

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

Hu, Chunyi; Almendros, Cristóbal; Nam, Ki Hyun; Costa, Ana Rita; Vink, Jochem N.A.; Haagsma, Anna C.; Bagde, Saket R.; Brouns, Stan J.J.; Ke, Ailong

DOI

[10.1038/s41586-021-03951-z](https://doi.org/10.1038/s41586-021-03951-z)

Publication date

2021

Document Version

Accepted author manuscript

Published in

Nature

Citation (APA)

Hu, C., Almendros, C., Nam, K. H., Costa, A. R., Vink, J. N. A., Haagsma, A. C., Bagde, S. R., Brouns, S. J. J., & Ke, A. (2021). Mechanism for Cas4-assisted directional spacer acquisition in CRISPR–Cas. *Nature*, 598(7881), 515–520. <https://doi.org/10.1038/s41586-021-03951-z>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17

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

Chunyi Hu^{1,#}, Cristóbal Almendros^{2,3#}, Ki Hyun Nam⁴, Ana Rita Costa^{2,3}, Jochem N.A. Vink^{2,3}, Anna C. Haagsma^{2,3}, Saket Rahul Bagde¹, 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

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

43 **Main Text**

44 Prokaryotes have a unique ability to acquire immunological memories against mobile genetic
45 elements by integrating short fragments of DNA (*i.e.* spacers) in between the CRISPR repeats
46 ^{7,8}. The array of repeat-spacers serves as a transcription template to generate guide RNAs that
47 can direct CRISPR effector protein complexes to find, bind and cleave DNA or RNA targets. To
48 support protection by all DNA-targeting CRISPR-Cas systems, spacers need to be compatible
49 with a universal DNA-targeting requirement called the protospacer adjacent motif (PAM) ⁹⁻¹¹.
50 This short sequence motif directly flanking the target site helps crRNA-guided complexes
51 distinguish true targets from the actual spacer in the CRISPR array, and thereby prevents lethal
52 self-targeting. Furthermore, the presence of a PAM dramatically speeds up the target-searching
53 process by the crRNA-guided effector complexes, by reducing the total number of candidate
54 sites within the DNA ¹². To ensure CRISPR spacers are only derived from PAM-flanking
55 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-
56 Cas systems ¹³ further encode a dedicated CRISPR adaptation protein Cas4 ¹⁴ that works in
57 conjunction with the core spacer acquisition machinery consisting of Cas1 and Cas2 ^{2-4,15-21}. A
58 number of studies have contributed to our understanding of the role of Cas4. While early studies
59 mainly showed that deletion of the *cas4* gene impaired spacer acquisition in type I-B systems in
60 *Haloarcula hispanica* ²² and type I-A in *Sulfolobus islandicus* ²³, recent studies using type I-A in
61 *Pyrococcus furiosus* ²⁴, I-D in *Synechocystis sp.* ²⁵ and I-G (previously I-U) in *Geobacter*
62 *sulfurreducens* ²⁶ established a critical role for Cas4 in acquiring spacers with a functional PAM.
63 On the protein level Cas4 was found to harbour an Fe-S cluster and to catalyze various exo-
64 and endonuclease activities ²⁷⁻²⁹. Only recently did it become clear from work in I-C *Bacillus*
65 *halodurans* that Cas4 uses its nuclease activity to cleave PAM sequences in spacer precursors
66 just before integration into the CRISPR array ^{30,31}. Further studies with this Cas4 variant showed
67 that Cas4 forms a complex with a dimer of Cas1 and associates with Cas2 upon prespacer
68 binding ^{30,31}. The emerging picture is that Cas4 is somehow involved in PAM-selection and
69 processing, and that it must be important for the directional integration of spacers into the
70 CRISPR array. Yet, the molecular mechanism of this key process has remained elusive.

71

72 **Cas4 is a dedicated PAM-cleaving endonuclease**

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

77 pair (bp)-long spacers (the majority are between 35-37 bp) into the CRISPR locus in a PAM-
78 dependent manner (5'-TTN, 3'-AAN at the 3'-overhang)²⁶. The enzymatic activity of Cas4 was
79 shown to be required for PAM processing²⁶. To derive rules governing the prespacer
80 processing and integration, we electroporated prespacers of various sequence and structure
81 compositions into *E. coli* cells containing a *G. sul cas4/cas1-cas2/CRISPR* genomic locus and
82 analyzed cells for newly acquired spacers using PCR and deep sequencing methods (**Fig. 1b**,
83 **c**, **Extended Data Fig. 1a**). Based on prior structural and biochemical work, it was hypothesized
84 that *GsuCas4/Cas1-Cas2* may preferentially integrate prespacers containing a 26-bp mid-
85 duplex, with 5-nt 3'-overhangs on each side^{18,20,26,30}. Such prespacers were indeed robustly
86 integrated in a directional and single-stranded PAM (ss-PAM) dependent fashion (**Fig. 1b-c**).
87 Prespacers lacking a ss-PAM were not integrated (**Fig. 1b**). The context surrounding PAM also
88 influenced the integration outcome. Whereas a ss-PAM 5-nt away from the mid-duplex were
89 efficiently integrated, the same ss-PAM immediately adjacent to the mid-duplex, or a ds-PAM in
90 the middle of a duplex, did not enable spacer integration (**Fig. 1b**). Dual-PAM containing
91 prespacers were integrated with scrambled directionality but a precise length distribution,
92 whereas the single-PAM containing prespacers were integrated directionally but with a 2-3 nt
93 length distribution (**Fig. 1c**). It is possible that the 3'-overhang trimming is precise at the PAM-
94 side but slightly distributive at the non-PAM side. These data converge in suggesting that
95 *GsuCas4/Cas1-Cas2* preferentially recognizes prespacers containing a correctly spaced PAM in
96 the 3'-overhang of a DNA duplex.

97
98 Next, we switched to biochemical reconstitution to understand the molecular basis of Cas4-
99 assisted spacer integration. The PAM-containing 3'-overhang of the prespacer was found to be
100 specifically cleaved by the recombinant *GsuCas4/Cas1-Cas2* complex; the non-PAM 3'-
101 overhang remained intact (**Fig. 1d**, **Extended Data Fig. 1b-i**). Cleavage was Mn²⁺-dependent
102 and took place precisely after PAM (3'-A₃A₂G₋₁↓; **Extended Data Fig. 1h-i**). While precise,
103 PAM processing was rather inefficient. Only ~5% of the PAM-containing overhang was
104 processed after 60 minutes of incubation in 37 °C, in 50-fold excess of *GsuCas4/Cas1-Cas2*
105 (**Extended Data Fig. 1h**). The underlining mechanism for the attenuated PAM processing only
106 became clear after structural analysis. Interestingly, extended exposure to air induced
107 promiscuous DNA cleavage activity from this complex (**Fig. 1e**), likely due to the oxidation of the
108 Fe-S cluster in Cas4. The various level of oxidation may explain the spectrum of reported endo

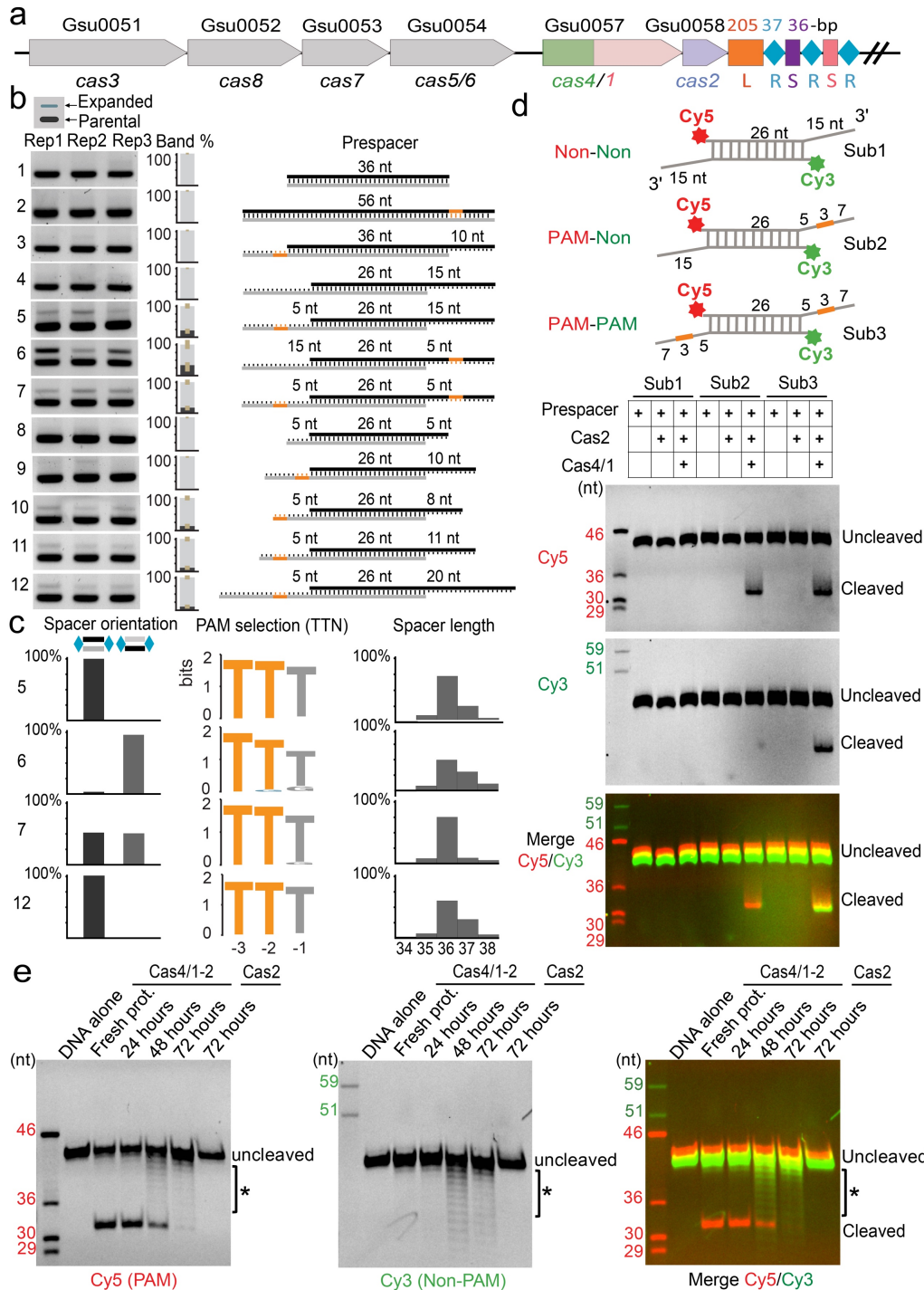


Figure 1. PAM-dependent prespacer processing and acquisition by *GsuCas4/Cas1-Cas2*. **a.** *cas* operon organization in Type I-G CRISPR in *G. sul*. Top: KEGG database identifier; Bottom: gene names; L, R, S, bp: Leader, repeat, spacer, base-pairs. **b.** *In vivo* acquisition of electroporated prespacers 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*.

110 and exonuclease activities for Cas4 in the literature²⁷⁻³¹.

111

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

113 Whereas physical interaction could be detected between *GsuCas4/Cas1* and *GsuCas2* in
114 affinity pull-down and size-exclusion chromatography (SEC) experiments, functional complex
115 formation was driven by the prespacer (**Extended Data Fig. 1g, 2**). A dual- or single-PAM
116 containing prespacer led to stable higher-order complex formation, as revealed by SEC and
117 electron microscopy (EM) analyses. In contrast, a PAM-less prespacer was not efficient at
118 organizing complex formation (**Extended Data Fig. 2**). EM analyses revealed the formation of
119 dumbbell-shaped particles characteristic of Cas1-Cas2 complexes. Because the dual-PAM
120 prespacer containing *GsuCas4/Cas1-Cas2* complex was especially homogeneous under
121 negative-staining and cryo-EM (**Extended Data Fig. 2c**), we first attempted to generate a high-
122 resolution reconstruction from this *quasi* symmetric PAM-recognition complex. The single
123 particle reconstruction reached 3.23 Å in resolution, which revealed significant more structural
124 details than the negative-staining EM reconstructions of related Cas4-Cas1-Cas2 complexes³⁰
125 (**Fig. 2; Extended Data Fig. 3a, 4**). The Cas1₄-Cas2₂ integrase core assumes its characteristic
126 dumbbell shape - the Cas2 dimer constitute the central handle, and two Cas1 dimers constitute
127 the two distal weights (**Fig. 2a**). In each dimer, only one Cas1 participates in spacer integration,
128 the other plays structural roles. The overall architecture and the detailed interactions leading to
129 *GsuCas1-Cas2* complex formation are more consistent with those found in the *Enterococcus*
130 *faecalis* rather than the *E. coli* complex^{18,20} (**Extended Data Fig. 3b-d**). For example, the C-
131 terminal tails of *GsuCas2* and *EfaCas2* stabilize the complex by mediating similar structural
132 contacts to the neighboring Cas1 and to the opposing Cas2 (**Extended Data Fig. 3c, 5**); the
133 contacts are mediated very differently in the *E. coli* complex. Surprisingly, Cas1-Cas2 was found
134 to specify a 22-bp mid-duplex rather than a 26-bp mid-duplex as defined by the integration
135 assay; an additional two base-pairs are unwound from each end, and the mid-duplex is end-
136 stacked by the N-terminal domain of the catalytic Cas1s on opposite ends (**Fig. 2a-b, 2e;**
137 **Extended Data Fig. 5b**). Indeed, re-designed prespacers containing a 22-bp mid-duplex
138 integrated as efficiently as the 26-bp version in the *in vivo* and *in vitro* assays (**Fig. 2d;**
139 **Extended Data Fig. 3e-f**). The 22-bp specification and the limited end-unwinding activity was
140 previously observed in *EfaCas1-Cas2* (**Fig. 2d**)^{19,20}. It is possible that Cas1-Cas2 has a
141 common preference for prespacers containing a 22-bp mid-duplex (occasionally 23-bp as in *E.*
142 *coli*), but has an idiosyncratic preference for 3'-overhang length (**Fig. 2e**).

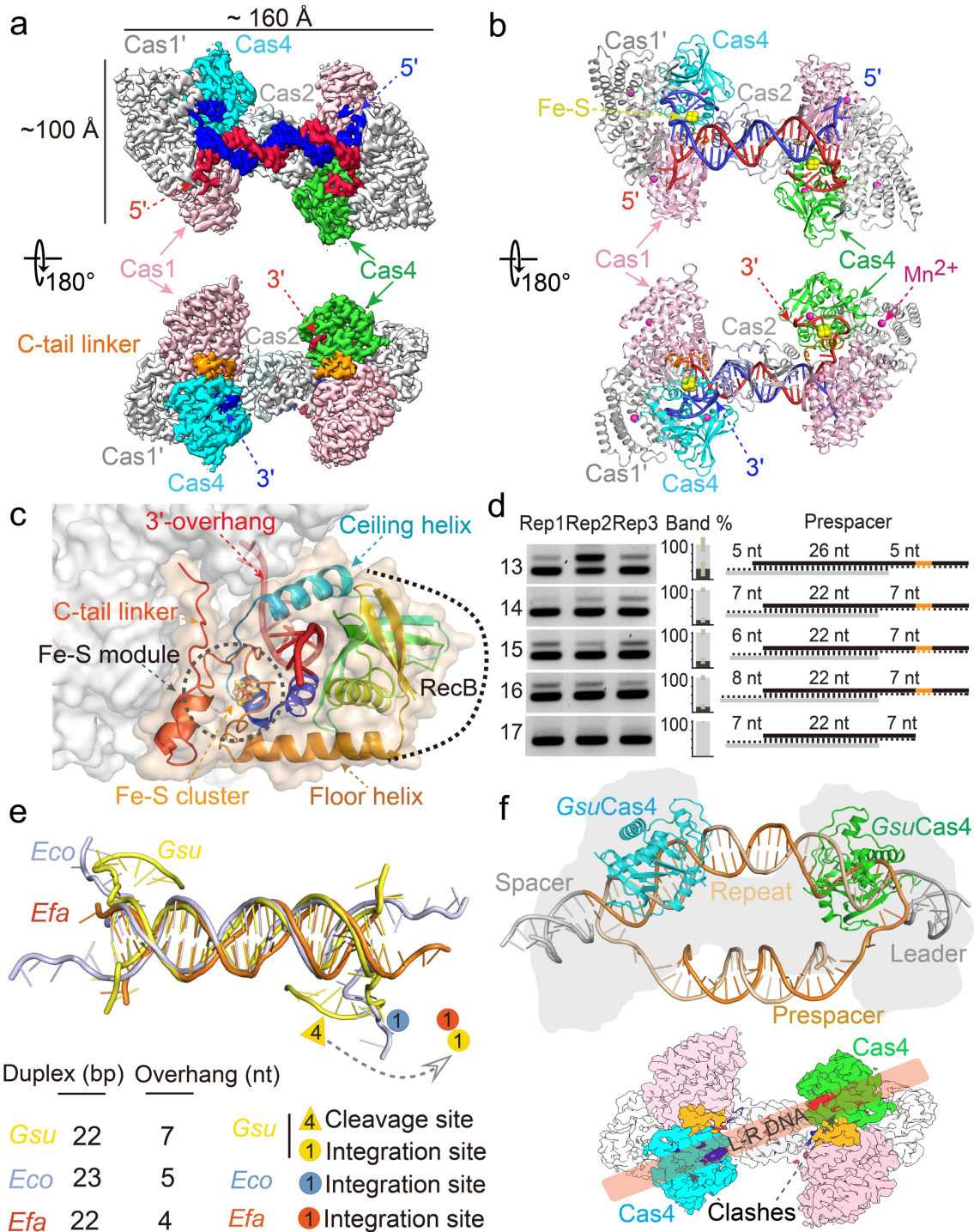


Figure 2. Insights from dual-PAM prespacer bound *GsuCas4/Cas1-Cas2* structure. **a, b.** Cryo-EM density and cartoon representation of the dual-PAM bound *GsuCas4/1-2* structure, respectively. **c.** Organization of Cas4 structural elements around the PAM-containing 3'-overhang. **d.** Validation that 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 *EfaCas1-Cas2* in the post-integration state. Note the PAM-recognizing Cas4 clashes with the repeat-spacer DNA entering into the integrase center in Cas1.

144 Among the four fused Cas4s, only the two non-catalytic Cas1-fused Cas4s are resolved in the
145 EM structure, due to their involvement in PAM recognition. The other two are missing from the
146 density presumably because they are not stably bound to the integrase core. Therefore, the
147 natural tethering between Cas4 and Cas1 in our system does not alter the dynamic nature of
148 the Cas4-Cas1-Cas2 interaction, and the mechanistic insights from this study are likely
149 applicable to all Cas4 systems. The EM density allows an unambiguous tracing of the entire
150 Cas4. Its structure aligns well with those of the stand-alone Cas4s^{27,28} and the nuclease
151 domains in helicase-nuclease fusion proteins AddAB³², AdnAB³³ and eukaryotic Dna2³⁴.
152 Interestingly, the Cas4 structure aligns poorly with the RecB nuclease in RecBCD; it agrees
153 better with the RecB-like fold in RecC instead (**Extended Data Fig. 6a-c**)³⁵. Cas4 organizes its
154 structural modules to form a narrow passage for the PAM-containing 3'-overhang. Its N-terminal
155 α -helical floor connects to the ceiling helix on the top, which reaches overhead to the RecB
156 nuclease center on the opposite side, which then weaves back through the floor helix, and the
157 remaining C-terminal region assembles with the N-terminal helical region to form the Fe-S
158 cluster module, a hallmark to all Cas4 nucleases (**Fig. 2c**). Cas4 connects to the non-catalytic
159 Cas1 through a 20-amino acid (aa) fusion linker, which mediates the dynamic docking and
160 dissociation of Cas4.

161

162 Importantly, the PAM-engaging Cas4s are wedged at the ventral side of the Cas1-Cas2
163 complex (**Fig. 2a-b**). Because this region of Cas1-Cas2 is responsible for recruiting the leader-
164 repeat DNA for spacer integration^{18,20}, it follows that the PAM-recognizing Cas4 sterically
165 blocks integration from the PAM-side Cas1 (**Fig. 2e-f**). Cas4 contacts both subunits of Cas1
166 through an extensive interface, many residues at the interface are conserved (**Extended Data**
167 **Fig. 5a-c, 6b**). The Cas4-Cas2 interface involves favorable polar contacts between the ceiling
168 helix in Cas4 (aa 39-50) and an outer helix in Cas2 (aa 42-53). It is difficult to identify key
169 interface residues that are universally conserved across all Cas4 branches. There may exist
170 evolutionary pressure to maintain idiosyncratic Cas4 and Cas1-Cas2 interactions in order to
171 avoid crosstalk among coexisting CRISPR systems. If true, this scheme would be analogous to
172 the highly selective binding relationship between Cas3 and Cascade³⁶.

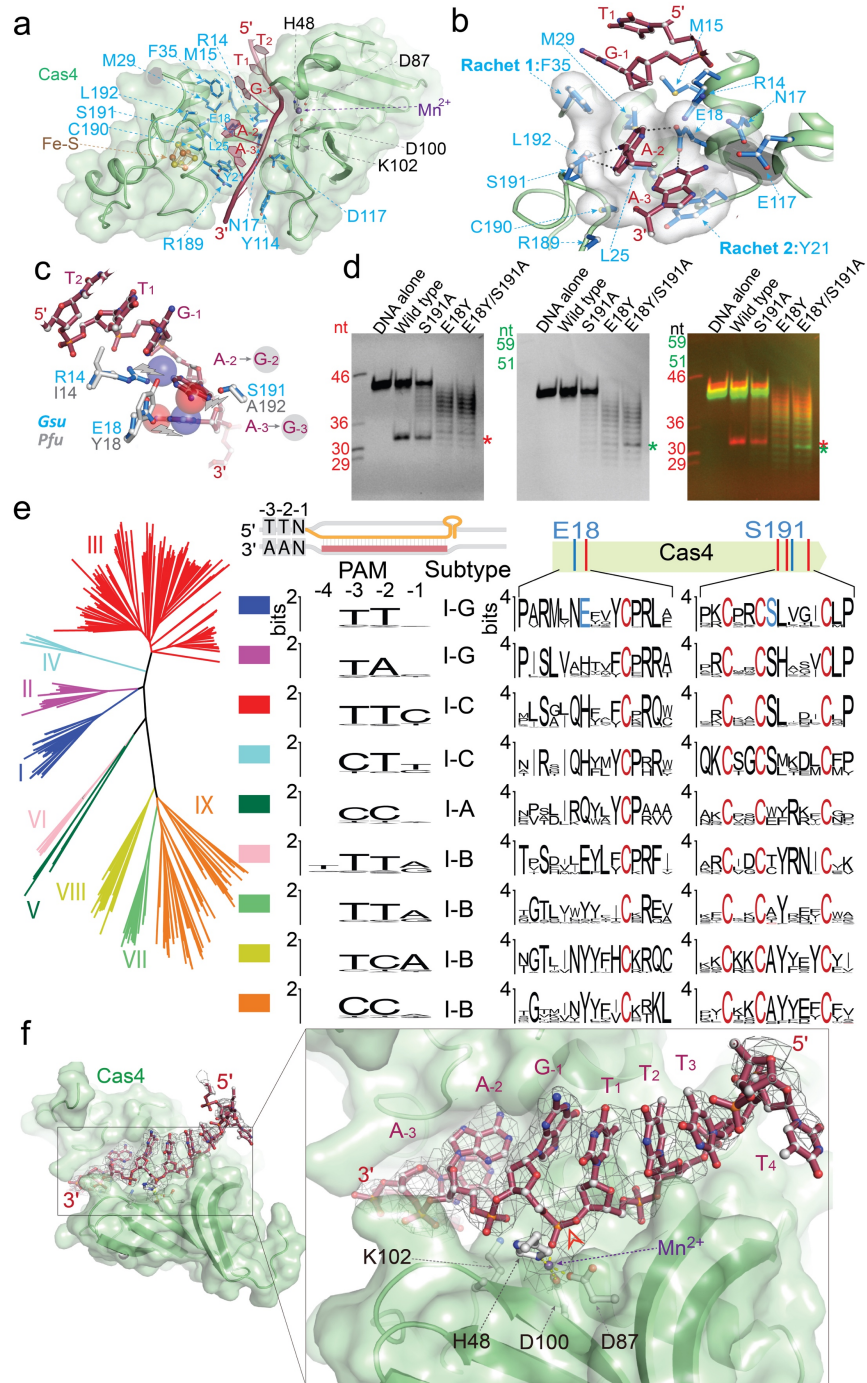


Figure 3. Cas4-mediated PAM-recognition and delayed overhang cleavage. **a.** PAM is caged inside a molecular ratchet in Cas4. Ceiling helix is omitted for better illustration of the narrow pathway for 3'-overhang. **b.** The di-adenosine PAM is surrounded by Van der Waals interactions that probe for 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.

174 **Structural basis for Cas4-mediated PAM recognition**

175 Despite extensive studies, the PAM recognition and cleavage mechanisms inside Cas4-Cas1-
176 Cas2 remain unresolved. This EM structure brings such mechanisms into focus. The substrate-
177 binding groove in Cas4 aligns with that in Cas1 to form a continuous 3'-overhang-binding
178 groove. The 11-nt 3'-overhang (5'-dA₇C₆T₅T₄T₃T₂T₁**G**₋₁**A**₋₂**A**₋₃T₋₄) travels deep inside, protected
179 from random nuclease cleavage. Stemming out of the mid-duplex, the first four nucleotides
180 travel more or less along the same path towards the Cas1 active site, as seen in the previous
181 Cas1-Cas2/prespacer structures^{15,18,20}. However, nucleotides 5-11 detour through Cas4. They
182 first travel on top of the RecB nuclease module, then enter into the narrow passage described
183 previously (**Fig. 3a**). Two hydrophobic residues F35 and Y21 interdigitates into the ssDNA
184 before and after the narrow passage, forming molecular ratchets that cage the di-
185 deoxyadenosine PAM (3'-A₋₃A₋₂) inside (**Fig. 3b**). They likely enforce a ratcheting motion to
186 slowly thread the 3'-overhang through, which allows the PAM sequence to be recognized and
187 captured. Inside the narrow passage, the edges of A₋₂ and A₋₃ are surrounded by hydrophobic
188 and long side chain residues (R14, M29, L25, L192, E117, N17, C190) that probe for shape
189 complementarity. Deoxyguanosines would not fit comfortably in the same cage because their
190 exocyclic N2 amines would cause steric clash; whereas the smaller-sized pyrimidines may slip
191 through without a chance to establish favorable contacts. Two Cas4 residues establish polar
192 contacts with PAM: E18 makes bidentate hydrogen-bonding interactions with A₋₂ and A₋₃, and
193 S191 forms a hydrogen bond with A₋₂ (**Fig. 3b**). They likely contribute significantly to the PAM
194 specificity. Consistent with the *in vivo* data²⁶, there is no sequence-specific recognition to the
195 first residue of PAM, G₋₁. This nucleotide is excluded from the PAM-recognition box and points
196 to the solvent (**Fig. 3b**).

197

198 Because Cas4 is responsible for PAM selection in a large fraction of CRISPR systems, we
199 attempted to rationalize the PAM code in other CRISPR systems. We first carried out a
200 structure-guided mutagenesis to explore the possibility of switching the PAM specificity of
201 *GsuCas4* to that of *Pyrococcus furiosus* Cas4 (**Fig. 3c**). *PfuCas4* share 17% sequence identity
202 with *GsuCas4* and specifies a 5'-CCN PAM (3'-GGN in the overhang). We substituted the two
203 sequence-specific PAM contacting residues in *GsuCas4* to their counterparts in *PfuCas4*. In
204 single substitutions, S191A retained *Gsu*-PAM specificity; cleavage activity was slightly
205 compromised. E18Y lost sequence specific cleavage activity on both PAMs, and cleaved
206 ssDNA distributively. Interestingly, the combination of these two substitutions resulted in a
207 cleavage preference for *Pfu*-PAM, even though the activity was quite distributive. These results

208 suggest E18 plays a more important role than S191 in PAM recognition (**Fig. 3c**). However, this
209 partial success in switching PAM specificity did not further extend into *in vivo* spacer acquisition
210 assays, which put further demand on prespacer/Cas4/Cas1-Cas2 stability and PAM cleavage
211 timing. While E18Y/S191A Cas4 showed compromised *Gsu*-PAM (TTN) prespacer integration,
212 it was able to support integration of *Pfu*-PAM (CCN) containing prespacers *in vivo* (**Extended**
213 **Data Fig. 5e**). These results suggest that while the hydrogen-bonding interactions are
214 important, a significant portion of the PAM specificity is likely conferred by the peripheral
215 residues mediating hydrophobic interactions.

216

217 Next, we attempted to use bioinformatics to establish a correlation between structural features
218 in Cas4 and PAM sequence variations. We first determined which PAM is used by different
219 Cas4-containing CRISPR systems by mapping spacers in annotated and metagenomic
220 databases. This led to a phylogenetic tree based on the alignment of Cas4s for which we could
221 reliably couple PAM code with clades of Cas4s, sometimes from different CRISPR types that
222 were using the same PAM ³⁷ (**Fig. 3d**). We expected that residues crucial for PAM selection
223 would be conserved within the clades, but would differ between groups selecting a different
224 PAM (**Fig. 3e**). The E18 residue that is in contact with A₋₂ and A₋₃ is one such discriminant
225 amino acid residue because it is highly conserved among Type I-G Cas4s specifying TTN PAMs
226 and among Type I-B Cas4s specifying a TTA or TTG PAM. S191, which contacts A₋₂, does not
227 appear to be a discriminant residue as it was also found in Type I-G Cas4s specifying TAN
228 PAMs. However, the highly conserved neighboring residue, L192, was exclusively found in
229 Cas4 groups specifying a T on the -2 position of the PAM, including the less closely related
230 Cas4s in Type I-C that either specify TTC or CTT. Therefore, the presence of L192 in Cas4 is a
231 good predictor of a T on PAM-2. Similarly, informatics identified R14 and L25 as good predictors
232 of T₋₂. The reverse argument is not necessarily true. For example, not all PAMs containing a T₋₂
233 predict L192 in the corresponding Cas4s. The structure reveals that PAM is specified at least
234 partially by hydrophobic contacts that select for shape complementarity (**Fig. 3b**). In such cases
235 a cluster of hydrophobic residues in Cas4 may be required to specify a PAM code, and their
236 identity may not be unique.

237

238 **PAM recognition delays 3'-overhang cleavage and prevents integration therein**

239 The most important mechanistic insight from the dual-PAM structure is the observation that the
240 PAM-containing 3'-overhang is recognized, sequestered, but not cleaved by Cas4 (**Fig. 3f**). The
241 labile phosphate of G₋₁ is correctly positioned into the active site, which consists a DEK motif

242 (D87, D100, K102) and a histidine residue (H48), all of which are highly conserved among Cas4
243 and RecB family of nucleases³⁸. These residues coordinate a catalytic metal ion, presumably
244 Mn²⁺, which is shown by the EM density to be tightly coordinated to the scissile phosphate. In
245 the AdnAB structure, such active site configuration was shown to promote efficient DNA
246 cleavage³³. However, here the EM density clearly argues for an intact DNA substrate at the
247 active site (**Fig. 3f**), which was subsequently confirmed by denaturing PAGE (**Extended Data**
248 **Fig. 5d**). The exact cleavage inhibition mechanism in Cas4 will require a more focused analysis
249 in the future. Among the many mechanistic possibilities, we speculate that it might be caused by
250 the sub-optimally placed K102 residue in the DEK motif, which has been implicated as essential
251 for Cas4 catalysis²⁶. Rather than pointing towards the labile phosphate, K102 is twisted away
252 by the residing β -strand. A minor conformational change in Cas4 may allow K102 to participate
253 in PAM cleavage. Without PAM cleavage, Cas4 is locked in place and integration is blocked
254 from taking place at the PAM side. This structural observation is in perfect agreement with the
255 spacer directionality requirement in Type I CRISPRs.

256

257 **Structure-guided reconstitution of directional integration**

258 Next, to investigate the status of the non-PAM 3'-overhang, we determined the cryo-EM
259 structure of the *Gsu*Cas4/Cas1-Cas2 complex programmed with a single-PAM containing
260 prespacer. This led to an asymmetric full complex structure at 3.57 Å resolution, and a 3.56 Å
261 assemble intermediate that will be discussed later (**Fig. 4; Extended Data Fig. 7**). Whereas the
262 PAM-side of *Gsu*Cas4/Cas1-Cas2 is blocked by a PAM-recognizing Cas4, 82.5% of the single-
263 PAM particles do not have a docked Cas4 at the non-PAM side (**Fig. 4a**); 17.5% contain a
264 docked Cas4 evidenced by weak densities, however, the non-PAM overhang is not captured
265 inside (**Extended Data Fig. 7c**). In both cases, the non-PAM side Cas4/1 dimer density is
266 weaker than the PAM-side counterpart, and a hinge motion is evident, anchored at the non-
267 catalytic Cas1. Only the first four nucleotides of the non-PAM 3'-overhang can be traced in the
268 density, along a similar path as in the PAM-side (**Extended Data Fig. 7c**). Because the non-
269 PAM overhang lacks Cas4 protection, we reasoned that it may be trimmed to the optimal
270 overhang length by certain host nucleases, then captured by the nearby Cas1 and preferentially
271 integrated to the leader-repeat DNA. This host nuclease-assisted integration mechanism would
272 lead to a fixed spacer directionality that is consistent with the CRISPR biology. We directly
273 tested this mechanistic model. Indeed, *E. coli* SbcB (ExoI) protein could trim the non-PAM 3'-
274 overhang to the preferred length of ~7-nt, (**Fig. 4b**). Even the distributive cleavage pattern was
275 categorically consistent with the spacer length distribution in the *G. sul* CRISPR systems (**Fig.**

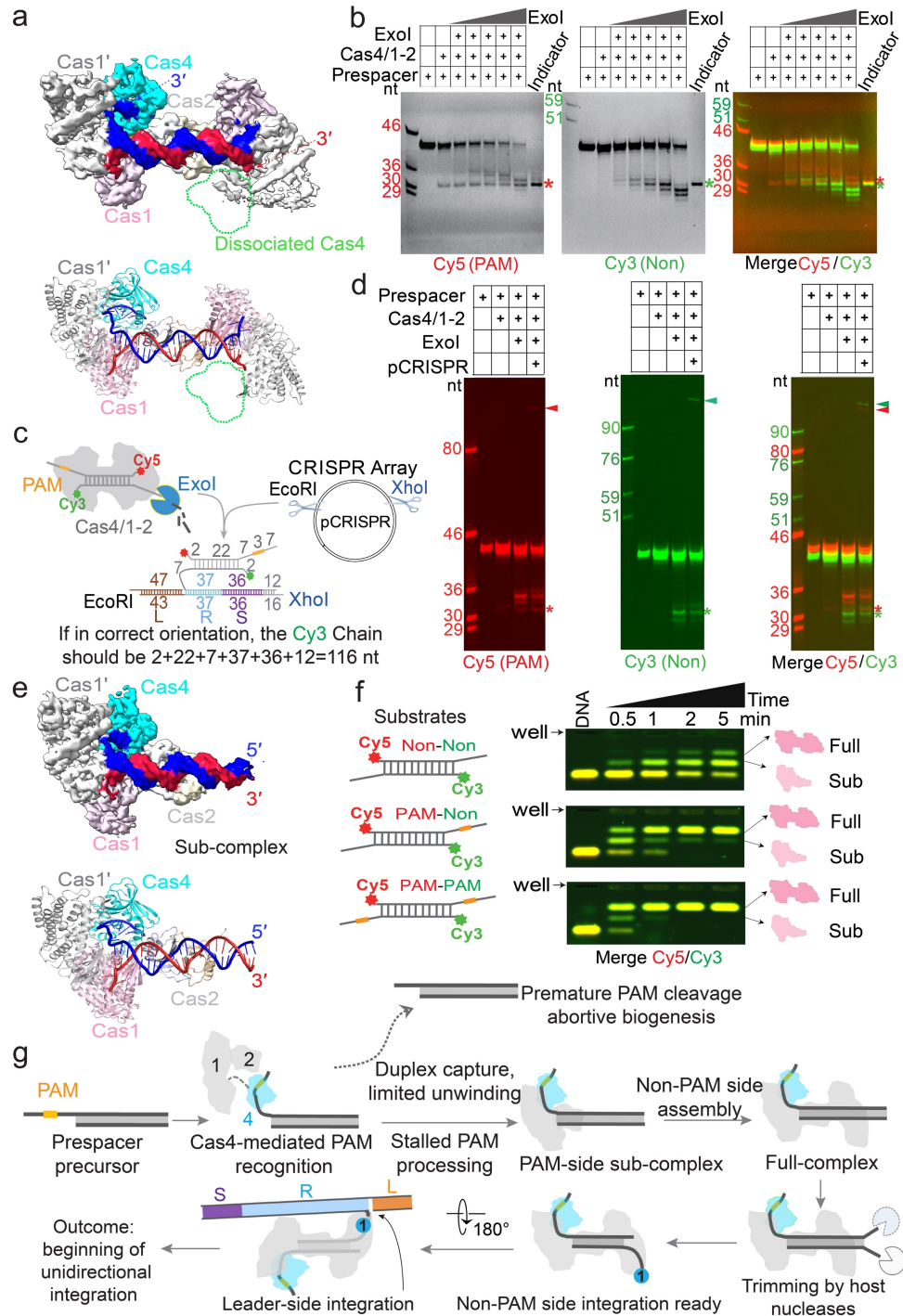


Figure 4. Mechanistic insights from the single-PAM prespacer bound GsuCas4/Cas1-Cas2 structure. **a.** Cryo-EM density (top) and structure (bottom) of the single-PAM prespacer bound GsuCas4/1-2 complex. Lack of Cas4 at the non-PAM side is highlighted. **b.** *E. coli* nuclease Exol is capable of trimming 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 Exol 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.

277 **1c)**²⁶. In contrast, the PAM-side 3'-overhang was protected by the footprint of Cas4 in the same
278 reaction (**Fig. 4b-c**). Next, we established an *in vitro* integration assay to test whether the Exol-
279 trimmed prespacer can be integrated unidirectionally. An obstacle to this effort is that although
280 *GsuCas4/Cas1-Cas2* readily integrated prespacers with optimal overhang length into a
281 negatively supercoiled leader-repeat containing plasmid, it failed to do so on a linear target
282 (**Extended Data Figs. 8a-d**). This behavior is similar to that of *E. coli* Cas1-Cas2, which was
283 later shown to rely on the host integration factor (IHF) to integrate into a linear target³⁹. Given
284 the limitation, in order to resolve the integration directionality, we first integrated a dual-
285 fluorescently labeled prespacer into the plasmid, then restriction-digested out the leader-repeat
286 region to determine the directionality based on the product size on denaturing polyacrylamide
287 gel (**Extended Data Figs. 8c-f**). In control experiments, we verified *GsuCas4/Cas1-Cas2*'s
288 preference to integrate first into the leader-proximal side and confirmed the ability of the setup to
289 distinguish integration directionality (**Extended Data Figs. 8e-f**). We went on to demonstrate
290 that Exol-trimming enabled the non-PAM side of the prespacer to specifically integrate into the
291 leader-proximal side of the repeat (**Fig. 4c-d**). This pattern is in agreement with the observed
292 spacer directionality in the *G. sul* CRISPR array.

293

294 **Intermediate structure generates insight about prespacer biogenesis**

295 The PAM/non-PAM cryo-EM reconstruction further captured an important functional state, which
296 corresponds to an intermediate assembly during prespacer biogenesis. The structure is of
297 sufficient resolution to reveal that a $(\text{Cas4/Cas1})_2\text{-Cas2}_2$ sub-complex has captured the PAM-
298 side overhang and the duplexed region of the prespacer (**Fig 4e; Extended Data Fig. 7**). While
299 the PAM-side arrangement is essentially the same as in the previous structures, $(\text{Cas4/Cas1})_2$
300 densities were absent from the non-PAM side. Using time-course and concentration-titration
301 based electrophoretic mobility shift assays (EMSA), we confirmed that the *GsuCas4/Cas1-Cas2*
302 integrase indeed assembled in a stepwise fashion, and the PAM-containing overhang strongly
303 promoted the assembly of the full-complex (**Fig 4f; Extended Data Fig. 5g**). Collectively, these
304 structural snapshots provide the much-needed temporal resolution for prespacer biogenesis.
305 We conclude that the $(\text{Cas4/Cas1})_2\text{-Cas2}_2$ sub-complex is capable of scouting for precursor
306 DNA with a PAM-containing 3'-overhang. Binding of such precursor triggers enzymatic stalling
307 in Cas4 and recruits a second $(\text{Cas4/Cas1})_2$ complex to the opposite side, leading to the
308 formation of an integration-competent $(\text{Cas4/Cas1})_4\text{-Cas2}_2$ full complex. The conditional
309 assembly process provides a quality-control mechanism to only recruit PAM-containing spacer
310 precursors for further processing and integration (**Fig. 4g; Supplementary Movie S1**). The

311 length of the precursor duplex is likely longer than the preferred length by Cas1₄-Cas2₂. In a
312 previous study we explored this scenario and found that the host nucleases are capable of
313 trimming dsDNA and ssDNA to the preferred prespacer specification as defined by the Cas1₄-
314 Cas2₂ footprint¹⁹.

315

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

317 Having established that Cas4 defines the spacer directionality by blocking the PAM-side
318 integrase center before integration, we next probed into the mechanism that relieves this
319 blockage after half-integration, since the PAM-side prespacer needs to be processed and
320 integrated to the opposite side of the CRISPR repeat to complete full integration. What serves
321 as the molecular switch? We hypothesized that the half-integration itself may stimulate PAM
322 cleavage and Cas4 dissociation. To test this, we programmed *Gsu*Cas4/Cas1-Cas2 to the half-
323 integration state using an annealed prespacer and leader-repeat DNA that mimics the half-
324 integration product¹⁸, and monitored the extent of PAM processing and half-to-full integration
325 transition at different conditions and over time (**Extended Data Fig. 9a-j**). Indeed, half-
326 integration led to faster and higher extent of PAM cleavage, and full integration quickly followed
327 (**Fig. 5a; Extended Data Fig. 9b**). As controls, PAM cleavage was much slower and weaker
328 when the leader-repeat DNA was absent (**Fig. 5a**), or when the half-integration did not take
329 place (**Extended Data Fig. 8a**).

330

331 Next, we sought to provide the structural basis for the observed mechanistic coupling. The
332 reacted sample in **Extended Data Fig. 9k-m** was snap-frozen for cryo-EM analysis (**Extended**
333 **Data Figs. 9k-m**). We were able to capture multiple conformational states from the single
334 particle reconstruction, which we interpret as representing three different functional states
335 during the half-to-full integration transition. The more populated state was solved at higher
336 resolution since more particles were available for 3D reconstruction, and vice versa (**Extended**
337 **Data Figs. 10**). The three states differ significantly in their spacer-side contacts and in Cas4 and
338 integration status. In what we interpret as an early state (5.83 Å in resolution), density clearly
339 reveals that Cas4 still blocks the PAM-side integration site and the PAM-containing 3'-overhang
340 is still sequestered in Cas4. Unable to dock into the integration site, the CRISPR repeat reaches
341 over from the leader-side Cas1 directly to the spacer-side counterpart, without contacting the
342 Cas2 dimer in the middle. The spacer-side CRISPR repeat contacts a positively-charged region
343 on Cas1, near Cas4 (**Fig. 5b-c; Extended Data Fig. 11**). The DNA density is weak, suggesting
344 that it may dynamically sample multiple conformations, some of these motions may involve

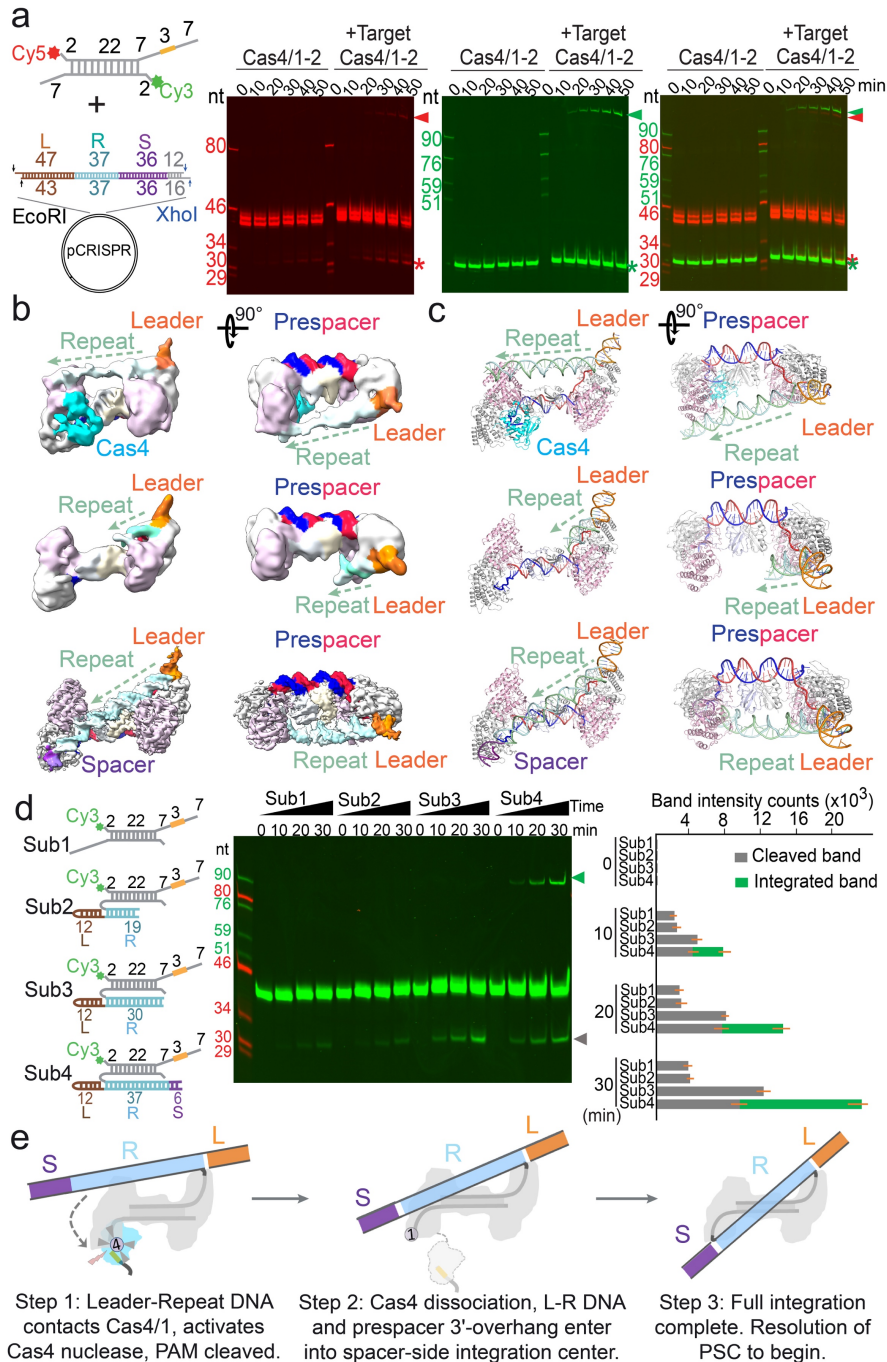


Figure 5. Snapshots of *GsuCas4/Cas1-Cas2* in coupling half-integration with PAM cleavage to achieve full-integration. **a.** Time-course experiments showing non-PAM side half-integration stimulates PAM cleavage. Full integration quickly follows. **b.** Three cryo-EM snapshots and **c.** corresponding 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.

346 Cas4 contacts. In the 5.76 Å intermediate state, the Cas4 density disappears, and the density
347 corresponding to the cleaved prespacer overhang appears to point towards the exposed Cas1
348 active site, although it is quite weak and choppy. With Cas4 out of the way, the CRISPR repeat
349 DNA projected from the leader-side Cas1 contacts the Cas2 dimer in the middle and appears to
350 further point towards the spacer-side integration center, however, its density is too degraded for
351 model building (**Fig. 5b-c; Extended Data Fig. 11**). This suggests that even with Cas4 out of
352 the way, spacer-side CRISPR DNA capture and integration is inefficient, presumably because
353 the favorable leader-sequence contacts are missing here ¹⁹. Lastly, we captured a 3.81 Å
354 snapshot of the full-integration state. EM densities clearly reveals that the CRISPR repeat DNA
355 has been docked into the spacer-side integration center, and a continuous density connects it
356 with the 3'-overhang, suggesting that full-integration has taken place (**Fig. 5b-c**). This snapshot
357 is architecturally similar to the previously determined post-integration complexes from *E. fae* ²⁰,
358 however, the leader-repeat DNA in the *G. sul* structure is not as sharply kinked at the Cas2
359 binding site as seen in the *E. fae* structure. The entire leader- repeat DNA is contacted in a
360 quasi-symmetric fashion at the following four regions (**Fig. 5b-c; Extended Data Figs. 10-11**).
361 The 4-bp leader region immediately upstream of the CRISPR repeat is favorably recognized
362 and significantly bent upwards by the DNA minor groove insertion of a glycine-rich α-helix in
363 Cas1. As previously revealed, this recognition leads to strong leader-proximal preference at the
364 first half-integration reaction ¹⁸⁻²⁰. Lacking such sequence at the spacer-side, DNA density is
365 degenerate and DNA bending is not observed. The α-helix insertion most likely does not take
366 place at the spacer side. The inverted repeats at the border region of the CRISPR repeat are
367 recognized at the major groove region by the catalytic Histidine-containing loop in Cas1 ²⁰. The
368 following minor groove is recognized by a conserved “PRPI” motif in the Cas4-Cas1 fusion
369 linker, which is not exposed when Cas4 is docked. Lastly, the backbone of the central dyad of
370 CRISPR repeat is contacted by the positive charges and a proline-rich motif on the ridge of the
371 Cas2 dimer (**Extended Data Fig. 11b-f**). Connecting the dots together, the three snapshots
372 define the order of molecular events and support a strong mechanistic coupling between the
373 leader-half integration and the Cas4-mediated PAM processing, which ensures PAM-specific
374 spacer-side integration.

375

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

379 physical contacts by the integrated leader-repeat DNA somehow activates Cas4. The allosteric
380 activation model was deemed unlikely because no significant conformational change in Cas1-
381 Cas2 was observed among apo, half- and full-integration structures, although we cannot
382 completely rule out the possibility that changes in the extent of hinge motions may play a role.
383 To further probe whether the physical contact by the leader-repeat DNA might activate Cas4,
384 we systematically shortened the leader-repeat DNA in the previous integration assay setup (**Fig.**
385 **5d**). Results revealed a strong correlation. When the leader-repeat was too short to reach
386 spacer-side Cas4/1 (Sub2: 19-bp CRISPR repeat), the extent of PAM cleavage was
387 indistinguishable from that in the prespacer-only control. When the leader-repeat is long enough
388 to reach the spacer-side Cas4/1 (Sub3: 30-bp CRISPR repeat), the PAM cleavage was
389 significantly enhanced, even without the spacer-side integration (**Fig. 5d**). We therefore
390 conclude that contacts by the half-integrated DNA efficiently stimulates the PAM cleavage
391 activity of Cas4. PAM cleavage leads to Cas4 dissociation, which exposes the spacer-side
392 integrase center and allows full integration (**Fig. 5e; Supplementary Movie S2**). It should be
393 noted that we are not able to define which specific DNA contact activates Cas4. This will require
394 even higher temporal and spatial resolutions to resolve.

395

396 **Discussion**

397 In summary, we provide a comprehensive set of mechanism to explain the PAM-dependent
398 spacer acquisition process in Cas4-containing CRISPR systems. Our study firmly establishes
399 that Cas4 is a dedicated PAM-cleaving endonuclease, whose activity is tightly regulated. In the
400 context of the Cas1-Cas2 integrase complex, Cas4 specifically recognizes but refrains from
401 cleaving the PAM-containing 3'-overhang in a prespacer. This unexpected molecular
402 constipation is the cornerstone for productive prespacer biogenesis and functional spacer
403 integration in Type I and V CRISPR systems. We provide direct and high-resolution evidence
404 that PAM recognition and the subsequent molecular constipation takes place early during
405 prespacer biogenesis, in essence it serves as a gatekeeper to channel only the productive
406 precursor into the biogenesis pathway. We further show that host nucleases can assist the
407 further processing of these precursors, and this eventually leads to a directional integration to
408 the leader-side CRISPR repeat. Moreover, we reveal that the leader-side integration efficiently
409 activates the PAM cleavage activity of Cas4 and causes Cas4 dissociation, which in turn
410 derepresses the PAM-side Cas1 integrase and allows the half-to-full integration transition.
411 Collectively, the series of structural snapshots depicts the entire directional integration process
412 for the Cas4-containing Type I and V CRISPR systems. Exactly how spacer directionality is

413 established in Cas4-less CRISPR systems requires further investigation^{15,16,40}. In Type I-E
414 CRISPR, such mechanism has been shown to involve Cas1-mediated PAM sequestration and
415 integration-dependent desequestration²¹. Therefore, the PAM-dependent blockage/activation of
416 the two integration centers in Cas1-Cas2 may be a universal theme to achieve directionally
417 spacer integration.

418
419 The structural similarity of Cas4 to the nuclease domains of AddAB/AdnAB and a structural
420 domain in the equivalent location in RecBCD shed light into the ancient function of Cas4 in
421 spacer acquisition. These helicase-nuclease machines not only play essential roles in
422 homology-directed repair, but also provide a line of innate immunity for bacteria by preferentially
423 degrading linear DNA lacking chi sites, which are more likely of foreign origin. Functional
424 interactions between RecBCD/AddAB and Cas1-Cas2 mediated spacer acquisition have been
425 noted in previous studies^{41,42}. Certain traits in the AdnA nuclease (and its structural equivalent
426 in RecBCD) may have made them particularly desirable by Cas1-Cas2. For example, the subtle
427 sequence preference and occasional enzymatic pausing may have been exploited by Cas1-
428 Cas2 to establish PAM-dependent directional integration. This dramatically increased the
429 productive spacer acquisition in the ancient CRISPR systems. It is possible that the ancient
430 Cas1-Cas2 relied on RecBCD or AddAB for spacer precursors so heavily, that it started to
431 establish a physical interaction with the nuclease domain to facilitate the process. It eventually
432 led to the hijacking of this host nuclease domain into the *cas* operon as *cas4*. A similar process
433 may have taken place for other nucleases such as *dnaQ*^{21,43,44}.

434

435 **Methods**

436 **PAM prediction**

437 221,089 unique spacers along with genome source, *cas* gene information, and repeat sequence
438 were obtained from CRISPRCasDb⁴⁵ in February 2020. These spacers were blasted against
439 our own sequence database containing all sequences from the NCBI nucleotide database^{46,47},
440 environmental nucleotide database⁴⁸, PHASTER⁴⁹, Mgnify⁵⁰, IMG/M⁵¹, IMG/Vr⁵², HuVirDb⁵³,
441 HMP database⁵⁴, and data from Pasolli *et al.*⁵⁵. All databases were accessed in February 2020.

442

443 Hits between spacers and sequences from the aforementioned nucleotide databases were
444 obtained using the BLASTN program⁵⁶ version 2.10.0, which was run with parameters
445 word_size 10, gap open 10, penalty 1 and an e-value cutoff of 1. Hits inside CRISPR arrays
446 were detected and filtered out by aligning the repeat sequence of the spacer to the flanking

447 regions of the spacer hit (23 nucleotides on both sides). To minimize the number of false
448 positive hits, we further filtered hits based on the fraction of spacer nucleotides that hit the target
449 sequence. In a first step, only hits with this fraction higher than 90% were kept. To find targets
450 for even more spacers while keeping the number of false positives low, we included a second
451 step where hits with a matching percentage higher than 80% were kept if another spacer from
452 the same phylogenetic genus hit the same sequence in the stringent first round. Finally, we
453 removed spacers that were shorter than 27 nucleotides.

454

455 Highly similar repeat sequences of the same length were clustered using CD-HIT⁵⁷ with a 90%
456 identity threshold. To increase the number of aligned sequences for PAM determination, we
457 hypothesized that similar repeat sequences would be used in the same orientation and would
458 correspond to the same PAM sequences, as coevolution of PAM, repeat and Cas1 sequences
459 has been shown previously^{58,59}. The PAM for each aligned repeat cluster was then determined
460 by aligning the flanking regions of the spacer hits in each cluster. To equally weigh each spacer
461 within the repeat cluster, irrespective of the number of blast hits, consensus flanks were
462 obtained per spacer. These consensus flanks contained the most frequent nucleotide per
463 position of the flanking regions. From the alignment of consensus flanks (for clusters with at
464 least 10 unique spacer hits) the nucleotide conservation in each flank was calculated.
465 Conserved nucleotides were considered part of the PAM in case nucleotide conservation was
466 higher than 0.5 bit score, and the bit score in that position was at least 5 times higher than the
467 median bit score of the two 23-nt flanks. This PAM database was manually curated to fix PAMs
468 determined incompletely when nucleotides that were slightly below the threshold did occur in
469 other repeat clusters of the same subtype. The orientation of the PAM was set to match the
470 overall orientations of experimentally determined PAMs in literature for different systems
471 (upstream of 5'-end of the protospacer in Type I systems and downstream of 3' of the
472 protospacer in Type II systems).

473

474 **Cas4 phylogenomics**

475 Cas4 sequences were retrieved from each Cas4-containing genome in the PAM database.
476 Cas4 sequences were discarded in case multiple Cas4 sequences of that subtype (subtypes
477 defined by CRISPRCasdb) were present in a single genome, or when Cas4 belonged to a
478 different subtype than the predicted subtype of the repeat cluster. The tree was generated with
479 PhyML⁶⁰ from a MAFFT alignment of all Cas4 sequences⁶¹. The sequence logos were
480 generated with Berkeley weblogo⁶² and were performed on each group of Cas4 sequences with

481 a similar PAM, where redundant sequences were removed by CD-hit (threshold 0.9). For groups
482 with a small amount of nonredundant sequences (I-G TTN, I-G TAN and I-C CTT), additional
483 Cas4 sequences were retrieved by BLAST search of repeat sequences of predetermined PAM
484 repeat clusters and retrieving adjacent Cas4 sequences in the NCBI nucleotide database.

485

486 **Bacterial strains and growth conditions**

487 *Escherichia coli* strains Dh5 α and BL21-AI were grown at 37 °C in Lysogenic Broth (LB) media
488 with shaking or on LB agar (LBA) plates containing 1.5% (w/v) agar. When required, media was
489 supplemented with 50 μ g/ml spectinomycin, 100 μ g/ml ampicillin, 50 μ g/ml Kanamycin, 1 mM
490 IPTG, and 0.2% (w/v) L-arabinose (see **Supplementary Table 1** for plasmids and their
491 corresponding selection markers).

492

493 **Plasmid construction**

494 Plasmids used in this work are listed in **Supplementary Table 1**. All cloning steps were
495 performed in *E. coli* Dh5 α . The type IG CRISPR-Cas acquisition module from *G. sulfurreducens*
496 DSMZ 12127 was amplified by PCR using the Q5 High-Fidelity Polymerase (New England
497 Biolabs) and primers BN462 and BN1196 (**Supplementary Table 2**). The amplicon was cloned
498 into the p13S-S ligation-independent (LIC) cloning vector
499 (<http://qb3.berkeley.edu/macrolab/addgene-plasmids/>) by TA cloning, generating plasmid
500 pCas4/1-2. For plasmid pCRISPR, a synthetic construct composed of T7 terminator, a CRISPR
501 array (leader-repeat-spacer1-repeat), the mCherry gene, and flanking 20-bp homology regions
502 to the vector, was introduced into pET cloning vector 2A-T amplified with primers BN1247 and
503 BN1650 by Gibson assembly. E18Y mutant of Cas41 (pCas4/1-2-E18Y) was generated by
504 mutagenesis using pCas4/1-2 as a template with primers BN3392 and BN3393. Double mutant
505 E18Y/S191A (pCas4/1-2-E18Y/S191A) was generated by mutagenesis using pCas4/1-2-E18Y
506 as a template with primers BN3394 and BN3395. All plasmids were verified by Sanger
507 sequencing (Macrogen Europe, Netherlands). Bacterial transformations were carried out by
508 electroporation (200 Ω , 25 μ F, 2.5 kV) using an ECM 630 electroporator (BTX Harvard
509 Apparatus), and transformants were selected on LBA supplemented with the appropriate
510 antibiotics.

511

512 **Spacer acquisition assay**

513 *Escherichia coli* BL21-AI was co-transformed with pCas4/1-2, pCas4/1-2-E18Y, or pCas4/1-2-
514 E18Y/S191A and pCRISPR. Colonies were grown in 5 ml of LB supplemented with

515 spectinomycin and ampicillin at 37 °C with shaking. After 2.5h of growth, the expression of *cas*
516 genes was induced with IPTG and L-arabinose, and the cultures were incubated for additional
517 2h. Cells were made electrocompetent and transformed with 5 µl of each 50 µM prespacer
518 prepared by mixing primers (**Supplementary Table 2**) at 1:1 from the 100 µM stock. Cells were
519 recovered in LB for 1h at 37 °C, 180 rpm, and then grown overnight in 10 ml of LB
520 supplemented with spectinomycin and ampicillin at 37 °C with shaking. Plasmids were extracted
521 from the overnight cultures (Thermo Scientific GeneJet Plasmid Extraction Kit) and digested
522 with EcoRI and NcoI to avoid amplification of larger products from the plasmid backbone.
523 Digested plasmids were used to detect spacer acquisition by PCR using OneTaq 2x MasterMix
524 (New England Biolabs) and a mix of three degenerate primers with different 3' nucleotides
525 (BN464, BN465, and BN1314) and primer BN1708²⁵. Samples were run on 2% agarose gels
526 and visualization for spacer acquisition using SYBR Safe. Unexpanded and expanded band
527 percentages were determined using the Analysis Tool Box of ImageLab software using
528 unmodified images. The expanded CRISPR DNA band was purified by automated size selection
529 and submitted to a second round of PCR using the degenerate primers and the internal reverse
530 primer BN1754^{25,63}.

531

532 **Expanded CRISPR array sequencing**

533 PCR amplicons of the expanded CRISPR arrays were purified using the GeneJET PCR
534 Purification kit (Thermo Fisher Scientific) and the DNA concentration was measured using Qubit
535 Fluorometric Quantification (Invitrogen). Samples were prepared for sequencing using the NEB
536 Next Ultra II DNA Library Prep Kit for Illumina and each library was individually barcoded with
537 the NEBNext Multiplex Oligos for Illumina (Index Primers Set1 and Set2). Sample size and
538 concentration were then assessed using the Agilent 2200 TapeStation D100 high sensitivity kit,
539 and samples were pooled with equal molarity. Pooled samples were denatured and diluted as
540 recommended by Illumina and spiked with 15% of PhiX174 control DNA (Illumina). Sequencing
541 was performed on a Nano flowcell (2 × 250 base paired-end) with an Illumina MiSeq. Image
542 analysis, base calling, de-multiplexing, and data quality assessments were performed on the
543 MiSeq instrument. Resulting FASTQ files were analyzed by pairing and merging the reads using
544 Geneious 9.0.5. Acquired spacers were extracted and analyzed as described previously²⁵.

545

546 **Cloning, expression and purification**

547 Full-length *GsuCas4/1* (*Gsu0057* in KEGG) gene was cloned from *Geobacter*
548 *sulfurreducens* genomic DNA into pET28a -His₆-Twin-Strep-SUMO vectors (Kan^R) or pGEX-41-

549 T-His₆-Flag-GST (Amp^R), between BamHI and XhoI sites. Sequence-verified plasmids were
550 transformed into *E. coli* BL21 (DE3) star cells under the appropriate antibiotic selection. A 6
551 liters cell culture was grown in LB medium at 37 °C until an optical density of 0.5 at 600 nm. The
552 culture temperature was then reduced to 16 °C and incubated for additional 2 hours. Expression
553 was induced with 0.5 mM isopropyl-β-D-thiogalactopyranoside (IPTG), 0.2 mg/mL ferrous
554 sulfate (Fisher) and 0.4 mg/mL L-cysteine (MP biomedical) at 16 °C overnight. Cells were
555 harvested by centrifugation and resuspended in 100 mL buffer A containing 50 mM HEPES pH
556 7.5, and 500 mM NaCl, 10% glycerol, and 5 mM TCEP. Cells were lysed by sonication, and the
557 lysate was centrifuged at 17,000 g for 50 min at 4 °C. The supernatant was transferred into
558 anaerobic conditioned glove box and applied onto the pre-equilibrated 4 mL Ni-NTA column
559 (SUMO tagged expression) or 5 mL GST column (GST tagged expression). After washing with
560 100 ml of buffer A, the protein was eluted with 20 ml buffer B (50 mM HEPES pH 7.5, 500 mM
561 NaCl, 10% glycerol, 300 mM imidazole, and 5 mM TCEP for SUMO tagged purification and
562 50 mM HEPES pH 7.5, 500 mM NaCl, 10% glycerol, 15 mM reduced GSH, and 5 mM TCEP for
563 GST tagged purification), then incubated with SUMO-protease or 3C protease at 4 °C for 2
564 hours. The sample was then concentrated to 2 ml and loaded onto a Superdex 200 16/60 size-
565 exclusion column (GE Healthcare) equilibrated with buffer C (10 mM HEPES pH 7.5, 500 mM
566 NaCl, and 5 mM TCEP), the peak fractions were pooled and snap-frozen in liquid nitrogen for
567 later usage.

568
569 Full-length *cas2* (Gsu0058 in KEGG) genes were cloned from *Geobacter*
570 *sulfurreducens* genomic DNA into His₆-Twin-Strep-SUMO-pET28a vectors (Kan^R) between
571 BamHI and XhoI sites. Sequence-verified plasmids were transformed into *E. coli* BL21 (DE3)
572 star cells. A 4 liters cell culture was grown in LB medium at 37 °C until an optical density of 0.8
573 at 600 nm. Expression was induced by adding IPTG to a final concentration of 0.5 mM at 25 °C
574 overnight. Cells were harvested by centrifugation and lysed by sonication in 80 ml buffer A
575 containing 50 mM HEPES pH 7.5, 20 mM imidazole and 500 mM NaCl, 10% glycerol, and 2 mM
576 B-ME. The lysate was centrifuged at 17,000 g for 50 min at 4 °C, and the supernatant was
577 applied onto the pre-equilibrated 4 mL Ni-NTA column. After washing with 100 ml of buffer A,
578 the protein was eluted with 20 ml buffer B (50 mM HEPES pH 7.5, 500 mM NaCl, 10% glycerol,
579 300 mM imidazole, and 2 mM B-ME), and incubated with SUMO-protease at 4 °C for 3 hours.
580 The tag cleaved Cas2 proteins were purified on Superdex 200 16/60 equilibrated with buffer C
581 (10 mM HEPES pH 7.5, 500 mM NaCl), the peak fractions were pooled and snap-frozen in liquid
582 nitrogen for later usage.

583

584 **Affinity pull-down assay**

585 15 µg GST-tagged Cas4/1 and 30 µg untagged Cas2 were mixed and incubated with 10 µL
586 GST resin at 4 °C for 30 min in different salt concentration buffer (50 mM HEPES pH7.5, 10%
587 glycerol, 5 mM TCEP, and 150/300/500 mM NaCl) in presence or absence of prespacer, in a
588 total assay volume of 50 µL. The GST resin was pelleted by centrifugation at ~100 g for 30
589 seconds, washed 3 times with 200 µL of the corresponding binding buffer, then eluted with 70
590 µL elution buffer (50 mM HEPES pH7.5, 500 mM NaCl, 5 mM TCEP, and 15 mM reduced
591 GSH). Eluted proteins were separated on 12% SDS-PAGE and stained by Coomassie blue.

592

593 **Fluorescently labeled prespacer substrate preparation**

594 Fluorescent DNA oligos (**Supplementary Table 2**) for biochemistry were synthesized
595 (Integrated DNA Technologies) with either a /5AmMC6/ or /3AmMO/ label, fluorescently labeled
596 in-house, and annealed at equimolar amount, and native PAGE purified to remove unannealed
597 ssDNA.

598

599 **Prespacer cleavage assays**

600 Prespacer cleavage assays were set up in 20 µL reactions containing 10 nM final concentration
601 of labelled prespacer, 500 nM Cas4/1, 250 nM Cas2 in a cleavage buffer containing 50 mM Tris
602 pH 8.0, 100 mM KCl, 10% Glycerol, 5 mM TCEP, and 5 mM metal ion MnCl₂ or different metal
603 ions in **Extended data Fig. 1h**. After 37 °C incubation for 60 min, reactions were quenched by
604 vortexing with 20µL of phenol-chloroform. The extracted aqueous phases were mixed with equi-
605 volume of 100% formamide and separated on 13% urea-PAGE. Signals from each fluorescent
606 dye were recorded at its corresponding excitation wavelength using a ChemiDoc imaging
607 system (Bio-Rad). The KMnO₄ foot printing assay was carried out following previously published
608 protocols ⁶⁴.

609

610 **Reconstitution of prespacer bound/integration Cas4/1-2 complex**

611 Complex was formed by mixing Cas4₂/Cas1₂, Cas2, and prespacer (or half-integration
612 mimicking substrate) at a final concentration of 30 µM, 60 µM, and 60 µM respectively in 500 µL
613 total volume with a reconstitute buffer containing 25 mM Tris pH 8.0, 300 mM NaCl, 5 mM
614 TCEP and 5 mM MnCl₂. After 37 °C incubation for 30 min, the complex was separated on
615 Superdex 200 16/30 column equilibrated in the same buffer. The full-complex peak was pooled

616 and concentrated to appropriate concentration and snap-frozen in liquid nitrogen for long-term
617 storage.

618

619 **Integration assays**

620 The *in vitro* integration assays were set up as follows. 10 nM of prespacer were incubated with
621 250 nM Cas4/1-2 complex in the integration buffer containing 50 mM Tris pH 8.0, 100 mM KCl, 5
622 mM TCEP and 5 mM MnCl₂ in 20 µL reaction volume. After an initial incubation at 37 °C for 5
623 minutes, 300 ng of pCRISPR plasmid was introduced into the reaction. Integration was allowed
624 at 37 °C for 60 min, after which 0.5 µL of EcoRI and XhoI restriction enzymes (NEB) were
625 introduced for 10 min more at 37 °C to digest out the leader-repeat region of the plasmid,
626 together with the integrated prespacer. Reactions were quenched by vortexing with 20 µL
627 phenol-chloroform solution. The extracted aqueous phase was mixed with equi-volume of
628 formamide, separated on 13% urea-PAGE, and scanned on ChemiDoc imaging system.

629

630 **Exol trimming and follow-up integration assays**

631 10 nM of prespacer were pre-incubated with 250 nM of Cas4/1-2 complex at 37 °C for 5 minutes
632 in 20 µL containing the trimming buffer (50 mM Tris pH 8.0, 100 mM KCl, 10% glycerol, 5 mM
633 TCEP, 5 mM MnCl₂ and 10 mM MgCl₂). The 2-fold Exol dilution series in Fig. 4b was prepared
634 by dilution *E. coli* Exol (NEB, 20 Units/µL) to a final concentration of 0.2, 0.1, 0.05, 0.025,
635 0.0125 Units/µL in each reaction. The 1/10 and 1/50 Exol concentrations in the Extended Data
636 Fig. 9a correspond to 0.1, 0.02 Units/µL. The Exol concentration in the Extended Data Fig. 9b
637 was 0.1 Units/µL across. In reactions where the trimming and integration were coupled, 300 ng
638 of pCRISPR plasmid (~ 5 nM final concentration) was introduced at the same time with Exol into
639 the reaction. After incubation, the reaction was quenched by mixing with equi-volume of a buffer
640 containing 95% formamide, 10 mM EDTA and 0.2% SDS, phenol-extracted, then separated on
641 13% urea-PAGE, and scanned on ChemiDoc imaging system (Bio-Rad), as described above.

642

643 **Electrophoretic mobility shift assay**

644 2 nM final concentration of fluorescently labeled prespacer DNA was incubated with an
645 increasing concentration of Cas4/1-2 complex for 15 minutes (in concentration titration
646 experiments), or with 50 nM Cas4/1-2 complex for 0.5, 1, 2, 5 minutes (in time-course
647 experiments) at 4 °C in a total 20 µL system containing 50 mM Tris pH 8.0, 100 mM KCl, 5 mM
648 TCEP, 5 mM MnCl₂ and 10% glycerol. After incubation, 15 µL of each sample was loaded onto
649 1% agarose gel equilibrated in 1x TG buffer (20 mM Tris pH 8.0, 200 mM Glycine) immediately.

650 Electrophoresis was performed at 60 V for 40 min. The fluorescent signals from the gel were
651 recorded using a ChemiDoc imaging system (Bio-Rad).

652

653 **Negative-stain electron microscopy**

654 4 μ L of 0.01 mg/mL prespacer-bound Cas4/1-2 complex was applied to a glow-discharged
655 copper 400-mesh continuous carbon grid. After a 30-second incubation, the grid was blotted on
656 a filter paper, immediately transferred carbon-face down on top of a 2% (w/v) uranyl acetate
657 solution for 60 seconds. The grid was then blotted on a filter paper again to remove residual
658 stain, then air-dried on bench for 5 min. The grid was examined under a Morgagni transmission
659 electron microscope operated at 100 keV with a direct magnification of $\times 140000$ (3.2 \AA° pixel
660 size) by AMT camera system. Each image was acquired using a 800 ms exposure time and -1
661 to -2 mm defocus setting. Data processing and 2D classification were performed on CyoSPARC
662 software.

663

664 **Cryo-EM data acquisition**

665 4 μ L of 0.6 mg/mL SEC-purified prespacer-bound or half-integration mimicking substrate-bound
666 Cas4/1-2 complexes were applied to a Quantifoil holey carbon grid (1.2/1.3, 400 mesh) which
667 had been glow-discharged for 30s. Grids were blotted for 4 s at 6°C , 100% humidity and
668 plunge-frozen in liquid ethane using a Mark IV FEI/Thermo Fisher Vitrobot. Cryo-EM images
669 were collected on a 200 kV Talos Arctica transmission microscope (Thermo Fisher) equipped
670 with a K3 Summit direct electron detector (Gatan). The total exposure time of each movie stack
671 was ~ 3.5 s, leading to a total accumulated dose of 50 electrons per Å° which fractionated into
672 50 frames. Dose fractionated super-resolution movie stacks collected from the K3 Summit direct
673 electron detector were 1x binned to a pixel size of 1.234 \AA° . The defocus value was set between
674 $-1.5 \mu\text{m}$ to $-3.5 \mu\text{m}$.

675

676 **Cryo-EM data processing**

677 Motion correction, CTF-estimation, blob particle picking, 2D classification, 3D classification and
678 non-uniform 3D refinement were performed in cryoSPARC v.2⁶⁵. Refinements followed the
679 standard procedure, a series of 2D and 3D classifications with *C1* symmetry were performed as
680 shown in Extended Data Fig. 4a, Extended Data Fig. 7 and Extended Data Fig. 10a, to
681 generate the final maps. A solvent mask was generated and was used for all subsequent
682 refinement steps. CTF post refinement was conducted to refine the beam-induced motion of the
683 particle set, resulting in the final maps. The final map 'CTF Post-refinement was used to

684 estimate resolution based on the Fourier shell correlation (FSC) = 0.143 criterion after
685 correcting for the effects of a soft shape mask using high-resolution noise substitution. We
686 noticed that the map of the full-integration complex was not homogeneous in both sides, so we
687 divided the map into two half parts from the middle site by Chimera UCSF. Then imported two
688 half maps into Relion 3.0⁶⁶ to make a mask for next masked local refinement respectively.
689 Finally imported these two masks into cryoSPARC again and did a local refinement to get two
690 half local refined maps and merged two maps to a final map in Extended data Fig.10. The
691 detailed data processing and refinement statistics for all cryo-EM structures are summarized in
692 Extended figures and Supplementary table 3.

693

694 **Data availability**

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

706

707 **Author Contributions**

708 A.K., S.J.J.B., C.H., and C.A. designed the research. C.H. is responsible for biochemistry and
709 cryo-EM reconstructions; C.A., J.N.A.V., A.R.C, A.C.H. for *in vivo* and bioinformatics analyses;
710 K.N., C.H. for structure building and refinement; and S.R.B. for assistance in cryo-EM work.
711 A.K., C.H. wrote the manuscript, with input from S.J.J.B., J.N.A.V. and A.R.C.

712

713 **Acknowledgements**

714 This work is supported by the Netherlands Organization for Scientific Research (NWO) VICI
715 grant [VI.C.182.027] to S.J.J.B. and the National Institutes of Health (NIH) grant [GM118174] to

716 A.K.. This work made use of the Cornell Center for Materials Research Shared Facilities which
717 are supported through the NSF MRSEC program (DMR-1719875). We thank Sebastian N.
718 Kieper, Romano Miojevic, Mariena Ramos, Gabriel Schuler and Katherine Spoth for helpful
719 discussions, advice and technical assistance. The authors declare no competing financial
720 interests. Correspondence should be addressed to A.K. and S.J.J.B. (ailong.ke@cornell.edu;
721 stanbrouns@gmail.com). Reprints and permissions information are available upon request.

722

723 References

- 724 1 Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in
725 prokaryotes. *Science* **315**, 1709-1712, doi:10.1126/science.1138140 (2007).
- 726 2 Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the
727 CRISPR adaptation process in *Escherichia coli*. *Nucleic acids research* **40**, 5569-5576
728 (2012).
- 729 3 Nuñez, J. K. *et al.* Cas1–Cas2 complex formation mediates spacer acquisition during
730 CRISPR–Cas adaptive immunity. *Nature structural & molecular biology* **21**, 528-534
731 (2014).
- 732 4 Nuñez, J. K., Lee, A. S., Engelman, A. & Doudna, J. A. Integrase-mediated spacer
733 acquisition during CRISPR-Cas adaptive immunity. *Nature* **519**, 193-198 (2015).
- 734 5 Mojica, F. J., Diez-Villasenor, C., Garcia-Martinez, J. & Almendros, C. Short motif
735 sequences determine the targets of the prokaryotic CRISPR defence system.
736 *Microbiology* **155**, 733-740, doi:10.1099/mic.0.023960-0 (2009).
- 737 6 Marraffini, L. A. & Sontheimer, E. J. Self versus non-self discrimination during CRISPR
738 RNA-directed immunity. *Nature* **463**, 568-571, doi:10.1038/nature08703 (2010).
- 739 7 Jackson, S. A. *et al.* CRISPR-Cas: Adapting to change. *Science* **356**, doi:ARTN
740 eaal505610.1126/science.aal5056 (2017).
- 741 8 McGinn, J. & Marraffini, L. A. Molecular mechanisms of CRISPR-Cas spacer acquisition.
742 *Nat Rev Microbiol* **17**, 7-12, doi:10.1038/s41579-018-0071-7 (2019).
- 743 9 Deveau, H. *et al.* Phage response to CRISPR-encoded resistance in *Streptococcus*
744 *thermophilus*. *J Bacteriol* **190**, 1390-1400, doi:10.1128/JB.01412-07 (2008).
- 745 10 Almendros, C., Guzman, N. M., Diez-Villasenor, C., Garcia-Martinez, J. & Mojica, F. J.
746 Target motifs affecting natural immunity by a constitutive CRISPR-Cas system in
747 *Escherichia coli*. *PLoS One* **7**, e50797, doi:10.1371/journal.pone.0050797 (2012).
- 748 11 Mojica, F. J. M., Diez-Villasenor, C., Garcia-Martinez, J. & Almendros, C. Short motif
749 sequences determine the targets of the prokaryotic CRISPR defence system.
750 *Microbiology (Reading)* **155**, 733-740, doi:10.1099/mic.0.023960-0 (2009).
- 751 12 Vink, J. N. A. *et al.* Direct Visualization of Native CRISPR Target Search in Live Bacteria
752 Reveals Cascade DNA Surveillance Mechanism. *Mol Cell* **77**, 39-50 e10,
753 doi:10.1016/j.molcel.2019.10.021 (2020).
- 754 13 Makarova, K. S. *et al.* Evolutionary classification of CRISPR-Cas systems: a burst of
755 class 2 and derived variants. *Nat Rev Microbiol* **18**, 67-83, doi:10.1038/s41579-019-
756 0299-x (2020).
- 757 14 Hudaiberdiev, S. *et al.* Phylogenomics of Cas4 family nucleases. *Bmc Evol Biol* **17**,
758 doi:ARTN 23210.1186/s12862-017-1081-1 (2017).
- 759 15 Nuñez, J. K., Harrington, L. B., Kranzusch, P. J., Engelman, A. N. & Doudna, J. A.
760 Foreign DNA capture during CRISPR–Cas adaptive immunity. *Nature* **527**, 535-538
761 (2015).

762 16 Wang, J. *et al.* Structural and mechanistic basis of PAM-dependent spacer acquisition in
763 CRISPR-Cas systems. *Cell* **163**, 840-853 (2015).

764 17 Wright, A. V. & Doudna, J. A. Protecting genome integrity during CRISPR immune
765 adaptation. *Nature Structural & Molecular Biology* (2016).

766 18 Wright, A. V. *et al.* Structures of the CRISPR genome integration complex. *Science* **357**,
767 1113-1118, doi:10.1126/science.aao0679 (2017).

768 19 Budhathoki, J. B. *et al.* Real-time observation of CRISPR spacer acquisition by Cas1-
769 Cas2 integrase. *Nat Struct Mol Biol* **27**, 489-499, doi:10.1038/s41594-020-0415-7
770 (2020).

771 20 Xiao, Y., Ng, S., Nam, K. H. & Ke, A. How type II CRISPR-Cas establish immunity
772 through Cas1-Cas2-mediated spacer integration. *Nature* **550**, 137-141,
773 doi:10.1038/nature24020 (2017).

774 21 Kim, S. *et al.* Selective loading and processing of prespacers for precise CRISPR
775 adaptation. *Nature* **579**, 141-+, doi:10.1038/s41586-020-2018-1 (2020).

776 22 Li, M., Wang, R., Zhao, D. & Xiang, H. Adaptation of the *Haloarcula hispanica* CRISPR-
777 Cas system to a purified virus strictly requires a priming process. *Nucleic Acids Res* **42**,
778 2483-2492, doi:10.1093/nar/gkt1154 (2014).

779 23 Liu, T. *et al.* Coupling transcriptional activation of CRISPR-Cas system and DNA repair
780 genes by Csa3a in *Sulfolobus islandicus*. *Nucleic Acids Res* **45**, 8978-8992,
781 doi:10.1093/nar/gkx612 (2017).

782 24 Shiimori, M., Garrett, S. C., Graveley, B. R. & Terns, M. P. Cas4 Nucleases Define the
783 PAM, Length, and Orientation of DNA Fragments Integrated at CRISPR Loci. *Molecular*
784 *Cell* **70**, 814-+, doi:10.1016/j.molcel.2018.05.002 (2018).

785 25 Kieper, S. N. *et al.* Cas4 Facilitates PAM-Compatible Spacer Selection during CRISPR
786 Adaptation. *Cell Reports* **22**, 3377-3384, doi:10.1016/j.celrep.2018.02.103 (2018).

787 26 Almendros, C., Nobrega, F. L., McKenzie, R. E. & Brouns, S. J. J. Cas4-Cas1 fusions
788 drive efficient PAM selection and control CRISPR adaptation. *Nucleic Acids Res* **47**,
789 5223-5230, doi:10.1093/nar/gkz217 (2019).

790 27 Lemak, S. *et al.* Toroidal structure and DNA cleavage by the CRISPR-associated [4Fe-
791 4S] cluster containing Cas4 nuclease SSO0001 from *Sulfolobus solfataricus*. *Journal of*
792 *the American Chemical Society* **135**, 17476-17487, doi:10.1021/ja408729b (2013).

793 28 Lemak, S. *et al.* The CRISPR-associated Cas4 protein Pcal_0546 from *Pyrobaculum*
794 *calidifontis* contains a [2Fe-2S] cluster: crystal structure and nuclease activity. *Nucleic*
795 *Acids Res* **42**, 11144-11155, doi:10.1093/nar/gku797 (2014).

796 29 Zhang, J., Kasciukovic, T. & White, M. F. The CRISPR Associated Protein Cas4 Is a 5 '
797 to 3 ' DNA Exonuclease with an Iron-Sulfur Cluster. *Plos One* **7**, doi:ARTN
798 e4723210.1371/journal.pone.0047232 (2012).

799 30 Lee, H., Dhingra, Y. & Sashital, D. G. The Cas4-Cas1-Cas2 complex mediates precise
800 prespacer processing during CRISPR adaptation. *Elife* **8**, doi:ARTN
801 e4424810.7554/eLife.44248 (2019).

802 31 Lee, H., Zhou, Y., Taylor, D. W. & Sashital, D. G. Cas4-Dependent Prespacer
803 Processing Ensures High-Fidelity Programming of CRISPR Arrays. *Molecular Cell* **70**,
804 48-+, doi:10.1016/j.molcel.2018.03.003 (2018).

805 32 Krajewski, W. W. *et al.* Structural basis for translocation by AddAB helicase-nuclease
806 and its arrest at chi sites. *Nature* **508**, 416-419, doi:10.1038/nature13037 (2014).

807 33 Jia, N. *et al.* Structures and single-molecule analysis of bacterial motor nuclease AdnAB
808 illuminate the mechanism of DNA double-strand break resection. *Proceedings of the*
809 *National Academy of Sciences of the United States of America* **116**, 24507-24516,
810 doi:10.1073/pnas.1913546116 (2019).

811 34 Zhou, C., Pourmal, S. & Pavletich, N. P. Dna2 nuclease-helicase structure, mechanism
812 and regulation by Rpa. *Elife* **4**, doi:10.7554/eLife.09832 (2015).

813 35 Singleton, M. R., Dillingham, M. S., Gaudier, M., Kowalczykowski, S. C. & Wigley, D. B.
814 Crystal structure of RecBCD enzyme reveals a machine for processing DNA breaks.
815 *Nature* **432**, 187-193, doi:10.1038/nature02988 (2004).

816 36 Xiao, Y., Luo, M., Dolan, A. E., Liao, M. & Ke, A. Structure basis for RNA-guided DNA
817 degradation by Cascade and Cas3. *Science* **361**, doi:10.1126/science.aat0839 (2018).

818 37 Shah, S. A., Erdmann, S., Mojica, F. J. & Garrett, R. A. Protospacer recognition motifs:
819 mixed identities and functional diversity. *RNA Biol* **10**, 891-899, doi:10.4161/rna.23764
820 (2013).

821 38 Yang, W. Nucleases: diversity of structure, function and mechanism. *Q Rev Biophys* **44**,
822 1-93, doi:10.1017/S0033583510000181 (2011).

823 39 Nuñez, J. K., Bai, L., Harrington, L. B., Hinder, T. L. & Doudna, J. A. CRISPR
824 immunological memory requires a host factor for specificity. *Molecular cell* **62**, 824-833
825 (2016).

826 40 Heler, R. *et al.* Cas9 specifies functional viral targets during CRISPR-Cas adaptation.
827 *Nature* **519**, 199-202, doi:10.1038/nature14245 (2015).

828 41 Levy, A. *et al.* CRISPR adaptation biases explain preference for acquisition of foreign
829 DNA. *Nature* **520**, 505-+, doi:10.1038/nature14302 (2015).

830 42 Modell, J. W., Jiang, W. & Marraffini, L. A. CRISPR-Cas systems exploit viral DNA
831 injection to establish and maintain adaptive immunity. *Nature* **544**, 101-104,
832 doi:10.1038/nature21719 (2017).

833 43 Drabavicius, G. *et al.* DnaQ exonuclease-like domain of Cas2 promotes spacer
834 integration in a type I-E CRISPR-Cas system. *EMBO Rep* **19**,
835 doi:10.15252/embr.201745543 (2018).

836 44 Ramachandran, A., Summerville, L., Learn, B. A., DeBell, L. & Bailey, S. Processing and
837 integration of functionally oriented pre-spacers in the Escherichia coli CRISPR system
838 depends on bacterial host exonucleases. *The Journal of biological chemistry* **295**, 3403-
839 3414, doi:10.1074/jbc.RA119.012196 (2020).

840 45 Pourcel, C. *et al.* CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays
841 and cas genes from complete genome sequences, and tools to download and query lists
842 of repeats and spacers. *Nucleic Acids Res* **48**, D535-D544, doi:10.1093/nar/gkz915
843 (2020).

844 46 Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq): a
845 curated non-redundant sequence database of genomes, transcripts and proteins.
846 *Nucleic Acids Res* **33**, D501-504, doi:10.1093/nar/gki025 (2005).

847 47 Benson, D. A. *et al.* GenBank. *Nucleic Acids Res* **46**, D41-D47, doi:10.1093/nar/gkx1094
848 (2018).

849 48 Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology
850 Information. *Nucleic Acids Res* **37**, D5-15, doi:10.1093/nar/gkn741 (2009).

851 49 Arndt, D. *et al.* PHASTER: a better, faster version of the PHAST phage search tool.
852 *Nucleic Acids Res* **44**, W16-21, doi:10.1093/nar/gkw387 (2016).

853 50 Mitchell, A. L. *et al.* MGnify: the microbiome analysis resource in 2020. *Nucleic Acids*
854 *Res* **48**, D570-D578, doi:10.1093/nar/gkz1035 (2020).

855 51 Chen, I. A. *et al.* IMG/M: integrated genome and metagenome comparative data analysis
856 system. *Nucleic Acids Res* **45**, D507-D516, doi:10.1093/nar/gkw929 (2017).

857 52 Paez-Espino, D. *et al.* IMG/VR v.2.0: an integrated data management and analysis
858 system for cultivated and environmental viral genomes. *Nucleic Acids Res* **47**, D678-
859 D686, doi:10.1093/nar/gky1127 (2019).

860 53 Soto-Perez, P. *et al.* CRISPR-Cas System of a Prevalent Human Gut Bacterium Reveals
861 Hyper-targeting against Phages in a Human Virome Catalog. *Cell Host Microbe* **26**, 325-
862 335 e325, doi:10.1016/j.chom.2019.08.008 (2019).

863 54 Group, N. H. W. *et al.* The NIH Human Microbiome Project. *Genome Res* **19**, 2317-
864 2323, doi:10.1101/gr.096651.109 (2009).

865 55 Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over
866 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell*
867 **176**, 649-662 e620, doi:10.1016/j.cell.2019.01.001 (2019).

868 56 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment
869 search tool. *Journal of molecular biology* **215**, 403-410, doi:10.1016/S0022-
870 2836(05)80360-2 (1990).

871 57 Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-
872 generation sequencing data. *Bioinformatics* **28**, 3150-3152,
873 doi:10.1093/bioinformatics/bts565 (2012).

874 58 Lange, S. J., Alkhnbashi, O. S., Rose, D., Will, S. & Backofen, R. CRISPRmap: an
875 automated classification of repeat conservation in prokaryotic adaptive immune systems.
876 *Nucleic Acids Res* **41**, 8034-8044, doi:10.1093/nar/gkt606 (2013).

877 59 Alkhnbashi, O. S. *et al.* CRISPRstrand: predicting repeat orientations to determine the
878 crRNA-encoding strand at CRISPR loci. *Bioinformatics* **30**, i489-496,
879 doi:10.1093/bioinformatics/btu459 (2014).

880 60 Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood
881 phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307-321,
882 doi:10.1093/sysbio/syq010 (2010).

883 61 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7:
884 improvements in performance and usability. *Mol Biol Evol* **30**, 772-780,
885 doi:10.1093/molbev/mst010 (2013).

886 62 Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo
887 generator. *Genome Res* **14**, 1188-1190, doi:10.1101/gr.849004 (2004).

888 63 McKenzie, R. E., Almendros, C., Vink, J. N. A. & Brouns, S. J. J. Using CAPTURE to
889 detect spacer acquisition in native CRISPR arrays. *Nat Protoc* **14**, 976-990,
890 doi:10.1038/s41596-018-0123-5 (2019).

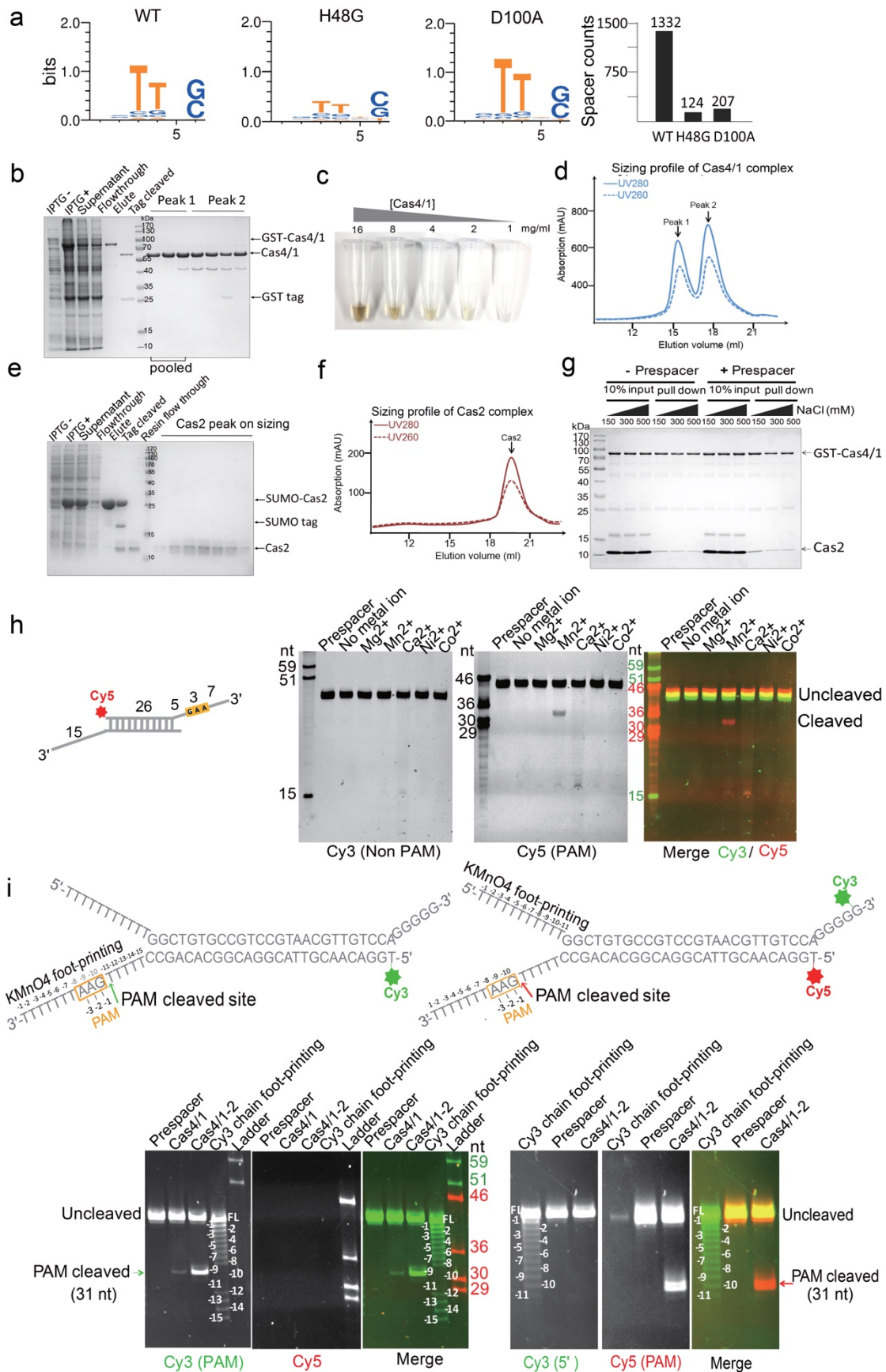
891 64 Xiao, Y. *et al.* Structure Basis for Directional R-loop Formation and Substrate Handover
892 Mechanisms in Type I CRISPR-Cas System. *Cell* **170**, 48-60 e11,
893 doi:10.1016/j.cell.2017.06.012 (2017).

894 65 Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for
895 rapid unsupervised cryo-EM structure determination. *Nat Methods* **14**, 290-296,
896 doi:10.1038/nmeth.4169 (2017).

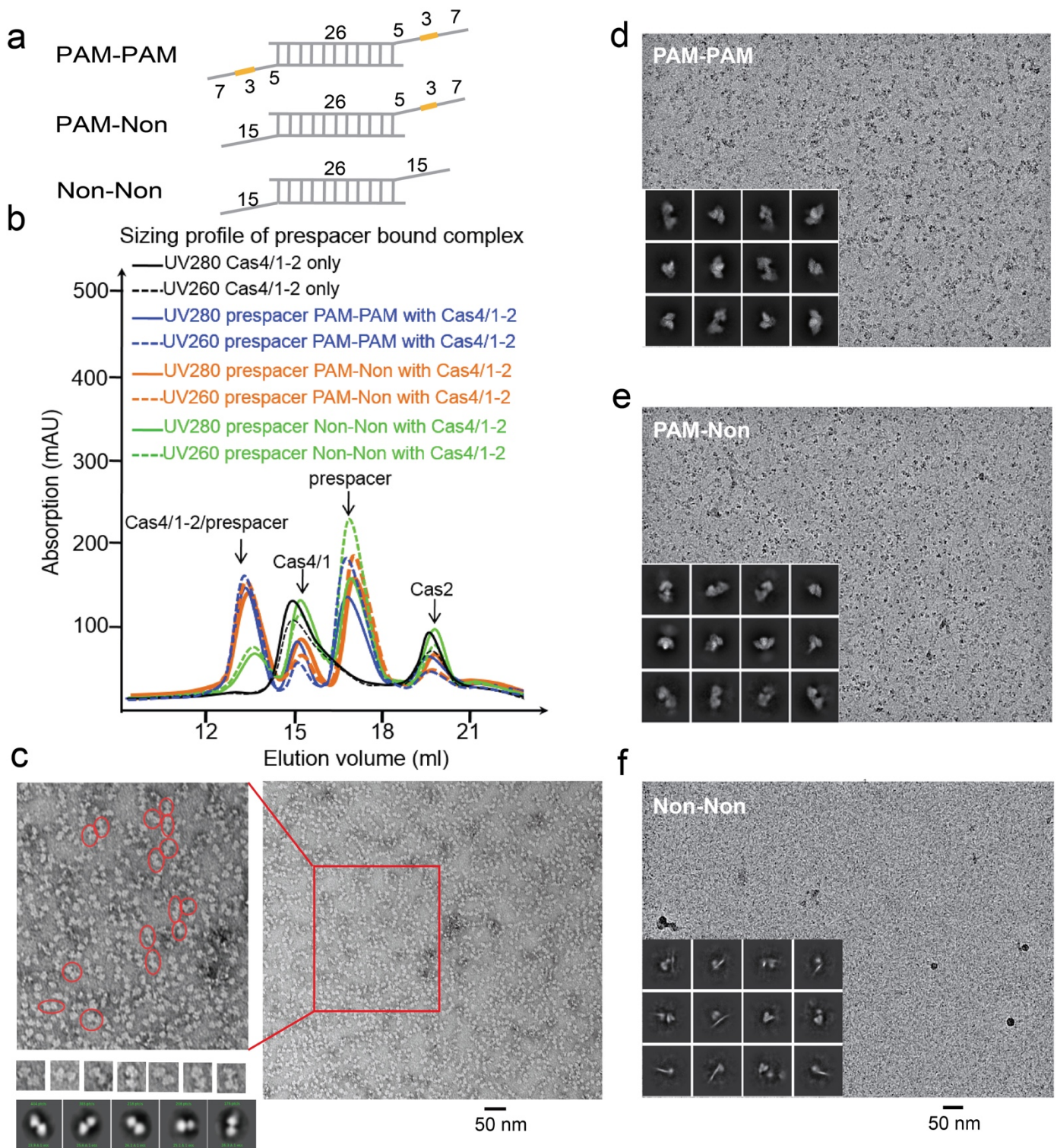
897 66 Xu, K., Zang, X., Peng, M., Zhao, Q. & Lin, B. Magnesium Lithospermate B
898 Downregulates the Levels of Blood Pressure, Inflammation, and Oxidative Stress in
899 Pregnant Rats with Hypertension. *Int J Hypertens* **2020**, 6250425,
900 doi:10.1155/2020/6250425 (2020).

901

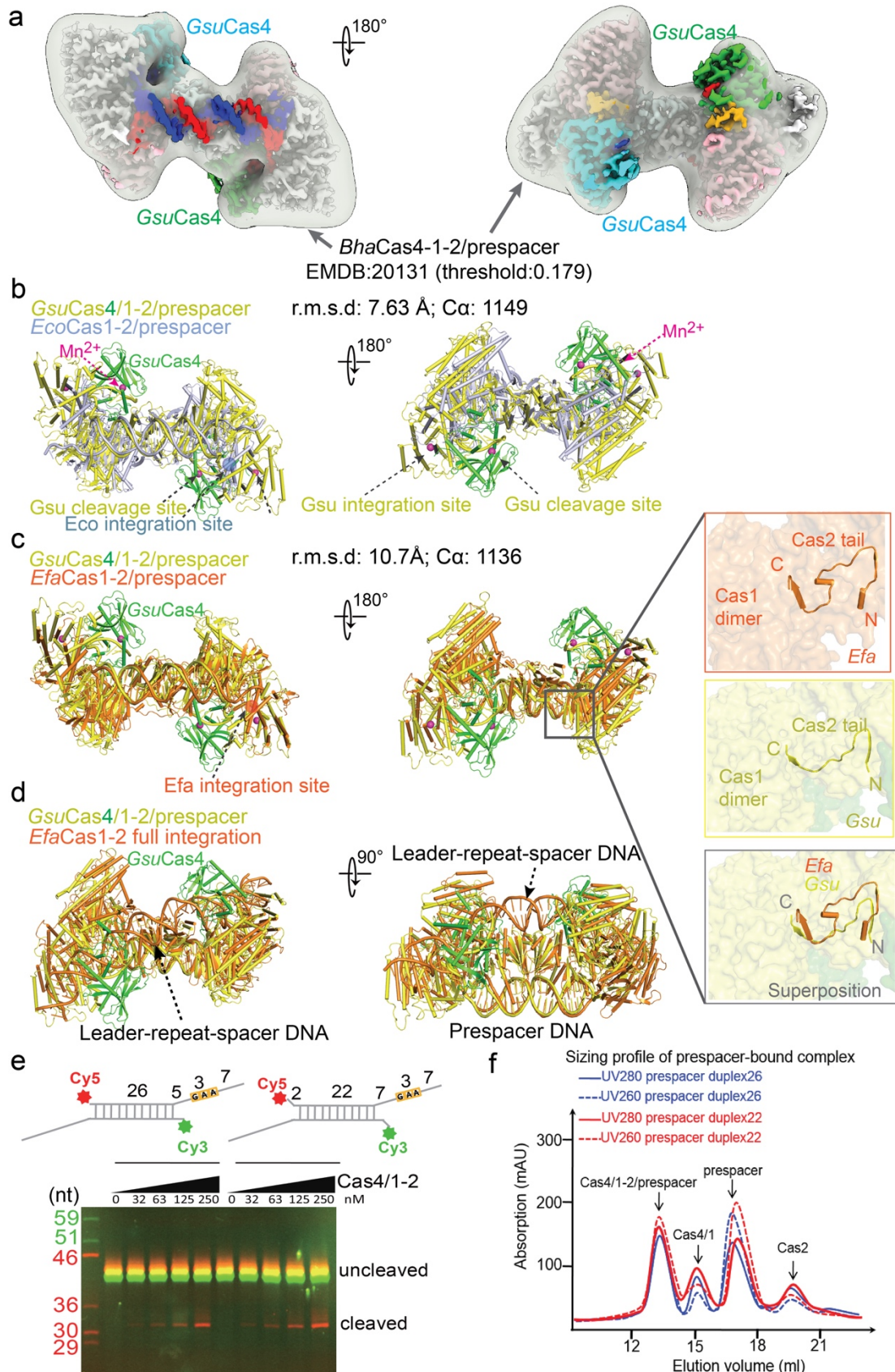
Extended Data Figures



Extended Data Figure 1. Reconstitution and characterization of the *GsuCas4/Cas1-Cas2* complex. **a.** Active site substitution in Cas4 nuclease center (H48G, D100A) reduced *in vivo* spacer acquisition efficiency dramatically. **b-d.** *GsuCas4/1* purification analyzed by SDS-PAGE, coloring from the Fe-S cluster, and SEC profile. **e,f.** Affinity *GsuCas2* purification analyzed by SDS-PAGE and SEC. **g.** GST pull-down results revealing the physical interaction between *GsuCas4/1* and *GsuCas2*, 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.

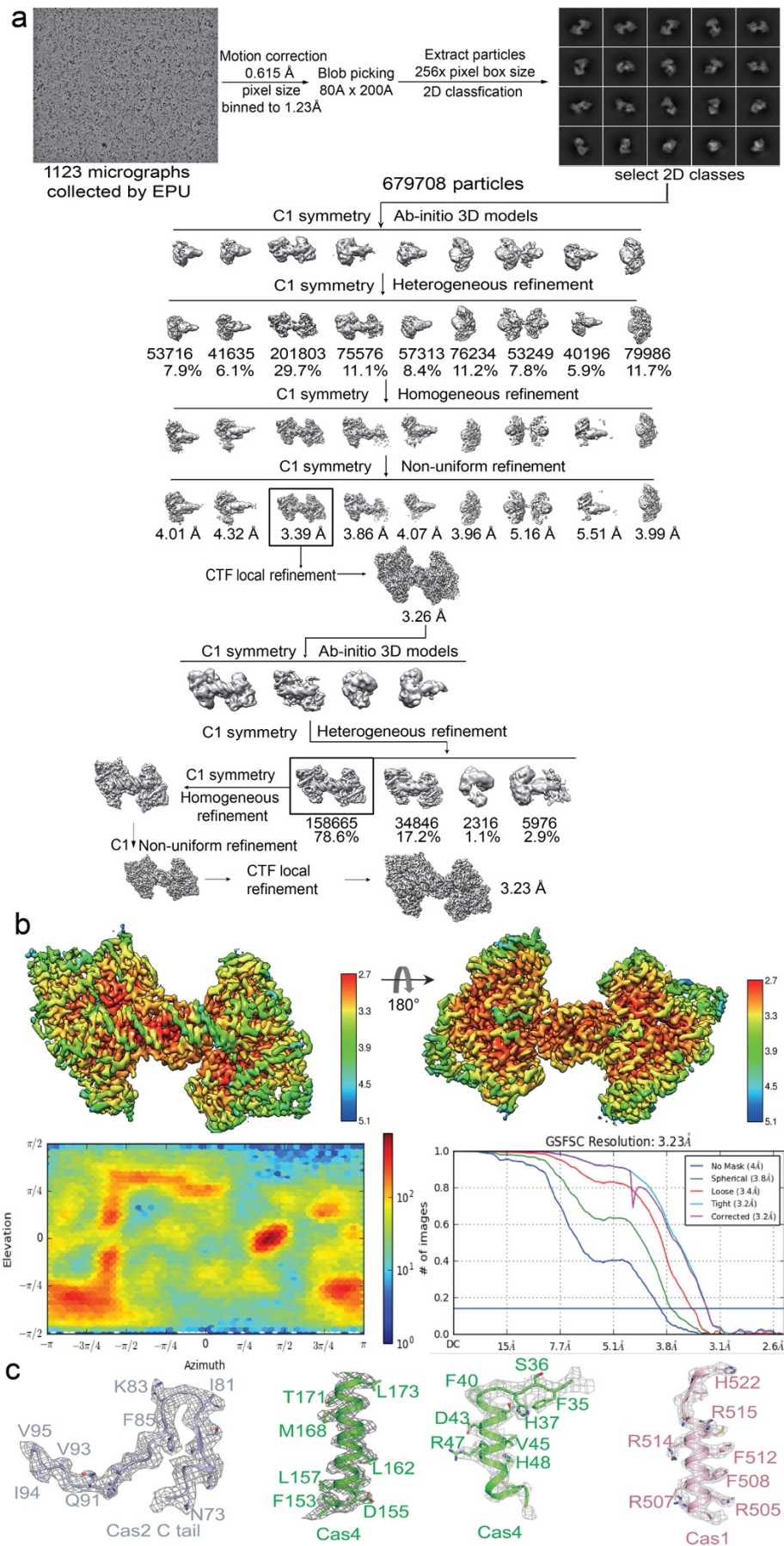


Extended Data Figure 2. PAM-dependent *GsuCas4/Cas1-Cas2* complex formation revealed by SEC and electron microscopy. **a.** Diagram of the prespacer substrates used in complex formation. **b.** SEC profile of *GsuCas4/Cas1-Cas2*, alone or programmed with different prespacer substrates. PAM-containing 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.

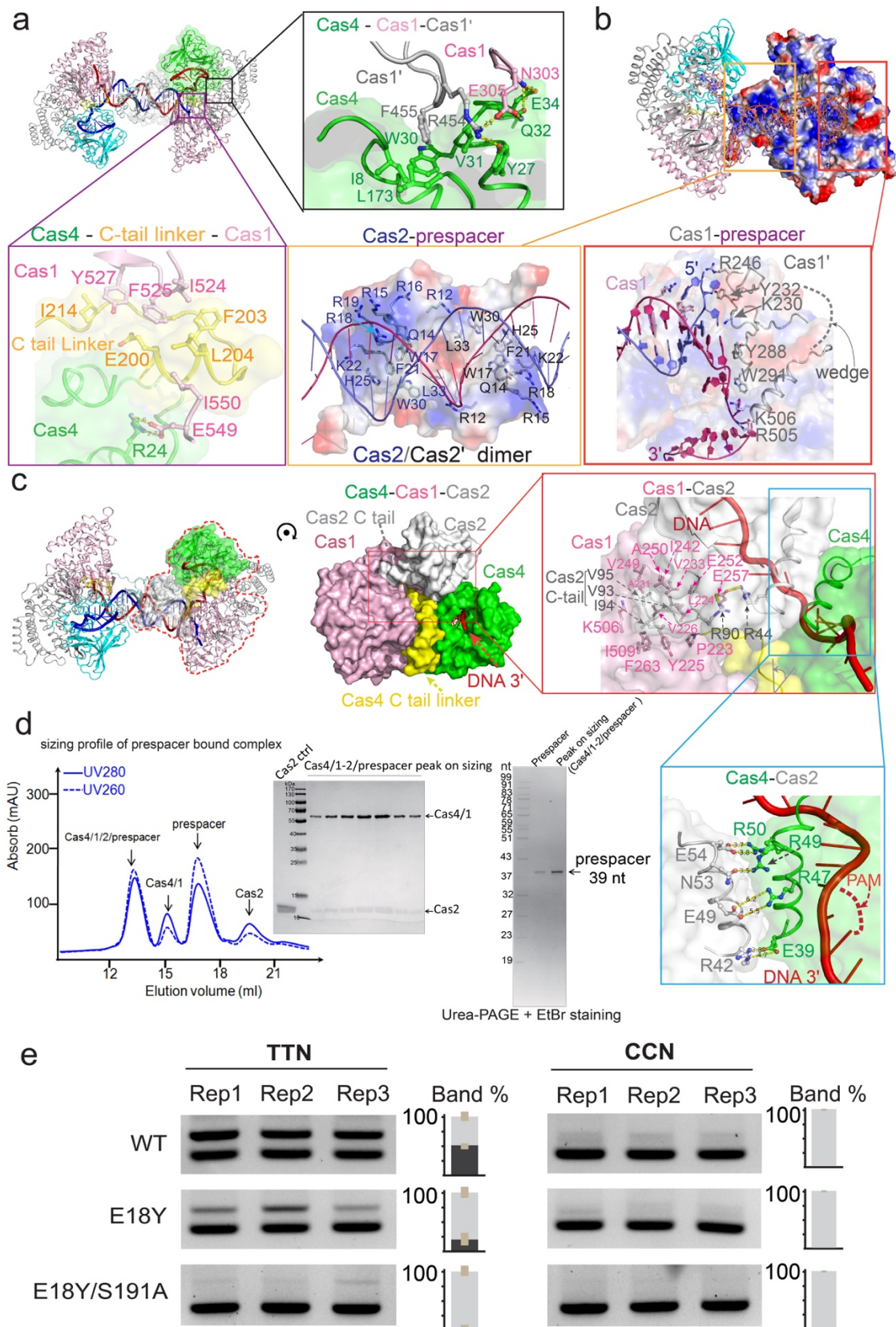


Extended Data Figure 3. Additional analysis of the dual-PAM prespacer bound *GsuCas4/Cas1-Cas2* structure.

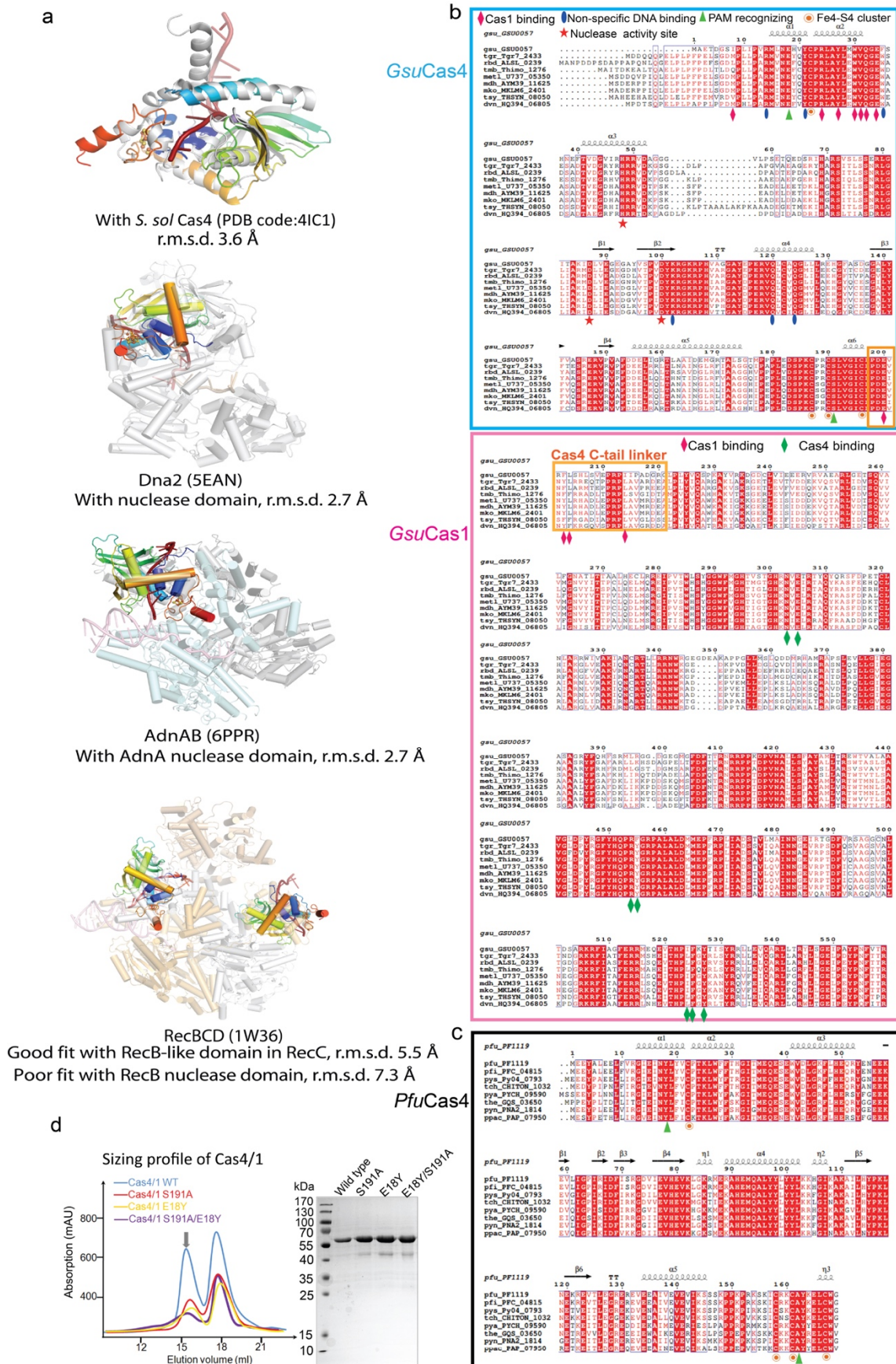
a. Comparison between the current 3.2 Å cryo-EM reconstruction with the previous negative staining reconstruction of the *B. hal* Cas4/1-2 complex (EMDB 20131)¹. **b-d.** Pairwise alignment between *GsuCas4/Cas1-Cas2/prespacer* and *EcoCas1-Cas2/prespacer* (PDB 5DS4), *EfaCas1-Cas2/prespacer* (PDB 5XVN), and *EfaCas1-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 *GsuCas4/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.

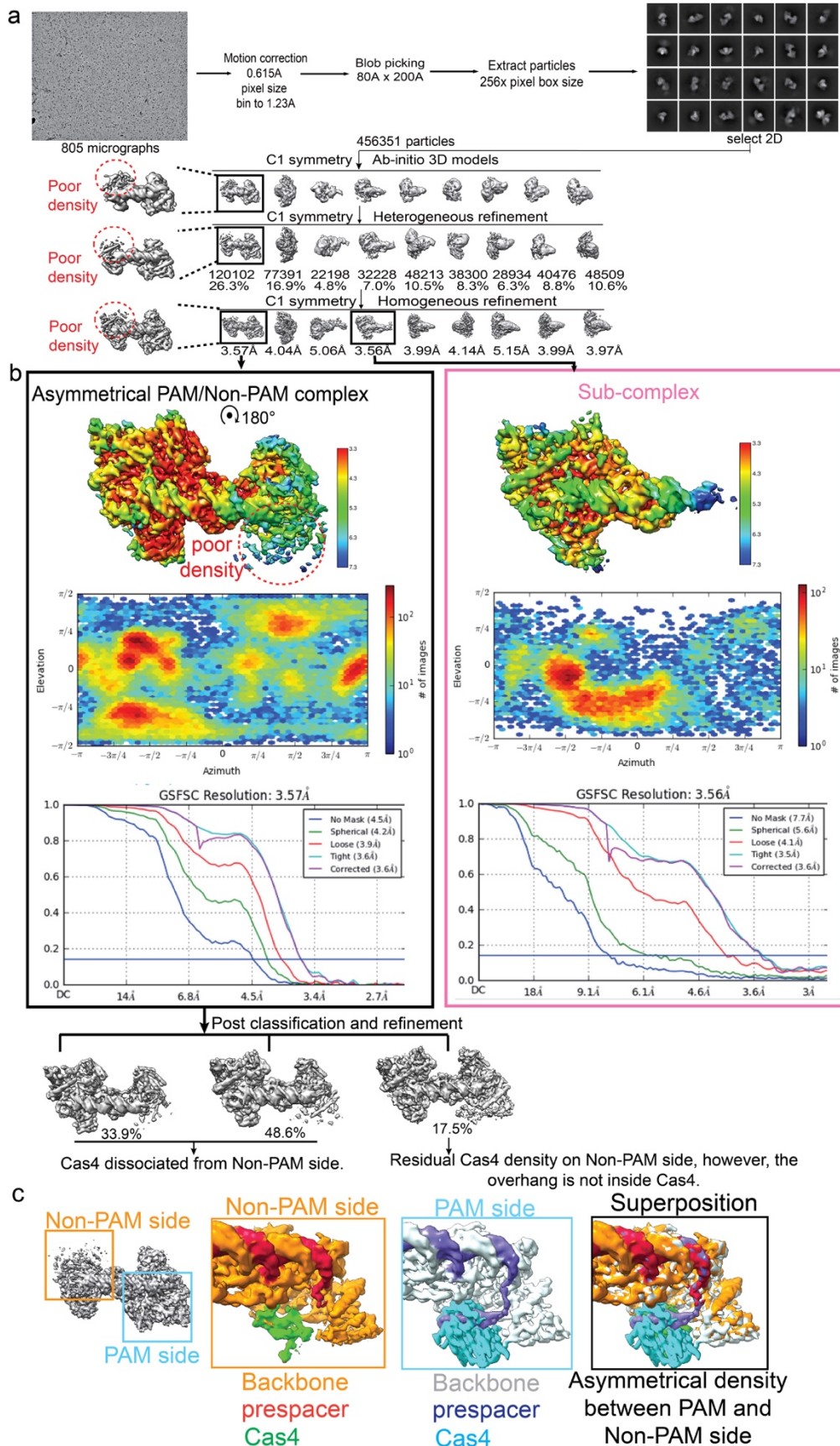


Extended Data Figure 5. In-depth interface analysis of the dual-PAM prespacer-bound *Gsu*Cas4/Cas1-Cas2 structure. **a.** Overall structure. Insets: zoom-ins of Cas4 interface with the neighboring Cas1s. **b.** Surface electrostatic potential. Left inset: Cas2 contacts to the mid-duplex; Right inset: 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.



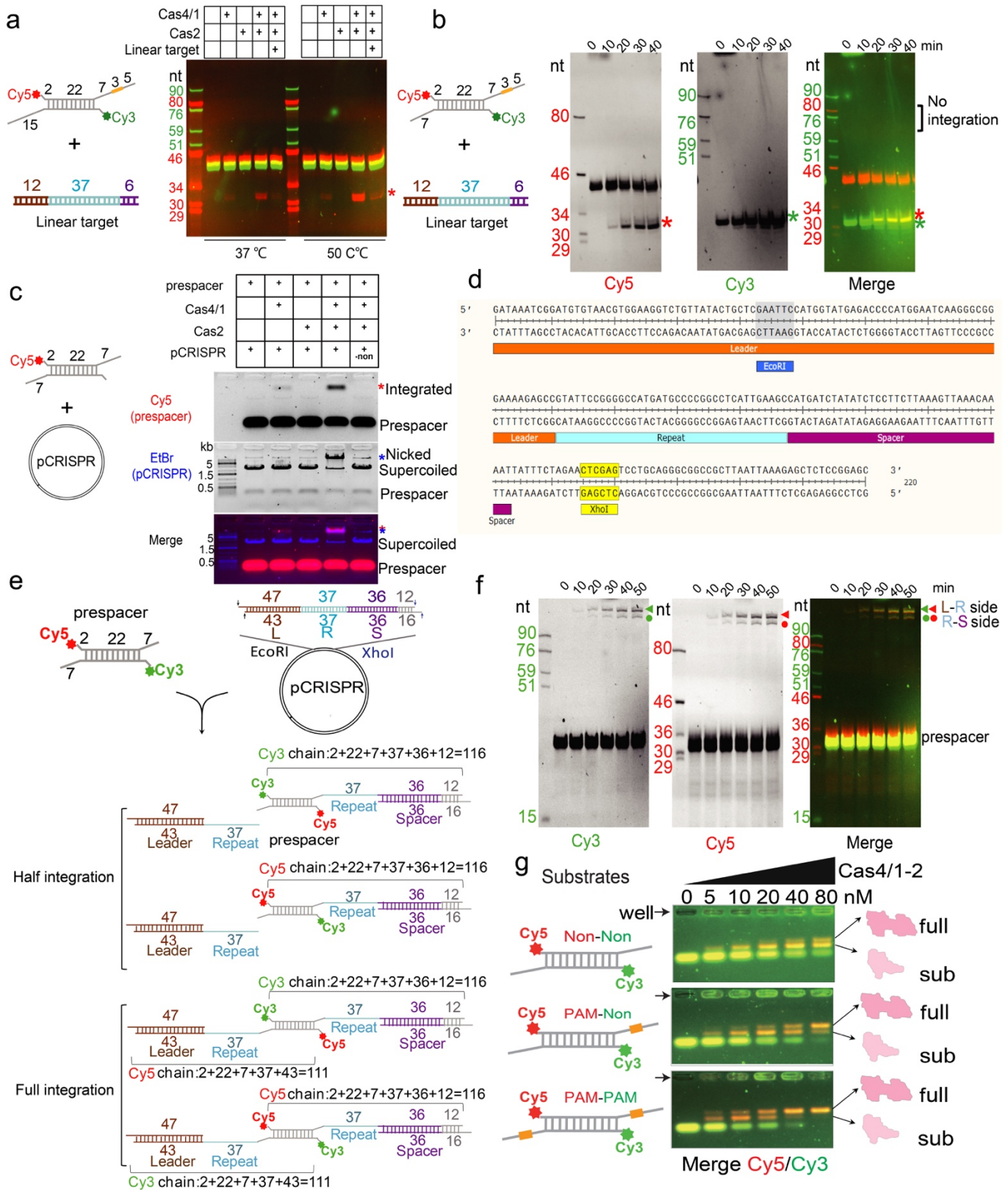
Extended Data Figure 6. In-depth analysis of the structure and sequence conservation in Cas4. a.

Superposition of *GsuCas4* 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 *GsuCas4*, *GsuCas1*, and *PfuCas4* 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 *GsuCas4* mutants that carry the PAM-recognition residues from *PfuCas4*. These mutants were used in the structure-guided PAM-switching experiment in Fig. 3d.

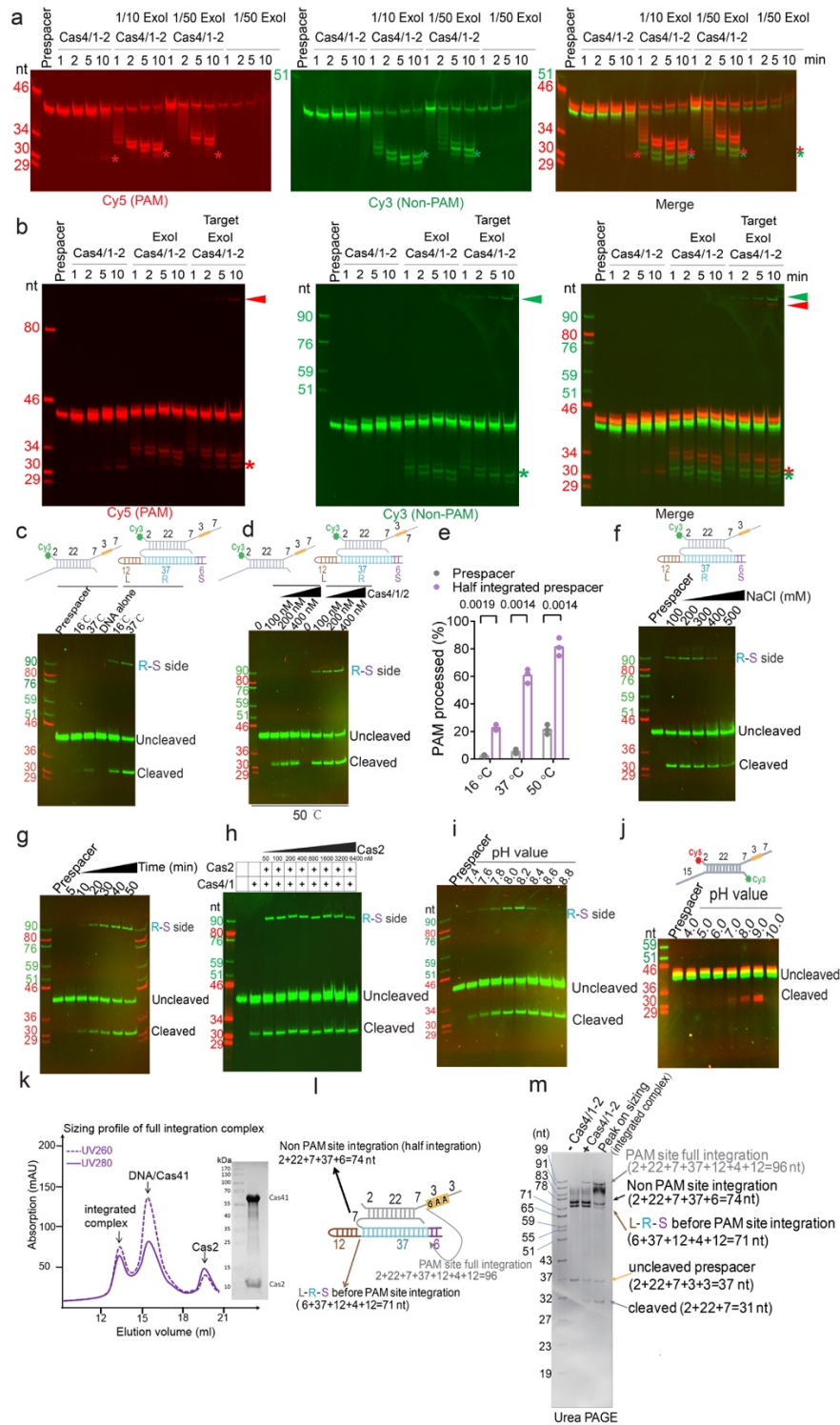


Extended Data Figure 7. Cryo-EM single particle reconstruction of the single-PAM prespacer bound

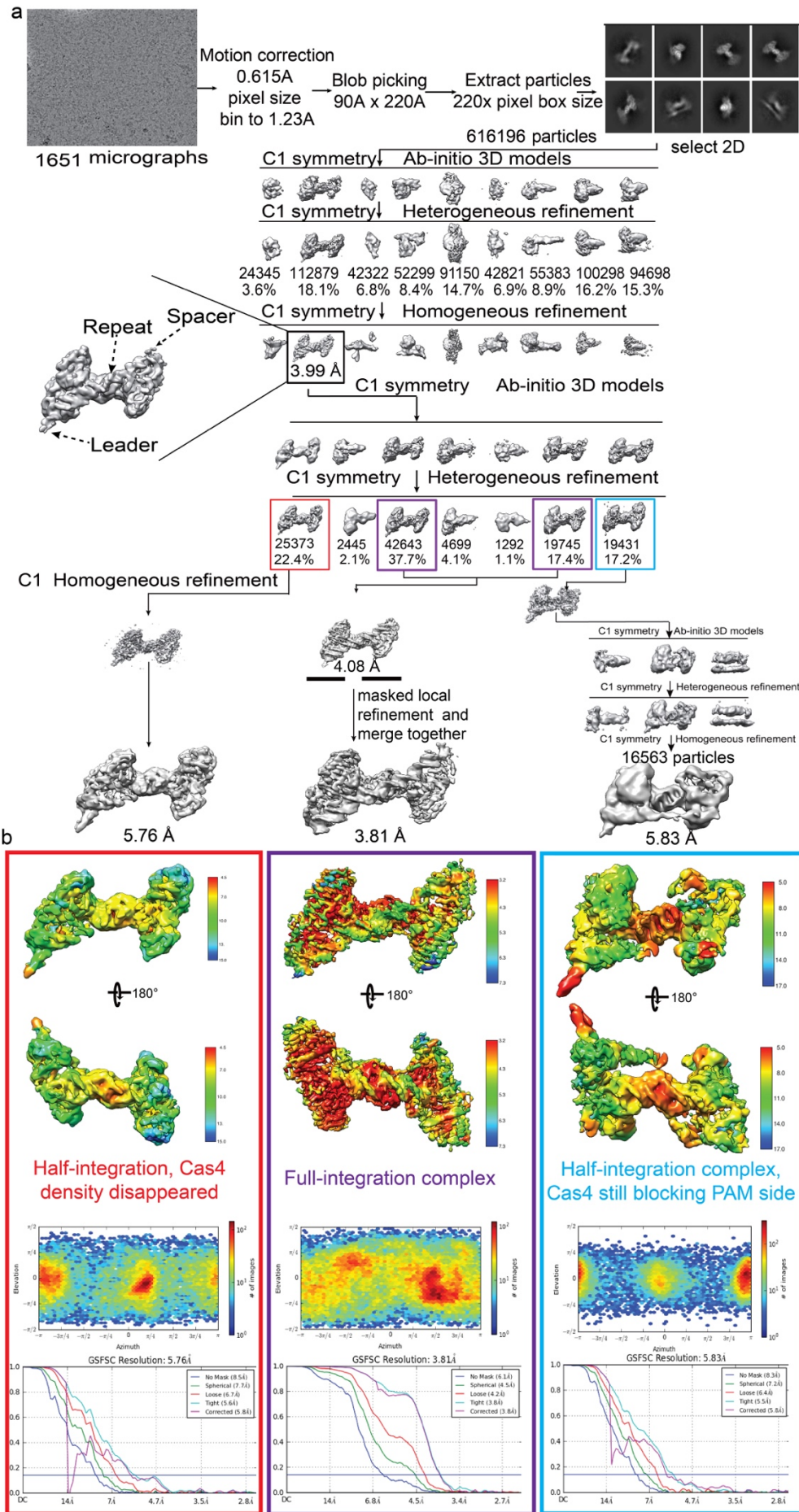
GsuCas4/Cas1-Cas2. **a.** Flow-chart of the cryo-EM single particle reconstruction process that led to the reconstruction of two major snapshots. Left: Asymmetrical PAM/Non-PAM prespacer bound Cas4/1-2 complex. Right: That of the sub complex lacking (Cas4/1)₂ on the non-PAM side. **b.** Cryo-EM density 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.



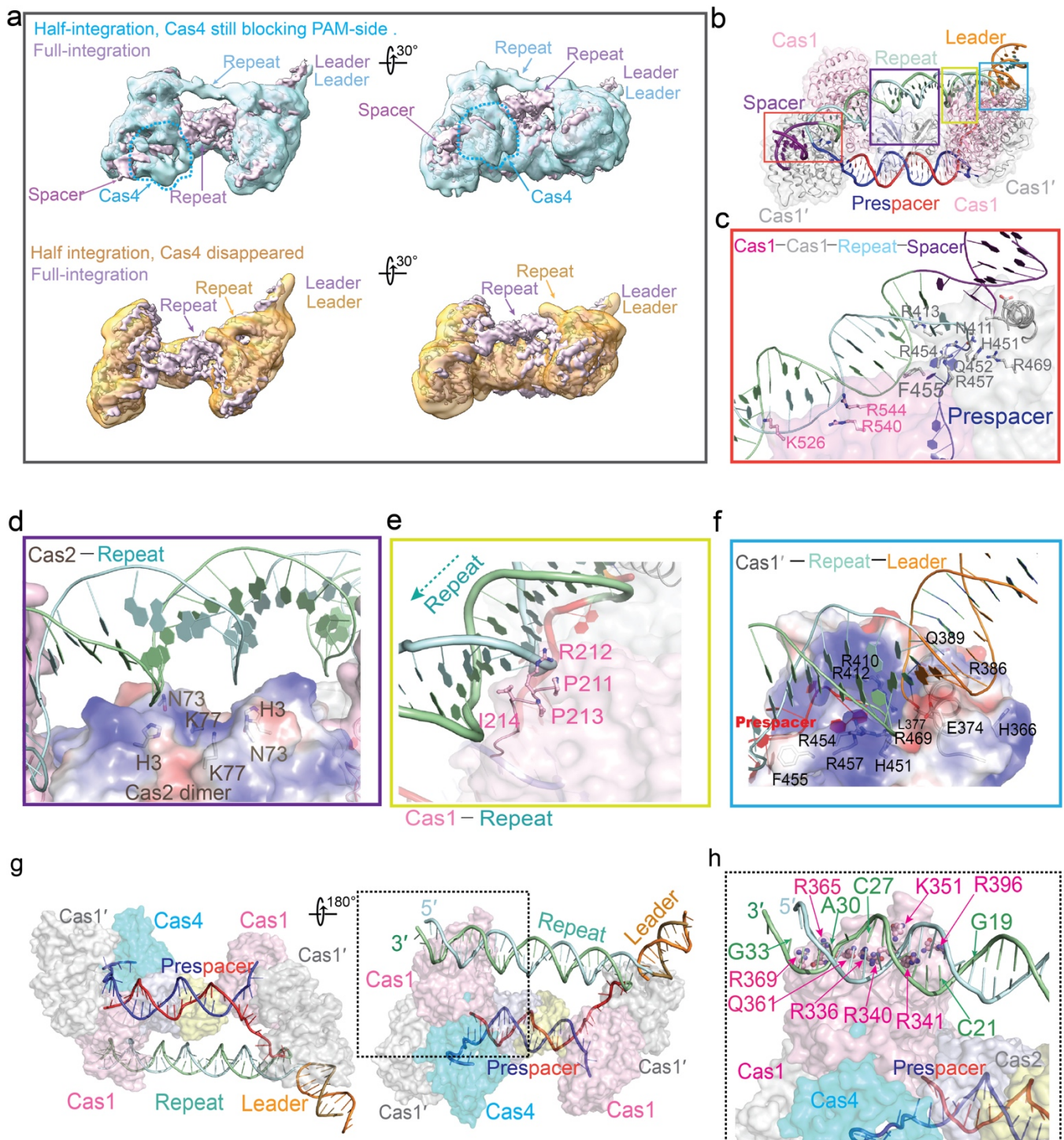
Extended Data Figure 8. *In vitro* integration assay to distinguish integration directionality. **a, b.** Biochemistry showing that Cas4/1-2 is unable to integrate prespacer into the linear form of leader-repeat DNA. **c.** Successful prespacer integration into a leader-repeat containing plasmid by Cas4/1-2. **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 Exol 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 Exol 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 *GsuCas4/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 *GsuCas4/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.