Mechanism for Cas4-assisted directional spacer acquisition in CRISPR–Cas

https://doi.org/10.1038/s41586-021-03951-z

Received: 22 December 2020

Accepted: 25 August 2021

Published online: 29 September 2021

Check for updates

Chunyi Hu[1,5], Cristóbal Almendros[2,3,5], Ki Hyun Nam[4], Ana Rita Costa[2,3], Jochem N. A. Vink[2,3], Anna C. Haagsma[2,3], Saket R. Bagde[1], Stan J. J. Brouns[2,3]✉ & Ailong Ke[1]✉

Prokaryotes adapt to challenges from mobile genetic elements by integrating spacers derived from foreign DNA in the CRISPR array[1]. Spacer insertion is carried out by the Cas1–Cas2 integrase complex[2–4]. A substantial fraction of CRISPR–Cas systems use a Fe–S cluster containing Cas4 nuclease to ensure that spacers are acquired from DNA flanked by a protospacer adjacent motif (PAM)[5,6] and inserted into the CRISPR array unidirectionally, so that the transcribed CRISPR RNA can guide target searching in a PAM-dependent manner. Here we provide a high-resolution mechanistic explanation for the Cas4-assisted PAM selection, spacer biogenesis and directional integration by type I-G CRISPR in *Geobacter sulfurreducens*, in which Cas4 is naturally fused with Cas1, forming Cas4/Cas1. During biogenesis, only DNA duplexes possessing a PAM-embedded 3′-overhang trigger Cas4/Cas1–Cas2 assembly. During this process, the PAM overhang is specifically recognized and sequestered, but is not cleaved by Cas4. This 'molecular constipation' prevents the PAM-side prespacer from participating in integration. Lacking such sequestration, the non-PAM overhang is trimmed by host nucleases and integrated to the leader-side CRISPR repeat. Half-integration subsequently triggers PAM cleavage and Cas4 dissociation, allowing spacer-side integration. Overall, the intricate molecular interaction between Cas4 and Cas1–Cas2 selects PAM-containing prespacers for integration and couples the timing of PAM processing with the stepwise integration to establish directionality.

Prokaryotes have a unique ability to acquire immunological memories against mobile genetic elements by integrating short fragments of DNA (spacers) between CRISPR repeats. The array of repeat-spacers is transcribed to generate guide RNAs that direct CRISPR effector complexes DNA or RNA targets for cleavage. DNA-targeting CRISPR–Cas systems further require the spacers to be acquired adjacent to the PAM. The PAM helps CRISPR RNA (crRNA)-guided complexes distinguish true targets from spacers in the CRISPR array, and thereby prevents lethal self-targeting. PAM also speeds up the target-searching process by markedly reducing the total number of candidate sites[7]. To ensure CRISPR spacers are derived only from PAM-flanking sequences, both class I (type I-A, I-B, I-C, I-D, I-G) and class II (type II-B, V-A, V-B) CRISPR–Cas systems further encode a dedicated CRISPR adaptation protein, Cas4, that works in conjunction with the core spacer-acquisition machinery comprising Cas1 and Cas2[2–4,8–13]. Early studies mainly showed that deletion of *cas4* impaired spacer acquisition in type I-B systems in *Haloarcula hispanica*[14] and type I-A in *Sulfolobus islandicus*[15]. More recent studies using type I-A in *Pyrococcus furiosus*[16], type I-D in *Synechocystis* sp.[17] and type I-G (previously known as I-U) in *G. sulfurreducens*[18] established a critical role for Cas4 in acquiring spacers with a functional PAM. Cas4 protein was found to contain a Fe–S cluster and to catalyse various exo- and endonuclease activities[19–21]. Recent work in I-C *Bacillus halodurans* has shown that Cas4 uses its nuclease activity to cleave PAM sequences in spacer precursors just before their integration in the CRISPR array[22,23]. Follow-up work showed that Cas4 forms a complex with a dimer of Cas1 and associates with Cas2 upon prespacer binding[22,23].

Results

Cas4 is a PAM-cleaving endonuclease

Geobacter sulfurreducens I-G CRISPR-Cas contains a highly active spacer-acquisition module, in which Cas4 is fused with Cas1[18] (Fig. 1a). This module acquires 34–40-base pair (bp)-long spacers for integration into the CRISPR locus in a PAM-dependent manner[18] (PAM code: 5′-TTN). To understand the prespacer processing and integration mechanisms, we electroporated prespacers of various sequence and structure compositions into *Escherichia coli* cells containing a *G. sulfurreducens cas4/cas1-cas2-CRISPR* genomic locus and analysed cells for newly acquired spacers using PCR and deep sequencing methods (Fig. 1b, c, Extended Data Fig. 1a). It has been hypothesized that *G. sulfurreducens* (*Gs*)Cas4/Cas1–Cas2 may preferentially integrate prespacers containing a 26-bp mid-duplex and 5-nucleotide (nt) 3′-overhangs[18,22]. Such prespacers were indeed robustly integrated in a single-stranded PAM (ss-PAM)-dependent fashion; prespacers lacking ss-PAM were not integrated (Fig. 1b). The context surrounding PAM also influenced the

[1]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, USA. [2]Department of Bionanoscience, Delft University of Technology, Delft, The Netherlands. [3]Kavli Institute of Nanoscience, Delft, The Netherlands. [4]Department of Life Science, Pohang University of Science and Technology, Pohang, Republic of Korea. [5]These authors contributed equally: Chunyi Hu, Cristóbal Almendros. ✉e-mail: stanbrouns@gmail.com; ailong.ke@cornell.edu

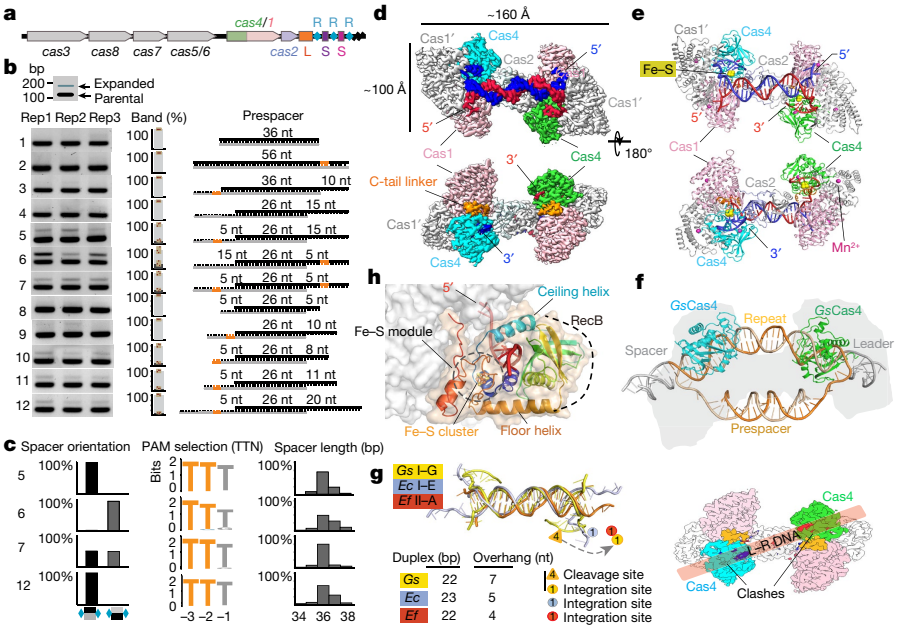


Fig. 1 | PAM-spacer acquisition and the dual-PAM prespacer-bound *Gs*Cas4/Cas1–Cas2 structure. a, Organization of the *G. sulfurreducens* CRISPR–Cas operon. L, leader; R, repeat; S, spacer. **b**, PAM dependency analysed using in vivo spacer-acquisition assay. $n = 3$ biologically independent assays detected by PCR are shown (Rep1–3), as well as relative percentages of expanded and non-expanded bands. Data are mean ± s.e.m. PAM is represented in orange. **c**, Deep-sequencing analysis of spacer orientation, length and PAM code for selected prespacers in **b**. PAM-1 appears conserved because a single prespacer was assayed. Spacer identities as assigned in **b** are marked at the left. Spacer orientation is illustrated underneath the *x* axis. **d**, **e**, Cryo-EM density (**d**) and structure (**e**) of the dual-PAM bound *Gs*Cas4/Cas1–Cas2 complex. **f**, Superposition with *E. faecalis* Cas1–Cas2 structure in full-integration state. Cas4 binding is incompatible with repeat-spacer docking into Cas1 integrase. **g**, Comparison of the 3′-overhang status among three prespacer-bound Cas1–Cas2 structures. The overhang is guided away from Cas1 and sequestered by Cas4 in *Gs*Cas4/Cas1–Cas2. *Ec*, *E. coli*. **h**, Organization of Cas4 structural elements around the PAM-containing 3′-overhang.

integration outcome. Whereas a ss-PAM 5 nt away from the mid-duplex were efficiently integrated, the same ss-PAM immediately adjacent to the mid-duplex or a double-stranded PAM in the middle of a duplex did not enable spacer integration (Fig. 1b). Dual-PAM-containing prespacers were integrated with scrambled directionality but a precise length distribution, whereas the single-PAM-containing prespacers were integrated directionally but with a 2–3-nt length distribution (Fig. 1c). These data suggest that *Gs*Cas4/Cas1-Cas2 preferentially recognizes prespacers containing a correctly spaced PAM in the 3′-overhang of a DNA duplex.

In biochemical reconstitutions (Extended Data Fig. 1b–k), the PAM-containing 3′-overhang of the prespacer was found to be specifically cleaved by recombinant *Gs*Cas4/Cas1-Cas2 complex, whereas the non-PAM 3′-overhang remained intact (Extended Data Fig. 1i). Cleavage was Mn^{2+}-dependent and took place precisely, if inefficiently, after the PAM (3′-$A_{-3}A_{-2}G_{-1}$↓; Extended Data Fig. 1h, i). Only about 5% of the PAM-containing overhang was processed after 1 h of incubation at 37 °C in a 50-fold excess of *Gs*Cas4/Cas1-Cas2 (Extended Data Fig. 1h). The underlying mechanism for the attenuated PAM processing became clear only after structural analysis. Notably, extended exposure to air induced promiscuous DNA cleavage (Extended Data Fig. 1j), probably owing to oxidation of the Fe–S cluster in Cas4. Various levels of oxidation may explain the spectrum of reported endo- and exonuclease activities of Cas4 in the literature[19–23].

Dual-PAM prespacer–Cas4/Cas1–Cas2 structure

Whereas a weak interaction was detected between *Gs*Cas4/Cas1 and *Gs*Cas2, formation of a functional complex required the presence of a prespacer. A dual- or single-PAM-containing prespacer led to stable higher-order complex formation, as revealed by size-exclusion chromatography (SEC) and electron microscopy analyses; a PAM-less prespacer was inefficient for complex formation (Extended Data Fig. 1g, k, l). The dual-PAM prespacer-bound *Gs*Cas4/Cas1–Cas2 complex was especially homogeneous, and its single-particle reconstruction reached 3.23 Å in resolution, revealing structural details that were not seen in previous studies[22] (Extended Data Figs. 2, 3). The $Cas1_4$–$Cas2_2$ integrase core assumes its characteristic dumbbell shape—the Cas2 dimer constitutes the central handle, and two Cas1 dimers constitute the two distal weights (Fig. 1d, e). In each dimer, only one Cas1 participates in spacer integration, and the other has structural roles. The architecture and interfaces are more consistent with *Enterococcus faecalis* $Cas1_4$–$Cas2_2$ than with *E. coli* $Cas1_4$–$Cas2_2$[10,12] (Extended Data Fig. 2b–d). Cas1–Cas2 was found to specify a 22-bp rather than 26-bp mid-duplex as defined by the integration assay—an additional two base-pairs are unwound from each end. Indeed, prespacers containing a 22-bp mid-duplex integrated as efficiently as the 26-bp version in various assays (Extended Data Fig. 2b–f). We predict that Cas1–Cas2 in different CRISPR systems are likely to share a preference for a 22-bp-long mid-duplex but specify an idiosyncratic 3′-overhang length in the prespacer[11,12] (Fig. 1g).

Among the four fused Cas4s, only the two PAM-engaging ones were resolved in the electron microscopy density; the other two Cas4s fused to the catalytic Cas1s were presumably too mobile (Fig. 1d, e). As the Cas4/Cas1 fusion does not alter the dynamic nature of the Cas4–Cas1–Cas2 interaction, the mechanistic insights from this study should apply to all Cas4 systems. This Cas4 structure aligns well with those of the standalone Cas4 proteins[19,20] and the nuclease domains in helicase–nuclease fusion proteins AddAB–AdnAB, RecBCD and eukaryotic Dna2 (Extended Data Fig. 5). Cas4 organizes its structural modules to form a narrow passage for the PAM-containing 3′-overhang. Its N-terminal α-helical floor connects to the ceiling helix on the top, which reaches overhead to the RecB nuclease centre on the opposite side, which then weaves back through the floor helix, and the remaining C-terminal region assembles with the N-terminal helical region to form the Fe–S cluster module, a hallmark of all Cas4 nucleases (Fig. 1h).

Notably, the Cas4 interface on Cas1–Cas2 overlaps with that of the leader-repeat DNA for spacer integration[10,12] (Fig. 1f). Cas4 binding therefore sterically blocks integration at the PAM-side Cas1. Cas4 contacts Cas1s through an extensive interface—many interface residues are conserved (Extended Data Figs. 4a–c, 5b). However, it is difficult to identify key interface residues that are universally conserved across all Cas4 branches. There may be evolutionary pressure to maintain idiosyncratic Cas4 and Cas1–Cas2 interactions to avoid crosstalk among coexisting CRISPR systems. If true, this scheme would be analogous to the highly selective binding relationship between Cas3 and Cascade[24].

Cas4-mediated PAM recognition

Despite extensive studies, the PAM recognition and cleavage mechanisms inside Cas4/Cas1–Cas2 remain unresolved. This electron microscopy structure brings such mechanisms into focus. The substrate-binding groove in Cas4 aligns with that in Cas1 to form a continuous 3′-overhang-binding groove. The 11-nt 3′-overhang (5′-$dA_7C_6T_5T_4T_3T_2T_1G_{-1}A_{-2}A_{-3}T_{-4}$) travels deep into the groove, protected from random nuclease cleavage. Nucleotides 1–4 travel along the previously described path towards the Cas1 active site[8,10,12]. However, nucleotides 5–11 move towards Cas4 (Fig. 1d, e), travelling through the RecB nuclease module and into a narrow passage, where PAM recognition takes place (Fig. 2a). Two hydrophobic residues, F35 and Y21, interdigitate into the single-stranded DNA (ssDNA) before and after the narrow passage, forming molecular ratchets that cage the di-deoxyadenosine PAM ($3'$-$A_{-3}A_{-2}$) in the passage (Fig. 2b). They probably enforce a ratcheting motion to slowly thread the 3′-overhang through. Inside the narrow passage, the edges of A_{-2} and A_{-3} are surrounded by hydrophobic or long sidechain residues (R14, M29, L25, L192, E117, N17 and C190) that probe nucleotides for shape complementarity. Deoxyguanosines would not fit in the same cage because their exocyclic N2 amines would cause steric clash; whereas the smaller pyrimidines may slip through without a chance to establish favourable contacts. Two Cas4 residues establish polar contacts with PAM: E18 makes bidentate hydrogen-bonding interactions with A_{-2} and A_{-3}, and S191 forms a hydrogen bond with A_{-2} (Fig. 2b). They probably contribute substantially to the PAM specificity. Consistent with the in vivo data[18], there is no sequence-specific recognition of the first residue of PAM, G_{-1}. This nucleotide is excluded from the PAM-recognition box and points towards the solvent (Fig. 2a, b).

Because Cas4 is responsible for PAM selection in a large fraction of CRISPR systems, we attempted to rationalize the PAM code in other CRISPR systems. Structure-guided mutagenesis was carried out to switch the PAM specificity of *Gs*Cas4 to that of *P. furiosus* (*Pf*)Cas4. *Pf*Cas4 shares 17% sequence identity with *Gs*Cas4 and specifies a 5′-CCN PAM (3′-GGN in the overhang). We substituted the two sequence-specific PAM-contacting residues in *Gs*Cas4 to their counterparts in *Pf*Cas4. In single substitutions, S191A retained *Gs*-PAM specificity; cleavage activity was slightly compromised. E18Y lost sequence-specific cleavage activity towards both PAMs and cleaved ssDNA distributively. Notably, the double substitution led to a cleavage preference for *Pf*-PAM on a distributive cleavage background. These results suggest that E18 has a more important role than S191 in PAM recognition (Extended Data Fig. 4e, f) . However, this partial success in switching PAM specificity did not further extend into in vivo spacer-acquisition assays, which put further demand on complex stability and PAM-cleavage timing. While E18Y/S191A Cas4 showed compromised *Gs*-PAM–prespacer integration, it was unable to support *Pf*-PAM–prespacer integration (Extended Data Fig. 4g). These results suggest that while the hydrogen-bonding interactions are important, a substantial portion of the PAM specificity is likely to be conferred by the peripheral residues mediating hydrophobic interactions.

Next, we used bioinformatics to establish a correlation between structural features in Cas4 and PAM sequence variations. A phylogenetic tree (Fig. 2c) was generated based on the alignment of Cas4s for which we could reliably couple PAM code with clades of Cas4s[25]. We expected that residues crucial for PAM selection would be conserved

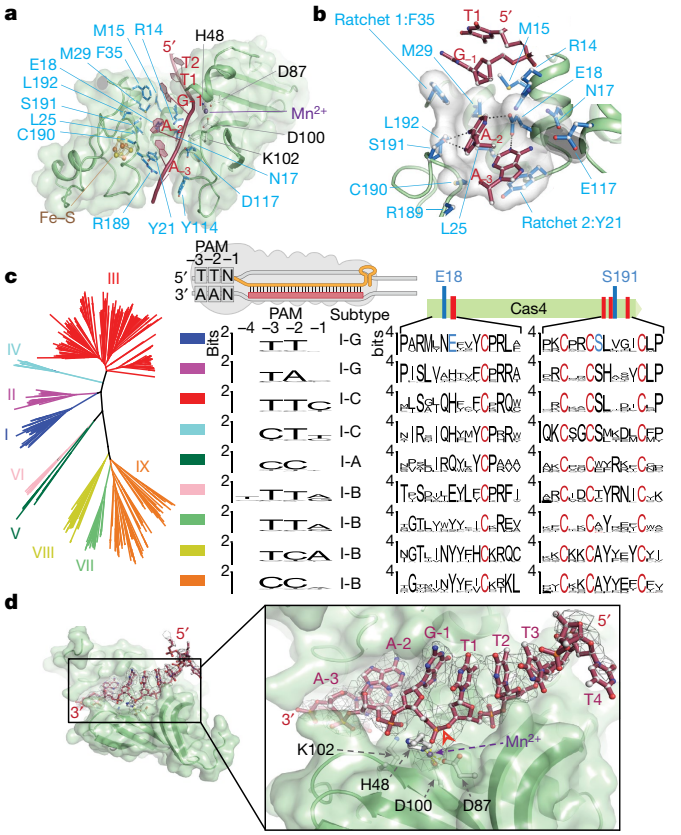


Fig. 2 | Cas4-mediated PAM recognition delays overhang cleavage. **a**, Cross-section of the narrow passage that sequesters the PAM-containing 3′-overhang. **b**, PAM ($A_{-2}A_{-3}$) is surrounded by Van der Waals interactions that probe for shape complementarity, and by sequence-specific hydrogen-bonding interactions from E18 and S191. **c**, Correlations between PAM code in Cas4-containing CRISPR systems and the recognition motif consensus in Cas4. **d**, Arrangement in the Cas4 nuclease centre. Cryo-EM density suggests the PAM-containing overhang is not cleaved. Red arrowhead indicates the labile bond.

within the clades, but differ between groups selecting a different PAM (Fig. 2c). Structure-defined E18 is one such discriminant residue because it is highly conserved among type I-G Cas4s specifying TTN PAMs and among type I-B Cas4s specifying a TTA or TTG PAM. S191 is not a discriminant residue as it was also found in Type I-G Cas4s specifying TAN PAMs. However, the highly conserved neighbouring residue, L192, was exclusively found in Cas4 groups specifying T_{-2} in PAM, including the distant Type I-C Cas4s that specify either TTC or CTT. Therefore, the presence of L192 in Cas4 is a good predictor of PAM-T_{-2}. Similarly, informatics identified R14 and L25 as good predictors of T_{-2}. The reverse argument is not necessarily true, as there is likely to be more than one evolutionary solution for Cas4 to specify a particular PAM.

PAM recognition prevents integration

The most important mechanistic insight from the dual-PAM structure is the observation that the PAM-containing 3′-overhang is recognized, sequestered, but not cleaved by Cas4 (Fig. 2c). The labile phosphate of G_{-1} is correctly positioned into the active site, which consists a DEK motif (D87, D100 and K102) and a histidine residue (H48), all of which are highly conserved among Cas4 and RecB family of nucleases. These residues coordinate a catalytic metal ion, presumably Mn^{2+}, which is shown by the electron microscopy density to be tightly coordinated to the scissile phosphate. In the AdnAB structure, this type of active site configuration was shown to cleave DNA efficiently[26]. However, in the case of Cas4, the electron microscopy density clearly argues for an

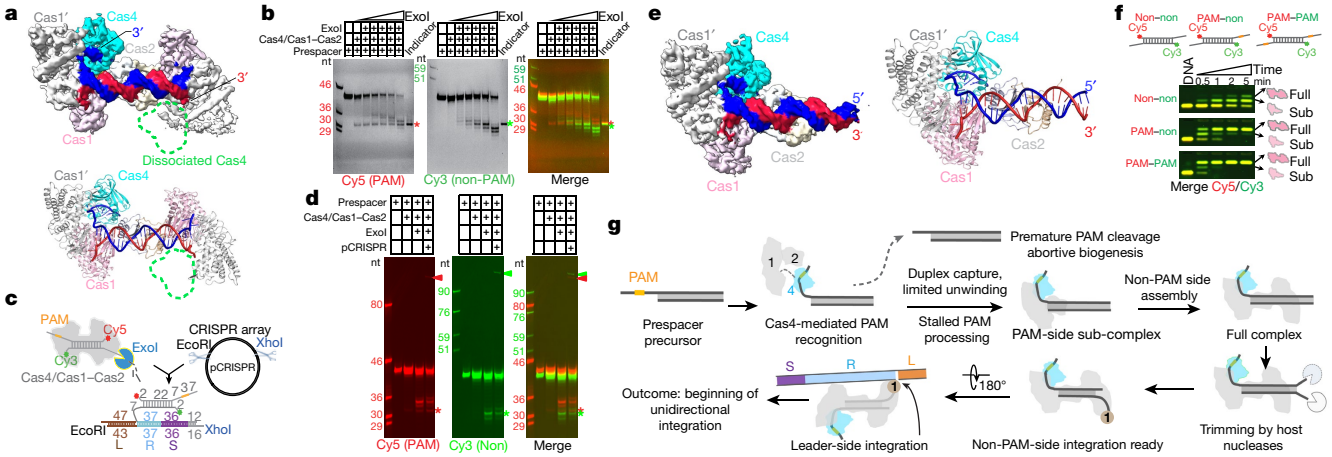


Fig. 3 | Mechanistic insights from the single-PAM prespacer-bound *Gs*Cas4/Cas1–Cas2 structure. a, Cryo-EM density and structure of the single-PAM prespacer-bound *Gs*Cas4/Cas1–Cas2 complex. Cas4 is absent from the non-PAM side. **b**, ExoI is capable of trimming the non-PAM overhang to the optimal length for integration. The PAM side is protected by Cas4. **c**, In vitro integration assay setup and the expected readout. If in the correct orientation, the Cy3 chain should be 2 + 22 + 7 + 37 + 36 + 12 = 116 nt in length. **d**, Non-PAM overhang is unidirectionally integrated to the leader-proximal end of the CRISPR repeat after ExoI trimming. Green and red arrows indicate integrated prespacer strands. Top to the leader-side target, bottom to the spacer side. **e**, Cryo-EM density and structure of a sub-complex. The Cas4/Cas1 dimer is missing from the non-PAM side. **f**, EMSA showing Cas4/Cas1–Cas2 is assembled sequentially and preferentially on PAM-containing prespacers. Non, non-PAM. **g**, A mechanistic model explaining Cas4-dependent prespacer biogenesis and directional integration.

intact DNA substrate at the active site (Fig. 2c). which was subsequently confirmed by denaturing PAGE (Extended Data Fig. 4d). The exact cleavage inhibition mechanism in Cas4 will require a more focused analysis in the future. Among the many mechanistic possibilities, we speculate that inhibition might be caused by the sub-optimally placed K102 residue, an essential catalytic residue in the DEK motif[18]. Rather than pointing towards the labile phosphate, K102 is twisted away by the residing β-strand. A minor conformational change in Cas4 may reorient K102 to participate in PAM cleavage. Without PAM cleavage, Cas4 is trapped in place and the adjacent integration centre is blocked. This structural observation agrees with the directionality requirement for the spacer in type I CRISPR systems.

Directional spacer integration reconstituted

Next, to investigate the status of the non-PAM 3′-overhang, we determined the cryo-electron microscopy (cryo-EM) structure of the *Gs*Cas4/Cas1–Cas2 complex with a single-PAM containing prespacer. We obtained an asymmetric full-complex structure at 3.57 Å resolution, and a 3.56 Å assembly intermediate (Fig. 3). Whereas Cas4 docks onto the PAM side of *Gs*Cas4/Cas1–Cas2, 82.5% of the particles do not contain a docked Cas4 at the non-PAM side (Fig. 3a, Extended Data Fig. 6); the remaining 17.5% contain a docked Cas4 at the non-PAM side, as evidenced by the weak densities. However, the non-PAM overhang is not retained inside Cas4 (Extended Data Fig. 6c). In both cases, the non-PAM side Cas4/Cas1 dimer density is weaker than the PAM-side counterpart, owing to a hinge motion around the non-catalytic Cas1. Only the first four nucleotides of the non-PAM 3′-overhang can be traced in the density, along a similar path as in the PAM-side (Extended Data Fig. 6c). Because the non-PAM overhang lacks Cas4 protection, we reasoned that it may be trimmed to the optimal overhang length by host nucleases, then captured by the nearby Cas1 and preferentially integrated to the leader-repeat DNA. This host nuclease-assisted integration mechanism would lead to a fixed spacer directionality that is consistent with the CRISPR biology. We tested this mechanistic model directly. Indeed, *E. coli* SbcB (ExoI) protein could trim the non-PAM 3′-overhang to the preferred length of around 7 nt (Fig. 4b). Even the distributive cleavage pattern was categorically consistent with the spacer length distribution in the *G. sulfurreducens* CRISPR system[18] (Fig. 1c). In the same reaction, the PAM-side 3′-overhang was protected by the footprint of Cas4 (Fig. 3b). Next, we established an in vitro integration assay to test whether the ExoI-trimmed prespacer can be integrated unidirectionally. An obstacle to this effort is that although *Gs*Cas4/Cas1–Cas2 readily integrated prespacers into a negatively supercoiled leader-repeat-containing plasmid, it did not do so on a linear double-stranded DNA (Extended Data Fig. 7a–d). This behaviour is similar to that of *E. coli* Cas1–Cas2, which was later shown to rely on the host integration factor (IHF) to integrate into a linear target[27]. Given this limitation, to resolve the integration directionality, we first integrated the fluorescently labelled prespacer into plasmid, then restriction digested out the leader-repeat region to determine the integration directionality on the basis of the product size by denaturing polyacrylamide gel electrophoresis (Extended Data Fig. 7c–f). In control experiments, we verified the preference of *Gs*Cas4/Cas1–Cas2s to integrate first into the leader-proximal side (Extended Data Fig. 7e, f). We went on to demonstrate that ExoI trimming enabled the non-PAM side of the prespacer to specifically integrate into the leader-proximal side of the repeat (Fig. 3c, d). This pattern is in agreement with the observed spacer directionality in the *G. sulfurreducens* CRISPR array.

Structural basis for prespacer biogenesis

The single-PAM cryo-EM reconstruction further captured an important functional state that corresponds to a prespacer-biogenesis intermediate. In this 3.6 Å structure, the PAM-side arrangement is essentially the same and the mid-duplex is protected by a Cas2 dimer, however, the non-PAM side lacks the protection from $(Cas4/Cas1)_2$ (Fig 3e, Extended Data Fig. 6). This structure raises the mechanistic possibility that components of the integration complex assemble onto prespacer in a stepwise manner. Indeed, in time-course and concentration-titration-based electrophoretic mobility shift assays (EMSA), the *Gs*Cas4/Cas1–Cas2 integrase was found to assemble in a stepwise fashion, and the PAM-containing overhang strongly promoted the assembly of the full complex (Fig 3f, Extended Data Fig. 7g). Collectively, these structural snapshots provide the necessary temporal resolution of prespacer biogenesis. We conclude that the $(Cas4/Cas1)_2$–$Cas2_2$ sub-complex is capable of searching for precursor DNA with a PAM-containing 3′-overhang. Binding of such precursor triggers enzymatic stalling in Cas4 and recruits a second $(Cas4/Cas1)_2$ complex to the opposite side, leading to the formation of an

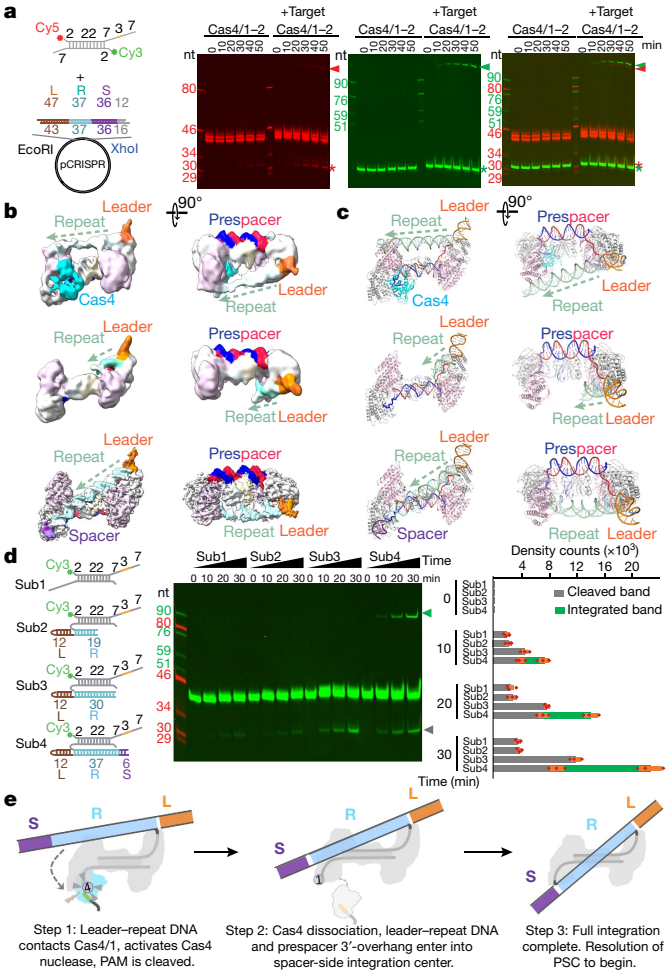


Fig. 4 | Structural basis for integration-coupled PAM cleavage by Cas4. **a**, Time-course experiments showing half-integration stimulates PAM cleavage by Cas4, enabling full integration. Green and red arrows indicate integrated prespacer strands. Top to the leader-side, bottom to the spacer side. Cas4/1–2, Cas4/Cas1–Cas2. **b**, **c**, Three cryo-EM reconstructions (**b**) and corresponding structures (**c**) captured from Cas4/Cas1–Cas2 incubated with half-integration-mimicking substrate. They depict the initial blockage of spacer-side integration by PAM-engaging Cas4 (top), the Cas4 dissociation after PAM cleavage (middle) and the full-integration state (bottom). Resolutions of the structures are 5.83, 5.76 and 3.81 Å, respectively. **d**, Time-course experiments showing that PAM cleavage is stimulated when leader-repeat DNA contacts the spacer-side Cas4/Cas1. Left, substrate design; middle, urea–PAGE; right, quantification of PAM cleavage and integration bands. Data were collected from $n = 3$ biologically independent experiments and presented as mean ± s.e.m. PSC, post-synaptic complex. **e**, Mechanistic model depicting the coupling between half integration, PAM cleavage, Cas4 dissociation and full integration.

integration-competent full $(Cas4/Cas1)_4$–$Cas2_2$ complex. The stepwise assembly process provides a quality control mechanism to selectively recruit PAM-containing precursors for further processing and integration (Fig. 3g, Supplementary Video 1). The length of the precursor duplex is probably longer than the duplex length preferred by $Cas1_4$–$Cas2_2$. A previous study explored this scenario and found that the host nucleases are capable of trimming the duplex and overhangs to optimal prespacer specifications as defined by the $Cas1_4$–$Cas2_2$ footprint[11].

Half-integration triggers PAM cleavage

Having established that Cas4 defines the spacer directionality by blocking the PAM-side integrase centre before integration, we next probed into the mechanism that relieves this blockage after half-integration, as the PAM-side prespacer needs to be processed and integrated to the opposite side of the CRISPR repeat to complete full integration. We hypothesized that the half-integration itself may stimulate PAM cleavage and Cas4 dissociation. To test this, we complexed *Gs*Cas4/Cas1–Cas2 to the half-integrated state using an annealed prespacer and leader-repeat DNA that mimics the half-integration product[10], and monitored the extent of PAM processing and half- to full-integration transition under different conditions (Extended Data Fig. 8a–j). Indeed, half-integration led to faster and greater extent of PAM cleavage, and full integration quickly followed (Fig. 4a, Extended Data Fig. 8b). PAM cleavage was much slower and weaker when the leader-repeat DNA was absent in the control condition (Fig. 4a).

To understand the structural basis for the observed mechanistic coupling, we snap-froze the reacted sample (Extended Data Fig. 8k–m) for cryo-EM analysis. We were able to capture three conformational states from the single-particle reconstruction, each depicting a distinct functional state during the half- to full-integration transition (Extended Data Fig. 9). The three states differ markedly in their spacer-side contacts and in Cas4 and integration status. In the 5.83 Å early-state reconstruction, the density clearly reveals that Cas4 still blocks the PAM-side integration site and the PAM-containing 3′-overhang is still sequestered in Cas4. Unable to dock into the integration site, the CRISPR repeat reaches over from the leader-side Cas1 directly to the spacer-side counterpart without contacting the Cas2 dimer in the middle. The spacer-side CRISPR repeat contacts a positively charged region on Cas1, near Cas4 (Fig. 4b, c, Extended Data Fig. 10g). The DNA density is weak, suggesting that it samples multiple conformations. In the 5.76 Å intermediate state reconstruction, the Cas4 density disappears and the CRISPR-repeat DNA points towards the spacer-side integration centre; however, the density is too weak for model building at the spacer side (Fig. 4b, c, Extended Data Fig. 10a). This suggests that even with Cas4 dissociation, the spacer-side CRISPR DNA capture and integration is inefficient owing to the lack of favourable leader-sequence contacts[11]. Finally, in the 3.81 Å full-integration-state reconstruction, densities clearly reveal that the CRISPR-repeat DNA has been accommodated into the spacer-side integration centre, and full integration has taken place (Fig. 4b, c). This snapshot is architecturally similar to the *E. faecalis* post-integration Cas1–Cas2 structure[12]—however, in the *G. sulfurreducens* structure, the leader-repeat DNA is not as sharply kinked at the Cas2 binding site as in the *E. faecalis* counterpart (Extended Data Fig. 10). These three snapshots define the order of molecular events and support a strong mechanistic coupling between the leader-half integration and the Cas4-mediated PAM processing, which enables PAM-specific spacer-side integration.

We considered how the leader-side integration activates spacer-side PAM cleavage remotely. There are at least two mechanistic possibilities: the leader-half integration may trigger a global conformational change that allosterically activates Cas4, or the physical contacts by the integrated leader-repeat DNA somehow activates Cas4. As no substantial conformational change in Cas1–Cas2 was observed among apo, half- and full-integration structures, we ruled out the allosteric activation model and probed deeper into the role of the leader-repeat DNA contact on Cas4 activation. We systematically shortened the leader-repeat DNA in the integration assay and observed a strong correlation with activation. When the leader repeat was too short to reach the spacer-side Cas4/Cas1 (sub2, 19-bp CRISPR repeat), the extent of PAM cleavage was indistinguishable from the prespacer-only control (Fig. 4d). When the leader repeat was long enough to reach the spacer-side Cas4/Cas1 but still too short to allow spacer-side integration (sub3, 30-bp CRISPR repeat), the PAM cleavage was markedly enhanced, approaching the extent observed in the positive control (sub4) (Fig. 4d). We therefore conclude that contacts by the half-integrated DNA efficiently stimulates the PAM-cleavage activity of Cas4. PAM cleavage leads to Cas4 dissociation, which exposes the spacer-side integrase centre and allows full integration (Fig. 4e, Supplementary Video 2).

Discussion

In this study, we provide a comprehensive set of mechanisms to explain the PAM-dependent spacer-acquisition process in Cas4-containing CRISPR systems. Our study firmly establishes that Cas4 is a dedicated PAM-cleaving endonuclease that is tightly regulated. In the context of the Cas1–Cas2 integrase complex, Cas4 specifically recognizes but refrains from cleaving the PAM-containing 3′-overhang in a prespacer. This 'molecular constipation' is the cornerstone for productive prespacer biogenesis and functional spacer integration in type I and type V CRISPR systems. We provide direct evidence that PAM recognition and the subsequent molecular constipation take place early during prespacer biogenesis. In essence, Cas4 serves as a gatekeeper to only channel productive precursors into the biogenesis pathway. We further show that host nucleases can assist the further processing of these precursors, and this eventually leads to a directional integration towards the leader-side CRISPR repeat. Moreover, we reveal that the leader-side integration efficiently activates the PAM-cleavage activity of Cas4, which causes Cas4 dissociation and allows the half- to full-integration transition. Exactly how spacer directionality is established in CRISPR systems lacking Cas4 requires further investigation. In type I-E CRISPR, the mechanism has been shown to involve Cas1-mediated PAM sequestration and integration-dependent desequestration[13,28]. Therefore, the PAM-dependent blockage and activation of the two integration centres in Cas1–Cas2 may be a universal feature to achieve directional spacer integration.

The structural similarity of Cas4 to the nuclease domains of AddAB, AdnAB and a structural domain in the equivalent location in RecBCD sheds light on the ancient function of Cas4 in spacer acquisition. These helicase–nuclease machines not only have essential roles in homology-directed repair, but also provide innate immunity for bacteria by preferentially degrading linear DNA lacking χ-sites, which are more probably of external origin. Functional interactions between RecBCD-, AddAB- and Cas1–Cas2-mediated spacer acquisition have been noted in previous studies[29,30]. Certain traits in the AdnA nuclease (and its structural equivalent in RecBCD) may have made them particularly desirable by Cas1–Cas2. For example, the subtle sequence preference and occasional enzymatic pausing may have been exploited by Cas1–Cas2 to establish PAM-dependent directional integration. This would have substantially increased the productive spacer acquisition in the ancient CRISPR systems. It is possible that the ancient Cas1–Cas2 relied so heavily on RecBCD or AddAB for spacer precursors that it began to establish a physical interaction with the nuclease domain to facilitate the process, eventually leading to the adoption of this host nuclease domain into the *cas* operon as *cas4*.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-03951-z.

1. Barrangou, R. et al. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).
2. Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res.* **40**, 5569–5576 (2012).
3. Nuñez, J. K. et al. Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. *Nat. Struct. Mol. Biol.* **21**, 528–534 (2014).
4. Nuñez, J. K., Lee, A. S., Engelman, A. & Doudna, J. A. Integrase-mediated spacer acquisition during CRISPR–Cas adaptive immunity. *Nature* **519**, 193–198 (2015).
5. Mojica, F. J., Diez-Villasenor, C., Garcia-Martinez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733–740 (2009).
6. Marraffini, L. A. & Sontheimer, E. J. Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature* **463**, 568–571 (2010).
7. Vink, J. N. A. et al. Direct visualization of native CRISPR target search in live bacteria reveals cascade DNA surveillance mechanism. *Mol. Cell* **77**, 39–50.e10 (2020).
8. Nuñez, J. K., Harrington, L. B., Kranzusch, P. J., Engelman, A. N. & Doudna, J. A. Foreign DNA capture during CRISPR–Cas adaptive immunity. *Nature* **527**, 535–538 (2015).
9. Wright, A. V. & Doudna, J. A. Protecting genome integrity during CRISPR immune adaptation. *Nat. Struct. Mol. Biol.* **23**, 876–883 (2016).
10. Wright, A. V. et al. Structures of the CRISPR genome integration complex. *Science* **357**, 1113–1118 (2017).
11. Budhathoki, J. B. et al. Real-time observation of CRISPR spacer acquisition by Cas1–Cas2 integrase. *Nat. Struct. Mol. Biol.* **27**, 489–499 (2020).
12. Xiao, Y., Ng, S., Nam, K. H. & Ke, A. How type II CRISPR-Cas establish immunity through Cas1–Cas2-mediated spacer integration. *Nature* **550**, 137–141 (2017).
13. Kim, S. et al. Selective loading and processing of prespacers for precise CRISPR adaptation. *Nature* **579**, 141–145 (2020).
14. Li, M., Wang, R., Zhao, D. & Xiang, H. Adaptation of the *Haloarcula hispanica* CRISPR–Cas system to a purified virus strictly requires a priming process. *Nucleic Acids Res.* **42**, 2483–2492 (2014).
15. Liu, T. et al. Coupling transcriptional activation of CRISPR-Cas system and DNA repair genes by Csa3a in *Sulfolobus islandicus*. *Nucleic Acids Res.* **45**, 8978–8992 (2017).
16. Shiimori, M., Garrett, S. C., Graveley, B. R. & Terns, M. P. Cas4 nucleases define the PAM, length, and orientation of DNA fragments integrated at CRISPR loci. *Mol. Cell* **70**, 814–824.e6 (2018).
17. Kieper, S. N. et al. Cas4 facilitates PAM-compatible spacer selection during CRISPR adaptation. *Cell Rep.* **22**, 3377–3384 (2018).
18. Almendros, C., Nobrega, F. L., McKenzie, R. E. & Brouns, S. J. J. Cas4–Cas1 fusions drive efficient PAM selection and control CRISPR adaptation. *Nucleic Acids Res.* **47**, 5223–5230 (2019).
19. Lemak, S. et al. Toroidal structure and DNA cleavage by the CRISPR-associated [4Fe-4S] cluster containing Cas4 nuclease SSO0001 from *Sulfolobus solfataricus*. *J. Am. Chem. Soc.* **135**, 17476–17487 (2013).
20. Lemak, S. et al. The CRISPR-associated Cas4 protein Pcal_0546 from *Pyrobaculum calidifontis* contains a [2Fe-2S] cluster: crystal structure and nuclease activity. *Nucleic Acids Res.* **42**, 11144–11155 (2014).
21. Zhang, J., Kasciukovic, T. & White, M. F. The CRISPR associated protein Cas4 is a 5′ to 3′ DNA exonuclease with an iron–sulfur cluster. *PLoS ONE* **7**, 0047232 (2012).
22. Lee, H., Dhingra, Y. & Sashital, D. G. The Cas4–Cas1–Cas2 complex mediates precise prespacer processing during CRISPR adaptation. *eLife* **8**, e44248 (2019).
23. Lee, H., Zhou, Y., Taylor, D. W. & Sashital, D. G. Cas4-dependent prespacer processing ensures high-fidelity programming of CRISPR arrays. *Mol. Cell* **70**, 48–59.e5 (2018).
24. Xiao, Y., Luo, M., Dolan, A. E., Liao, M. & Ke, A. Structure basis for RNA-guided DNA degradation by Cascade and Cas3. *Science* **361**, aat0839 (2018).
25. Shah, S. A., Erdmann, S., Mojica, F. J. & Garrett, R. A. Protospacer recognition motifs: mixed identities and functional diversity. *RNA Biol.* **10**, 891–899 (2013).
26. Jia, N. et al. Structures and single-molecule analysis of bacterial motor nuclease AdnAB illuminate the mechanism of DNA double-strand break resection. *Proc. Natl Acad. Sci. USA* **116**, 24507–24516 (2019).
27. Nuñez, J. K., Bai, L., Harrington, L. B., Hinder, T. L. & Doudna, J. A. CRISPR immunological memory requires a host factor for specificity. *Mol. Cell* **62**, 824–833 (2016).
28. Ramachandran, A., Summerville, L., Learn, B. A., DeBell, L. & Bailey, S. Processing and integration of functionally oriented prespacers in the *Escherichia coli* CRISPR system depends on bacterial host exonucleases. *J. Biol. Chem.* **295**, 3403–3414 (2020).
29. Levy, A. et al. CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* **520**, 505–510 (2015).
30. Modell, J. W., Jiang, W. & Marraffini, L. A. CRISPR–Cas systems exploit viral DNA injection to establish and maintain adaptive immunity. *Nature* **544**, 101–104 (2017).

Methods

PAM prediction

The 221,089 unique spacers along with genome source, *cas* gene information[31,32], and repeat sequence were obtained from CRISPRCasDb[33] in February 2020. These spacers were analysed by Blast search against our own sequence database containing all sequences from the NCBI nucleotide database[34,35], environmental nucleotide database[36], PHASTER[37], Mgnify[38], IMG/M[39], IMG/Vr[40], HuVirDb[41], HMP database[42], and data from Pasolli et al.[43]. All databases were accessed in February 2020.

Hits between spacers and sequences from the aforementioned nucleotide databases were obtained using the BLASTN program[44] version 2.10.0, which was run with parameters word_size 10, gap open 10, penalty 1 and an *e*-value cut-off of 1. Hits inside CRISPR arrays were detected and filtered out by aligning the repeat sequence of the spacer to the flanking regions of the spacer hit (23 nucleotides on both sides). To minimize the number of false positive hits, we further filtered hits based on the fraction of spacer nucleotides that hit the target sequence. In a first step, only hits with this fraction >90% were kept. To find targets for even more spacers while keeping the number of false positives low, we included a second step where hits with a matching percentage >80% were kept if another spacer from the same phylogenetic genus hit the same sequence in the stringent first round. Finally, we removed spacers that were shorter than 27 nt.

Highly similar repeat sequences of the same length were clustered using CD-HIT[45] with a 90% identity threshold. To increase the number of aligned sequences for PAM[5,46,47] determination, we hypothesized that similar repeat sequences would be used in the same orientation and would correspond to the same PAM sequences, as coevolution of PAM, repeat and Cas1 sequences has previously been shown[48,49]. The PAM for each aligned repeat cluster was then determined by aligning the flanking regions of the spacer hits in each cluster. To equally weigh each spacer within the repeat cluster, irrespective of the number of blast hits, consensus flanks were obtained per spacer. These consensus flanks contained the most frequent nucleotide per position of the flanking regions. From the alignment of consensus flanks (for clusters with at least 10 unique spacer hits) the nucleotide conservation in each flank was calculated. Conserved nucleotides were considered part of the PAM in case nucleotide conservation was higher than 0.5 bit score, and the bit score in that position was at least 5 times higher than the median bit score of the two 23-nt flanks. This PAM database was manually curated to fix PAMs determined incompletely when nucleotides that were slightly below the threshold did occur in other repeat clusters of the same subtype. The orientation of the PAM was set to match the overall orientations of experimentally determined PAMs in literature for different systems (upstream of 5′-end of the protospacer in type I systems and downstream of 3′ of the protospacer in type II systems).

Cas4 phylogenomics

Cas4 sequences were retrieved from each Cas4-containing genome in the PAM database. Cas4 sequences were discarded in case multiple Cas4 sequences of that subtype (subtypes defined by CRISPRCasdb) were present in a single genome, or when Cas4 belonged to a different subtype than the predicted subtype of the repeat cluster. The tree was generated with PhyML[50] from a MAFFT alignment of all Cas4 sequences[51]. The sequence logos were generated with Berkeley weblogo[52] and were performed on each group of Cas4 sequences with a similar PAM, where redundant sequences were removed by CD-hit (threshold 0.9). For groups with a small amount of nonredundant sequences (I-G TTN, I-G TAN and I-C CTT), additional Cas4 sequences were retrieved by BLAST search of repeat sequences of predetermined PAM repeat clusters and retrieving adjacent Cas4 sequences in the NCBI nucleotide database.

Bacterial strains

See Supplementary Table 1 for plasmids and their corresponding selection markers.

Plasmid construction

Plasmids used in this work are listed in Supplementary Table 1. The type IG CRISPR-Cas acquisition module from *G. sulfurreducens* DSMZ 12127 was amplified by PCR using the Q5 High-Fidelity Polymerase (New England Biolabs) and primers BN462 and BN1196 (Supplementary Information Table 2). The amplicon was cloned into the p13S-S ligation-independent (LIC) cloning vector (http://qb3.berkeley.edu/macrolab/addgene-plasmids/) by TA cloning, generating plasmid pCas4/1-2. For plasmid pCRISPR, a synthetic construct composed of T7 terminator, a CRISPR array (leader-repeat-spacer1-repeat), the mCherry gene, and flanking 20-bp homology regions to the vector, was introduced into pET cloning vector 2A-T amplified with primers BN1247 and BN1650 by Gibson assembly. E18Y mutant of Cas41 (pCas4/1-2-E18Y) was generated by mutagenesis using pCas4/1-2 as a template with primers BN3392 and BN3393. Double mutant E18Y/S191A (pCas4/1-2-E18Y/S191A) was generated by mutagenesis using pCas4/1-2-E18Y as a template with primers BN3394 and BN3395. All plasmids were verified by Sanger sequencing.

Spacer-acquisition assay

Escherichia coli BL21-AI was co-transformed with pCas4/1-2, pCas4/1-2-E18Y, or pCas4/1-2-E18Y/S191A and pCRISPR. Colonies were grown in 5 ml of LB supplemented with spectinomycin and ampicillin at 37 °C with shaking. After 2.5 h of growth, the expression of *cas* genes was induced with IPTG and L-arabinose, and the cultures were incubated for additional 2 h. Cells were made electrocompetent and transformed with 5 µl of each 50 µM prespacer prepared by mixing primers (Supplementary Table 2) at 1:1 from the 100 µM stock. Cells were recovered in LB for 1h at 37 °C, 180 rpm, and then grown overnight in 10 ml of LB supplemented with spectinomycin and ampicillin at 37 ˚C with shaking. Plasmids were extracted from the overnight cultures (Thermo Scientific GeneJet Plasmid Extraction Kit) and digested with EcoRI and NcoI to avoid amplification of larger products from the plasmid backbone. Digested plasmids were used to detect spacer acquisition by PCR using OneTaq 2x MasterMix (New England Biolabs) and a mix of three degenerate primers with different 3′ nucleotides (BN464, BN465 and BN1314) and primer BN1708[17]. Samples were run on 2% agarose gels and visualized for spacer acquisition using SYBR Safe. Unexpanded and expanded band percentages were determined using the Analysis Tool Box of ImageLab software using unmodified images. The expanded CRISPR DNA band was purified by automated size selection and submitted to a second round of PCR using the degenerate primers and the internal reverse primer BN1754[17,53].

Expanded CRISPR array sequencing

PCR amplicons of the expanded CRISPR arrays were purified using the GeneJET PCR Purification kit (Thermo Fisher Scientific) and the DNA concentration was measured using Qubit Fluorometric Quantification (Invitrogen). Samples were prepared for sequencing using the NEB Next Ultra II DNA Library Prep Kit for Illumina and each library was individually barcoded with the NEBNext Multiplex Oligos for Illumina (Index Primers Set1 and Set2). Sample size and concentration were then assessed using the Agilent 2200 TapeStation D100 high sensitivity kit, and samples were pooled with equal molarity. Pooled samples were denatured and diluted as recommended by Illumina and spiked with 15% of PhiX174 control DNA (Illumina). Sequencing was performed on a Nano flow cell (2 × 250 base paired-end) with an Illumina MiSeq. Image analysis, base calling, de-multiplexing, and data quality assessments were performed on the MiSeq instrument. Resulting FASTQ files were analysed by pairing and merging the reads using Geneious 9.0.5. Acquired spacers were extracted and analysed as described previously[17].

Cloning, expression and purification

The *Gsu_cas4/1* (Gsu0057 in KEGG) gene was cloned into pET28a - His_6-Twin-Strep-SUMO vector or pGEX-41-T-His_6-Flag-GST, between

BamHI and XhoI sites and expressed in *E. coli* BL21 (DE3) star cells. A 6 l cell culture was grown in LB medium at 37 °C until optical density at 600 nm (OD_{600}) reached 0.5. Expression was induced with 0.5 mM IPTG, 0.2 mg ml^{-1} ferrous sulfate and 0.4 mg ml L-cysteine at 16 °C overnight. Collected cells were resuspended in 100 ml buffer A containing 50 mM HEPES pH 7.5, and 500 mM NaCl, 10% glycerol and 5 mM TCEP, lysed by sonication, and centrifuged at 17,000*g* for 50 min at 4 °C. The supernatant was transferred into anaerobic conditioned glove box and applied onto the pre-equilibrated 4 ml Ni-NTA column (SUMO tagged expression) or 5 ml glutathione (GSH) column (for glutathione-*S*-transferase (GST)-tagged protein expression). After washing with 100 ml of buffer A, the protein was eluted with 20 ml buffer A plus 300 mM imidazole for SUMO-tagged purification and buffer A plus 15 mM reduced GSH for GST-tagged purification, then incubated with SUMO-protease or 3C protease at 4 °C for 2 h. Two millilitres of concentrated eluate was loaded onto a Superdex 200 16/60 SEC column (GE Healthcare) equilibrated with buffer C (10 mM HEPES pH 7.5, 500 mM NaCl, and 5 mM TCEP), the peak fractions were pooled and snap-frozen in liquid nitrogen for later use.

Gsu_cas2 (Gsu0058 in KEGG) gene was cloned into His_6-Twin-Strep-SUMO-pET28a vectors (Kan^R) between BamHI and XhoI sites. Protein expression, Ni-NTA purification, and SUMO-tag cleavage were carried out in similar conditions as for His–SUMO–Cas4/Cas1. After tag cleavage, Cas2 was purified on Superdex 200 16/60. The peak fractions were pooled and snap-frozen in liquid nitrogen for later use.

Affinity pull-down assay

Fifteen micrograms of GST-tagged Cas4/Cas1 and 30 μg untagged Cas2 were mixed and incubated with 10 μl GSH resin at 4 °C for 30 min in different salt concentration buffer (50 mM HEPES pH7.5, 10% glycerol, 5 mM TCEP, and 150, 300 or 500 mM NaCl) in the presence or absence of prespacer, in a total assay volume of 50 μl. The GSH resin was pelleted by centrifugation at about 100*g* for 30 s, washed 3 times with 200 μl of the corresponding binding buffer, then eluted with 70 μl elution buffer (50 mM HEPES pH7.5, 500 mM NaCl, 5 mM TCEP, and 15 mM reduced GSH). Eluted proteins were separated on 12% SDS–PAGE and stained by Coomassie blue.

Fluorescently labelled prespacer substrate preparation

Fluorescent DNA oligonucleotides (Supplementary Information Table 2) for biochemistry were synthesized (Integrated DNA Technologies) with either a /5AmMC6/ or /3AmMO/ label, fluorescently labelled in-house, annealed at equimolar amount, and purified by native PAGE to remove unannealed ssDNA.

Prespacer cleavage assays

Prespacer cleavage assays were set up in 20 μl reactions containing 10 nM final concentration of labelled prespacer, 500 nM Cas4/Cas1 and 250 nM Cas2 in a cleavage buffer containing 50 mM Tris pH 8.0, 100 mM KCl, 10% glycerol, 5 mM TCEP, and 5 mM metal ion $MnCl_2$ or different metal ions in Extended Data Fig. 1h. After 37 °C incubation for 1 h, reactions were quenched by vortexing with 20 μl of phenol/chloroform. The extracted aqueous phases were mixed with an equal volume of 100% formamide and separated on 13% urea–PAGE. Signals from each fluorescent dye were recorded using ChemiDoc (Bio-Rad). The $KMnO_4$ footprinting assay was carried out following previously published protocols[12].

Reconstitution of prespacer bound/integration Cas4/Cas1–Cas2 complex

Complex was formed by mixing $Cas4_2/Cas1_2$, Cas2 and prespacer (or half-integration-mimicking substrate) at a final concentration of 30 μM, 60 μM and 60 μM, respectively, in 500 μl total volume with a reconstitution buffer containing 25 mM Tris pH 8.0, 300 mM NaCl, 5 mM TCEP and 5 mM $MnCl_2$. After 37 °C incubation for 30 min, the complex was separated on Superdex 200 16/30 column equilibrated in the same buffer. The full-complex peak was pooled and concentrated to appropriate concentration and snap-frozen in liquid nitrogen for long-term storage.

Integration assays

The in vitro integration assays were set up as follows. Ten nanomolar prespacer was incubated with 250 nM Cas4/Cas1–Cas2 complex in the integration buffer containing 50 mM Tris pH 8.0, 100 mM KCl, 5 mM TCEP and 5 mM $MnCl_2$ in 20 μl reaction volume. After an initial incubation at 37 °C for 5 min, 300 ng of pCRISPR plasmid was introduced into the reaction. Integration was allowed to proceed at 37 °C for 1 h, after which 0.5 μl of EcoRI and XhoI restriction enzymes (NEB) were introduced for 10 min more at 37 °C to digest out the leader-repeat region of the plasmid, together with the integrated prespacer. Reactions were quenched by vortexing with 20 μl phenol–chloroform solution. The extracted aqueous phase was mixed with an equal volume of formamide, separated on 13% urea–PAGE, and scanned on ChemiDoc imaging system.

ExoI trimming and follow-up integration assays

Ten nanomolar prespacer was pre-incubated with 250 nM of Cas4/Cas1–Cas2 complex at 37 °C for 5 min in 20 μl containing the trimming buffer (50 mM Tris pH 8.0, 100 mM KCl, 10% glycerol, 5 mM TCEP, 5 mM $MnCl_2$ and 10 mM $MgCl_2$). The twofold ExoI dilution series in Fig. 3b was prepared by dilution of *E. coli* ExoI (NEB, 20 U μl^{-1}) to a final concentration of 0.2, 0.1, 0.05, 0.025 or 0.0125 U μl^{-1} in each reaction. The 1/10 and 1/50 ExoI concentrations in the Extended Data Fig. 9a correspond to 0.1 and 0.02 U μl^{-1}, respectively. The ExoI concentration in the Extended Data Fig. 8b was 0.1 U μl^{-1} across. In reactions in which the trimming and integration were coupled, 300 ng of pCRISPR plasmid (about 5 nM final concentration) was introduced at the same time with ExoI into the reaction. After incubation, the reaction was quenched by mixing with an equal volume of a buffer containing 95% formamide, 10 mM EDTA and 0.2% SDS, phenol-extracted, then separated on 13% urea–PAGE, and scanned on a ChemiDoc imaging system (Bio-Rad), as described above.

Electrophoretic mobility shift assay

Two nM final concentration of fluorescently labelled prespacer DNA was incubated with an increasing concentration of Cas4/Cas1–Cas2 complex for 15 min (in concentration titration experiments), or with 50 nM Cas4/Cas1–Cas2 complex for 0.5, 1, 2, 5 min (in time-course experiments) at 4 °C in a system with a total volume of 20 μl containing 50 mM Tris pH 8.0, 100 mM KCl, 5 mM TCEP, 5 mM $MnCl_2$ and 10% glycerol. After incubation,15 μl of each sample was loaded onto 1% agarose gel equilibrated in 1× TG buffer (20 mM Tris pH 8.0, 200 mM glycine) immediately. Electrophoresis was performed at 60 V for 40 min. The fluorescent signals from the gel were recorded using a ChemiDoc imaging system (Bio-Rad).

Negative-stain electron microscopy

Four microlitres of 0.01 mg ml^{-1} prespacer-bound Cas4/Cas1–Cas2 complex was applied to a glow-discharged copper 400-mesh continuous carbon grid. After a 30-s incubation, the grid was blotted on a filter paper, immediately transferred carbon-face down on top of a 2% (w/v) uranyl acetate solution for 1 min. The grid was then blotted on a filter paper again to remove residual stain, then air-dried on bench for 5 min. The grid was examined under a Morgagni transmission electron microscope operated at 100 kV with a direct magnification of 140,000× (3.2 Å pixel size) with an AMT camera system. Each image was acquired using an 800 ms exposure time and −1 to −2 μm defocus setting. Data processing and 2D classification were performed on cryoSPARC software.

Cryo-EM data acquisition

Four microlitres of 0.6 mg ml^{-1} SEC-purified prespacer-bound or half-integration-mimicking substrate-bound Cas4/Cas1–Cas2 complexes were applied to a Quantifoil holey carbon grid (1.2/1.3, 400 mesh) which had been glow-discharged for 30 s. Grids were blotted for 4 s

at 6 °C, 100% humidity and plunge-frozen in liquid ethane using a Mark IV FEI/Thermo Fisher Vitrobot. Cryo-EM images were collected on a 200-kV Talos Arctica transmission electron microscope (Thermo Fisher) equipped with a K3 Summit direct electron detector (Gatan). The total exposure time of each movie stack was about 3.5 s, leading to a total accumulated dose of 50 electrons per $Å^2$ which fractionated into 50 frames. Dose-fractionated super-resolution movie stacks collected from the K3 Summit direct electron detector were binned to a pixel size of 1.234 Å. The defocus value was set between −1.5 µm to −3.5 µm.

Cryo-EM data processing

Motion correction, contrast transfer function (CTF) estimation, blob particle picking, 2D classification, 3D classification and non-uniform 3D refinement were performed in cryoSPARC v.2[54]. Refinements followed the standard procedure, a series of 2D and 3D classifications with C_1 symmetry were performed as shown in Extended Data Figs. 4a, 7, 10a to generate the final maps. A solvent mask was generated and was used for all subsequent refinement steps. CTF post-refinement was conducted to refine the beam-induced motion of the particle set, resulting in the final maps. The final map CTF post-refinement was used to estimate resolution based on the Fourier shell correlation (FSC) = 0.143 criterion after correcting for the effects of a soft shape mask using high-resolution noise substitution. We noticed that the map of the full-integration complex was not homogeneous in both sides, so we divided the map into two half parts from the middle site by Chimera UCSF. Then imported two half maps into Relion 3.0[55] to make a mask for next masked local refinement respectively. Finally imported these two masks into cryoSPARC again and did a local refinement to get two half local refined maps and merged two maps to a final map in Extended Data Fig. 10. The detailed data processing and refinement statistics for all cryo-EM structures are summarized in Extended Data figures and Extended Data Table 1.

Statistics and reproducibility

We typically drew biochemistry conclusions on the basis of the best-quality gels. Such gels typically were repeated multiple times during the optimization stage to ensure reproducibility, albeit they may not have been repeated in the exact same format or loading sequence. When a conclusion was drawn on the basis of the band intensity changes or differences in a gel, we typically carried out $n = 3$ biologically independent assays to ensure reproducibility and statistical significance (for example, Fig. 4d; Extended Data Fig. 8e). In vivo assays were carried out in $n = 3$ biologically independent assays for quantification. All data points are displayed on the figure panels.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The cryo-EM density maps that support the findings of this study have been deposited in the Electron Microscopy Data Bank (EMDB) under accession numbers EMD-23839 (PAM/PAM prespacer bound), EMD-23840 (PAM/non-PAM prespacer bound), EMD-23843 (full-integration complex), EMD-23845 (half-integration complex, Cas4 still blocking the PAM side), EMD-23849 (half-integration complex, Cas4 dissociated) and EMD-23847 (sub-complex). The coordinates have been deposited in the Protein Data Bank (PDB) under accession numbers 7MI4 (PAM/PAM prespacer-bound), 7MI5 (PAM/non-PAM prespacer-bound), 7MI9 (full integration), 7MIB (half integration, Cas4 still blocking the PAM side), 7MID (sub-complex). MiSeq sequencing data that support analysis of in vivo prespacer integration have been deposited in the European Nucleotide Archive (ENA) under accession number PRJEB41616. Plasmids used in this study are available upon request.

31. Makarova, K. S. et al. Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.* **18**, 67–83 (2020).
32. Hudaiberdiev, S. et al. Phylogenomics of Cas4 family nucleases. *BMC Evol. Biol.* **17**, 232 (2017).
33. Pourcel, C. et al. CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and *cas* genes from complete genome sequences, and tools to download and query lists of repeats and spacers. *Nucleic Acids Res.* **48**, D535–D544 (2020).
34. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33**, D501–D504 (2005).
35. Benson, D. A. et al. GenBank. *Nucleic Acids Res.* **46**, D41–D47 (2018).
36. Sayers, E. W. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **37**, D5–D15 (2009).
37. Arndt, D. et al. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* **44**, W16–W21 (2016).
38. Mitchell, A. L. et al. MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.* **48**, D570–D578 (2020).
39. Chen, I. A. et al. IMG/M: integrated genome and metagenome comparative data analysis system. *Nucleic Acids Res.* **45**, D507–D516 (2017).
40. Paez-Espino, D. et al. IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes. *Nucleic Acids Res.* **47**, D678–D686 (2019).
41. Soto-Perez, P. et al. CRISPR–Cas system of a prevalent human gut bacterium reveals hyper-targeting against phages in a human virome catalog. *Cell Host Microbe* **26**, 325–335.e325 (2019).
42. Group, N. H. W. et al. The NIH Human Microbiome Project. *Genome Res.* **19**, 2317–2323 (2009).
43. Pasolli, E. et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* **176**, 649–662.e620 (2019).
44. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
45. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
46. Deveau, H. et al. Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J. Bacteriol.* **190**, 1390–1400 (2008).
47. Almendros, C., Guzman, N. M., Diez-Villasenor, C., Garcia-Martinez, J. & Mojica, F. J. Target motifs affecting natural immunity by a constitutive CRISPR–Cas system in *Escherichia coli*. *PLoS ONE* **7**, e50797 (2012).
48. Lange, S. J., Alkhnbashi, O. S., Rose, D., Will, S. & Backofen, R. CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems. *Nucleic Acids Res.* **41**, 8034–8044 (2013).
49. Alkhnbashi, O. S. et al. CRISPRstrand: predicting repeat orientations to determine the crRNA-encoding strand at CRISPR loci. *Bioinformatics* **30**, i489–i496 (2014).
50. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
51. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
52. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
53. McKenzie, R. E., Almendros, C., Vink, J. N. A. & Brouns, S. J. J. Using CAPTURE to detect spacer acquisition in native CRISPR arrays. *Nat. Protoc.* **14**, 976–990 (2019).
54. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
55. Xu, K., Zang, X., Peng, M., Zhao, Q. & Lin, B. Magnesium lithospermate B downregulates the levels of blood pressure, inflammation, and oxidative stress in pregnant rats with hypertension. *Int. J. Hypertens.* **2020**, 6250425 (2020).

Acknowledgements This work is supported by the Netherlands Organization for Scientific Research (NWO) VICI grant (VI.C.182.027) to S.J.J.B. and the National Institutes of Health (NIH) grant (GM118174) to A.K. This work made use of the Cornell Center for Materials Research Shared Facilities which are supported through the NSF MRSEC program (DMR-1719875). We thank S. N. Kieper, R. Miojevic, M. Ramos, G. Schuler and K. Spoth for helpful discussions, advice and technical assistance.

Author contributions A.K., S.J.J.B., C.H. and C.A. designed the research. C.H. is responsible for biochemistry and cryo-EM reconstructions; C.A., J.N.A.V., A.R.C. and A.C.H. are responsible for in vivo and bioinformatics analyses; K.H.N. and C.H. are responsible for structure building and refinement; and S.R.B. assisted with cryo-EM work. A.K. and C.H. wrote the manuscript with input from S.J.J.B., J.N.A.V. and A.R.C.

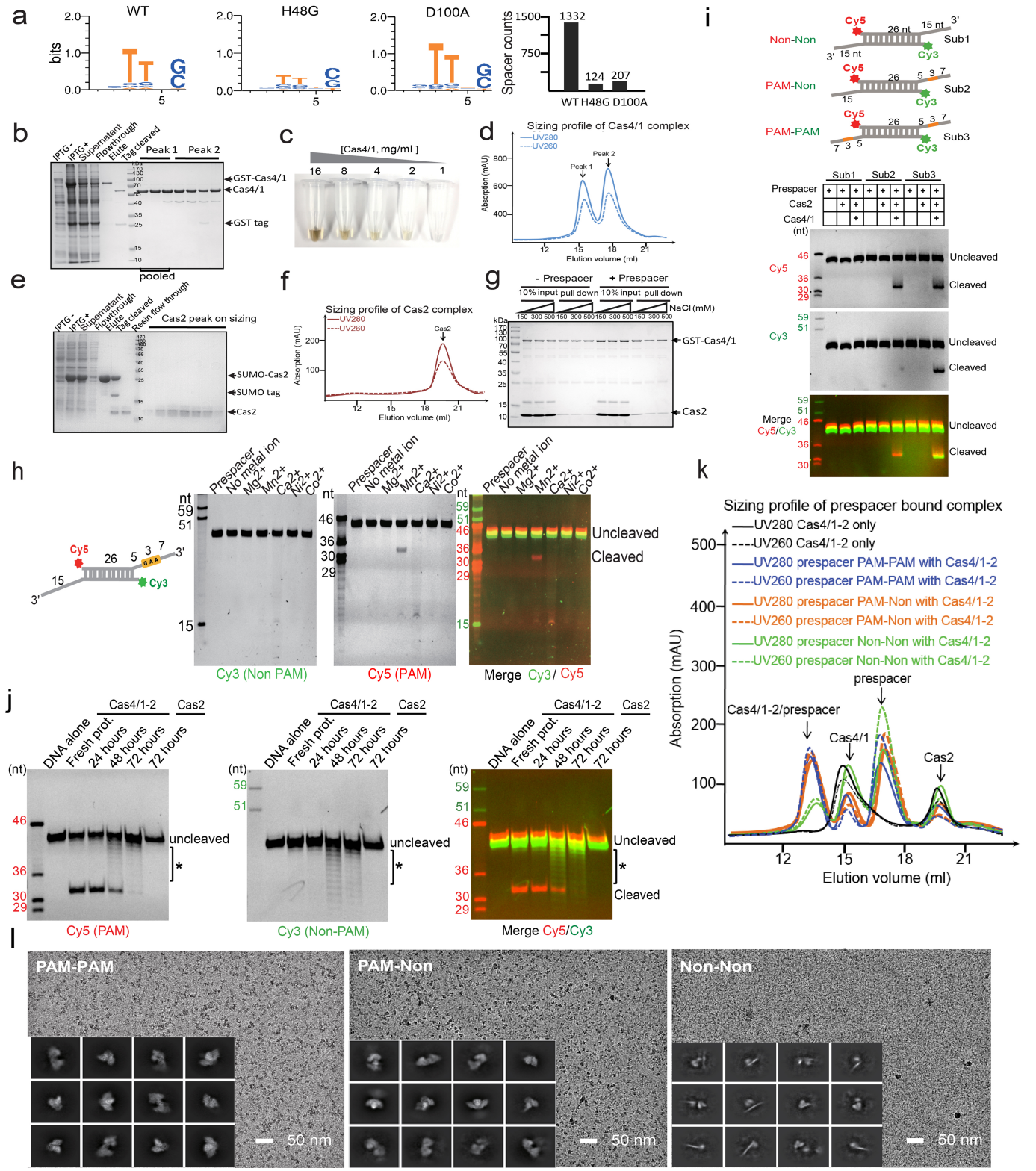
Competing interests The authors declare no competing interests.

Additional information

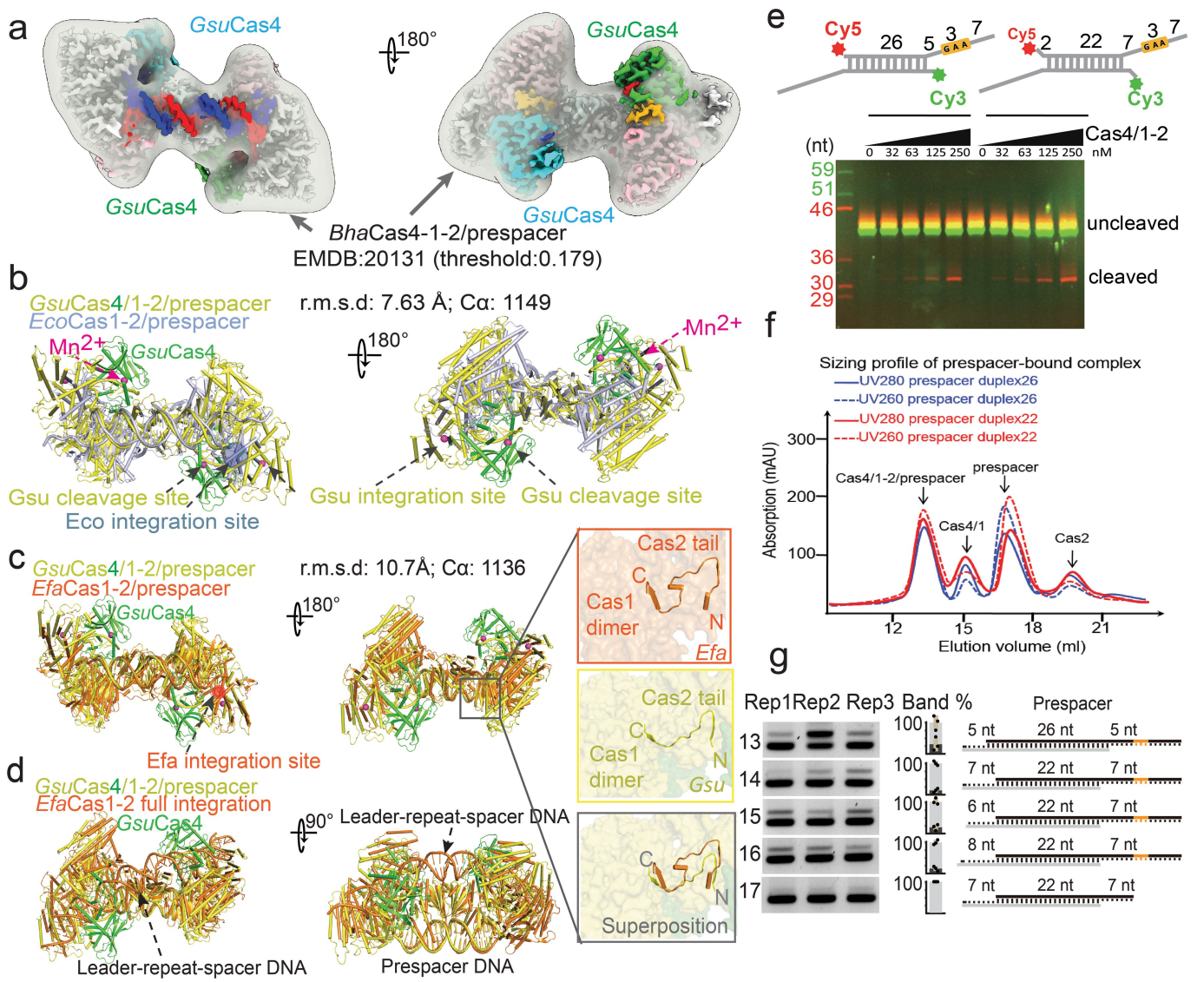
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03951-z.

Correspondence and requests for materials should be addressed to Stan J. J. Brouns or Ailong Ke.

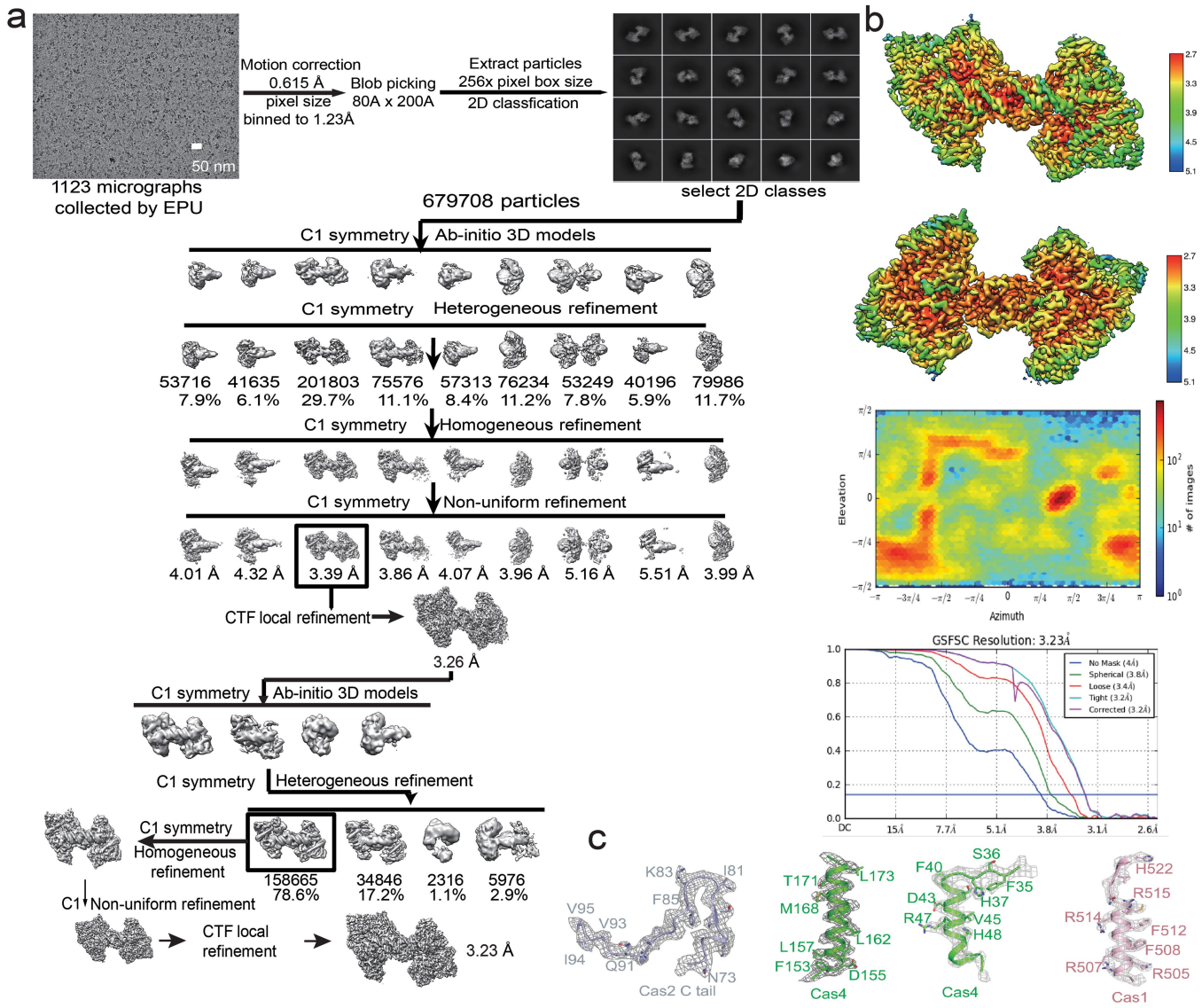
Peer review information *Nature* thanks Martin Jinek, Lennart Randau and Malcolm White for their contribution to the peer review of this work.



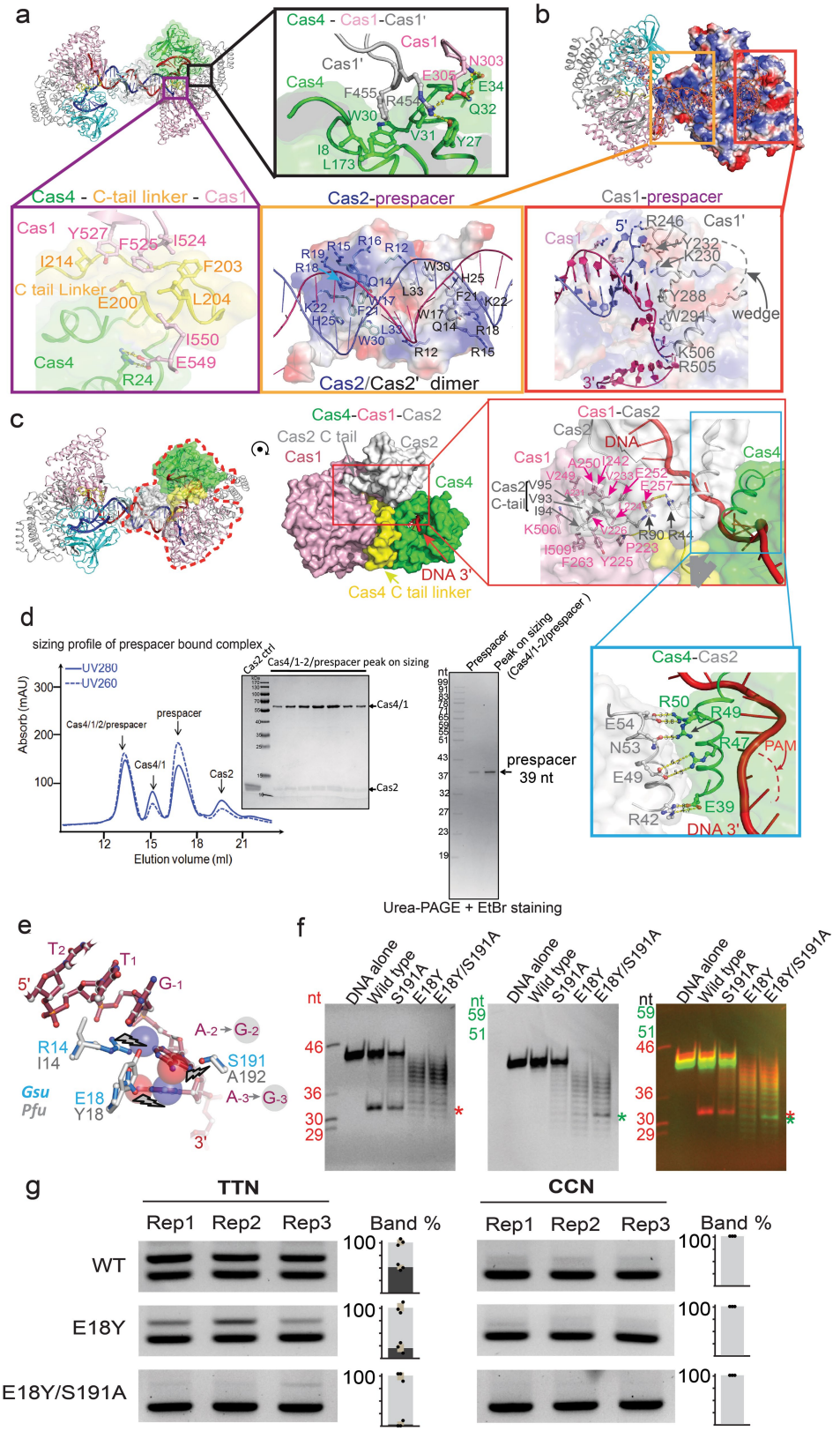
Extended Data Fig. 1 | Reconstitution and characterization of the *Gsu*Cas4/Cas1-Cas2 complex. a. Active site substitution in Cas4 nuclease center (H48G, D100A) reduced *in vivo* spacer acquisition efficiency dramatically. Left three panels display the WebLogo of PAM code from spacers integrated by each Cas4/1-2 variant. Rightmost panel displays the number of deep-sequencing reads that confirm spacer integration. **b**–**d**. *Gsu*Cas4/1 purification analyzed by SDS-PAGE, coloring from the Fe-S cluster, and SEC profile, respectively. **e,f**. Affinity purification of *Gsu*Cas2, SDS-PAGE, and SEC analysis, respectively. **g**. GST pull-down experiments revealing the physical interaction between *Gsu*Cas4/1 and *Gsu*Cas2, with or without prespacer present. **h**. Metal ion dependency in PAM cleavage reaction. **i**. Biochemistry showing Cas4/1-2 specifically cleaves the PAM-embedded 3′-overhang in prespacer. **j**. PAM-cleavage specificity is lost over time, presumably due to Fe-S oxidation in Cas4. **k**. SEC profile of *Gsu*Cas4/Cas1-Cas2, alone or programmed with different prespacer substrates. PAM-containing prespacers drive high-order complex formation. **l**. Cryo-electron micrographs of three different complexes, with corresponding preliminary 2D averages to investigate sample quality.



Extended Data Fig. 2 | In-depth analysis of the dual-PAM prespacer bound *Gsu*Cas4/Cas1-Cas2 structure. a. Comparison between the current 3.2 Å cryo-EM reconstruction with the previous negative staining reconstruction of the *B. hal* Cas4/1-2 complex (EMDB 20131)[22]. **b**–**d.** Pairwise alignment between *Gsu*Cas4/Cas1-Cas2/prespacer and *Eco*Cas1-Cas2/prespacer[8,31] (PDB 5DS4), *Efa*Cas1-Cas2/prespacer[12] (PDB 5XVN), and *Efa*Cas1-Cas2/full-integration[12] (PDB 5XVO), respectively. Alignments details are noted on the figure panel. Inset: the C-terminal tail of Cas2 plays similar roles in *G. sul* and *E. fae* structures in mediating edge-stacking with both Cas2 and Cas1. **e.** PAM was processed similarly in 22-bp or 26-bp mid-duplex containing prespacer by *Gsu*Cas4/Cas1-Cas2. **f.** SEC profile was similar when the two different prespacers were used to assemble the complex. **g.** Validation that prespacers containing a 22-bp mid-duplex are actively acquired *in vivo*. N=3 biologically independent assays were evaluated by PCR detection as shown, as well as relative percentages of expanded and non-expanded bands. Data presented as mean ± s.e.m.

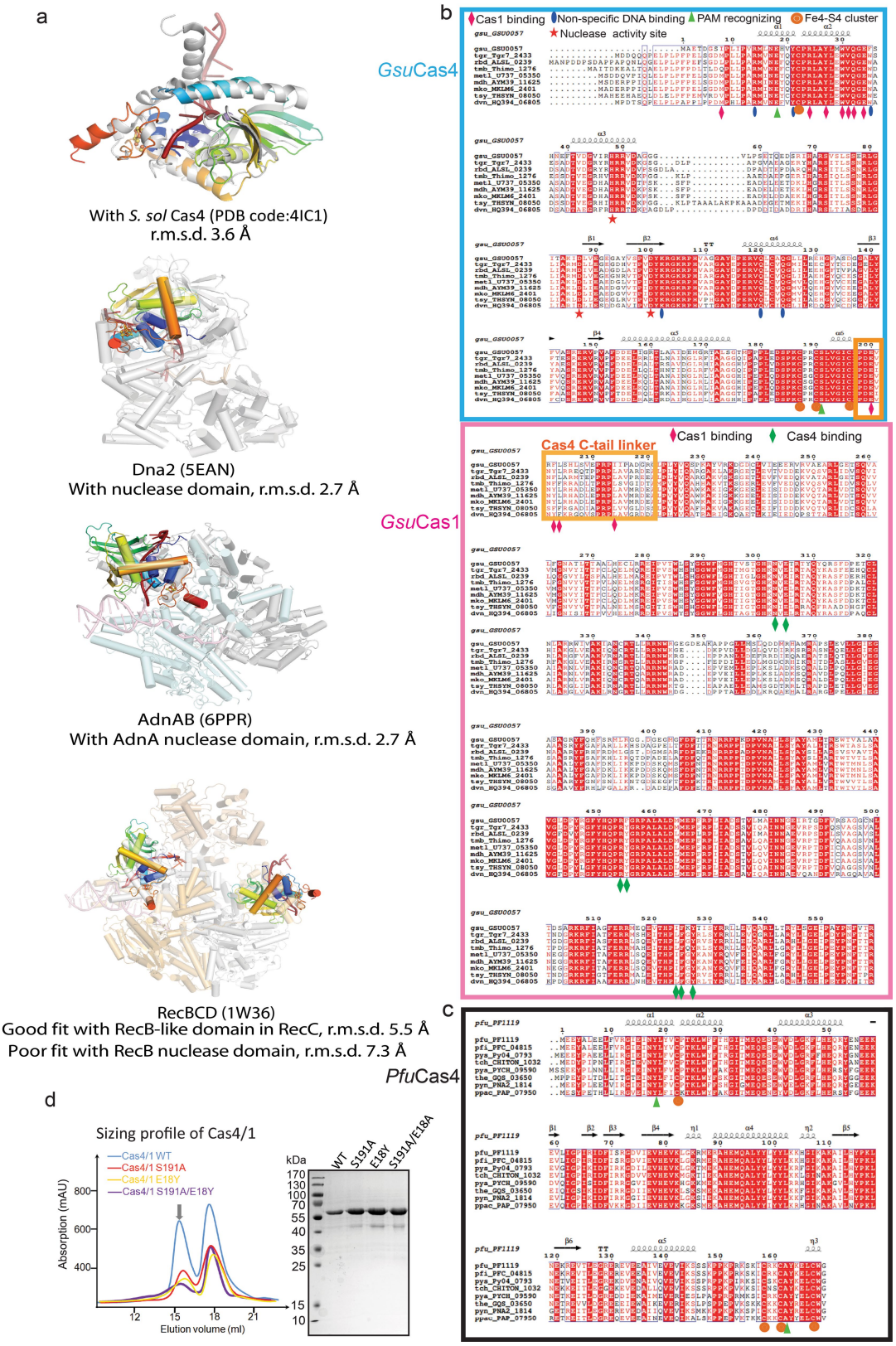


Extended Data Fig. 3 | Flow-chart of the cryo-EM single particle reconstruction of the dual-PAM prespacer bound GsuCas4/Cas1-Cas2. **a**. Cryo-EM reconstruction workflow for the dual-PAM prespacer bound Cas4/1-2 complex. **b**. Cryo-EM density of the dual-PAM prespacer bound Cas4/1-2 complex, colored according to local resolution (top). The viewing direction distribution plot (middle) and FSC curves (bottom) for data processing. **c**. Representative EM densities for Cas2, Cas4, and Cas1, superimposed with their corresponding structural model.

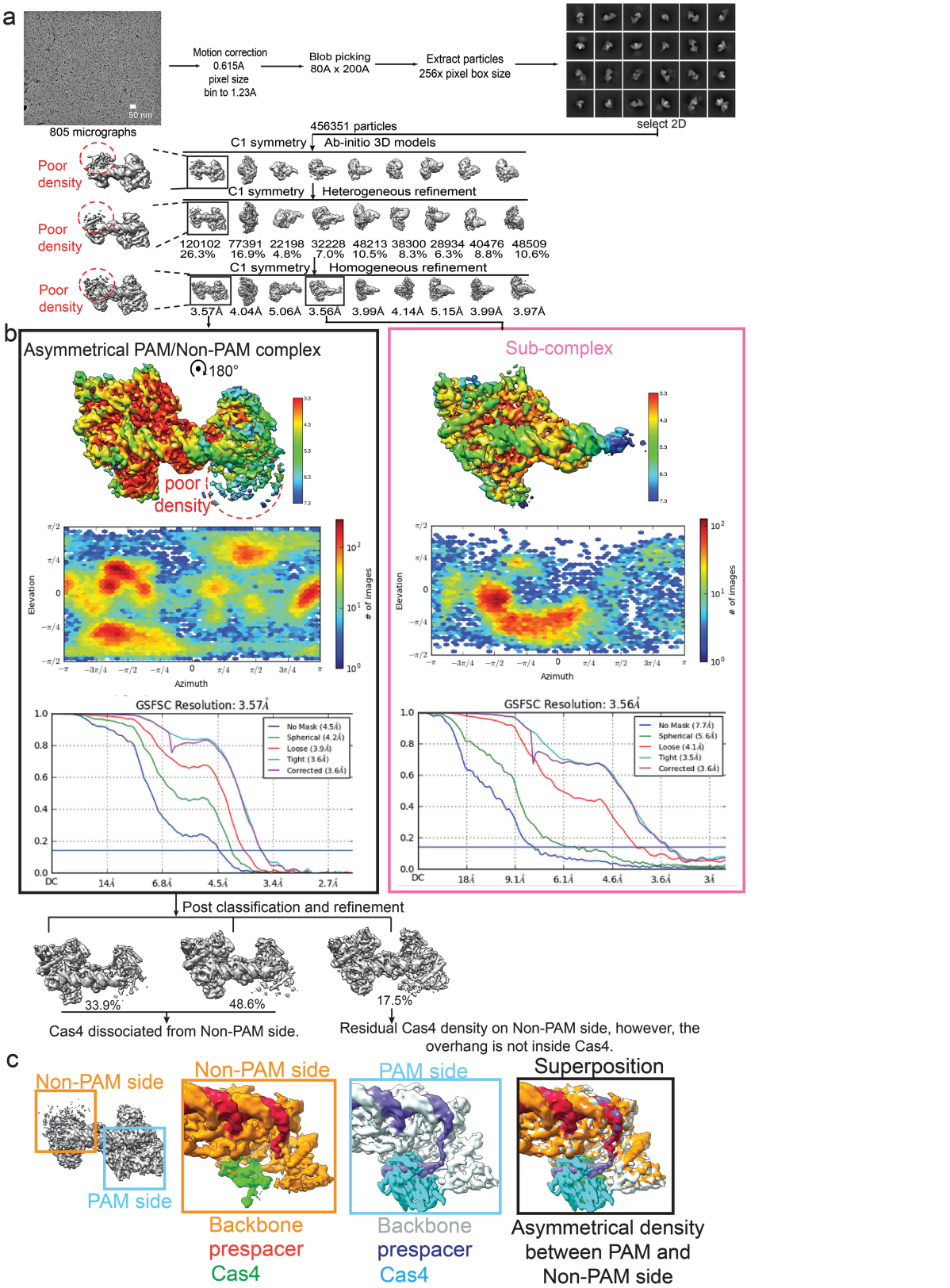


Extended Data Fig. 4 | See next page for caption.

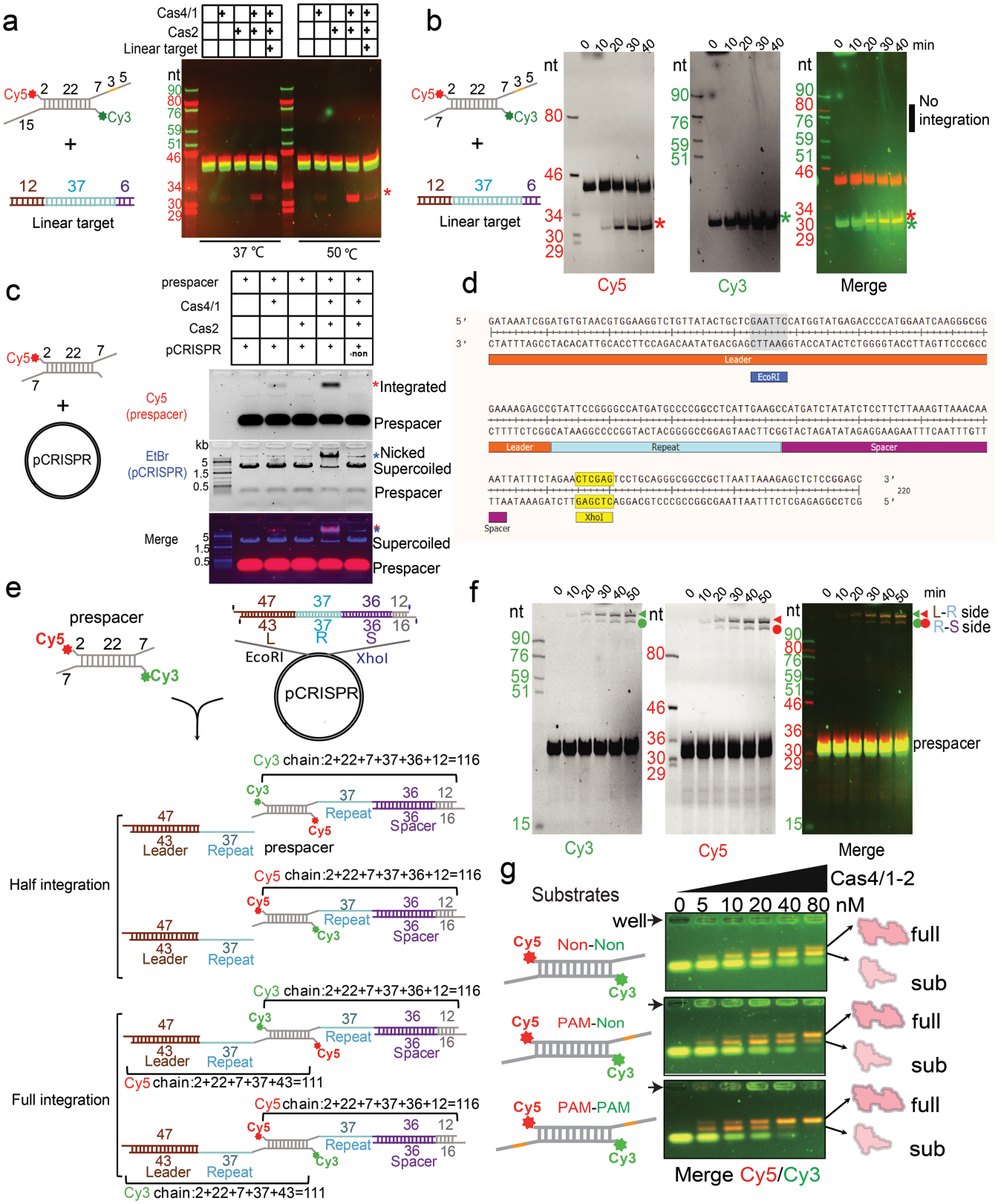
Extended Data Fig. 4 | In-depth *Gsu*Cas4/Cas1-Cas2 interface analysis and structure-guided mutagenesis attempt to switch PAM specificity. **a**. Overall dual-PAM structure. Insets: zoom-ins of interface between Cas4 and the two neighboring Cas1s. Cas4 connects to the non-catalytic Cas1 through a 20-amino acid fusion linker (colored in yellow), which mediates the dynamic docking and dissociation of Cas4. **b**. Surface electrostatic potential. Left inset: Cas2 contacts to the mid-duplex; Right inset: Cas1 end-stacking to the mid-duplex. Residues responsible for guiding the 3′-overhang are also shown. Cas1-Cas2 was found to specify a 22-bp mid-duplex rather than a 26-bp mid-duplex as defined by the integration assay; an additional two base-pairs are unwound from each end, and the mid-duplex is end-stacked by the N-terminal domain of the catalytic Cas1s on opposite ends. The 22-bp specification and the limited end-unwinding activity was previously observed in *Efa*Cas1-Cas2[11,12]. **c**. Cas1-Cas2 and Cas4-Cas2 interfaces. Top inset: the highly conserved C-terminus of Cas2 inserting into a hydrophobic pocket in Cas1, stabilizing complex formation. Bottom inset: the ceiling helix of Cas4 (aa 39–50) makes extensive polar contacts with a helix in Cas2 (aa 42–53). **d**. SEC, SDS-PAGE, and urea-PAGE analyses of the prespacer-bound complex used in cryo-EM analysis. They reveal the molecular weight, protein integrity, and prespacer integrity, respectively. For example, urea-PAGE reveals the PAM-overhang is not cleaved inside the Cas4/1-2 complex. **e**. Modeling the impact on PAM recognition by introducing the equivalent residues of E18 and S191 in *P. fur* Cas4 into *G. sul* Cas4 (E18Y and S191A substitutions). Specific atom changes in A-to-G switching (N6O substitution and N2 amine addition) are highlighted in colored balls. The steric clashes (lightening arrows) to *Pfu*PAM (3′-GGN in the 3′-overhang) are expected to be partially relieved when substitutions are in place. **f**. Impact of E18Y and S191A substitutions on PAM cleavage activity. **g**. *In vivo* spacer acquisition assay results for the wild type and PAM-specificity Cas4 mutants. While E18Y/S191A Cas4 showed compromised *Gsu*-PAM (TTN) prespacer integration, it was able to support integration of *Pfu*-PAM (CCN) containing prespacers *in vivo*. $N = 3$ biological independent assays were analyzed by PCR and the band quantification revealed integration efficiency. Data presented as mean ± s.e.m.



Extended Data Fig. 5 | In-depth analysis of the structure and sequence conservation in Cas4. a. Superposition of *Gsu*Cas4 with a standalone Cas4[19,20], and the nuclease domains in helicase-nuclease fusion proteins AddAB[32], AdnAB[26], RecBCD[33], and eukaryotic Dna2[34]. The caging of the ssDNA substrate and the arrangement of the Fe-S cluster and the catalytic triad are conserved themes. Interestingly, the Cas4 structure aligns poorly with the RecB nuclease in RecBCD; it agrees better with the RecB-like fold in RecC instead. **b, c**. Sequence alignment of *Gsu*Cas4, *Gsu*Cas1, and *Pfu*Cas4 with their close homologs. Based on the structural analysis, we marked the residues important for subunit interaction, substrate binding, catalysis and Fe-S cluster formation. **d**. Quality of the purified *Gsu*Cas4 mutants that carry the PAM-recognition residues from *Pfu*Cas4. These mutants were used in the structure-guided PAM-switching experiments in Extended Data Fig. 4.

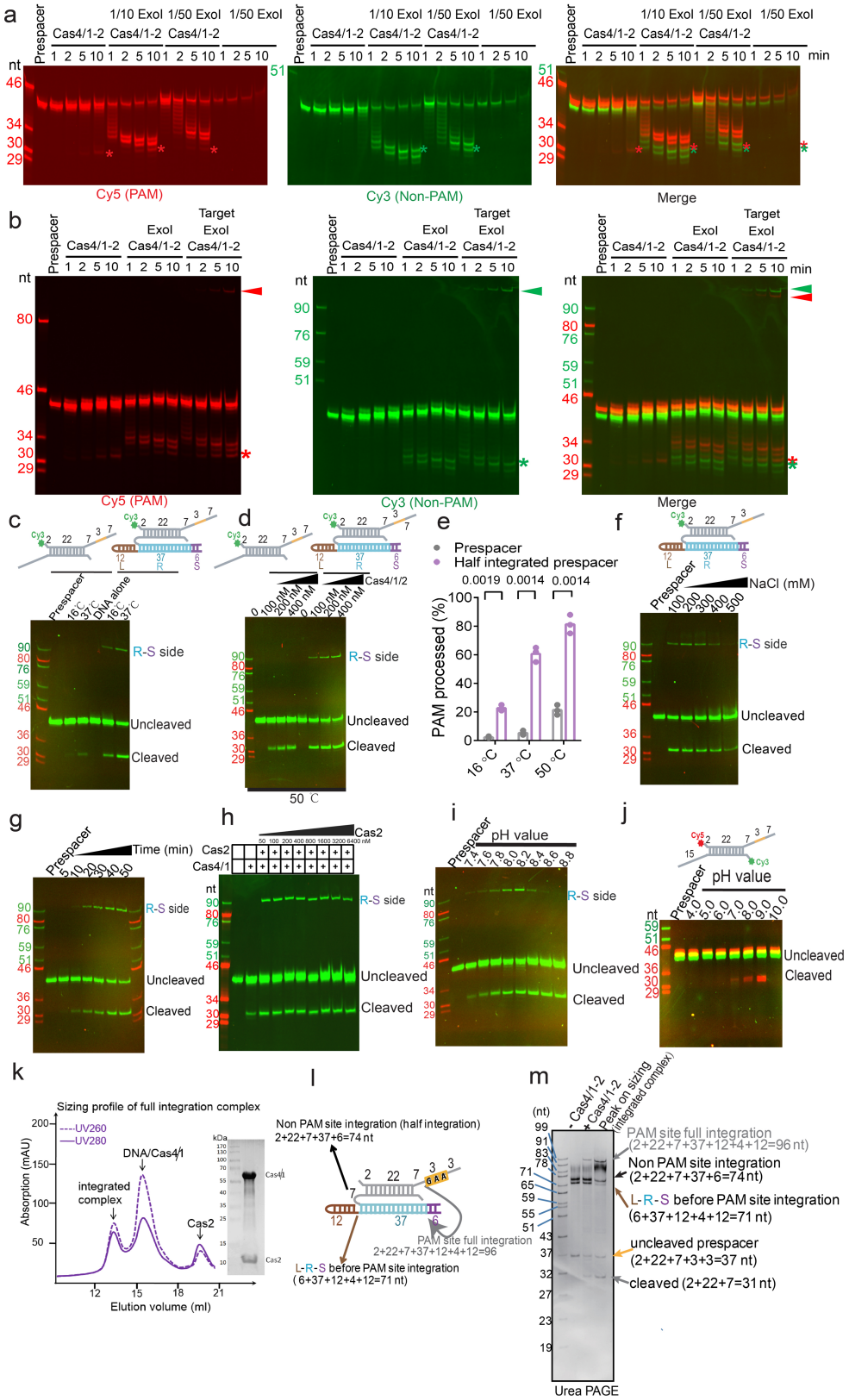


Extended Data Fig. 6 | Cryo-EM single particle reconstruction of the single-PAM prespacer bound *Gsu*Cas4/Cas1-Cas2. a. Flow-chart of the cryo-EM single particle reconstruction process that led to the reconstruction of two major snapshots. Left: Asymmetrical PAM/Non-PAM prespacer bound Cas4/1-2 complex. Right: That of the sub complex lacking $(Cas4/1)_2$ on the non-PAM side. **b**. Cryo-EM density of the two reconstructions colored according to local resolution (top); viewing direction distribution plot (middle); and FSC curves (bottom). **c**. Superposition of the PAM side and non-PAM side densities showing that Cas4 density is largely missing at the non-PAM side, and the non-PAM 3′-overhang is largely disordered. Only the first four nucleotides of the non-PAM 3′-overhang can be traced in the density, along a similar path as in the PAM-side.



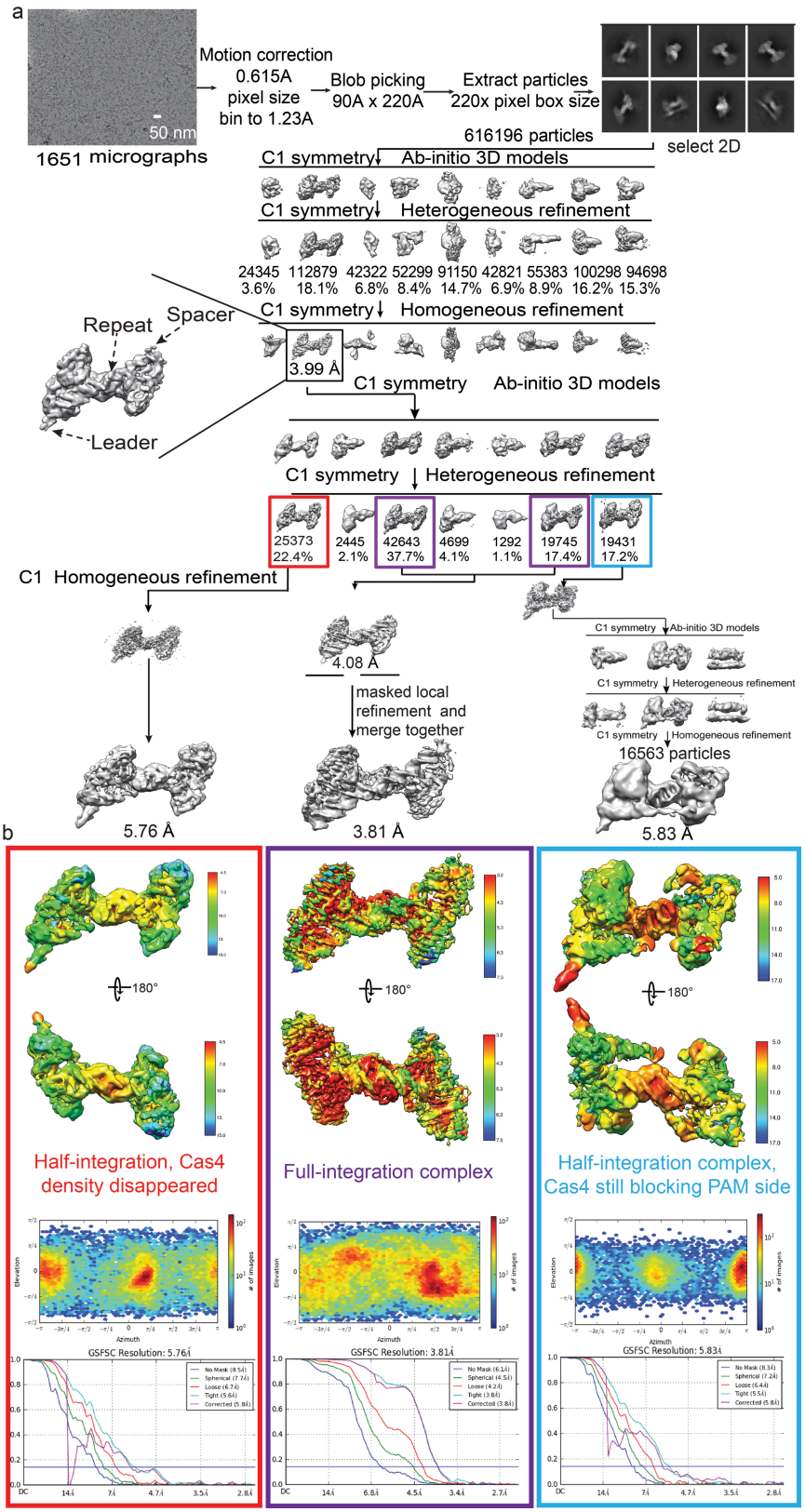
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | *In vitro* assays to distinguish integration directionality. a, b. Biochemistry showing that *Gsu*Cas4/1-2 is unable to integrate prespacer into the linear form of leader-repeat DNA. **c**. Successful prespacer integration into a leader-repeat containing plasmid by Cas4/1-2. **d**. The leader-repeat sequence cloned into the plasmid. We cleaved the leader-repeat sequence via the EcoRI and XhoI sites after the integration assay to further resolve the integration directionality on urea-PAGE. **e**. Schematic diagram explaining how the integration directionality can be resolved based on the fluorescent ssDNA sizes. **f**. Integration profile in urea-PAGE when both overhangs are integration-ready (7-nt long). Results showed that from the leader-repeat point of view, integration preferentially initiates from the leader-side, as the spacer-side integration trails after the leader-side integration in the time-course experiment. From the prespacer point of view, the integration directionality is scrambled. Each integration band contains two overlapping fluorescent signals. **g**. Native PAGE showing that in the concentration-gradient experiment, complex formation between Cas4/1-2 and prespacer takes place in a stepwise and PAM-dependent fashion.

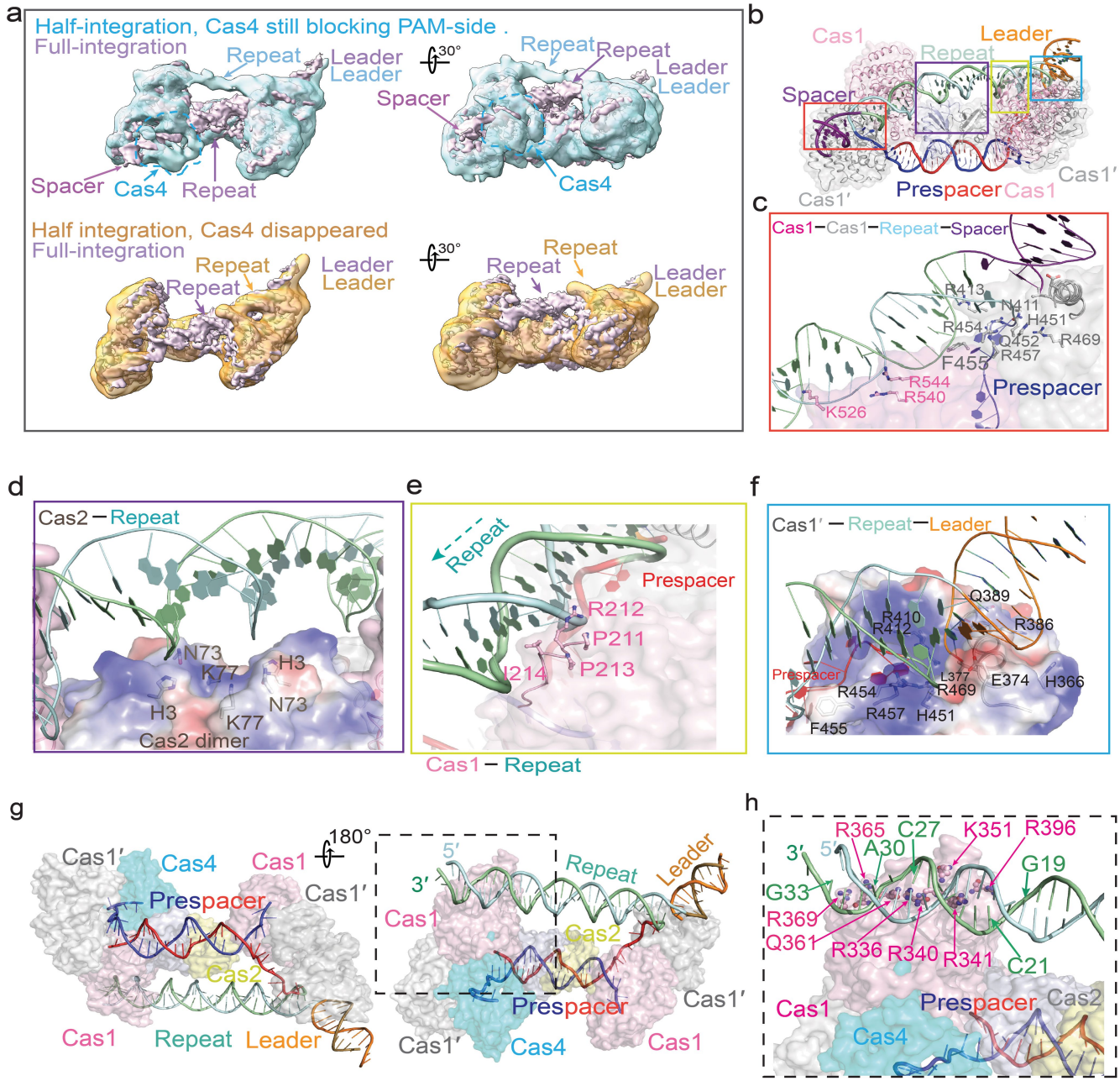


Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | In-depth analysis of the mechanistic coupling between half-integration and PAM cleavage by Cas4. a. Time-course experiment showing ExoI trims PAM and non-PAM overhangs differently. The non-PAM 3′-overhang was trimmed to within one nucleotide of the preferred length, 7 nt. The PAM-side 3′-overhang was protected by the footprint of Cas4 in the same reaction. **b**. Time-course experiment resolving the order of events from prespacer processing to full integration. Using the Cas4/1-2 (left set) and Cas4/1-2 plus ExoI (middle set) lanes as controls, the right set of experiment shows ExoI trimming triggers the integration of the non-PAM overhang into the leader-proximal target DNA. This is followed by the stimulation of PAM cleavage, and then the full integration from PAM-overhang to spacer-side target. **c**. Temperature-dependency of PAM cleavage and spacer-side integration. **d**. Side-by-side comparison of PAM cleavage at 50 °C, prespacer alone or programmed to the half-integrated state. **e**. Quantification of the cleaved band in **c. and d**. revealing the elevated PAM cleavage and full integration when leader-side integration already took place. Data were collected from $N = 3$ biologically independent experiments and presented with mean ± s.e.m. Statistical significance was assessed by two-tailed t-test, with the exact P values displayed. **f**. Salt-dependency of PAM cleavage and full integration. **g**–**i**. Optimization of full integration reaction by defining its time course, Cas2-dependency, and pH-dependency, respectively. **j**. Defining pH-dependency of PAM cleavage by Cas4. **k**. SEC analysis of the Cas4/1-2 complex programmed with the half-integration product mimic. Samples in the integrated complex peak was used for cryo-EM data collection and single particle reconstruction. **l**, Schematics of the half-integration product mimic annealed from oligonucleotides. **m**. Urea-PAGE analysis of the SEC peak in **k**. revealing that Cas4/1-2 further catalyzed the full-integration reaction after binding to the half-integration mimic.



Extended Data Fig. 9 | Cryo-EM single particle reconstruction of *Gsu*Cas4/Cas1-Cas2 programmed with a half-integration mimic. a. Workflow of cryo-EM data processing. **b.** Overall cryo-EM density showing resolution distribution, viewing direction distribution plot, and FSC curves of three different snapshots. Left: half-integration, Cas4 disappeared; Middle: full-integration; Right: half-integration, Cas4 still blocking PAM-side.



Extended Data Fig. 10 | In-depth analysis of the three snapshots captured from *Gsu*Cas4/Cas1-Cas2 programmed with a half-integration mimic. **a**. Superposition of cryo-EM reconstructions to reveal the structural differences among three functional states. **b**. Orientation view of the full integration snapshot for additional interface analysis. The entire leader-repeat DNA is contacted in a quasi-symmetric fashion at the following four regions. **c**. Contacts from the two Cas1 subunits to the spacer-repeat DNA. The spacer-side DNA density is degenerate and DNA bending is not significant. The leader-recognition α-helix in the catalytic Cas1 is not inserted into the minor groove of the spacer-side DNA. **d**. The backbone of the central dyad of CRISPR repeat is contacted by the positive charges and a proline-rich motif on the ridge of the Cas2 dimer. **e**. Immediately adjacent to the catalytic loop, the linker connecting Cas4 to Cas1 is involved in DNA contact. A conserved PRPI motif is exposed upon Cas4 dissociation and is involved in DNA minor groove contact. **f**. The 4-bp leader region immediately upstream of the CRISPR repeat is favorably recognized and significantly bent upwards by the DNA minor groove insertion of a glycine-rich α-helix in Cas1. As previously revealed, this recognition leads to strong leader-proximal preference at the first half-integration reaction[10–12]. A pair of inverted repeats is found at the border region of the CRISPR repeat. This inverted repeat is recognized at the major groove region by the catalytic Histidine-containing loop in Cas1[12]. **g**. Overall structure of the "Half-integration, Cas4 still blocking PAM-side" snapshot. This represents an early state, when Cas4 is still engaged in PAM recognition and the spacer-side leader-repeat is not allowed to enter into the integration site. **h**. The low-resolution EM density defines that the leader-repeat DNA preferentially contact a positively charged patch in Cas1. It should be noted that we are not able to define which specific DNA contact activates Cas4. This will require even higher temporal and spatial resolutions to resolve.

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

	PAM/PAM bound complex (EMDB-23839) (PDB 7MI4)	PAM/non-PAM bound complex (EMDB-23840) (PDB 7MI5)	Sub-complex (EMDB-23847) (PDB 7MID)	Full integration (EMDB-23843) (PDB 7MI9)	Half-int, PAM intact, Cas4 remains. (EMDB-23845) (PDB 7MIB)	Half-int, PAM cleaved, Cas4 dissociated. (EMDB-23849) (PDB N/A)
Data collection and processing						
Magnification	63,000	63,000	63,000	63,000	63,000	63,000
Voltage (kV)	200	200	200	200	200	200
Electron exposure (e–/Å^2)	50	50	50	50	50	50
Defocus range (μm)	1.5 ~ 3.5	1.5 ~ 3.5	1.5 ~ 3.5	1.5 ~ 3.5	1.5 ~ 3.5	1.5 ~ 3.5
Pixel size (Å)	1.23	1.23	1.42	1.31	2.18	1.32
Symmetry imposed	C1	C1	C1	C1	C1	C1
Initial particle images (no.)	1214203	896858	896858	1711962	1711962	1711962
Final particle images (no.)	158665	120102	32228	62074	16563	25373
Map resolution (Å)	3.2	3.6	3.6	3.8	5.8	5.8
FSC threshold	0.143	0.143	0.143	0.143	0.143	0.143
Map resolution range (Å)	20-2.8	20-3.0	20-3.2	20-3.5	20-5.0	20-5.0
Refinement						
Initial model used (PDB code)	N/A	N/A	N/A	N/A	N/A	N/A
Model resolution (Å)	3.2	3.6	3.6	3.8	5.8	-
FSC threshold	0.143	0.143	0.143	0.143		
Model resolution range (Å)	20-3.1	20-3.6	20-3.6	20-3.8	20-5.8	-
Map sharpening *B* factor (Å^2)	-50	-50	-50	-50	-	-
Model composition						
Non-hydrogen atoms	17162	15469	9789	15924	17216	-
Protein residues	2048	1852	1137	1706	1922	-
Ligands	10	8	6	4	6	-
DNA base	70	60	57	170	121	-
B factors (Å^2)						
Protein	71.98	146.39	71.96	54.24	54.24	-
Ligand						
R.m.s. deviations						
Bond lengths (Å)	0.009	0.009	0.009	0.006	0.008	-
Bond angles (°)	0.952	0.932	0.945	0.840	0.899	-
Validation						
MolProbity score	2.7	2.8	2.7	2.9	2.8	-
Clashscore	13	33	14	30	57	-
Poor rotamers (%)	5.8	8.1	5.7	2.08	0.81	-
Ramachandran plot						-
Favored (%)	91.45	90.47	91.32	86.9	87.54	-
Allowed (%)	8.44	9.42	8.68	12.97	12.45	-
Disallowed (%)	0.1	0.1	0.1	0.13	0.1	-

This table documents the data collection parameters and the refinement statistics for the six cryo-EM reconstructions analyzed in this mechanistic study.

nature research

Corresponding author(s): Ailong Ke, Stan Brouns

Last updated by author(s): Jun 25, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Cas4 sequences were retrieved from each Cas4-containing genome in the PAM database. A tree was generated with PhyML from a MAFFT alignment of all Cas4 sequences.
Data analysis	The Cas4 sequence logos were generated with Berkeley weblogo and were performed on each group of Cas4 sequences with a similar PAM, where redundant sequences were removed by CD-hit (threshold 0.9).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The cryo-EM density maps that support the findings of this study have been deposited in the Electron Microscopy Data Bank (EMDB) under accession numbers of EMD-23839 (PAM/PAM prespacer bound), EMD-23840 (PAM/Non-PAM prespacer bound), EMD-23843 (full integration complex), EMD-23845(half integration, Cas4 still blocking the PAM side), EMD-23849 (half integration, Cas4 dissociated), and EMD-23847 (sub-complex). The coordinates have been deposited in the Protein Data Bank (PDB) under accession numbers of 7MI4 (PAM/PAM prespacer-bound), 7MI5 (PAM/non-PAM prespacer-bound), 7MI9 (full integration), 7MIB (half integration, Cas4 still blocking the PAM side), 7MID (sub-complex). MiSeq sequencing data that support analysis of in vivo prespacer integration have been deposited in the European Nucleotide Archive (ENA) under accession number PRJEB41616. Plasmids used in this study are available upon request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Three biological replicates were performed for in vivo spacer acquisition assays. Biochemistry gels in the publication were done once in its final form. We made sure the results were reproducible using pilot experiments.
Data exclusions	No data exclusion performed. Cropped version of the gels may be shown in figures. Uncropped gel figures are provided.
Replication	Three Biological replicates whenever possible.
Randomization	N/A
Blinding	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging