9938–9952 Nucleic Acids Research, 2021, Vol. 49, No. 17 Published online 24 August 2021
https://doi.org/10.1093/nar/gkab725

Sequence specific integration by the family 1 casposase from Candidatus *Nitrosopumilus koreensis* AR1

Xiaoke Wang[1,2], Qinling Yuan[1,2], Wenxuan Zhang[1,2], Suyu Ji[1,2], Yang Lv[1,2], Kejing Ren[1,2], Meiling Lu[3,*] and Yibei Xiao[1,2,*]

[1]State Key Laboratory of Natural Medicines, China Pharmaceutical University, Nanjing 210009, China, [2]Department of Pharmacology, School of Pharmacy, China Pharmaceutical University, Nanjing 210009, China and [3]Department of Biochemistry, School of Life Science and Technology, China Pharmaceutical University, Nanjing 210009, China

Received June 10, 2021; Revised August 05, 2021; Editorial Decision August 06, 2021; Accepted August 10, 2021

ABSTRACT

Casposase, a homolog of Cas1 integrase, is encoded by a superfamily of mobile genetic elements known as casposons. While family 2 casposase has been well documented in both function and structure, little is known about the other three casposase families. Here, we studied the family 1 casposase lacking the helix-turn-helix (HTH) domain from Candidatus *Nitrosopumilus koreensis* AR1 (Ca. *N. koreensis*). The determinants for integration by Ca. *N. koreensis* casposase were extensively investigated, and it was found that a 13-bp target site duplication (TSD) sequence, a minimal 3-bp leader and three different nucleotides of the TSD sequences are indispensable for target specific integration. Significantly, the casposase can site-specifically integrate a broad range of terminal inverted repeat (TIR)-derived oligonucleotides ranging from 7-nt to ∼4000-bp, and various oligonucleotides lacking the 5′-TTCTA-3′ motif at the 3′ end of TIR sequence can be integrated efficiently. Furthermore, similar to some Cas1 homologs, the casposase utilizes a 5′-ATAA-3′ motif in the TSD as a molecular ruler to dictate nucleophilic attack at 9-bp downstream of the end of the ruler during the spacer-side integration. By characterizing the family 1 Ca. *N. koreensis* casposase, we have extended our understanding on mechanistic similarities and evolutionary connections between casposons and the adaptation elements of CRISPR-Cas immunity.

INTRODUCTION

Bacteria and archaea are under constant threat of invading genetic elements and have generated sophisticated strategies, such as CRISPR (Clustered regularly interspaced short palindromic repeats) -Cas (CRISPR-associated proteins) system for defense (1–3). CRISPR-Cas systems provide prokaryotes with an adaptive immunity through three phases: adaptation, CRISPR RNA (crRNA) biogenesis, and interference (4–6). During adaptation, a new molecular memory represented by a spacer is acquired when a short foreign DNA-derived prespacer is captured and incorporated into the CRISPR array (6,7). Upon reinvasion by the previous foreign genomes, the CRISPR cluster is transcribed into an RNA molecule, and then processed to serve as a guide for a distinct complex of Cas proteins to recognize and initiate the destruction of homologous infecting nucleic acids (8,9).

Although various CRISPR-Cas systems with different Cas proteins and genomic architectures have been identified, the acquisition elements, Cas1 and Cas2, are highly conserved in nearly all CRISPR systems (10–12). They work in concert to select, process and integrate foreign nucleic acids into the CRISPR array (13,14). Cas1 functions as the integrase to catalyze spacer integration into the CRISPR repeat, while Cas2 appears to primarily serve a structural role to bridge the two Cas1 dimers and to stabilize the bound prespacer DNA and the CRISPR repeat during integration (15–17). Two type V CRISPR-Cas subtypes, type V-C and V-D, lack the *cas2* gene but can catalyze integration of shorter prespacers into a shorter repeat, suggesting the preservation of function of the ancestral Cas1 prior to Cas2 adoption (18). However, the origin of Cas1, as well as the evolution of CRISPR-Cas systems, remains unclear.

Recent phylogenetic analysis of Cas1 proteins has led to the discovery that they potentially originated from casposons, a novel class of transposons, in which a Cas1 homolog (termed 'casposase') functions as the transposase (19–21). Unlike canonical *cas1* genes from CRISPR-Cas

[*]To whom correspondence should be addressed. Tel: +86 25 86185335; Email: yibei.xiao@cpu.edu.cn
Correspondence may also be addressed to Meiling Lu. Email:lumeiling@cpu.edu.cn

systems, the casposases are not associated with CRISPR loci or other *cas* genes, although *cas4*-like genes have been identified in some casposons (22,23). Instead, they are associated with a variable set of transposase-related genes such as a family B DNA polymerase (Poly B) (24,25). These clusters are flanked by terminal inverted repeats (TIRs) and further encompassed by direct repeats termed target site duplications (TSDs), which is a common feature in various DNA transposons (20,26).

Based on the gene content, taxonomic distribution, phylogeny and specific relationships with the Cas1 proteins, casposons are classified into four families (24). Family 2 casposases from *Aciduliprofundum boonei* and *Methanosarcina mazei* possess distinct C-terminal helix-turn-helix (HTH) domain and exhibit integration activity *in vitro* (27,28). These casposases are able to integrate TIR-derived oligonucleotides and artificial mini-casposons flanked by the TIR into a target site (29,30). Notably, casposon integration bears some striking resemblance to prespacer integration by Cas1, as both of them adopt two-step transesterification reaction with site specificity during integration (30,31). Furthermore, their corresponding target sites require two key components: a sequence that duplicates upon insertions of the selected DNA, and an upstream sequence that orientates the integration (32,33). The recently resolved *M. mazei* casposase structure revealed putative architectural changes that enable casposase to acquire Cas2-binding preference concomitant to the loss of tetramerization upon target binding (28). These experimental evidences combined with phylogenetic analysis, potentiates the evolutionary scenario that the adaptation module of CRISPR-Cas systems is originated from casposons (22,34,35).

However, our knowledge on the integration mechanisms of casposons from other families remains limited. Previous studies suggested a family 1 casposase from Candidatus *Nitrosopumilus koreensis* AR1 can mediate concerted integration of TIR-derived double-stranded oligonucleotides into plasmids containing the reconstituted target site (32). Moreover, the detailed integration mechanism of Ca. *N. koreensis* casposon remains unclear, including its preference for the TIR substrates, the exact motif of integration target and the sequential order of the nucleophilic attacks. This has prompted us to further examine and characterize the integration characteristics of this casposon. In this study, by deletion mapping and site-directed mutagenesis, we determined the TSD sequences and the sequence motif that contribute to the recognition of the target by the Ca. *N. koreensis* casposase. Taking advantages of TIR-derived oligonucleotides, we obtained the optimal TIR sequence and its specific nucleotides crucial to the efficient integration. Furthermore, our results indicated that, similar to some Cas1 homologs (36,37), the casposase recognizes a molecular ruler in the TSD motif, which shows that leader-distal insertion sites are determined by active sites of integrase regardless of the specific sequence. Altogether, by addressing the integration properties of the family 1 casposon, we provide the evidence to further support the mechanistic similarities and evolutionary connections between casposon and the adaptation elements of CRISPR-Cas adaptive immunity.

MATERIALS AND METHODS

Cloning, expression, and purification

The codon-optimized Ca. *N. koreensis* AR1 casposase genes (GenBank: AFS80663.1, synthesized by Genscript) were cloned into His_6-Twin-Strep-SUMO-pET28a vectors (KanR), using BamHI and XhoI sites. Sequence verified plasmids were transformed into *Escherichia coli* BL21 (DE3) star cells. The cell culture was grown in LB medium at 37°C until the OD_{600} reached 0.8–1.0. Expression was induced by adding isopropyl-β-D-thiogalactopyranoside (IPTG) to a final concentration of 0.5 mM at 25°C overnight. Cells were harvested by centrifugation (5000 rpm for 10 min) and lysed by high-pressure homogeniser (800 bar) in buffer A (50 mM HEPES, pH 7.5, 10% (v/v) glycerol, 20 mM imidazole and 1 M NaCl). The lysate was centrifuged at 15 000 rpm for 60 min at 4°C, and the supernatant was applied onto the Ni-NTA column (GE Healthcare) pre-equilibrated with buffer A. After washing with 20 column volumes of buffer A, the Ni-NTA column was washed at room temperature with 10 column volumes of buffer B (20 mM HEPES, pH 7.5, 4.5 M NaCl) to remove the nucleic acids retained by the casposase, followed by 5 volumes of buffer A. The protein was eluted with buffer C (50 mM HEPES, pH 7.5, 1 M NaCl, 10% (v/v) glycerol and 300 mM imidazole) and incubated with SUMO-protease at 4°C overnight. The His_6-Twin-Strep-SUMO tag cleaved casposase was further purified by size-exclusion chromatography (SEC, HiLoad 16/60 Superdex 200; GE Healthcare) equilibrated with buffer D (20 mM HEPES, pH 7.5, 150 mM NaCl and 10% (v/v) glycerol), the peak fractions were pooled, concentrated by ultrafiltration and then snap-frozen in liquid nitrogen for later usage.

All mutants were generated by mutagenesis PCR and sequence verified. Expression and purification conditions of the mutants were identical to that of the wildtype, except for the R113A/Y117A mutant, which was induced at 18°C. All primers used for mutations are listed in Supplementary Table S1. The purity of all the proteins was analyzed by SDS-PAGE (Supplementary Figure S4).

Oligonucleotides and DNA substrate preparation

Oligonucleotides used for deletions, insertions or site-directed mutagenesis were ordered from Genscript; 6-FAM and Cy5 labeled oligonucleotides used in casposase integration assays were purchased from Sangon Biotech. Sequences of all oligonucleotides are listed in Supplementary Table S1. Red indicates change for the mutational nucleotide. The TIR and leader-TSD-spacer target duplexes were annealed by heating to 95°C and slow cooling to room temperature in annealing buffer containing 20 mM HEPES, pH 7.5, 150 mM NaCl.

Integration assays using labeled oligonucleotide substrates

To study the TIR sequence, the two strands of the leader-TSD-spacer target DNA were labeled with 3′ 6-FAM and 3′ Cy5, respectively. To study the target DNA sequence, one strand of the TIR was labeled with 5′ 6-FAM. The reactions were incubated with 200 nM TIR, 500 nM casposase and 200 nM cold target site DNA duplexes in an integration buffer containing 20 mM HEPES, pH 7.5, 150 mM NaCl, 5 mM $MnCl_2$ and 50 μg/ml bovine serum albumin (BSA) for 30 min at 37°C unless otherwise stated in the figure legends. The reaction was quenched by adding 25 mM (final) EDTA and digested with 1 mg/ml proteinase K (Sangon Biotech, 40 U/mg) for 30 min. Samples were separated on 8% urea–PAGE in 0.5× TBE buffer. The fluorescent signals were visualized on Tanon 5200 Multi imager.

Sanger and high-throughput sequencing

To investigate the exact integration site, Sanger and high-throughput sequencing were performed using unlabeled 59-nt single-stranded TIR (ssTIR) and 58-bp target substrate comprised of 22-bp leader,13-bp TSD and 23-bp spacer. The integration products were amplified by PCR. The PCR products were purified with the DNA Clean-Up Kit (Axygen), then Illumina barcodes and adapter sequences were added via PCR, after which the resulting library was separated on a 1% agarose gel. DNA in a 100–400 bp size range was selected and isolated using the Hieff NGSTM DNA Selection Beads Kit (Yeasen). Sequencing was performed on an Illumina MiSeq. After sequencing, samples were demultiplexed by barcode and mapped to determine the sites of integration.

Integration of the mini-casposon into leader-TSD-spacer target

The mini-casposon in different length containing a kanamycin resistance gene flanked by the 20-bp TIR from Ca. *N. koreensis* casposon was generated by PCR using primers Mini1000-F and relevant Mini1000/2000/3000/4000-R. The long fragment mini-casposon were purified by the Gel DNA Extraction Mini Kit (Vazyme). The short chain target integration assay was performed with 200 nM long fragment mini-casposon, 500 nM casposase and 200 nM 45-bp fluorescent leader-TSD-spacer target (each strand in the dsDNA being labelled at the 3′ end by 6-FAM and Cy5 respectively) in the same integration buffer as described above for 60 min at 37°C. Integration products were loaded on an ethidium bromide-free 1.5% agarose gel, scanned for 6-FAM, Cy5 fluorescence. For the full site integration assay of pUC19 plasmid target bearing leader-TSD-spacer, 30 ng/μl plasmid was employed in the integration reaction, the products were precipitated with isopropanol at –20°C for three hour and resuspended in sterile water, then electroporated to the ElectroMAX™ DH10B cells (ThermoFisher) and plated on ampicillin plus kanamycin plates. The sites of integration transformants were determined by the Sanger sequencing.

RESULTS

The TSD motif of Ca. *N. koreensis* Casposon is only 13-bp in length

The well-studied family 2 casposons, such as these from *A. boonei* and *M. mazei*, have a ∼14-bp TSD in their integration target. However, the putative TSD segment in the target of the family 1 Ca. *N. koreensis* casposase, which lacks the C-terminal HTH domain, is exceptionally long (22-bp) (Figure 1A) (24,32). To verify the TSD sequences of Ca. *N. koreensis* casposon, we incubated Ca. *N. koreensis* casposase with a 59-nt ssTIR, and a 58-bp natural target consisting of a presumed 18-bp leader, 22-bp TSD and 18-bp spacer sequences (Unless otherwise stated, the leader or spacer refer to the casposon elements). The leader- and spacer-integration products were amplified by corresponding primers and then analyzed by Sanger sequencing. Unexpectedly, the sequencing results indicated that the first four bp and the last five bp of the putative 22-bp TSD actually belongs to the leader and spacer sequence, respectively. In other words, the precise TSD motif of the Ca. *N. koreensis* casposon is 5′-ATTGATAAAGAGT-3′, a 13-bp long oligonucleotide (Figure 1B).

The previous plasmid integration assay demonstrated that Ca. *N. koreensis* casposase can mediate concerted integration of double-stranded TIR oligonucleotides into plasmids containing the reconstituted target site (32). However, the data did not allow us to confirm whether the proteins could integrate oligonucleotides sequences site-specifically into short oligonucleotide targets, define the definite motif of the integration target, distinguish between half-site and full-site TIR integration, or reveal the sequential order of the two nucleophilic attacks. To address these questions, we performed an assay involving the integration of a 5′ fluorescently labeled TIR oligonucleotide into a linear target consisting of leader-TSD-spacer sequence *in vitro* (Figure 1C).

Given that the precise nature of the casposon ends involved in integration had not been determined, and some Cas1-Cas2 integrases showed apparent preference for prespacers with short 3′ overhang (15,38), we first tested 6-FAM labeled 30-nt long single-stranded TIR (ssTIR30), 30-bp long blunt-end double-stranded TIR (dsTIR30) and dsTIR with 3′ overhang (3′ overhang TIR) as the potential TIR substrates. The integration reaction of Ca. *N. koreensis* casposase was performed with aforementioned TIR substrate and a 36-bp target consisting of 9-bp leader, 13-bp TSD and 14-bp spacer. In agreement with Sanger sequencing, all the TIR substrates were integrated into the expected position (i.e. products of L-int 59-nt and S-int 57-nt, respectively) (Figure 1D). TIR with 3′ overhang exhibited the highest level of leader- and spacer-side integration efficiency, while both dsTIR30 and ssTIR30 were less efficient (Figure 1D and Supplementary Figure S1, S3A). In conclusion, we demonstrated that Ca. *N. koreensis* casposase is able to specifically integrate TIR substrate into the target site containing a leader sequence and a 13-bp TSD sequence.

Ca. *N. koreensis* casposase prefers double-stranded TIR substrate with 3-nt 3′ overhangs and catalyzes full-site integration

Having confirmed that 3′ overhang dsTIR is the most efficient substrate for Ca. *N. koreensis* casposase, we then tested a series of single-forked dsTIR duplexes which are 30–21 bp in length and flanked by 0–9 nt overhangs at the 3′ end to determine the optimal TIR substrate. The results showed that the integration efficiency gradually elevated along the increase in length of overhang, and reached its maximum when dsTIR was flanked by 3-nt 3′ overhang. Longer over-

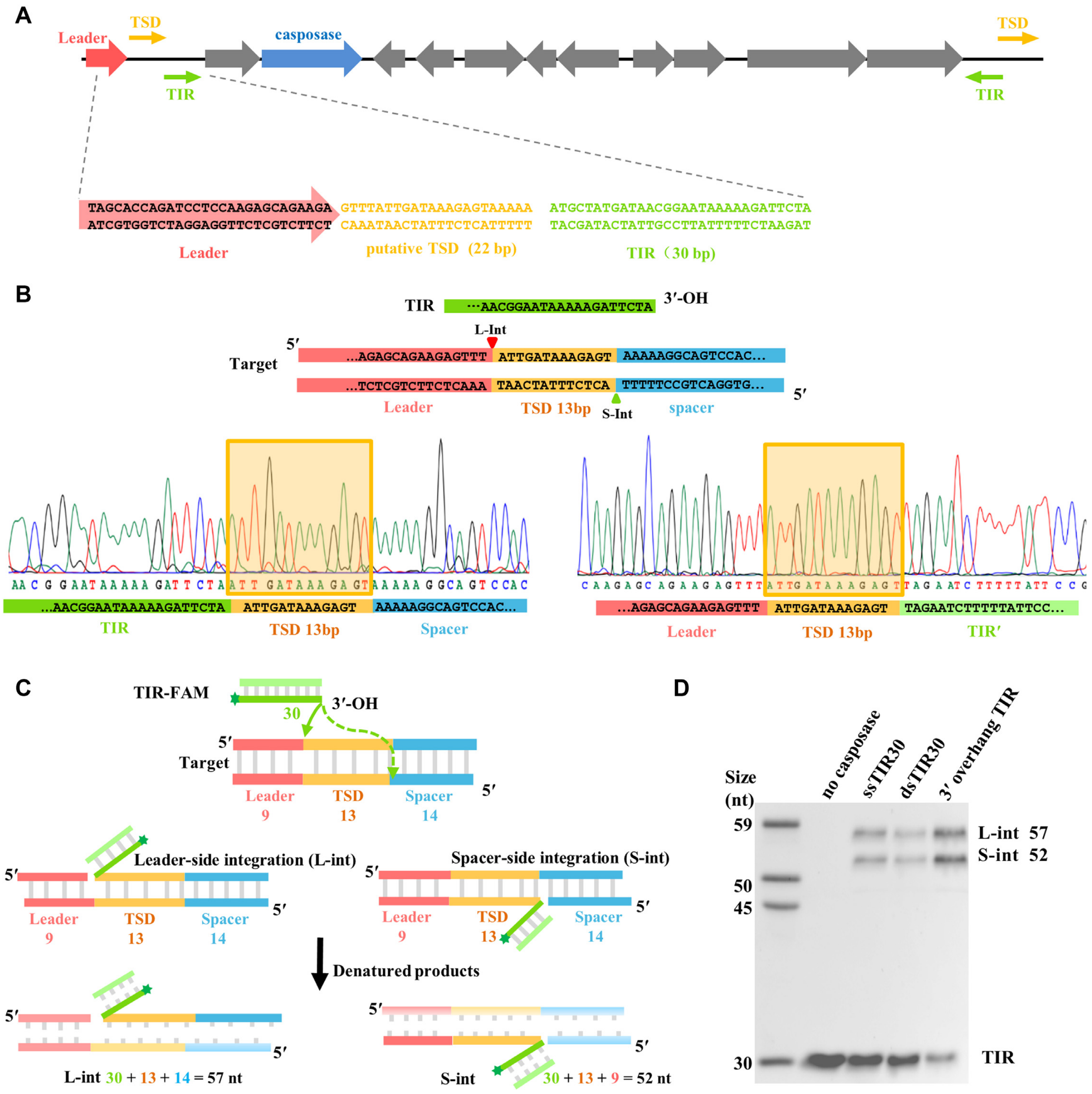

Figure 1. Defining the TSD motif in the integration target of Ca. *N. koreensis* casposon as a 13-bp oligonucleotide. (**A**) Genome architecture and the sequence of key elements of the Candidatus *Nitrosopumilus koreensis* AR1. Terminal inverted repeats (TIR), target site duplication (TSD) and casposase are depicted with green arrow, orange arrow, and a blue arrow, respectively. The leader, TIR and TSD are not drawn to scale, and the sequences are shown below. The putative 22-bp TSD seqeuence started from 5′-GTTT-3′ motif at the 5′-end. (**B**) Sanger sequencing analysis of integration products defines the TSD motif in the target of Ca. *N. koreensis* casposon. Integration sites are indicated by arrowheads. (**C**) Schematic of *in vitro* integration reaction of TIR into a short linear duplex mimicking the *in vivo* target site (hereinafter referred to as 'target' for short). The target consists of a leader sequence (pink), a TSD sequence (orange) and a spacer sequence (blue). Expected products of leader-side integration (L-int) and spacer-side integration (S-int) along with their lengths are indicated. Green star indicates the 6-carboxyfluorescein (6-FAM) label. (**D**) Integration assay of fluorescent 30-nt ssDNA (ssTIR30), dsDNA (dsTIR30) and a 27-bp duplex DNA TIR flanked by 3-nt overhangs at 3′ end (3′ overhang TIR) into a 36-bp linear target that consists of a 9-bp leader, a 13-bp TSD and a 14-bp spacer. Biochemistry was done in triplicates, and representative gels are shown.

hang will lead to reduction in integration efficiency, which dropped dramatically when dsTIR was flanked by 7-nt and 9-nt 3′ overhang (Figure 2A and Supplementary Figure S3B). These results indicated that the length of overhang of dsTIR is precisely selected by Ca. *N. koreensis* casposase, and 3-nt 3′ overhang is mostly preferred. Poor preference for longer overhang is possibly due to the overshoot or undershoot at the integration active site, or the repulsion by the casposase on account of a conformation clash.

We next explored if the Ca. *N. koreensis* casposase was capable of producing full-integration products similar to some Cas1–Cas2 integrases (13,39). A spacer-side hairpin substrate was used as the target for the 5′ 6-FAM labeled TIR integration, which enabled full-site products to be distin-

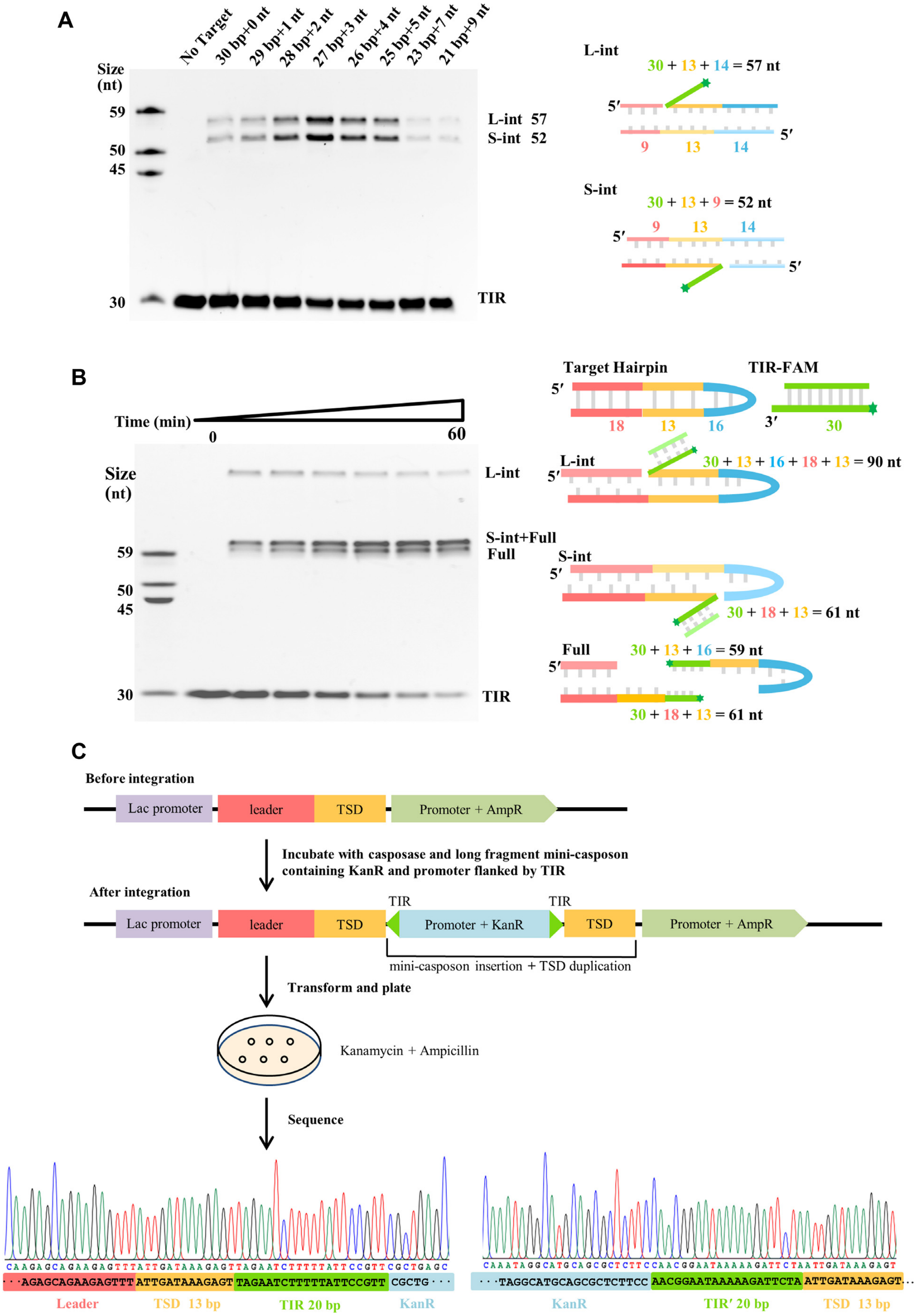

Figure 2. Ca. *N. koreensis* casposase prefers 3-nt 3′ overhangs double-stranded TIR substrate and catalyzes full site integration. (**A**) Determination of the integration preference of Ca. *N. koreensis* casposase for TIR duplexes with 3′ overhang. 'X bp + Y nt' denots that the TIR contains X-bp TIR duplex and Y-nt 3′ overhangs. The target comprises of 9-bp of the leader, 13-bp of the TSD and 14-bp of the spacer. (**B**) Time course of full-site and half-site integration products of spacer hairpin targets and fluorescent TIR. For the hairpin target, expected product sizes are 90-nt for the leader-side integration, 61-nt for the spacer-side half site and 59-nt for the full-site product. For each reaction, time points are: 0, 10 s, 0.5, 2, 10, 30 and 60 min. (**C**) Schematic representation of the kanamycin resistance screen to detect full-site integration near the leader-TSD junction. The pUC19 before integration contained a leader (pink) and TSD (orange) sequence of Ca. *N. koreensis* casposon, and the mini-casposon designed with a kanamycin resistance gene (light blue) flanked by 20-bp left and right TIR (green). After integration, transforming and plating on kanamycin and ampicillin plate allowed for positive selection of full-site integration clones that have the inserted mini-casposon. Representative Sanger sequencing was presented to confirm full-site integration.

guished from half-site intermediates on the basis of product size. Three integration products, from largest to smallest, corresponded to leader-side half-integration (90-nt), a mixture of full-integration and spacer-side half-integration (61-nt) as well as individual full-integration (59-nt), were observed respectively (Figure 2B). The time course analysis showed that leader-side half-integration products diminished gradually, while the full-site products accumulated steadily (Figure 2B and Supplementary Figure S3D), illustrating that the half-site integration product may be an intermediate that would be converted to a full-site product. To further verify the full-site TIR integration, Ca. *N. koreensis* casposase was employed to integrate a ∼1000-bp mini-casposon with a kanamycin resistance gene flanked by the 20-bp TIR into the pUC19 bearing leader-TSD, the constructs obtained were then transformed and plated on the ampicillin plus kanamycin plates. The transformants that survived resistance screen were expected to be the products of full-site integration, which was further confirmed by Sanger sequencing (Figure 2C). The sequencing alignments show insertion of the expected ∼1000-bp mini-casposon and duplication of the 13-bp TSD occur directly adjacent to the mini-casposon (Figure 2C), which is similar to full-site integration in CRISPR-Cas systems. These results exemplify that Ca. *N. koreensis* casposase prefers double-stranded TIR substrate with 3-nt 3′ overhangs and exhibits full-site integration capability similar to some Cas1-Cas2 integrases.

Substrates of various types, lengths and sequences can be integrated by Ca. *N koreensis* casposase

To identify recognition sites in TIR sequences, we next tested casposase integration activity using a 3′ 6-FAM, 3′ Cy5 double-labeled fluorescent target composed of a 19-bp leader, a 13-bp TSD and a 13-bp spacer with the unlabeled single-stranded TIR or their double-stranded counterparts. ssTIR block mutation assays showed that ssTIR (1–25), with the first 25 nucleotides scrambled, can be incorporated with a comparable efficiency to the wildtype (WT) (Figure 3A), confirming that the nucleotides preceding the 5′-TTCTA-3′ motif do not play a significant role in TIR recognition. To determine which nucleotide in 5′-TTCTA-3′ motif was pivotal for efficient incorporation, ssTIR with a series of point mutations were tested. The ssTIR mutants with TTCTT and TTCTG, but not TTCTC at the 3′-end slightly impaired integration (Figure 3B and Supplementary Figure S3C). Although no obvious effects of other point mutations were observed, the scrambled 5′-TTCTA-3′ motif (namely mutant 26–30 in the Figure 3A) was poorly incorporated, suggesting that these nucleotides affect integration specificity and/or efficiency potentially via a cooperative or interactive mechanism. To rule out the possibility that 3′ end nucleotides in the complementary sequences of the dsTIR result in different integration efficiencies, we tested a special double-stranded TIR consisting of a palindromic sequence with a uniform 3′ end. The palindromic dsTIR substrates exhibited a profile similar to that of the ssTIR in the integration. The last 5-bp in the 3′ terminal of dsTIR was crucial for the reaction. In addition to A20T and A20G, T19C, T19G and A20C mutations also reduced integration efficiency (Figure 3C).

It has been reported that the integration activity of casposases from *A. boonei* and *M. mazei* casposase greatly rely on TIR sequence (28,32). In our research, the weak integration efficiency of ssTIR with the scrambled 5′-TTCTA-3′ motif raised a possibility that the random oligonucleotides may also be the substrate for the Ca. *N. koreensis* casposase. To validate this hypothesis, a series of random ssDNA substrates were employed to the integration assay. As shown in Figure 3D, the random ssDNA substrates showed diverse integration activities, suggesting that different from *A. boonei* and *M. mazei* casposase, integration by Ca. *N. koreensis* casposase did not strictly rely on TIR. Strikingly, the 3′ terminal of poor ssDNA substrates are seemingly GC-rich (e.g. CGGCG, CGGCT and CGGAT) compared with the wild type 5′-TTCTA-3′ TIR sequence. Next, to study the insertion capacity of casposase for short- and long-fragment DNA substrates *in vitro*, short TIRs and four long artificial mini-casposons (∼1000, ∼2000, ∼3000 and ∼4000-bp) flanked by 20-bp TIR at both their left and right ends were employed in integration assays with previously mentioned specific 45-bp fluorescent target. For the short TIR substrate, as the length decrease, the integration efficiency reduces, and the shortest incorporated substrate is a 7-nt ssTIR, which may be explained by that a too short substrate cannot contact with the active site of casposase to initiate the transesterification (Figure 3E). When using mini-casposons, clear inserted products were observed with the specific target, while there was no incorporation with the non-specific target (Supplementary Figure S2). Meanwhile, the integration efficiency decreases with the length of mini-casposon increasing, which could be ascribe to the steric hindrance and the weak binding between the long-fragment substrate and casposase. The presence of integration bands demonstrates that the Ca. *N. koreensis* casposase can potentially insert DNA substrates of broad length scales into the specific target site *in vitro*.

Together with the data from the TIR assay, the results indicate that the Ca. *N. koreensis* casposase has a wide tolerance for the type, length and sequence of integrating oligonucleotides, but possesses an apparent preference for TIRs with a 3-nt 3′ overhangs.

The 5′-TTT-3′ motif recognized by the Ca. *N. koreensis* casposase is the minimal leader for the integration

To explore leader recognition by this casposase, we employed an integration assay using 5′ 6-FAM labeled 27-bp double-stranded TIR with 3-nt 3′ overhangs and the leader-TSD-spacer target containing a series of deletion mutations in the leader's elements. Truncation of the leader from 32-bp to 3-bp did not substantially reduce leader-integration products (namely L-int 57-nt) (Figure 4A). Remarkably, deletions that removed all but 2-bp of the leader strongly impaired the tandem integration, as only dispersive spacer-integration bands could be observed (Figure 4A). This indicates that the minimal leader for the integration of Ca. *N. koreensis* casposase is 3-bp in length.

Next, to further identify the nucleotide residues involved in leader recognition, we assayed block substitution mutations consisting of each possible combination of nucleotides in the leader sequences adjacent to the TSD. A reduction of integration efficiency at both the Leader-TSD (LT) and

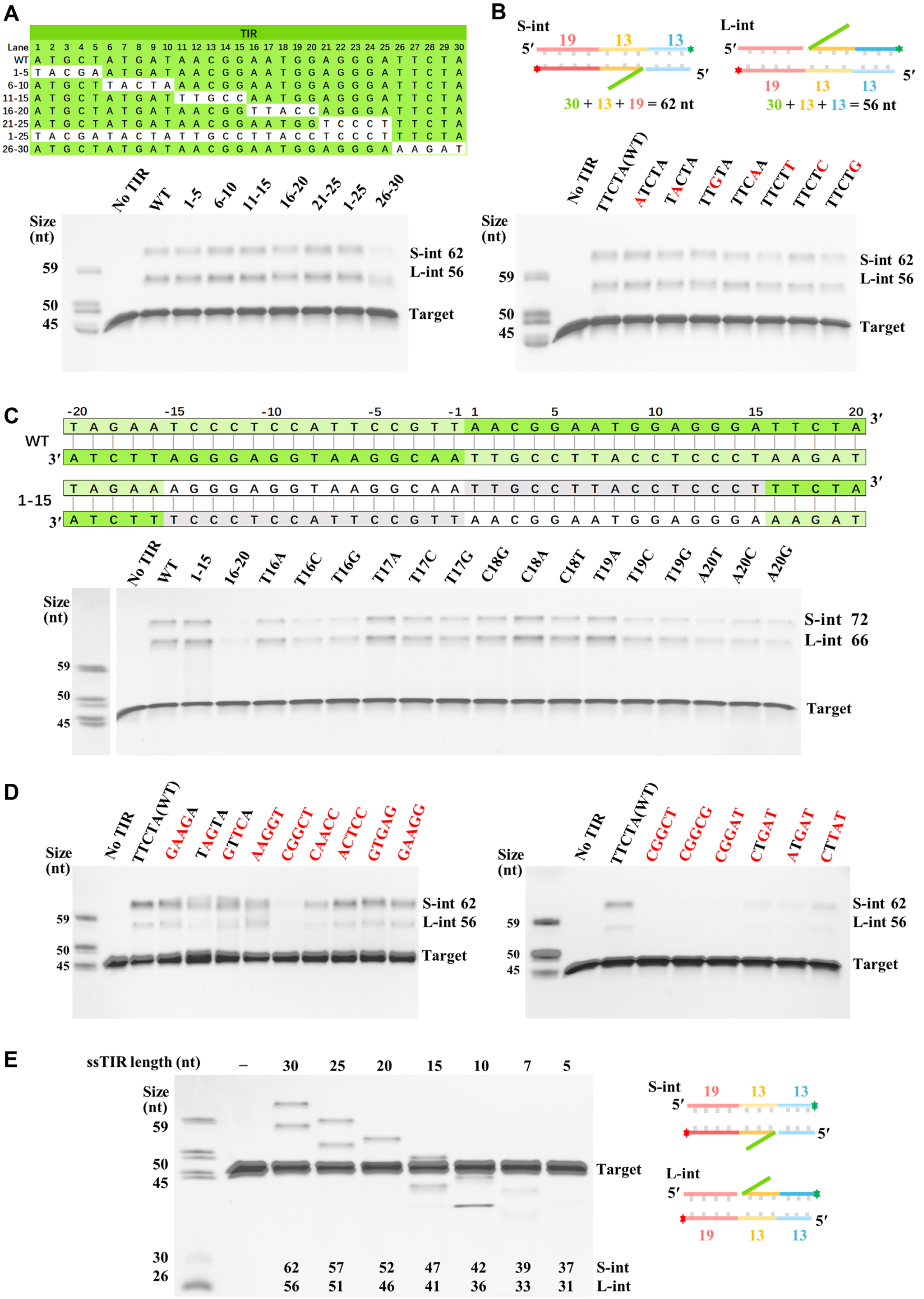
A
TIR
Lane 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
WT A T G C T A T G A T A A C G G A A T G G A G G G A T T C T A
1-5 T A C G A A T G A T A A C G G A A T G G A G G G A T T C T A
6-10 A T G C T T A C T A A A C G G A A T G G A G G G A T T C T A
11-15 A T G C T A T G A T T T G C C A A T G G A G G G A T T C T A
16-20 A T G C T A T G A T A A C G G T T A C C A G G G A T T C T A
21-25 A T G C T A T G A T A A C G G A A T G G T C C C T T T C T A
1-25 T A C G A T A C T A T T G C C T T A C C T C C C T T T C T A
26-30 A T G C T A T G A T A A C G G A A T G G A G G G A A A G A T
No TIR
WT
1-5
6-10
11-15
16-20
21-25
1-25
26-30
Size (nt)
59
50
45
S-int 62
L-int 56
Target
B
S-int
5′
19
13
13
5′
30 + 13 + 19 = 62 nt
L-int
5′
19
13
13
5′
30 + 13 + 13 = 56 nt
No TIR
TTCTA(WT)
ATCTA
TACTA
TTGTA
TTCAA
TTCTT
TTCTC
TTCTG
Size (nt)
59
50
45
S-int 62
L-int 56
Target
C
-20 -15 -10 -5 -1 1 5 10 15 20
WT
T A G A A T C C C T C C A T T C C G T T A A C G G A A T G G A G G G A T T C T A 3′
3′ A T C T T A G G G A G G T A A G G C A A T T G C C T T A C C T C C C T A A G A T
1-15
T A G A A A G G G A G G T A A G G C A A T T G C C T T A C C T C C C T T T C T A 3′
3′ A T C T T T C C C T C C A T T C C G T T A A C G G A A T G G A G G G A A A G A T
No TIR
WT
1-15
16-20
T16A
T16C
T16G
T17A
T17C
T17G
C18G
C18A
C18T
T19A
T19C
T19G
A20T
A20C
A20G
Size (nt)
59
50
45
S-int 72
L-int 66
Target
D
No TIR
TTCTA(WT)
GAAGA
TAGTA
GTTCA
AAGGT
CGGCT
CAACC
ACTCC
GTGAG
GAAGG
Size (nt)
59
50
45
S-int 62
L-int 56
Target
No TIR
TTCTA(WT)
CGGCT
CGGCG
CGGAT
CTGAT
ATGAT
CTTAT
Size (nt)
59
50
45
S-int 62
L-int 56
Target
E
ssTIR length (nt)
–
30
25
20
15
10
7
5
Size (nt)
59
50
45
30
26
Target
62 57 52 47 42 39 37 S-int
56 51 46 41 36 33 31 L-int
S-int
19
13
13
5′
5′
L-int
5′
5′
19
13
13

Figure 3. Diverse DNA can be the substrate of Ca. *N. koreensis* casposase integration. The block (**A**) and single point (**B**) mutagenesis mapping of sequence determinants in the single-stranded TIR (ssTIR). The green and red star denots the 6-FAM and Cy5 labeled at 3′ end of the double-stranded target, respectively. The target was scaled by 19-bp of the leader, 13-bp of the TSD and 13-bp of the spacer. The nucleotides are represented by the last five bases of random ssDNA at its 3′ end, and red base indicates the mutational position relative to the WT. (**C**) Mutagenesis mapping of sequence determinants in the palindromic double-stranded TIR (dsTIR). (**D**) Insertion of various non-labeled random single-stranded DNA of 30-nt into the fluorescent labeled target. (**E**) Ca. *N. koreensis* casposase integration of ssTIR with variable length and the invariable 3′ fluorescent double labeled target.

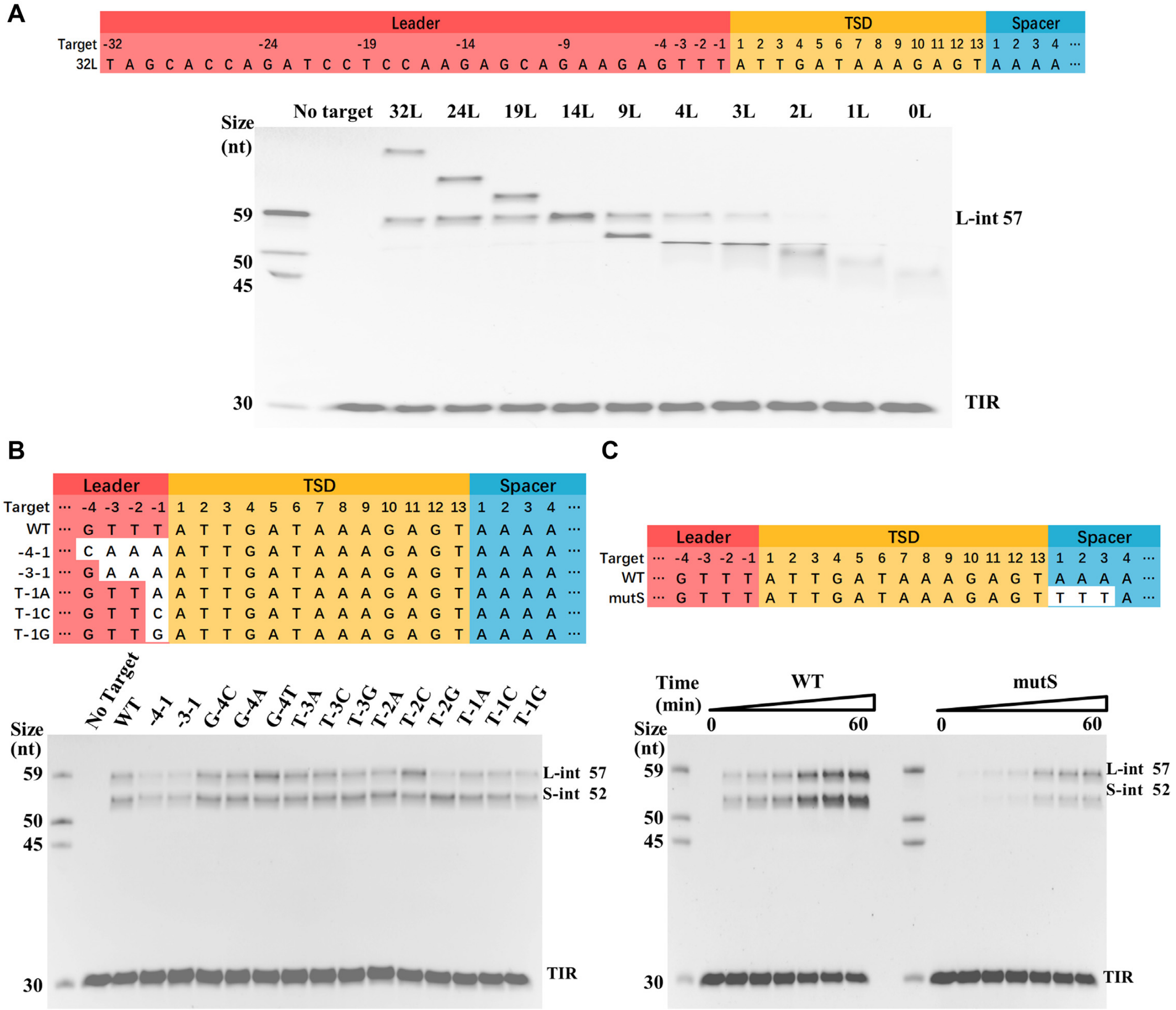

Figure 4. The 5′-TTT-3′ motif is the shortest leader sequence required for the integration of Ca. *N. koreensis* casposase. (**A**) Effect of deletions in the leader segment (32–0L) in the target of Ca. *N. koreensis* casposon. Numbering starts at the border between the leader sequence and the TSD, with + 1 and –1 referring to the first TSD nucleotide and the last leader nucleotide, respectively. For the 'no target' controls, the casposase was incubated for 60 min in the presence of fluorescent TIR and metal ion, but without the addition of target. The target was constituted by different length of leader, 13-bp of TSD and 14-bp of spacer. (**B**) Integration with mutated leader substrates. Transversion mutants were made for every single nucleotide in the indicated regions. The mutant sites are shown in white. Products as well as free TIR are indicated next to each band. (**C**) Integration assay with the wildtype target (WT) and the spacer mutant target (mutS) taken at time points 0, 5 s, 20 s, 1, 15, 30 and 60 min.

TSD-Spacer (TS) junction was observed for both leader block mutants –4 to –1 and –3 to –1 (Figure 4B). Leader mutation of the nucleotide in position –4 and –3 did not significantly impact integration efficiency or specificity at either LT or TS junction. For the thymine in position -2, a moderate loss in specificity at LT junctions was observed in the guanine substitution (T-2G), but not in the adenine (T-2A) or cytosine (T-2C) substitution. Notably, at the last position of the leader, mutation from a thymine to any other nucleotides (T-1A, T-1C and T-1G) did not significantly affect spacer integration but resulted in a loss of integration efficiency at the LT border (Figure 4B). Together, the results support a role for the terminal nucleotides, especially the thymine in position –1 situated at the junction of the leader and TSD sequences, in determining the efficiency and/or specificity of integrations by casposase.

It is worth noting that the first 3-bp (5′-AAA-3′) of the natural spacer sequences in Ca. *N. koreensis* casposon are complementary with the last 3-bp of leader sequences, so we hypothesized that this first 3-bp of the spacer had a leader-like role, dictating the integration preference for the LT and TS junction. Therefore, we mutated the first 3-bp of spacer in target substrate to produce the 'mutS' target substrate. A time course analysis of the integration reaction with wild-type and 'mutS' was performed. As shown in Figure 4C, the spacer mutation of 3-bp 5′-TTT-3′ motif ('mutS') acts as a poor integration target, severely reducing the products of both leader and spacer integration. Unlike some Cas1–

Cas2 integrases in CRISPR-Cas systems (e.g. *E. coli* of type I-E and *Enterococcus faecalis* of type II-A), for which leader integration is preferred and reaches completion faster than spacer-side integration (16,17), no obvious insertion preference has been observed in the 'mutS' target of Ca. *N. koreensis* casposase, hence indicating that the two attacks are not strictly coordinated or ordered. Collectively, these results suggested that integration by the Ca. *N. koreensis* casposase is target-specific and the 5′-TTT-3′ motif at the leader terminal is necessary for efficient casposon integration.

The A1, A11 and T13 serve as significant elements of the TSD in the target of Ca. *N. koreensis* casposon

The repeat sequence in CRISPR-Cas systems plays a key role in guiding Cas1–Cas2 integration (7,38). Therefore, we introduced a series of block and single mutations to the TSD (Figure 5A) and evaluated the effects of each mutation on the integration efficiency and specificity in comparison with the wildtype TSD sequence. Strikingly, mutations in close proximity to each integration site not only completely abolished integration at the proximal site, but also severely affected integration at the other integration site (Figure 5A and Supplementary Figure S3E). Mutation of a mid-TSD sequence block close to the leader (mutant TSD 4–6) did not significantly affect integration preference at the LT or TS junction, but led to a moderate reduction in efficiency on both sides. Meanwhile, mid-TSD sequence mutations proximal to the spacer (mutant TSD 7–9) generated a more asymmetric effect, that is, a significant reduction of spacer-side integration but a slight decrease in integration rate at the leader-TSD junction.

To further characterize the key nucleotides of the TSD during TIR incorporation, systematic single transversion mutants (A→T and C→G) were made for all nucleotides in the TSD regions. Mutating the leader-proximal end of the TSD (i.e. A1T on the top strand) nearly abolished integration at this site, similar to the effect of T13A at the spacer-proximal end (Figure 5B and Supplementary Figure S3F). The point mutations in the internal sequence of the TSD appeared to be less critical for incorporation. Generally, mutation of any single nucleotide of the TSD was insufficient to produce an obvious difference in incorporation efficiency, with the exception of A11T mutation which almost eliminated all detectable integration (Figure 5B and Supplementary Figure S3F).

For detailed mapping of integration site, more point mutations were performed at positions 1, 11, 13 of the TSD, as well as for T3 which is symmetric with A11 (Figure 5C). At the first and last position of the TSD, mutation to the complementary base (A1T and T13A) resulted in a loss of the integration preference at either leader-TSD or TSD-spacer junction, whereas other mutations (A1C, A1G, T13C and T13G) triggered a slight attenuation of the integration preference. A diminution in integration efficiency was observed for T3G mutation, but not for T3A and T3C. Significantly, mutating A11 to any other nucleotide virtually eliminated the efficiency of integration at both the leader and spacer site (Figure 5C), suggesting a significant base-specific contact between A11 and the Ca. *N. koreensis* casposase. These data demonstrate that A1 and T13, the bookends of the TSD, serve as the important sites of nucleophilic attack during TIR integration by directing casposase integration efficiently and/or specifically, and other sites in TSD may also constructively contribute to integration, as the scrambled TSD with only position 1, 3, 11 and 13 unchanged also showed no detectable integration ability.

A molecular ruler in the TSD controls site specificity of Ca. *N. koreensis* casposon

As the TSD harbors asymmetric elements, such as A11, that determine integration efficiency and yet are proximal to the spacer, we postulated the existence of a molecular ruler in the TSD, which has previously been identified in the repeat sequence of CRISPR-Cas systems (36,37,39). We set out to investigate this by generating a panel of targets with TSD insertions and deletions. Considering that the leader sequences can direct the integration site in some CRISPR-Cas system (17,39,40), we speculated this first 3 bp in spacer (i.e. 5′-TTT-3′) may play a leader-like role in dictating the integration site. To rule out the influence of this leader-like spacer and investigate the ruler mechanism, we chose the target with the 5′-GGT-3′ motif in the spacer as the leader-TSD-spacer substrates for the molecular ruler of Ca. *N. koreensis* casposon. Single-nucleotide deletion mutants at positions from 1 to 11 strongly diminished leader-side integration, while Del4 to Del13 strongly reduced spacer-side integration (Figure 6A). A similar but relatively milder attenuation in integration efficiency was observed in these insertion mutants, with the exception of Ins12 (Figure 6B). That mutations in the edge positions of the TSD (1, 2, 12 and 13) typically do not interfere with integration efficiency of the distal side is consistent with previous observations that these junction nucleotides are important for integration efficiency at their proximal side (Figure 5A, B), and that integration is non-preferential to the leader- or the spacer-side.

Moreover, we noticed that a deletion (Del1 or Del2) or an insertion (Ins1 or Ins2) at the leader-TSD (LT) border would result in shorter (12-bp TSD) and longer (14-bp TSD) spacer-side integration products respectively, while the leader-side integration products remained unchanged (Figure 6A, B). A similar situation was observed with the deletion (Del13) or the insertion (Ins11 and Ins12) at the spacer-TSD (ST) border, which led to the product band of leader-side integration presenting at a lower (Del13) or a higher (Ins11 and Ins12) position. Lengths of spacer-side integration products remain almost unchanged. These results suggest that there is a molecular ruler-based mechanism in the TSD governing the specificity of spacer-side integration.

To determine the location of molecular ruler within the TSD of the Ca. *N. koreensis* casposon, a single cytosine (C) residue was inserted sequentially across the 13-bp TSD, and the sites of integration for each mutant were determined by sequencing (Figure 6C). None of the C insertions impacted accurate integration at the leader-TSD border (site 1). However, significant differences were observed in the TSD-spacer integration site before and after site 5–7. Introducing a cytidine upstream of the fifth nucleotide typically leads to a one-nucleotide rightward shift in the spacer-side integration site (i.e. site 15) relative to WT TSD (site

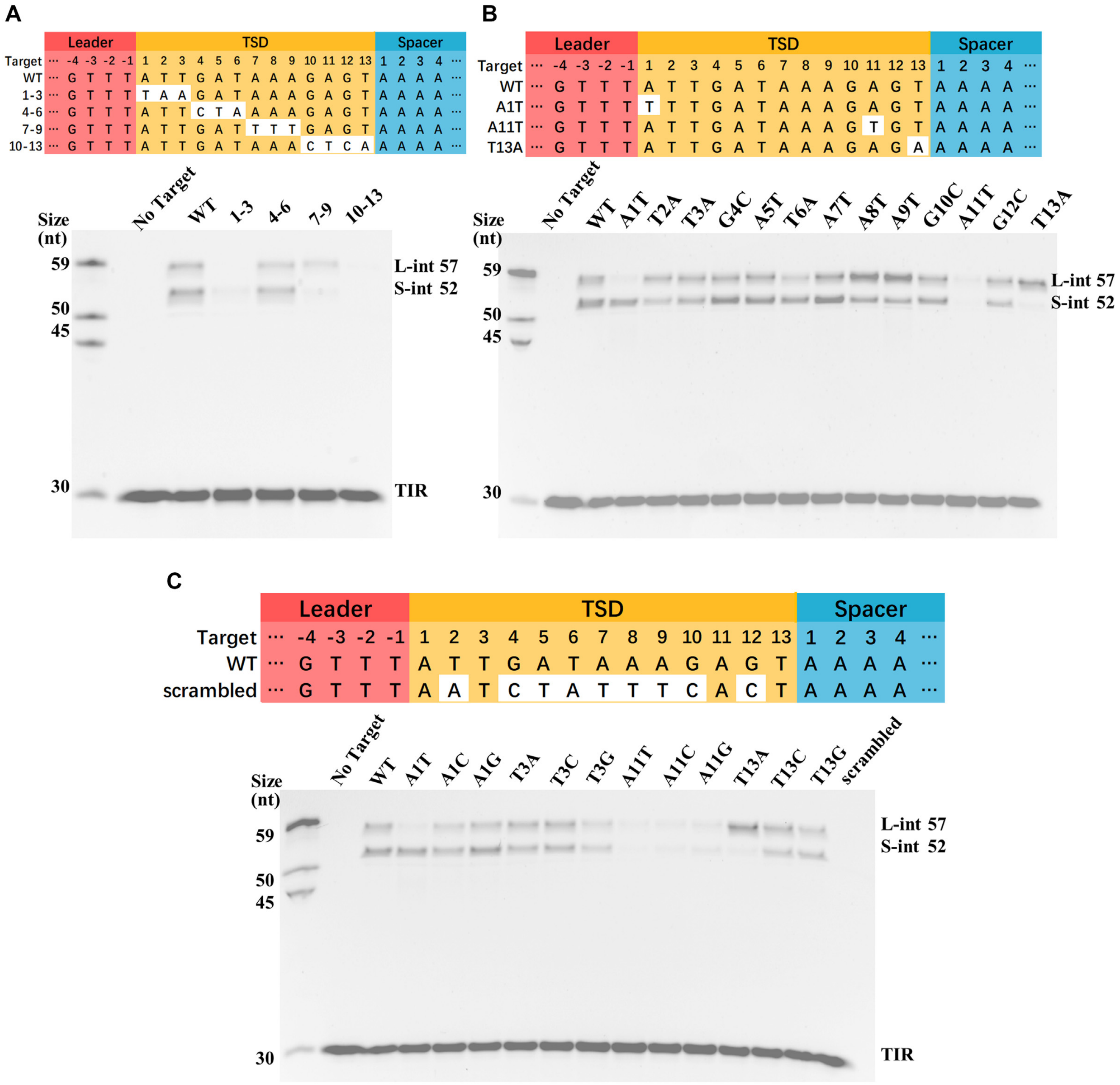

Figure 5. Identifying the key TSD site for the integration by Ca. *N. koreensis* casposase using the TSD variants. (**A**) Integration with block substitution mutants in the TSD, with the same color-coding as in Figure 4 and mutated residues shown in white background. (**B**) The effect of sequential single-point mutant in the TSD sequence on the TIR Integration. (**C**) Integration with systematic point mutations in position 1, 3, 11 and 13 of the TSD.

14). Interestingly, we found that when cytidine was inserted at location downstream of site 8 (mutants Ins8-Ins12, except Ins10), the spacer integration occurred at the same site as the WT repeat (site 14). Moreover, except the mutants Ins5-Ins7, all the mutants had an almost exclusive (>90%) insertion of spacer-side. Specifically, aberrant overlapping Sanger sequencing peaks were observed in spacer-side integration products of Ins5-Ins7 mutants. To determine the actual integration site of the above mutants, high-throughput DNA sequencing was performed. It turned out that the insertion fluctuated between site 14 and 15 with a relatively equal level, suggesting these mutants may disable the molecular ruler and then result in the loss of the exclusiveness of integration at the spacer side. Therefore, we propose that the 5′-ATAA-3′ motif in the site 5–8 of TSD serves as a molecular ruler, with insertion downstream of the ruler resulting in a regular-sized 13-bp TSD, indicating that the molecular ruler would measure a 9-bp distance downstream to this motif, regardless of the insertion.

To further verify the molecular ruler, we constructed a series of single and double insertions (mutants Ins8-9, Ins11-12C) or deletions (mutants Del2, Del4, Del12 and Del12-13) in the TSD and monitored the integration sites of both the leader and spacer sides in the products (Figure 6D). As

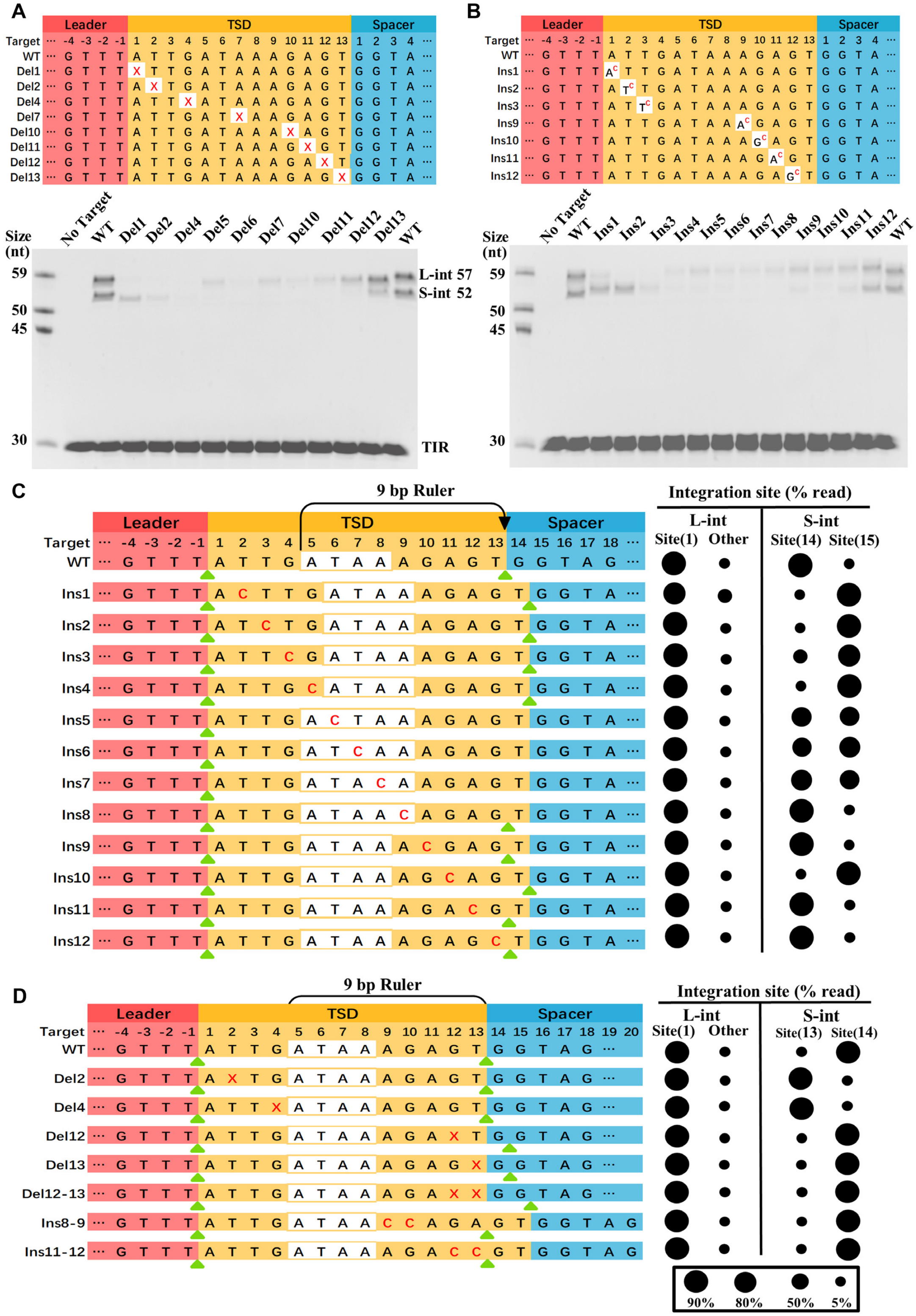

Figure 6. The spacer-side integration of Ca. *N. koreensis* casposon is directed by a molecular ruler harboring in the TSD. Detecting the TIR integration position with a panel of targets with TSD deletions (**A**, the red cross denotes the deleted nucleotide) and insertions (**B**, the superscript C denotes the inserted cytosine after the corresponding position). Mutated residues are shown in white background. The expected products of WT target in Figure B are the same as that shown in Figure A. Integration sites for systematic single insertion (**C**) and single or double deletion and double insertion (**D**) mutations in TSD motif mapped to the linear target at nucleotide resolution as verified by Sanger or high-throughput DNA sequencing. Predicted 5′-ATAA-3′ ruler motif controlling the spacer-side nucleophilic attack 9-bp from the 5′-A is marked. Green triangles indicate the preferred site of integration in every target. Integration site represented as percent of total mapped reads at the corresponding positions of the leader-side integration (L-int, left panel) and the spacer-side integration (S-int, right panel) qualitatively. The deleted base (the red cross in Del2 and Del4) were not counted in the integration site of S-int.

predicted by the molecular ruler scenario, single nucleotide deletion (Del2 or Del4, since site 2 and 3 have the same T nucleotide, Del2 is the same as Del3) at the upstream of ruler motif resulted the loss of one nucleotide from the 3′ end of the TSD, while insertions (Ins8-9, Ins11-12) or deletions (Del12, Del13 and Del12-13) downstream of ruler motif did not affect integration site at the spacer-side. This may be explained by that the molecular ruler measures the length of nine nucleotides from the site the spacer inserted, regardless of the downstream sequence. Taken together, a molecular ruler in the TSD is adopted to measure distinct distance and thus to precisely control integration site of Ca. *N. koreensis* casposon.

Key amino acid residues involved in the integration

To investigate the potential functional residues of Ca. *N. koreensis* casposase involved in the integration activity, we generated a panel of casposase mutants around the catalytic core which may have specific or non-specific interactions with the TIR and leader-TSD sequences, based on its homology model of family 2 *M. mazei* casposase (PDB: 6opm), and sequence alignment of casposases and Cas1 proteins (Supplementary Figure S5). The TIR integration activity of the mutants was tested, and as shown in Figure 7A, the double-mutants, R113A/K117A, and the single mutant K184A, Y197A, Y234A presented decreased integration activity, most notably for Y197A. In light of the mutagenesis results, we proposed a potential integration scenario based on the homologue structure (Figure 7B): the residues Y197 and K184 interact with the phosphate backbone of casposon DNA, while Y197 and Y234 form a stacking interaction with their corresponding TIR base. The residue R113, located at the edge of a α-helix of the casposase monomer, forms a hydrogen bond with D110 from the other monomer, which may be significant for the tetramerization of casposase. By combining the homologue structural analysis and biochemical evidence, we assumed that the aforementioned residues may play key roles in the integration by Ca. *N. koreensis* casposase.

DISCUSSION

The phylogenetic tree analysis splits Cas1 family proteins into two major branches, one consists of the CRISPR-associated Cas1 and the other one includes the casposases (21,24,41). The Cas1 phylogeny thus has led to an evolutionary hypothesis that casposon was the ancestor of the adaptation module of CRISPR-Cas systems. Previous biochemical analysis of the *A. boonei* casposase from family 2, which contains a C-terminal HTH domain, indeed suggested similar biochemical mechanisms between the casposons and the adaptation module of the prokaryotic CRISPR-Cas systems (29,30,32).

In this work, we characterized a family 1 Ca. *N. koreensis* casposase in detail, which lacks the C-terminal HTH domain and makes it more similar to the CRISPR-associated Cas1 proteins. In contrast to previous preliminary characterization (32), our result showed that the TSD motif of Ca. *N. koreensis* casposon is only 13-bp in length. The 5′-GTTTT-3′ sequence flanking the putative TSD motif was actually its leader sequence. Notably, the 5′-GTTTT-3′ sequence is conserved and indispensable at the leader-repeat junction of type II CRISPR-Cas systems (42,43). In addition, both Ca. *N. koreensis* casposase and type II-A Cas1 recognize short leader sequence (17). This raised an interesting possibility that type II CRISPR may evolve from family 1 casposon. Similar to family 2 casposase and Cas1-Cas2 integrase, family 1 Ca. *N. koreensis* casposase also prefers dsTIR substrate with 3-nt 3′ overhang. Unlike family 2 *A. boonei* casposase which prefer C-rich TIR sequence (29,32), the family 1 Ca. *N. koreensis* casposase has a much wider tolerance for the TIR sequence. This feature if interesting and provides the opportunity to engineer the Ca. *N. koreensis* casposase as a tool for site specific insertion of long DNA fragments.

We also observed that the 5′-ATAA-3′ motif situated in the TSD of Ca. *N. koreensis* casposon can control integration site precisely at spacer side. This finding reinforces the mechanism that repeats of some CRISPR-Cas systems (e.g. type II-A in *Streptococcus thermophilus* and type I-E in *E. coli*) harbor a motif which serves as the molecular ruler directing the site of integration (36,37,39). The ability of ruler motifs to maintain constant repeat or TSD length is critical for generating a functional and stable CRISPR or casposon array, which is capable of subsequent immune interference against new viral invaders.

In brief, there are some noteworthy differences between the family 1 casposase we characterized and the well-studied family 2 casposase. First, the HTH domain which is indispensable for the family 2 casposon integration, does not existed in Ca. *N. koreensis* casposase (26,32). This deficiency makes Ca. *N. koreensis* casposase smaller (347 residues) than *A. boonei* casposase (404 residues) and *M. mazei* casposase (405 residues) but possess an approximate residue number as type V-D Cas1 (341 residues) (18,28,29). Second, integration with Ca. *N. koreensis* casposase is strictly dependent on the sequence of leader-TSD target, for instance a scrambled target cannot serve as an effective substrate for Ca. *N. koreensis* casposase. By contrast, *A. boonei* casposase can insert TIR into the plasmid which contains no target site (30,32). Third, *A. boonei* casposase integration is TIR-specific (27,32), whereas various oligonucleotides without TIR sequence can still be efficient integration substrate by Ca. *N. koreensis* casposase. The lack of sequence specificity of TIR substrates is reminiscent of the prespacer of the CRISPR-Cas system. Considering from the biochemical mechanisms, the family 1 casposase may be evolutionary closer to the CRISPR-Cas1 than the family 2 casposase.

An important feature of casposon is its absence of Cas2 protein. It has been shown that the type V-C Cas1 protein can also integrate prespacer in the absence of Cas2, suggesting that the function of the ancestral Cas1 prior to Cas2 adoption (18). The Ca. *N. koreensis* casposase, as well as well characterized *A. boonei* casposase, shows similar integrase activity and target specificity as the type V-C Cas1 proteins but differs in some significant characteristics. First, other than the short TIR sequences, the long (∼4000-bp) DNA fragments flanked by the TIR can also be integrated by casposase. In contrast, the CRISPR-Cas systems are characterized by short spacers, e.g. ∼30-bp or ∼18-bp

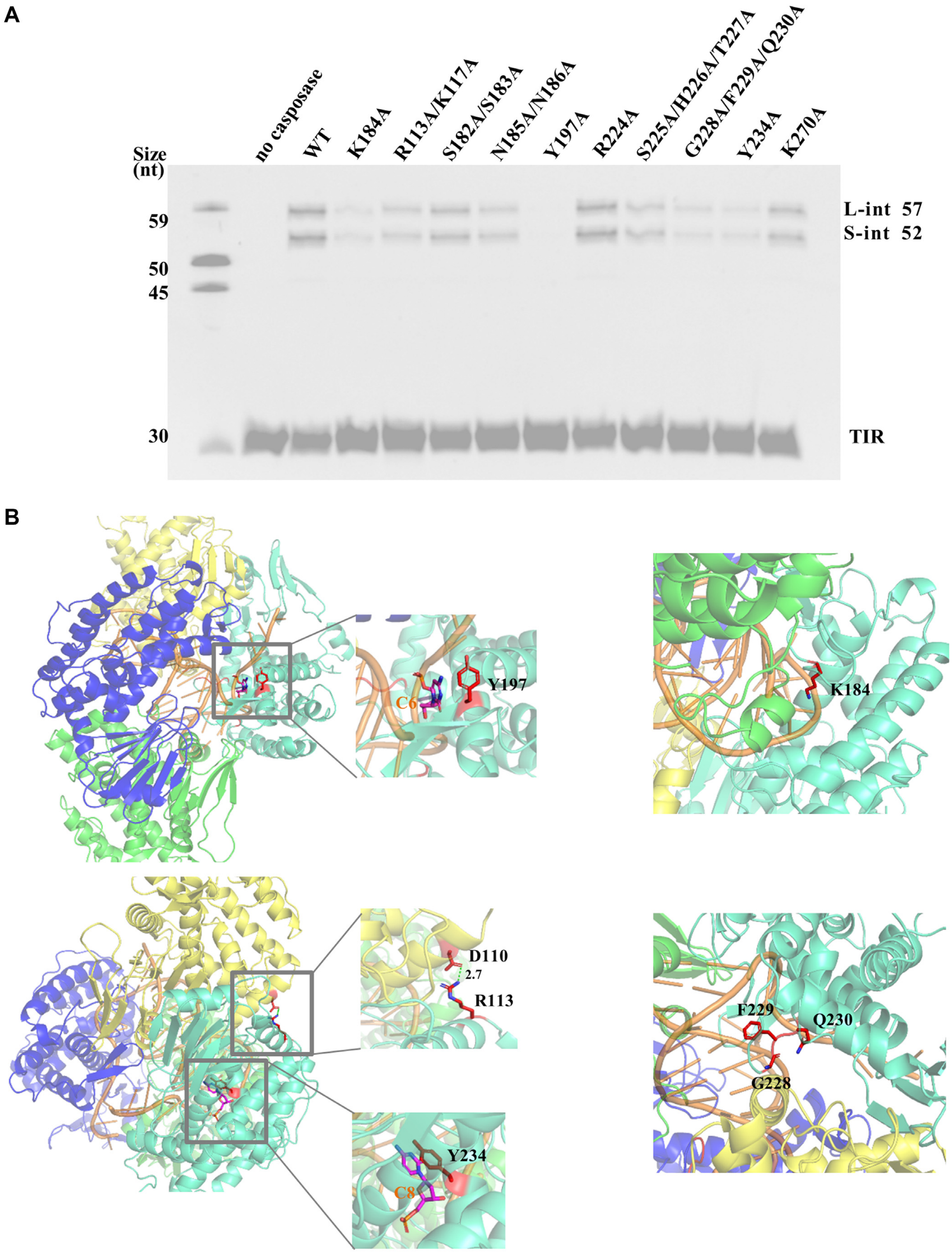

Figure 7. Detection of the potential interaction between the amino acid residues and DNA substrates in the integration by Ca. *N. koreensis* casposase. (**A**) Integration of 5′ fluorescent labeled 3′ overhang TIR into the linear target consisting of a 9-bp leader, a 13-bp TSD and a 14-bp spacer by mutant Ca. *N. koreensis* casposase. (**B**) Close-up view of the potential interaction showing the mutated residues of Ca. *N. koreensis* casposase.

spacer for the *Enterococcus faecalis* Cas1–Cas2 complex or type V-C Cas1, respectively (17,18). Second, Ca. *N. koreensis* casposase presents the preference for the short TSD of 13-bp. Without a Cas2 dimer acting as a bridge, the V-C Cas1 proteins prefer a ∼25-bp repeat, which is shorter than the average of CRISPR-Cas systems but longer than the TSD of casposase (18). Why casposase is able to integrate long DNA sequence while type V-C Cas1 only integrate short fragments remains unclear. Even though recent structural studies have provided insight into how tetramerized *M. mazei* casposase recognizes single-stranded TIR and its target sequence (28), the DNA used in that study does not contain all the elements required for site-specific integration. Further detailed biochemical and structural investigation may elucidate its substrate selection and loading mechanism.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr. Dongfei Liu for providing the fluorescent imager. We would like to acknowledge Yanshan Tao, Xinye Zhang and Zi Ye for their help for the partial language editing. We also express gratitude to Dr. Meirong Chen and Raymond Zhang for critical reading of the manuscript.
Author contributions: X.K.W., M.L.L. and Y.B.X. conceived the project and designed the experiments. X.K.W., Q.L.Y., W.X.Z. and S.Y.J. carried out the experiments. X.K.W., Q.L.Y. and Y.B.X. analyzed the data. X.K.W., Q.L.Y. and Y.B.X. wrote the manuscript. All authors discussed the results and contributed to the final manuscript.

FUNDING

National Key Research and Development Program of China [2018YFA0902000]; National Science Foundation of China [31970547]; Natural Science Foundation of Jiangsu Province [BK20190552]; Project Program of State Key Laboratory of Natural Medicines, China Pharmaceutical University [SKLNMZZ202014]. Funding for open access charge: National Key Research and Development Program of China.
Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. Garneau,J.E., Dupuis,M.E., Villion,M., Romero,D.A., Barrangou,R., Boyaval,P., Fremaux,C., Horvath,P., Magadan,A.H. and Moineau,S. (2010) The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature*, **468**, 67–71.
3. 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**, 6333–6343.
4. Gasiunas,G., Sinkunas,T. and Siksnys,V. (2014) Molecular mechanisms of CRISPR-mediated microbial immunity. *Cell. Mol. Life Sci.*, **71**, 449–465.
5. Bikard,D., Euler,C.W., Jiang,W.Y., Nussenzweig,P.M., Goldberg,G.W., Duportet,X., Fischetti,V.A. and Marraffini,L.A. (2014) Exploiting CRISPR-Cas nucleases to produce sequence-specific antimicrobials. *Nat. Biotechnol.*, **32**, 1146–1150.
6. Hille,F., Richter,H., Wong,S.P., Bratovic,M., Ressel,S. and Charpentier,E. (2018) The biology of CRISPR-Cas: backward and forward. *Cell*, **172**, 1239–1259.
7. 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.
8. Marraffini,L.A. (2015) CRISPR-Cas immunity in prokaryotes. *Nature*, **526**, 55–61.
9. Xue,C.Y., Seetharam,A.S., Musharova,O., Severinov,K., Brouns,S.J.J., Severin,A.J. and Sashital,D.G. (2015) CRISPR interference and priming varies with individual spacer sequences. *Nucleic Acids Res.*, **43**, 10831–10847.
10. Nunez,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.
11. Fagerlund,R.D., Wilkinson,M.E., Klykov,O., Barendregt,A., Pearce,F.G., Kieper,S.N., Maxwell,H.W.R., Capolupo,A., Heck,A.J.R., Krause,K.L. *et al.* (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.
12. McGinn,J. and Marraffini,L.A. (2019) Molecular mechanisms of CRISPR-Cas spacer acquisition. *Nat. Rev. Microbiol.*, **17**, 7–12.
13. Wright,A.V. and Doudna,J.A. (2016) Protecting genome integrity during CRISPR immune adaptation. *Nat. Struct. Mol. Biol.*, **23**, 876–883.
14. Lee,H., Dhingra,Y. and Sashital,D.G. (2019) The Cas4-Cas1-Cas2 complex mediates precise prespacer processing during CRISPR adaptation. *Elife*, **8**, e44248
15. Wang,J.Y., Li,J.Z., Zhao,H.T., Sheng,G., Wang,M., Yin,M.L. and Wang,Y.L. (2015) Structural and mechanistic basis of PAM-dependent spacer acquisition in CRISPR-Cas systems. *Cell*, **163**, 840–853.
16. 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
17. 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.
18. 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**, 727–740
19. Hickman,A.B. and Dyda,F. (2014) CRISPR-Cas immunity and mobile DNA: a new superfamily of DNA transposons encoding a Cas1 endonuclease. *Mob DNA*, **5**, 23.
20. Krupovic,M., Makarova,K.S., Forterre,P., Prangishvili,D. and Koonin,E.V. (2014) Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. *BMC Biol.*, **12**, 36–47.
21. Koonin,E.V. and Makarova,K.S. (2019) Origins and evolution of CRISPR-Cas systems. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **374**, 20180087.
22. Koonin,E.V. and Krupovic,M. (2015) Evolution of adaptive immunity from transposable elements combined with innate immune systems. *Nat. Rev. Genet.*, **16**, 184–192.
23. Koonin,E.V., Makarova,K.S. and Zhang,F. (2017) Diversity, classification and evolution of CRISPR-Cas systems. *Curr. Opin. Microbiol.*, **37**, 67–78.
24. Krupovic,M., Shmakov,S., Makarova,K.S., Forterre,P. and Koonin,E.V. (2016) Recent mobility of casposons, self-synthesizing transposons at the origin of the CRISPR-Cas immunity. *Genome Biol Evol*, **8**, 375–386.
25. Kazlauskas,D., Krupovic,M., Guglielmini,J., Forterre,P. and Venclovas,C. (2020) Diversity and evolution of B-family DNA polymerases. *Nucleic Acids Res.*, **48**, 10142–10156.
26. Koonin,E.V. and Makarova,K.S. (2017) Mobile genetic elements and evolution of CRISPR-Cas systems: all the way there and back. *Genome Biol Evol*, **9**, 2812–2825.

27. Hickman,A.B. and Dyda,F. (2015) The casposon-encoded Cas1 protein from *Aciduliprofundum boonei* is a DNA integrase that generates target site duplications. *Nucleic Acids Res.*, **43**, 10576–10587.
28. Hickman,A.B., Kailasan,S., Genzor,P., Haase,A.D. and Dyda,F. (2020) Casposase structure and the mechanistic link between DNA transposition and spacer acquisition by CRISPR-Cas. *Elife*, **9**, e50004
29. Hickman,A.B. and Dyda,F. (2015) The casposon-encoded Cas1 protein from *Aciduliprofundum boonei* is a DNA integrase that generates target site duplications. *Nucleic Acids Res.*, **43**, 10576–10587.
30. Beguin,P., Charpin,N., Koonin,E.V., Forterre,P. and Krupovic,M. (2016) Casposon integration shows strong target site preference and recapitulates protospacer integration by CRISPR-Cas systems. *Nucleic Acids Res.*, **44**, 10367–10376.
31. Sasnauskas,G. and Siksnys,V. (2020) CRISPR adaptation from a structural perspective. *Curr Opin Struc Biol*, **65**, 17–25.
32. Beguin,P., Chekli,Y., Sezonov,G., Forterre,P. and Krupovic,M. (2019) Sequence motifs recognized by the casposon integrase of *Aciduliprofundum boonei*. *Nucleic Acids Res.*, **47**, 6386–6395.
33. Krupovic,M., Makarova,K.S., Wolf,Y.I., Medvedeva,S., Prangishvili,D., Forterre,P. and Koonin,E.V. (2019) Integrated mobile genetic elements in Thaumarchaeota. *Environ. Microbiol.*, **21**, 2056–2078.
34. Krupovic,M., Beguin,P. and Koonin,E.V. (2017) Casposons: mobile genetic elements that gave rise to the CRISPR-Cas adaptation machinery. *Curr. Opin. Microbiol.*, **38**, 36–43.
35. Makarova,K.S., Wolf,Y.I., Shmakov,S.A., Liu,Y., Li,M. and Koonin,E.V. (2020) Unprecedented diversity of unique CRISPR-Cas-related systems and Cas1 homologs in asgard archaea. *CRISPR J*, **3**, 156–163.
36. 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.
37. Wang,R., Li,M., Gong,L.Y., Hu,S.N. and Xiang,H. (2016) DNA motifs determining the accuracy of repeat duplication during CRISPR adaptation in Haloarcula hispanica. *Nucleic Acids Res.*, **44**, 4266–4277.
38. Nunez,J.K., Lee,A.S.Y., Engelman,A. and Doudna,J.A. (2015) Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature*, **519**, 193–200
39. Kim,J.G., Garrett,S., Wei,Y., 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.
40. Grainy,J., Garrett,S., Graveley,B.R. and M,P.T. (2019) CRISPR repeat sequences and relative spacing specify DNA integration by *Pyrococcus furiosus* Cas1 and Cas2. *Nucleic Acids Res.*, **47**, 7518–7531.
41. Makarova,K.S., Wolf,Y.I., Iranzo,J., Shmakov,S.A., Alkhnbashi,O.S., Brouns,S.J.J., Charpentier,E., Cheng,D., Haft,D.H., Horvath,P. *et al.* (2020) Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.*, **18**, 67–83.
42. Chylinski,K., Makarova,K.S., Charpentier,E. and Koonin,E.V. (2014) Classification and evolution of type II CRISPR-Cas systems. *Nucleic Acids Res.*, **42**, 6091–6105.
43. Van Orden,M.J., Newsom,S. and Rajan,R. (2020) CRISPR type II-A subgroups exhibit phylogenetically distinct mechanisms for prespacer insertion. *J. Biol. Chem.*, **295**, 10956–10968.