RESEARCH ARTICLE

CRISPR BIOLOGY

Structures of the CRISPR genome integration complex

Addison V. Wright,[1]* Jun-Jie Liu,[1,2]* Gavin J. Knott,[1] Kevin W. Doxzen,[3] Eva Nogales,[1,2,4] Jennifer A. Doudna[1,2,3,4,5,6,7]†

CRISPR-Cas systems depend on the Cas1-Cas2 integrase to capture and integrate short foreign DNA fragments into the CRISPR locus, enabling adaptation to new viruses. We present crystal structures of Cas1-Cas2 bound to both donor and target DNA in intermediate and product integration complexes, as well as a cryo–electron microscopy structure of the full CRISPR locus integration complex, including the accessory protein IHF (integration host factor). The structures show unexpectedly that indirect sequence recognition dictates integration site selection by favoring deformation of the repeat and the flanking sequences. IHF binding bends the DNA sharply, bringing an upstream recognition motif into contact with Cas1 to increase both the specificity and efficiency of integration. These results explain how the Cas1-Cas2 CRISPR integrase recognizes a sequence-dependent DNA structure to ensure site-selective CRISPR array expansion during the initial step of bacterial adaptive immunity.

CRISPR-Cas (clustered regularly interspaced short palindromic repeats–CRISPR associated) bacterial adaptive immune systems store fragments of viral DNA in the CRISPR array, a genomic locus comprising direct sequence repeats of ~20 to 50 base pairs, separated by virally derived spacer sequences of similar length (*1*–*4*). In most systems, a transcriptional promoter located in an AT-rich leader sequence preceding the first CRISPR repeat gives rise to precursor CRISPR transcripts that are processed and used to recognize viral nucleic acids by base-pairing with complementary sequences. Bacteria acquire immunity to new viruses when the CRISPR integrase, a heterohexameric complex of four Cas1 and two Cas2 proteins, inserts new viral DNA at the first CRISPR repeat after the leader sequence (*5*–*7*). Integration involves nucleophilic attack by the 3′ ends of the viral DNA fragment, called a protospacer, at each end of the repeat (Fig. 1A) (*7*). Half-site intermediates form when one of the two protospacer DNA ends attacks the CRISPR locus integration site, and these can either progress to full-site integration products or be disintegrated, leaving the target sequence intact (*7*, *8*).

To ensure effective acquisition of new immunity and avoid deleterious insertions into the genome, integration by Cas1-Cas2 must be highly specific for the CRISPR locus. In the type I CRISPR system from *Escherichia coli*, acquisition requires sequences spanning the leader-repeat junction, as well as an inverted repeat motif in the repeat (*8*–*11*). IHF (integration host factor), a histone-like protein, binds in the leader and assists in recruiting Cas1-Cas2 to the leader-proximal repeat, possibly involving a secondary upstream binding site (*10*, *12*, *13*). The mechanism by which Cas1-Cas2 recognizes these sequences has thus far been unknown.

Here we present structures of the Cas1-Cas2 CRISPR integrase bound to both substrate and target DNA in intermediate and product integration states. We also present a structure of the entire natural integration complex, including Cas1-Cas2, the DNA substrate, and a 130–base pair DNA target sequence in complex with IHF. These structures show how specificity for the CRISPR repeat relies on target DNA deformation to allow access to both Cas1 integrase active sites. In addition to recruiting a secondary recognition site, IHF sharply bends the target DNA adjacent to the integration site, favoring integrase binding to this locus and thereby suppressing off-target integration.

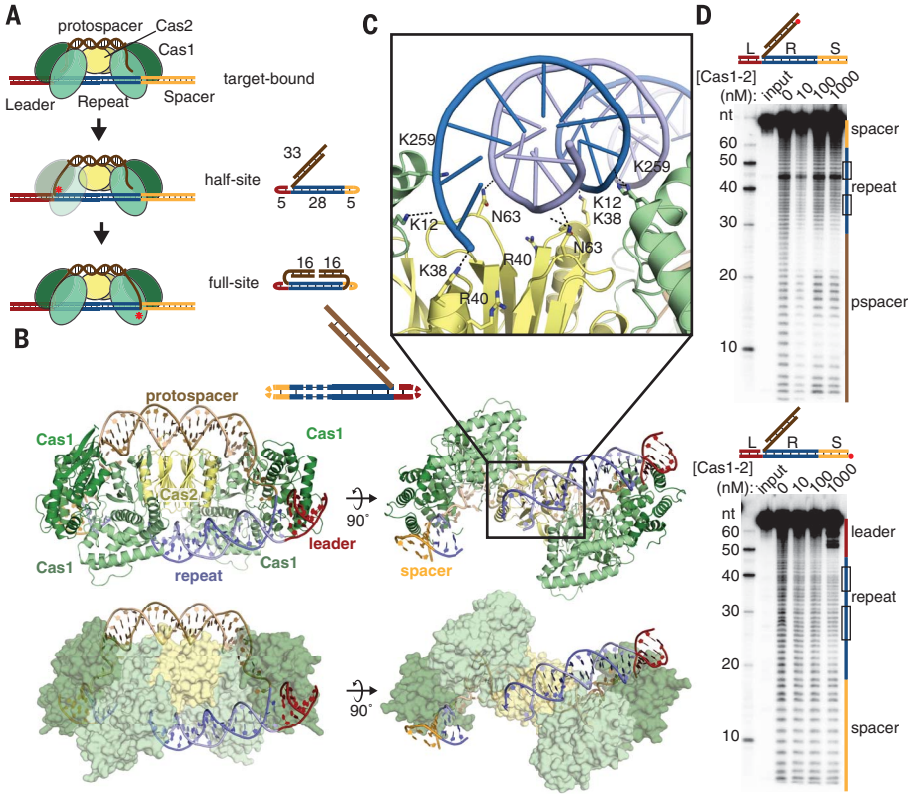

Fig. 1. Half-site binding by Cas1-Cas2. (**A**) Cartoon of integration by Cas1-Cas2. Crystallography substrates are shown next to the corresponding reaction intermediate, with nucleotide lengths indicated. Red asterisks represent integration events. (**B**) Cartoon and surface representations of the half-site substrate bound by Cas1-Cas2. DNA is colored as in (A). A substrate schematic is shown above, with disordered regions shown as dashed lines. (**C**) Close-up of backbone interactions between Cas1-Cas2 and half-site repeat DNA. Polar contacts are shown as dotted lines. (**D**) Hydroxyl-radical footprinting of radiolabeled half-site DNA. The input is untreated DNA. The substrates are shown above the gel, with the radiolabel indicated with a red circle (L, leader; R, repeat, S, spacer). Regions of the gel corresponding to the leader, repeat, spacer, and protospacer (pspacer) are indicated alongside the gel. The inverted repeat regions of the repeat are boxed. nt, nucleotides. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; D, Asp; E, Glu; F, Phe; H, His; K, Lys; N, Asn; Q, Gln; R, Arg; S, Ser.

These results suggest an unexpected mechanism of target recognition with implications for the engineering of the CRISPR integrase as a genome-tagging tool.

Target binding in the half-site intermediate

To determine the mechanism by which Cas1-Cas2 recognizes its target sequence, we crystallized the integrase bound to DNA substrates representing a half-site integration intermediate and the full-site integration product (Fig. 1A). The full-site product mimic, which we term the pseudo–full-site substrate, was designed with a break in the middle of the protospacer to allow Cas1-Cas2 to access the repeat (Fig. 1A). Both substrates bound to Cas1-Cas2 with high affinity (fig. S1). The half-site–bound structure, refined at 3.9-Å resolution, revealed an overall complex architecture similar to that of the previously solved protospacer-bound structures (Fig. 1B, fig. S2, and table S1) (*14*, *15*). A Cas2 dimer sits at the center of two Cas1 dimers, with the protospacer DNA stretching across the flat back of the complex. The first 18 base pairs of the repeat sequence bind across a central channel formed by Cas2 and the noncatalytic Cas1 monomers, with the leader-repeat junction positioned across a Cas1 active site (Fig. 1B and fig. S3, A and B). Seven nucleotides of the spacer-proximal repeat are unresolved, whereas the repeat-spacer junction binds at the distal Cas1 active site. Basic residues on both Cas2 (K38 and R40) and the noncatalytic Cas1 monomers (K12 and K259) are positioned to contact the phosphate backbone of the midrepeat DNA (Fig. 1, B and C) (*15*). Charge-swap mutations of these residues reduce or eliminate acquisition of new spacers in vivo, confirming their importance for the CRISPR integration reaction (fig. S4A).

Although earlier work suggested that inverted sequence motifs in the repeat might form a cruciform structure during target recognition, our structure shows that the center of the repeat remains a canonical duplex at this intermediate stage of integration (*7*, *16*, *17*). Although the inverted repeat sequences are critical for spacer acquisition, we found no evidence of sequence-specific contacts in these motifs (Fig. 1C) (*9*, *11*, *18*). Contacts between the midrepeat DNA and the integrase proteins are limited to nonspecific backbone interactions, with no regions of Cas1 or Cas2 positioned to insert into either the major or minor groove. To test for contacts in solution, we performed hydroxyl-radical footprinting of the half-site substrate bound by the complex (Fig. 1D). Protection of the backbone is evident in the protospacer, including in the single-stranded end where the DNA binds in a channel of Cas1. Only weak protection occurs near the ends of the repeat on the nonintegrated target strand and largely does not overlap with the inverted repeats. Several hypersensitive nucleotides are apparent at the beginning of the second inverted repeat even in the absence of protein, suggesting that these nucleotides exhibit increased flexibility or a distorted conformation in solution. Although direct sequence readout could involve a distinct but transient binding mode before half-site integration, our data suggest that integrase recognition of the repeat sequence likely relies on a mechanism other than base-specific hydrogen-bonding.

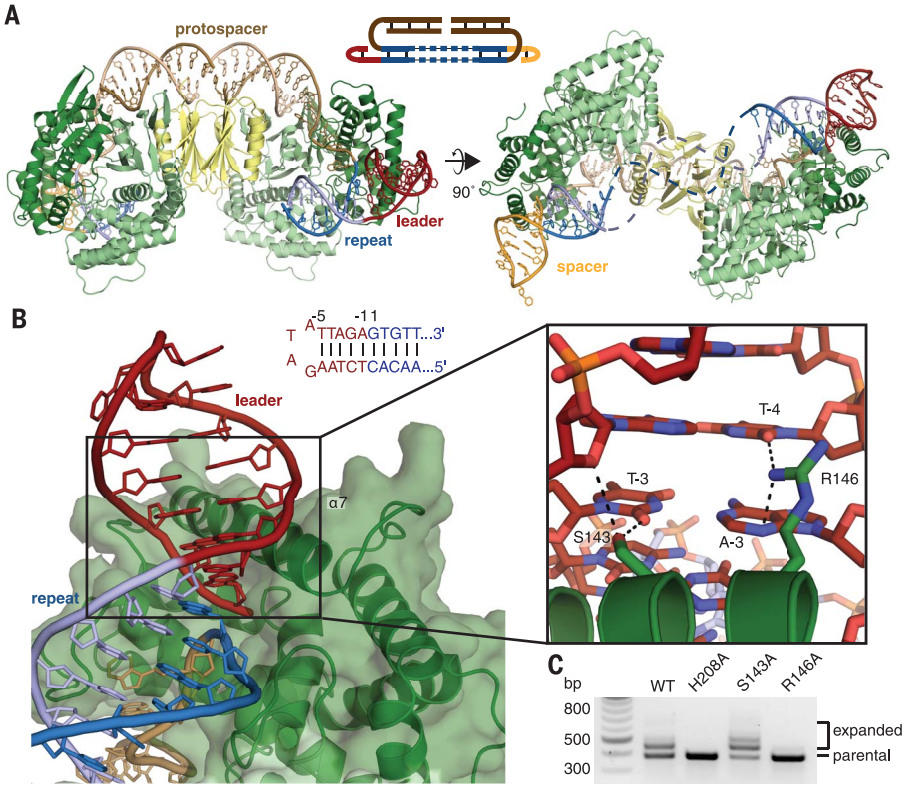

Fig. 2. Pseudo–full-site binding by Cas1-Cas2. (**A**) Overview of pseudo–full-site substrate binding by Cas1-Cas2. In the second view, the expected path of the disordered DNA is shown as dashed lines. A schematic of the substrate is shown above, with the disordered region as dashed lines. (**B**) A view of minor groove insertion by α-helix 7. Dotted lines in the close-up show polar contacts. The sequence of the leader-repeat junction and residue numbering are shown above. Residues are numbered such that the final residue of the leader is –1 and the first residue of the repeat is 1. (**C**) Agarose gel of a representative in vivo acquisition assay with indicated Cas1 mutants and wild-type Cas2. Acquisition results in expansion of the CRISPR array, which is visible as larger bands above the parental locus. The H208A active-site mutant is a negative control. bp, base pairs; WT, wild type.

Leader-sequence recognition in the pseudo–full-site structure

The pseudo–full-site–bound structure was solved at 2.9 Å and reveals more details of the interaction between Cas1 and the target DNA (table S1). The nucleotides at both the leader-adjacent and spacer-adjacent integration sites are clearly resolved, whereas the middle of the repeat is disordered, suggesting that the repeat disengages from Cas2 after full integration (Fig. 2A and fig. S3, C and D). Previous crystal structures have suggested that the Cas1 α-helix 7 might interact with target DNA, and we indeed observed insertion of this helix into the minor groove of both the leader and spacer regions of the target DNA (Fig. 2B) (*14*). The terminal residues of the leader sequence contribute to integration efficiency, and our structure reveals that several residues make hydrogen bonds with the minor-groove face of leader bases (*8*, *18*–*20*). Cas1 R146 hydrogen-bonds with A-3 and T-4 and is essential for integration in vivo, suggesting that it may also stabilize binding through interactions with the phosphate backbone (Fig. 2, B and C, and fig. S4B). Cas1 S143 interacts with T-3 of the nonintegrated target strand, although it is dispensable for in vivo activity (Fig. 2, B and C, and fig. S4B).

Integration requires DNA distortion

Both the half-site and the pseudo–full-site structures show substantial distortion of the target DNA. The DNA exhibits a sharp kink at both

[1]Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA 94720, USA. [2]Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [3]Biophysics Graduate Group, University of California, Berkeley, Berkeley, CA 94720, USA. [4]Howard Hughes Medical Institute, University of California, Berkeley, Berkeley, CA 94720, USA. [5]Department of Chemistry, University of California, Berkeley, Berkeley, CA 94720, USA. [6]Innovative Genomics Institute, University of California, Berkeley, Berkeley, CA 94720, USA. [7]Center for RNA Systems Biology, University of California, Berkeley, Berkeley, CA 94720, USA.
*These authors contributed equally to this work. **†Corresponding author. Email: doudna@berkeley.edu**

integration sites, with the bases on either side of the leader-repeat and repeat-spacer junction forming a nearly 30° angle (Fig. 3A). The repeat-spacer junction of the half-site substrate exhibits a similar kink, which indicates that the distortion occurs not as a result of integration but instead upon Cas1-Cas2 binding to the target. Binding across the Cas2 dimer surface also forces a bend in the repeat, which is mostly localized to the region directly over Cas2 (Fig. 3B).

Both structures show that the repeat must also undergo twist deformation to be properly positioned in both active sites. Modeling B-form DNA into the disordered regions of the repeat results in the incorrect backbone being positioned in the spacer-side active site (Fig. 3C). Connecting the resolved regions of DNA requires that the missing region be underwound by about one-third of a turn relative to canonical B-form DNA. It is unclear how this distortion is distributed across the disordered region, and the lack of order might indicate that the DNA adopts a range of conformations to accommodate the strain. The required bending and underwinding of the repeat, together with the lack of sequence-specific contacts in the repeat, suggest that Cas1-Cas2 recognizes the target through indirect readout based on the repeat's sequence-dependent deformability. In particular, the poly-G stretches in the inverted repeat motifs may facilitate the adoption of strained conformations to allow binding across both active sites (*21*, *22*).

To investigate whether these motifs are required for the DNA to be coordinated at opposing active sites, we performed in vitro integration assays using repeats with mutations known to prevent acquisition in vivo (Fig. 3D) (*9*). The mutations did not noticeably affect leader-side integration, but they prevented integration at the repeat-spacer junction. Half-site substrates bearing the same mutations were unable to be converted to full-site products, despite supporting binding and disintegration, whereas wild-type half-sites were readily converted to full-site products (Fig. 3E and fig. S5). These results confirm that the repeat sequence is important not for binding and recruitment of Cas1-Cas2 but instead for determining the ability of the target to reach the spacer-side active site.

To further investigate the importance of DNA deformation for spacer-side integration, we performed integration assays using targets with single- or double-base mismatches between the inverted repeats (Fig. 3F). We expected that the introduction of a mismatch would disrupt the DNA duplex and generate a flexible hinge in the middle of the repeat. Mismatches immediately before the second inverted repeat increased the rate of spacer-side integration, indicating that increasing the deformability of the repeat at specific sites enhances full-site integration. These data support the model that sequence-dependent distortion is necessary for recognition and integration at the repeat. Both G→C and G→A transitions in the inverted repeats prevented full-site integration, suggesting that the necessary deformation of the repeat depends on factors other than or in

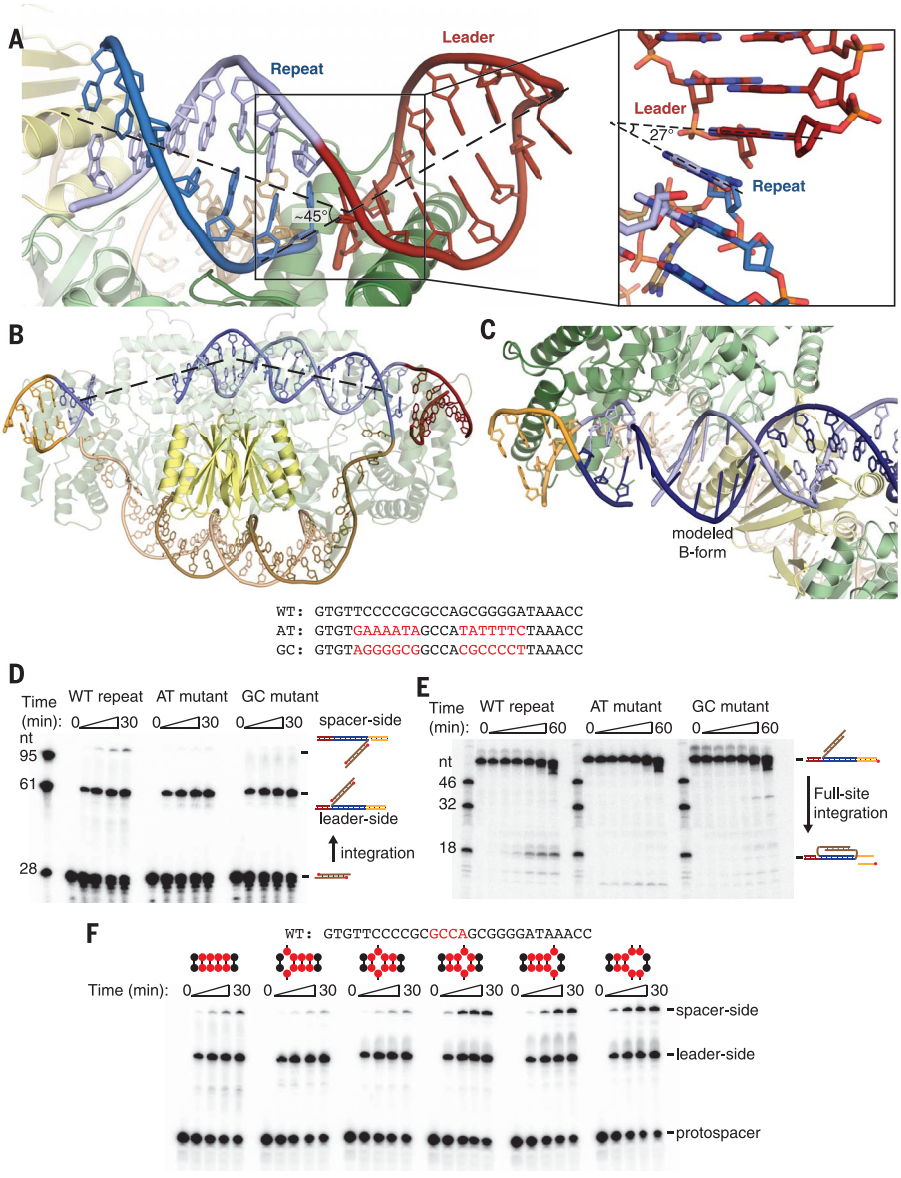

Fig. 3. Integration involves DNA distortion. (**A**) View of the kink introduced at the leader-repeat junction in the pseudo–full-site structure. The kink is highlighted with a dashed line showing the central axis of the DNA. The inset shows the bases before and after the integration site. Part of the backbone is omitted for clarity, and the angle formed by adjacent bases is shown with dashed lines. (**B**) Representation of the half-site repeat bending over the Cas2 dimer. The DNA trajectory is fit with a dashed line to show the localized bending. (**C**) Modeled B-form DNA fails to connect resolved regions of the half-site repeat. Modeled bases are shown with bases as sticks rather than rings. The (+) and (−) strands are shown in dark and light blue, respectively, to show that the modeled DNA does not properly join with the spacer-proximal DNA. (**D**) Urea-PAGE (polyacrylamide gel electrophoresis) gel of an integration assay with a radiolabeled protospacer. The substrate and expected products are shown as cartoons, with the radiolabel represented by a red circle. Their expected positions are indicated. The repeat sequences are shown above, with the mutated regions highlighted in red. Time points were taken at 0, 1, 5, 15, and 30 min. (**E**) Urea-PAGE gel of a second-site integration assay using mutant repeat sequences. The substrate and expected product are schematized, with the radiolabel represented by a red circle, and their expected positions are indicated on the gel. The mutant repeats are the same as in (D). Time points were taken at 0, 0.5, 1, 2, 10, 30, and 60 minutes. (**F**) Integration assay with a radiolabeled protospacer and mismatched repeats. Mismatches were introduced in the region of the repeat highlighted in red in the wild-type sequence above the gel. The positions of the mismatches are schematized above each time course, with red circles representing the highlighted midrepeat nucleotides. Time points were taken at 0, 1, 5, 15, and 30 min.

addition to GC content, such as specific purine-pyrimidine steps in the region where mismatches favor integration.

Active-site geometry

To better understand Cas1 active-site geometry, we grew pseudo–full-site–bound crystals in the presence of Ni^{2+}, which does not support catalysis but should allow for Mg^{2+}-like coordination geometry, and solved the structure to 3.3-Å resolution (fig. S6 and table S1). We observed density and peaks in the anomalous difference map for a single Ni^{2+} located at each of the four Cas1 active sites, although the metals are at lower occupancy in the substrate-engaged active sites, potentially because of lower solvent accessibility at these sites (Fig. 4A and fig. S7, A to D). At the noncatalytic active sites, the metal is coordinated by H208 and D221, as previously described (*14*, *23*). In the postintegration active sites, the phosphate of the newly formed phosphodiester bond bridging the protospacer and the repeat coordinates the metal, and the free 3′ OH of the cleaved leader or spacer is in close proximity. E141, which has been identified as a metal-coordinating residue, has poor side-chain density in all monomers and appears to be outside the range of a favorable interaction with the metal (fig. S7, E and F). The absolute requirement of E141 for activity suggests that it may play another role in catalysis, perhaps acting as a proton donor for the departing 3′ hydroxyl (*6*, *23*).

In vivo CRISPR integration assays to test the role of basic residues in the integrase that might contact either side of the DNA integration site showed that alanine mutants of Cas1 R132, R138, and R163 eliminate or nearly eliminate acquisition (Fig. 4, B and C). The R112A Cas1 mutant maintained some activity, but the R112E mutation prevented acquisition. The importance of all of these residues may reflect the need for a strong network of favorable contacts to capture the DNA in a strained conformation. To test this hypothesis, we performed disintegration and second-site integration assays with an R138A Cas1 mutant. This mutation reduced the rate of second-site integration by 50%, but R138A Cas1 exhibited wild type–like binding and enhanced disintegration activity, likely because of faster product release or the reduced rate of the competing forward reaction (Fig. 4D and fig. S8). These data confirm that R138 is dispensable for catalysis but important for trapping the DNA at the distal active site.

IHF sharply bends the integration locus and recruits an upstream binding site

To investigate the mechanism by which IHF recruits Cas1-Cas2 to the leader-proximal repeat, we purified the Cas1-Cas2 and IHF bound to a half-site substrate with an extended leader sequence (fig. S9). Negatively stained samples were used to generate an initial low-resolution reconstruction that showed additional density attached to the Cas1-Cas2 module that we could assign to IHF (fig. S10). We then used cryo–electron microscopy (cryo-EM) to solve the structure at a final resolution of 3.6 Å (figs. S11 to S13). We generated a complete model of the Cas1-Cas2–IHF-DNA holocomplex by first fitting the crystal structure of half-site–bound Cas1-Cas2 solved in this work and the published atomic model of the IHF module (Protein Data Bank ID, 1IHF) into the cryo-EM map, then manually rebuilding the

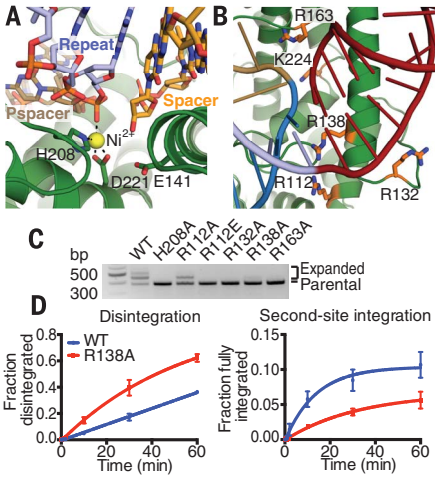

Fig. 4. Full-site integration requires a basic clamp around the active site. (**A**) Metal coordination in the spacer-side active site. Active-site residues, repeat, spacer, and protospacer (pspacer) are labeled, and coordination is shown as dotted lines. (**B**) View of basic residues surrounding the leader-repeat junction. Basic residues in close proximity to the target DNA backbone on either side of the integration site are shown as sticks and colored orange. (**C**) Agarose gel of an in vivo acquisition assay with indicated Cas1 mutants. H208A Cas1 is a negative control. (**D**) Quantification of disintegration and second-site integration time-course assays with wild-type and R138A Cas1. Means and standard deviations of three independent experiments are plotted. Representative gels are shown in fig. S8.

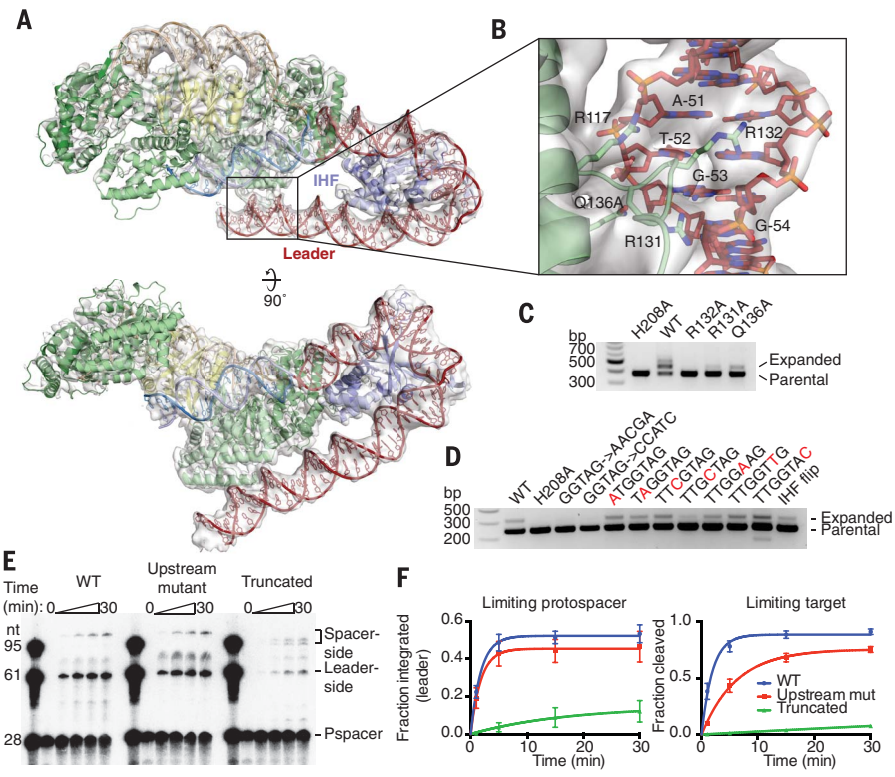

Fig. 5. Upstream sequence recognition by Cas1. (**A**) Cryo-EM structure of Cas1-Cas2 with IHF and extended leader. The atomic model is shown as a cartoon, and the electron density is shown as a transparent surface using an 8σ threshold. (**B**) View of upstream sequence readout by Cas1. Electron density is shown as in (A). Relevant Cas1 residues are labeled. Bases in the conserved recognition sequence are also labeled. (**C**) Acquisition assay with wild-type Cas2 and the indicated Cas1 mutants. H208A Cas1 is a negative control. (**D**) Acquisition assay with wild-type proteins and the noted mutations in the leader sequence. Single-nucleotide mutations in the conserved recognition region are highlighted in red. "IHF flip" denotes the leader sequence with the IHF-binding sequence reversed in place. H208A Cas1 is a negative control. (**E**) Integration assay with radiolabeled protospacer and targets with variable leaders. The upstream mutant substrate has the GGTAG→CCATC mutation in the conserved recognition motif; the truncated substrate begins at residue −46, after the recognition motif. Time points were taken at 0, 1, 5, 15, and 30 min. (**F**) Quantification of integration assays with limiting protospacer and limiting target. Means and standard deviations of three independent experiments are shown. A representative gel of the limiting target experiment is shown in fig. S14.

models to fit the density. The DNA substrates were manually built ab initio, and the resulting complete model was improved by real-space refinement (fig. S14).

In the holocomplex, Cas1-Cas2 and the repeat are overall in the same conformation as in the half-site crystal structure, and disorder of the spacer end of the complex again prevented building the DNA across to the distal active site. The structure shows how IHF binds the leader immediately upstream of Cas1-Cas2 and induces a 180° turn in the DNA, directing it back toward the Cas1-Cas2 complex (Fig. 5A) (*24*). The upstream binding motif interacts with one of the noncatalytic Cas1 protomers, with the loop between α6 and α7 inserting into the minor groove. R117 and Q136 interact with the phosphate backbone, and R131 and R132 are positioned to hydrogen-bond with the minor groove face of bases in the conserved recognition region (Fig. 5B). R132 is essential for integration in vivo, but it is difficult to assess the importance of its role in upstream readout, given that R132 on the catalytic Cas1 protomer is implicated in the basic clamp described above (Figs. 4B and 5C). R131 and Q136 also contribute to DNA binding: Alanine mutations of either reduce acquisition. Mutation of the conserved upstream sequence as a block eliminated acquisition, as previously noted, and single-nucleotide mutations revealed G-*53*, which is recognized by R131, as particularly important for recognition (Fig. 5D) (*12*).

To determine how much the IHF-dependent recruitment of Cas1-Cas2 depends on upstream sequence recognition, as opposed to nonspecific stabilizing interactions, we performed in vitro integration assays with targets containing leaders with mutations in the upstream binding region or leaders truncated before the upstream interaction region (Fig. 5E). Mutations in the binding site reduced the rate of leader-side integration by a factor of 3 when the target was limiting (Fig. 5F and fig. S15). The rate effect is masked when the target is in excess over the protospacer-bound complex, but a higher level of off-target integration is observed (fig. S15). The increased importance of the upstream sequence for in vivo acquisition suggests that it may be important for initial identification of the target in the context of genomic DNA, whereas it is dispensable when the correct target is saturating and no competitor is present. Truncation of the leader had a much more consequential effect, with the rate of leader-side integration reduced by a factor of ~100 when the target was limiting (Fig. 5F). Spacer-side integration was also affected by the truncation, as indicated by the appearance of a second band consistent with misplaced integration within the repeat (Fig. 5E). These results show that nonspecific interactions with the leader DNA are critical for robust Cas1-Cas2 activity and specificity, whereas the sequence-specific interactions aid in efficient recognition.

Suppression of off-target integration by IHF

We also investigated whether IHF contributes to Cas1-Cas2 recruitment by mechanisms other than juxtaposition of the upstream binding site. Our structure reveals that Cas1 and the α-protomer of IHF (IHF-α) are in close proximity, with a solvent-inaccessible surface of 200 Å^2 between the two proteins (Fig. 6A). However, there is little continuous electron density between the proteins. Mutations of IHF-α residues near the interface with Cas1 identified E10 and D14 as important for acquisition (Fig. 6B). These residues might interact favorably with Cas1 R131 or R132 to aid in Cas1 recruitment. However, reversing the orientation of the IHF-binding site in the leader, which should position IHF-β rather than IHF-α to interact with Cas1, did not dramatically affect acquisition, suggesting that any interaction that occurs is not highly specific (Fig. 5D).

To further investigate the role of IHF, we performed integration assays with and without IHF, using a truncated leader to prevent contribution from upstream interactions (Fig. 6C). In the absence of IHF, off-target integration occurs in the leader, indicating a role for IHF in limiting spurious integration events. Shifting the IHF-binding site one to five nucleotides farther away from the leader-repeat junction led to a modest decrease in the efficiency of leader-side integration, although the site of integration was unaltered (Fig. 6, C and D). This supports the model that contacts between IHF and Cas1 contribute to specific and efficient CRISPR locus expansion, although recruitment of the upstream binding site appears to be the more important contribution.

Conclusions

These data show that the type I Cas1-Cas2 from *E. coli* relies heavily on active-site positioning and structural features of the DNA, rather than direct sequence recognition, to localize DNA integration to the CRISPR locus (fig. S16). The ability of the DNA substrate duplex to access both Cas1 active sites regulates recognition of the CRISPR repeat, with the GC-rich inverted repeats allowing for twist deformation and the midrepeat sequence acting as a hinge, and IHF aids in recruitment at the leader by providing a secondary binding surface for the complex. The lack of direct sequence recognition might reflect the evolutionary origins of Cas1 as a more promiscuous transposase (*25*–*27*). Bending of the DNA target site is a common feature in transposases and integrases, where it disfavors the disintegration reaction by ejecting DNA from the integrase active sites once integration is achieved (*28*, *29*). Although Cas1-Cas2 may use a similar mechanism, as suggested by the displacement of the midrepeat upon full-site integration, CRISPR systems appear to have exploited the requirement for DNA bending to provide sequence specificity for the integration reaction. The role played by

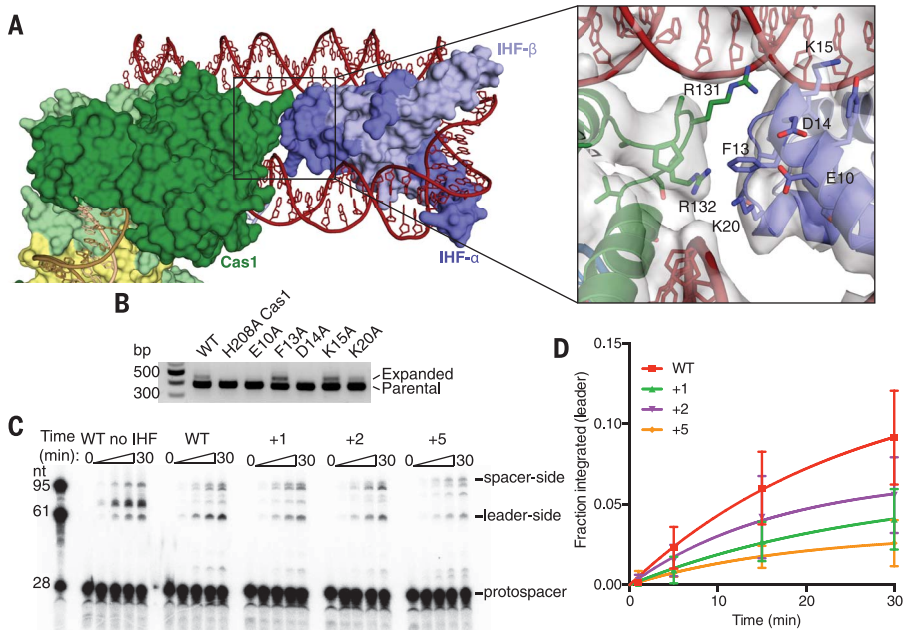

Fig. 6. Interactions between Cas1 and IHF. (**A**) Surface and cartoon representations of the interface between Cas1 and IHF-α. In the inset, residues at the interaction surface are shown as sticks, and residues of interest are labeled. Electron density is shown as a surface with an 8σ threshold. (**B**) Acquisition assay with wild-type Cas1 and Cas2 and the indicated IHF-α mutants. H208A Cas1 is a negative control. (**C**) Integration assays with radiolabeled protospacer and targets with truncated leaders. IHF is included unless otherwise noted. Mutant substrates have 1, 2, or 5 base pairs inserted between the IHF recognition sequence and the Cas1 recognition sequence of the leader. Time points were taken at 0, 1, 5, 15, and 30 min. (**D**) Quantification of leader-side integration with radiolabeled protospacer and truncated targets. Means and standard deviations of three independent replicates are shown.

IHF also represents an unexpected variation on a feature sometimes seen in transposases. In both λ and μ phage mobilization pathways, IHF or the related protein HU is involved in bringing recognition sequences on the viral DNA into contact with the integrase (*29*, *30*). In the phage pathways, IHF aids in the recognition of donor DNA, whereas in CRISPR acquisition, it is important for recognition of the target DNA, highlighting the shift in substrate selectivity from donor to target that was essential for the "domestication" of Cas1 for use in immunity (*25*, *26*).

The distinctive substrate preferences of the CRISPR integrase could make it useful as a molecular recording device for barcoding genomes or generating locus-specific sequence insertions (*31*). Bacterial transposases such as Tn*5* and MuA are robust tools for DNA tagging, insertion, and deletion, but they are promiscuous in their target selection and require sequence-specific interactions with the donor DNA that limit their use in some systems (*32*–*34*). Although the CRISPR integrase shares the reaction chemistry of other transposases, its substrate sequence independence, coupled with its selectivity for target DNA sequences, may enable a complementary set of applications. The architecture of the CRISPR integration complexes presented here suggests that subtle adjustment of the distance between Cas1 active sites could reprogram the CRISPR integrase to recognize different integration target sites. Changes in integrase architecture could thereby be exploited for genome tagging applications and may also explain the natural divergence of CRISPR arrays in bacteria.

REFERENCES AND NOTES

1. F. J. M. Mojica, C. Díez-Villaseñor, J. García-Martínez, E. Soria, *J. Mol. Evol.* **60**, 174–182 (2005).
2. A. Bolotin, B. Quinquis, A. Sorokin, S. D. Ehrlich, *Microbiology* **151**, 2551–2561 (2005).
3. C. Pourcel, G. Salvignol, G. Vergnaud, *Microbiology* **151**, 653–663 (2005).
4. R. Barrangou *et al.*, *Science* **315**, 1709–1712 (2007).
5. I. Yosef, M. G. Goren, U. Qimron, *Nucleic Acids Res.* **40**, 5569–5576 (2012).
6. J. K. Nuñez *et al.*, *Nat. Struct. Mol. Biol.* **21**, 528–534 (2014).
7. J. K. Nuñez, A. S. Y. Lee, A. Engelman, J. A. Doudna, *Nature* **519**, 193–198 (2015).
8. C. Rollie, S. Schneider, A. S. Brinkmann, E. L. Bolt, M. F. White, *eLife* **4**, e08716 (2015).
9. M. G. Goren *et al.*, *Cell Rep.* **16**, 2811–2818 (2016).
10. J. K. Nuñez, L. Bai, L. B. Harrington, T. L. Hinder, J. A. Doudna, *Mol. Cell* **62**, 824–833 (2016).
11. C. Moch, M. Fromant, S. Blanquet, P. Plateau, *Nucleic Acids Res.* **45**, 2714–2723 (2017).
12. K. N. R. Yoganand, R. Sivathanu, S. Nimkar, B. Anand, *Nucleic Acids Res.* **45**, 367–381 (2017).
13. R. D. Fagerlund *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **114**, E5122–E5128 (2017).
14. J. K. Nuñez, L. B. Harrington, P. J. Kranzusch, A. N. Engelman, J. A. Doudna, *Nature* **527**, 535–538 (2015).
15. J. Wang *et al.*, *Cell* **163**, 840–853 (2015).
16. M. Babu *et al.*, *Mol. Microbiol.* **79**, 484–502 (2011).
17. Z. Arslan, V. Hermanns, R. Wurm, R. Wagner, Ü. Pul, *Nucleic Acids Res.* **42**, 7884–7893 (2014).
18. R. Wang, M. Li, L. Gong, S. Hu, H. Xiang, *Nucleic Acids Res.* **44**, 4266–4277 (2016).
19. J. McGinn, L. A. Marraffini, *Mol. Cell* **64**, 616–623 (2016).
20. A. V. Wright, J. A. Doudna, *Nat. Struct. Mol. Biol.* **23**, 876–883 (2016).
21. W. K. Olson, A. A. Gorin, X. J. Lu, L. M. Hock, V. B. Zhurkin, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 11163–11168 (1998).
22. E. J. Gardiner, C. A. Hunter, M. J. Packer, D. S. Palmer, P. Willett, *J. Mol. Biol.* **332**, 1025–1035 (2003).
23. B. Wiedenheft *et al.*, *Structure* **17**, 904–912 (2009).
24. P. A. Rice, S. Yang, K. Mizuuchi, H. A. Nash, *Cell* **87**, 1295–1306 (1996).
25. P. Béguin, N. Charpin, E. V. Koonin, P. Forterre, M. Krupovic, *Nucleic Acids Res.* **44**, 10367–10376 (2016).
26. M. Krupovic, K. S. Makarova, P. Forterre, D. Prangishvili, E. V. Koonin, *BMC Biol.* **12**, 36 (2014).
27. A. B. Hickman, F. Dyda, *Nucleic Acids Res.* **43**, 10576–10587 (2015).
28. G. N. Maertens, S. Hare, P. Cherepanov, *Nature* **468**, 326–329 (2010).
29. S. P. Montaño, Y. Z. Pigli, P. A. Rice, *Nature* **491**, 413–417 (2012).
30. G. Laxmikanthan *et al.*, *eLife* **5**, 1–23 (2016).
31. S. L. Shipman, J. Nivala, J. D. Macklis, G. M. Church, *Science* **353**, aaf1175 (2016).
32. I. Y. Goryshin, J. Jendrisak, L. M. Hoffman, R. Meis, W. S. Reznikoff, *Nat. Biotechnol.* **18**, 97–100 (2000).
33. D. C. Nadler, S.-A. Morgan, A. Flamholz, K. E. Kortright, D. F. Savage, *Nat. Commun.* **7**, 12266 (2016).
34. A. Adey, J. Shendure, *Genome Res.* **22**, 1139–1143 (2012).

ACKNOWLEDGMENTS

We thank G. Meigs and the staff of the Advanced Light Source 8.3.1 beamline and D. Tzanko, A. Gonzalez, and the staff of the Stanford Synchrotron Radiation Light Source 9-2 beamline for assistance with data collection. Beamline 8.3.1 at the Advanced Light Source is operated by the University of California Office of the President, Multicampus Research Programs and Initiatives (grant MR-15-328599), and the Program for Breakthrough Biomedical Research, which is partially funded by the Sandler Foundation. Use of the Stanford Synchrotron Radiation Lightsource (SSRL), a directorate of SLAC National Accelerator Laboratory, is supported by the U.S. Department of Energy (DOE), Office of Science, Office of Basic Energy Sciences, under contract no. DE-AC02-76SF00515. The EM data were collected in the Howard Hughes Medical Institute EM facility located at the University of California, Berkeley. We thank D. B. Toso and P. Grob for expert EM assistance, A. Chintangal for computational support, and members of the Nogales laboratory for helpful discussions about EM data processing. We thank A. East-Seletsky for input on the manuscript. The SSRL Structural Molecular Biology Program is supported by the U.S. DOE, Office of Biological and Environmental Research, and the National Institutes of Health, National Institute of General Medical Sciences (including grant no. P41GM103393). This project was funded by U.S. National Science Foundation (NSF) grant no. 1244557 (to J.A.D.) and National Institute of General Medical Sciences grant no. 1P50GM102706-01 (to J. H. Cate). A.V.W. and K.W.D. are supported by a U.S. NSF Graduate Research Fellowship, and G.J.K. is funded by the Howard Hughes Medical Institute. J.A.D. and E.N. are investigators of the Howard Hughes Medical Institute and members of the Center for RNA Systems Biology. Atomic coordinates and structure factors for the reported crystal structures have been deposited in the Protein Data Bank under accession codes 5VVJ (half-site–bound), 5VVK (pseudo–full-site–bound), and 5VVL (pseudo–full-site–bound with Ni^{2+}). The cryo-EM structure and map have been deposited in the Protein Data Bank under accession code 5WFE and the Electron Microscopy Data Bank under accession code EMD-8827. A patent has been filed by the University of California for the use of Cas1-Cas2 for integrating DNA into genomes. J.A.D. is a cofounder and Scientific Advisory Board member of Caribou Biosciences and Intellia Therapeutics and a cofounder of Editas Medicine, all of which develop CRISPR-based technologies.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/357/6356/1113/suppl/DC1
Materials and Methods
Figs. S1 to S16
Tables S1 to S3
References (*35*–*51*)

10 June 2017; accepted 13 July 2017
Published online 20 July 2017
10.1126/science.aao0679

Structures of the CRISPR genome integration complex

Addison V. Wright, Jun-Jie Liu, Gavin J. Knott, Kevin W. Doxzen, Eva Nogales and Jennifer A. Doudna

Science **357** (6356), 1113-1118.
DOI: 10.1126/science.aao0679originally published online July 20, 2017

Host factor drives the big bend

Bacteria have a highly adaptable DNA-detecting and -editing machine called CRISPR-Cas to ward off virus attack. The Cas1-Cas2 integrase, with the help of an accessory protein called IHF (integration host factor), captures foreign DNA motifs into bacterial CRISPR loci. These motifs then act as sensors of any further invaders. By analyzing the integrase complex structure, Wright *et al.* show how Cas1-Cas2 recognizes the CRISPR array for site-specific integration (see the Perspective by Globus and Qimron). IHF sharply bends DNA, which allows DNA to access two active sites within the integrase complex to ensure sequence specificity for the integration reaction. The features of the CRISPR integrase complex may explain the natural divergence of CRISPR arrays in bacteria and can be exploited for genome-tagging applications.

Science, this issue p. 1113; see also p. 1096