








Structure reveals why genome folding is necessary for site-specific integration of foreign DNA into CRISPR arrays

Received: 22 March 2023

Accepted: 15 August 2023

Published online: 14 September 2023

 Check for updates

Andrew Santiago-Frangos¹ , William S. Henriques¹, Tanner Wiegand¹ , Colin C. Gauvin^{2,3} , Murat Buyukyoruk¹, Ava B. Graham¹, Royce A. Wilkinson¹, Lenny Triem¹, Kasahun Neselu⁴, Edward T. Eng⁴ , Gabriel C. Lander⁵  & Blake Wiedenheft¹  

Bacteria and archaea acquire resistance to viruses and plasmids by integrating fragments of foreign DNA into the first repeat of a CRISPR array. However, the mechanism of site-specific integration remains poorly understood. Here, we determine a 560-kDa integration complex structure that explains how *Pseudomonas aeruginosa* Cas (Cas1–Cas2/3) and non-Cas proteins (for example, integration host factor) fold 150 base pairs of host DNA into a U-shaped bend and a loop that protrude from Cas1–2/3 at right angles. The U-shaped bend traps foreign DNA on one face of the Cas1–2/3 integrase, while the loop places the first CRISPR repeat in the Cas1 active site. Both Cas3 proteins rotate 100 degrees to expose DNA-binding sites on either side of the Cas2 homodimer, which each bind an inverted repeat motif in the leader. Leader sequence motifs direct Cas1–2/3-mediated integration to diverse repeat sequences that have a 5′-GT. Collectively, this work reveals new DNA-binding surfaces on Cas2 that are critical for DNA folding and site-specific delivery of foreign DNA.

Vertebrates, bacteria and archaea have domesticated transposases (for example, RAG1 and Cas1) for adaptive immunity^{1,2}. Transposases, integrases and recombinases often co-opt additional DNA-bending proteins (for example, IHF, HU, H-NS or HMGB1) that facilitate DNA integration and excision^{3–7}. However, the structural role of DNA folding during this mobilization of DNA remains largely enigmatic.

CRISPRs are essential components of an adaptive immune system that stores DNA-based molecular memories of past infections⁸. CRISPR-associated proteins, Cas1 and Cas2, integrate fragments of foreign DNA ('spacers') into CRISPRs. Integration duplicates a repeat sequence, which thereby maintains the characteristic repeat–spacer–repeat architecture (Fig. 1a). Cas1 and Cas2 form a heterohexameric complex that consists of two Cas1 homodimers (Cas1a–a* and

Cas1b–b*) flanking a Cas2 homodimer (Fig. 1a,b)^{9–11}. Foreign DNA fragments bind across one face of the Cas2 homodimer, which positions the 3′ ends into Cas1 active sites on either end of the complex (that is, Cas1a* and Cas1b*)^{8–10,12}. The CRISPR repeat sequence wraps around the opposing face of Cas2, sandwiching the Cas2 homodimer between the foreign and repeat DNA duplexes. Opposing Cas1 subunits (Cas1a* and Cas1b*) catalyze two successive strand-transfer reactions, linking the 3′ ends of the foreign DNA to opposite ends of the repeat^{5,11}. CRISPR integration complexes sense a 2–5 base-pair (bp) 3′ overhang, called a protospacer-adjacent motif (PAM), in the foreign DNA to determine the integration orientation⁹. Correct spacer orientation is necessary to produce a functional CRISPR RNA that guides the CRISPR interference machinery (that is, Cascade) to

¹Department of Microbiology and Cell Biology, Montana State University, Bozeman, MT, USA. ²Department of Chemistry and Biochemistry, Montana State University, Bozeman, MT, USA. ³Thermal Biology Institute, Montana State University, Bozeman, MT, USA. ⁴Simons Electron Microscopy Center, National Resource for Automated Molecular Microscopy, New York Structural Biology Center, New York, NY, USA. ⁵Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, USA. ✉e-mail: bwiedenheft@gmail.com

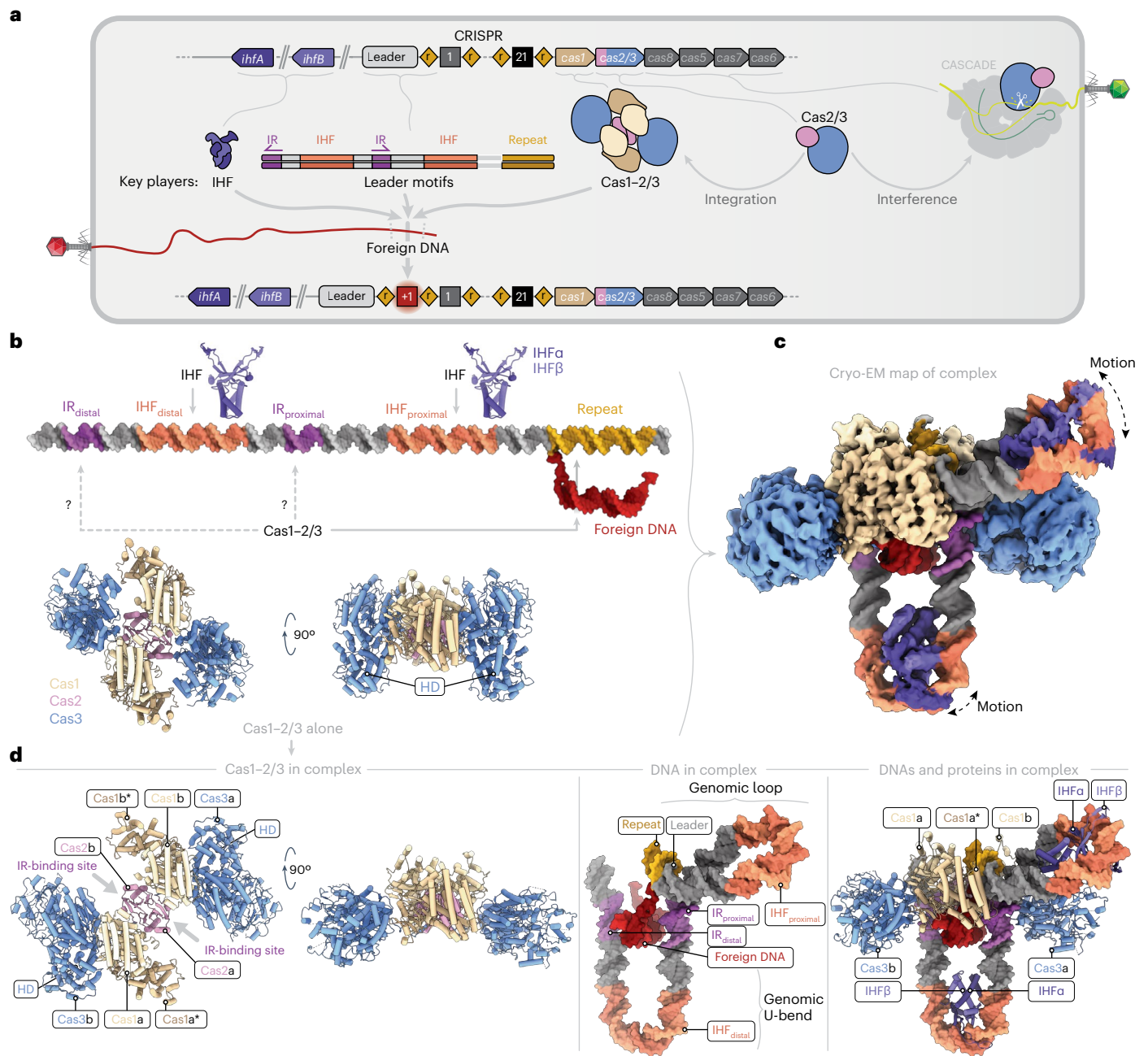


Fig. 1 | Cryo-EM structure of the type I-F CRISPR integration complex.

a, Scheme of the I-F CRISPR system of *P. aeruginosa* PA14. A CRISPR is composed of repeated DNA sequences (diamonds) interspersed with unique spacer sequences (black squares). The CRISPR is adjacent to six *cas* genes (arrows). Four Cas1 and two Cas2/3 proteins assemble into a heterohexamer in which the Cas3 and Cas1 subunits surround the central Cas2 homodimer like petals of a closed flower (Cas1₄–Cas2/3₂). Cas1–2/3 and IHF proteins cooperate with DNA upstream of the CRISPR (leader) to integrate foreign DNA at the first repeat. The leader sequence contains two IHF-binding sites and two IRs that are necessary for integration of foreign DNA at the leader-repeat junction. In addition to playing a central role in integration, the Cas2/3 fusion is recruited to DNA-bound Cascade CRISPR

surveillance complex to degrade foreign genetic parasites. **b**, The Cas1–2/3 heterohexamer and IHF proteins were mixed with a half-site DNA integration intermediate consisting of a foreign DNA linked to one strand of the CRISPR DNA at the leader-repeat junction. **c**, Cryo-EM density map of the type I-F CRISPR integration complex at -3.5 \AA resolution (Extended Data Fig. 1g–k and Table 1). **d**, Atomic model of the type I-F CRISPR integration complex. Cas1–2/3 proteins alone (left) are shown in cartoon representation. Cas3 domains rotate by 100° simulating the motion of a bloomed flower and exposing DNA-binding sites on Cas2 that interact with each of the IRs (Extended Data Fig. 2a and Supplementary Video 1). DNA alone is shown in the middle (surface representation). Proteins (cartoon representations) and DNAs of the integration complex are shown on the right.

complementary targets^{8,13}. Integration occurs in a stepwise manner. First, the non-PAM end of the foreign DNA is integrated at the leader side of the repeat¹⁴. Second, the PAM is cleaved by Cas or non-Cas nucleases and the trimmed 3' end is integrated at the spacer side of the repeat^{14–16}. These integration events tie a noncovalent knot around the Cas2 homodimer (foreign DNA on one side and repeat DNA on

the other), which is held together by complementary base pairing in the foreign DNA.

New foreign DNA is preferentially integrated at the first repeat in a CRISPR locus, ensuring efficient transcription and processing of CRISPR RNAs that target the most recently encountered genetic parasites (Fig. 1a)^{17,18}. Cas1–2 is thought to recognize a palindromic sequence

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

I-F integration complex (EMD-29280) (PDB 8FLJ)	
Data collection and processing	
Magnification	×46,860
Voltage (kV)	300
Electron exposure (e ⁻ /Å ²)	69.09
Defocus range (μm)	0.7–2.1
Pixel size (Å)	1.067
Symmetry imposed	C1
Initial particle images (no.)	5,846,923
Final particle images (no.)	366,794
Map resolution (Å)	3.47
FSC threshold	0.143
Map resolution range (Å)	2.3–9.9
Refinement	
Initial model used (PDB code)	1IHF; 3GOD
Model resolution (Å)	2.50; 2.17
FSC threshold	–
Model resolution range (Å)	–
Map sharpening B factor (Å ²)	–55
Model composition	
Non-hydrogen atoms	31,794
Protein residues	3,549
DNA nucleotides	350
Ligands	–
B factor (Å ²)	51.4
R.m.s deviations	
Bond lengths (Å)	0.005
Bond angles (°)	0.835
Validation	
MolProbity score	1.54
Clashscore	4.67
Poor rotamers (%)	0.26
Ramachandran plot	
Favored (%)	95.72
Allowed (%)	4.28
Disallowed (%)	0

within the CRISPR repeat, similar to target-site recognition by many DNA transposases^{5,11,19–21}. However, Cas1–2 recognition of the palindromic repeat does not explain how the first repeat is differentiated from downstream repeat sequences in a CRISPR. Thus, polarized integration often relies on additional proteins and DNA sequence motifs upstream of the CRISPR (that is, leader)^{3–5,11,18,22–27}. Integration host factor (IHF) facilitates polarized integration in the type I-E CRISPR system from *Escherichia coli*³. A structure of the I-E integration complex revealed that IHF bends the leader DNA to bring an upstream sequence motif into contact with Cas1, and IHF further stabilizes the Cas1–2 integrase at the first repeat through direct Cas1–IHF interactions^{3,5}.

Cas1 and Cas2 are conserved components of CRISPR-mediated immune systems. However, the type I-F CRISPR system has a unique

fusion of the Cas2 subunit to the Cas3 nuclease/helicase found in many type I systems (Fig. 1a,b)^{28,29}. Cas3 degrades Cascade-bound DNA into fragments with PAM-containing termini that are captured by Cas1–2 and integrated into the CRISPR locus in a process called ‘primed acquisition’^{4,30–37}. However, the structural mechanism for primed acquisition is unclear. Additionally, in contrast to the *E. coli* I-E system, which relies on one IHF to bend DNA and recruit a DNA motif found ~50 bp upstream of the CRISPR repeat, most I-F, I-C and some I-E CRISPR leaders contain multiple IHF-binding sites and multiple subtype-specific DNA motifs found up to 200 bp upstream of the CRISPR repeat (Fig. 1a and Extended Data Fig. 1a)²².

To determine how DNA sequence motifs in the CRISPR leader regulate the Cas1–2/3 integrase, we determined the structure of an ~560-kDa CRISPR integration complex from *P. aeruginosa*. The structure reveals that Cas1–2/3 cannot interact with the leader without first undergoing a large conformational change, which may be induced by foreign DNA binding. Further, the structure explains how the I-F leader and IHF proteins guide Cas1–2/3 to deliver and integrate foreign DNA at the first repeat in the CRISPR array (Fig. 1). Cas1–2/3 and IHF interact with all five DNA sequence motifs (that is, two inverted repeats (IRs), two IHF-binding sites and the CRISPR repeat) primarily through a shape-based readout^{38,39}. The shape of the folded I-F CRISPR leader is similar to that of the lambda-phage excision complex, suggesting that DNA is often used as a flexible scaffold to regulate DNA mobilization^{3–7,40–53}. The structure suggests that site-specific integration relies on protein-induced folding of the upstream DNA rather than sequence-specific recognition of the repeat. To test this idea, we perform a series of integration reactions demonstrating that efficient integration relies on conserved sequences in the leader and a 5'-GT dinucleotide in the repeat. We show that 5'-GT dinucleotides are broadly conserved in repeats derived from different CRISPR types, suggesting that they play a conserved role in integration across diverse CRISPR systems. In addition, the I-F CRISPR integration complex suggests a structural mechanism for interactions of the Cas1–2/3 integrase with the Cascade surveillance complex, which may be necessary for rapid adaptation to phage escape mutants^{4,30–37}.

Results

Cryo-EM structure of type I-F CRISPR integration complex

To understand how the Cas1–2/3 integrase cooperates with IHF and CRISPR leader motifs to integrate foreign DNA at the first CRISPR repeat, we purified the heterohexameric Cas1–2/3 integrase and the IHFα–β heterodimer, incubated these proteins with a DNA substrate representing a half-site integration intermediate and isolated the assembled complex using size-exclusion chromatography (SEC) (Extended Data Fig. 1a–f). The purified I-F integration complex was applied to cryo-electron microscopy (cryo-EM) grids and vitrified. We recorded 10,740 movies and picked 366,794 particles to determine an ~3.48-Å resolution structure of the integration complex. The reconstructed density was sufficient to model 90.7% of the 10 polypeptides and 88.4% of the 396 nucleotides of DNA (Fig. 1b–d, Extended Data Fig. 1g–k and Table 1). The model explains how the Cas1–2/3 subunits cooperate with two IHF heterodimers to kink and twist ~150 base pairs of host DNA into a structure that precisely positions foreign DNA for integration at the first repeat of the CRISPR (Fig. 1b–d).

The Cas1 and Cas2 subunits adopt a familiar quaternary arrangement that binds a foreign DNA on one face of the Cas2 homodimer and CRISPR repeat DNA on the other face (Figs. 1d, 2a and 3a)⁸. A previously determined structure of Cas1–2/3 alone revealed that the Cas3 and Cas1 domains surround the central Cas2 homodimer like petals of a closed flower (Cas1₄–Cas2/3₂) (Fig. 1b)⁵⁴. While this structure explained how Cas1 regulates the Cas3 nuclease, the role of Cas3 during integration remained unclear⁵⁴. Here, we show that the addition of DNA drives a series of conformational changes in both the DNA and proteins. The Cas3 domains rotate ~100° to align in a

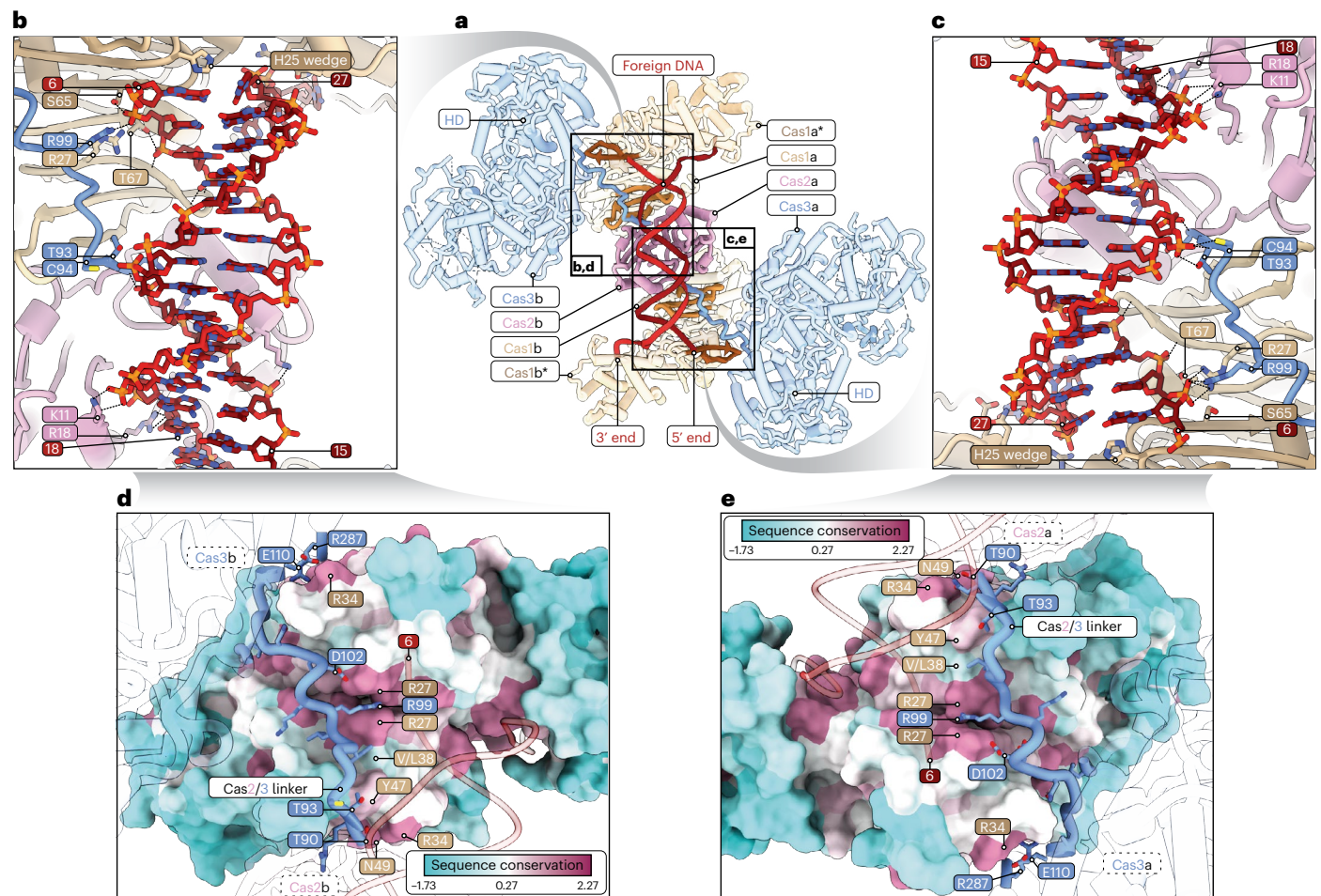


Fig. 2 | Foreign DNA constrains the Cas2/3 linker against conserved Cas1 residues. **a**, View of the foreign DNA-bound face of Cas1–2/3. The foreign DNA, Cas2 subunits, Cas2/3 linker and Cas1 beta hairpins that contact the start and end of the Cas2/3 linker are shown in solid while other parts of the complex are shown at 40% transparency for clarity. Insets outline locations of close-up views shown in panels **b–e**. **b,c**, The foreign DNA constrains the Cas2/3 linker against each Cas1 subunit (Extended Data Fig. 2b). Cas2, the Cas2/3 linker and Cas1 cooperate to bind the foreign DNA body and to splay the ends of the foreign DNA. Histidine wedges in Cas1 measure out a central foreign DNA duplex of 22 base pairs.

Most DNA-binding residues are conserved or undergo conservative mutations (Extended Data Fig. 3). Close-up view of Cas1a and Cas2a interface (**b**) and close-up view of Cas2b and Cas1b interface (**c**). **d,e**, Conserved Cas2/3 linker residues (blue, sticks) contact residues conserved in Cas1 proteins from type I-F CRISPR systems (mauve, surface) (Extended Data Fig. 2b,c). Cas2 and Cas3 domains are shown at 90% transparency for clarity. Close-up view of Cas1a and Cas1a* interface (**d**) and close-up view of Cas2b and Cas1b* interface (**e**). Inset shows the Cas1 sequence conservation color key.

planar configuration with Cas2, simulating the motion of a bloomed flower and exposing equivalent surfaces on opposite sides of the Cas2 homodimer that recognize an IR that is conserved in I-F leaders (Fig. 1d, Extended Data Fig. 2a and Supplementary Video 1)⁵⁴. Thus, the new planar conformation of Cas1–2/3 enables the simultaneous coordination of four DNA helices (IR_{distal}, IR_{proximal}, foreign DNA and CRISPR repeat) around the central Cas2 homodimer (Figs. 1d and 3). Further, this Cas3 rotation flips the nuclease domain from an interaction with Cas1 that suppresses the Cas3 nuclease activity to the opposite side of the complex, where the back of the Cas3 nuclease domain docks onto a groove created at the Cas1–Cas1 interface (Fig. 1b,d and Extended Data Fig. 2a,b).

The structure reveals two prominent DNA bends that protrude at right angles from Cas1–2/3 (Fig. 1c,d). An IHF heterodimer is wedged at the apex of each DNA bend, consistent with the well-defined role of IHF in DNA bending³⁸. These two DNA protrusions extend ~75 Å from the Cas1–2 core. Flexibility of these DNA extensions limits the resolution of the regions to 4–8 Å (Fig. 1c,d and Supplementary Video 2). IHF-mediated bending of the IHF_{distal} site positions the flanking IR sequences as symmetrical DNA pillars, which are recognized by

equivalent surfaces on opposite sides of the Cas2 homodimer (Fig. 3 and Extended Data Figs. 3b and 4c)²². Cas2 binding to these DNA pillars traps foreign DNA on one face of the Cas1–2/3 integrase. Further, Cas2 bends the IRs and steers downstream DNA away from Cas1–2/3, which would project the downstream CRISPR repeat away from the Cas1–2/3 integrase (Fig. 1d). However, Cas1–2/3 and IHF cooperate to constrict the DNA around the IHF_{proximal} site, forming a loop that places the CRISPR repeat into the Cas1a* active site (Figs. 1d and 3).

Foreign DNA constrains Cas2/3 linker against Cas1

The type I-F Cas2 and Cas3 subunits are connected by a 21 amino acid disordered linker (residues 90–110)^{4,28,29,55}. The structure explains how foreign DNA constrains the Cas2/3 linker against conserved surfaces of Cas1, which suggests that foreign DNA binding either initiates, or stabilizes, the Cas3 rotation (Fig. 2a)^{54,55}. The constrained Cas2/3 linker positions the HD nuclease domain of Cas3 (residues 111–374) against the Cas1–Cas1 interface, and facilitates Cas3 interactions with the IRs (Fig. 2a and Extended Data Figs. 2 and 4c). The foreign DNA and amino acids in the Cas2/3 linker contact conserved residues in type I-F Cas1 proteins (Fig. 2d,e and Extended Data Fig. 2c). Polar residues in the Cas2/3 linker may assist the binding

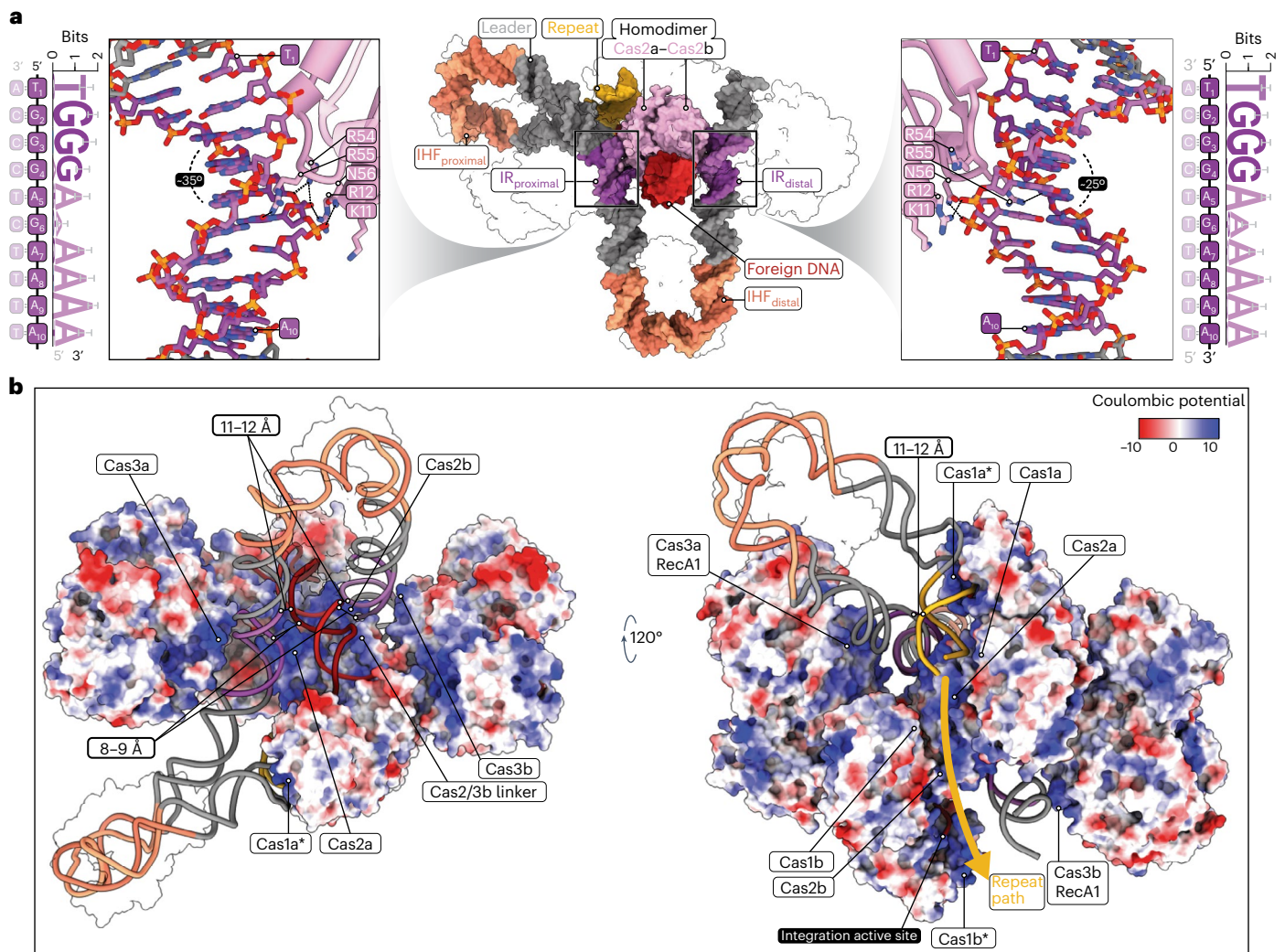


Fig. 3 | The Cas2 homodimer simultaneously coordinates four dsDNA helices critical to CRISPR integration. **a**, The Cas2 homodimer (pink surface) is flanked by DNA on four sides. Previous structures have shown that the CRISPR repeat (yellow) and foreign DNA (red) are bound to opposite faces of Cas2. Here, we show that symmetrical surfaces on Cas2 also bind IR (left and right) motifs in the leader. Surface representations of the IHF heterodimers, Cas1 homodimers, Cas2/3 linkers, Cas3 domains and the 3' overhang of the foreign DNA are shown in 100% transparency for clarity. Each Cas2 inserts an arginine (R55) into the center of the IRs, which stack between deoxyribose sugars, and additional polar residues (R54, N56, R12 and K11) contact the DNA backbone. Cas2 induces 25–35°

bends in the DNA (Extended Data Fig. 4c,e). The sequence logos of the type I-F IR_{proximal} (left) and IR_{distal} motifs (right), and the IR sequences present in the *P. aeruginosa* PA14 CRISPR leader, are shown. **b**, Views of the foreign DNA- (left) and repeat-bound (right) faces of Cas1–2/3 are shown in surface representation and colored by Coulombic potential. For clarity the highly electronegative DNA is shown in cartoon representation. Labels highlight highly basic and conserved surfaces of each Cas1–2/3 subunit that accommodate the packing of four dsDNA helices in proximity around the Cas2 homodimer (Extended Data Fig. 3). IHF heterodimers are shown in 100% transparency for clarity. The phosphate-to-phosphate distances of DNA helices packed around Cas2 are noted.

or splaying of the foreign DNA duplex at the conserved histidine wedge (H25) in Cas1 (Fig. 2b,c and Extended Data Fig. 3). Mutation of the histidine wedge (Cas1^{H25A}) decreases Cas1–2/3 integration activity of foreign DNA that has either fully complementary or splayed DNA ends (Extended Data Figs. 5 and 6a,b). The integration defect on substrates with splayed ends suggest that H25 is more than a simple wedge that pries apart the ends for foreign DNA¹⁰. The histidine steers the 3' ends down a positively charged channel that positions each 3'-hydroxyl into Cas1 active sites on opposite ends of the complex (Figs. 2 and 4a,d), whereas the 5' ends of the protospacer DNA are directed towards the back face of the Cas3 HD domain (Fig. 2)^{9,10,56}.

Cas2 homodimers recognize and bend inverted repeat sequences

The structure of the type I-F CRISPR integration complex reveals that Cas2 is the homodimer that binds the IRs (Fig. 3a). Mutations

that scramble the order of nucleotides in either the IR_{distal} or IR_{proximal} motifs limit Cas1–2/3-mediated integration²². While Cas2 does not make extensive sequence-specific contacts with nucleobases of the IR, a single residue (Cas2^{R55}) intercalates in the minor groove, and may participate in recognizing two conserved bases in the 10-bp long motif (Fig. 3a and Extended Data Fig. 4c,e). However, there is insufficient density for the R55 side chain to confidently assign contacts. Other conserved Cas2 residues (that is, K11, R12 and N56) form additional hydrogen bonds with the phosphate backbone of one DNA strand in each IR (Fig. 3a and Extended Data Fig. 4c). Mutation of these Cas2 residues (Cas2^{K11D,R12E}, Cas2^{R55E,N56D}, Cas2^{K11D,R12E,R55E,N56D}) prevents Cas1–2/3-mediated DNA integration (Extended Data Figs. 5 and 6c,d). Cas2 acts as a wedge that induces a 25–35° bend in the DNA upstream of IR_{distal} and downstream of IR_{proximal} (Fig. 3a). These flared IRs lean against basic residues (K381, R393, K397) on the back surface of Cas3 (Fig. 3 and Extended Data Figs. 3 and 4c). In sum, these observations reveal that

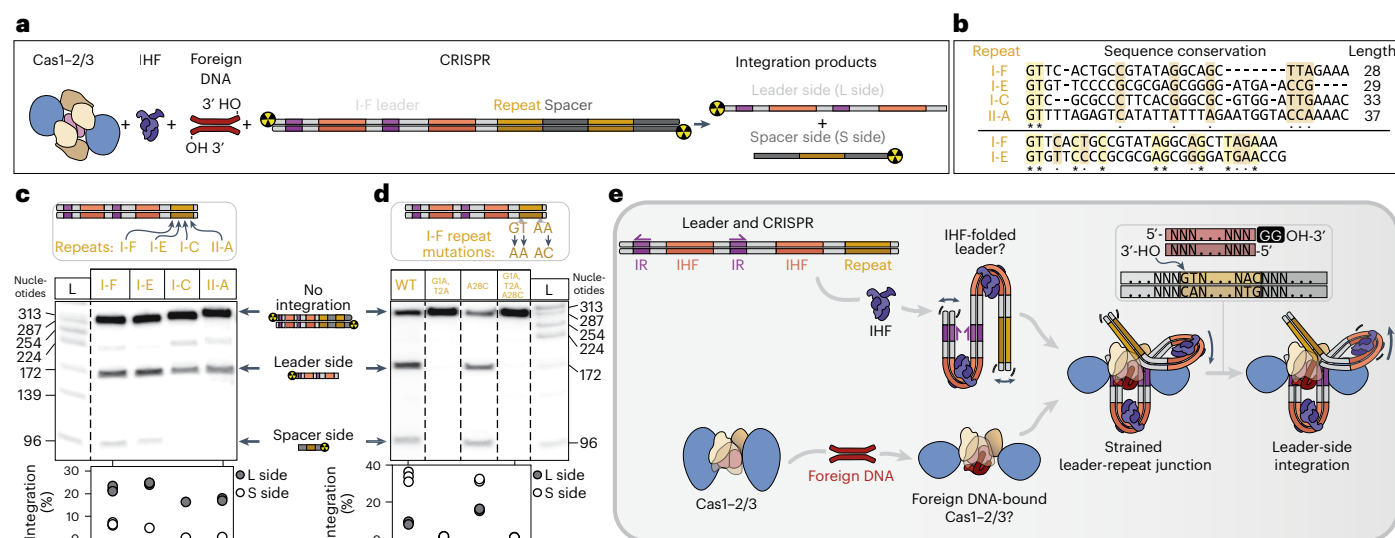


Fig. 4 | Sequence motifs in the leader and IHF proteins facilitate Cas1-2/3-based integration into diverse repeat sequences. **a**, Schematic of reactants and products of in vitro CRISPR integration assays (Extended Data Figs. 6, 7 and 9). **b**, Four CRISPR repeats used in the integration assays. A gapped sequence alignment highlights two identical (asterisks) and six similar (dots) positions. An ungapped sequence alignment reveals nine identical nucleotide positions between the I-F and I-E repeats. All four repeats have different internal palindromes and GC content (Extended Data Fig. 8b). **c**, Endpoint integration reactions with CRISPR repeat-swapped mutants, resolved on denaturing polyacrylamide gels. One of three representative gel images is shown (Extended Data Fig. 7). Quantification of leader- (gray circles) or spacer-side (white circles) integration events from all three replicate gels (Extended Data Fig. 7). The reactions were performed in triplicate, each dot represents one reaction, and some dots overlap. **d**, Four-minute time point of time-course integration

reactions with I-F repeat mutants, resolved on denaturing polyacrylamide gels. One of three representative images is shown (Extended Data Fig. 9). Quantification of leader- (gray circles) or spacer-side (white circles) integration events from all three replicate gels (Extended Data Fig. 9). **e**, CRISPR integration model. IHF-mediated folding of the genome presents IRs as symmetric DNA pillars that recruit foreign DNA-bound Cas1-2/3. Cas3 domains of Cas1-2/3 must rotate away from Cas2 to expose IR-binding sites on Cas2. Cas1-2/3 and IHF cooperate to fold DNA into a loop, docking the leader-repeat junction at the Cas1 active site. Foreign DNA integration at the leader-repeat junction nicks the DNA duplex, releasing tension in the DNA duplex and inhibiting the reverse disintegration reaction (Extended Data Fig. 4)^{61,62}. 5'-GT dinucleotides are required for efficient leader- and spacer-side integration, but no strict sequence requirements are necessary in the rest of the repeat.

the IR DNA sequences are primarily recognized by Cas1-2/3 through shape readout rather than base readout³⁹.

Cas2 homodimer is surrounded by four DNA helices

Cas2 is a cube-shaped homodimer at the center of the Cas integrase. The Cas2 cube is flanked by Cas1 homodimers to form an elongated DNA-binding platform that interacts with the CRISPR repeat on one face and the foreign DNA on the other (Fig. 3). Unique to the type I-F Cas1-2/3 integration complex, the IRs occupy the last two accessible surfaces of the Cas2 cube (Fig. 3a). Positively charged surfaces on Cas1-2/3 bind and shield negatively charged DNA, which enables the packing of four DNA helices around the small Cas2 homodimer (Fig. 3b and Extended Data Fig. 3). The foreign DNA-binding face of Cas2 has two electronegative pillars of leader DNA that straddle the foreign DNA, such that major grooves of the leader DNA pillars are clamped against major grooves of the foreign DNA. The two DNA pillars continue past Cas2 to flank the Cas1 active sites (Fig. 3b). At the IHF_{proximal} loop, Cas3 packs the leader against the Cas1-bound repeat, decreasing the phosphate-to-phosphate distances between these helices to ~11–12 Å. Although the latter two-thirds of the CRISPR repeat could not be resolved, the trajectory of the repeat suggests it will follow a path that threads between the distal leader DNA duplex and the 3'-hydroxyl of the foreign DNA that rests in the Cas1b* active site (Fig. 3b). Collectively, Cas1, the Cas2/3 linker and Cas3 accommodate four DNA helices (IR_{distal}, IR_{proximal}, foreign DNA and repeat) around the central Cas2 homodimer to facilitate site-specific integration.

IHF and the leader direct integration into diverse sequences

The structure suggests that Cas1-2/3 is guided to the first repeat of the CRISPR by IHF-mediated folding of the I-F leader, rather than direct

recognition of the repeat sequence (Fig. 1d and Extended Data Fig. 4b,d). To determine whether or how the repeat sequence impacts integration, we measured the efficiency of Cas1-2/3-catalyzed integration into DNAs containing either a I-F, I-E, I-C or II-A repeat downstream of a I-F leader (Fig. 4a–c). The type I-F leader supports Cas1-2/3-catalyzed leader-side integration at repeats derived from I-E, I-C and II-A CRISPR loci (Fig. 4a–c and Extended Data Figs. 7 and 8). Integration efficiency at non-native repeats is not correlated with sequence similarity to the I-F repeat or with GC content (Extended Data Fig. 8b). Instead, integration efficiency is correlated to the length of the repeat. I-F and I-E repeats are similar in length (28 and 29 bp, respectively), whereas the I-C and II-A repeats are 0.5 to 1 full DNA turns longer than the I-F repeat (33 and 37 bp, respectively) (Fig. 4b and Extended Data Fig. 9a,b). While leader-side integration is robust with different repeats, spacer-side integration is ~3.5-fold slower for the I-E repeat, and undetectable for the longer repeats (Extended Data Figs. 7 and 9a,b).

As expected, Cas1-2/3 does not catalyze integration at a I-F repeat downstream of a scrambled I-F leader, nor does Cas1-2/3 catalyze integration at I-E, I-C or II-A repeats downstream of their respective leaders²² (Extended Data Fig. 7). Cas1-2 sequences are diverse, such that leader-interacting residues are only conserved in a subset of proteins within a given CRISPR subtype²². For example, the I-F Cas1 protein lacks residues required to interact with the I-E leader²². Further, the I-E, I-C and I-F leaders have distinct nucleotide spacings between the leader motifs and the repeat. These nucleotide spacings impart a unique shape to the IHF-folded leader DNA, which is critical for integration.

These experiments were performed with foreign DNA substrates, either with or without a PAM (Extended Data Fig. 7). Cas1-2/3 integration of PAM-containing DNA is more specific, but the conclusions are otherwise consistent between the two substrates. The PAM must

be trimmed by an ancillary nuclease before Cas1–2/3 can catalyze spacer-side integration, therefore we focused our discussion on results from the trimmed foreign DNA to compare differences in spacer-side integration (Extended Data Fig. 7a–d)^{14–16}. Collectively, these integration experiments indicate that leader sequences and host factors dictate site-specific integration of foreign DNA at diverse DNA target sites.

5′-GT is critical for Cas1-mediated integration

Repeat sequences are strongly conserved within CRISPR subtypes, but vary in sequence and length between subtypes^{57,58}. However, in the small subset of repeats tested above, we noticed that the 5′-GT is conserved. To determine whether conservation of the 5′-GT is a coincidence or a more widely conserved feature of repeats, we performed a bioinformatic analysis consisting of 24,940 CRISPRs. This bioinformatic analysis reveals that a 5′-GT dinucleotide is broadly conserved at the leader side of the repeat, and conserved in some CRISPR systems at the spacer side of the repeat (Extended Data Fig. 4g). Therefore, we hypothesized that the 5′-GT dinucleotide is a base-specific determinant for leader-side integration. To test this hypothesis, we mutated the 5′-GT (G1A, T2A) and repeated the integration assays. The 5′-GT to AA mutation ablates both leader- and spacer-side integration, indicating that the 5′-GT is essential and that leader-side integration is a prerequisite for spacer-side integration (Fig. 4d). Since Cas1 requires a 5′-GT at the leader side of the repeat, we hypothesized that introducing a 5′-GT at the spacer side of the repeat would increase spacer-side integration efficiency. To test this hypothesis, we replaced adenosine 28 of the I-F repeat with cytosine (A28C) and repeated the integration assays. The A28C mutation increases the rate and amount of spacer-side integration approximately twofold, relative to the wild type (WT) I-F repeat (Fig. 4d). We do not detect integration into a I-F repeat that lacks a 5′-GT at the leader side, even if the repeat contains a 5′-GT at the spacer side. This result further supports our conclusion that leader-side integration is a prerequisite for spacer-side integration (Fig. 4d). We examined the structure of the Cas1 active site to determine whether the 5′-G is directly recognized by protein contacts. The Cas1 residue E184 is within 4 Å of the 5′-G (Extended Data Fig. 4b,d). However, a Cas1^{E184A} mutation destabilizes the complex, decreasing the amount of Cas1 subunits per Cas1–2/3 complex (Extended Data Fig. 5). Therefore, the decrease in integration activity of the Cas1^{E184A}–2/3 complex cannot be solely attributed to a decrease in recognition of the repeat (Extended Data Fig. 9d,f). Collectively, these data reveal that the 5′-GT is a conserved feature necessary for integration in most CRISPR systems, although no available structure provides a mechanism for direct recognition of the 5′-GT of the repeat^{5,11,15,59}.

Discussion

Here we demonstrate that Cas1–2/3 and IHF fold DNA into a structure that is necessary for site-specific integration of foreign DNA into CRISPRs. IHF proteins are highly expressed and most IHF-binding sites are thought to be occupied *in vivo*⁶⁰. Therefore, IHF may prefold the CRISPR leader into a ‘landing pad’ that recruits foreign DNA-bound Cas1–2/3 (Fig. 4e)^{22,38}. The Cas3 and Cas1 domains of the Cas1–2/3 complex are arranged like petals of a closed flower around the central Cas2 homodimer, such that the Cas3 domains occlude two of the four DNA-binding surfaces on Cas2, which precludes interactions with the leader (Fig. 1b,d). Foreign DNA binding to Cas1–2/3 physically constrains the Cas2/3 linker against the Cas1 homodimer, pulling the Cas3 HD domain against the Cas1–Cas1 interface (Fig. 2, Extended Data Fig. 2 and Supplementary Video 1). The 100° rotation of each Cas3 simulates the motion of a bloomed flower and exposes DNA-binding sites on Cas2 that interact with each of the IR motifs in the leader (Figs. 1 and 3 and Extended Data Figs. 2 and 3). Cas1–2/3 and IHF proteins fold DNA around the leader-repeat junction into a 260° loop that docks the first CRISPR repeat into the Cas1 active site under tension (Figs. 1 and 4 and Extended Data Figs. 3 and 4). The structure suggests that Cas1-mediated

strand transfer releases tension in this DNA loop, which may prevent disintegration of an otherwise isoenergetic strand-transfer reaction, and thereby favor complete integration (Extended Data Fig. 4f). A similar mechanism has been proposed to favor complete integration in other systems, where both strand-transfer events occur simultaneously^{61,62}.

We show that Cas1–2/3, IHF and the I-F leader facilitate leader-side integration at four different repeat sequences (Fig. 4b,c). These repeats are diverse in sequence identity, length, palindrome and GC content, but they share a 5′-GT (Fig. 4b and Extended Data Fig. 8b). To determine whether the 5′-GT is a universal feature of CRISPR repeats we analyzed 24,940 CRISPRs. This analysis reveals that CRISPR repeats contain a strongly conserved 5′-GT at the leader end, and that a 5′-GT is also conserved at the spacer end of repeats from several CRISPR systems (Extended Data Fig. 4g). We demonstrate that the 5′-GT is critical for leader-side integration and that introducing a 5′-GT at the spacer end of the repeat increases spacer-side integration. The broad conservation of 5′-GT is consistent with previous reports that type I-A, II-A and I-E systems require a 5′-G for integration^{11,63}. Similarly, the putative evolutionary ancestor to Cas1 enzymes, casposase, requires a conserved 5′ dinucleotide in target-site DNA for integration⁶⁴. Collectively, these data suggest that Cas1 proteins retain a shared sequence preference for a 5′-G or a 5′-GT, and lack strict sequence requirements for the central body of the repeat (Fig. 4b,c)⁶⁵. The lack of strict sequence requirements may be advantageous because the CRISPR repeat is at the nexus of foreign DNA integration, processing of the transcribed CRISPR and loading the mature CRISPR RNA (crRNA) into the surveillance complexes (for example, Cascade, Cas9).

Genetic parasites commonly escape CRISPR-based immunity through point mutations⁶⁶. To counter escape mutants, many CRISPR–Cas systems use existing spacer sequences to enhance the acquisition of new spacers from the same foreign genetic element via ‘primed’ acquisition^{4,30–37}. In some examples of primed acquisition, the Cas3 nuclease/helicase degrades CRISPR-targeted DNAs into single-stranded (ss) DNA fragments enriched in PAM-containing termini⁶⁷. Cas1–2 has been proposed to anneal complementary ssDNA fragments and integrate these into CRISPRs^{14,30,31,67}. Single-molecule colocalization and bulk immunoprecipitation suggest that the type I-E Cas1–2 integrase is recruited to a Cas3–Cascade–target DNA complex to facilitate primed acquisition^{34,68}. The structure of the type I-F integration complex reveals conformational changes in Cas3 that may enable interactions with DNA-bound Cascade (Fig. 5a)^{36,69}. Cascade improves integration efficiency and fidelity *in vivo*^{69,70}, and the structure suggests a model for the formation of a primed acquisition complex (Cas1–2/3–Cascade–target DNA) that transfers new foreign DNA fragments to the integrase. Additional structures will be necessary to clarify the mechanism(s) of primed adaptation.

A comparison of the I-E and I-F integration complexes with a structure of the lambda-phage excision complex, reveals structural similarities and differences. In I-E systems, the IHF protein folds the leader DNA to present an upstream motif to a lobe of Cas1. In contrast, the I-F structure highlights extensive cooperation between IHF and Cas1–2/3 in bending the leader into an energetically strained conformation that may increase the specificity of CRISPR recognition (Fig. 6a,b). This cooperation includes Cas1–2/3 kinking the IR DNA motifs presented as parallel DNA pillars by IHF, and Cas1–2/3 constricting the IHF_{proximal} 180° bend into a 260° bend (Fig. 6b). Further, the sequestration of the IR-binding surface of Cas2 suggests a unique structural mechanism that prevents Cas1–2/3 interactions with the leader until foreign DNA binding induces a rotation of Cas3 (Fig. 4e and Supplementary Videos 1 and 3)⁵⁴.

Many CRISPR leaders contain multiple IHF-binding sites and subtype-specific motifs that are reminiscent of motifs found at the ends of transposons, phages and plasmids, which facilitate integration, excision or recombination complex assembly (Extended Data Fig. 8a)^{3–7,22,40–53}. For example, the left and right ends of the lambda-phage genome

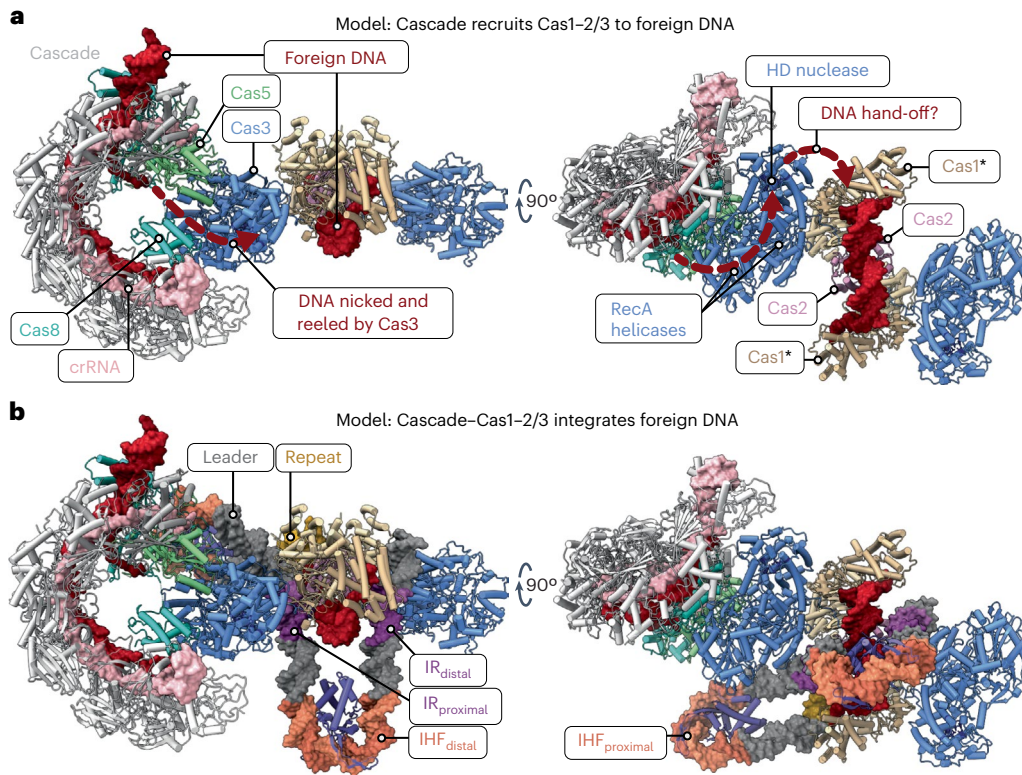


Fig. 5 | I-F CRISPR integration complex suggests a mechanism for primed acquisition and Cascade impact on integration. **a**, Cascade (gray) bound to foreign DNA (red) displays a Cas8 helical bundle (turquoise) that recruits Cas2/3. Cascade-recruited Cas3 degrades foreign DNA into fragments that are captured by the Cas1-2 integrase for subsequent integration^{14,30,31,54,55,67,76-78}. The Cas1-2/3 integration complex can be docked onto double-stranded (ds) DNA-bound Cascade with minimal clashing and suggests a model for the formation of a

primed acquisition complex that facilitates rapid adaptation to genetic parasite variants^{44,68,76}. **b**, Recruitment of the Cascade-Cas1-2/3 complex to IHF-folded CRISPR leader DNA is not predicted to form any new clashes, suggesting a mechanism for the role of Cascade in facilitating integration *in vivo*^{69,70}. A total of two Cascade complexes can be docked onto the Cas1-2/3 integration complex (one per Cas3) without introducing additional clashing; a single Cascade is shown above for clarity.

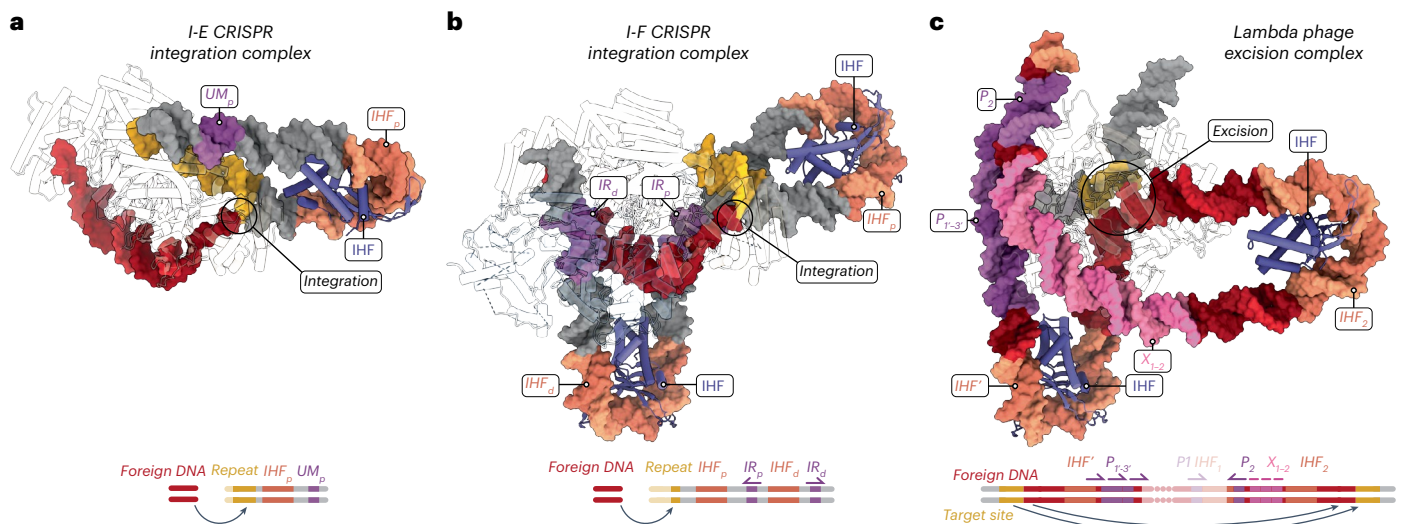


Fig. 6 | DNA is a flexible scaffold that controls DNA mobilization. **a–c**, Structures for the I-E CRISPR integration complex (**a**), I-F CRISPR integration complex (**b**) and lambda-phage excision complex (**c**). DNA shown as a surface, IHF

(purple) and all other proteins shown as transparent cartoons. Integration and excision sites, along with DNA motifs that regulate DNA mobilization, are labeled and colored according to the schematic (bottom).

contain three IHF-binding sites, five copies of a second DNA motif (P1-2,1'-3') and three copies of a third DNA motif (X11.5,2) that recruit lambda-phage (Int, Xis) and non-lambda-phage (IHF, Fis) proteins

(Fig. 6c)⁷. These proteins use DNA as a flexible scaffold that organizes enzyme active sites and DNA substrates to facilitate either the integration, or excision, of lambda-phage DNA from the bacterial genome

(Fig. 6c). Similarly, the I-F CRISPR leader motifs recruit Cas (Cas1–2/3) and non-Cas (IHF) proteins, which bend DNA into a flexible scaffold to organize enzyme active sites and DNA substrates in space, facilitating the integration of foreign DNA into the bacterial genome (Fig. 6b). Diverse systems thus use DNA as a flexible scaffold to regulate the isoenergetic mobilization of DNA (Fig. 6a–c).

Diverse DNA-mobilizing enzymes across the tree of life co-opt DNA folding to regulate DNA mobilization^{3–7}. In sum, these data provide a mechanistic understanding for the role of DNA as a flexible scaffold that controls DNA mobilization. These insights are critical to developing applications of DNA-mobilizing enzymes in gene therapy, genetic engineering and chronological DNA recordings^{6,53,71–75}.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41594-023-01097-2>.

References

- Koonin, E. V. & Krupovic, M. Evolution of adaptive immunity from transposable elements combined with innate immune systems. *Nat. Rev. Genet.* **16**, 184–192 (2015).
- McClintock, B. The origin and behavior of mutable loci in maize. *Proc. Natl Acad. Sci. USA* **36**, 344–355 (1950).
- Nuñez, J. K., Bai, L., Harrington, L. B., Hinder, T. L. & Doudna, J. A. CRISPR immunological memory requires a host factor for specificity. *Mol. Cell* **62**, 824–833 (2016).
- Fagerlund, R. D. et al. Spacer capture and integration by a type I-F Cas1–Cas2–3 CRISPR adaptation complex. *Proc. Natl Acad. Sci. USA* **114**, 201618421 (2017).
- Wright, A. V. et al. Structures of the CRISPR genome integration complex. *Science* **357**, 1113–1118 (2017).
- Obergfell, K. P. & Seifert, H. S. Mobile DNA in the pathogenic *Neisseria*. *Microbiol. Spectrum* **3**, <https://doi.org/10.1128/microbiolspec.mdna3-0015-2014> (2014).
- Laxmikanthan, G. et al. Structure of a Holliday junction complex reveals mechanisms governing a highly regulated DNA transaction. *eLife* **5**, e14313 (2016).
- Lee, H. & Sashital, D. G. Creating memories: molecular mechanisms of CRISPR adaptation. *Trends Biochem. Sci.* **47**, 464–476 (2022).
- Wang, J. et al. Structural and mechanistic basis of PAM-dependent spacer acquisition in CRISPR–Cas systems. *Cell* **163**, 840–853 (2015).
- Nuñez, J. K., Harrington, L. B., Kranzusch, P. J., Engelman, A. N. & Doudna, J. A. Foreign DNA capture during CRISPR–Cas adaptive immunity. *Nature* **527**, 535–538 (2015).
- Xiao, Y., Ng, S., Nam, K. H. & Ke, A. How type II CRISPR–Cas establish immunity through Cas1–Cas2-mediated spacer integration. *Nature* **550**, 137–141 (2017).
- Jackson, S. A. et al. CRISPR–Cas: adapting to change. *Science* **356**, eaal5056 (2017).
- Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733–740 (2009).
- Kim, S. et al. Selective loading and processing of prespacers for precise CRISPR adaptation. *Nature* **579**, 141–145 (2020).
- Hu, C. et al. Mechanism for Cas4-assisted directional spacer acquisition in CRISPR–Cas. *Nature* **598**, 515–520 (2021).
- Ramachandran, A., Summerville, L., Learn, B. A., DeBell, L. & Bailey, S. Processing and integration of functionally oriented prespacers in the *Escherichia coli* CRISPR system depends on bacterial host exonucleases. *J. Biol. Chem.* **295**, 3403–3414 (2020).
- Liao, C. et al. Spacer prioritization in CRISPR–Cas9 immunity is enabled by the leader RNA. *Nat. Microbiol.* **7**, 530–541 (2022).
- McGinn, J. & Marraffini, L. A. CRISPR–Cas systems optimize their immune response by specifying the site of spacer integration. *Mol. Cell* **64**, 616–623 (2016).
- Wang, R., Li, M., Gong, L., Hu, S. & Xiang, H. DNA motifs determining the accuracy of repeat duplication during CRISPR adaptation in *Haloarcula hispanica*. *Nucleic Acids Res.* **44**, 4266–4277 (2016).
- Goren, M. G. et al. Repeat size determination by two molecular rulers in the type I-E CRISPR array. *Cell Rep.* **16**, 2811–2818 (2016).
- Linheiro, R. S. & Bergman, C. M. Testing the palindromic target site model for DNA transposon insertion using the *Drosophila melanogaster* P-element. *Nucleic Acids Res.* **36**, 6199–6208 (2008).
- Santiago-Frangos, A., Buyukyoruk, M., Wiegand, T., Krishna, P. & Wiedenheft, B. Distribution and phasing of sequence motifs that facilitate CRISPR adaptation. *Curr. Biol.* **31**, 3515–3524 (2021).
- Kieper, S. N., Almendros, C. & Brouns, S. J. J. Conserved motifs in the CRISPR leader sequence control spacer acquisition levels in type I-D CRISPR–Cas systems. *FEMS Microbiol. Lett.* **366**, 2016–2020 (2019).
- Rollie, C., Graham, S., Rouillon, C. & White, M. F. Prespacer processing and specific integration in a type I-A CRISPR system. *Nucleic Acids Res.* **46**, 1007–1020 (2018).
- Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res.* **40**, 5569–5576 (2012).
- Wei, Y., Chesne, M. T., Terns, R. M. & Terns, M. P. Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in *Streptococcus thermophilus*. *Nucleic Acids Res.* **43**, 1749–1758 (2015).
- Wright, A. V. & Doudna, J. A. Protecting genome integrity during CRISPR immune adaptation. *Nat. Struct. Mol. Biol.* **23**, 876–883 (2016).
- Westra, E. R. et al. Parasite exposure drives selective evolution of constitutive versus inducible defense. *Curr. Biol.* **25**, 1043–1049 (2015).
- Makarova, K. S. et al. Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.* **18**, 67–83 (2020).
- Richter, C. et al. Priming in the type I-F CRISPR–Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. *Nucleic Acids Res.* **42**, 8516–8526 (2014).
- Datsenko, K. A. et al. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat. Commun.* **3**, 945 (2012).
- Xiao, Y. et al. Structure basis for RNA-guided DNA degradation by Cascade and Cas3. *Science* **361**, eaat0839 (2018).
- Nicholson, T. J. et al. Bioinformatic evidence of widespread priming in type I and II CRISPR–Cas systems. *RNA Biol.* **16**, 566–576 (2019).
- Dillard, K. E. et al. Assembly and translocation of a CRISPR–Cas primed acquisition complex. *Cell* **175**, 934–946.e15 (2018).
- Li, M., Wang, R., Zhao, D. & Xiang, H. Adaptation of the *Haloarcula hispanica* CRISPR–Cas system to a purified virus strictly requires a priming process. *Nucleic Acids Res.* **42**, 2483–2492 (2014).
- Semenova, E. et al. Highly efficient primed spacer acquisition from targets destroyed by the *Escherichia coli* type I-E CRISPR–

- Cas interfering complex. *Proc. Natl Acad. Sci. USA* **113**, 7626–7631 (2016).
37. Fineran, P. C. et al. Degenerate target sites mediate rapid primed CRISPR adaptation. *Proc. Natl Acad. Sci. USA* **111**, E1629–E1638 (2014).
38. Rice, P. A., Yang, S., Mizuuchi, K. & Nash, H. A. Crystal structure of an IHF–DNA complex: a protein-induced DNA U-turn. *Cell* **87**, 1295–1306 (1996).
39. Rohs, R. et al. Origins of specificity in protein–DNA recognition. *Annu. Rev. Biochem.* **79**, 233–269 (2010).
40. Zayed, H. The DNA-bending protein HMGB1 is a cellular cofactor of *Sleeping Beauty* transposition. *Nucleic Acids Res.* **31**, 2313–2322 (2003).
41. Little, A. J., Corbett, E., Ortega, F. & Schatz, D. G. Cooperative recruitment of HMGB1 during V(D)J recombination through interactions with RAG1 and DNA. *Nucleic Acids Res.* **41**, 3289–3301 (2013).
42. Nash, H. A. & Robertson, C. A. Purification and properties of the *Escherichia coli* protein factor required for lambda integrative recombination. *J. Biol. Chem.* **256**, 9246–9253 (1981).
43. Lavoie, B. D. & Chaconas, G. Site-specific HU binding in the Mu transpososome: conversion of a sequence-independent DNA-binding protein into a chemical nuclease. *Genes Dev.* **7**, 2510–2519 (1993).
44. Chalmers, R., Guhathakurta, A., Benjamin, H. & Kleckner, N. IHF modulation of Tn10 transposition: sensory transduction of supercoiling status via a proposed protein/DNA molecular spring. *Cell* **93**, 897–908 (1998).
45. Haniford, D. B. Transpososome dynamics and regulation in Tn10 transposition. *Crit. Rev. Biochem. Mol. Biol.* **41**, 407–424 (2006).
46. Whitfield, C. R., Wardle, S. J. & Haniford, D. B. The global bacterial regulator H-NS promotes transpososome formation and transposition in the Tn5 system. *Nucleic Acids Res.* **37**, 309–321 (2009).
47. Liu, D., Haniford, D. B. & Chalmers, R. M. H-NS mediates the dissociation of a refractory protein–DNA complex during Tn10/IS10 transposition. *Nucleic Acids Res.* **39**, 6660–6668 (2011).
48. van Gent, D. C., Hiom, K., Paull, T. T. & Gellert, M. Stimulation of V(D)J cleavage by high mobility group proteins. *EMBO J.* **16**, 2665–2670 (1997).
49. Rowland, S.-J., Stark, W. M. & Boocock, M. R. Sin recombinase from *Staphylococcus aureus*: synaptic complex architecture and transposon targeting. *Mol. Microbiol.* **44**, 607–619 (2002).
50. Alonso, J. C., Weise, F. & Rojo, F. The *Bacillus subtilis* histone-like protein Hbsu is required for DNA resolution and DNA inversion mediated by the β recombinase of plasmid pSM19135. *J. Biol. Chem.* **270**, 2938–2945 (1995).
51. Petit, M.-A., Ehrlich, D. & Janni re, L. pAM β 1 resolvase has an atypical recombination site and requires a histone-like protein HU. *Mol. Microbiol.* **18**, 271–282 (1995).
52. Rojo, F. & Alonso, J. C. The β recombinase of plasmid pSM19035 binds to two adjacent sites, making different contacts at each of them. *Nucleic Acids Res.* **23**, 3181–3188 (1995).
53. Walker, M. W. G., Klompe, S. E., Zhang, D. J. & Sternberg, S. H. Transposon mutagenesis libraries reveal novel molecular requirements during CRISPR RNA-guided DNA integration. Preprint at *bioRxiv* <https://doi.org/10.1101/2023.01.19.524723> (2023).
54. Rollins, M. F. et al. Cas1 and the Csy complex are opposing regulators of Cas2/3 nuclease activity. *Proc. Natl Acad. Sci. USA* **114**, 201616395 (2017).
55. Wang, X. et al. Structural basis of Cas3 inhibition by the bacteriophage protein AcrF3. *Nat. Struct. Mol. Biol.* **23**, 868–870 (2016).
56. Wiedenheft, B. et al. Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. *Structure* **17**, 904–912 (2009).
57. Kunin, V., Sorek, R. & Hugenholtz, P. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol.* **8**, R61 (2007).
58. Nethery, M. A. et al. CRISPRclassify: repeat-based classification of CRISPR loci. *CRISPR J.* **4**, 558–574 (2021).
59. Dhingra, Y., Suresh, S. K., Juneja, P. & Sashital, D. G. PAM binding ensures orientational integration during Cas4-Cas1-Cas2-mediated CRISPR adaptation. *Mol. Cell* **82**, 4353–4367.e6 (2022).
60. Ali Azam, T., Iwata, A., Nishimura, A., Ueda, S. & Ishihama, A. Growth phase-dependent variation in protein composition of the *Escherichia coli* nucleoid. *J. Bacteriol.* **181**, 6361–6370 (1999).
61. Monta o, S. P., Pigli, Y. Z. & Rice, P. A. The Mu transpososome structure sheds light on DDE recombinase evolution. *Nature* **491**, 413–417 (2012).
62. Maertens, G. N., Hare, S. & Cherepanov, P. The mechanism of retroviral integration from X-ray structures of its key intermediates. *Nature* **468**, 326–329 (2010).
63. Rollie, C., Schneider, S., Brinkmann, A. S., Bolt, E. L. & White, M. F. Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition. *eLife* **4**, e08716 (2015).
64. B guin, P., Chekli, Y., Sezonov, G., Forterre, P. & Krupovic, M. Sequence motifs recognized by the casposon integrase of *Aciduliprofundum boonei*. *Nucleic Acids Res.* **47**, 6386–6395 (2019).
65. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Classification and nomenclature of CRISPR-Cas systems: where from here? *CRISPR J.* **1**, 325–336 (2018).
66. Deveau, H. et al. Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J. Bacteriol.* **190**, 1390–1400 (2008).
67. K nne, T. et al. Cas3-derived target DNA degradation fragments fuel primed CRISPR adaptation. *Mol. Cell* **63**, 852–864 (2016).
68. Musharova, O. et al. Pespacers formed during primed adaptation associate with the Cas1–Cas2 adaptation complex and the Cas3 interference nuclease–helicase. *Proc. Natl Acad. Sci. USA* **118**, e2021291118 (2021).
69. Wiegand, T. et al. Reproducible antigen recognition by the type I-F CRISPR–Cas System. *CRISPR J.* **3**, 378–387 (2020).
70. Vorontsova, D. et al. Foreign DNA acquisition by the I-F CRISPR–Cas system requires all components of the interference machinery. *Nucleic Acids Res.* **43**, 10848–10860 (2015).
71. Cavazzana-Calvo, M. et al. Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. *Science* **288**, 669–672 (2000).
72. Strecker, J. et al. RNA-guided DNA insertion with CRISPR-associated transposases. *Science* **365**, 48–53 (2019).
73. Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S. & Sternberg, S. H. Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration. *Nature* **571**, 219–225 (2019).
74. Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. Molecular recordings by directed CRISPR spacer acquisition. *Science* **353**, aaf1175 (2016).
75. Schmidt, F., Cherepkova, M. Y. & Platt, R. J. Transcriptional recording by CRISPR spacer acquisition from RNA. *Nature* **562**, 380–385 (2018).
76. Rollins, M. F. et al. Structure reveals a mechanism of CRISPR-RNA-guided nuclease recruitment and anti-CRISPR viral mimicry. *Mol. Cell* **74**, 132–142.e5 (2019).

77. Sinkunas, T. et al. Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *EMBO J.* **30**, 1335–1342 (2011).
78. Huo, Y. et al. Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. *Nat. Struct. Mol. Biol.* **21**, 771–777 (2014).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2023

Methods

Nucleic acid preparation

Four single-stranded DNAs (Supplementary Table 1) were synthesized (IDT) and resuspended in 1× TE buffer (10 mM Tris–HCl pH 8, 1 mM EDTA) before being used to assemble the structure of the type I-F integration complex, the assembly is detailed in a section below. The splayed foreign DNAs used in integration assays (Supplementary Table 1) were synthesized and resuspended in 1× TE buffer before use. To make ³²P-labeled CRISPR integration substrates, the sequences consisting of the leader and CRISPR arrays were first synthesized and cloned into pUC57 (Genscript). These plasmids have been made available on Addgene (Supplementary Table 1). These plasmids were transformed into chemically competent *E. coli* DH5α cells and the transformed cells were plated onto LB agar plates containing 100 μg ml⁻¹ ampicillin. These cells were cultured in LB medium and plasmids were purified using ZymoPURE II Plasmid Midiprep kit (Zymo Research). Each plasmid was then digested with EcoRI-HF and BamHI-HF (NEB) restriction enzymes, and the 294–383 bp inserts of interest were separated from the vector backbone by agarose gel electrophoresis. The gel segments containing the DNA inserts of interest were excised and DNA was purified using a Zymoclean Gel DNA Recovery kit (Zymo Research) (Supplementary Table 1). The 5′ ends of the CRISPR leader and array fragments were dephosphorylated using Quick calf intestinal alkaline phosphatase (NEB), and the DNAs were purified away from protein using a DNA Clean and Concentrator kit (Zymo Research). Both 5′ ends of 1 pmol of the CRISPR leader and array fragments were then labeled with ³²P, by incubation with 4 pmol of [γ -³²P]ATP (PerkinElmer) by polynucleotide kinase (NEB) in 1× PNK buffer at 37 °C for 45 minutes. PNK was heat denatured by incubation at 65 °C for 20 minutes. Spin column purification (G-25, GE Healthcare) was used to remove unincorporated radioactive nucleotides and to buffer exchange DNAs into 1× TE buffer.

Cas1 and Cas2/3 mutagenesis

The plasmid used to express Cas1 and Cas2/3 (Addgene, plasmid no. 89240) was PCR amplified with mutagenic primer pairs using Q5 polymerase (NEB) (Supplementary Table 1). The parental template plasmid was digested with DpnI (NEB) and the PCR-amplified products were purified using the DNA Clean and Concentrate kit (Zymo). The purified DNA was 5′ phosphorylated with T4 polynucleotide kinase (NEB) and ligated with home-made T4 DNA ligase. The ligation reactions were transformed into DH5α cells. The Cas2 K11D, R12E, R55E, N56D mutant was prepared using the Cas2 K11D, R12E mutant as the template with the PCR primers for Cas2 R55E, N56D in the mutagenic PCR reaction. Plasmids for expressing the mutant Cas1–2/3 complexes, Cas1H25A–2/3, Cas1E184A–2/3, Cas1–2K11D, R12E/3, Cas1–2R55E, N56D/3 and Cas1–2K11D, R12E, R55E, N56D/3 have been deposited at Addgene (plasmid nos. 200213, 200214, 200216, 200217 and 200218).

Protein purification

P. aeruginosa IHF heterodimer was purified as previously described²². Briefly, 6× His-tagged IHFα and StrepII-tagged IHFβ were coexpressed in *E. coli* BL21(DE3). Cell pellets were lysed by sonication in IHF lysis buffer (25 mM HEPES–NaOH pH 7.5, 500 mM NaCl, 10 mM imidazole, 1 mM TCEP, 5% glycerol) supplemented with 0.3× Halt Protease Inhibitor Cocktail (ThermoFisher) at 4 °C. Lysate was clarified by two rounds of centrifugation at 10,000g for 15 minutes at 4 °C. His-tagged IHF was captured on HisTrap HP resin (Cytiva) and eluted with 500 mM imidazole. Affinity tags were cleaved using PreScission protease, and the PreScission protease and remaining 6× His-IHFα were removed by affinity chromatography using HisTrap HP resin (Cytiva). Untagged IHF heterodimer was then further purified on Heparin Sepharose (Cytiva) and eluted with a linear gradient to a buffer containing 2 M NaCl. Fractions containing IHF heterodimer were concentrated and further purified by SEC on a Superdex 75 column (Cytiva) equilibrated in IHF buffer (25 mM HEPES–NaOH pH 7.5, 200 mM NaCl, 5% glycerol)

(Extended Data Fig. 1a). IHF overexpression plasmids have been previously described and deposited at Addgene (plasmid nos. 149384 and 149385)²².

P. aeruginosa Cas1–2/3 heterohexameric complexes (wild type and mutants) were purified as previously described²². Briefly, StrepII-tagged Cas1–Cas2/3 was overexpressed in *E. coli* BL21(DE3). Cell pellets were lysed via sonication in Cas1–2/3 lysis buffer (50 mM HEPES pH 7.5, 500 mM KCl, 10% glycerol, 1 mM DTT) supplemented with 0.3× Halt Protease Inhibitor Cocktail (ThermoFisher) at 4 °C. Lysate was clarified as above. StrepII-tagged Cas1–Cas2/3 complexes were affinity purified on StrepTrap HP resin (GE Healthcare) and eluted with Cas1–2/3 lysis buffer containing 3 mM desthiobiotin (Sigma-Aldrich). Eluate was concentrated at 4 °C (Corning Spin-X concentrators), before purification over a Superdex 200 size-exclusion column (Cytiva) equilibrated in 10 mM HEPES pH 7.5, 500 mM potassium glutamate and 10% glycerol (Extended Data Figs. 1c and 5e).

In vitro integration assays

Endpoint integration reactions were performed in triplicate using 300 nM of splayed foreign DNA fragments, containing or lacking a PAM (IDT), 200 nM of Cas1–2/3, 300 nM of IHF heterodimer and roughly 1 nM of a given ³²P-labeled CRISPR variant fragment in integration buffer (20 mM HEPES pH 7.5, 150 mM potassium glutamate, 5 mM MnCl₂, 1 mM TCEP, 1% glycerol) (Supplementary Table 1). Reactions were assembled on ice and then incubated at 35 °C for 20 minutes. Time-course integration reactions were performed in triplicate using 300 nM of splayed, or fully complementary, foreign DNA fragments lacking a PAM (IDT), 200 nM of Cas1–2/3, 300 nM of IHF heterodimer and roughly 1 nM of a given ³²P-labeled CRISPR variant fragment in modified integration buffer (20 mM HEPES pH 7.5, 150 mM potassium glutamate, 5 mM MnCl₂, 10 mM TCEP, 1% glycerol) (Supplementary Table 1). We noticed that IHF and Cas1–2/3 protein stocks exhibited a propensity to precipitate when diluted into prechilled buffer to form working dilution stocks. Therefore, all working dilutions of IHF and Cas1–2/3 prepared for time-course assays were made by first diluting the proteins into room-temperature buffer, mixed and then chilled on ice. Reactions were assembled on ice and then incubated at 20 minutes. Time points were taken at 0, 1, 2, 4 and 8 minutes. Reactions were stopped by the addition of phenol. The aqueous (nucleic acid containing) layer was mixed 1:1 with 2× formamide loading buffer (95% formamide, 20 mM EDTA, 0.05% bromophenol blue, 0.05% xylene cyanol) and then denatured at 95 °C for 5 minutes, before resolving the ³²P-labeled CRISPR substrates and integration products on a 7% (w/v) (29:1 mono:bis) polyacrylamide urea gel in 1× TBE (100 mM Tris–borate pH 8.3, 2 mM EDTA). Gels were dried and quantified using a Typhoon phosphorimager (GE Healthcare). The intensities of full-length CRISPR variant, leader-side integration fragments and spacer-side integration fragments were quantified with Multi Gauge v.3 (Fujifilm). These readings were then used to calculate leader- and spacer-side integration events as percentages of all events. Images of all gels that resolved integration reactions are shown Extended Data Figs. 6, 7 and 9), and additional control gels show that Cas1–2/3 is required for integration, and show how the custom ³²P-labeled ladder was generated by restriction enzyme digestion of ³²P-labeled CRISPR variant DNAs (Extended Data Fig. 8). Time-course integration data were fit to a plateau followed by one-phase association (GraphPad Prism v.10).

Assembly and purification of I-F integration complex

A total of four ssDNAs synthesized to mimic a half-site integration intermediate were annealed in a stepwise manner. Two nanomoles of ssDNAs, mostly corresponding to the sense and antisense strands of the CRISPR leader (‘strand_1’ and ‘strand_2’) were denatured at 100 °C and then slowly annealed using a PCR program that cooled the samples to 25 °C, in 5 °C steps for 5 minutes each, in 100 μl of hybridization buffer (20 mM Tris–HCl pH 7.5, 100 mM monopotassium glutamate, 5 mM EDTA,

1 mM TCEP). Two nanomoles of ssDNAs mostly corresponding to the sense and antisense of the strands of the foreign DNA ('strand_3' and 'strand_4') were slow annealed using the same protocol (Extended Data Fig. 1a and Supplementary Table 1). The two sets of annealed DNAs (tube 1: 'strand_1' and 'strand_2'; tube 2: 'strand_3' and 'strand_4') were mixed together, heated to 80 °C and then slow annealed using a PCR program that cooled the samples to 25 °C, in 5 °C steps for 5 minutes each, to anneal the complementary sense and antisense regions of the CRISPR repeat included in strand_2 and strand_3 together. Next, 6 nanomoles of IHF heterodimer in 50 µl of hybridization buffer was warmed to 25 °C and then mixed and incubated with the annealed DNAs at 25 °C for 10 minutes. Next, 3 nanomoles of Cas1–2/3 in 250 µl of hybridization buffer was warmed to 25 °C and mixed with the prepared DNA and IHF mixture, and incubated at 25 °C for 10 minutes. The total concentration of monopotassium glutamate in the mixture at this stage was ~200 mM, due to carryover from the stored protein stocks. This sample was centrifuged at 22,000g at 4 °C for 20 minutes to remove precipitates. The type I-F CRISPR integration complex was then purified on a Superdex 200 10/300 column (Cytiva) equilibrated in SEC buffer (20 mM Tris–HCl pH 7.5, 200 mM monopotassium glutamate, 5 mM EDTA, 1 mM TCEP, 2% glycerol). Then 0.5-ml fractions were individually concentrated and stored. The sixth SEC fraction contained all DNAs and proteins of interest and was further analyzed by cryo-EM (Extended Data Fig. 1d–f).

Cryo-EM sample preparation and data acquisition

Purified integration complex was diluted to a concentration of 1 µM in SEC buffer lacking glycerol (20 mM Tris–HCl pH 7.5, 200 mM monopotassium glutamate, 5 mM EDTA, 1 mM TCEP), such that the final glycerol concentration was 0.2% within 1 hour of freezing. Sample was applied to Quantifoil R2/2 Cu 200 mesh grids that were glow discharged using 15 mA for 15 seconds with a 10 second hold (easiGlow, Pelco). A 4-µl portion of diluted integration complex was applied to the grids, and then the grids were blotted for 5–6 seconds using Vitrobot filter paper (Electron Microscopy Sciences) with a blot force of 6, at 100% humidity, 8 °C, followed by plunge freezing into liquid ethane using a Vitrobot (Mk. IV, ThermoFisher Scientific). A preliminary dataset of 230 movies was collected on Montana State University's Talos Arctica transmission electron microscope (ThermoFisher Scientific), with a field-emission gun operating at an acceleration voltage of 200 kV using parallel illumination conditions⁷⁹. Movies were acquired using a Gatan K3 direct electron detector operated in electron counting mode, applying a total electron exposure of 50 e⁻/Å² over 50 frames (3.995 s exposure, 0.08 s frame time). The SerialEM data collection software was used to collect micrographs at 36,000-fold nominal magnification (1.152 Å per pixel at the specimen level) with a nominal defocus set to 0.5 µm–2.0 µm (ref. 80). Stage movement was used to target the center of four 2.0-µm holes for focusing, and image shift was used to acquire high-magnification images in the center of each of the holes. A preliminary reconstruction was determined from a curated set of 160 images that had CTF fits less than 9 Å and a full-frame motion less than 40 pixels. Briefly, a round of blob picking (150–270 Å) followed by two-dimensional (2D) classification was used to identify 2D classes used as templates for template picking in cryoSPARC⁸¹. Template picking identified 33,403 initial particles from the above 160 images. The 2D classification of these 33,403 particles into 50 classes was used to identify 5 classes with strong structural features containing 4,002 particles. Nonuniform refinement of these 4,002 particles resulted in an ~14.7 Å resolution reconstruction that appeared to contain a complete integration complex; therefore, new grids were prepared as described above and shipped to the National Center for CryoEM Access and Training (NCCAT) and the Simons Electron Microscopy Center located at the New York Structural Biology Center (NYSBC) for additional data collection (Table 1). At NCCAT, grids were imaged using a 300 kV Titan Krios G3i (ThermoFisher Scientific) equipped with a GIF BioQuantum and K3 camera (Gatan). A total of 10,740 images

were recorded with Legikon v.3.5 (ref. 82) with a calibrated pixel size of 0.5335 Å per pixel (micrograph dimension of 11,520 × 8,184 pixels) over a nominal defocus range of –0.7 µm to –2.1 µm and 20 eV slit. Movies were recorded in 'super-resolution mode' (native K3 camera binning 1) with subframes of 50 ms over a 2.5 s exposure (50 frames) to give a total exposure of ~69 e⁻/Å² (Table 1).

Cryo-EM image processing

Patch motion correction and patch CTF correction were performed in cryoSPARC⁸¹. First, 3,792 of 10,740 total images (CTF < 8 Å, full-frame motion < 30 Å) were processed to build an initial template. Blob picking was used to pick particles with diameters ranging from 120 to 280 Å. These ~1.8 million particles were extracted, Fourier-binned 2 × 2 and then subjected to 2D classification (custom parameters: initial classification uncertainty factor = 3; number of online-EM iterations = 30; batchsize per class = 200) (Extended Data Fig. 1g). Particles from 82 of the 200 2D classes were selected for an initial round of ab initio reconstruction and heterogeneous refinement. Particles from one of five of these classes were selected for a second round of ab initio reconstruction and heterogeneous refinement. Particles from one of three of these classes (174,000 particles) were selected for nonuniform refinement (custom parameters: optimize per-particle defocus = true; optimize per-group CTF params = true) to create an initial reconstruction with a resolution of ~3.7 Å, which was used to calculate templates⁸³. These templates were used to choose particles from 9,858 of 10,740 total images (CTF < 8 Å cutoff). These ~5.85 million particles were classified into a total of six classes by heterogeneous refinement, which were seeded with 1 good volumes and 5 junk volumes taken from the above heterogeneous refinement analyses. The ~1.31 million particles in the single selected class, were passed through a round of 2D classification (custom parameters: batchsize per class = 200) (Extended Data Fig. 1h). The ~1.29 million particles from 49 of the 50 2D classes were selected for a round of ab initio reconstruction followed by heterogeneous refinement into two classes. The ~1.1 million particles from one of these two classes were subjected to three-dimensional (3D) classification into four classes (custom parameters: batchsize per class = 20,000; initialization mode = PCA; target resolution = 2 Å; particles per reconstruction = 500; class similarity = 0.3), followed by separate nonuniform refinements of particles from each of these four classes (custom parameters: optimize per-particle defocus = true; optimize per-group CTF params = true)⁸³. The final set of 366,794 particles were re-extracted and recentered, Fourier-binned 2 × 2 and subjected to nonuniform refinement to generate a final reconstruction refined to a global resolution of 3.48 Å on the basis of 0.143 FSC cutoff (Extended Data Fig. 1h–k and Table 1)⁸⁴. The 3D FSC was calculated using the webserver 3dfsc.salk.edu (ref. 85).

Model building and validation

The map was sharpened from two half-maps using the local anisotropic sharpening job in Phenix⁸⁶. The published structure of the *P. aeruginosa* Cas1 homodimer was used as the starting model⁵⁶ because Colabfold consistently failed to predict the alternative fold that one Cas1 subunit adopts to form the asymmetric homodimer interface⁸⁷, even when provided with template structures. Whereas, the Colabfold-predicted models for the *P. aeruginosa* IHF heterodimer and the *P. aeruginosa* Cas2/3 subunit were used as starting models. The conformation of the DNA sequences within the *E. coli* IHF–DNA cocrystal structure (PDB 1IHF)³⁸ was used as starting model for DNA segments within the IHF_{distal} and IHF_{proximal} DNA bends. For all other double-stranded DNA segments, B-form DNA was used as a starting model. Single-stranded DNA segments were built in de novo. Protein and DNA segments were individually rigid-body fitted into the EM density map. The relative orientations of the Cas2, Cas2/3 linker and Cas3 domains were corrected by real-space refinement into the EM density map in WinCoot⁸⁸. The ReadySet job in Phenix was used to generate hydrogens on all proteins

and nucleic acids and prepare the model for further refinement. Then, protein and DNA segments were real-space refined in WinCoot⁸⁸, restrained to ideal geometry, secondary structure and German McClure distance restraints generated in ProSMART from the input models⁸⁹. The models were iteratively real-space refined in WinCoot and in Phenix using Ramachandran and secondary structure restraints^{86,88}. The starting model was used as a reference model, and harmonic restraints on the starting coordinates were enabled. MolProbity⁹⁰ and the PDB validation service server (<https://validate-rcsb-1.wwpdb.org/>) were used to identify problem regions subsequently corrected in WinCoot⁸⁸. For regions of the reconstruction where side chains are not visible (resolution >4.0 Å) the atomic model was truncated to the peptide backbone. For regions of the reconstruction where the backbone was ambiguous the sections of the peptide or DNA model were removed. Contacts and hydrogen bonds between residues were identified by ChimeraX v.1.4 using the 'contacts' and 'hbonds' commands, respectively, with default parameters^{91,92}. The DNAPRODB webserver (<https://dnaprodb.usc.edu/>) was further used to analyze DNA–protein contacts (Extended Data Fig. 4d,e)⁹³. Structure-guided mutagenesis was used to further validate key Cas1–2/3–DNA contacts in the above biochemical assays.

Cas1, Cas2/3 and repeat conservation analysis

To build a list of type I-F Cas1 sequences, CRISPRDetect v.2.4 with default parameters was used to identify CRISPR arrays within a total of 18,225 bacterial and 376 archaeal complete genomes accessed from the NCBI Assembly database on 10 June 2019, as previously described^{22,94}. The 15,274 high-confidence CRISPR arrays were classified with a CRISPR subtype by CRISPRDetect v.2.4 (by matching to a list of repeats with known subtype annotations) and by genetic proximity to subtype-specific *cas* genes (within 20,000 bp). To identify *cas* genes, the 20,000 bp flanking the CRISPR were submitted to PRODIGAL v.2.6.3 (default parameters) to predict all potential open reading frames (ORFs)⁹⁵. This ORF database was then used as input to search for *cas* gene clusters with MasyFinder v.1.0.5 (ref. 96). The following parameters were used: 'masyfinder --sequence-db<peptide_database> --db-type gembase -d<CRISPR_subtype_definitions> -p<HMM_profiles> -w 50 -vv all'. HMM profiles and classification definitions used in MasyFinder were acquired from the local version of CRISPRCasFinder v.4.2.20 (ref. 97). Next, the first repeat and 200 nucleotides upstream of CRISPR arrays (leader), which were classified as type I-F (1,683 arrays), were collected. A nonredundant list of I-F CRISPR leaders (536 leaders) was generated using CD-HIT v.4.8.1 with a 95% identity cutoff⁹⁸. A local copy of FIMO was used to identify matches to the position weight matrix representing the I-F IHF-binding site, as previously described^{22,99}. I-F CRISPR arrays that possess more than one IHF site (IHF_{proximal} and/or IHF_{distal}) in the leader sequences were extracted for downstream analyses. Cas1 homologs were identified within the 20,000-bp flanking regions of extracted 444 I-F CRISPR arrays by using PRODIGAL and MasyFinder with the same parameters described above. A total of 371 Cas1 homologs associated with type I-F CRISPRs and possessing at least one IHF site in the leader sequences were identified. A nonredundant list of Cas1 sequences was generated with CD-HIT v.4.8.1 with a 95% identity cutoff, resulting in 222 sequences⁹⁸. Sequences smaller than 200 residues and larger than 500 residues were removed, and the remaining 205 sequences were further curated with MaxAlign, which selected a list of 144 unique type I-F Cas1 sequences¹⁰⁰. The *P. aeruginosa* PA14 Cas1 sequence was then added to a final list of 145 type I-F Cas1 sequences. To build a list of type I-F Cas2/3 sequences, the *P. aeruginosa* PA14 Cas2/3 sequence was used as an input for HHMER for a search for homologs using three iterations, an *E* value cutoff of 0.0001, against the UNIREF-90 database^{101,102}. A list of 500 representative sequences was further curated with MaxAlign, to generate a final list of 458 unique Cas2/3 sequences. Type I-F Cas1 and Cas2/3 sequences were aligned using the MAFFT webserver with the E-INS-I iterative refinement methods to result in alignments with the highest number of gap-free sites¹⁰³.

To build an updated list of CRISPR repeat sequences, CRISPRDetect v.3.0 with default parameters was used to identify CRISPR arrays within a total of 25,502 bacterial and 398 archaeal complete genomes and chromosomes accessed from the NCBI RefSeq Assembly database (accessed on 10 June 2021)⁹⁴. This search identified CRISPR loci within 58,864 genomic and plasmid sequences, resulting in 24,940 high-confidence CRISPR loci predictions (array quality score >3). Similar to above, CRISPRDetect annotated the subtype of 14,446 of these CRISPR loci, on the basis of the sequence similarity of the repeats in these loci to known CRISPR repeats. The subtypes of the remaining 10,494 CRISPR loci were determined by their proximity to subtype-specific *cas* genes as described above. A total of 5,321 of the 10,494 unclassified CRISPR loci were assigned a subtype using this protocol, such that 5,173 CRISPR loci remained unclassified. The consensus repeats for each of the 24,940 CRISPR loci, as reported by CRISPRDetect, were used for downstream analyses. To ensure the repeats were arranged in the correct orientation, the 24,940 repeats were grouped by subtype, and each group was individually aligned by MAFFT using the '--adjustdirection' parameter. Sequence logos of the first and last three base pairs of CRISPR repeats were made using Weblogo v.3.7.1 for CRISPR subtypes and across all subtypes^{104,105} (Extended Data Fig. 4g).

Reporting summary

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

Data availability

The data that support the findings of this study are available from the corresponding author upon request. Curated raw multi-frame movies have been deposited along with a reference gain file under accession number EMPIAR-I1659. Cryo-EM maps were deposited in the Electron Microscopy Data Bank under accession number EMD-29280. The atomic model of the type I-F integration complex was deposited in the PDB under accession number 8FLJ. Plasmids generated in this study are available from Addgene. Source data are provided with this paper.

Code availability

Code is available at <https://github.com/WiedenheftLab/>.

References

- Herzik, M. A., Wu, M. & Lander, G. C. High-resolution structure determination of sub-100 kDa complexes using conventional cryo-EM. *Nat. Commun.* **10**, 1032 (2019).
- Mastrorarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51 (2005).
- Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. CryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
- Suloway, C. et al. Automated molecular microscopy: the new Legimon system. *J. Struct. Biol.* **151**, 41–60 (2005).
- Punjani, A., Zhang, H. & Fleet, D. J. Non-uniform refinement: adaptive regularization improves single-particle cryo-EM reconstruction. *Nat. Methods* **17**, 1214–1221 (2020).
- Scheres, S. H. W. & Chen, S. Prevention of overfitting in cryo-EM structure determination. *Nat. Methods* **9**, 853–854 (2012).
- Tan, Y. Z. et al. Addressing preferred specimen orientation in single-particle cryo-EM through tilting. *Nat. Methods* **14**, 793–796 (2017).
- Liebschner, D. et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr. D Struct. Biol.* **75**, 861–877 (2019).
- Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).

88. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).
89. Nicholls, R. A. *Conformation-independent Comparison of Protein Structures*. PhD thesis, Univ. of York (2011).
90. Williams, C. J. et al. MolProbity: more and better reference data for improved all-atom structure validation. *Protein Sci.* **27**, 293–315 (2018).
91. Goddard, T. D. et al. UCSF ChimeraX: meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14–25 (2018).
92. Pettersen, E. F. et al. UCSF ChimeraX: structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).
93. Sagendorf, J. M., Markarian, N., Berman, H. M. & Rohs, R. DNAProDB: an expanded database and web-based tool for structural analysis of DNA–protein complexes. *Nucleic Acids Res.* **48**, D277–D287 (2020).
94. Biswas, A., Staals, R. H. J., Morales, S. E., Fineran, P. C. & Brown, C. M. CRISPRDetect: a flexible algorithm to define CRISPR arrays. *BMC Genomics* **17**, 356 (2016).
95. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **11**, 119 (2010).
96. Abby, S. S., Néron, B., Ménager, H., Touchon, M. & Rocha, E. P. C. MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. *PLoS ONE* **9**, e110726 (2014).
97. Couvin, D. et al. CRISPRCasFinder, an update of CRISPRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res.* **46**, W246–W251 (2018).
98. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
99. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
100. Gouveia-Oliveira, R., Sackett, P. W. & Pedersen, A. G. MaxAlign: maximizing usable data in an alignment. *BMC Bioinform.* **8**, 312 (2007).
101. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).
102. Suzek, B. E. et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
103. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
104. Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18**, 6097–6100 (1990).
105. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).

Acknowledgements

Thanks to members of the B.W. laboratory for feedback and discussions. We thank M. Matyszewski and J. Jeliakov for helpful discussions. Thanks to C. Hophan-Nichols for computational support. A.S.-F. is a postdoctoral fellow of the Life Science Research Foundation that is supported by the Simons Foundation. A.S.-F. is supported by the Postdoctoral Enrichment Program Award from the Burroughs Wellcome Fund. This work was supported by National Institutes of Health, United States grant 1K99GM147842

(A.S.-F.). L.T. and A.B.G. are supported by Montana State University's Undergraduate Scholars Program, and by the NIH NIGMS IDeA program (P20GM103474). This work was performed using the cryo-EM facility at Montana State University (NSF 1828765 and the M.J. Murdock Charitable Trust). Microscopy was also performed at the National Center for CryoEM Access and Training (NCCAT) and the Simons Electron Microscopy Center located at the New York Structural Biology Center, supported by the NIH Common Fund Transformative High Resolution Cryo-Electron Microscopy program (U24 GM129539), and by grants from the Simons Foundation (SF349247) and NY State Assembly. Research in the Wiedenheft laboratory is supported by the NIH (R35GM134867), the M.J. Murdock Charitable Trust, a young investigator award from Amgen, the Montana State University Agricultural Experimental Station (USDA NIFA) and a sponsored research agreement from VIRIS Detection Systems. Molecular graphics and analyses performed with UCSF ChimeraX, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from National Institutes of Health R01-GM129325 and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases. Funders had no role in the conceptualization, designing, data collection, analysis, decision to publish or preparation of the manuscript.

Author contributions

A.S.-F. carried out conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, supervision, validation, visualization and writing (original draft). W.S.H., T.W. and M.B. performed data curation, investigation, methodology, visualization and writing (review and editing). A.B.G. and R.A.W. undertook investigation and methodology. L.T. carried out visualization. C.C.G. contributed to software, resources and writing (review and editing). K.N. and E.T.E. performed investigation and resources. G.C.L. undertook methodology, supervision, visualization and writing (review and editing). B.W. was responsible for funding acquisition, project administration, resources, supervision, visualization and writing (review and editing).

Competing interests

B.W. is the founder of SurGene and VIRIS Detection Systems. B.W. and A.S.-F. are inventors on patent applications related to CRISPR-Cas systems and applications thereof. The remaining authors declare no competing interests.

Additional information

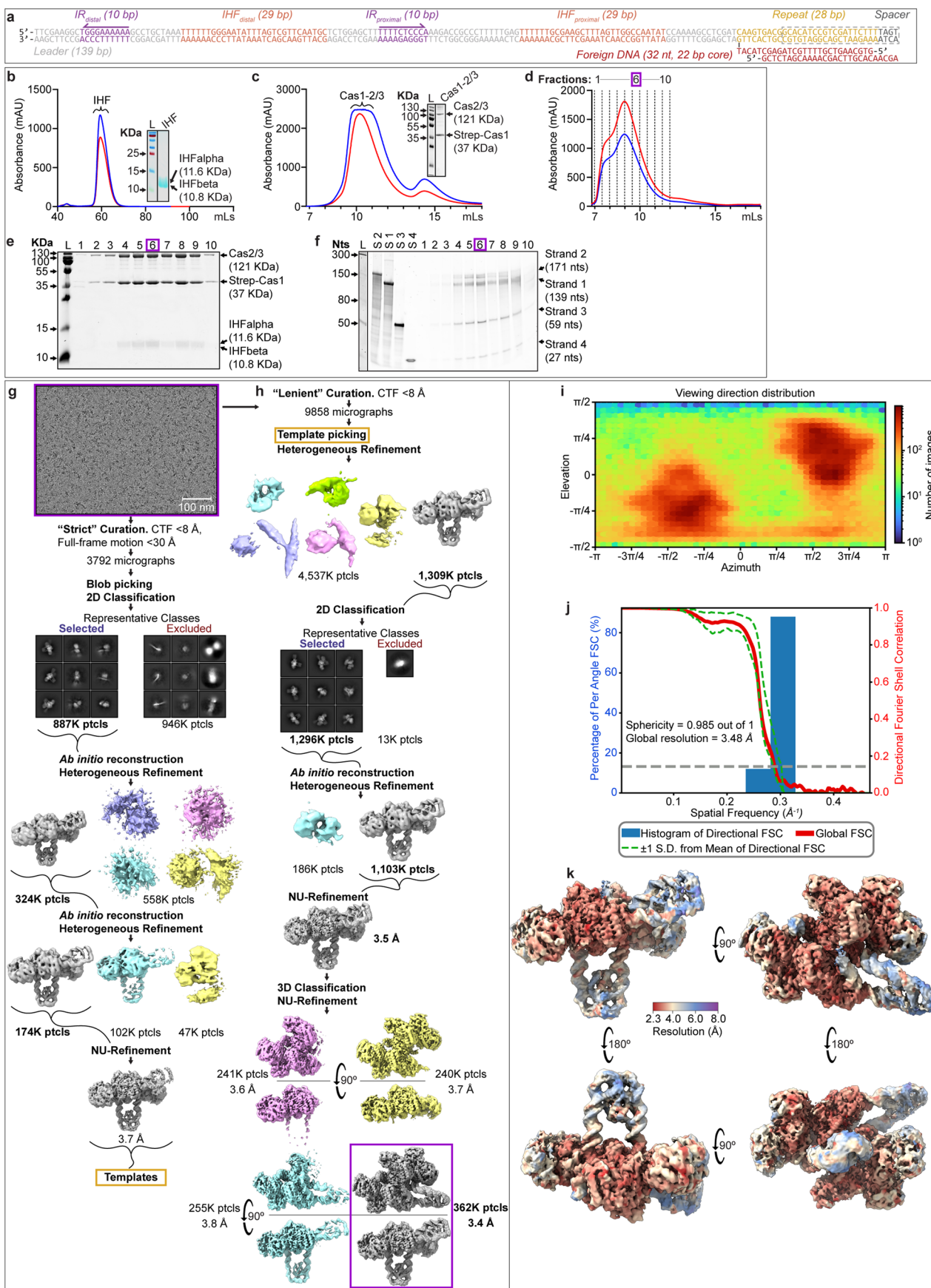
Extended data is available for this paper at <https://doi.org/10.1038/s41594-023-01097-2>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41594-023-01097-2>.

Correspondence and requests for materials should be addressed to Blake Wiedenheft.

Peer review information *Nature Structural & Molecular Biology* thanks Elizabeth Kellogg for her contribution to the peer review of this work. Dimitris Typas was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. Peer reviewer reports are available.

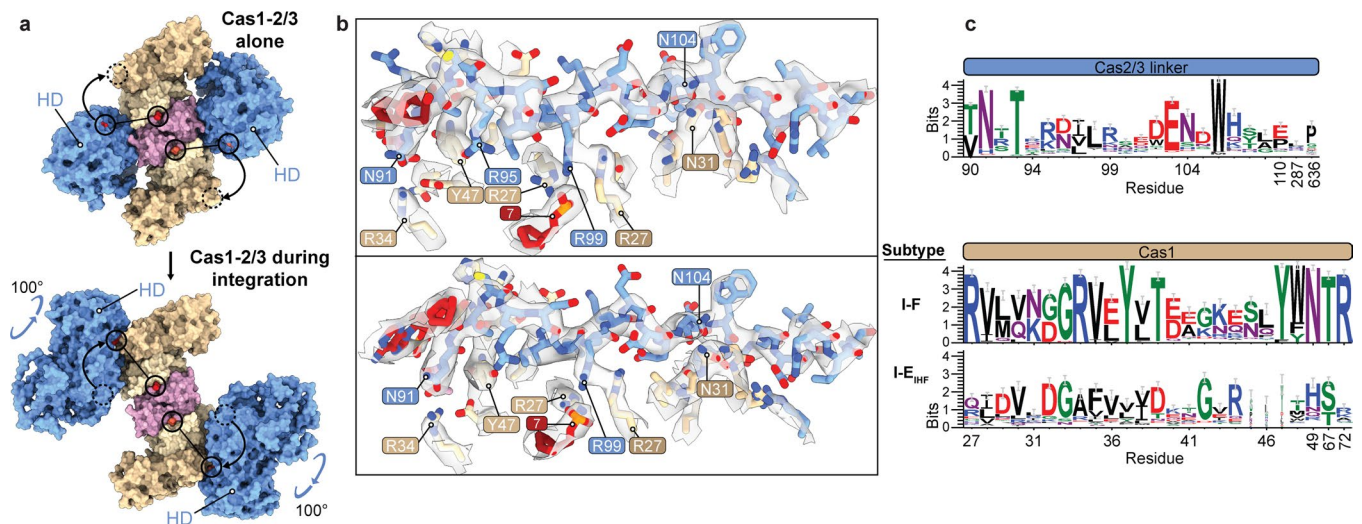
Reprints and permissions information is available at www.nature.com/reprints.



Extended Data Fig. 1 | See next page for caption.

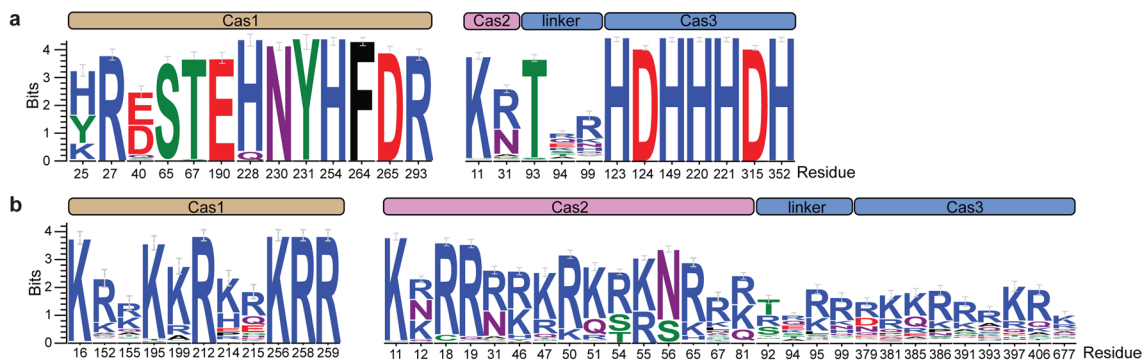
Extended Data Fig. 1 | Cryo-EM sample preparation, imaging and processing for type I-F integration complex. **a**, Sequence-level schematic of DNA used to assemble the integration complex. The length of each motif is listed. The latter two thirds of the CRISPR repeat and second spacer (grey dashed box) could not be resolved in the cryo-EM reconstruction. See also Supplementary Table 1. **b**, Size-exclusion chromatography (SEC) profile (Superdex 75 16/600, Cytiva) of IHF heterodimer purified as described in methods section, and SDS-PAGE gel (inset). **c**, SEC profile (Superdex 200 10/300, Cytiva) of Cas1-2/3 heterohexamer purified as described in methods section, and SDS-PAGE gel (inset). **d**, I-F integration complex was assembled from purified DNAs, IHF and Cas1-2/3 as described in the methods section, and the assembled complex was further purified by size-exclusion chromatography (SEC) (Superdex 200 10/300, Cytiva). Individual fractions were collected along the elution profile, and were concentrated and stored separately for further analysis and imaging. **e**, Individual SEC fractions were analyzed by SDS-PAGE to determine which fractions contained all the proteins necessary for a complete complex.

f, Individual SEC fractions were phenol-chloroform extracted, and the aqueous layer was analyzed by Urea-PAGE to determine which fractions contained all four DNA strands necessary for a complete complex. The fraction chosen for cryo-EM analysis is indicated with a dotted purple box. **g**, Image processing pipeline for a small subset of 10,740 total micrographs for the type I-F integration complex, to generate an initial model for template picking. Scale bar represents 100 nm. **h**, Final image processing pipeline for the type I-F integration complex. **i**, Viewing direction distribution plot depicting particle orientations present in final reconstruction. More populated views are shown in red, and less populated views are shown in blue. **j**, 3D Fourier Shell Correlation (3DFSC) of the final I-F integration complex reconstruction. The global resolution at 0.143 is indicated by a dashed line, 3.48 Å. **k**, Local resolution estimation of the cryo-EM reconstruction calculated by cryoSPARC⁸¹. The purification of proteins, assembly of the integration complex, and analysis of these samples by SDS-PAGE or Urea page was performed once. Micrographs were collected on two separate occasions with similar results.



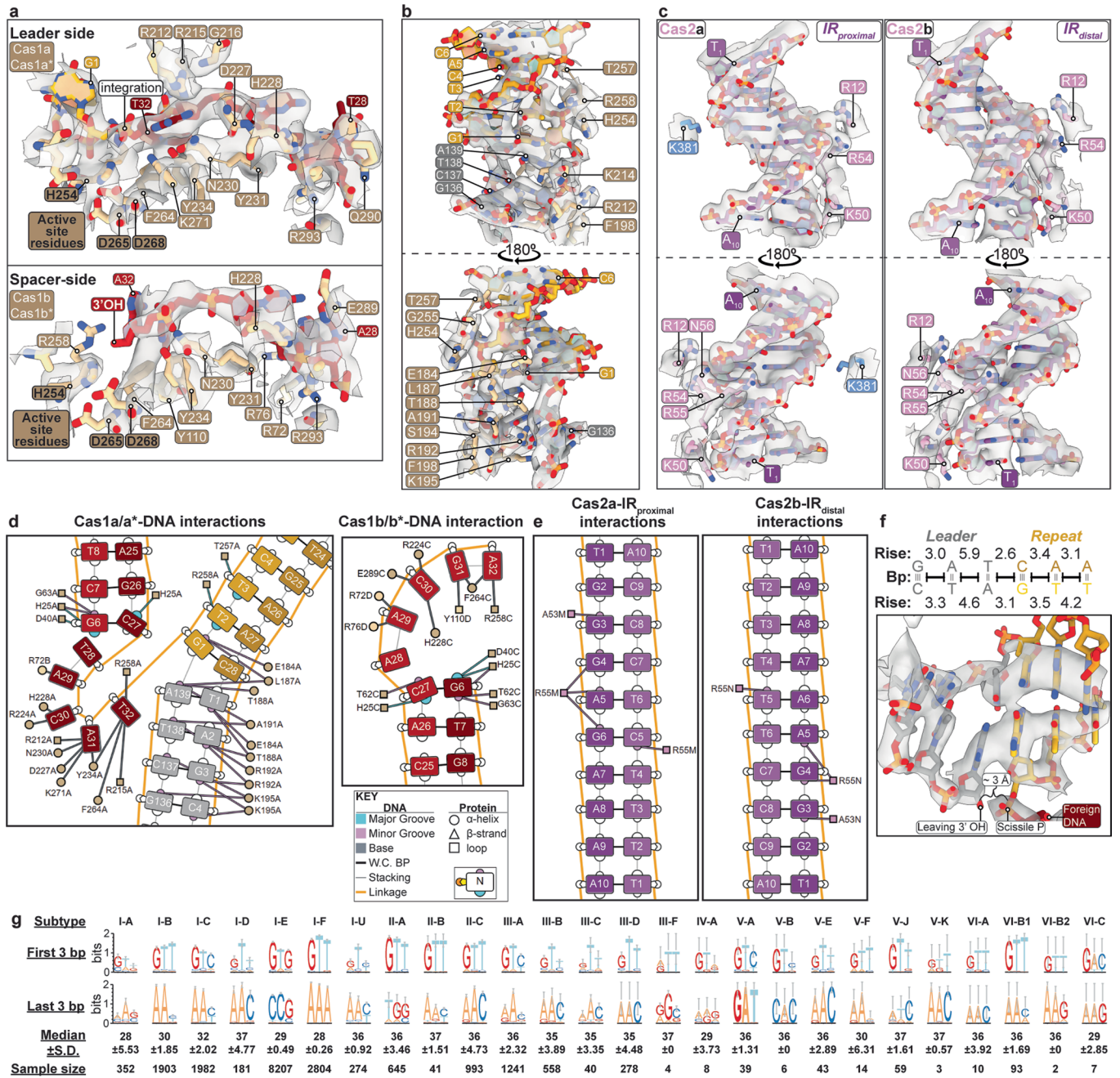
Extended Data Fig. 2 | Cas1-2/3 undergoes a large structural rearrangement during integration. **a**, The Cas3 domains of the Cas1-2/3 complex have undergone a $\sim 100^\circ$ rotation in the structure of the integration complex as compared to Cas1-2/3 alone⁵⁴. The positions of the first (90) and last residues (110) of the Cas2/3 linker are shown (Cas2/3a, red; Cas2/3b, tomato). The linker residues were not resolved for the previously determined pseudo-atomic model of Cas1-2/3 alone, and so are not shown for either complex for clarity⁵⁴. The rotation of the Cas3 domains outwards unveils new DNA binding sites on two opposing faces of the Cas2 dimer. **b**, Zoom-in on the atomic fit of the Cas2/3

linker to the cryo-EM map. The Cas2/3 linker is disordered in the absence of Cas1 (PDB:5B71)⁵⁵, but has become ordered in the I-F integration complex structure due to packing by the foreign DNA against the Cas1 beta hairpins. **c**, A sequence logo depicting the conservation of Cas2/3 linker residues (top). The Cas1 residues that contact the Cas2/3 linker are conserved in Cas1 proteins associated with type I-F CRISPR loci (middle), but are not conserved in the closely related Cas1 proteins associated with type I-E CRISPR loci that have similar IHF motif-containing leaders²². Residues are numbered according to *P. aeruginosa* PA14 Cas1 and Cas2/3 proteins.



Extended Data Fig. 3 | Conservation analysis of Cas1-2/3 residues involved in DNA binding and integration. a, Conservation of Cas1 and Cas2/3 residues involved in binding the foreign DNA, or catalyzing the strand transfer reaction, or catalyzing the degradation of nucleic acids. See Fig. 2. **b**, Conservation of

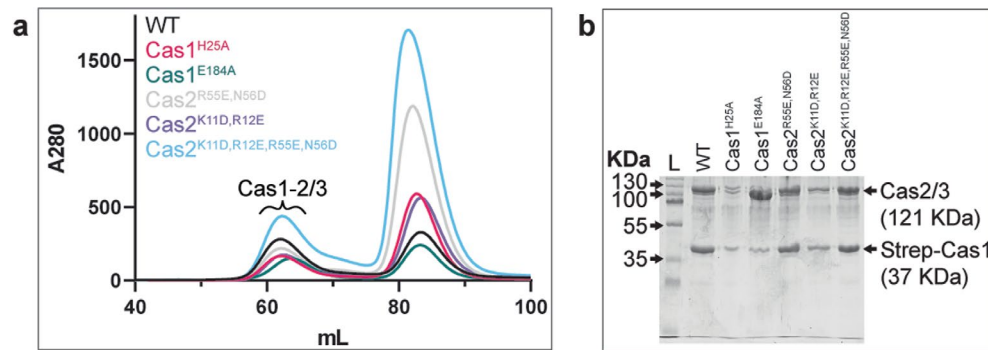
basic and polar Cas1 and Cas2/3 residues involved in accommodating the DNA duplexes bound by the Cas1-2/3 complex during integration. See Fig. 3. Residues are numbered according to *P. aeruginosa* Cas1 and Cas2/3 proteins.



Extended Data Fig. 4 | Cas1-2/3 predominantly recognizes IR motifs, CRISPR repeat and foreign DNA through non-sequence specific interactions.

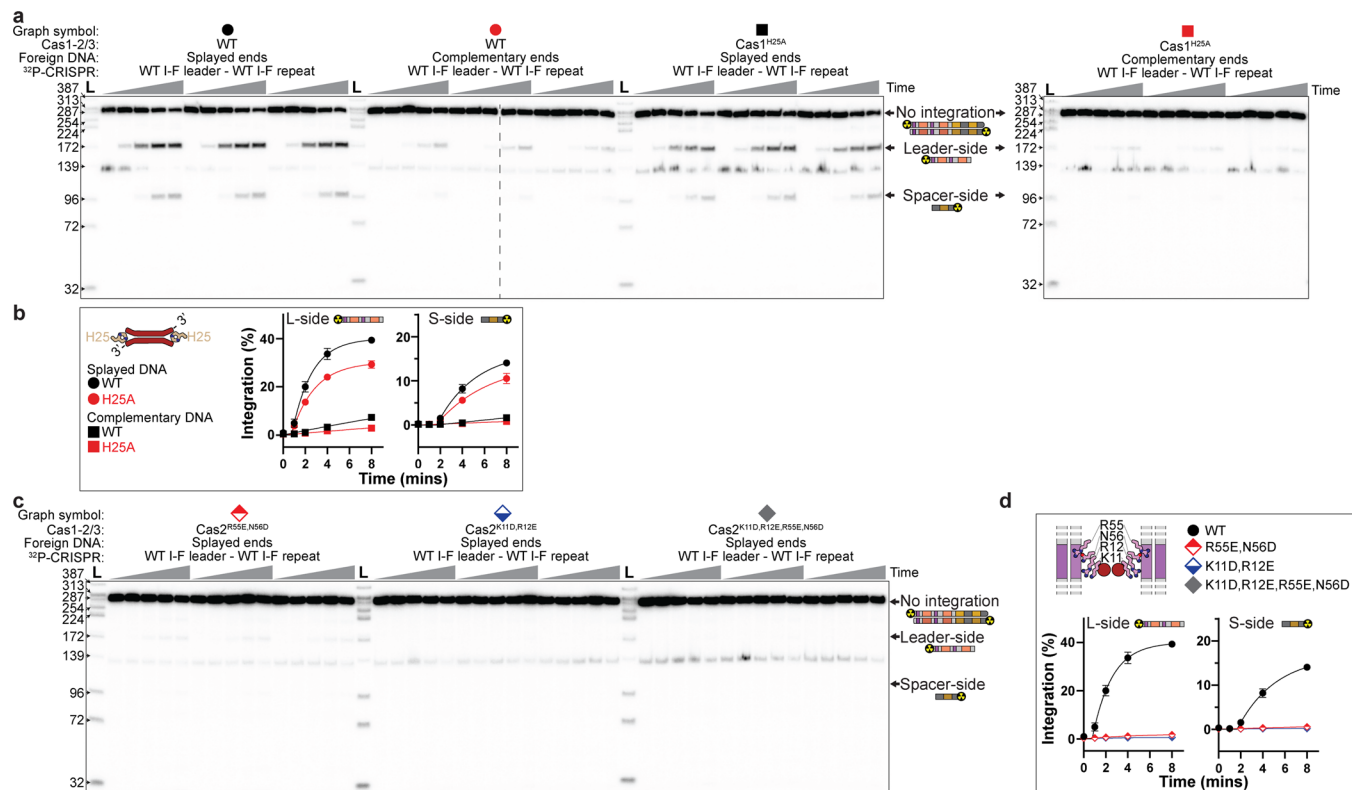
a, Splayed 3' ends of the foreign DNA are directed into the Cas1 transesterification active site. The product of the first strand-transfer reaction is shown in the Cas1a* active site (top), and the 3' OH of the other end of the foreign DNA is positioned in the Cas1b* active site (bottom). The cryo-EM map is shown in transparent grey. **b**, Zoom-in on the Cas1-2/3 contacts to the CRISPR repeat (ChimeraX contacts command with default parameters). Most protein contacts occur to the DNA backbone and minor groove. Cas1 residue E184 appears to probe nucleotide G1 of the repeat. **c**, Zoom-in on the Cas1-2/3 contacts to the IR leader motifs. Most protein contacts occur to the DNA backbone and minor groove. **d**, DNAProDB analysis of Cas1-2/3 interactions with the 3' ends of foreign DNA and the CRISPR leader-repeat junction. For clarity, only protein interactions to the nucleobases are shown⁹³. **e**, DNAProDB analysis of Cas1-2/3 interactions with the IR leader

motifs⁹³. For clarity, only protein interactions to the nucleobases are shown. **f**, Zoom-in on the atomic fit of the base-pairs around the leader-repeat junction (-3 and +3 bps, coordinated by Cas1a*) to the cryo-EM map. Tension in the DNA loop at the leader-repeat junction has been released in the post-integration structure by a physical separation of base-pairs, as measured by an increase in base step rise. This tension may further pull the leaving 3'OH out of the Cas1 transesterification active site, to inhibit disintegration of the foreign DNA from the repeat. The approximate local base step rise was calculated using the <http://web.x3dna.org/webserver>. **g**, A bioinformatic analysis of the first repeat from 24,940 CRISPR loci reveals that a 5' GT dinucleotide is strongly conserved across most CRISPR subtypes. Similarly, a 5' GT is present at the spacer-end of the repeat (seen as AC-3' on the sense strand) within certain CRISPR subtypes (I-D, II-C, III-C, III-D, V-B, V-E, V-K, VI-A), but it is not broadly conserved.



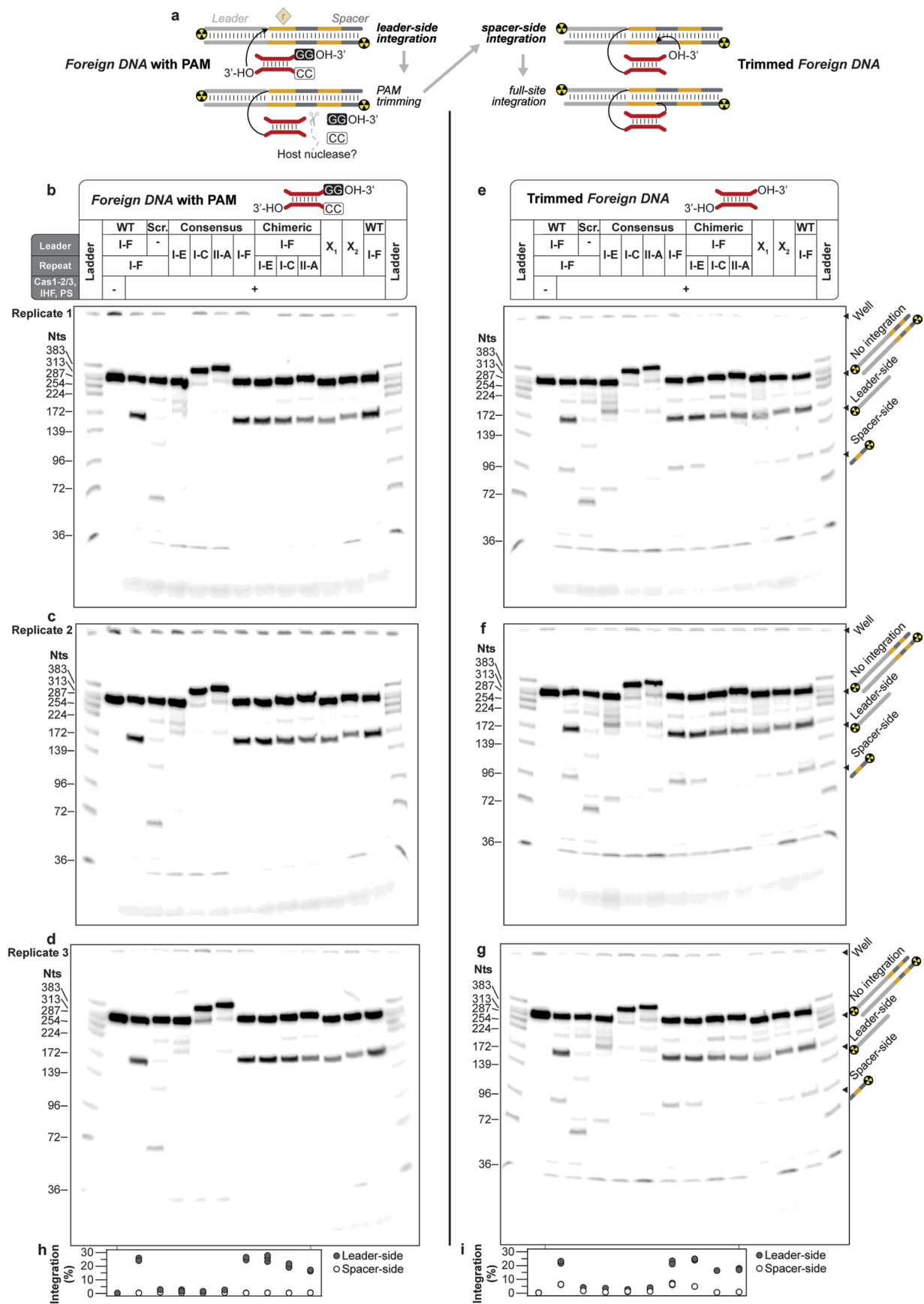
Extended Data Fig. 5 | Purification of structure-guided mutants of Cas1-2/3. **a**, SEC profile of a new preparation of wildtype Cas1-2/3 and all variants purified in the same manner on a Superdex 200 16/600 (Cytiva). An excess of free Strep-

tagged Cas1 elutes at approximately 82 mL. **b**, SDS-PAGE gel of the Cas1-2/3 hetero-hexamers for all purified Cas1-2/3 variants. The SDS-PAGE gel of all Cas1-2/3 samples was run twice with similar results.



Extended Data Fig. 6 | Validation of Cas1-2/3 interactions with the foreign DNA and IR motifs. a, Time-course integration reactions to test the role of Cas1^{H25} in playing the foreign DNA ends. Integration reactions were performed with trimmed foreign DNA (lacking a PAM) in triplicate, resolved on denaturing polyacrylamide gels. Timepoints were taken at 0, 1, 2, 4 and 8 minutes. Reactions were stopped by the addition of phenol. A ³²P-labelled DNA that is shorter (140-160 bp) than the full length CRISPR is present in some DNA preparations (also see Extended Data Fig. 6c). Full-length CRISPR DNA, leader- and spacer-side integration products, do not overlap with this band. Further, Cas1-2/3, foreign DNA and IHF are in excess over the ³²P-labelled DNA. The 140-160 bp band does not interfere with the quantification or generation of integration products. **b**, Quantification of time-course experiments to determine the role of Cas1 residue H25 in integration. The mean and standard deviation of three replicate experiments are shown. The Cas1^{H25A} mutant integrates splayed and

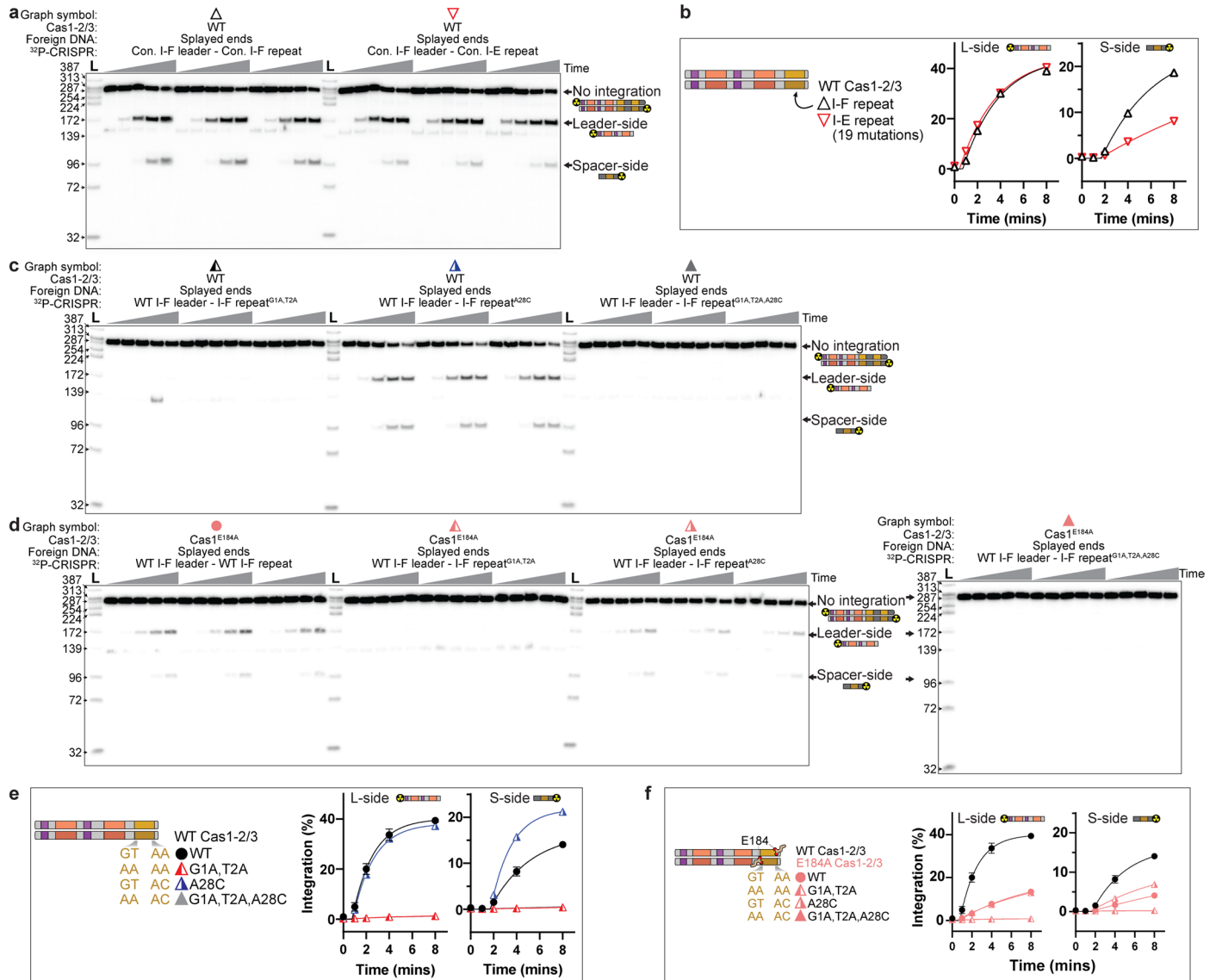
fully complementary foreign DNA fragments less efficiently than WT Cas1-2/3, suggesting that H25 steers the non-nucleophilic DNA strand away from the Cas1 active site. These results mirror the previously published effect of type I-E Cas1-2 tyrosine wedge mutation¹⁰. **c**, Time-course integration reactions to test the role of Cas2 residues in recognition of the IR motifs in the leader, performed as in panel a. **d**, Quantification of time-course experiments to determine the role of Cas2 residues K11, R12, R55 and N56 in integration. The mean and standard deviation of three replicate experiments are shown. The Cas2^{R55E,N56D}/3 mutant retains a small amount of integration activity. The Cas2^{K11D,R12E}/3 and Cas2^{K11D,R12E,R55E,N56D}/3 mutants do not integrate DNA into the I-F CRISPR. Quantification of leader- (grey circles) or spacer-side (white circles) integration events from all three replicate gels. Individual dots for each triplicate reaction are shown, and some dots overlap.



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

Extended Data Fig. 7 | PAM blocks Cas-mediated integration of foreign DNA into CRISPR repeat. **a**, Schematic summarizing the step-wise strand-transfer reactions catalyzed by a Cas integrase. In the absence of putative host nucleases that trim PAMs, the foreign DNA fragments that contain a PAM stall the reaction at leader-side integration. However, the foreign DNA fragments that have been trimmed (no PAM) proceed through leader- and spacer-side integration. **b-d**, Endpoint integration reactions performed with a PAM-containing foreign DNA in triplicate, resolved on denaturing polyacrylamide gels. The X1 and X2 lanes signify lanes that were not further analyzed for this manuscript.

e-g, Endpoint integration reactions performed with a trimmed foreign DNA in triplicate, resolved on denaturing polyacrylamide gels. The X1 and X2 lanes signify integration substrates that were not analyzed for this manuscript. **h, i**, Quantification of leader- (grey circles) or spacer-side (white circles) integration events from all three replicate gels. Individual dots for each triplicate reaction are shown, and some dots overlap. Three independent gels were run for PAM-containing or trimmed Foreign DNA integration reactions with similar results.



Extended Data Fig. 9 | Validation of Cas1-2/3 interactions with the repeat.

a, Time-course integration reactions to compare rate of integration into I-E and I-F repeats downstream of a I-F leader. **b**, Quantification of time-course experiments to determine the impact of 19 mutations associated with swapping the I-F repeat for the I-E repeat, on integration. Leader-side integration is indistinguishable. But spacer-side integration is slower into the I-E repeat. The mean and standard deviation of three replicate experiments are shown. **c**, Time-course integration reactions to measure the impact of I-F repeat mutations on integration rate. **d**, Time-course integration reactions to measure the impact of I-F repeat mutations on integration rate, in the context of a Cas1^{E184A} mutation. The Cas1^{E184A} mutation is expected to disrupt 5' G recognition, but also impacts

stability of the Cas1-2/3 complex (Extended Data Fig. 5). **e**, Quantification of time-course experiments to determine the impact of I-F repeat mutations on integration rate, in the context of WT Cas1-2/3. The mean and standard deviation of three replicate experiments are shown. **f**, Quantification of time-course experiments to determine the impact of I-F repeat mutations on integration rate, in the context of Cas1^{E184A}-2/3. In panels e and f, no integration occurs into either the 'G1A,T2A,A28C' or 'G1A,T2A' repeat mutations, and the datapoints for these plots overlaps at roughly Y = 0 over the time course. The mean and standard deviation of three replicate experiments are shown. Each Urea-PAGE gel was run once.

Reporting Summary

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

Statistics

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

n/a | Confirmed

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

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection Commercial software used to collect data: SerialEM v3.8, Legion v3.5 and Amersham Typhoon.

Data analysis Commercial software used to analyze data: Multi Gauge v3, cryoSPARC v3.3.2, WinCoot v0.9.8.1, MolProbit v4.5, PDB validation service V5.19.1/0.43.4, Phenix v1.20.1, Colabfold v1.2.0, CRISPRDetect v2.4, PRODIGAL v2.6.3, CRISPRCasFinder v4.2.20, MacsFinder v1.0.5, CD-HIT v4.8.1, FIMO v5.5.3, MaxAlign v1.1, MAFFT v7.490, GraphPad Prism v10, Excel, DNAProDB, ChimeraX v1.4, Weblogo v3.7.1.

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

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The data that support the findings of this study are available from the corresponding author Blake Wiedenheft upon request. Curated raw multi-frame movies have been deposited along with a reference gain file under accession number EMPIAR-11659. Cryo-EM maps were deposited in the Electron Microscopy Data Bank under

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender	<input type="text" value="This information has not been collected."/>
Reporting on race, ethnicity, or other socially relevant groupings	<input type="text" value="See above."/>
Population characteristics	<input type="text" value="See above."/>
Recruitment	<input type="text" value="See above."/>
Ethics oversight	<input type="text" value="See above."/>

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

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

- Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://doi.org/10.1126%2Fscience.aao0679)

Life sciences study design

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

Sample size	<input type="text" value="No sample size calculation was performed. Experiments were performed in triplicate, as is standard practice in biochemical assays for integration assays (https://doi.org/10.1126%2Fscience.aao0679). Statistical methods were not used to determine sample size, but were used to calculated standard deviations."/>
Data exclusions	<input type="text" value="Variants of the I-F CRISPR, termed X1 and X2 represent mutants that were synthesized to test an early hypothesis based on a misinterpretation of the I-F integration complex structure. Integration assays were performed with the X1 and X2 mutants, but these assays were not analyzed further when the error in the structure interpretation was realized. No other data were excluded from the analyses."/>
Replication	<input type="text" value="Biochemical experiments to measure integration were replicated 3 times."/>
Randomization	<input type="text" value="There were no experimental subjects in this study."/>
Blinding	<input type="text" value="No analysis that required subjective assessment was performed."/>

Reporting for specific materials, systems and methods

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

Materials & experimental systems

- | n/a | Included in the study |
|-------------------------------------|--|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Antibodies |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Eukaryotic cell lines |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Palaeontology and archaeology |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Animals and other organisms |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Clinical data |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Dual use research of concern |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Plants |

Methods

- | n/a | Included in the study |
|-------------------------------------|---|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> ChIP-seq |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Flow cytometry |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> MRI-based neuroimaging |