High-resolution mechanism for Cas4-assisted PAM-selection and directional spacer acquisition in CRISPR-Cas

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

[1] Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, U.S.A.

[2] Department of Bionanoscience, Delft University of Technology, Van der Maasweg 9, 2629 HZ Delft, Netherlands.

[3] Kavli Institute of Nanoscience, Delft, Netherlands

[4] Department of Life Science, Pohang University of Science and Technology, Pohang, Gyeongbuk, Republic of Korea.

[#] These authors contributed equally to the publication.

[*] Correspondence: ailong.ke@cornell.edu, stanbrouns@gmail.com

Prokaryotes adapt to challenges from mobile genetic elements by acquiring foreign DNA-derived spacers into the CRISPR array to update the RNA-guided CRISPR immunity [1]. Spacer insertion is carried out by the Cas1-Cas2 integrase complex [2-4]. A significant fraction of CRISPR-Cas systems further utilize an Fe-S cluster containing nuclease Cas4 to ensure spacers are acquired from a DNA flanked by a protospacer adjacent motif (PAM) [5,6] and inserted into the CRISPR array directionally, so that the resulting CRISPR RNA can guide target-searching in PAM-dependent fashion. Focusing on Type I-G CRISPR in *Geobacter sulfurreducens* where Cas4 is naturally fused with Cas1, here we provide a complete and high-resolution mechanistic explanation for the Cas4-assisted PAM-selection, spacer biogenesis and directional integration. The Fe-S cluster region is an integral component of the PAM-recognition module in Cas4. During biogenesis, only DNA duplexes possessing a PAM-containing 3′-overhang trigger the stable assembly of an intact Cas4/Cas1-Cas2 complex. Importantly, throughout this process the PAM-containing 3′-overhang is specifically recognized, sequestered, but not cleaved by the Cas4 nuclease. This molecular constipation prevents the PAM-end of the prespacer from participating in integration. Lacking such recognition and sequestration, the non-PAM end of the prespacer is trimmed by host nucleases and preferentially integrated by Cas1 to the leader-side CRISPR repeat. Importantly, when the half-integrated CRISPR repeat DNA reaches over to contact the spacer-side Cas4/1-2, it activates Cas4 to cleave PAM and dissociate from Cas1-Cas2. This in turn exposes the Cas1 integrase center to allow spacer-side integration to take place. Overall, the intricate molecular interaction between Cas4 and Cas1-Cas2 dictates the type of prespacers eligible for integration, and couples the timing of PAM processing with the stepwise integration to establish directionality, so that the newly acquired spacers are productive in guiding PAM-dependent CRISPR interference.

Main Text

Prokaryotes have a unique ability to acquire immunological memories against mobile genetic elements by integrating short fragments of DNA (*i.e.* spacers) in between the CRISPR repeats [7,8]. The array of repeat-spacers serves as a transcription template to generate guide RNAs that can direct CRISPR effector protein complexes to find, bind and cleave DNA or RNA targets. To support protection by all DNA-targeting CRISPR-Cas systems, spacers need to be compatible with a universal DNA-targeting requirement called the protospacer adjacent motif (PAM) [9-11]. This short sequence motif directly flanking the target site helps crRNA-guided complexes distinguish true targets from the actual spacer in the CRISPR array, and thereby prevents lethal self-targeting. Furthermore, the presence of a PAM dramatically speeds up the target-searching process by the crRNA-guided effector complexes, by reducing the total number of candidate sites within the DNA [12]. To ensure CRISPR spacers are only derived 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 [13] further encode a dedicated CRISPR adaptation protein Cas4 [14] that works in conjunction with the core spacer acquisition machinery consisting of Cas1 and Cas2 [2-4,15-21]. A number of studies have contributed to our understanding of the role of Cas4. While early studies mainly showed that deletion of the *cas4* gene impaired spacer acquisition in type I-B systems in *Haloarcula hispanica* [22] and type I-A in *Sulfolobus islandicus* [23], recent studies using type I-A in *Pyrococcus furiosus* [24], I-D in *Synechocystis sp.* [25] and I-G (previously I-U) in *Geobacter sulfurreducens* [26] established a critical role for Cas4 in acquiring spacers with a functional PAM. On the protein level Cas4 was found to harbour an Fe-S cluster and to catalyze various exo- and endonuclease activities [27-29]. Only recently did it become clear from work in I-C *Bacillus halodurans* that Cas4 uses its nuclease activity to cleave PAM sequences in spacer precursors just before integration into the CRISPR array [30,31]. Further studies with this Cas4 variant showed that Cas4 forms a complex with a dimer of Cas1 and associates with Cas2 upon prespacer binding [30,31]. The emerging picture is that Cas4 is somehow involved in PAM-selection and processing, and that it must be important for the directional integration of spacers into the CRISPR array. Yet, the molecular mechanism of this key process has remained elusive.

Cas4 is a dedicated PAM-cleaving endonuclease

A highly active and robust Cas4-containing spacer acquisition system from the *Geobacter sulfurreducens* I-G CRISPR-Cas was identified in the screening of a suitable system for biochemical and structural characterizations. Cas4 is naturally fused with Cas1 in the *G. sul* acquisition module **(Fig. 1a)**. Together with Cas2 they were capable of acquiring 34-40 base

pair (bp)-long spacers (the majority are between 35-37 bp) into the CRISPR locus in a PAM-dependent manner (5′-TTN, 3′-AAN at the 3′-overhang) [26]. The enzymatic activity of Cas4 was shown to be required for PAM processing [26]. To derive rules governing the prespacer processing and integration, we electroporated prespacers of various sequence and structure compositions into *E. coli* cells containing a *G. sul cas4/cas1-cas2/CRISPR* genomic locus and analyzed cells for newly acquired spacers using PCR and deep sequencing methods **(Fig. 1b, c, Extended Data Fig. 1a)**. Based on prior structural and biochemical work, it was hypothesized that *Gsu*Cas4/Cas1-Cas2 may preferentially integrate prespacers containing a 26-bp mid-duplex, with 5-nt 3′-overhangs on each side [18,20,26,30]. Such prespacers were indeed robustly integrated in a directional and single-stranded PAM (ss-PAM) dependent fashion **(Fig. 1b-c)**. Prespacers lacking a ss-PAM were not integrated **(Fig. 1b)**. The context surrounding PAM also influenced the 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 ds-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)**. It is possible that the 3'-overhang trimming is precise at the PAM-side but slightly distributive at the non-PAM side. These data converge in suggesting that *Gsu*Cas4/Cas1-Cas2 preferentially recognizes prespacers containing a correctly spaced PAM in the 3′-overhang of a DNA duplex.

Next, we switched to biochemical reconstitution to understand the molecular basis of Cas4-assisted spacer integration. The PAM-containing 3′-overhang of the prespacer was found to be specifically cleaved by the recombinant *Gsu*Cas4/Cas1-Cas2 complex; the non-PAM 3′-overhang remained intact **(Fig. 1d, Extended Data Fig. 1b-i)**. Cleavage was Mn^{2+}-dependent and took place precisely after PAM (3′-$A_{-3}A_{-2}G_{-1}$↓; **Extended Data Fig. 1h-i**). While precise, PAM processing was rather inefficient. Only ~5% of the PAM-containing overhang was processed after 60 minutes of incubation in 37 ℃, in 50-fold excess of *Gsu*Cas4/Cas1-Cas2 **(Extended Data Fig. 1h)**. The underlining mechanism for the attenuated PAM processing only became clear after structural analysis. Interestingly, extended exposure to air induced promiscuous DNA cleavage activity from this complex **(Fig. 1e)**, likely due to the oxidation of the Fe-S cluster in Cas4. The various level of oxidation may explain the spectrum of reported endo

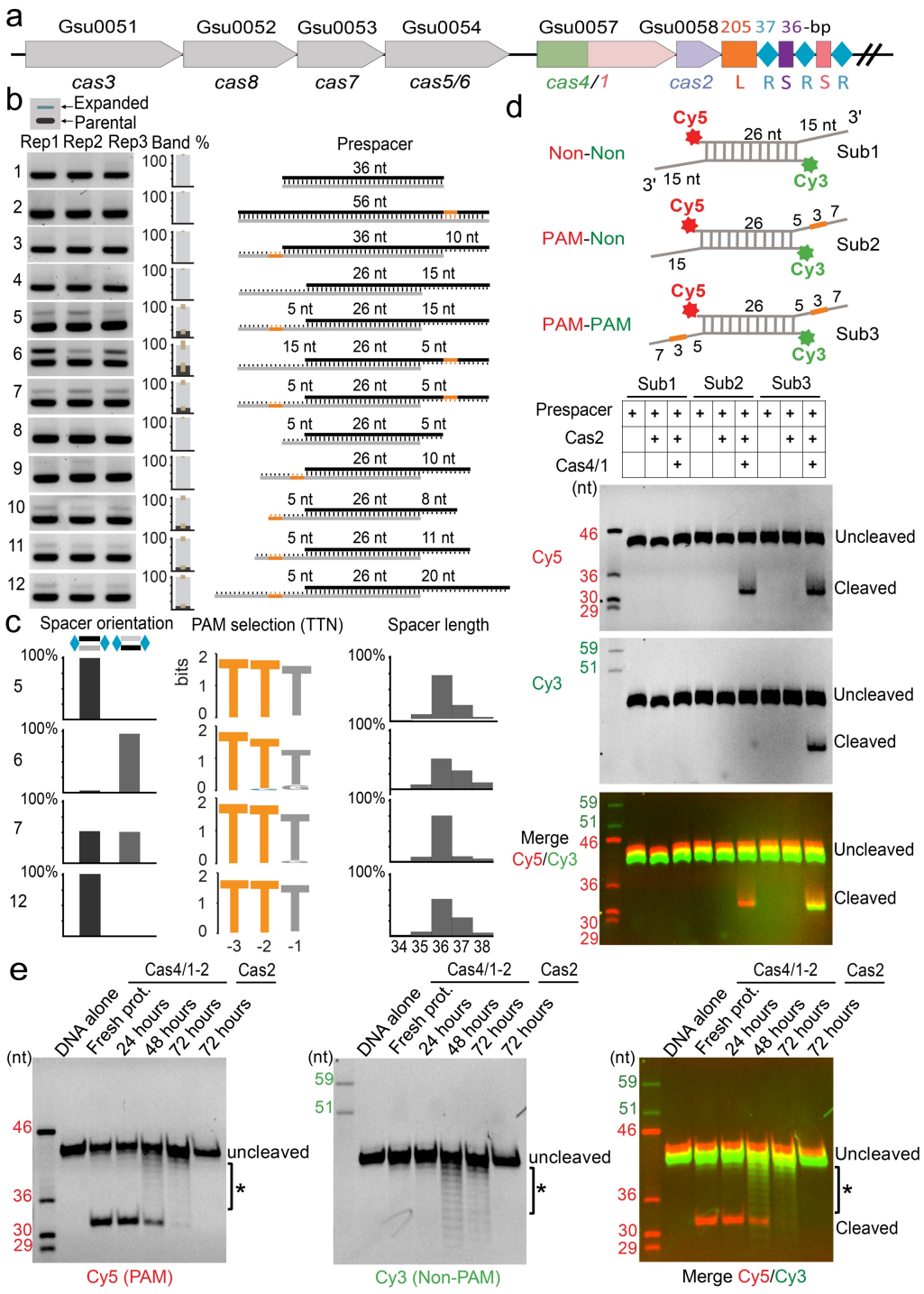


Figure 1. PAM-dependent prespacer processing and acquisition by *Gsu*Cas4/Cas1-Cas2. a. *cas* operon organization in Type I-G CRISPR in *G. sul*. Top: KEGG database identifier; Bottom: gene names; L, R, S, bp: Leader, repeat, spacer, base-pairs. **b.** *In vivo* acquisition of electroporated prespacers with different sequence and structural compositions. Three replicates of PCR detection are shown, as well as relative percentages of expanded and non-expanded bands. PAM is represented in orange. PAM-1 appears conserved because a single prespacer was used in the assay. **c.** Analysis of spacer orientation, PAM code and length for a subset of prespacers in **b. d.** Biochemistry showing Cas4/1-2 specifically cleaves PAM-embedded 3′-overhang in prespacer. **e.** PAM-cleavage specificity is lost over time, presumably due to Fe-S oxidation in Cas4.

109

and exonuclease activities for Cas4 in the literature [27-31].

Architecture of the dual-PAM prespacer bound Cas4/Cas1-Cas2 complex

Whereas physical interaction could be detected between *Gsu*Cas4/Cas1 and *Gsu*Cas2 in affinity pull-down and size-exclusion chromatography (SEC) experiments, functional complex formation was driven by the prespacer **(Extended Data Fig. 1g, 2)**. A dual- or single-PAM containing prespacer led to stable higher-order complex formation, as revealed by SEC and electron microscopy (EM) analyses. In contrast, a PAM-less prespacer was not efficient at organizing complex formation **(Extended Data Fig. 2)**. EM analyses revealed the formation of dumbbell-shaped particles characteristic of Cas1-Cas2 complexes. Because the dual-PAM prespacer containing *Gsu*Cas4/Cas1-Cas2 complex was especially homogeneous under negative-staining and cryo-EM **(Extended Data Fig. 2c)**, we first attempted to generate a high-resolution reconstruction from this *quasi* symmetric PAM-recognition complex. The single particle reconstruction reached 3.23 Å in resolution, which revealed significant more structural details than the negative-staining EM reconstructions of related Cas4-Cas1-Cas2 complexes [30] **(Fig. 2; Extended Data Fig. 3a, 4)**. The $Cas1_4$-$Cas2_2$ integrase core assumes its characteristic dumbbell shape - the Cas2 dimer constitute the central handle, and two Cas1 dimers constitute the two distal weights **(Fig. 2a)**. In each dimer, only one Cas1 participates in spacer integration, the other plays structural roles. The overall architecture and the detailed interactions leading to *Gsu*Cas1-Cas2 complex formation are more consistent with those found in the *Enterococcus faecalis* rather than the *E. coli* complex [18,20] **(Extended Data Fig. 3b-d)**. For example, the C-terminal tails of *Gsu*Cas2 and *Efa*Cas2 stabilize the complex by mediating similar structural contacts to the neighboring Cas1 and to the opposing Cas2 **(Extended Data Fig. 3c, 5**); the contacts are mediated very differently in the *E. coli* complex. Surprisingly, 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 **(Fig. 2a-b, 2e; Extended Data Fig. 5b)**. Indeed, re-designed prespacers containing a 22-bp mid-duplex integrated as efficiently as the 26-bp version in the *in vivo* and *in vitro* assays **(Fig. 2d; Extended Data Fig. 3e-f)**. The 22-bp specification and the limited end-unwinding activity was previously observed in *Efa*Cas1-Cas2 **(Fig. 2d)** [19,20]. It is possible that Cas1-Cas2 has a common preference for prespacers containing a 22-bp mid-duplex (occasionally 23-bp as in *E. coli*), but has an idiosyncratic preference for 3′-overhang length **(Fig. 2e)**.

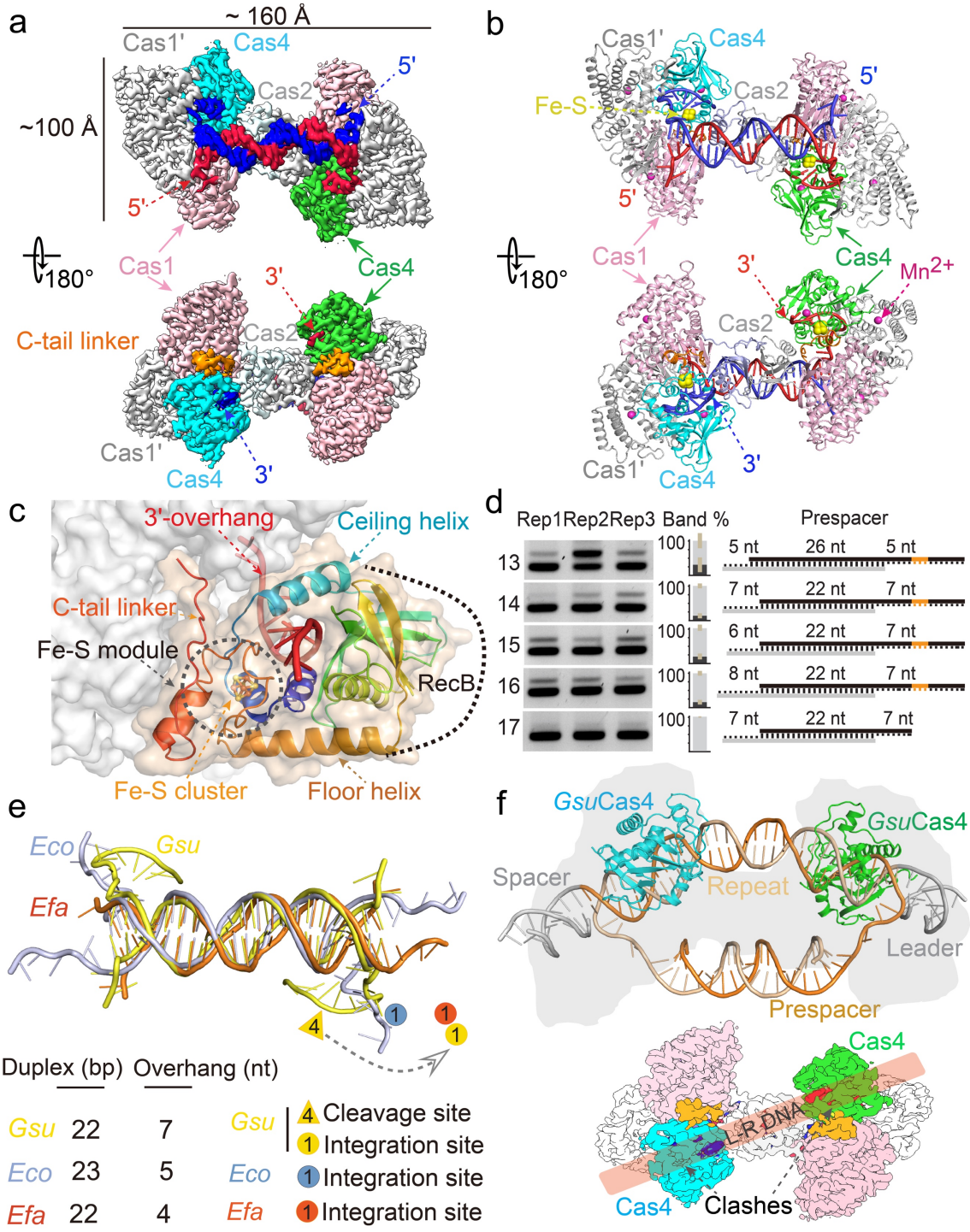


Figure 2. Insights from dual-PAM prespacer bound *Gsu*Cas4/Cas1-Cas2 structure. a, b. Cryo-EM density and cartoon representation of the dual-PAM bound *Gsu*Cas4/1-2 structure, respectively. **c.** Organization of Cas4 structural elements around the PAM-containing 3′-overhang. **d.** Validation that prespacers containing a 22-bp mid-duplex are actively acquired *in vivo*. **e.** Comparison of the 3′-overhang status among three prespacer-bound Cas1-Cas2 structures. The overhang is sequestered from the Cas1 integrase center by Cas4 in our structure. **f.** Superposition of our structure with *Efa*Cas1-Cas2 in the post-integration state. Note the PAM-recognizing Cas4 clashes with the repeat-spacer DNA entering into the integrase center in Cas1.

143

Among the four fused Cas4s, only the two non-catalytic Cas1-fused Cas4s are resolved in the EM structure, due to their involvement in PAM recognition. The other two are missing from the density presumably because they are not stably bound to the integrase core. Therefore, the natural tethering between Cas4 and Cas1 in our system does not alter the dynamic nature of the Cas4-Cas1-Cas2 interaction, and the mechanistic insights from this study are likely applicable to all Cas4 systems. The EM density allows an unambiguous tracing of the entire Cas4. Its structure aligns well with those of the stand-alone Cas4s [27,28] and the nuclease domains in helicase-nuclease fusion proteins AddAB [32], AdnAB [33] and eukaryotic Dna2 [34]. Interestingly, the Cas4 structure aligns poorly with the RecB nuclease in RecBCD; it agrees better with the RecB-like fold in RecC instead **(Extended Data Fig. 6a-c)** [35]. 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 center 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 to all Cas4 nucleases **(Fig. 2c)**. Cas4 connects to the non-catalytic Cas1 through a 20-amino acid (aa) fusion linker, which mediates the dynamic docking and dissociation of Cas4.

Importantly, the PAM-engaging Cas4s are wedged at the ventral side of the Cas1-Cas2 complex **(Fig. 2a-b)**. Because this region of Cas1-Cas2 is responsible for recruiting the leader-repeat DNA for spacer integration [18,20], it follows that the PAM-recognizing Cas4 sterically blocks integration from the PAM-side Cas1 **(Fig. 2e-f)**. Cas4 contacts both subunits of Cas1 through an extensive interface, many residues at the interface are conserved (**Extended Data Fig. 5a-c, 6b**). The Cas4-Cas2 interface involves favorable polar contacts between the ceiling helix in Cas4 (aa 39-50) and an outer helix in Cas2 (aa 42-53). It is difficult to identify key interface residues that are universally conserved across all Cas4 branches. There may exist evolutionary pressure to maintain idiosyncratic Cas4 and Cas1-Cas2 interactions in order to avoid crosstalk among coexisting CRISPR systems. If true, this scheme would be analogous to the highly selective binding relationship between Cas3 and Cascade [36].

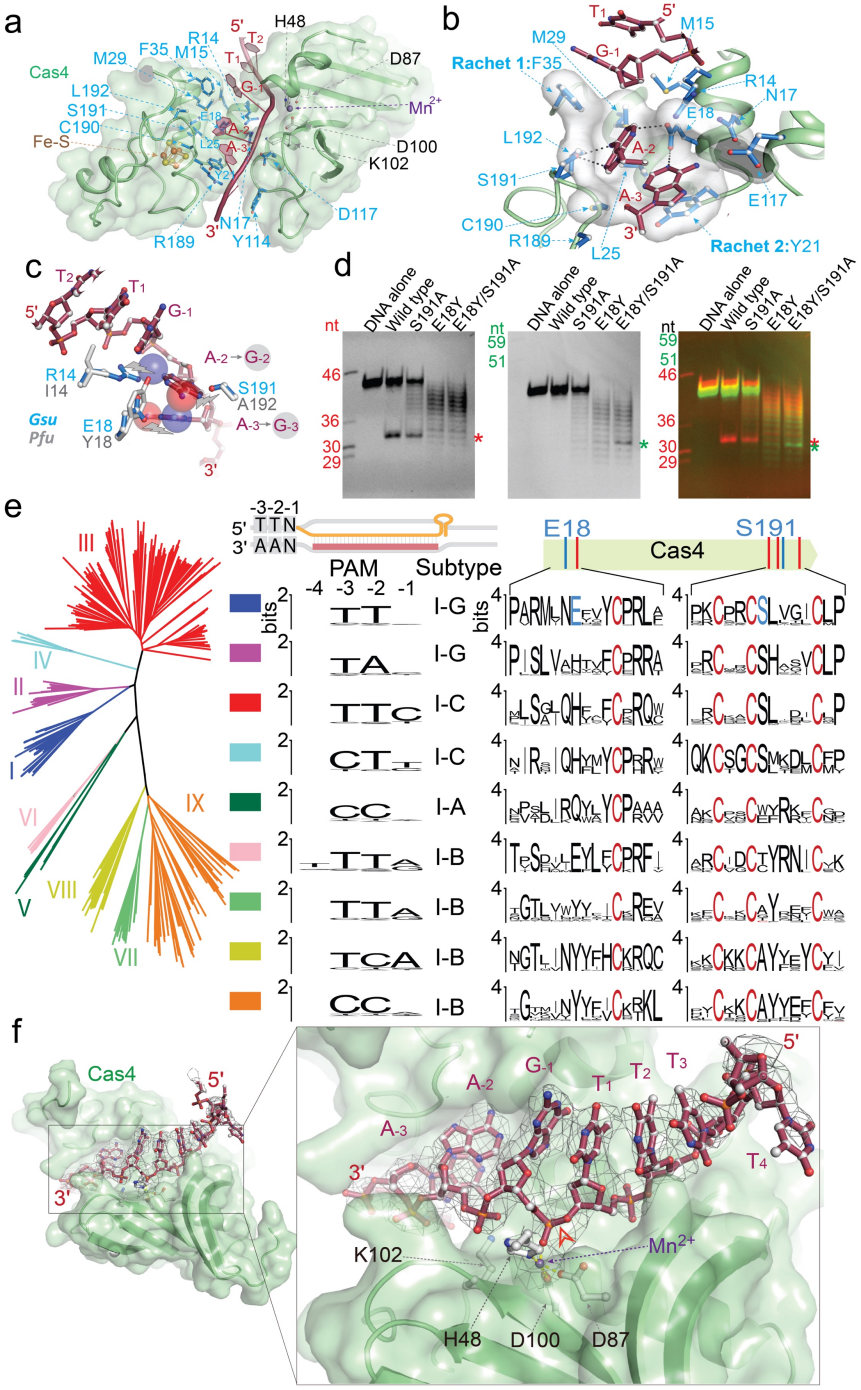


Figure 3. Cas4-mediated PAM-recognition and delayed overhang cleavage. a. PAM is caged inside a molecular ratchet in Cas4. Ceiling helix is omitted for better illustration of the narrow pathway for 3′-overhang. **b.** The di-adenosine PAM is surrounded by Van der Waals interactions that probe for shape complementarity, and by sequence-specific hydrogen-bonding interactions from E18 and S191. **c.** Modeling the impact of E18Y and S191A substitutions on recognizing *P. fur* instead of *G. sul* PAM. Specific atom changes in A-to-G switching (N6O substitution and N2 amine addition) are highlighted in colored balls. The steric clashes to *Pfu*PAM (lightening arrows) are partially relieved when substitutions are in place. **d.** Impact of E18Y and S191A substitutions on PAM cleavage activity. **e.** Correlations between PAM code in Cas4-containing CRISPR systems and the recognition motif consensus in Cas4. **f.** Arrangement of the Cas4 nuclease center. Cryo-EM density of the prespacer backbone is continuous, suggesting that the PAM-containing overhang is sequestered but not cleaved. Red arrow: labile bond.

Structural basis for Cas4-mediated PAM recognition

Despite extensive studies, the PAM recognition and cleavage mechanisms inside Cas4-Cas1-Cas2 remain unresolved. This EM 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′-d$A_7C_6T_5T_4T_3T_2T_1$**$G_{-1}A_{-2}A_{-3}$**T_{-4}) travels deep inside, protected from random nuclease cleavage. Stemming out of the mid-duplex, the first four nucleotides travel more or less along the same path towards the Cas1 active site, as seen in the previous Cas1-Cas2/prespacer structures [15,18,20]. However, nucleotides 5-11 detour through Cas4. They first travel on top of the RecB nuclease module, then enter into the narrow passage described previously **(Fig. 3a)**. Two hydrophobic residues F35 and Y21 interdigitates into the ssDNA before and after the narrow passage, forming molecular ratchets that cage the di-deoxyadenosine PAM (3′-$A_{-3}A_{-2}$) inside **(Fig. 3b)**. They likely enforce a ratcheting motion to slowly thread the 3′-overhang through, which allows the PAM sequence to be recognized and captured. Inside the narrow passage, the edges of A_{-2} and A_{-3} are surrounded by hydrophobic and long side chain residues (R14, M29, L25, L192, E117, N17, C190) that probe for shape complementarity. Deoxyguanosines would not fit comfortably in the same cage because their exocyclic N2 amines would cause steric clash; whereas the smaller-sized pyrimidines may slip through without a chance to establish favorable 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. 3b)**. They likely contribute significantly to the PAM specificity. Consistent with the *in vivo* data [26], there is no sequence-specific recognition to the first residue of PAM, G_{-1}. This nucleotide is excluded from the PAM-recognition box and points to the solvent **(Fig. 3b)**.

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. We first carried out a structure-guided mutagenesis to explore the possibility of switching the PAM specificity of *Gsu*Cas4 to that of *Pyrococcus furiosus* Cas4 **(Fig. 3c)**. *Pfu*Cas4 share 17% sequence identity with *Gsu*Cas4 and specifies a 5′-CCN PAM (3′-GGN in the overhang). We substituted the two sequence-specific PAM contacting residues in *Gsu*Cas4 to their counterparts in *Pfu*Cas4. In single substitutions, S191A retained *Gsu*-PAM specificity; cleavage activity was slightly compromised. E18Y lost sequence specific cleavage activity on both PAMs, and cleaved ssDNA distributively. Interestingly, the combination of these two substitutions resulted in a cleavage preference for *Pfu*-PAM, even though the activity was quite distributive. These results

suggest E18 plays a more important role than S191 in PAM recognition **(Fig. 3c)**. However, this partial success in switching PAM specificity did not further extend into *in vivo* spacer acquisition assays, which put further demand on prespacer/Cas4/Cas1-Cas2 stability and PAM cleavage timing. 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* **(Extended Data Fig. 5e)**. These results suggest that while the hydrogen-bonding interactions are important, a significant portion of the PAM specificity is likely conferred by the peripheral residues mediating hydrophobic interactions.

Next, we attempted to use bioinformatics to establish a correlation between structural features in Cas4 and PAM sequence variations. We first determined which PAM is used by different Cas4-containing CRISPR systems by mapping spacers in annotated and metagenomic databases. This led to a phylogenetic tree based on the alignment of Cas4s for which we could reliably couple PAM code with clades of Cas4s, sometimes from different CRISPR types that were using the same PAM [37] **(Fig. 3d)**. We expected that residues crucial for PAM selection would be conserved within the clades, but would differ between groups selecting a different PAM **(Fig. 3e)**. The E18 residue that is in contact with A_{-2} and A_{-3} is one such discriminant amino acid 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, which contacts A_{-2}, does not appear to be a discriminant residue as it was also found in Type I-G Cas4s specifying TAN PAMs. However, the highly conserved neighboring residue, L192, was exclusively found in Cas4 groups specifying a T on the -2 position of the PAM, including the less closely related Cas4s in Type I-C that either specify TTC or CTT. Therefore, the presence of L192 in Cas4 is a good predictor of a T on PAM-2. Similarly, informatics identified R14 and L25 as good predictors of T_{-2}. The reverse argument is not necessarily true. For example, not all PAMs containing a T_{-2} predict L192 in the corresponding Cas4s. The structure reveals that PAM is specified at least partially by hydrophobic contacts that select for shape complementarity **(Fig. 3b)**. In such cases a cluster of hydrophobic residues in Cas4 may be required to specify a PAM code, and their identity may not be unique.

PAM recognition delays 3′-overhang cleavage and prevents integration therein

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. 3f)**. The labile phosphate of G_{-1} is correctly positioned into the active site, which consists a DEK motif

(D87, D100, K102) and a histidine residue (H48), all of which are highly conserved among Cas4 and RecB family of nucleases [38]. These residues coordinate a catalytic metal ion, presumably Mn^{2+}, which is shown by the EM density to be tightly coordinated to the scissile phosphate. In the AdnAB structure, such active site configuration was shown to promote efficient DNA cleavage [33]. However, here the EM density clearly argues for an intact DNA substrate at the active site **(Fig. 3f)**. which was subsequently confirmed by denaturing PAGE **(Extended Data Fig. 5d)**. The exact cleavage inhibition mechanism in Cas4 will require a more focused analysis in the future. Among the many mechanistic possibilities, we speculate that it might be caused by the sub-optimally placed K102 residue in the DEK motif, which has been implicated as essential for Cas4 catalysis [26]. Rather than pointing towards the labile phosphate, K102 is twisted away by the residing β-strand. A minor conformational change in Cas4 may allow K102 to participate in PAM cleavage. Without PAM cleavage, Cas4 is locked in place and integration is blocked from taking place at the PAM side. This structural observation is in perfect agreement with the spacer directionality requirement in Type I CRISPRs.

Structure-guided reconstitution of directional integration

Next, to investigate the status of the non-PAM 3′-overhang, we determined the cryo-EM structure of the *Gsu*Cas4/Cas1-Cas2 complex programmed with a single-PAM containing prespacer. This led to an asymmetric full complex structure at 3.57 Å resolution, and a 3.56 Å assemble intermediate that will be discussed later **(Fig. 4; Extended Data Fig. 7)**. Whereas the PAM-side of *Gsu*Cas4/Cas1-Cas2 is blocked by a PAM-recognizing Cas4, 82.5% of the single-PAM particles do not have a docked Cas4 at the non-PAM side **(Fig. 4a)**; 17.5% contain a docked Cas4 evidenced by weak densities, however, the non-PAM overhang is not captured inside **(Extended Data Fig. 7c)**. In both cases, the non-PAM side Cas4/1 dimer density is weaker than the PAM-side counterpart, and a hinge motion is evident, anchored at 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. 7c)**. Because the non-PAM overhang lacks Cas4 protection, we reasoned that it may be trimmed to the optimal overhang length by certain 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 directly tested this mechanistic model. Indeed, *E. coli* SbcB (ExoI) protein could trim the non-PAM 3′-overhang to the preferred length of ~7-nt, **(Fig. 4b)**. Even the distributive cleavage pattern was categorically consistent with the spacer length distribution in the *G. sul* CRISPR systems **(Fig.**

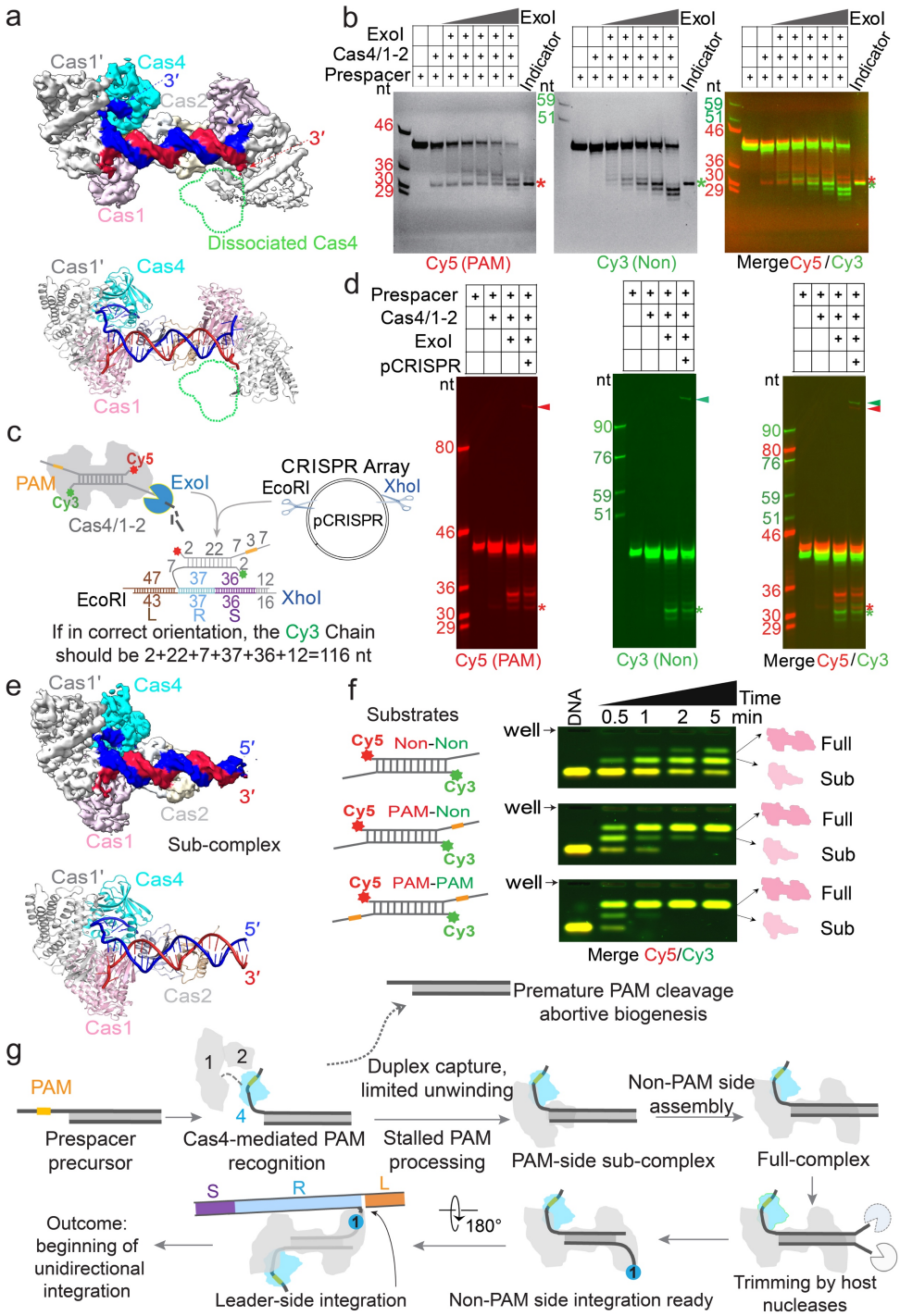


Figure 4. Mechanistic insights from the single-PAM prespacer bound *Gsu*Cas4/Cas1-Cas2 structure. a. Cryo-EM density (top) and structure (bottom) of the single-PAM prespacer bound GsuCas4/1-2 complex. Lack of Cas4 at the non-PAM side is highlighted. **b.** *E. coli* nuclease ExoI is capable of trimming the non-PAM overhang to the optimal length for integration. The PAM-side is protected. **c.** *In vitro* integration assay setup and the expected readout. **d.** Non-PAM overhang is unidirectionally integrated to the leader-proximal end of the leader-repeat upon ExoI trimming. **e.** Cryo-EM density (top) and structure (bottom) of a sub-complex specifically bound to the PAM-side prespacer. Cas4/1 dimer is missing from the non-PAM side. **f.** EMSA showing Cas4/1-2 is assembled sequentially and preferentially on PAM-containing prespacers. **g.** Mechanistic model explaining Cas4-dependent prespacer biogenesis and directional integration. See **Supplementary Movie S1** for details.

1c) [26]. In contrast, the PAM-side 3′-overhang was protected by the footprint of Cas4 in the same reaction **(Fig. 4b-c)**. 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 *Gsu*Cas4/Cas1-Cas2 readily integrated prespacers with optimal overhang length into a negatively supercoiled leader-repeat containing plasmid, it failed to do so on a linear target **(Extended Data Figs. 8a-d)**. This behavior 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 [39]. Given the limitation, in order to resolve the integration directionality, we first integrated a dual-fluorescently labeled prespacer into the plasmid, then restriction-digested out the leader-repeat region to determine the directionality based on the product size on denaturing polyacrylamide gel (**Extended Data Figs. 8c-f**). In control experiments, we verified *Gsu*Cas4/Cas1-Cas2's preference to integrate first into the leader-proximal side and confirmed the ability of the setup to distinguish integration directionality **(Extended Data Figs. 8e-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. 4c-d)**. This pattern is in agreement with the observed spacer directionality in the *G. sul* CRISPR array.

Intermediate structure generates insight about prespacer biogenesis

The PAM/non-PAM cryo-EM reconstruction further captured an important functional state, which corresponds to an intermediate assembly during prespacer biogenesis. The structure is of sufficient resolution to reveal that a $(Cas4/Cas1)_2$-$Cas2_2$ sub-complex has captured the PAM-side overhang and the duplexed region of the prespacer **(Fig 4e; Extended Data Fig. 7)**. While the PAM-side arrangement is essentially the same as in the previous structures, $(Cas4/Cas1)_2$ densities were absent from the non-PAM side. Using time-course and concentration-titration based electrophoretic mobility shift assays (EMSA), we confirmed that the *Gsu*Cas4/Cas1-Cas2 integrase indeed assembled in a stepwise fashion, and the PAM-containing overhang strongly promoted the assembly of the full-complex **(Fig 4f; Extended Data Fig. 5g)**. Collectively, these structural snapshots provide the much-needed temporal resolution for prespacer biogenesis. We conclude that the $(Cas4/Cas1)_2$-$Cas2_2$ sub-complex is capable of scouting 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 integration-competent $(Cas4/Cas1)_4$-$Cas2_2$ full complex. The conditional assembly process provides a quality-control mechanism to only recruit PAM-containing spacer precursors for further processing and integration **(Fig. 4g; Supplementary Movie S1)**. The

length of the precursor duplex is likely longer than the preferred length by $Cas1_4$-$Cas2_2$. In a previous study we explored this scenario and found that the host nucleases are capable of trimming dsDNA and ssDNA to the preferred prespacer specification as defined by the $Cas1_4$-$Cas2_2$ footprint [19].

Structural basis for mechanistic coupling between half-integration and PAM-cleavage

Having established that Cas4 defines the spacer directionality by blocking the PAM-side integrase center before integration, we next probed into the mechanism that relieves this blockage after half-integration, since the PAM-side prespacer needs to be processed and integrated to the opposite side of the CRISPR repeat to complete full integration. What serves as the molecular switch? We hypothesized that the half-integration itself may stimulate PAM cleavage and Cas4 dissociation. To test this, we programmed *Gsu*Cas4/Cas1-Cas2 to the half-integration state using an annealed prespacer and leader-repeat DNA that mimics the half-integration product [18], and monitored the extent of PAM processing and half-to-full integration transition at different conditions and over time **(Extended Data Fig. 9a-j)**. Indeed, half-integration led to faster and higher extent of PAM cleavage, and full integration quickly followed **(Fig. 5a; Extended Data Fig. 9b)**. As controls, PAM cleavage was much slower and weaker when the leader-repeat DNA was absent **(Fig. 5a)**, or when the half-integration did not take place **(Extended Data Fig. 8a)**.

Next, we sought to provide the structural basis for the observed mechanistic coupling. The reacted sample in **Extended Data Fig. 9k-m** was snap-frozen for cryo-EM analysis **(Extended Data Figs. 9k-m)**. We were able to capture multiple conformational states from the single particle reconstruction, which we interpret as representing three different functional states during the half-to-full integration transition. The more populated state was solved at higher resolution since more particles were available for 3D reconstruction, and vice versa **(Extended Data Figs. 10)**. The three states differ significantly in their spacer-side contacts and in Cas4 and integration status. In what we interpret as an early state (5.83 Å in resolution), 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. 5b-c; Extended Data Fig. 11)**. The DNA density is weak, suggesting that it may dynamically sample multiple conformations, some of these motions may involve

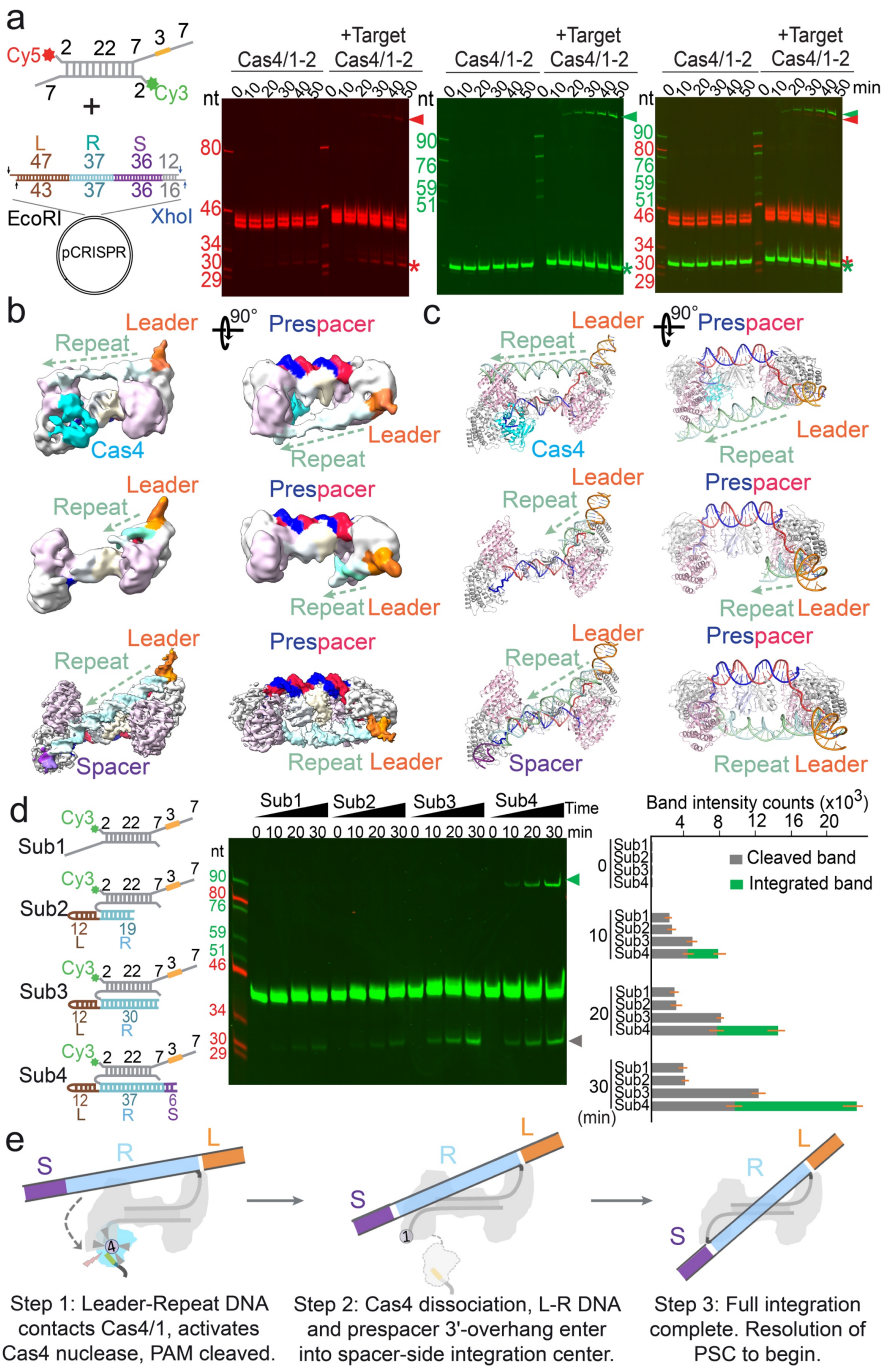


Figure 5. Snapshots of *Gsu*Cas4/Cas1-Cas2 in coupling half-integration with PAM cleavage to achieve full-integration. a. Time-course experiments showing non-PAM side half-integration stimulates PAM cleavage. Full integration quickly follows. **b.** Three cryo-EM snapshots and **c.** corresponding structure models captured from Cas4/1-2 incubated with half-integration mimicking substrate. They represent sequence of events from the initial blockage of spacer-side integration site by PAM-bound Cas4 (top), PAM cleavage triggered Cas4 dissociation (middle), and the post full integration state (bottom). Resolutions of the three cryo-EM reconstructions are 5.83, 5.76, and 3.81 Å, respectively. **d.** Biochemistry showing that PAM cleavage is stimulated by leader-repeat DNA contacting the spacer-side Cas4/1. Left: substrate design; middle: urea-PAGE; right: quantification of PAM cleavage bands. **e.** Diagram explaining the mechanistic coupling between half-integration, PAM cleavage, Cas4 dissociation, and full-integration. See **Supplementary Movie S2** for details.

Cas4 contacts. In the 5.76 Å intermediate state, the Cas4 density disappears, and the density corresponding to the cleaved prespacer overhang appears to point towards the exposed Cas1 active site, although it is quite weak and choppy. With Cas4 out of the way, the CRISPR repeat DNA projected from the leader-side Cas1 contacts the Cas2 dimer in the middle and appears to further point towards the spacer-side integration center, however, its density is too degraded for model building **(Fig. 5b-c; Extended Data Fig. 11)**. This suggests that even with Cas4 out of the way, spacer-side CRISPR DNA capture and integration is inefficient, presumably because the favorable leader-sequence contacts are missing here [19]. Lastly, we captured a 3.81 Å snapshot of the full-integration state. EM densities clearly reveals that the CRISPR repeat DNA has been docked into the spacer-side integration center, and a continuous density connects it with the 3′-overhang, suggesting that full-integration has taken place **(Fig. 5b-c)**. This snapshot is architecturally similar to the previously determined post-integration complexes from *E. fae* [20], however, the leader-repeat DNA in the *G. sul* structure is not as sharply kinked at the Cas2 binding site as seen in the *E. fae* structure. The entire leader- repeat DNA is contacted in a quasi-symmetric fashion at the following four regions **(Fig. 5b-c; Extended Data Figs. 10-11)**. 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 [18-20]. Lacking such sequence at the spacer-side, DNA density is degenerate and DNA bending is not observed. The α-helix insertion most likely does not take place at the spacer side. The inverted repeats at the border region of the CRISPR repeat are recognized at the major groove region by the catalytic Histidine-containing loop in Cas1 [20]. The following minor groove is recognized by a conserved “PRPI” motif in the Cas4-Cas1 fusion linker, which is not exposed when Cas4 is docked. Lastly, 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 **(Extended Data Fig. 11b-f)**. Connecting the dots together, the 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 ensures PAM-specific spacer-side integration.

How does the leader-side integration activate the PAM-cleavage by Cas4? The two active sites are located ~120 Å apart. There are at least two mechanistic possibilities: 1) the leader-half integration may trigger a global conformational change that allosterically activates Cas4; 2) the

physical contacts by the integrated leader-repeat DNA somehow activates Cas4. The allosteric activation model was deemed unlikely because no significant conformational change in Cas1-Cas2 was observed among apo, half- and full-integration structures, although we cannot completely rule out the possibility that changes in the extent of hinge motions may play a role. To further probe whether the physical contact by the leader-repeat DNA might activate Cas4, we systematically shortened the leader-repeat DNA in the previous integration assay setup **(Fig. 5d)**. Results revealed a strong correlation. When the leader-repeat was too short to reach spacer-side Cas4/1 (Sub2: 19-bp CRISPR repeat), the extent of PAM cleavage was indistinguishable from that in the prespacer-only control. When the leader-repeat is long enough to reach the spacer-side Cas4/1 (Sub3: 30-bp CRISPR repeat), the PAM cleavage was significantly enhanced, even without the spacer-side integration **(Fig. 5d)**. 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 center and allows full integration **(Fig. 5e; Supplementary Movie S2)**. 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.

Discussion

In summary, we provide a comprehensive set of mechanism 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, whose activity 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 unexpected molecular constipation is the cornerstone for productive prespacer biogenesis and functional spacer integration in Type I and V CRISPR systems. We provide direct and high-resolution evidence that PAM recognition and the subsequent molecular constipation takes place early during prespacer biogenesis, in essence it serves as a gatekeeper to channel only the productive precursor 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 to the leader-side CRISPR repeat. Moreover, we reveal that the leader-side integration efficiently activates the PAM cleavage activity of Cas4 and causes Cas4 dissociation, which in turn derepresses the PAM-side Cas1 integrase and allows the half-to-full integration transition. Collectively, the series of structural snapshots depicts the entire directional integration process for the Cas4-containing Type I and V CRISPR systems. Exactly how spacer directionality is

established in Cas4-less CRISPR systems requires further investigation[15,16,40]. In Type I-E CRISPR, such mechanism has been shown to involve Cas1-mediated PAM sequestration and integration-dependent desequestration[21]. Therefore, the PAM-dependent blockage/activation of the two integration centers in Cas1-Cas2 may be a universal theme to achieve directionally spacer integration.

The structural similarity of Cas4 to the nuclease domains of AddAB/AdnAB and a structural domain in the equivalent location in RecBCD shed light into the ancient function of Cas4 in spacer acquisition. These helicase-nuclease machines not only play essential roles in homology-directed repair, but also provide a line of innate immunity for bacteria by preferentially degrading linear DNA lacking chi sites, which are more likely of foreign origin. Functional interactions between RecBCD/AddAB and Cas1-Cas2 mediated spacer acquisition have been noted in previous studies [41,42]. 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 dramatically increased the productive spacer acquisition in the ancient CRISPR systems. It is possible that the ancient Cas1-Cas2 relied on RecBCD or AddAB for spacer precursors so heavily, that it started to establish a physical interaction with the nuclease domain to facilitate the process. It eventually led to the hijacking of this host nuclease domain into the *cas* operon as *cas4*. A similar process may have taken place for other nucleases such as *dnaQ* [21,43,44].

Methods

PAM prediction

221,089 unique spacers along with genome source, *cas* gene information, and repeat sequence were obtained from CRISPRCasDb [45] in February 2020. These spacers were blasted against our own sequence database containing all sequences from the NCBI nucleotide database [46,47], environmental nucleotide database [48], PHASTER [49], Mgnify [50], IMG/M [51], IMG/Vr [52], HuVirDb [53], HMP database [54], and data from Pasolli *et al.*[55]. All databases were accessed in February 2020.

Hits between spacers and sequences from the aforementioned nucleotide databases were obtained using the BLASTN program [56] version 2.10.0, which was run with parameters word_size 10, gap open 10, penalty 1 and an e-value cutoff 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 higher than 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 higher than 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 nucleotides.

Highly similar repeat sequences of the same length were clustered using CD-HIT [57] with a 90% identity threshold. To increase the number of aligned sequences for PAM 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 been shown previously [58,59]. 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 [60] from a MAFFT alignment of all Cas4 sequences [61]. The sequence logos were generated with Berkeley weblogo [62] 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 and growth conditions

Escherichia coli strains Dh5α and BL21-AI were grown at 37 ˚C in Lysogenic Broth (LB) media with shaking or on LB agar (LBA) plates containing 1.5% (w/v) agar. When required, media was supplemented with 50 µg/ml spectinomycin, 100 µg/ml ampicillin, 50 µg/ml Kanamycin, 1 mM IPTG, and 0.2% (w/v) L-arabinose (see **Supplementary Table 1** for plasmids and their corresponding selection markers).

Plasmid construction

Plasmids used in this work are listed in **Supplementary Table 1**. All cloning steps were performed in *E. coli* Dh5α. 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 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 (Macrogen Europe, Netherlands). Bacterial transformations were carried out by electroporation (200 Ω, 25 µF, 2.5 kV) using an ECM 630 electroporator (BTX Harvard Apparatus), and transformants were selected on LBA supplemented with the appropriate antibiotics.

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 2.5h of growth, the expression of *cas* genes was induced with IPTG and L-arabinose, and the cultures were incubated for additional 2h. 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 [25]. Samples were run on 2% agarose gels and visualization 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 [25,63].

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 flowcell (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 analyzed by pairing and merging the reads using Geneious 9.0.5. Acquired spacers were extracted and analyzed as described previously [25].

Cloning, expression and purification

Full-length *Gsu*Cas4/1 (Gsu0057 in KEGG) gene was cloned from *Geobacter sulfurreducens* genomic DNA into pET28a -His_6-Twin-Strep-SUMO vectors (Kan^R) or pGEX-41-

T-His_6-Flag-GST (Amp^R), between BamHI and XhoI sites. Sequence-verified plasmids were transformed into *E. coli* BL21 (DE3) star cells under the appropriate antibiotic selection. A 6 liters cell culture was grown in LB medium at 37 ˚C until an optical density of 0.5 at 600 nm. The culture temperature was then reduced to 16 ˚C and incubated for additional 2 hours. Expression was induced with 0.5 mM isopropyl-β-D-thiogalactopyranoside (IPTG), 0.2 mg/mL ferrous sulfate (Fisher) and 0.4 mg/mL L-cysteine (MP biomedicals) at 16 ˚C overnight. Cells were harvested by centrifugation and resuspended in 100 mL buffer A containing 50 mM HEPES pH 7.5, and 500 mM NaCl, 10% glycerol, and 5 mM TCEP. Cells were lysed by sonication, and the lysate was 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 GST column (GST tagged expression). After washing with 100 ml of buffer A, the protein was eluted with 20 ml buffer B (50 mM HEPES pH 7.5, 500 mM NaCl, 10% glycerol, 300 mM imidazole, and 5 mM TCEP for SUMO tagged purification and 50 mM HEPES pH 7.5, 500 mM NaCl, 10% glycerol, 15 mM reduced GSH, and 5 mM TCEP for GST tagged purification), then incubated with SUMO-protease or 3C protease at 4 ˚C for 2 hours. The sample was then concentrated to 2 ml and loaded onto a Superdex 200 16/60 size-exclusion 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 usage.

Full-length *cas2* (Gsu0058 in KEGG) genes were cloned from *Geobacter sulfurreducens* genomic DNA into His_6-Twin-Strep-SUMO-pET28a vectors (Kan^R) between BamHI and XhoI sites. Sequence-verified plasmids were transformed into *E. coli* BL21 (DE3) star cells. A 4 liters cell culture was grown in LB medium at 37 ˚C until an optical density of 0.8 at 600 nm. Expression was induced by adding IPTG to a final concentration of 0.5 mM at 25 ˚C overnight. Cells were harvested by centrifugation and lysed by sonication in 80 ml buffer A containing 50 mM HEPES pH 7.5, 20 mM imidazole and 500 mM NaCl, 10% glycerol, and 2 mM B-ME. The lysate was centrifuged at 17,000 *g* for 50 min at 4 ˚C, and the supernatant was applied onto the pre-equilibrated 4 mL Ni-NTA column. After washing with 100 ml of buffer A, the protein was eluted with 20 ml buffer B (50 mM HEPES pH 7.5, 500 mM NaCl, 10% glycerol, 300 mM imidazole, and 2 mM B-ME), and incubated with SUMO-protease at 4 ˚C for 3 hours. The tag cleaved Cas2 proteins were purified on Superdex 200 16/60 equilibrated with buffer C (10 mM HEPES pH 7.5, 500 mM NaCl), the peak fractions were pooled and snap-frozen in liquid nitrogen for later usage.

Affinity pull-down assay

15 µg GST-tagged Cas4/1 and 30 µg untagged Cas2 were mixed and incubated with 10 µL GST resin at 4 ˚C for 30 min in different salt concentration buffer (50 mM HEPES pH7.5, 10% glycerol, 5 mM TCEP, and 150/300/500 mM NaCl) in presence or absence of prespacer, in a total assay volume of 50 µL. The GST resin was pelleted by centrifugation at ~100 *g* for 30 seconds, 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 labeled prespacer substrate preparation

Fluorescent DNA oligos **(Supplementary Table 2)** for biochemistry were synthesized (Integrated DNA Technologies) with either a /5AmMC6/ or /3AmMO/ label, fluorescently labeled in-house, and annealed at equimolar amount, and native PAGE purified 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 Cas4/1, 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 60 min, reactions were quenched by vortexing with 20µL of phenol-chloroform. The extracted aqueous phases were mixed with equi-volume of 100% formamide and separated on 13% urea-PAGE. Signals from each fluorescent dye were recorded at its corresponding excitation wavelength using a ChemiDoc imaging system (Bio-Rad). The $KMnO_4$ foot printing assay was carried out following previously published protocols [64].

Reconstitution of prespacer bound/integration Cas4/1-2 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 reconstitute 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. 10 nM of prespacer were incubated with 250 nM Cas4/1-2 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 minutes, 300 ng of pCRISPR plasmid was introduced into the reaction. Integration was allowed at 37 ˚C for 60 min, 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 equi-volume of formamide, separated on 13% urea-PAGE, and scanned on ChemiDoc imaging system.

ExoI trimming and follow-up integration assays

10 nM of prespacer were pre-incubated with 250 nM of Cas4/1-2 complex at 37 ˚C for 5 minutes 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 MgCl2). The 2-fold ExoI dilution series in Fig. 4b was prepared by dilution *E. coli* ExoI (NEB, 20 Units/µL) to a final concentration of 0.2, 0.1, 0.05, 0.025, 0.0125 Units/µL in each reaction. The 1/10 and 1/50 ExoI concentrations in the Extended Data Fig. 9a correspond to 0.1, 0.02 Units/µL. The ExoI concentration in the Extended Data Fig. 9b was 0.1 Units/µL across. In reactions where the trimming and integration were coupled, 300 ng of pCRISPR plasmid (~ 5 nM final concentration) was introduced at the same time with ExoI into the reaction. After incubation, the reaction was quenched by mixing with equi-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 ChemiDoc imaging system (Bio-Rad), as described above.

Electrophoretic mobility shift assay

2 nM final concentration of fluorescently labeled prespacer DNA was incubated with an increasing concentration of Cas4/1-2 complex for 15 minutes (in concentration titration experiments), or with 50 nM Cas4/1-2 complex for 0.5, 1, 2, 5 minutes (in time-course experiments) at 4 ˚C in a total 20 µL system 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 1x 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

4 µL of 0.01 mg/mL prespacer-bound Cas4/1-2 complex was applied to a glow-discharged copper 400-mesh continuous carbon grid. After a 30-second 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 60 seconds. 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 keV with a direct magnification of x140000 (3.2 A˚ pixel size) by AMT camera system. Each image was acquired using a 800 ms exposure time and -1 to -2 mm defocus setting. Data processing and 2D classification were performed on CyoSPARC software.

Cryo-EM data acquisition

4 µL of 0.6 mg/mL SEC-purified prespacer-bound or half-integration mimicking substrate-bound Cas4/1-2 complexes were applied to a Quantifoil holey carbon grid (1.2/1.3, 400 mesh) which had been glow-discharged for 30s. 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 microscope (Thermo Fisher) equipped with a K3 Summit direct electron detector (Gatan). The total exposure time of each movie stack was ~ 3.5 s, leading to a total accumulated dose of 50 electrons per A˚ which fractionated into 50 frames. Dose fractionated super-resolution movie stacks collected from the K3 Summit direct electron detector were 1x binned to a pixel size of 1.234 A˚. The defocus value was set between −1.5 µm to −3.5 µm.

Cryo-EM data processing

Motion correction, CTF-estimation, blob particle picking, 2D classification, 3D classification and non-uniform 3D refinement were performed in cryoSPARC v.2 [65]. Refinements followed the standard procedure, a series of 2D and 3D classifications with *C1* symmetry were performed as shown in Extended Data Fig. 4a, Extended Data Fig. 7 and Extended Data Fig. 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 [66] 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 figures and Supplementary table 3.

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.

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, A.C.H. for *in vivo* and bioinformatics analyses; K.N., C.H. for structure building and refinement; and S.R.B. for assistance in cryo-EM work. A.K., C.H. wrote the manuscript, with input from S.J.J.B., J.N.A.V. and A.R.C.

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 Sebastian N. Kieper, Romano Miojevic, Mariena Ramos, Gabriel Schuler and Katherine Spoth for helpful discussions, advice and technical assistance. The authors declare no competing financial interests. Correspondence should be addressed to A.K. and S.J.J.B. (ailong.ke@cornell.edu; stanbrouns@gmail.com). Reprints and permissions information are available upon request.

References

1 Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709-1712, doi:10.1126/science.1138140 (2007).

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

3 Nuñez, J. K. *et al.* Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. *Nature structural & molecular biology* **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, doi:10.1099/mic.0.023960-0 (2009).

6 Marraffini, L. A. & Sontheimer, E. J. Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature* **463**, 568-571, doi:10.1038/nature08703 (2010).

7 Jackson, S. A. *et al.* CRISPR-Cas: Adapting to change. *Science* **356**, doi:ARTN eaal505610.1126/science.aal5056 (2017).

8 McGinn, J. & Marraffini, L. A. Molecular mechanisms of CRISPR-Cas spacer acquisition. *Nat Rev Microbiol* **17**, 7-12, doi:10.1038/s41579-018-0071-7 (2019).

9 Deveau, H. *et al.* Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *J Bacteriol* **190**, 1390-1400, doi:10.1128/JB.01412-07 (2008).

10 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, doi:10.1371/journal.pone.0050797 (2012).

11 Mojica, F. J. M., Diez-Villasenor, C., Garcia-Martinez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology (Reading)* **155**, 733-740, doi:10.1099/mic.0.023960-0 (2009).

12 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, doi:10.1016/j.molcel.2019.10.021 (2020).

13 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, doi:10.1038/s41579-019-0299-x (2020).

14 Hudaiberdiev, S. *et al.* Phylogenomics of Cas4 family nucleases. *Bmc Evol Biol* **17**, doi:ARTN 23210.1186/s12862-017-1081-1 (2017).

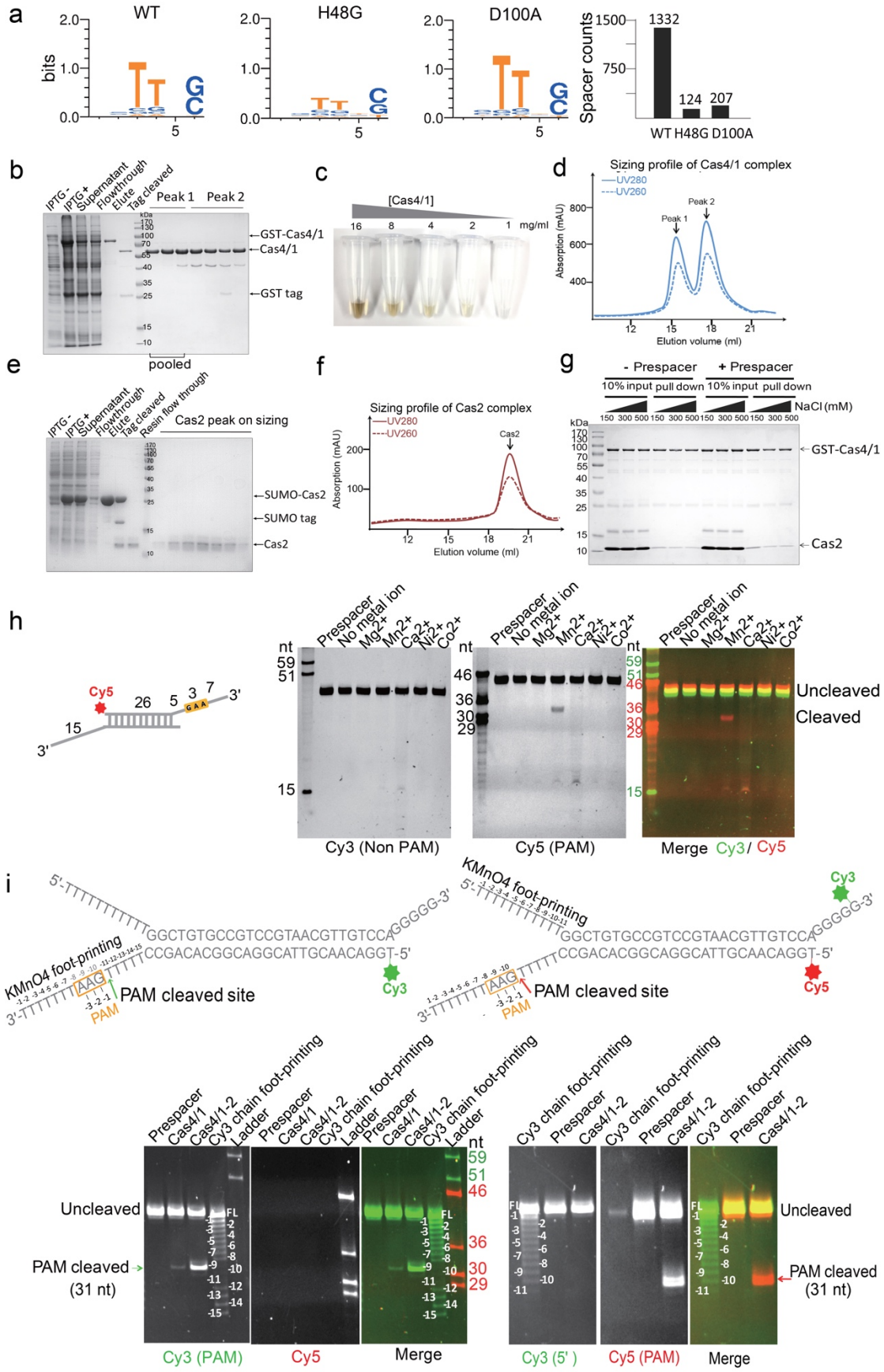
15 Nuñez, J. K., Harrington, L. B., Kranzusch, P. J., Engelman, A. N. & Doudna, J. A. Foreign DNA capture during CRISPR–Cas adaptive immunity. *Nature* **527**, 535-538 (2015).

16 Wang, J. *et al.* Structural and mechanistic basis of PAM-dependent spacer acquisition in CRISPR-Cas systems. *Cell* **163**, 840-853 (2015).
17 Wright, A. V. & Doudna, J. A. Protecting genome integrity during CRISPR immune adaptation. *Nature Structural & Molecular Biology* (2016).
18 Wright, A. V. *et al.* Structures of the CRISPR genome integration complex. *Science* **357**, 1113-1118, doi:10.1126/science.aao0679 (2017).
19 Budhathoki, J. B. *et al.* Real-time observation of CRISPR spacer acquisition by Cas1-Cas2 integrase. *Nat Struct Mol Biol* **27**, 489-499, doi:10.1038/s41594-020-0415-7 (2020).
20 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, doi:10.1038/nature24020 (2017).
21 Kim, S. *et al.* Selective loading and processing of prespacers for precise CRISPR adaptation. *Nature* **579**, 141-+, doi:10.1038/s41586-020-2018-1 (2020).
22 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, doi:10.1093/nar/gkt1154 (2014).
23 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, doi:10.1093/nar/gkx612 (2017).
24 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. *Molecular Cell* **70**, 814-+, doi:10.1016/j.molcel.2018.05.002 (2018).
25 Kieper, S. N. *et al.* Cas4 Facilitates PAM-Compatible Spacer Selection during CRISPR Adaptation. *Cell Reports* **22**, 3377-3384, doi:10.1016/j.celrep.2018.02.103 (2018).
26 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, doi:10.1093/nar/gkz217 (2019).
27 Lemak, S. *et al.* 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, doi:10.1021/ja408729b (2013).
28 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, doi:10.1093/nar/gku797 (2014).
29 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**, doi:ARTN e4723210.1371/journal.pone.0047232 (2012).
30 Lee, H., Dhingra, Y. & Sashital, D. G. The Cas4-Cas1-Cas2 complex mediates precise prespacer processing during CRISPR adaptation. *Elife* **8**, doi:ARTN e4424810.7554/eLife.44248 (2019).
31 Lee, H., Zhou, Y., Taylor, D. W. & Sashital, D. G. Cas4-Dependent Prespacer Processing Ensures High-Fidelity Programming of CRISPR Arrays. *Molecular Cell* **70**, 48-+, doi:10.1016/j.molcel.2018.03.003 (2018).
32 Krajewski, W. W. *et al.* Structural basis for translocation by AddAB helicase-nuclease and its arrest at chi sites. *Nature* **508**, 416-419, doi:10.1038/nature13037 (2014).
33 Jia, N. *et al.* Structures and single-molecule analysis of bacterial motor nuclease AdnAB illuminate the mechanism of DNA double-strand break resection. *Proceedings of the National Academy of Sciences of the United States of America* **116**, 24507-24516, doi:10.1073/pnas.1913546116 (2019).
34 Zhou, C., Pourmal, S. & Pavletich, N. P. Dna2 nuclease-helicase structure, mechanism and regulation by Rpa. *Elife* **4**, doi:10.7554/eLife.09832 (2015).

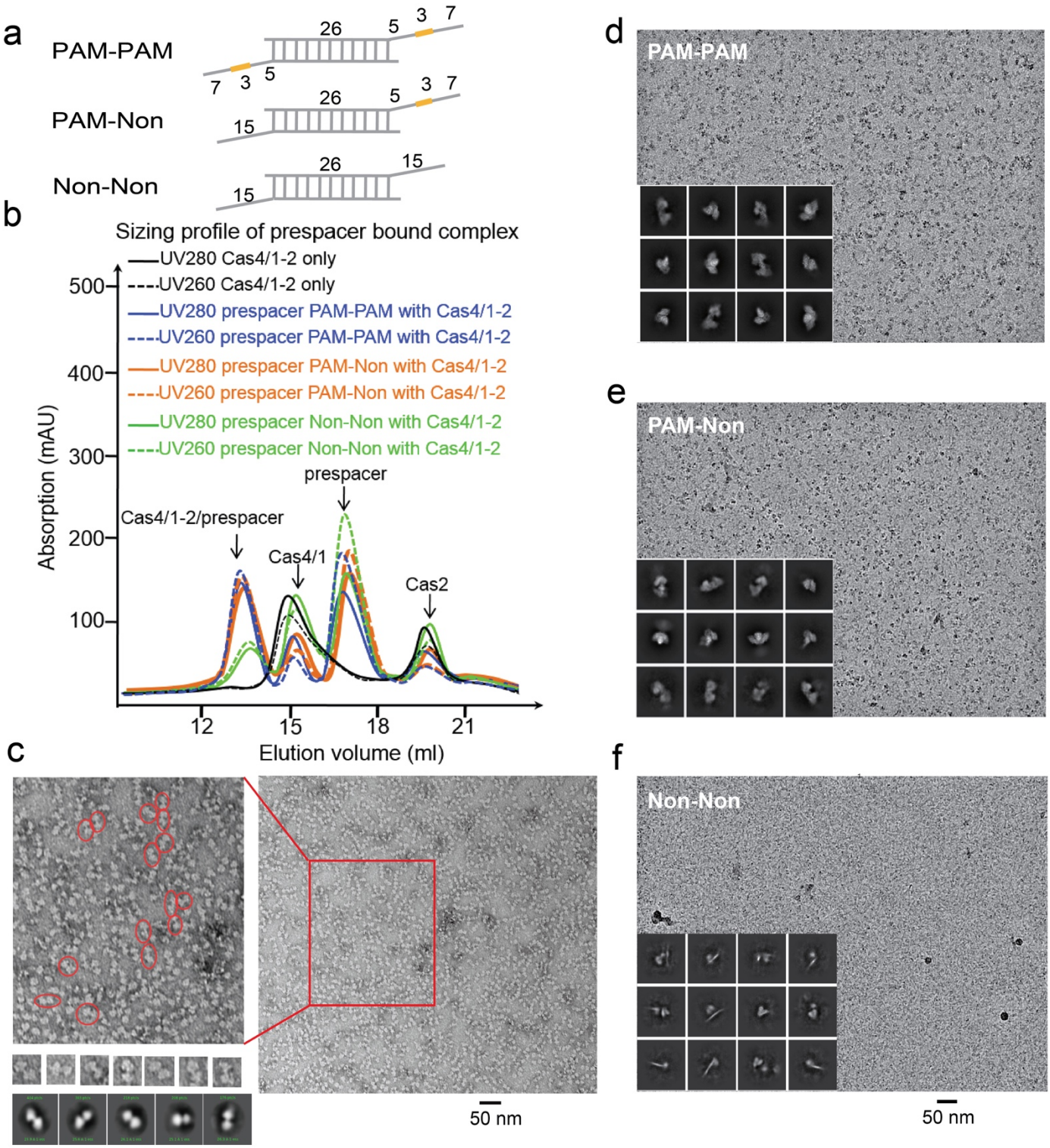
35 Singleton, M. R., Dillingham, M. S., Gaudier, M., Kowalczykowski, S. C. & Wigley, D. B. Crystal structure of RecBCD enzyme reveals a machine for processing DNA breaks. *Nature* **432**, 187-193, doi:10.1038/nature02988 (2004).
36 Xiao, Y., Luo, M., Dolan, A. E., Liao, M. & Ke, A. Structure basis for RNA-guided DNA degradation by Cascade and Cas3. *Science* **361**, doi:10.1126/science.aat0839 (2018).
37 Shah, S. A., Erdmann, S., Mojica, F. J. & Garrett, R. A. Protospacer recognition motifs: mixed identities and functional diversity. *RNA Biol* **10**, 891-899, doi:10.4161/rna.23764 (2013).
38 Yang, W. Nucleases: diversity of structure, function and mechanism. *Q Rev Biophys* **44**, 1-93, doi:10.1017/S0033583510000181 (2011).
39 Nuñez, J. K., Bai, L., Harrington, L. B., Hinder, T. L. & Doudna, J. A. CRISPR immunological memory requires a host factor for specificity. *Molecular cell* **62**, 824-833 (2016).
40 Heler, R. *et al.* Cas9 specifies functional viral targets during CRISPR-Cas adaptation. *Nature* **519**, 199-202, doi:10.1038/nature14245 (2015).
41 Levy, A. *et al.* CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* **520**, 505-+, doi:10.1038/nature14302 (2015).
42 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, doi:10.1038/nature21719 (2017).
43 Drabavicius, G. *et al.* DnaQ exonuclease-like domain of Cas2 promotes spacer integration in a type I-E CRISPR-Cas system. *EMBO Rep* **19**, doi:10.15252/embr.201745543 (2018).
44 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. *The Journal of biological chemistry* **295**, 3403-3414, doi:10.1074/jbc.RA119.012196 (2020).
45 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, doi:10.1093/nar/gkz915 (2020).
46 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**, D501-504, doi:10.1093/nar/gki025 (2005).
47 Benson, D. A. *et al.* GenBank. *Nucleic Acids Res* **46**, D41-D47, doi:10.1093/nar/gkx1094 (2018).
48 Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **37**, D5-15, doi:10.1093/nar/gkn741 (2009).
49 Arndt, D. *et al.* PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* **44**, W16-21, doi:10.1093/nar/gkw387 (2016).
50 Mitchell, A. L. *et al.* MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res* **48**, D570-D578, doi:10.1093/nar/gkz1035 (2020).
51 Chen, I. A. *et al.* IMG/M: integrated genome and metagenome comparative data analysis system. *Nucleic Acids Res* **45**, D507-D516, doi:10.1093/nar/gkw929 (2017).
52 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, doi:10.1093/nar/gky1127 (2019).
53 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, doi:10.1016/j.chom.2019.08.008 (2019).

54 Group, N. H. W. *et al.* The NIH Human Microbiome Project. *Genome Res* **19**, 2317-2323, doi:10.1101/gr.096651.109 (2009).
55 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, doi:10.1016/j.cell.2019.01.001 (2019).
56 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403-410, doi:10.1016/S0022-2836(05)80360-2 (1990).
57 Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152, doi:10.1093/bioinformatics/bts565 (2012).
58 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, doi:10.1093/nar/gkt606 (2013).
59 Alkhnbashi, O. S. *et al.* CRISPRstrand: predicting repeat orientations to determine the crRNA-encoding strand at CRISPR loci. *Bioinformatics* **30**, i489-496, doi:10.1093/bioinformatics/btu459 (2014).
60 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, doi:10.1093/sysbio/syq010 (2010).
61 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780, doi:10.1093/molbev/mst010 (2013).
62 Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res* **14**, 1188-1190, doi:10.1101/gr.849004 (2004).
63 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, doi:10.1038/s41596-018-0123-5 (2019).
64 Xiao, Y. *et al.* Structure Basis for Directional R-loop Formation and Substrate Handover Mechanisms in Type I CRISPR-Cas System. *Cell* **170**, 48-60 e11, doi:10.1016/j.cell.2017.06.012 (2017).
65 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, doi:10.1038/nmeth.4169 (2017).
66 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, doi:10.1155/2020/6250425 (2020).

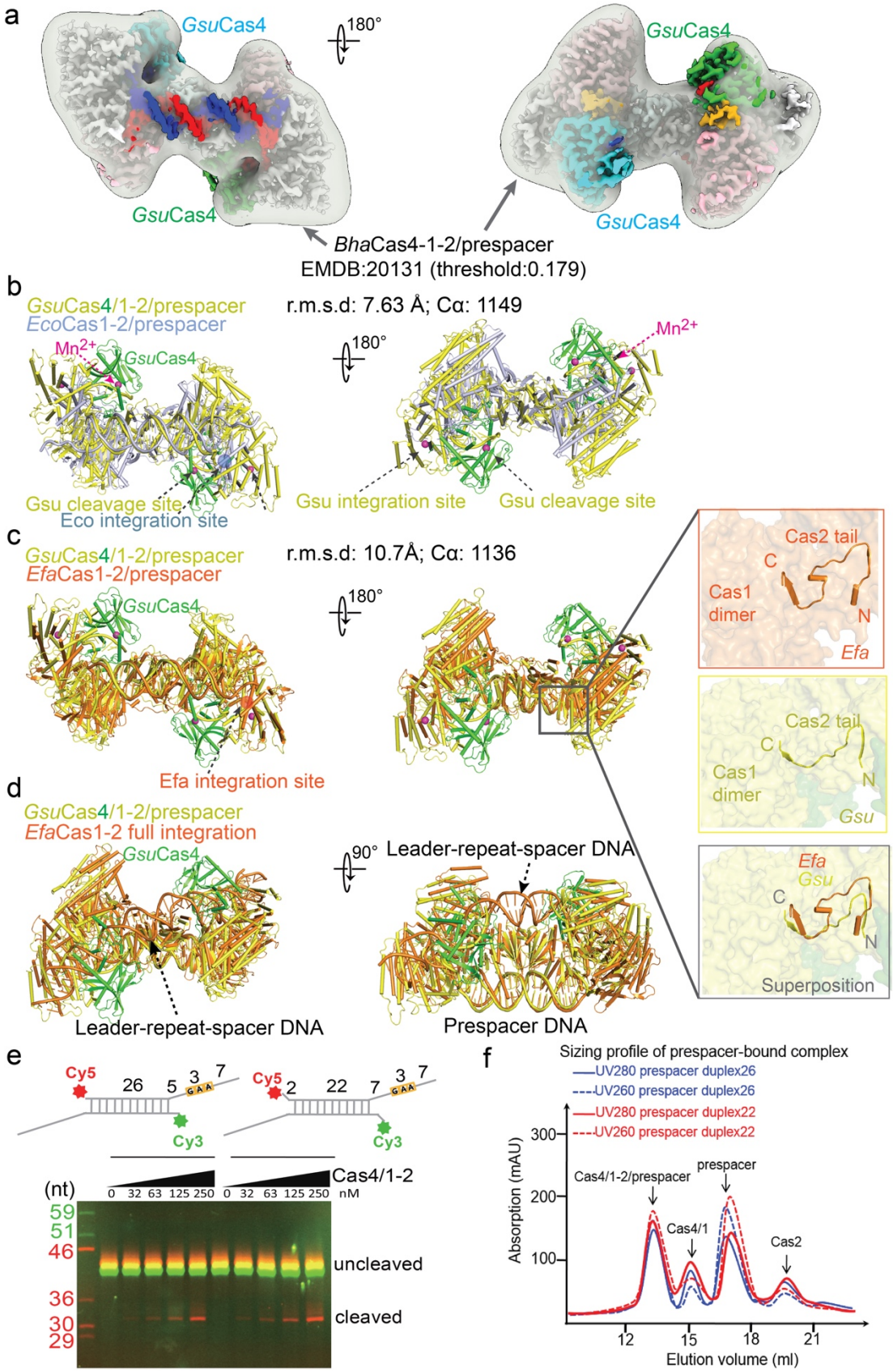
Extended Data Figures



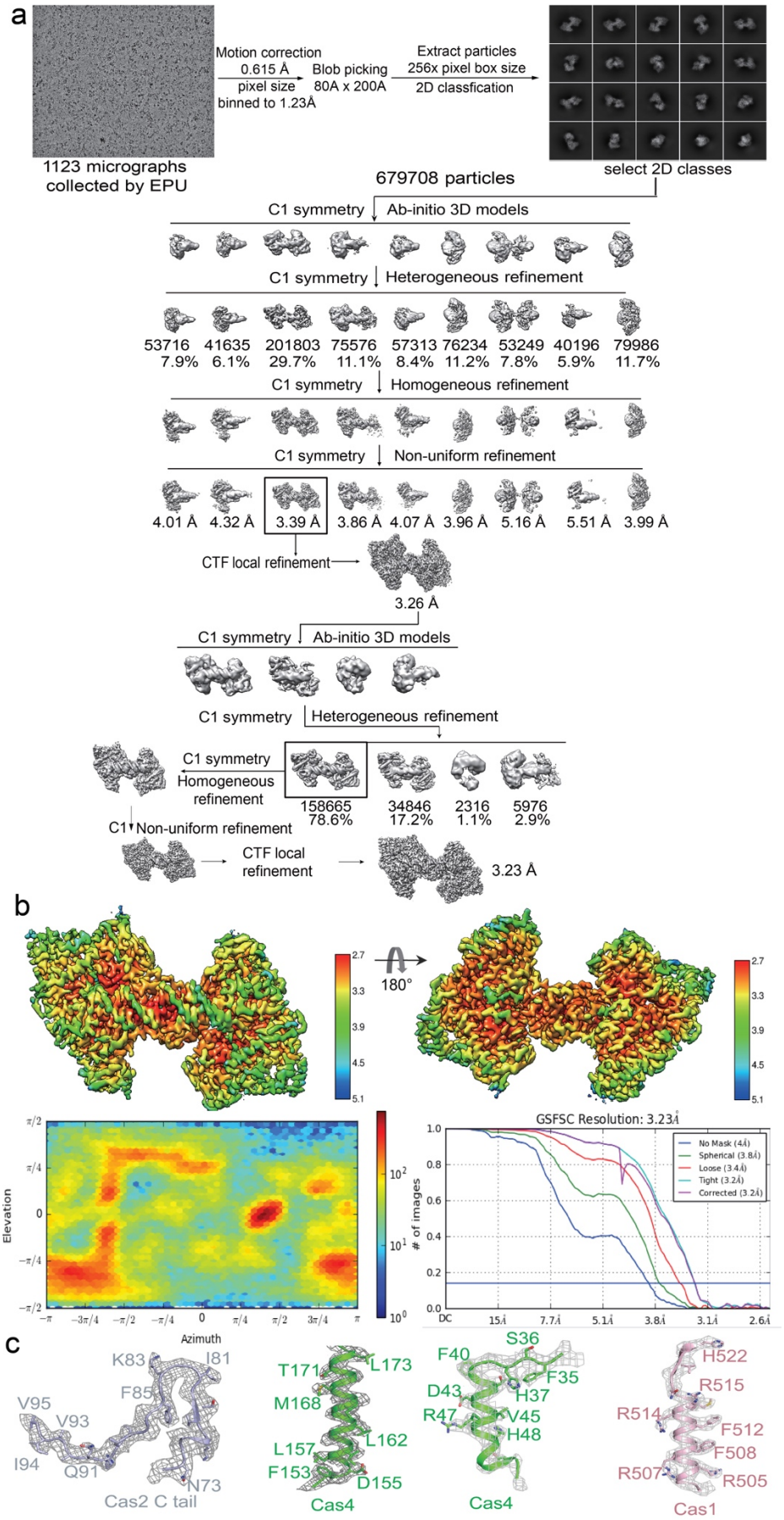
Extended Data Figure 1. Reconstitution and characterization of the *Gsu*Cas4/Cas1-Cas2 complex. a. Active site substitution in Cas4 nuclease center (H48G, D100A) reduced *in vivo* spacer acquisition efficiency dramatically. **b-d.** *Gsu*Cas4/1 purification analyzed by SDS-PAGE, coloring from the Fe-S cluster, and SEC profile. **e,f.** Affinity *Gsu*Cas2 purification analyzed by SDS-PAGE and SEC. **g.** GST pull-down results 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.** Definition of the Cas4 cleavage site in reference to the potassium permanganate foot-printing ladder.



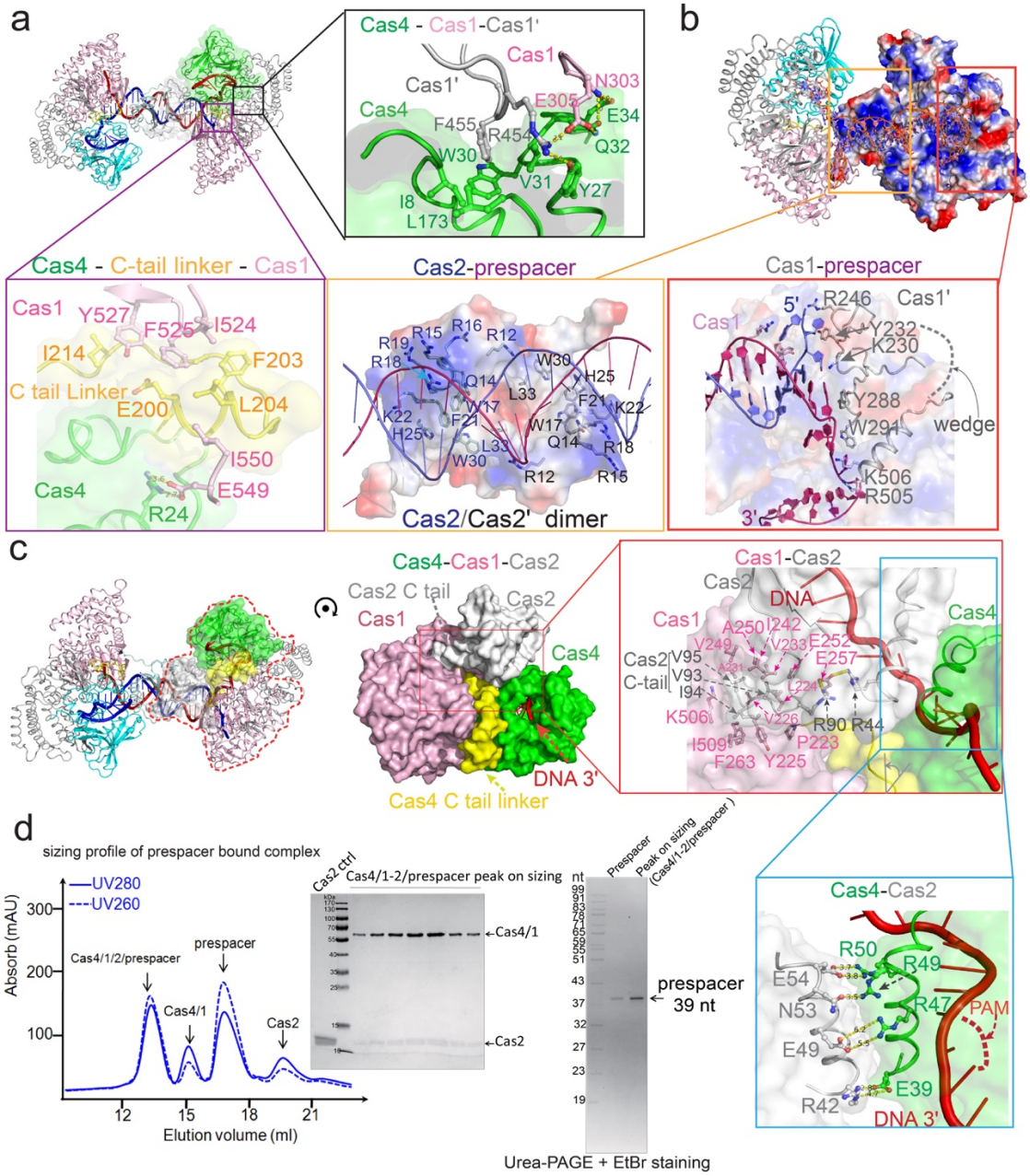
Extended Data Figure 2. PAM-dependent *Gsu*Cas4/Cas1-Cas2 complex formation revealed by SEC and electron microscopy. a. Diagram of the prespacer substrates used in complex formation. **b.** SEC profile of *Gsu*Cas4/Cas1-Cas2, alone or programmed with different prespacer substrates. PAM-containing prespacers drive high-order complex formation. **c.** Negative-staining electron micrograph of dual-PAM bound complex and 2D averages (bottom). **d-f.** Cryo- electron micrographs of three different complexes, with corresponding preliminary 2D averages to investigate sample quality.



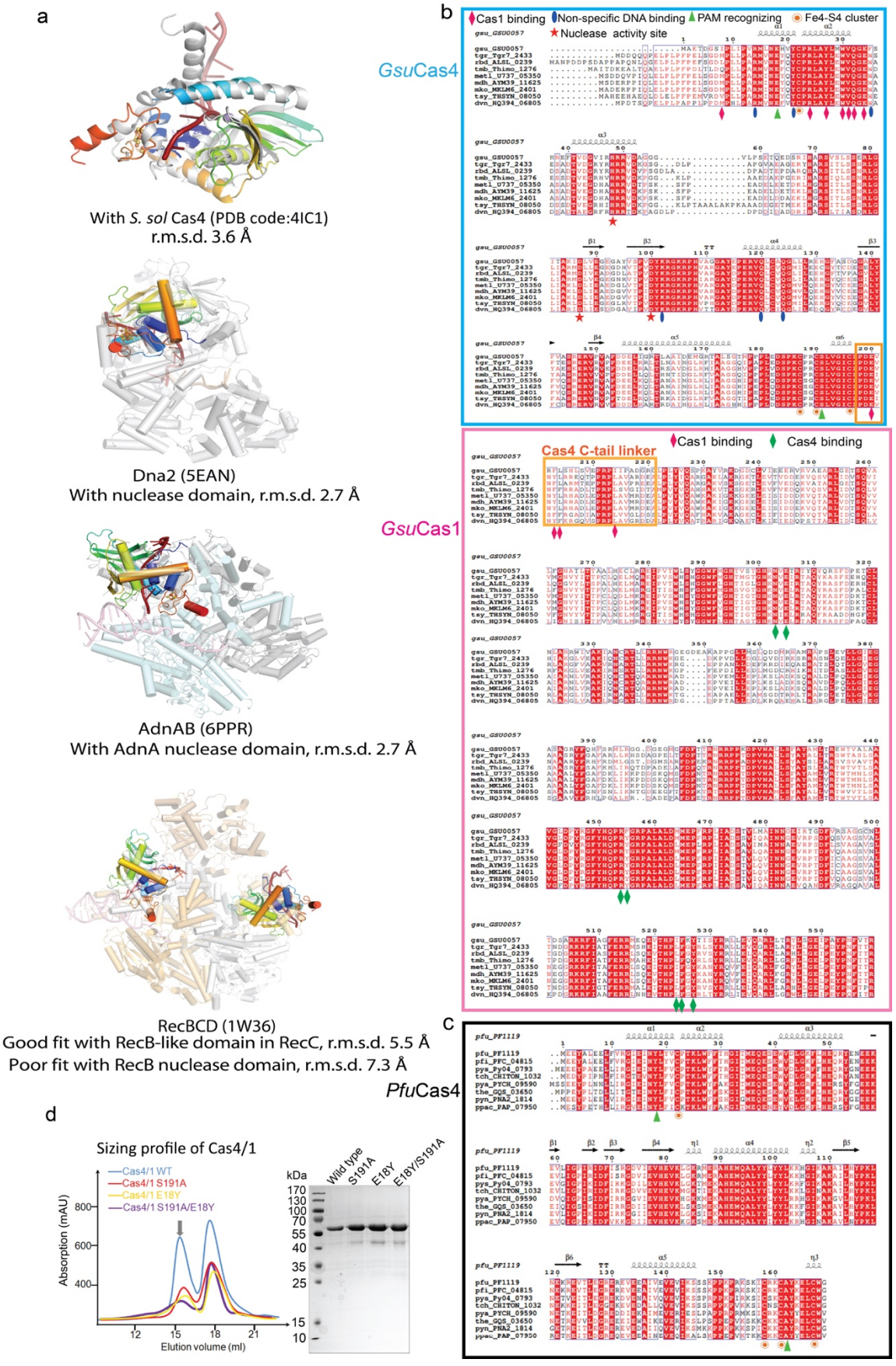
Extended Data Figure 3. Additional analysis of the dual-PAM prespacer bound *Gsu*Cas4/Cas1-Cas2 structure. **a.** Comparison between the current 3.2 Å cryo-EM reconstruction with the previous negative staining reconstruction of the *B. hal* Cas4/1-2 complex (EMDB 20131)[1]. **b-d.** Pairwise alignment between *Gsu*Cas4/Cas1-Cas2/prespacer and *Eco*Cas1-Cas2/prespacer (PDB 5DS4), *Efa*Cas1-Cas2/prespacer (PDB 5XVN), and *Efa*Cas1-Cas2/full-integration (PDB 5XVO), respectively. Alignments details are noted on the 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.



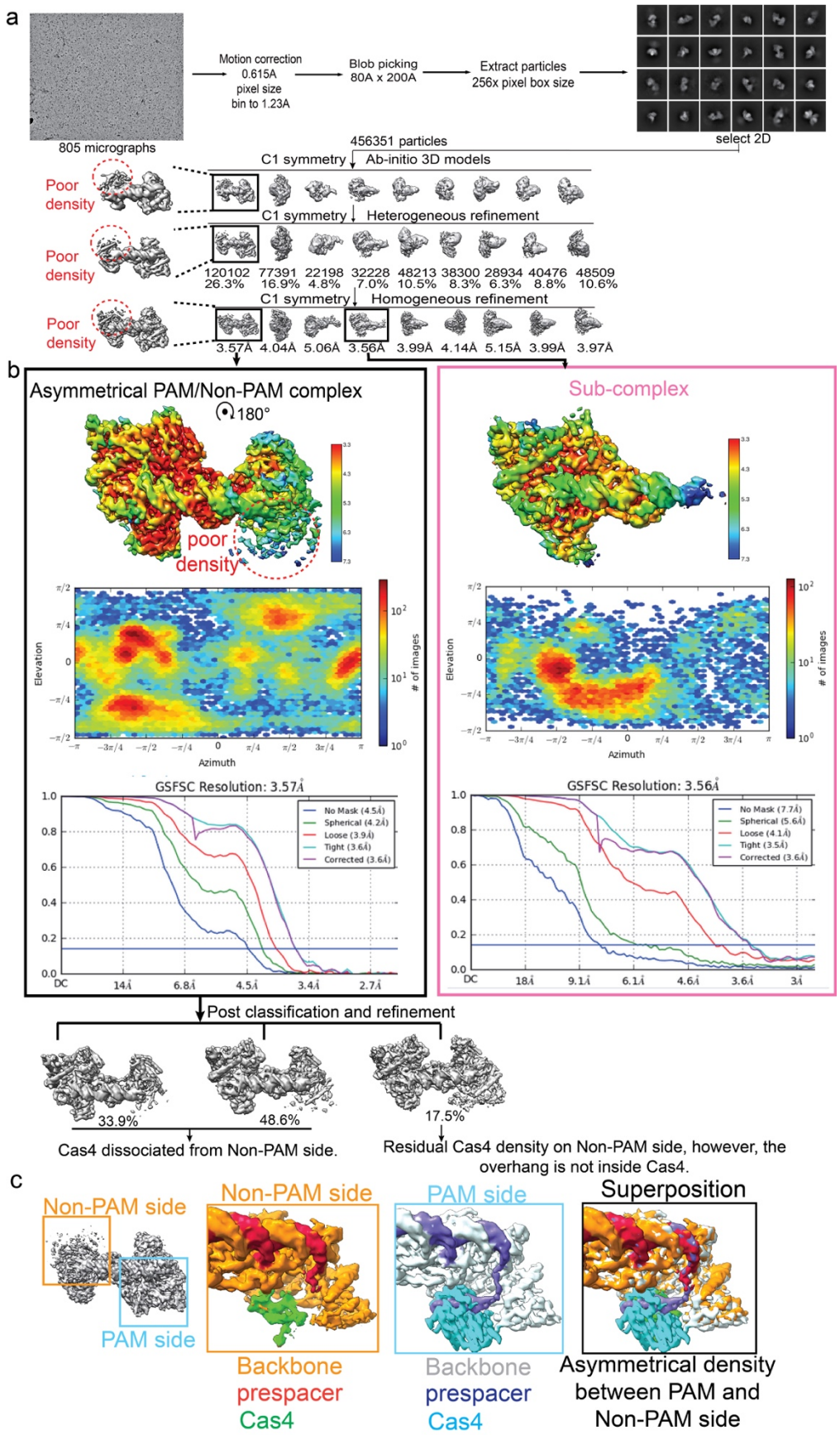
Extended Data Figure 4. Flow-chart of the cryo-EM single particle reconstruction of the dual-PAM prespacer bound GsuCas4/Cas1-Cas2. a. workflow of data processing for the dual-PAM prespacer bound Cas4/1-2 complex. **b.** Cryo-EM density of the dual-PAM prespacer bound Cas4/1-2 complex, colored according to local resolution (top). The viewing direction distribution plot (bottom left) and FSC curves (bottom right) for data processing. **c.** Representative EM densities for Cas2, Cas4, and Cas1, superimposed with their corresponding structural model.



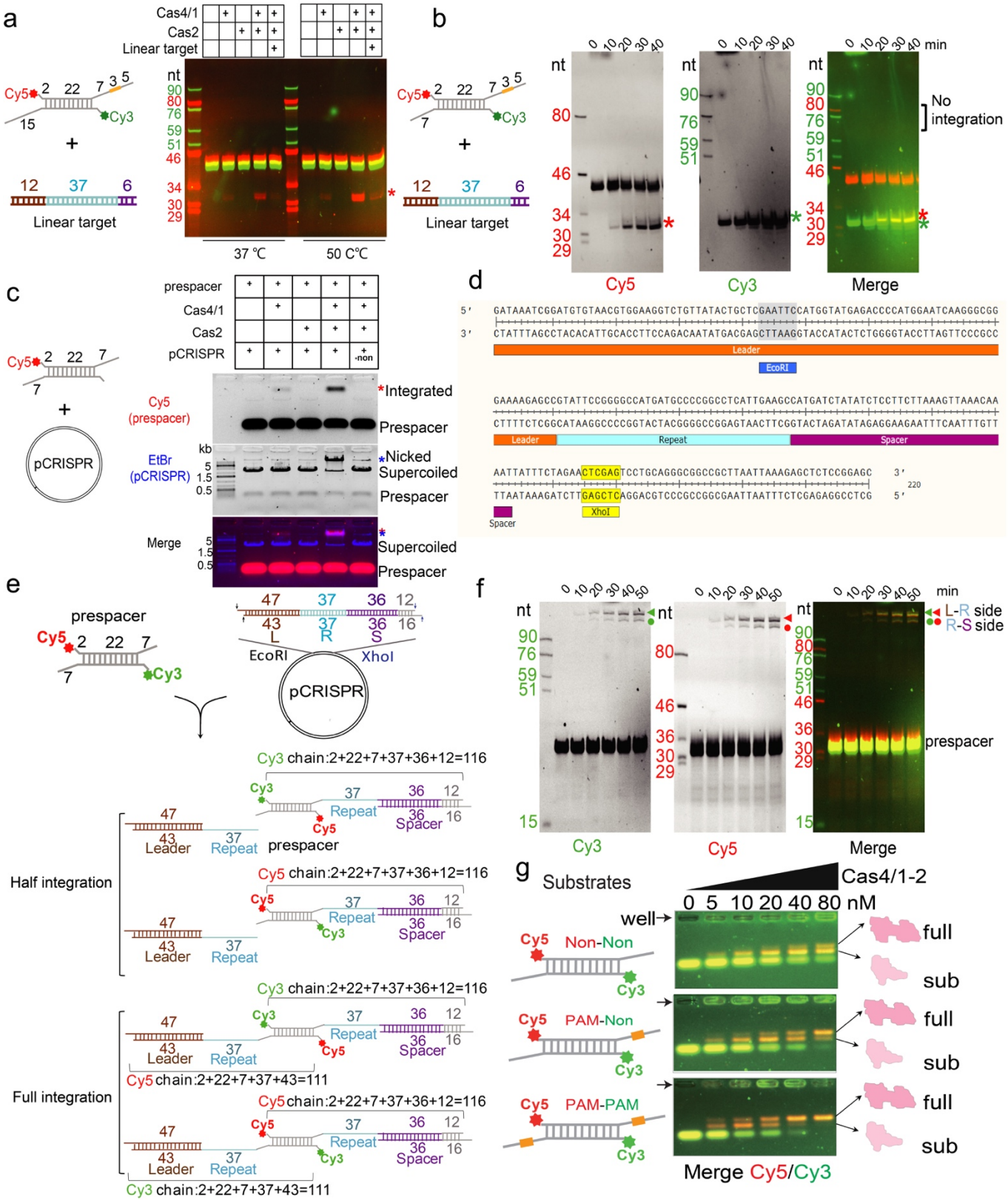
Extended Data Figure 5. In-depth interface analysis of the dual-PAM prespacer-bound *Gsu*Cas4/Cas1-Cas2 structure. a. Overall structure. Insets: zoom-ins of Cas4 interface with the neighboring Cas1s. **b.** Surface electrostatic potential. Left inset: Cas2 contacts to the mid-duplex; Right inset: Cas1 end-stacking to the mid-duplex. Residues responsible for guiding the 3'-overhang are also shown. **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. Right inset: favorable coiled coil interaction between Cas4 and Cas2. **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. **e.** *In vivo* spacer acquisition assay results for the wild type and PAM-specificity Cas4 mutants. Three biological replicates were analyzed by PCR and the band quantification revealed integration efficiency.



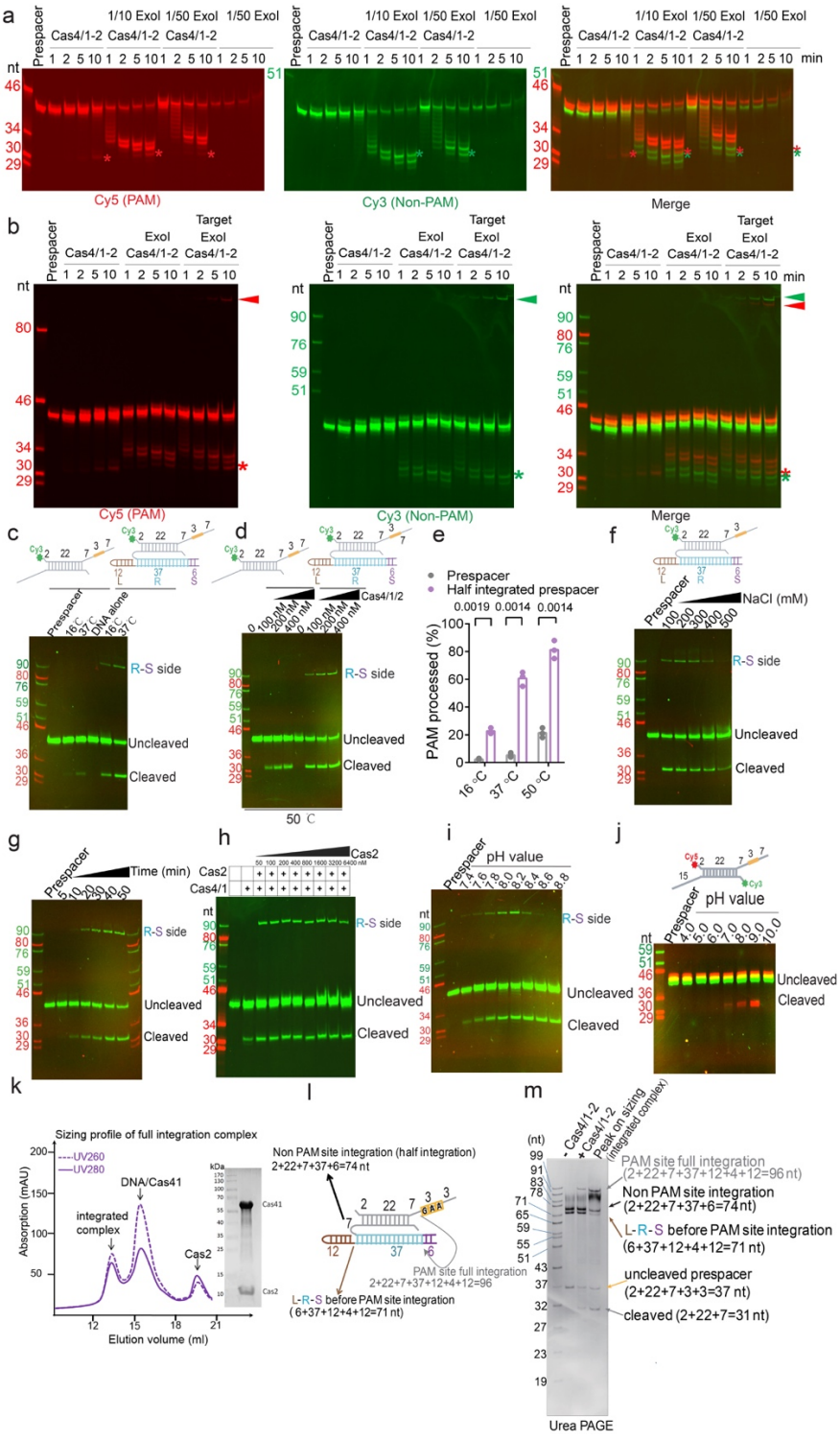
Extended Data Figure 6. In-depth analysis of the structure and sequence conservation in Cas4. a. Superposition of *Gsu*Cas4 with a standalone Cas4, and three different kinds of RecB-fold containing helicase-nuclease machines. The caging of the ssDNA substrate and the arrangement of the Fe-S cluster and the catalytic triad are conserved themes. **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 experiment in Fig. 3d.



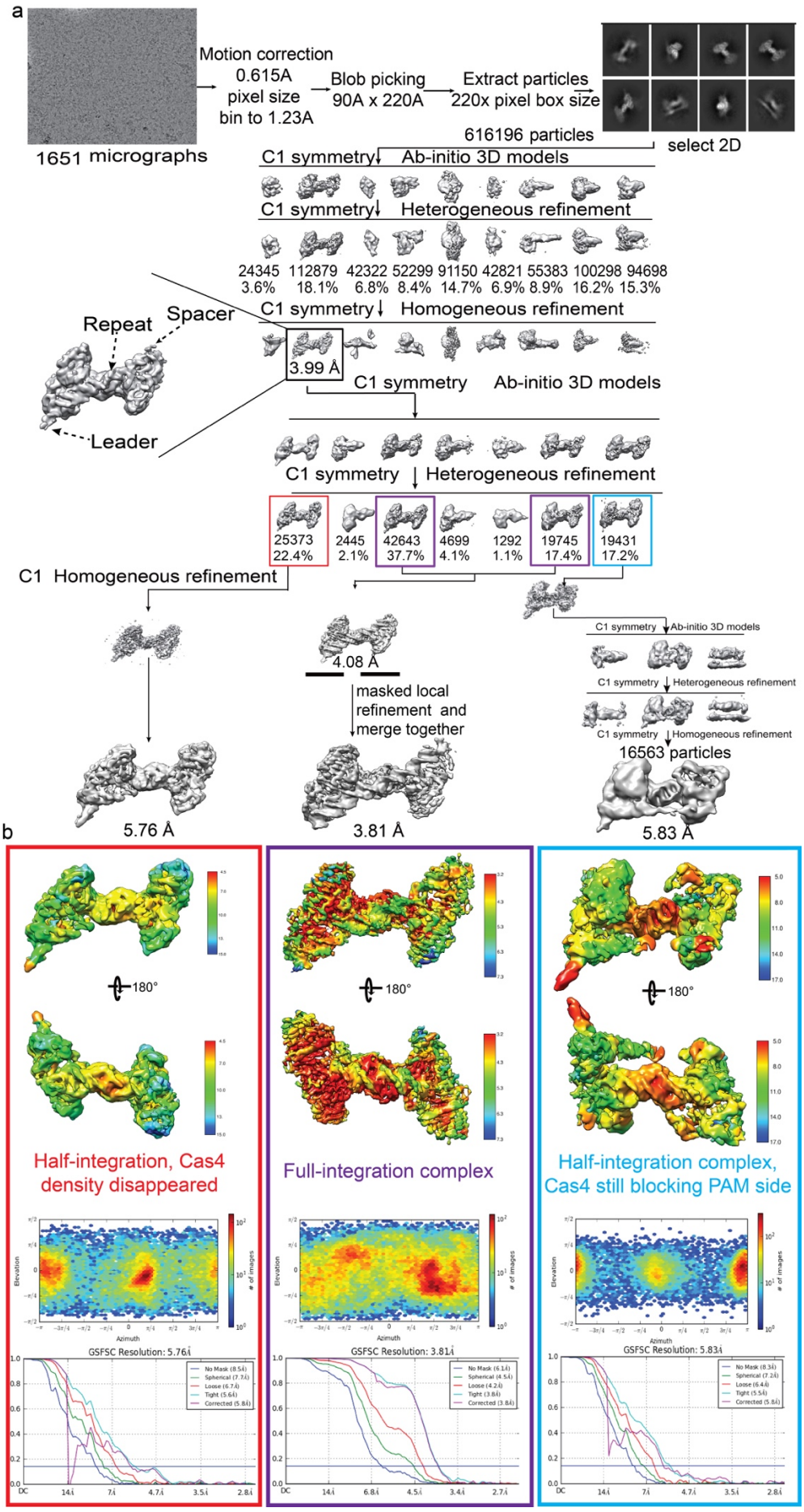
Extended Data Figure 7. Cryo-EM single particle reconstruction of the single-PAM prespacer bound *Gsu*Cas4/Cas1-Cas2. a. Flow-chart of the cryo-EM single particle reconstruction process that led to the reconstruction of two major snapshots. Left: Asymmetrical PAM/Non-PAM prespacer bound Cas4/1-2 complex. 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.



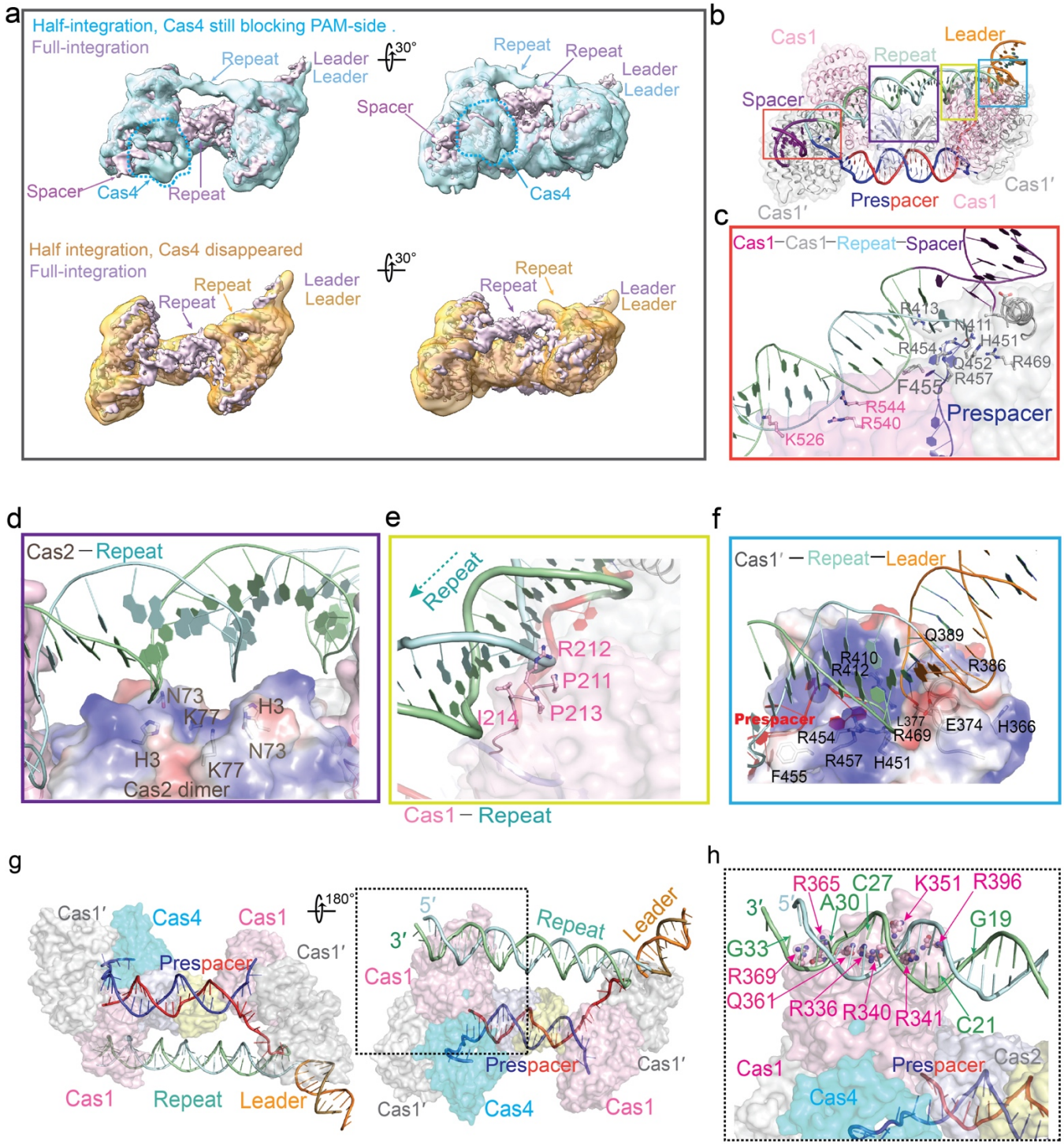
Extended Data Figure 8. ***In vitro*** **integration assay to distinguish integration directionality. a, b.** Biochemistry showing that Cas4/1-2 is unable to integrate prespacer into the linear form of leader-repeat DNA. **c.** Successful prespacer integration into a leader-repeat containing plasmid 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.** 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 afterwards. From the prespacer point of view, integration directionality is scrambled. Each integration band contains two fluorescent signals. **g.** Native PAGE showing that full Cas4/1-2 complex formation with prespacer takes place in a stepwise and PAM-dependent fashion.



Extended Data Figure 9. In-depth analysis of half-integration triggered PAM cleavage by Cas4. a. Time-course experiment showing ExoI trims PAM and non-PAM overhangs differently. **b.** Time-course experiment resolving the order of events from prespacer processing to full integration. Using the left and middle sets of experiments 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 a stimulation of Cas4-mediated cleavage of PAM-side overhang, and the full integration from PAM-overhang to spacer-side target quickly follows. **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.** Band quantification of results in **c.** revealing elevated PAM cleavage and full integration when leader-side integration already took place. **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 complex mimicking the half-integration complex that was used for cryo-EM analysis. **l, m.** Expected and observed ssDNA sizes due to PAM cleavage and full integration, respectively.



Extended Data Figure 10. Flow-chart of the 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 Figure 11. 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. **c.** Recognition of the leader-repeat junction by Cas1. The leader sequence is recognized at the DNA minor groove by the insertion of a Glycine-rich helix in Cas1. The repeat sequence immediately inside the integration site is recognized at the major groove by the hydrophobic and charged residues in a loop that contains the catalytic Histidine in Cas1. **d.** 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. **e.** The ridge of Cas2 further contacts the central dyad of the CRISPR repeat. **f.** A quasi-symmetric set of contacts are present at the spacer side for the fully integrated structure, albeit the contacts are less-well resolved due to elevated hinge motion, and the helix insertion and DNA bending at the flanking region does not take place. **g.** Orientational view 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. The residual density revealed that the leader-repeat DNA preferentially contact a positively charged patch in Cas1.