cas1-Cas2-mediated spacer integration. *Nature* **550**, 137-141.
Mechanism of prespacer capture and processing in E. coli

68. Wright, A. V., and Doudna, J. A. (2016) Protecting genome integrity during CRISPR immune adaptation. *Nat. Struct. Mol. Biol.* **23**, 876-883

69. Nuñez, J. K., Kranzusch, P. J., Noeske, J., Wright, A. V., Davies, C. W., and Doudna, J. A. (2014) Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. *Nat. Struct. Mol. Biol.* **21**, 528-534

70. Künne, T., Kieper, S. N., Bannenberg, J. W., Vogel, A. I. M., Miellet, W. R., Klein, M., Depken, M., Suarez-Diez, M., and Brouns, S. J. J. (2016) Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation. *Mol. Cell* **63**, 852-864

71. Sinkunas, T., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P., and Siksnys, V. (2011) Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *EMBO J.* **30**, 1335-1342

72. Musharova, O., Klimuk, E., Datsenko, K. A., Metlitskaya, A., Logacheva, M., Semenova, E., Severinov, K., and Savitskaya, E. (2017) Spacer-length DNA intermediates are associated with Cas1 in cells undergoing primed CRISPR adaptation. *Nucleic Acids Res.* **45**, 3297-3307

73. Babu, M., Beloglazova, N., Flick, R., Graham, C., Skarina, T., Nocek, B., Gagarinova, A., Bogoutse, O., Brown, G., Binkowski, A., Phanse, S., Joachimiak, A., Koonin, E. V., Savchenko, A., Emili, A., Greenblatt, J., Edwards, A. M., and Yakunin, A. F. (2011) A dual function of the CRISPR–Cas system in bacterial antiviral immunity and DNA repair. *Mol. Microbiol.* **79**, 484-502

74. Kim, S., Loeff, L., Colombo, S., Brouns, S. J. J., and Joo, C. (2019) Selective Prespacer Processing Ensures Precise CRISPR-Cas Adaptation. *bioRxiv*, 10.1101/608976

75. Savitskaya, E., Semenova, E., Dedkov, V., Metlitskaya, A., and Severinov, K. (2013) High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in E. coli. *RNA Biol.* **10**, 716-725

76. Rao, C., Chin, D., and Ensminger, A. W. (2017) Priming in a permissive type I-C CRISPR-Cas system reveals distinct dynamics of spacer acquisition and loss. *RNA* **23**, 1525-1538

77. Shmakov, S., Savitskaya, E., Semenova, E., Logacheva, M. D., Datsenko, K. A., and Severinov, K. (2014) Pervasive generation of oppositely oriented spacers during CRISPR adaptation. *Nucleic Acids Res.* **42**, 5907-5916

78. Li, M., Gong, L., Zhao, D., Zhou, J., and Xiang, H. (2017) The spacer size of I-B CRISPR is modulated by the terminal sequence of the protospacer. *Nucleic Acids Res.* **45**, 4642-4654

79. Jackson, S. A., Birkholz, N., Malone, L. M., and Fineran, P. C. (2019) Imprecise Spacer Acquisition Generates CRISPR-Cas Immune Diversity through Primed Adaptation. *Cell Host Microbe* **25**, 250-260.e4

80. Musharova, O., Vyhovskiy, D., Medvedeva, S., Guzina, J., Zhitnyuk, Y., Djordjevic, M., Severinov, K., and Savitskaya, E. (2018) Avoidance of Trinucleotide Corresponding to Consensus Protospacer Adjacent Motif Controls the Efficiency of Prespacer Selection during Primed Adaptation. *mBio* **9**, e02169-18

81. Kim, T. Y., Shin, M., Huynh Thi Yen, L., and Kim, J. S. (2013) Crystal structure of Cas1 from Archaeoglobus fulgidus and characterization of its nucleolytic activity. *Biochem. Biophys. Res. Commun.* **441**, 720-725

82. Zhang, Y. (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **9**, 40

83. Wright, A. V., Wang, J. Y., Burstein, D., Harrington, L. B., Paez-Espino, D., Kyrpides, N. C., Iavarone, A. T., Banfield, J. F., and Doudna, J. A. (2019) A Functional Mini-Integrase in a Two-Protein-type V-C CRISPR System. *Mol. Cell* **73**, 727-737.e3

84. Guerrero, F., Ciragan, A., and Iwaï, H. (2015) Tandem SUMO fusion vectors for improving soluble protein expression and purification. *Protein Expr. Purif.* **116**, 42-49

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

86. Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10-12
87. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402

88. Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004) WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188-1190

89. Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. L. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422-1423

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

91. Notredame, C., Higgins, D. G., and Heringa, J. (2000) T-coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205-217

92. Robert, X., and Gouet, P. (2014) Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* **42**, W320-W324

93. Goddard, T. D., Huang, C. C., Meng, E. C., Pettersen, E. F., Couch, G. S., Morris, J. H., and Ferrin, T. E. (2018) UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14-25

**Footnotes**

Reads obtained from high-throughput sequencing have been deposited in the Sequence Read Archive (SRA) under accession number: PRJNA527928.
Mechanism of prespacer capture and processing in E. coli

Figures

Figure 1. Prespacer length regulates the fate of spacer integration at CRISPR locus. (A) Schema of CRISPR DNA (CD) and prespacer (P) used in Cas1-2 mediated prespacer integration assay is presented. Regions corresponding to 69 bp leader (L in blue), 28 bp repeats (R1-R2 in red), 33 bp spacer 1 (S1 in cyan) and 19 bp spacer 2 (S2 in magenta) of the CRISPR DNA are indicated. Two integration events resulted by 3’-OH nucleophilic attack of Cas1-2 bound prespacer at the top strand (L-R1 junction) and bottom strand (R1-S1 junction) and their respective denatured DNA fragments are shown. The design of integration assays to determine the positions of nucleophilic attack (top strand (i) and bottom strand (ii)) and prespacer ligation (top strand (iii) and bottom strand (iv)) are displayed. (B) Pictures depicting various prespacers that are employed in the integration assay are shown. Prespacers with an overall length of 33 nt are coloured black, whereas, those with >33 nt are represented in orange colour. Sizes of the respective DNA fragments due to top strand integration (P+R) and bottom strand integration (L’+P) are indicated. (C-E) Representative denaturing PAGE gels displaying the prespacer integration at the top strand (C) or at the bottom strand (D) or both (E) are shown. Absence (–) or presence (+) of each reaction product is indicated.
Mechanism of prespacer capture and processing in E. coli

component and the type of prespacer used in each sample are indicated on top of the respective lanes. DNA molecular weight marker (M) positions are shown on the right side. The positions of intermediate products of integration (L, R’, P+R and L’+P) are displayed at the left side of the respective gels. (F) The equilibrium disassociation constant values (K_d) of Cas1-2 with each type of prespacer substrate are displayed. The success (Yes) or failure (No) of integration for each prespacer is shown. (G) Cartoon depicting the relationship between prespacer length (33 nt and >33 nt) and Cas1-2 (in blue and brown) mediated binding and integration is presented. Possibility of binding and integration of each prespacer is denoted as ‘Yes’ or ‘No’.
Mechanism of prespacer capture and processing in *E. coli*

Figure 2. Tailoring of Cas1-2 bound large DNA fragments by exonucleases generates integration competent prespacers. (A) Denaturing PAGE depicting the nuclease treatment of Cas1-2 bound P63 DNA fragments is displayed. Presence (+) or absence (-) of each reaction component is labelled on top of each lane. Positions corresponding to the substrate (P63) and T5 exo/ExoIII digested DNA fragments (P63exo+) are indicated on the left side, whereas, oligo marker (M) positions are shown on the right. (B) Denaturing PAGE displaying the integration reactions employing various prespacers (P23[3’-5] (lanes 1-2 and 5-6), P63 (lanes 3 and 7) and Cas1-2 protected DNA fragments (P63exo+) (lanes 4 and 8)) and CRISPR DNA substrates (CD-T* (lanes 1-4); CD-B* (lanes 5-8)) is shown. Presence (+) or absence (-) of each reaction component is labelled on top of each lane. Positions corresponding to labelled DNA products that are resultant of prespacer nucleophilic attack (L and R’) are displayed on the left. The DNA molecular weight marker (M) positions are shown on the right. (C) Schema illustrating the mechanism of Cas1-2 mediated protection of prespacer boundaries is displayed. Cas1-2 (in blue and brown), T5 exonuclease (magenta pie), Exonuclease III (Cyan pie) and prespacer P63 (red ladder) are portrayed.
Figure 3. Cas1-2 complex is predominantly localized around PAM region. (A) Schematic representation of various fluorescein labelled prespacer substrates (P63T*, P63mPAMT*, P63B* and P63mPAMB* as grey ladder) used in the assay is shown. Labelled 3' -end of each prespacer is highlighted as green circles, whereas, PAM and its mutated variant are depicted as red and black boxes, respectively. Numbering on the DNA represents the distance (in nt) of particular position from the labelled end. T5 exo (magenta pie) is positioned at susceptible 5' -ends of DNA substrate. Positions of T5exo stalling points (magenta triangles) and binding sites of Cas1-2 (blue and brown blobs) that are estimated from nuclease footprinting assay performed in (B) are displayed. (B) Denaturing PAGE depicting the T5exo treatment of Wt Cas1-2 bound fluorescein labelled P63 variants (P63T*, P63mPAMT*, P63B* and P63mPAMB*) is shown. Presence (+) or absence (-) of each reaction component is labelled on top of each lane. Positions corresponding to the DNA fragments of oligo marker (M) are shown on the left.
Figure 4. Structural features of Cas1-2 that determine the prespacer selection. (A-B) Structural comparison of apo-Cas1-2 integrase (PDB ID: 4P6I) (A) and Cas1-2-prespacer complex (PDB: 5DQZ) (B) is presented. Four protomers of Cas1 (indicated as Cas1.a-d.) and two protomers of Cas2 (indicated as Cas2.a-b) are shown in different colours. The stacking of Cas2 tryptophan residues (W44 and W60 in cyan) with I291 of Cas1.a C-terminal tail (G275-S305 in magenta) in apo-Cas1-2 (close-up view in (A)) and the interactions of Cas1.a I291 and Q287 residues with PAM residues (in navy blue) Cytosine 34 (dC34) and Thymidine 35 (dT35) in Cas1-2-prespacer complex (close-up view in (B)) are shown. The Cas1 Y22 residues (in dark green spheres) that stack the nucleic acid bases at the prespacer duplex ends are denoted (B). Amino acid residues corresponding to Cas1 mutations (Q24H, P202Q, G241D, E276D and L297Q) in 5M variant are displayed (green spheres) as part of the Cas1.a and Cas1.c protomers. For clarity, E276 and L297 of Cas1.a and Q24, P202 and G241 of Cas1.c are labelled at their respective positions.
Figure 5. Intrinsic specificity of Cas1-2 integrase directs the uniformity in spacer length and PAM preference during CRISPR adaptation. (A) Denaturing PAGE depicting the T5exo treatment of Cas1-2 (Wt (lanes 1-4) or ΔC (lanes 5-6) or Y22A (lanes 7-8) or 5M (lanes 9-10)) bound fluorescein labelled P63B* is displayed. Presence (+) or absence (-) of each reaction component is indicated on top of each lane. Position of P63B* labelled DNA fragment is shown on the left, whereas, oligo marker (M) positions are indicated on the right. (B) Schematic illustration of the footprinting assay performed in (A) is displayed. DNA substrate P63B* (grey ladder), positions of 3' fluorescein label (green circle) and PAM region (red rectangle) are represented. Numbering on the DNA represents the distance (in nt) of particular position from the labelled end. T5 exo (magenta pie) is positioned at susceptible 5'-ends of DNA substrate. Positions of T5exo stalling points (magenta triangles) and binding sites of each variant of Cas1-2 (Wt or ΔC or Y22A or 5M in blue and brown blobs) that are estimated from nuclease footprinting assay performed in (A) are displayed. (C) Agarose gel depicting the PCR products from spacer acquisition assay performed in E. coli harbouring the plasmids that express the Cas1-2 variants (Wt (lanes 1–2), ΔC (lanes 3–4), 5M (lanes 5–6) and Y22A (lanes 7–8)) is shown. Absence (−) or presence (+) of inducers is indicated on top of each lane. Positions corresponding to parental and expanded arrays (CRISPR 2.1 array) are indicated on left. The percentage of integration is displayed on top of the respective lanes (indicated in blue). DNA marker (M) positions are represented on the right. (D) Overlay of plots
mechanism of prespacer capture and processing in E. coli

depicting length distribution of the newly acquired spacers that are incorporated into CRISPR 2.1 array by Cas1-2 variants (Wt, ΔC, 5M and Y22A) during in vivo integration assay (C) is shown. X-axis depicts length of the spacer (nt) whereas normalised frequency (density) is indicated on the Y-axis. (E) Illustration depicting the PAM preference of Cas1-2 variants (Wt, ΔC, 5M and Y22A) during in vivo integration assay (C) is displayed. +1 to +33 sequence of each spacer (grey ladder) was extracted from high-throughput sequencing data. Subsequently, sequence information of -1 and -2 positions of each spacer was derived from the respective plasmid/genome sequence. The conservation profile of PAM sequences (-2, -1 and +1 in red) corresponding to the respective Cas1-2 variant is shown as a sequence logo.
Figure 6. Cas1/I-E harbours extended C-terminal tail. (A-D) Structures highlighting Cas1 C-terminal tail (in magenta) of Archeoglobus fulgidus (type I-A; PDB ID: 4N06) (A), Pyrococcus horikoshii (type I-B; PDB ID: 4WJ0) (B), Bacillus halodurans (type I-C; predicted model) and Escherichia coli (type I-E; PDB ID: 5DQZ) (D) are portrayed. The amino acids corresponding to start and end position of C-terminal tail are displayed at the bottom of the respective structures. (E) Scatter plot representing the length differences among Cas1 of type I-A, I-B, I-C and I-E is displayed. Each Cas subtype is shown in a different colour and the average length of amino acids in C-terminal tail (Mean±SD) of each subtype is indicated at the respective position. ‘n’ corresponds to the number of Cas1 sequences from each subtype that are considered for the analysis (vide methods section).
Mechanism of prespacer capture and processing in E. coli

Figure 7. Model depicting the mechanism of Cas4 dependent and independent prespacer processing in type I CRISPR-Cas systems. The prespacer production in CRISPR systems encompass a subset of two events: (i) PAM directed prespacer capture and (ii) processing of selected prespacer to a defined length for integration at CRISPR locus. Generally, Cas1-2 captures long dsDNA fragments that are produced by reannealing of ssDNA products (in brown) derived from DNA repair pathways (such as RecBCD in E. coli) or CRISPR interference. Though in type I-E it is clear that the DNA capture by Cas1-2 precedes the prespacer processing event, the order of DNA capture and processing is yet to be understood in other CRISPR-Cas subtypes. The Cas4 (in green) in CRISPR-Cas subtypes I-A, I-C, I-D and I-G or the Cas4 domain of Cas4-Cas1 fusion in type I-U trims the DNA upon recognizing the PAM region (in magenta) (25,45-51). A second copy of Cas4 in type I-G was shown to trim the non-PAM end upon recognizing a short motif (47), whereas, in other subtypes it is not clear whether Cas4 processes this end. Here, the dual role of PAM recognition and prespacer processing by Cas4 propels CRISPR adaptation. Unlike this, in type I-E system, the intrinsic affinity of Cas1-2 integrase towards PAM itself is sufficient to define the potential prespacer regions. This aspect of Cas1-2/I-E precludes the involvement of any PAM specific Cas nucleases (such as Cas4) in prespacer selection. As the Cas1-2/I-E protects the prespacer boundaries efficiently upon recognizing the PAM, any common cellular non-Cas nucleases (in cyan) could trim the exposed ends to generate aptly sized prespacers for integration into the CRISPR locus.
Fidelity of prespacer capture and processing is governed by the PAM mediated interactions of Cas1-2 adaptation complex in CRISPR-Cas type I-E system
Kakimani Nagarajan Yoganand, Manasasri Muralidharan, Siddharth Nimkar and Baskaran Anand

*J. Biol. Chem.* published online November 20, 2019

Access the most updated version of this article at doi: 10.1074/jbc.RA119.009438

Alerts:
- When this article is cited
- When a correction for this article is posted

Click here to choose from all of JBC's e-mail alerts