Cas4-Cas1 Is a Protospacer Adjacent Motif-Processing Factor Mediating Half-Site Spacer Integration during CRISPR Adaptation

Kieper, S.N.; Almendros Romero, C.; van Eijkeren-Haagsma, A.C.; Barendregt, Arjan; Heck, Albert J.R.; Brouns, S.J.J.

DOI
10.1089/crispr.2021.0011

Publication date
2021

Document Version
Accepted author manuscript

Published in
CRISPR Journal

Classification: BIOLOGICAL SCIENCES: Microbiology

Research article

Cas4-Cas1 is a PAM-processing factor mediating half-site spacer integration during CRISPR adaptation

Sebastian N. Kieper[1,2], Cristobal Almendros[1,2], Anna C. Haagsma[1,2], Arjan Barendregt[3,4], Albert J.R. Heck[3,4], Stan J.J. Brouns[1,2§]

[1] Department of Bionanoscience, Delft University of Technology, Delft, Netherlands

[2] Kavli Institute of Nanoscience, Delft, Netherlands.

[3] Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research, Utrecht Institute of Pharmaceutical Sciences, Utrecht University, Utrecht, Netherlands.

[4] Netherlands Proteomics Center, Utrecht, Netherlands.

§ Corresponding author: Brouns, S.J.J. (stanbrouns@gmail.com, Tel +31 15 278 3920)

Keywords: CRISPR adaptation, Cas4, Spacer acquisition, PAM selection

Abstract

The immunization of bacteria and archaea against invading viruses via CRISPR adaptation is critically reliant on the efficient capture, accurate processing and integration of CRISPR spacers into the host genome. The adaptation proteins Cas1 and Cas2 are sufficient for successful spacer acquisition in some CRISPR-Cas systems. However, many CRISPR-Cas systems additionally require the Cas4 protein for efficient adaptation. Cas4 has been implied in selection and processing of spacer precursors, but the detailed mechanistic understanding of how Cas4 contributes to CRISPR adaptation is lacking. Here we biochemically reconstitute the CRISPR-Cas type I-D adaptation system and show two functionally distinct adaptation complexes, Cas4-Cas1 and Cas1-Cas2. The Cas4-Cas1 complex recognizes and cleaves PAM sequences in 3' overhangs in a sequence-specific manner, while the Cas1-Cas2 complex defines the cleavage of non-PAM sites via host factor nucleases. Both sub-complexes are capable of mediating half-site integration, facilitating the integration of processed spacers in the correct, interference-proficient orientation. We provide a model in which an asymmetric adaptation complex differentially acts on PAM and non-PAM containing overhangs, providing cues for the correct orientation of spacer integration.

Introduction

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and their associated genes (*cas*) provide adaptive and inheritable immunity against mobile genetic elements (MGEs) in bacteria and archaea (Nussenzweig and Marraffini, 2020). The CRISPR array is composed of palindromic repeats interspersed by sequences derived from MGEs and serves as a template for the biogenesis of CRISPR RNAs (crRNAs) (Barrangou, 2013; van der Oost et al., 2014). Cas proteins subsequently assemble around the crRNA to form effector complexes that mediate the recognition and destruction of invading MGEs that have been recorded in the bacterial genome during previous infections (Brouns et al., 2008). Therefore, the main requirement for the establishment of immunity is the memorization of foreign genetic material in a step called CRISPR adaptation (Jackson et al., 2017; McGinn and Marraffini, 2019). The core machinery responsible for adaptation is composed of the Cas1 and Cas2 proteins that assemble into the adaptation complex (Nuñez et al., 2015a; Nuñez et al., 2014; Yosef et al., 2012). Among the first identified *cas* genes were the adaptation genes *cas1* and *cas2*, the *cas3* gene encoding the nuclease-helicase Cas3 and the *cas4* gene encoding a protein, the function of which has been unknown until recently (Hou and Zhang, 2018; Jansen et al., 2002). The *cas4* gene is widespread among several sub-types of type I, type II and type V systems and therefore present in the majority of CRISPR-Cas systems (Hudaiberdiev et al., 2017). Predictions of Cas4 function existed early on, based on the frequent colocalization of the CRISPR adaptation genes *cas1* and *cas2* and the *cas4* gene. This co-localization suggested that Cas4 could be contributing to the adaptation stage and the early studies indeed found supporting evidence for this hypothesis: Adaptation of the type I-A system of *Sulfolobus islandicus* was severely impaired upon deletion of *cas4* (Liu et al., 2017). Similarly, deleting *cas4* from the type I-B system of *Haloarcula hispanica* abrogated CRISPR adaptation

against HHPV-2 (Li et al., 2014). Biochemical evidence was provided by Plagens et al. showing a protein-protein interaction *in vitro* between Cas4 and the type I-A adaptation fusion protein Cas1/2 and Csa1 demonstrating that Cas4 directly interacts with the adaptation machinery (Plagens et al., 2012). Recently, several studies defined the role of Cas4 in more detail, finding that the presence of Cas4 increases the fidelity of spacer integration. Specifically, the Cas4 protein of a cyanobacterial type I-D CRISPR-Cas system facilitated the integration of interference-proficient spacers that carry the consensus PAM of the type I-D CRISPR-Cas system. Spacers acquired in the presence of Cas4 displayed shorter lengths compared to those acquired in the absence of Cas4 or in the presence of catalytically inactive variant of the protein (Kieper et al., 2018). Additionally, two Cas4 variants (Cas4-1 and Cas4-2) encoded in the type I-A system of *Pyrococcus furiosus* were shown to define the upstream protospacer adjacent motif (PAM) and a downstream NW motif *in vivo* (Shiimori et al., 2018). Deletion of *cas4-1* and *cas4-2* resulted in incorrect processing of prespacers with respect to up- and downstream motifs, random orientation of the integrated spacer as well as large deviations from the consensus spacer length (Shiimori et al., 2018). Previously, biochemical studies of the Cas4 protein, containing an iron-sulfur cluster and a RecB domain, found several nuclease activities, demonstrating endo- and exonuclease activities (Lemak et al., 2013; Lemak et al., 2014; Zhang et al., 2012). The requirement of Cas4 for prespacer processing was therefore in accordance with the previously described biochemical activities. Indeed, Lee et al. provided the first mechanistic details of how Cas4 proteins ensure PAM-processing and correct spacer orientation (Lee et al., 2019; Lee et al., 2018). It was shown that Cas4 tightly interacts with the Cas1 integrase, forming a heterohexameric complex composed of two Cas1 dimers and two Cas4 subunits (Lee et al., 2018). This complex would interact with

double-stranded prespacer substrates and endonucleolytically cleave PAM sequences in long 3' overhangs, ensuring that only PAM-processed spacers would be eventually integrated into the CRISPR array (Lee et al., 2018). Interestingly, the authors did not see interaction of the Cas4-Cas1 complex with Cas2 in their initial experiments, suggesting the possibility that a prespacer is required for assembly of the full complex. After supplying a dsDNA substrate to the three adaptation proteins, Lee et al. could demonstrate the assembly of the full Cas4-Cas1-Cas2 complex (Lee et al., 2019). This complex was shown to assemble in a mixture of symmetric and asymmetric architectures as shown by negative-staining Electron Microscopy, in which the asymmetric complex would contain only a Cas4 monomer associated with one of the Cas1 dimers (Lee et al., 2019). The authors suggested that this asymmetry might aid in the differential processing of prespacer substrates in which the Cas4 containing half of the complex would interact with the PAM-containing overhang. This hypothesis is supported by the findings in the type I-A system, in which two independent Cas4 homologs are dedicated processing factors for the PAM- and NW motif containing prespacer overhangs (Shiimori et al., 2018). However, how this asymmetric processing is orchestrated in CRISPR systems containing only a single *cas4* gene is currently unknown. In this work we provide mechanistic insights of an asymmetric complex, specifically how the Cas4-Cas1 complex is able to recognize and sequence specifically process the PAM sequence of the type I-D CRISPR-Cas system. Previously we have shown that the type I-D Cas4 protein facilitates the integration of PAM-compliant spacers *in vivo* (Kieper et al., 2018). We demonstrate that Cas4 strongly interacts with the Cas1 integrase forming a heteromeric $Cas4_1$-$Cas1_2$ complex. This heteromeric complex does not require the Cas2 protein for processing and half-site integration of PAM-containing prespacer substrates. The catalytic activity of Cas4 is required for

prespacer cleavage and is crucially dependent on the presence of Cas1 in order to recognize and process the PAM overhang. We show that this Cas4-Cas1 complex does not cleave the non-PAM containing overhang. Processing of the non-PAM containing overhang potentially relies on the Cas1-Cas2 complex and likely requires host-factor nucleases. We provide a model in which an asymmetric adaptation complex differentially acts on PAM and non-PAM containing overhangs, providing cues for the correct orientation of spacer integration. This correct PAM processing as well as a functional orientation explains the importance and hence the strong conservation of the *cas4* gene, increasing the integration of interference-proficient spacers.

Material & Methods

Bacterial strains and growth conditions

E. coli strains DH5α and BL21 were grown in Lysogeny Broth (LB) at 37ºC and continuous shaking at 180 rpm or grown on LB agar plates (LBA) containing 1.5% (wt/vol) agar. When required, the media were supplemented with 100 µg ml^{-1} ampicillin, 50 µg ml^{-1} spectinomycin, 25 µg ml^{-1} chloramphenicol (see Table S1 for plasmids and corresponding selection markers).

Plasmid construction and transformation

Plasmids used in this study are listed in Table S1. All cloning steps were performed in *E. coli* DH5α. Primers described in Table S2 were used for PCR amplification of the type I-D CRISPR-Cas locus (*cas4*, *cas1*, *cas2* and leader-repeat-spacer1) from *Synechocystis* cell material using the Q5 high-fidelity Polymerase (New England Biolabs). PCR amplicons were subsequently cloned into Berkeley MacroLab LIC vectors (https://qb3.berkeley.edu/facility/qb3-macrolab/) using either ligation-independent cloning (LIC), or into the pACYCDuet-1 vector system (Novagen (EMD Millipore) using conventional restriction-ligation cloning. The *cas4*$^{D76A+K91A}$ mutant was obtained using a PCR-based mutagenesis of pCas4^{D76A} using primers listed in Table S2. All plasmids were verified by Sanger-sequencing (Macrogen Europe, Amsterdam, The Netherlands). Bacterial transformations were either carried out by electroporation (2.5 kV, 25 mF, 200 V) using an ECM 630 electroporator (BTX Harvard Apparatus) or chemically competent cells prepared according to manufacturer's manual (Mix&Go, Zymo research). Electrocompetent cells were prepared following a protocol adapted from (Gonzales et al., 2013). Transformants were selected on LBA supplemented with appropriate antibiotics.

Protein expression and purification

Plasmid encoded *cas* genes were either co-expressed or expressed individually in *E. coli* BL21 AI cells (Invitrogen). Pre-cultures were grown from individual colonies and used for inoculation of pre-warmed (37ºC) LB medium at an initial OD_{600}=0.05. Protein expression was induced at OD_{600}=0.5 by addition of IPTG and L-arabinose preceded by a 30-minute cold-shock. Cultures were subsequently grown overnight at 20ºC with continuous shaking. Cells were harvested by centrifugation (10 min, 4ºC, 2400xg) and subsequently resuspended in lysis buffer (50 mM HEPES pH 7.5, 300 mM KCl, 5% Glycerol, 1 mM DTT, 25 mM Imidazole, 0.1% Triton-X 100) supplemented with cOmplete™, EDTA-free Protease Inhibitor Cocktail (Roche). Cells were lysed by two passages through a CF1 cell disruptor (Constant Systems Ltd.) equilibrated with lysis buffer at a constant pressure of 1 kbar. Lysates were cleared by centrifugation (45 min, 4ºC, 25000xg) and filtered through 0.45 µm filter. Protein was bound in batch to HIS-Select (Sigma Aldrich) IMAC resin for 30 min at 4ºC and rotary shaking. IMAC resin was then loaded onto Pierce gravity-flow columns (Thermo Scientific) and washed with 10 CV wash buffer (50 mM HEPES pH 7.5, 300 mM KCl, 5% Glycerol, 1 mM DTT, 50 mM Imidazole). Proteins were subsequently block eluted in 0.5 ml elution buffer (50 mM HEPES pH 7.5, 300 mM KCl, 5% Glycerol, 1 mM DTT, 250 mM Imidazole). Protein concentration and purity was determined by NanoDrop A280 spectroscopy and SDS PAGE analysis. Protein elution fractions were pooled and subjected to size exclusion chromatography using Superdex 200 10/300 GL (GE Healthcare) column with 0.5 ml/min flow rate using elution buffer as mobile phase. Cas1-Cas2 complex IMAC elution fractions used for integration assays were prepared for ion-exchange chromatography by adjusting the KCL concentration to 30 mM and subsequently loaded onto HiTrap Heparin HP column (GE Healthcare). Cas1-Cas2 complexes were eluted by gradually increasing KCL concentration to 1 M. Resulting fractions

were analyzed by SDS-PAGE and appropriate fractions pooled, snap frozen and stored at -80ºC.

Native Mass Spectrometry

Cas4-Cas1 and Cas1-Cas2 complexes were buffer exchanged into 500 mM ammonium acetate (pH 7.5) using seven sequential steps on a centrifugal filter with a molecular weight cut-off of 10 kDa (Sartorius) at 4°C. MS measurements were performed in positive mode by directly infusing the individual complexes at a concentration of 1 μM using an LCT electrospray time-of-flight (Waters, United Kingdom) adjusted for optimal performance in high mass detection (Tahallah et al., 2001; van den Heuvel et al., 2006). The needles used for electrospray were prepared in house from borosilicate capillaries (Kwik-Fil, World Precision Instruments, Sarasota, FL) on a P97 puller (SutterInstruments, Novato, USA) and gold coated by using an Edwards Scancoat Six Pirani 501 Sputter Coater (Edwards Laboratories, Milpitas, USA). During the measurement the capillary voltage was kept at 1200V, cone voltage between 80-150V and the source pressure was increased to ≈8mbar. Exact mass measurements of the individual Cas proteins were acquired under denaturing conditions by adding formic acid to a final

concentration of 5%. All spectra were mass calibrated by using an aqueous solution of cesium iodide (25 mg/ml). Mass spectra were accumulated, averaged, smoothened and centered, using the software MassLynx 4.1 (Waters, United Kingdom).

Nuclease assays

Oligo-nucleotide sequences used in this study are indicated in Table S2. Oligo-nucleotides with C6-Amino modifications on the 5' terminus were obtained from ELLA Biotech (Planegg, Germany). Cy5 or Cy3 (GE Healthcare) labelling of 5' termini was done in 100 mM Sodium-bicarbonate buffer as described by (Joo and Ha, 2012). Unlabeled oligo-nucleotides were obtained from Integrated DNA Technologies (IDT). Nuclease assays were performed in buffer R (5 mM HEPES pH 7.5, 100 mM Sodium-Glutamate supplemented with 2 mM $MnCl_2$ and 10 mM $MgCl_2$. Annealed and Cy5 and Cy3 labelled oligo-nucleotides (Cy3-BN1829+Cy5-BN1830) were added to a final concentration of 125 nM and purified protein complexes to a final concentration of 500 nM. Reactions were incubated for 1 hour at 30ºC after which reactions were quenched by addition of Proteinase K (Thermo Fischer) and incubation for 1 hour at 37ºC. The resulting products were analyzed on denaturing PAGE (10% acrylamide, 8M Urea) and analyzed with Amersham Typhoon fluorescence gel scanner (GE Healthcare).

In vitro spacer integration assays

Oligo-nucleotide integrations with either Cy5 labeled or unlabeled oligo-nucleotides were performed by pre-incubating indicated protein complexes (500 nM) with oligo-nucleotides (250 nM) on ice for 15 min. Following pre-incubation, either linear CRISPR substrate (obtained by Q5 high-fidelity PCR from pCRISPR using primers BN015+BN1398) or supercoiled pCRISPR were added to a final concentration of 7.5

nM. Reaction mixtures were incubated at 30ºC for 1 hour after which reactions were quenched by addition of Proteinase K (Thermo Fischer) and incubation for 1 hour at 37ºC. Reactions were run on 1% native agarose gels for 45 min and gels subsequently stained with SYBR gold (Sigma Aldrich). Gels were scanned for Cy5 and SYBR gold using Amersham Typhoon fluorescence gel scanner (GE Healthcare). For PCR analysis of *in vitro* integration, unlabeled oligo-nucleotides were used in the reaction. Open-circular plasmid DNA was gel isolated and DNA purified using Zymoclean gel recovery kit (ZymoResearch) after which integration was assessed by PCR using primers BN1711+BN1713 (leader distal integration; correct spacer orientation), BN1711+BN1714 (leader distal integration; incorrect spacer orientation), BN1712+BN1713 (leader proximal integration; incorrect spacer orientation) and BN1712+BN1714 (leader proximal integration; correct spacer orientation). Purified PCR amplicons were subjected to MiSeq sequencing (Illumina).

Next generation sequencing and statistical analysis

After validation of PCR amplicons by gel electrophoresis and clean up with the GeneJET PCR Purification kit (Thermo Fisher Scientific) the samples were analyzed using Qubit fluorometric quantification (Invitrogen). Samples were prepared for sequencing with the Nextera XT DNA Library Preparation Kit (Illumina) and each library individually barcoded with the Nextera XT Index Kit v2 SetA (Illumina). Libraries were pooled equally and spiked with ~5% of the PhiX control library (Illumina) to artificially increase the genetic diversity before sequencing on a Nano flowcell (250 nt paired-end) with an Illumina MiSeq. Image analysis, base calling, de-multiplexing and data quality assessments were performed on the MiSeq instrument. FASTAQ files generated by the MiSeq were analyzed by pairing and merging the reads using Geneious 9.0.5 and subsequently extracting the oligo-nucleotide sequences used in

the in vitro integration assay. Overhang processing was analyzed by annotating the primers used for amplification and comparing the overhangs post-integration to the initial oligo-nucleotide sequence.

In vivo spacer integration assays

E. coli BL21 AI cells were co-transformed with either pCas1-Cas2 and pEmpty or pCas1-2 and pCas4 (wild-type Cas4 or Cas4D76A+K91A). One transformant for each combination was grown in LB at 37ºC and continuous shaking (180 rpm) to OD_{600}=0.3 and made electrocompetent after which pCRISPR was transformed. For each treatment three individual colonies were grown in SOB medium (LB supplemented with 10 mM MgSO4 and 10 mM MgCl2) at 37ºC and continuous shaking (180 rpm) to OD_{600}=0.3 after which protein expression was induced by addition of 0.2% L-arabinose and 0.5 mM IPTG. Induced cultures were grown for additional 2 hours at 37ºC and continuous shaking (180 rpm). Cells were made electrocompetent and annealed pre-spacer oligo-nucleotides (BN1763+BN1768) electroporated at a final concentration of 1 µM. After 30 min recovery cells were harvested and plasmid DNA extracted using GeneJET plasmid miniprep kit (Thermo Scientific). Extracted plasmid DNA was normalized to 0.5 ng μl^{-1} and subsequently 2 µl used in half-site integration PCRs using primers BN1711+BN1713 (leader distal integration; correct spacer orientation), BN1711+BN1714 (leader distal integration; incorrect spacer orientation), BN1712+BN1713 (leader proximal integration; incorrect spacer orientation) and BN1712+BN1714 (leader proximal integration; correct spacer orientation). PCR amplicons were validated by agarose gel electrophoresis and purified with the GeneJET PCR Purification kit (Thermo Scientific). Purified PCR amplicons were subjected to MiSeq sequencing (Illumina).

Results

PAM-containing overhang processing depends on orientation of spacer integration

We have previously demonstrated that the type I-D Cas4 protein facilitates the integration of PAM-compatible spacers *in vivo* (Kieper et al., 2018). In these experiments we looked at the total pool of spacers that were acquired from cytosolic DNA, obscuring the detailed mechanism that governs the processing of prespacer substrates. In order to obtain more detailed insights into processing of PAM and non-PAM containing substrates and how spacer orientation affects overhang processing, we electroporated an idealized prespacer substrate into *E. coli* cells overexpressing either Cas1-Cas2 or Cas4-Cas1-Cas2 proteins (Fig. 1A&B). Cas4 was either expressed as the wild-type protein or the catalytically inactive mutant (D76A, K91A). In addition to the adaptation genes, the cells were carrying a plasmid containing the type I-D leader and a single repeat (pCRISPR). By employing half-site integration PCRs followed by high-throughput sequencing, we analyzed the 3' overhangs after processing and integration *in vivo* (Fig. 1C). This approach allowed us to differentiate between correct and incorrect spacer orientation, as well as correct and incorrect PAM processing of their 3' end (Fig. 1D&E).

We observed that prespacer overhangs were trimmed by at least 5 nt in all cases, regardless of the presence or absence of Cas4. However, processing of PAM and non-PAM containing overhangs differed depending on the orientation in which the spacer was integrated. In particular, spacers integrated in the correct orientation (Fig. 1D) were more precisely processed in the PAM-containing overhang in the presence of Cas4. Although cells expressing only Cas1-Cas2 or a combination of Cas1-Cas2 with a catalytically inactive Cas4 double mutant (D76A, K91A) also displayed 30% to 35% of correct processing of the PAM overhang, their spacer size distributions were

typically broader and shifted towards longer overhang lengths. Analyzing the non-PAM containing overhangs of correctly oriented spacers did not display any differences between the conditions (with and without Cas4), suggesting that prespacer overhangs without PAM are not processed by Cas4, but rather by endogenous *E. coli* nucleases. Spacers integrated in the incorrect orientation (Fig. 1E) showed similar 3' overhangs under all conditions. We observed most accurate processing when Cas4 was present in its active form. In those samples the presence of Cas4 led to an increase in the shortening of overhangs, with a predominant overhang length of 6 nucleotides. In the *E. coli* model system, host factor nucleases potentially act on both PAM and non-PAM containing 3' overhangs that remain unprotected by the core Cas1-Cas2 complex holding the prespacer. Cas4 does not specifically cleave non-PAM containing overhangs, but requires the presence of the PAM in order to engage in sequence-specific processing.

Cas4 forms a strong heteromeric complex with Cas1

In order to assess whether the overhang processing connected to the presence of Cas4 was a result of Cas4 specifically interacting with the Cas1-Cas2 integration complex, we first investigated the formation of a Cas1-Cas2 complex. The Cas1 protein was N-terminally His_6-tagged and co-expressed with Cas2 in *E. coli* BL21-AI cells. After the initial nickel-affinity pull-down from cleared cell lysate, the elution fraction was subjected to size exclusion chromatography (SEC), which resulted in one peak containing aggregated protein and another peak species (Fig. 2A). This peak contained three proteins (Fig. 2A) for which the tagged-Cas1, untagged Cas1 (due to autoproteolysis of the tag) and Cas2 protein identity was confirmed by mass spectrometry. Next, we co-expressed the His6-tagged Cas1 with untagged Cas4 and observed strong co-purification of both proteins (Fig. 2B). In order to verify the Cas4-

Cas1 interaction, a reverse tagging strategy was used (His6-tagged Cas4 co-expressed with untagged Cas1), which again confirmed the presence of a Cas4-Cas1 complex (Fig. S1). When tagged-Cas1 and Cas2 were co-expressed along with Cas4, we observed a strong co-purification of Cas4 and Cas1 that abolished formation of the Cas1-Cas2 complex since Cas2 eluted separately as a low molecular weight species. This fraction also contained minor amounts of Cas1 and Cas4 that did not assemble into higher order complexes. Our results demonstrate that under these conditions Cas1 can form complexes with Cas4 or Cas2, and that these complexes appear to be mutually exclusive. In the presence of both Cas4 and Cas2, Cas1 strongly favors the interaction with Cas4 over the interaction with Cas2.

Cas4 associates with Cas1 in a 1:2 ratio

Next, we determined the stoichiometry of the formed Cas4-Cas1 complex. Previously, Lee et al. demonstrated that the heteromeric complex consists of two Cas1 dimers that each associate with a single Cas4 monomer (Lee et al., 2018). To gain insight into the composition of the untagged Cas4-Cas1 complex (Fig. S2) native protein mass spectrometry analysis was performed (van den Heuvel et al., 2006). The mass spectrum (Fig. 2D) revealed a distribution of different complex species, with the most abundant mass-over-charge (m/z) peaks consisting of either Cas1 dimers (73.6 ± 1.6 kDa) or the Cas4-Cas1 complex consisting of a single Cas1 dimer and a Cas4 monomer resulting in a $Cas4_1$-$Cas1_2$ complex of 96.3 kDa (Fig. 2D). Even though we observed co-purification of Cas1 and Cas2 in the SEC analysis, the native mass spectrum of the Cas1-Cas2 complex resulted in mainly Cas1 dimers (Fig. S3) with Cas2 likely being lost during the native MS sample preparation.

The Cas4-Cas1 complex sequence specifically processes PAM-containing 3' overhangs

The acquisition of functional spacers not only requires appropriate prespacer selection, but also PAM-compliant processing. We have previously shown that the presence of Cas4 in addition to the core adaptation proteins Cas1 and Cas2 significantly increases the integration of spacers with a correctly processed PAM *in vivo* (Kieper et al., 2018). Due to the strong interaction of Cas4 and Cas1 we aimed to test whether PAM processing is mediated only by Cas4, or if the heteromeric Cas4-Cas1 complex is required. In order to address this question, we performed prespacer cleavage assays with a dual-labelled model prespacer (Fig. 3A). This model prespacer consisted of a 25 bp duplex flanked by 13 nucleotide 3' overhangs on each side and fluorescent labels at their 5' ends. The top strand was labelled with Cy3 and did not contain a PAM sequence in its 3' overhang of the, while the bottom strand was labelled with Cy5 and contained the I-D consensus PAM. We found that neither free Cas4 nor Cas1-Cas2 was able to catalyze 3' overhang cleavage. However, the addition of the Cas4-Cas1 complex resulted in a defined band corresponding to processing of the PAM sequence within the PAM-containing overhang (Fig. 3A). This result suggests that PAM recognition is mediated by the interactions within the Cas4-Cas1 complex, where Cas4 acts as the catalytic subunit of the complex.

We have previously shown that mutating D76 in the conserved RecB domain of Cas4 abolished integration of PAM-proficient spacers *in vivo* (Kieper et al. (2018). Using the D76A mutant in our in vitro cleavage assay fully abolished processing activity of the Cas4-Cas1 complex, demonstrating that the RecB domain of Cas4 is indeed the catalytically active site required for PAM processing. Interestingly, although Cas4 did not show processing activity on its own, combining Cas1-Cas2 and Cas4 fully restored processing to similar levels as the Cas4-Cas1 complex. The addition of both, the Cas1-

Cas2 and the Cas4-Cas1 complex, resulted in processing of the PAM overhang as observed with the Cas4-Cas1 complex alone or combination of Cas4 and Cas1-2 complex. All conditions that showed cleavage of the substrate resulted in a single defined band, suggesting that cleavage occurred via an endonuclease mechanism which is in line with previous studies (Lee et al., 2018). The processing of the non-PAM containing overhang was not observed in any of the conditions, indicating that the processing of the non-PAM site presumably relies on host factor nucleases such as DnaQ-like exonucleases or Exonuclease T as recently found in in the I-E system (Kim et al., 2020; Ramachandran et al., 2020). Our results demonstrate that sequence specific Cas4 activity requires the presence of Cas1 and that the Cas4-Cas1 complex is the core processing complex that sequence-specifically recognizes and processes the PAM sequence before integration.

The Cas4-Cas1 complex integrates new spacers into both linear and supercoiled DNA

Next, we tested whether the Cas4-Cas1 complex not only processes prespacer substrates but also catalyzes their integration into the CRISPR array. We performed adaptation assays using supercoiled plasmid DNA containing the type I-D leader and a single repeat (pCRISPR; Fig. 1A) as well as linear CRISPR array substrates generated by PCR (Fig. 3B). Both linear and plasmid CRISPR loci were incubated with a Cy5-labelled prespacer, the Cas4-Cas1 and Cas1-2 complexes. Intriguingly, we observed coupling of the labelled prespacer by the Cas4-Cas1 complex to both CRISPR substrates, showing that this sub complex is proficient in catalyzing at least half-site spacer integration (Fig. 3B&C). Spacer integration into plasmid DNA resulted in the formation of open-circular (OC) plasmid conformations. Merging the Cy5 signal of the prespacer and the plasmid DNA signal confirmed that the prespacer was indeed coupled to the OC form of the plasmid. The Cas1-2 complex was able to integrate the

prespacer into both linear and supercoiled arrays similar to the Cas4-Cas1 complex. Our observation demonstrates that at least two different sub-complexes exist, which are both capable of catalyzing half-site spacer integration. Taken together, based on the selective PAM-overhang processing of the Cas4-Cas1 complex, we hypothesize that the Cas4-Cas1 complex processes and integrates the PAM containing overhang and the Cas1-Cas2 complex the non-PAM containing overhang.

Correct spacer orientation requires overhang processing prior to integration

In order to analyze the accuracy of spacer integration by the Cas4-Cas1 complex in more detail, OC plasmid resulting from the integration reaction was gel purified and subjected to half-site integration PCRs as described previously (Fig. 1C). PCR products were subjected to Illumina MiSeq sequencing and prespacer sequences were extracted. This approach allowed us to assess 3' overhang processing before integration at the leader-proximal or leader-distal integration site. Interestingly, PAM-containing overhangs only showed sequence specific processing when the spacer was correctly oriented with respect to the PAM (Fig. 3D). The Cas4-Cas1 complex cleaved 65% of correctly oriented spacers exactly downstream of the PAM, however, we also observed incorrect removal of a single nucleotide in 25% of sequences and removal of 2 or more nucleotides in 10% of the sequences. Surprisingly, incorrectly oriented spacers did not show any processing of the PAM-containing overhang (Fig. 3E), indicating that integration in the correct orientation is preceded by the processing of the overhang. As predicted from the bulk cleavage assays, we did not observe any processing of the non-PAM containing overhangs regardless of the spacer orientation. Our data show that integration of new spacers in the correct orientation by Cas4-Cas1 requires PAM recognition and processing before a spacer can be integrated.

Spacer integration preferentially initiates with the non-PAM overhang

In type I CRISPR-Cas systems spacer integration initiates by first integrating the non-PAM end of the spacer at the leader repeat junction and proceeding with the coupling of the PAM-end of the spacer at the Repeat-Spacer boundary (Arslan et al., 2014; Nuñez et al., 2015b; Rollie et al., 2015). In the Type I-E CRISPR-Cas system that is lacking Cas4, directionality of spacer integration is dictated by the prespacer processing kinetics (Kim et al., 2020; Ramachandran et al., 2020). We therefore hypothesized that the prespacer processing of our Cas4 containing system could influence the orientation of the integrated spacer. To test the effect of the processed and unprocessed prespacer overhangs on integration, we assayed spacer integration using the Cy5-labelled prespacer substrates. Prespacers were either fully processed (5 nt 3' overhangs), with an unprocessed non-PAM overhang (13 nt) or with a processed, but integration-deficient (Fagerlund et al., 2017; Rollie et al., 2015) 3' phosphorylated non-PAM overhang (Fig. 4A). The fully processed substrate was efficiently coupled by Cas1-Cas2, Cas4-Cas1 as well as the combination of both. Similarly, the prespacer with an unprocessed non-PAM overhang was coupled efficiently in all three treatments. However, when the processed non-PAM overhang was blocked for integration by 3'-phosphorylation, neither of the protein complexes was able to efficiently couple the spacer, indicating that coupling of the PAM-overhang requires prior integration of the non-PAM overhang. Altogether, this mechanism ensures that integration of the PAM site of the spacer is halted until integration of the non-PAM site has occurred, resulting in the correct orientation of the spacer with respect to the PAM.

Discussion

Although Cas4 proteins have been recognized as part of the core *cas* gene machinery almost two decades ago (Jansen et al., 2002), its role in acquiring PAM-compatible

spacers has been revealed only in the recent years (Kieper et al., 2018; Lee et al., 2019; Lee et al., 2018; Shiimori et al., 2018; Zhang et al., 2019). Here we provide a new mechanistic understanding of how Cas4-dependent PAM selection is achieved during CRISPR adaptation, and specifically how asymmetry of the adaptation complex drives the selection, processing and integration of PAM-compatible spacers. We present a model in which two independent subcomplexes, Cas4-Cas1 and Cas1-Cas2, selectively process the two 3' overhangs of a prespacer (Fig. 5). The interaction of Cas4 with the Cas1 integrase protein is central to the recognition and processing of PAM-containing prespacer substrates. Formation of this Cas4-Cas1 complex is mutually exclusive with formation of the Cas1-Cas2 complex, which may suggest distinct roles of both subcomplexes. We found that the Cas4-Cas1 subcomplex displays prespacer cleavage activity only on PAM-containing 3' overhangs. Cas4-Cas1 removes the PAM via endonuclease cleavage while Cas1-2 defines overhang trimming likely through host factor nucleases. Subsequently, Cas1-Cas2 initiates coupling of the non-PAM overhang to the leader-repeat junction followed by integration of the processed PAM-site at the repeat-spacer junction.

Our findings are consistent with previous studies that established the existence of two mutually exclusive Cas4-Cas1 and Cas1-2 complexes in type I-C CRISPR-Cas systems (Lee et al., 2019; Lee et al., 2018), and expand our understanding of the roles of these subcomplexes. Moreover, the RecB-domain mediated activity of Cas4 is dependent on the presence of Cas1, since Cas4 alone is not able to recognize and process the PAM sequence. This observation suggests that the Cas4-Cas1 interaction is essential for sequence specific recognition of the PAM. It remains to be determined whether the PAM sequence recognition domain is located within Cas4 or Cas1. Interestingly, PAM selection in the type I-E system is mediated by the C-terminal tail of

Cas1 (Kim et al., 2020), however, this C-terminal proportion is not conserved in the type I-D Cas1 protein. Future structural and biochemical studies will have to address how PAM selection is achieved.

Cas4-Cas1 did not display activity on non-PAM containing overhangs *in vitro*, however, processing of the non-PAM overhang was observed in our *in vivo* setup, suggesting that processing involves other non-Cas proteins. This finding is in line with the Cas4-deficient type I-E system in which host factors such as the ExoT and DnaQ-like exonucleases are required for processing both, PAM-containing and non-PAM overhangs. (Kim et al., 2020; Ramachandran et al., 2020). We propose that, in analogy to the *E. coli* type I-E system, 3'-5' exonucleases act as trimming factors for non-PAM 3' overhangs in the native *Synechocystis* PCC6803 host. The Cas1 protein of the type I-E system recognizes and protects the PAM from premature trimming, causing a delayed processing of the PAM end that ensures correct orientation (Kim et al., 2020). Upon activation, the Cas4-Cas1 complex sequence specifically removes the type I-D PAM, although incorrect processing was observed *in vivo* and *in vitro* that would result in single-nucleotide slipped spacers. Recently, it was observed in the type I-F system that slipped spacers increase primed adaptation which enhances the spacer diversity of the population (Jackson et al., 2019). Our results suggest the possibility that such erroneous PAM processing could promote the integration of slipped spacers and by extension, primed adaptation as found in other type I CRISPR-Cas systems (Nussenzweig and Marraffini, 2020).

Cryo-EM structures of the type I-C Cas4-Cas1-Cas2 complex revealed that the complex might undergo a conformational change (e.g. causing dissociation of Cas4 from the complex) in order to allow for Cas1 mediated integration of the PAM-end site of the spacer. Lee et al. showed that 50% of their Cas4-Cas1-Cas2 complex structures

lacked the Cas4 density on one site of the complex, resulting in an asymmetric complex. Our observation of two integrase complexes (Cas4-Cas1 and Cas1-Cas2) that are independently capable of at least half-site spacer integration points towards a similar asymmetrical organization of the full adaptation complex, in which Cas4-Cas1 is involved in PAM-site and Cas1-Cas2 in non-PAM site integration. By testing asymmetric spacer precursors, we demonstrate that integrase activity of the type I-D Cas4-Cas1 complex is potentially halted until integration of the non-PAM overhang has occurred. We propose a model for the type I-D system that relies on a delayed PAM-site integration by Cas4 in order to result in a correctly oriented spacer. In summary, we propose a mechanism in which two functionally independent complexes, Cas4-Cas1 and Cas1-Cas2, sequentially process and integrate prespacer substrates. This mechanism ensures correct spacer orientation as well as correct PAM-processing, thereby resulting in interference-proficient CRISPR adaptation.

Acknowledgements

We would like to thank Marre Niessen for early contributions. We thank Dr. Viktorija Globyte for critical reading of the manuscript. We thank Rob B.M. Koehorst for providing the *Synechocystis* PCC6803 strain.

Author contributions: S.N.K, C.A. and S.J.J.B. designed the experiments. S.N.K, A.C.H. and A.B. performed the experiments. S.N.K., C.A., A.B., A.J.H.R and S.J.J.B analyzed the data. S.N.K, C.A. and S.J.J.B. wrote the paper with input from all authors.

FUNDING: FOM (Projectruimte 15PR3188-2); Netherlands Organisation for Scientific Research [VICI VI.C.182.027]

Conflict of interest statement. None declared.

References

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

Barrangou, R. (2013). CRISPR-Cas systems and RNA-guided interference. WIREs RNA *4*, 267-278.

Brouns, S.J.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J.H., Snijders, A.P.L., 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.

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.

Gonzales, M.F., Brooks, T., Pukatzki, S.U., and Provenzano, D. (2013). Rapid Protocol for Preparation of Electrocompetent Escherichia coli and Vibrio cholerae. Journal of Visualized Experiments : JoVE, 50684.

Hou, Z., and Zhang, Y. (2018). Insights into a Mysterious CRISPR Adaptation Factor, Cas4. Mol Cell *70*, 757-758.

Hudaiberdiev, S., Shmakov, S., Wolf, Y.I., Terns, M.P., Makarova, K.S., and Koonin, E.V. (2017). Phylogenomics of Cas4 family nucleases. BMC Evolutionary Biology *17*, 232-232.

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 and Microbe *25*, 250-260.e254.

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*.

Jansen, R., Embden, J.D.A.V., Gaastra, W., and Schouls, L.M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. Molecular Microbiology *43*, 1565-1575.

Joo, C., and Ha, T. (2012). Labeling DNA (or RNA) for single-molecule FRET. Cold Spring Harb Protoc *2012*, 1005-1008.

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 Reports *22*, 3377-3384.

Kim, S., Loeff, L., Colombo, S., Jergic, S., Brouns, S.J.J., and Joo, C. (2020). Selective loading and processing of prespacers for precise CRISPR adaptation. Nature.

Lee, H., Dhingra, Y., and Sashital, D.G. (2019). The Cas4-Cas1-Cas2 complex mediates precise prespacer processing during CRISPR adaptation. eLife *8*, 1-84.

Lee, H., Zhou, Y., Taylor, D.W., and Sashital, D.G. (2018). Cas4-Dependent Prespacer Processing Ensures High-Fidelity Programming of CRISPR Arrays. Molecular Cell *70*, 48-59.e45.

Lemak, S., Beloglazova, N., Nocek, B., Skarina, T., Flick, R., Brown, G., Popovic, A., Joachimiak, A., Savchenko, A., and Yakunin, A.F. (2013). Toroidal Structure and DNA Cleavage by the CRISPR-Associated [4Fe-4S] Cluster Containing Cas4 Nuclease SSO0001 from Sulfolobus solfataricus. Journal of the American Chemical Society *135*, 17476-17487.

Lemak, S., Nocek, B., Beloglazova, N., Skarina, T., Flick, R., Brown, G., Joachimiak, A., Savchenko, A., and Yakunin, A.F. (2014). The CRISPR-associated Cas4 protein Pcal_0546 from Pyrobaculum calidifontis contains a [2Fe-2S] cluster: crystal structure and nuclease activity. Nucleic Acids Research *42*, 11144-11155.

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 Research *42*, 2483-2492.

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 Research *45*, 8978-8992.

McGinn, J., and Marraffini, L.A. (2019). Molecular mechanisms of CRISPR–Cas spacer acquisition. Nature Reviews Microbiology *17*, 7-12.

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

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. Nature Structural & Molecular Biology *21*, 528-534.

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

Nussenzweig, P.M., and Marraffini, L.A. (2020). Molecular Mechanisms of CRISPR-Cas Immunity in Bacteria. Annu Rev Genet *54*, 93-120.

Plagens, A., Tjaden, B., Hagemann, A., Randau, L., and Hensel, R. (2012). Characterization of the CRISPR/Cas Subtype I-A System of the Hyperthermophilic Crenarchaeon Thermoproteus tenax. Journal of Bacteriology *194*, 2491-2500.

Ramachandran, A., Summerville, L., Learn, B.A., DeBell, L., and Bailey, S. (2020). Processing and integration of functionally oriented prespacers in the Escherichia coli CRISPR system depends on bacterial host exonucleases. J Biol Chem *295*, 3403-3414.

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.

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. Molecular cell *70*, 814-824.e816.

Tahallah, N., Pinkse, M., Maier, C.S., and Heck, A.J. (2001). The effect of the source pressure on the abundance of ions of noncovalent protein assemblies in an electrospray ionization orthogonal time-of-flight instrument. Rapid Commun Mass Spectrom *15*, 596-601.

van den Heuvel, R.H., van Duijn, E., Mazon, H., Synowsky, S.A., Lorenzen, K., Versluis, C., Brouns, S.J., Langridge, D., van der Oost, J., Hoyes, J.*, et al.* (2006). Improving the performance of a quadrupole time-of-flight instrument for macromolecular mass spectrometry. Anal Chem *78*, 7473-7483.

van der Oost, J., Westra, E.R., Jackson, R.N., and Wiedenheft, B. (2014). Unravelling the structural and mechanistic basis of CRISPR–Cas systems. Nature Reviews Microbiology *12*, 479-492.

Yosef, I., Goren, M.G., and Qimron, U. (2012). Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucleic Acids Research *40*, 5569-5576.

Zhang, J., Kasciukovic, T., and White, M.F. (2012). The CRISPR Associated Protein Cas4 Is a 5′ to 3′ DNA Exonuclease with an Iron-Sulfur Cluster. PLoS ONE *7*, e47232-e47232.

Zhang, Z., Pan, S., Liu, T., Li, Y., and Peng, N. (2019). Cas4 Nucleases Can Effect Specific Integration of CRISPR Spacers. Journal of Bacteriology *201*.

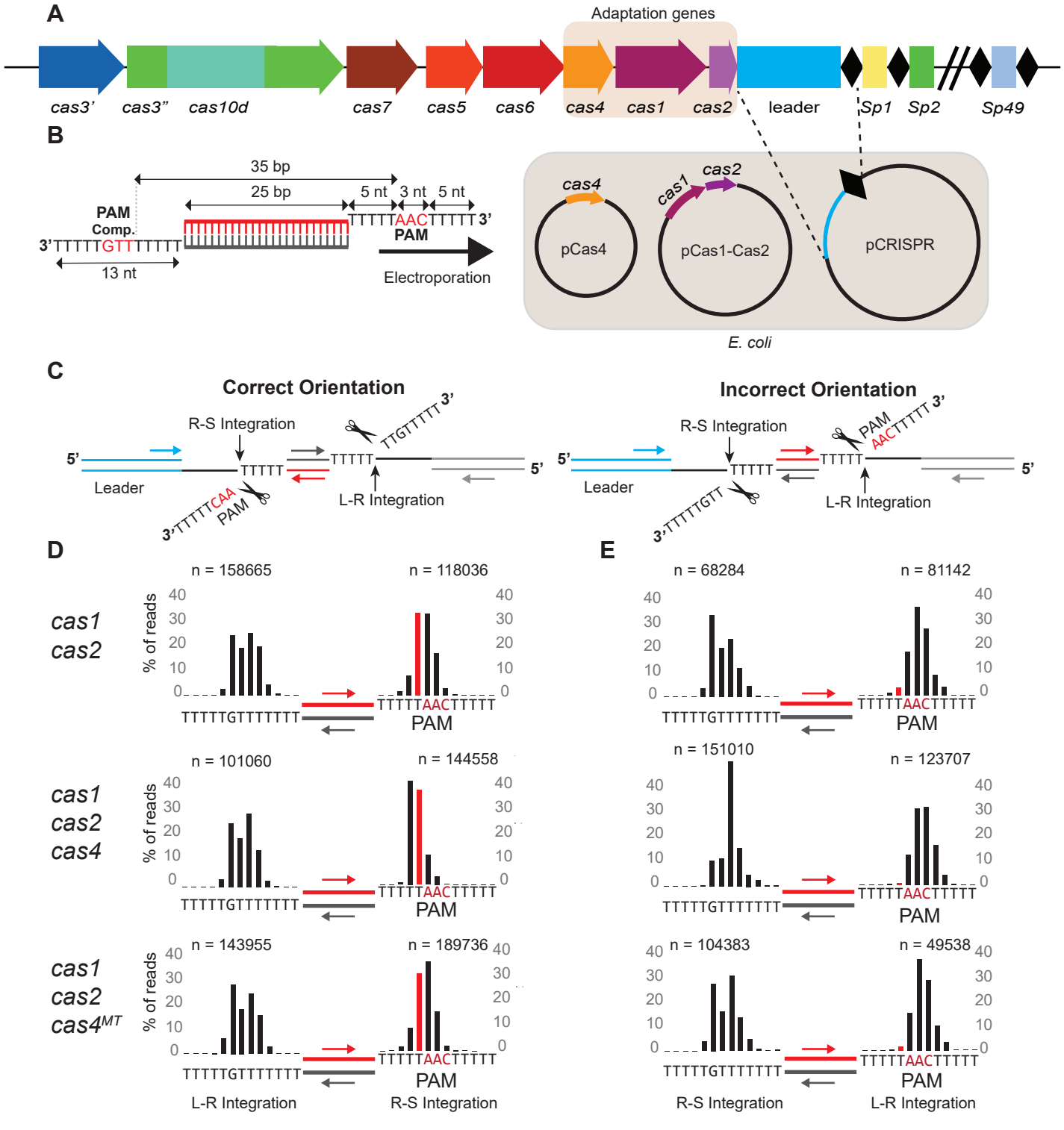

A
Adaptation genes
cas3'
cas3"
cas10d
cas7
cas5
cas6
cas4
cas1
cas2
leader
Sp1
Sp2
Sp49
B
35 bp
25 bp
5 nt
3 nt
5 nt
PAM
Comp.
3'TTTTTGTTTTTTT
TTTTTAACTTTTT 3'
PAM
13 nt
Electroporation
cas4
cas1
cas2
pCas4
pCas1-Cas2
pCRISPR
E. coli
C
Correct Orientation
R-S Integration
5'
Leader
TTTTT
TTGTTTTT 3'
L-R Integration
3'TTTTTCAA
PAM
Incorrect Orientation
R-S Integration
PAM
AACTTTTT 3'
L-R Integration
3'TTTTTGTT
D
n = 158665
n = 118036
cas1
cas2
% of reads
TTTTTGTTTTTTT
TTTTTAACTTTTT
PAM
n = 101060
n = 144558
cas1
cas2
cas4
n = 143955
n = 189736
cas1
cas2
cas4MT
L-R Integration
R-S Integration
E
n = 68284
n = 81142
n = 151010
n = 123707
n = 104383
n = 49538
R-S Integration
L-R Integration

Figure 1

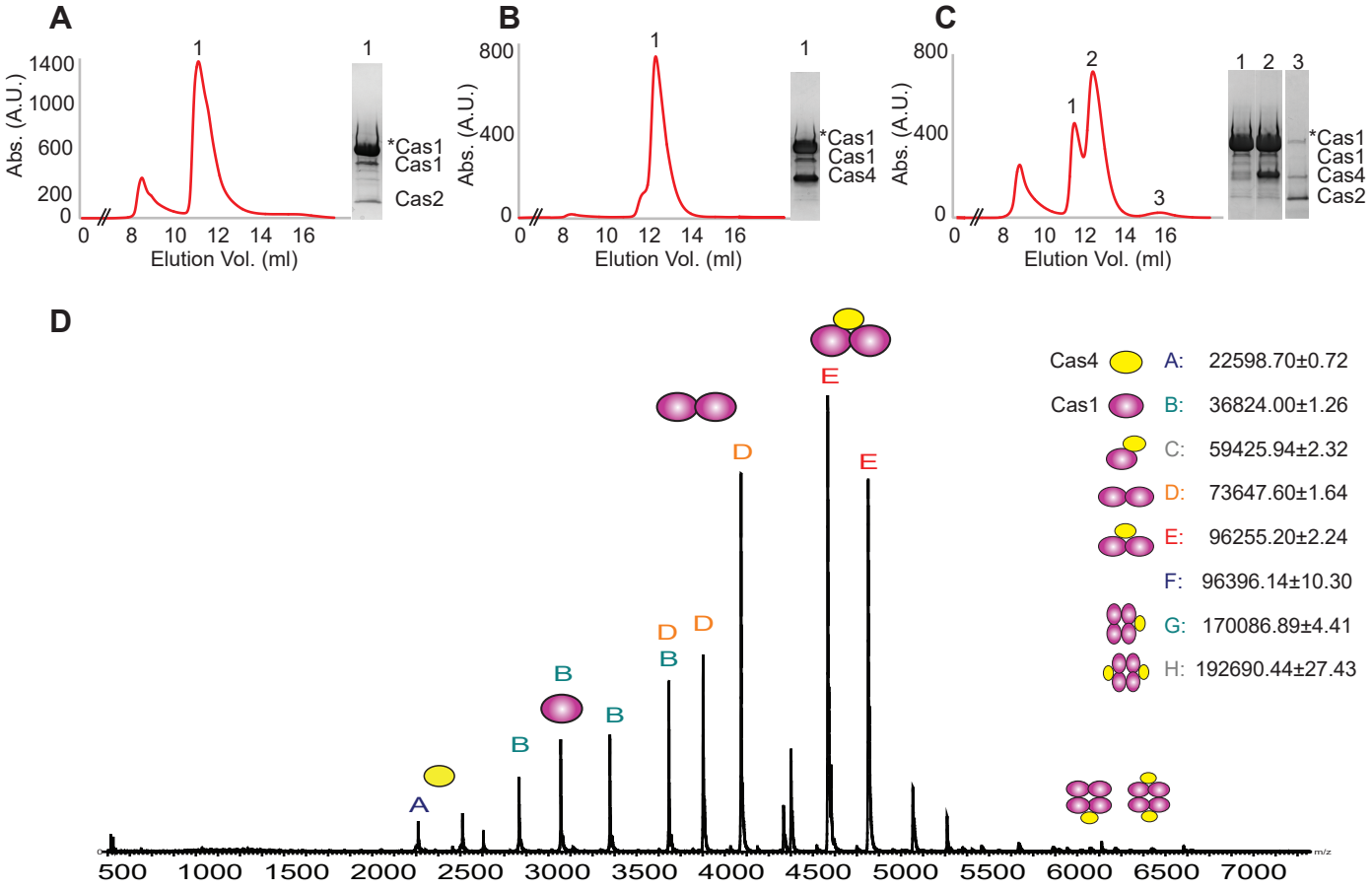
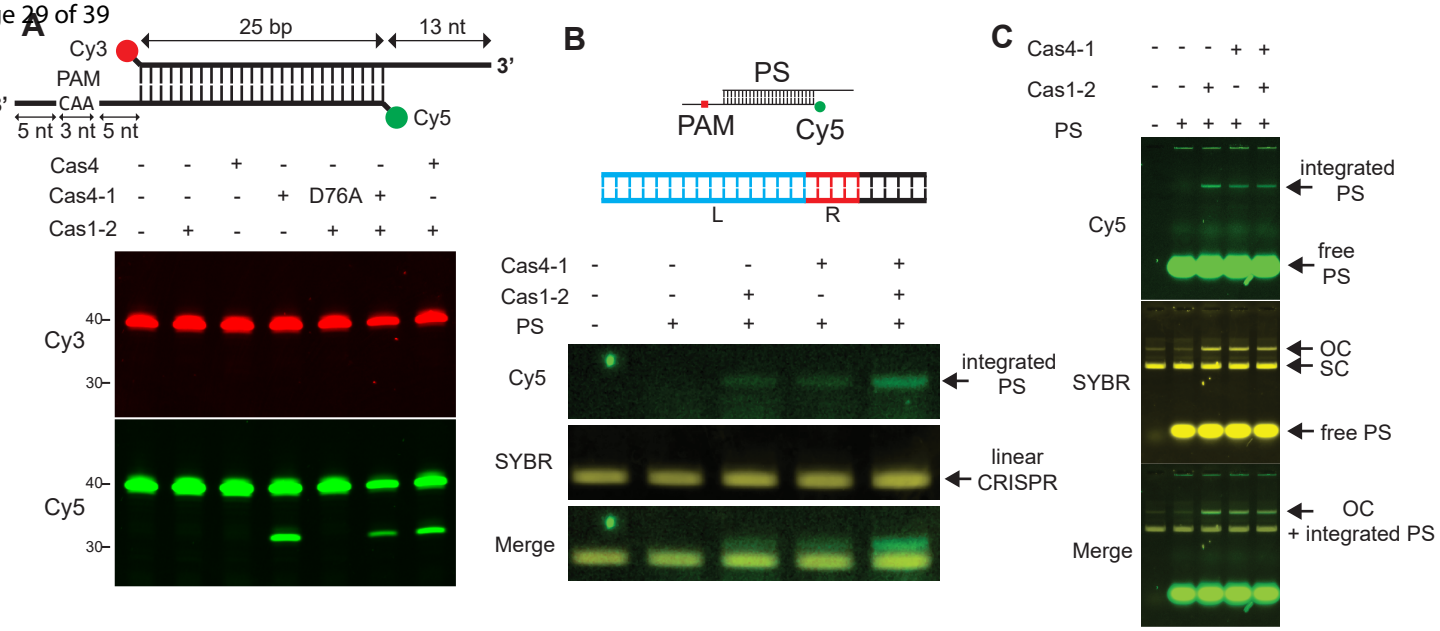
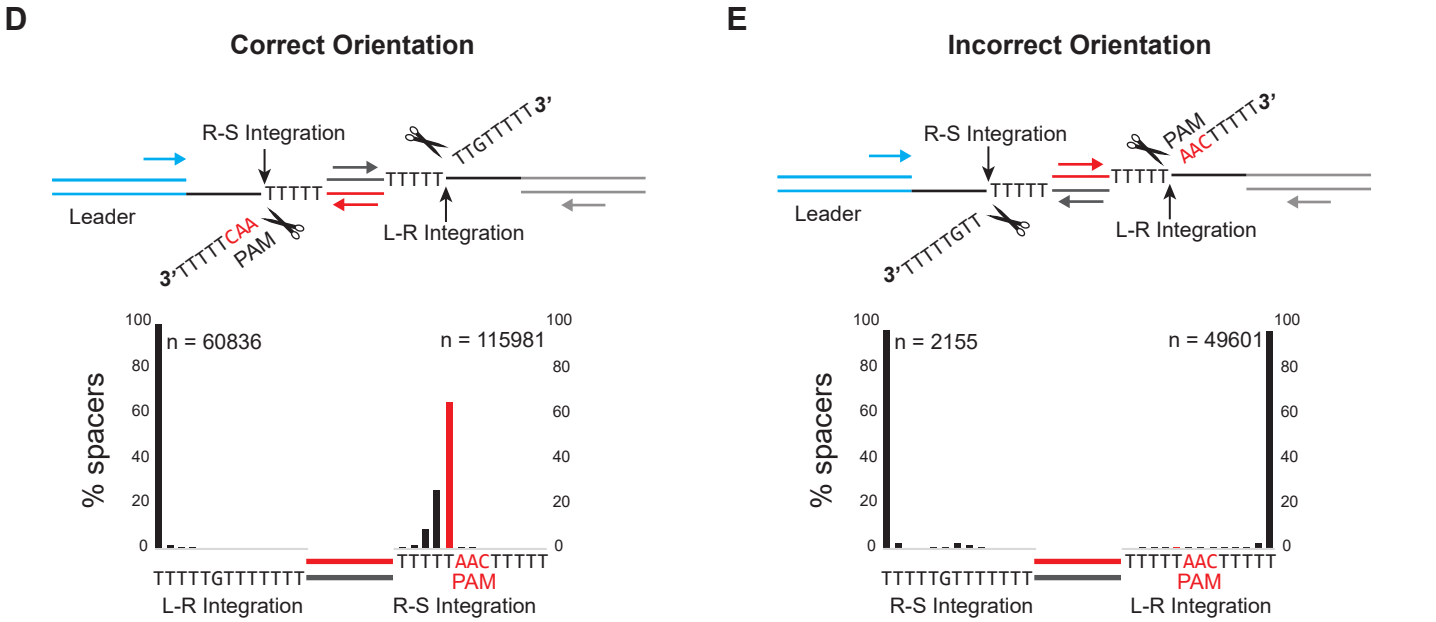

A
B
C
D
Abs. (A.U.)
Elution Vol. (ml)
*Cas1
Cas1
Cas2
Cas4
Cas4 A: 22598.70±0.72
Cas1 B: 36824.00±1.26
C: 59425.94±2.32
D: 73647.60±1.64
E: 96255.20±2.24
F: 96396.14±10.30
G: 170086.89±4.41
H: 192690.44±27.43
m/z

Figure 2


A
25 bp
13 nt
Cy3
PAM
CAA
Cy5
3'
5 nt 3 nt 5 nt
Cas4 - - + - - - +
Cas4-1 - - - + D76A + -
Cas1-2 - + - - + + +
Cy3
Cy5
40
30
B
PS
PAM
Cy5
L
R
Cas4-1 - - - + +
Cas1-2 - - + - +
PS - + + + +
Cy5
SYBR
Merge
integrated PS
linear CRISPR
C
Cas4-1 - - - + +
Cas1-2 - - + - +
PS - + + + +
Cy5
SYBR
Merge
integrated PS
free PS
OC
SC
free PS
OC
+ integrated PS


D
Correct Orientation
R-S Integration
TTTTT
TTGTTTTT 3'
Leader
3'TTTTTCAA
PAM
L-R Integration
n = 60836
n = 115981
% spacers
100
80
60
40
20
0
TTTTTGTTTTTTT
L-R Integration
TTTTTAACTTTTT
PAM
R-S Integration
E
Incorrect Orientation
R-S Integration
TTTTT
PAM
AACTTTTT 3'
Leader
3'TTTTTGTT
L-R Integration
n = 2155
n = 49601
% spacers
TTTTTGTTTTTTT
R-S Integration
TTTTTAACTTTTT
PAM
L-R Integration

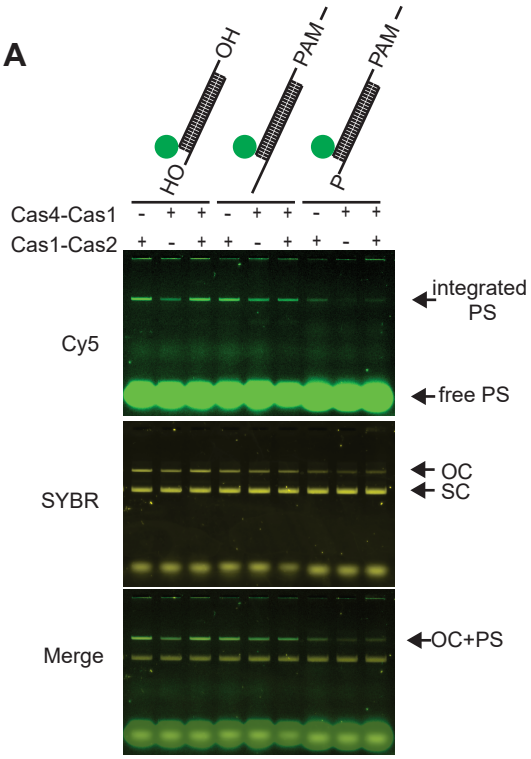

A
OH
HO
PAM
PAM
P
Cas4-Cas1
Cas1-Cas2
Cy5
SYBR
Merge
integrated PS
free PS
OC
SC
OC+PS

Figure 4

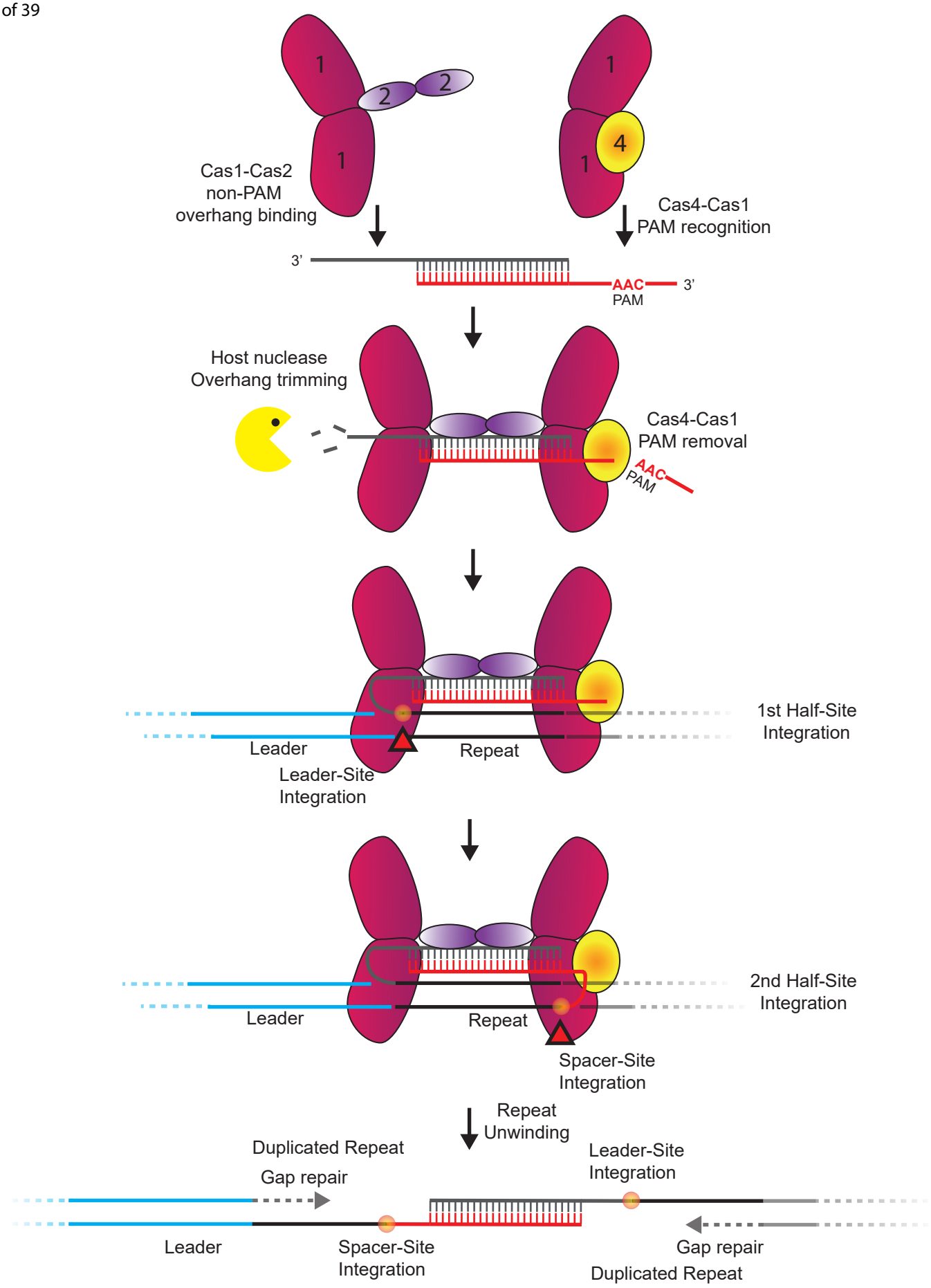

1
2
2
1
Cas1-Cas2
non-PAM
overhang binding
1
4
1
Cas4-Cas1
PAM recognition
3'
AAC
PAM
3'
Host nuclease
Overhang trimming
Cas4-Cas1
PAM removal
AAC
PAM
1st Half-Site
Integration
Leader
Repeat
Leader-Site
Integration
2nd Half-Site
Integration
Leader
Repeat
Spacer-Site
Integration
Repeat
Unwinding
Duplicated Repeat
Gap repair
Leader-Site
Integration
Leader
Spacer-Site
Integration
Gap repair
Duplicated Repeat

Figure 5

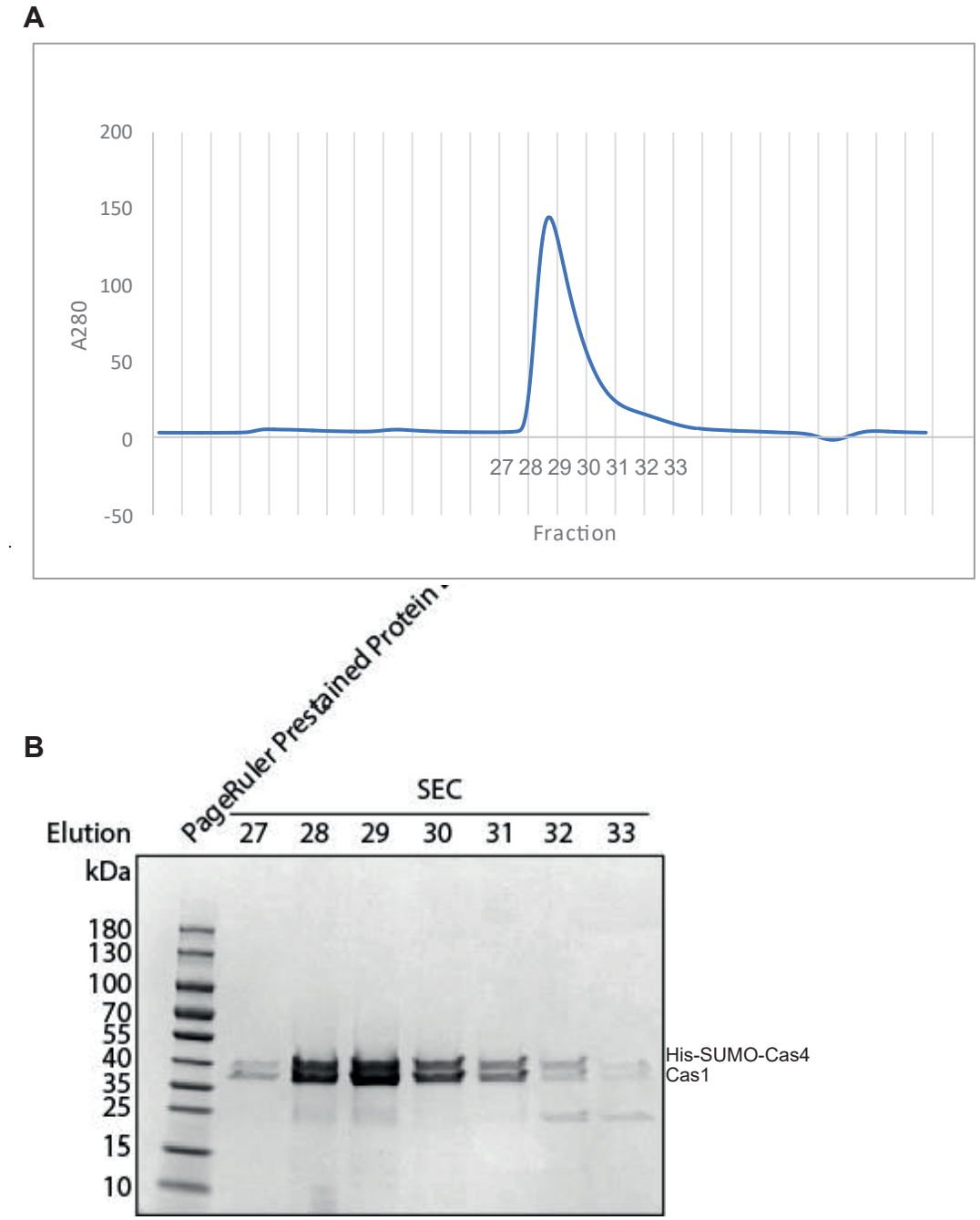
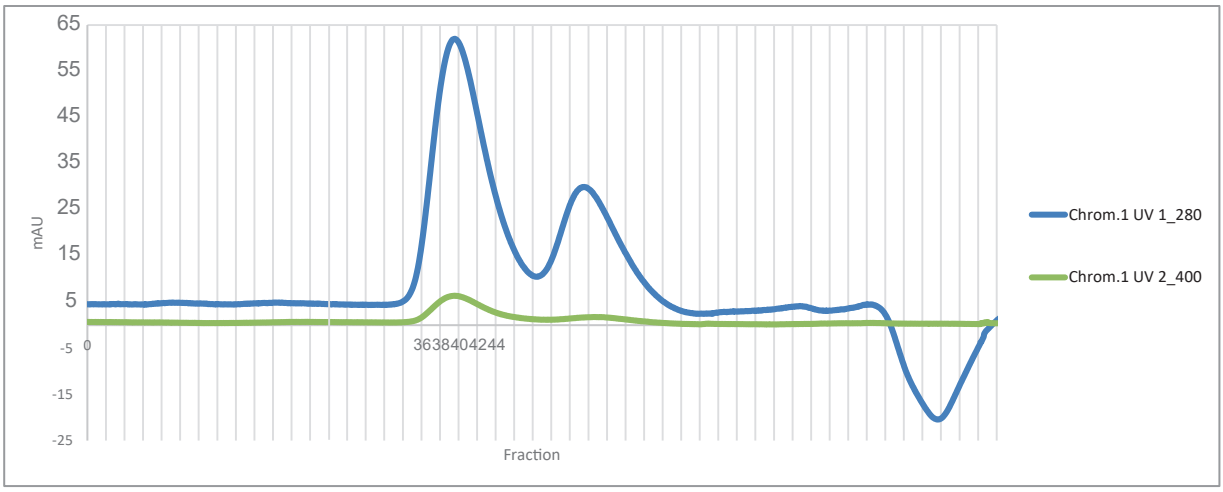

A
200
150
100
50
0
-50
A280
27 28 29 30 31 32 33
Fraction
B
PageRuler Prestained Protein
SEC
Elution
27
28
29
30
31
32
33
kDa
180
130
100
70
55
40
35
25
15
10
His-SUMO-Cas4
Cas1

Figure S1

A


65
55
45
35
25
15
5
-5
-15
-25
mAU
0
3638404244
Fraction
Chrom.1 UV 1_280
Chrom.1 UV 2_400

B

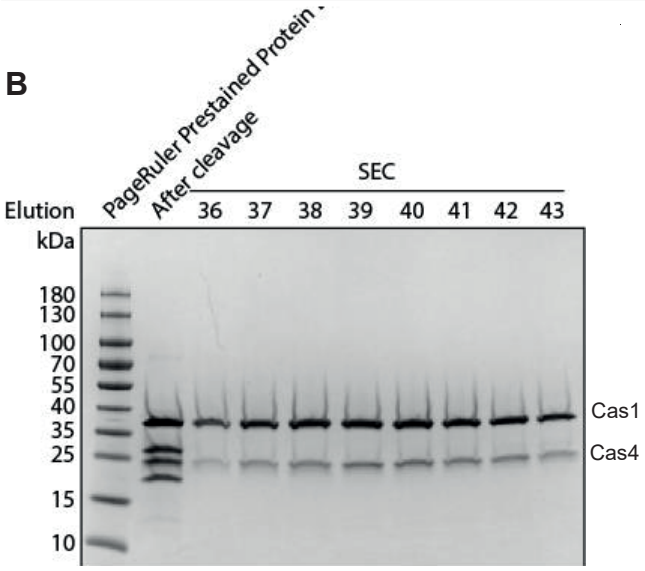

PageRuler Prestained Protein
After cleavage
SEC
Elution
36
37
38
39
40
41
42
43
kDa
180
130
100
70
55
40
35
25
15
10
Cas1
Cas4

Figure S2

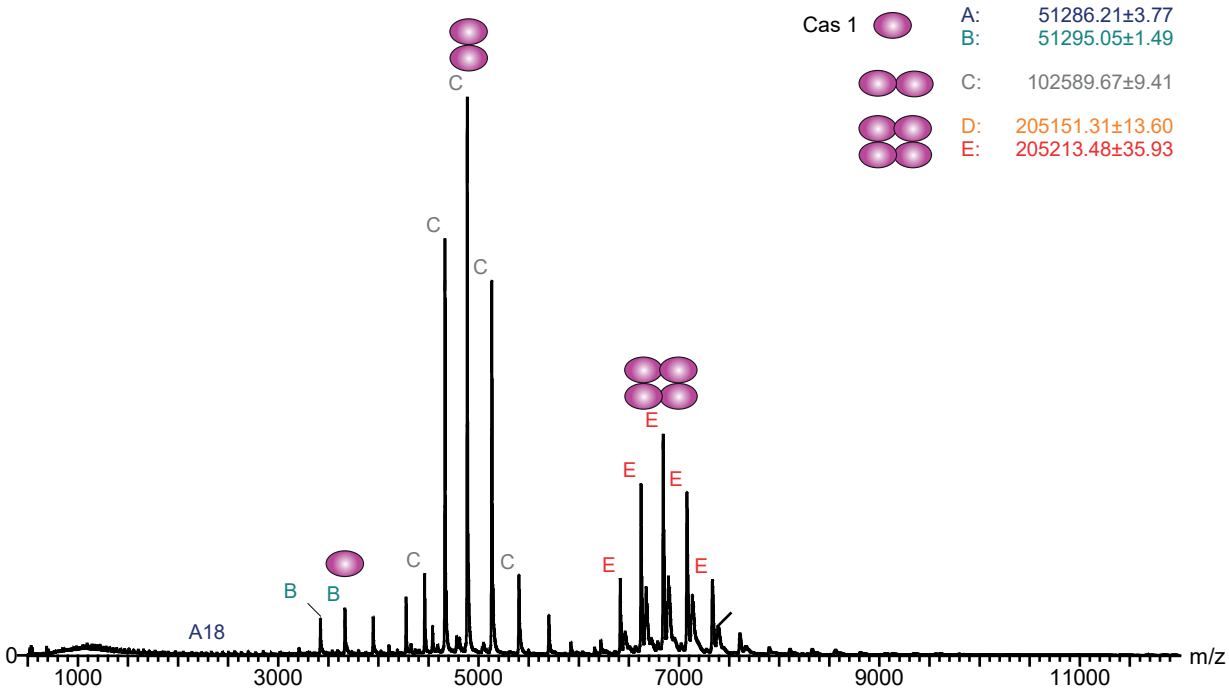

Cas 1
A: 51286.21±3.77
B: 51295.05±1.49
C: 102589.67±9.41
D: 205151.31±13.60
E: 205213.48±35.93
A18
B
B
C
C
C
C
E
E
E
E
E
0
1000
3000
5000
7000
9000
11000
m/z

Figure S3

Figure Legends

Figure 1 – Prespacer processing and half-site integration in vivo

A Genetic organization of the type I-D CRISPR-locus. Genes constituting the interference machinery are located upstream of the adaptation complex. The adaptation complex consisting of *cas4*, *cas1* and *cas2* is highlighted in purple. Downstream of *cas2* is the leader sequence followed by the type I-D array.

B Experimental design of spacer processing in vivo assay. Idealized pre-spacer substrates were electroporated into *E. coli* cells carrying the plasmid encoded minimalized type I-D array and expressing the adaptation genes *cas1* and *cas2*. The Cas4 protein was either omitted or co-expressed as the wild-type or D76A+K91A mutant protein. Half-site integration was assessed by PCR as depicted in **C**.

C PCR scheme allowing differentiation of spacer orientation, integration site and prespacer processing. PCR amplicons of correct (PAM overhang integrated at Repeat-Spacer junction) and incorrect spacer orientations and Leader-Repeat (L-R) or Repeat-Spacer (R-S) integration site were subjected to high-throughput sequencing.

D-E Overhang processing resulting from high-throughput sequencing of *in vivo* half-site integration PCR. Bar charts indicate the overhang nucleotide left after processing and integration as a percentage of the total number of sequenced spacer integration events. Red bars represent correctly trimmed PAM-containing 3’ overhangs. Integration was assessed in correct (D) or incorrect (E) spacer orientation resulting from either Cas1-Cas2 or Cas1-Cas2 co-expressed with Cas4 wild-type or Cas4 mutant (MT) background. n = number of sequenced integration events.

Figure 2 - Size exclusion chromatograms and SDS-PAGE analysis of peak fractions

A N-terminally tagged *Cas1 associates with untagged Cas2 in the absence of Cas4. Unnumbered peak contains protein aggregates.

B Complex formation of N-terminally tagged *Cas1 and untagged Cas4.

C Co-expression of *Cas1, Cas2 and Cas4. *Cas1 elutes separately (peak 1) from the Cas4-Cas1 complex (peak 2). Cas2 together with dissociated Cas1 and Cas4 elutes as a low molecular weight peak (peak 3). Unnumbered peak contains protein aggregates.

D Native Mass Spectrometry of Cas4-Cas1 complex as shown in B (native spectrum obtained after removal of His-SUMO tag from Cas1 by TEV protease cleavage (Fig. S2)). Cas4 monomers assemble with Cas1 dimers into a $Cas4_1$-$Cas1_2$ complex. Cas1 dimers are also frequently observed. Free monomers of Cas1 and Cas4 are less frequent in the measured sample.

Figure 3 – In vitro pre-spacer cleavage and integration

A Pre-spacer model substrate containing a 25 bp duplex region flanked by 13 nt 3’ overhangs incubated with different protein combinations. The non-PAM strand is 5’ Cy3 labelled and the PAM-containing strand 5’ Cy5 labelled. Cleavage of PAM containing 3’ overhang results in Cy5 labelled fragment of 30 nt. Protein samples consist of co-purified Cas4-Cas1 and Cas4D76A-Cas1 complexes, combined Cas4-Cas1 and Cas1-Cas2 complex and individually purified Cas4 in the presence of Cas1-Cas2.

B Integration of labelled pre-spacer (PS) into linear CRISPR DNA consisting of the type I-D leader sequence (L) and a single Repeat (R). Labelled spacer imaged via Cy5, total DNA via SYBR gold stain. Merge of Cy5 and SYBR gold channels indicates integration of Cy5 labelled pre-spacer resulting in a higher molecular weight band.

C Labelled pre-spacer (PS) integration into pCRISPR DNA. Similar to B, both adaption complexes facilitate integration into plasmid encoded CRISPR locus. Integration reaction is accompanied by nicking of supercoiled (SC) plasmid DNA, resulting in formation of open-circular (OC) plasmid conformation. Merge image of Cy5 and SYBR channels shows co-localization of OC plasmid species and Cy5 labelled spacer substrate.

D-E High-throughput sequencing of Cas4-Cas1 integrated pre-spacers. Reaction was performed similar to the assay shown in **C** using unlabeled pre-spacer DNA. OC plasmids were gel extracted followed by PCRs specific for the leader-repeat (L-R) and repeat-spacer (R-S) integration as well as correct and incorrect spacer orientation. Bar graphs represent the percentage of spacers with specific overhang length depending on integration site and orientation (D-correct; E-incorrect). n = number of sequenced integration events.

Figure 4 - Spacer overhang preferences of Cas1-Cas2 and Cas4-Cas1 complexes.

A - Integration activity with respect to 3' overhang requirements. Pre-spacer substrates were 5' Cy5 labelled in order to follow coupling to pCRISPR. Phosphorylated (P) 3' overhangs were used to block integration of one of the DNA strands.

Figure 5 – Model of Cas4-Cas1 and Cas1-Cas2 assisted spacer selection, processing and integration. Prespacers with long PAM- and non-PAM containing 3' overhangs are bound by Cas4-Cas1 (PAM overhang) and Cas1-Cas2 (non-PAM overhang). Following processing by host-factor nucleases, the non-PAM site of the spacer is integrated at the leader-repeat site (first half-site integration). Subsequently, the second half-site integration (spacer-site integration) of the

PAM-site occurs, likely orchestrated by release of the Cas4-processed overhang into the integrase site of Cas1. Unwinding of the repeat followed by gap-repair completes repeat duplication and full-site spacer integration.

Supplementary Figures Legends

Figure S1 – Related to Fig. 2B – Co-purification of His_6-SUMO-Cas4 and untagged Cas1. **A** Size exclusion chromatogram of His_6-SUMO-Cas4 and untagged Cas1. **B** SDS PAGE analysis of SEC purified proteins.

Figure S2 – Related to Fig. 2D – Co-purification of untagged Cas4 and untagged Cas1 after TEV-protease cleavage of His_6-SUMO-TEV tag. **A** Size exclusion chromatogram of Cas4 and Cas1 (A280 – total protein; A400 – Cas4 FeS-cluster). **B** SDS PAGE analysis of SEC purified proteins.

Figure S3 – Related to Fig 2A – Native Mass Spectrometry of Cas1-Cas2 complex as shown in Fig. 2A.

Table S1 – Plasmids used in this study

Name in this study	Name	Insert	Vector	Resistance	Source
pCas2	pTU084	Synechocystis PCC6803 Type I-D *cas2* (deltaCas1)	pET-T7	Amp	Kieper et al. (2018)
pCas1	pTU085	Synechocystis PCC6803 Type I-D *cas1* (deltaCas2)	pET-T7	Amp	Kieper et al. (2018)
pCas1	pTU092	Synechocystis PCC6803 Type I-D *cas1*	pET-T7	Spec	This study
pCas4^{D76A}	pTU086	Synechocystis PCC6803 Type I-D *cas4* (D76A)	pET-T7	Spec	Kieper et al. (2018)
pCas4$^{D76A+K91A}$	pTU411	Synechocystis PCC6803 Type I-D *cas4*(D76A+K91A)	pET-T7	Spec	This study
pCas4	pTU130	Synechocystis PCC6803 Type I-D *cas4*	pET-T7	Spec	Kieper et al. (2018)
pCRISPR	pTU134	Synechocystis PCC6803 Type I-D Leader-R-S1	pACYCDuet1	Cm	Kieper et al. (2018)
pCas1-2	pTU70	Synechocystis PCC6803 Type I-D *cas1-cas2*	pET-T7	Amp	Kieper et al. (2018)
pEmp	pTU116	NA	pET-T7	Spec	Addgene Plasmid #48329

Table S2 - Oligonucleotides used in this study

Name	Sequence	Description
BN015	CGTCCATGGGAAGTCATTCTTCAAATTTTGGC	Leader Fw
BN277	GTGGAATACGCAAAAGGC	cas4 mutagenesis K91A Fw
BN278	AGGAATTAATAAGCCATCACTTTC	cas4 mutagenesis K91A Rv
BN1398	GCTAGTTATTGCTCAGCGG	pCRISPR bb Rv
BN1711	GGAAGGTTTGCCAAAGTC	Leader Distal Half-Site Integration
BN1712	CTGTTCGACTTAAGCATTATGC	Leader Proximal Half-Site Integration
BN1713	ATCGACACCACCACG	OligoSpecific Primer Fw (PAM overhang)
BN1714	CGTGGTGGTGTCGAT	OligoSpecific Primer Rv (non-PAM overhang)
BN1763	CTACCATCGACACCACCACGCTGGCTTTTTAACTTTTT	25 nt duplex 13nt PAM 3' ovhng
BN1768	GCCAGCGTGGTGGTGTCGATGGTAGTTTTTTTGTTTTT	25 nt duplex 13nt RvC PAM 3' ovhng
BN1829	CTACCATCGACACCACCACGCTGGCTTTTTTTGTTTTT	25 nt duplex 13nt RvC PAM 3' ovhng (5' C6-Amino)
BN1830	GCCAGCGTGGTGGTGTCGATGGTAGTTTTTAACTTTTT	25 nt duplex 13nt PAM 3' ovhng (5' C6-Amino)