Science
AAAS

Supplementary Materials for

Structures of the CRISPR genome integration complex

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

Affiliations:

[1]Department of Molecular and Cell Biology

[2]Biophysics Graduate Group

[3]Department of Chemistry

[4]Innovative Genomics Institute

[5]Center for RNA Systems Biology

[6]Howard Hughes Medical Institute

University of California, Berkeley, Berkeley, California 94720, USA.

[7]Molecular Biophysics & Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.

*Correspondence to: doudna@berkeley.edu

†These authors contributed equally to this work

This PDF file includes:

Materials and Methods
Figs. S1 to S15
Tables S1 to S3

Materials and Methods

Protein and DNA preparation

Cas1 and Cas2 proteins from *E. coli* K12 (MG1655) were individually purified as previously described (*6*). IHF from *E. coli* K12 (MG1655) was purified as a heterodimer as previously described (*10*). DNA oligonucleotides were purchased from Integrated DNA Technologies or Dharmacon and were purified using urea-PAGE. DNA substrates for crystallography were prepared by mixing the appropriate ssDNA oligonucleotides in 20 mM HEPES-NaOH, pH 7.5, 25 mM KCl, 10 mM $MgCl_2$, incubating at 95°C for 5 minutes, slow-cooling to room temperature, and purifying over an 8% native polyacrylamide gel. Radiolabeled substrates were prepared by labeling with T4 polynucleotide kinase (New England Biolabs) and [γ-^{32}P]-ATP (Perkin Elmer) and annealing with a two-fold excess of the unlabeled strands, with the exception of radiolabeled protospacers, which were annealed and purified using native PAGE prior to radiolabeling the duplex. Substrates used for hydroxyl radical footprinting were further purified on an 8% native polyacrylamide gel. Sequences for all substrates are shown in Table S3.

Complex formation, crystallization, and data collection

Purified Cas1 and Cas2 were incubated at 50 μM each (monomer concentration) in Complex Buffer (10 mM HEPES-NaOH, pH 7.5, 150 mM KCl, 10 mM EDTA, 1 mM DTT) at room temperature for 1 hour while dialyzing against Complex Buffer. DNA substrates were also dialyzed against Complex Buffer and added to the Cas1-Cas2 complex to a final substrate concentration of 10 μM, such that Cas1-Cas2 complex was in 1.25-fold excess. Cas1-Cas2 was incubated with the target DNA for 30 minutes before purifying over a Superdex 200 Increase 10/300 column (GE Healthcare). For the half-site-bound complex, the major peak was collected and concentrated to an A_{280} of 9.0 AU as measured by Nanodrop. Crystals were initially grown by hanging drop diffusion at 16°C in drops containing 100 mM MES, pH 6.5, 10% (w/v) poly(ethylene glycol) (PEG) monomethyl ether (MME) 5000, and 12% (v/v) propanol. The resulting crystals were used to microseed drops containing equal volumes of protein-DNA complex at A_{280}=7.0 AU and solution containing 100 mM MES, pH 6.5, 8% (w/v) PEG MME 5000, and 12% (v/v) propanol. The final crystals were cryoprotected in reservoir solution supplemented with 15% (v/v) glycerol. For the pseudo-full-site-bound complex, the second major peak was collected and concentrated to an A_{280} of 4.0 AU. Crystals were grown by sitting drop diffusion at 16°C in a solution containing 100 mM MES pH 6.4, 20% (w/v) PEG MME 2000, and 0.2 M NaCl and cryoprotected in reservoir solution supplemented with 15% (v/v) glycerol. For crystals grown in the presence of Ni^{2+}, Complex Buffer with 1 mM EDTA was used and the reservoir solution described above was supplemented with 3 mM $NiCl_2$. Crystals were soaked in reservoir solution with 10 mM $NiCl_2$ and 15% glycerol for 5 minutes prior to flash freezing.

X-ray diffraction data for the half-site and pseudo-full-site were collected under cryogenic conditions at beamline 8.3.1 at the Lawrence Berkeley National Laboratory Advanced Light Source with a wavelength of 1.1158 Å. Native X-ray diffraction data for the pseudo-full-site with nickel was collected under cryogenic conditions at beamline 9-2 at the Stanford Synchrotron Radiation Lightsource with a wavelength of 0.9795 Å. All data were collected with a Pilatus3 S 6M detector (Dectris). Anomalous data were collected from crystals grown in the presence of Ni^{2+} at 8345.7 eV, an energy between the inflection point and peak anomalous energies. All data were indexed in *XDS* and scaled in *XSCALE* before merging in *AIMLESS* (*35*, *36*). Resolution cut-offs were determined using correlation-coefficient threshold of 0.5 (*37*).

Negative-staining EM microscopy and image processing

Cas1-Cas2-DNA-IHF complexes were assembled by co-incubating Cas1 and Cas2 at 50 µM each in buffer containing 20 mM HEPES, pH 7.5, 150 mM KCl, 5 mM EDTA, and 1 mM DTT. IHF and half-site DNA were incubated in the same buffer at 20 µM and 10 µM, respectively. After an hour, equal volumes Cas1-Cas2 and IHF-DNA were combined, such that Cas1-Cas2 complex was in 1.25-fold excess over DNA, and allowed to complex for 30 minutes before purifying over a Superose 6 10/300 column (GE Healthcare). The complexes were diluted to a final concentration of 50~80 nM and negatively stained in a 2% (w/v) solution of uranyl acetate (Electron Microscopy Sciences) following the standard deep-staining procedure on glow-discharged holey carbon-coated EM copper grids covered with a thin layer of continuous carbon. The negatively stained specimen was then mounted onto a transmission electron microscope holder and examined by an FEI Tecnai Spirit electron microscope operated at 120-kV. Magnified digital micrographs of the specimen were automatically taken at a nominal magnification of 80,000 on a Gatan Ultrascan 4000 CCD camera with a pixel size of 1.5 Angstroms at the specimen level within Leginon. The defocus values used were about -1.0 to -1.8 µm, and the total accumulated dose at the specimen was about 60 electrons per Å^2. The particles were automatically picked, CTF corrected and then 2D-classified without reference in Appion (*38*). 10 good 2D classaverages were imported into EMAN2 for generating the initial 3D model based on common line method (*39*). Good particles sorted by the 2D classification were futher refined against the initial model with SPIDER (*40*).

Cryo-EM microscopy

Cas1-Cas2-DNA-IHF complexes in a buffer containing 20 mM HEPES, pH 7.5, 150 mM KCl, 5 mM EDTA, 1 mM DTT, and 0.1% glycerol were used for cryo-EM sample preparation. Immediately after glow-discharging the grid for 14 seconds using a Solaris plasma cleaner, 3.6 µl droplets of the sample (~1µM) were placed onto C-flat grids with 2 µm holes and 2 µm spacing between holes (Protochips Inc.). The grids were rapidly plunged into liquid ethane using an FEI Vitrobot MarkIV maintained at 8 °C and 100% humidity, after being blotted for 4.5 seconds with a blot force of 12. Data were acquired using an FEI Titan Krios transmission electron microscope operated at 300 keV, at a nominal magnification of ×24,500 (1.07 Å pixel size), and with defocus ranging from −1.2 to −2.8 µm. A total of ~3,000 micrographs were recorded using SerialEM on a Gatan K2 Summit direct electron detector operated in super-resolution mode (*41*). We collected a 6.0s exposure fractionated into 30, 200 ms frames with a dose of 6.8 e- $Å^{-2}s^{-1}$.

Images processing and reconstruction for Cryo-EM

The 28 frames (we skipped the first 2 frames) of each image stack in super-resolution model were aligned, decimated, and summed and dose-weighted using Motioncor2 (*42*). CTF values of the summed-micrographs were determined using CTFFIND4 and then applied to dose-weighted summed-micrographs for further processing (*43*). Initial particle picking to generate template images was performed using EMAN2. About 20,000 particles were stacked and then imported into Relion2.0 for reference-free 2D classification (*44*). Particle picking for the complete dataset was carried out using Gautomatch (http://www.mrc-lmb.cam.ac.uk/kzhang/) with templates generated in previous 2D classification. About 650,000 good particles were selected in total. Due to the preferred orientation issue, random half of the particles in preferred orientations were thrown away, and then only 410,000 particles were left for further processing. Using the 3D model got from negative staining and low-pass filtered to 60Å as a reference, we performed 3D classification using RELION2.0. 3D refinements of the 2 best classes were performed in Cryosparc by importing the 3d models generated from 3D classification (*45*). The local resolution was calculated by Relion2.0. The reported resolution was based on the gold standard FSC criterion using two independent half-maps. The model resolution in different orientations was calculated using two independent half-maps with the ThreeDFSC script shared by Philip Baldwin (https://github.com/nysbc/Anisotropy).

Model building and refinement

Initial phases for the half-site and pseudo-full-site crystal structures were calculated by molecular replacement with the protospacer-bound Cas1-Cas2 complex (Protein Data Bank accession number 5DS5) in *PHASER* (*46*). The low resolution of the half-site-bound structure and the disorder of the DNA in particular precluded confident placement of individual nucleotides. DNA from the leader-repeat and repeat-spacer junction generated from the pseudo-full-site structure were used to generate initial models at the corresponding density in the half-site structure. Regular B-form DNA was used as an initial model for the early and mid-repeat regions of the half-site DNA and modified to fit the trajectory and helical pitch of the visible density. The structures were completed through iterative model-building in *COOT* and refinement in *PHENIX* (*47*, *48*). The pseudo-full-site structure was refined using NCS and reference-model restraints until the final rounds of refinement. For the lower resolution half-site structure, refinement was carried out using reference-model, NCS, and secondary structure restraints. Anomalous difference maps to identify Ni^{2+} sites were generated using data truncated to 6.2 Å. Web 3DNA was used to analyze structural parameters of the DNA (*49*).

To generate a complete model for the cryo-EM map, the crystal structure of half-site-bound Cas1-Cas2 solved in this work and published atomic model of IHF module (PDB accession code 1IHF) were first fitted into the refined 3D-reconstruction map using UCSF Chimera (*18*) and then manually rebuilt in Coot to fit the density. The DNA substrates were manually built ab initio in Coot based on the EM density. To improve backbone geometry, the atomic model of Cas1-Cas2-IHF-DNA model was subjected to PHENIX real space refinement (global minimization and ADP refinement) with Ramachandran, rotamer, and nucleic-acid restraints. The final model was validated using Molprobity (*50*). Structural analysis was performed in Coot and figures were prepared using PyMOL (Schrodinger LLC) and UCSF Chimera. Data collection and refinement statistics are in Table S2.

Electrophoretic mobility shift assays

Cas1 and Cas2 (or Cas1 alone, where indicated) were co-incubated in equimolar concentrations in EMSA buffer (20 mM HEPES pH 7.5, 50 mM KCl, 10 mM EDTA, 0.01% Tween, 100 μg/mL heparin, 100 μg/mL BSA, 5% glycerol, 1 mM DTT) on ice for 30 minutes. Reported concentrations are that of Cas1 and Cas2 monomers. The appropriate radiolabeled DNA substrate was added to a final concentration of <0.2 nM. Binding was carried out at room temperature for one hour, and bound and unbound species were separated on a 5% native polyacrylamide gel in 0.5X TBE. The gel was dried and visualized with phosphorimaging. Bands were quantified with ImageQuant (GE Healthcare) and analyzed with Prism using a single-site saturation binding model (GraphPad). Only bands corresponding to the intact unbound substrate and the full-complex-bound substrate were used for quantification.

Hydroxyl radical footprinting

Cas1 and Cas2 were coincubated in buffer containing 20 mM HEPES pH 7.5, 50 mM KCl, 10 mM EDTA, and 100 μg/mL BSA on ice for 30 minutes. Cas1-Cas2 was added to 1 nM DNA at a final concentration of 0, 10, 100, or 1000 nM and allowed to bind at room temperature for one hour. Hydroxyl radical cleavage was carried out as previously described, except that additional EDTA was not added (*51*). The DNA was ethanol precipitated and resuspended in loading buffer containing 95% formamide and 10 mM EDTA, incubated at 95° for 5 minutes, and resolved on a 12% denaturing polyacrylamide gel. The gel was dried and visualized with phosphorimaging.

***In vivo* acquisition assays**

In vivo acquisition assays using Cas1 or Cas2 mutants were performed as previously described(*6*). Assays involving mutations in the leader region were performed using our pCDF-Cas1-Cas2 expression plasmid with the leader and single repeat from the BL21 CRISPR-I locus cloned into the XbaI site. Amplification was performed with primers specific to the plasmid-based locus. Complete leader sequences are shown in Table S3. Assays using IHF-α point mutants were performed in a IHF-α knockout strain with the mutant IHF-α and wild-type IHF-β expressed off a plasmid, as previously described (*10*). Both IHF mutant assays and leader mutant assays were grown under induction for two 24-hour cycles before analysis to allow for higher levels of integration (*10*).

Integration, second-site integration, and disintegration assays

Integration assays with plasmid target were performed largely as previously described (*7*). pCRISPR was used as a target, Cas1-Cas2 were at 100 nM, protospacer at 100 nM, and plasmid at 7.5 nM, and the reaction was carried out for one hour. Metal was omitted from the reaction buffer or added at 10 mM where indicated. Integration assays with radiolabeled protospacer were performed as previously described, except that protospacer concentration was changed to 10 nM and target concentration was 100 nM, and reactions were carried out at room temperature (*10*). IHF was included at 200 nM unless otherwise noted. Integration assays with radiolabeled leaders were carried out with 200 nM Cas1-Cas2, 100 nM unlabeled protospacer, 50 nM IHF, and 10 nM labeled target. Reactions were carried out at room temperature. Samples were run on a 12% denaturing polyacrylamide gel. Progression of half-site substrates to full-site and disintegration assays were both performed as previously described for Cas1-Cas2 from *Streptococcus pyogenes*, with 100 nM protein and 1 nM radiolabeled DNA substrate (*20*). Reactions were performed at room temperature, time points were taken at 0.5, 1, 2, 10, 30, and 60 minutes, and samples were run on a 12% denaturing polyacrylamide gel. Gels was dried and visualized with phosphorimaging. Bands were quantified with ImageQuant (GraphPad) and data were analyzed with Prism and fit with a one-phase association model (GraphPad).

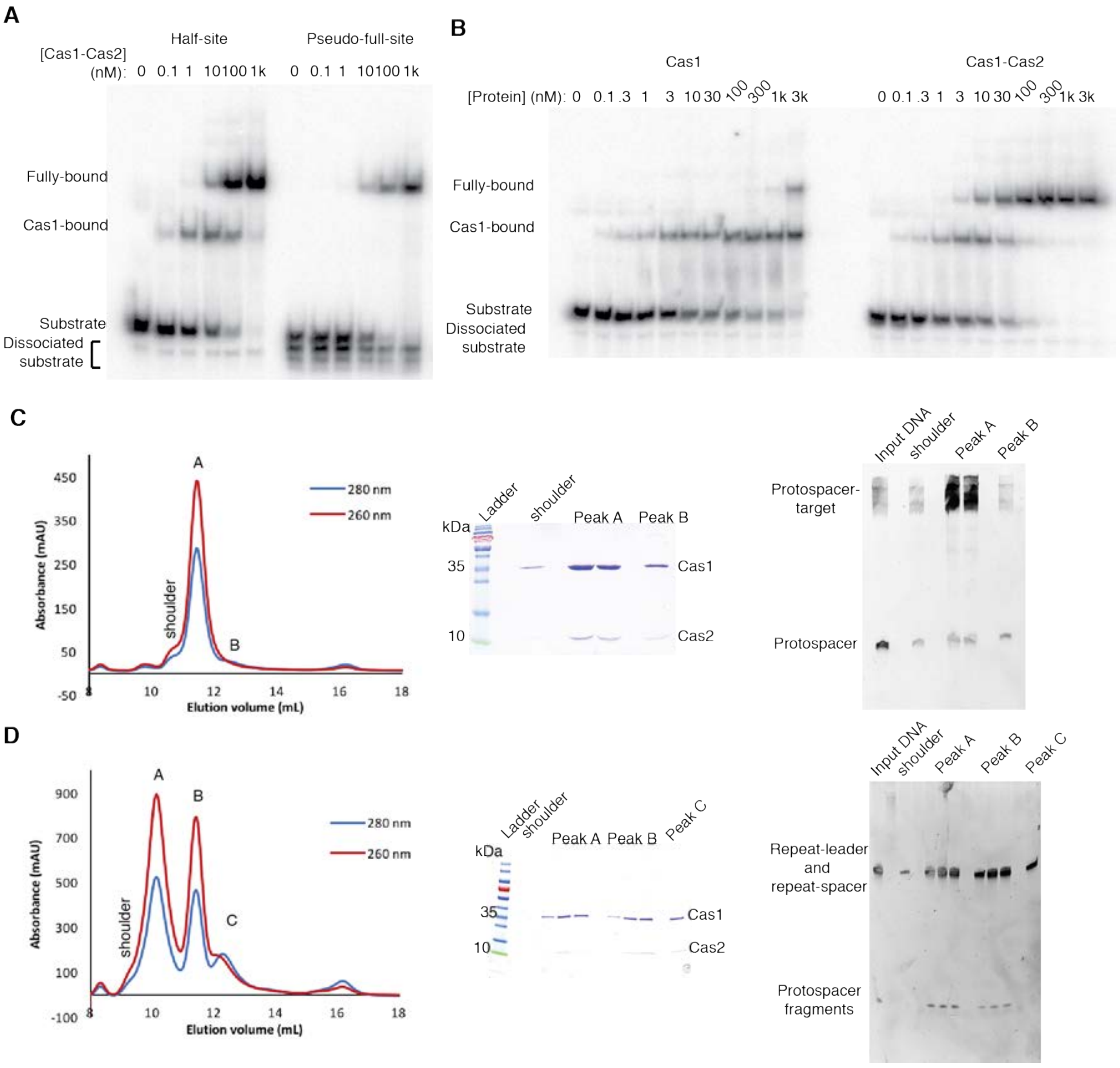

Fig. S1. Half-site and pseudo-full-site binding by Cas1 and Cas1-Cas2.
(A) EMSA of radiolabeled half-site and pseudo-full-site substrates with Cas1-Cas2. Protein concentrations are indicated, and bands are labeled. **(B)** EMSA of half-site substrate by either Cas1 alone or Cas1-Cas2. The intermediate band observed in half-site binding appears to result from binding of free Cas1 dimer. **(C)** Purification of half-site-bound complex by size exclusion chromatography. A representative S200 size exclusion trace is shown. Samples were taken from the labeled peaks and analyzed on SDS-PAGE with Coomassie brilliant blue and urea-PAGE with SybrGold. The smear for the protospacer-target DNA strand results from partial renaturation of the hairpin structure. Peak A was used for crystallography. **(D)** Purification of pseudo-full-site-bound complex by size exclusion chromatography, with gels as described for (C). Peak B was used for crystallography

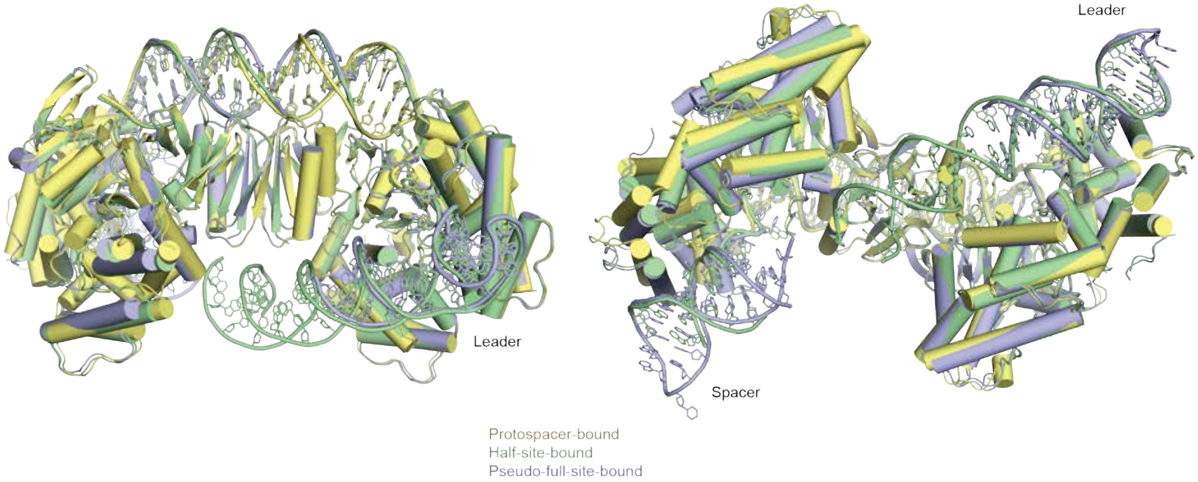

Fig. S2. Superposition of protospacer-bound, half-site-bound, and pseudo-full-site-bound Cas1-Cas2.

Structural alignment of our target-bound structures with a previously-solved protospacer-bound structure (PDB code 5DS5). Alignments were made using the Cas2 dimer as the reference. The protospacer-bound structure is shown in yellow, half-site bound in green, and pseudo-full-site-bound in blue. Only modest structural rearrangements occur upon target binding, with the spacer-side Cas1 dimer (left side) rotating slightly to position the active site closer to the central channel.

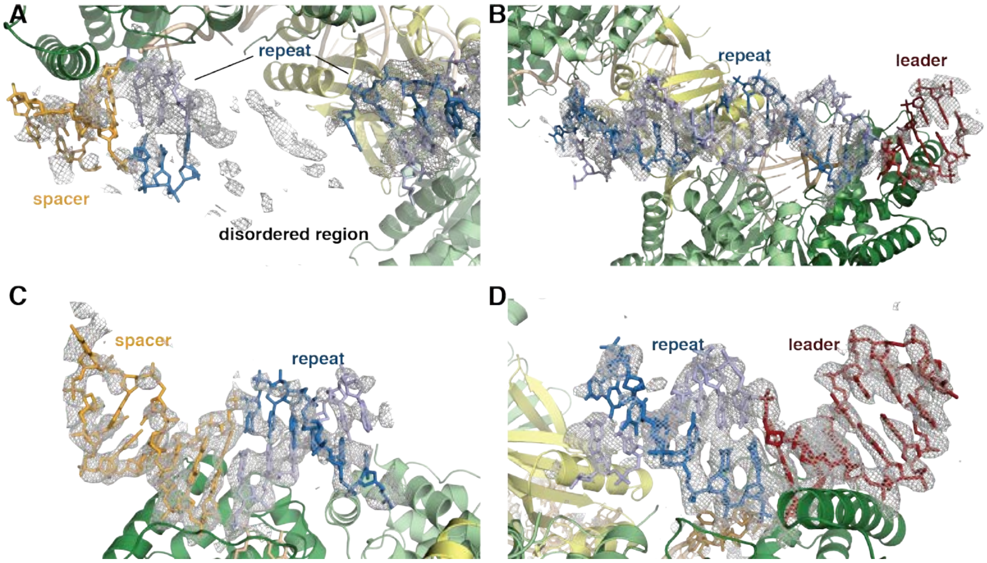

Fig. S3. Simulated annealing omit maps for Cas1-Cas2 bound to half-site and pseudo-full-site DNA.

(A) F_o-F_c omit map for the entire target DNA using half-site map and model, showing the leader and early to mid-repeat DNA. **(B)** F_o-F_c omit map showing the repeat-spacer junction and the unresolved region. **(C)** F_o-F_c omit map of the target DNA using the pseudo-full-site map and model. Leader-repeat region is shown. **(D)** F_o-F_c omit map showing the spacer-repeat region. Maps are contoured at 2.0 σ.

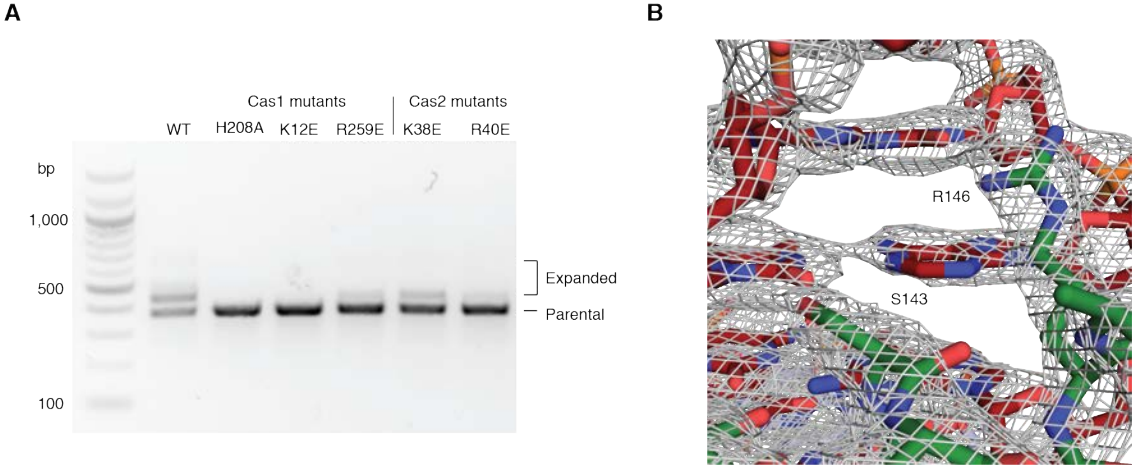

Fig. S4. Residues involved in mid-repeat and leader interactions.
(A) Agarose gel of *in vivo* acquisition assays performed with the indicated Cas1 or Cas2 mutants. Cas1 H208A is used as a negative control. **(B)** Feature-enhanced map of the leader and interacting residues. Map is shown as mesh at 2.0 σ.

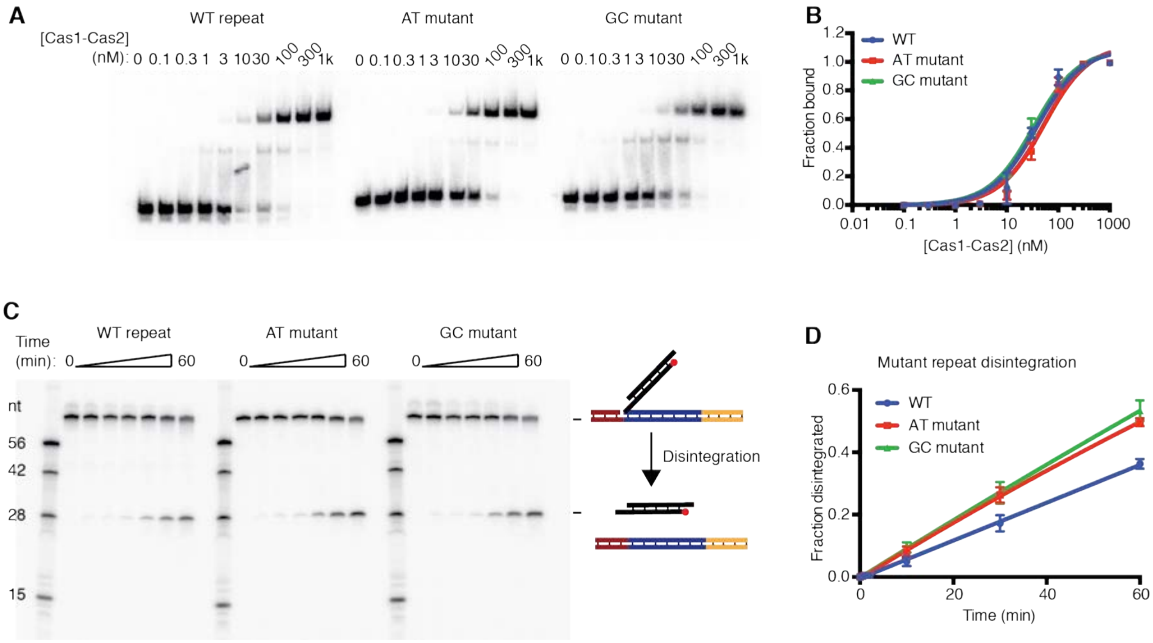

Fig. S5. Half-site substrates with mutant inverted repeats support robust binding and disintegration.
(A) Representative gel showing EMSA of radiolabeled half-site substrates with wild-type repeats or mutant repeats as described in the text. **(B)** Quantification of EMSAs of half-site substrates. Mean and standard deviation of three independent experiments are plotted. **(C)** Representative urea-PAGE gel of disintegration assays performed with radiolabeled wild-type or mutant repeat half-site substrates. Substrate and expected product are schematized, with the radiolabel indicated as a red circle. **(D)** Quantification of disintegration assays of half-site substrates. Mean and standard deviation of three independent experiments are plotted, and rates were fitted as a pseudo-first-order reaction.

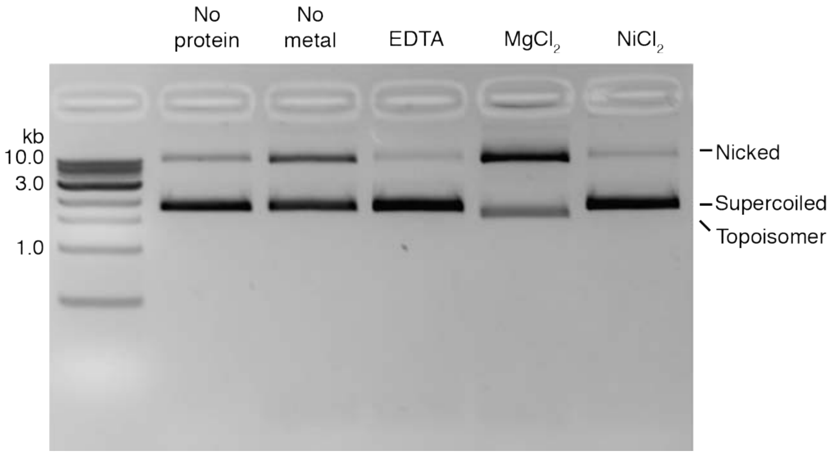

Fig. S6. Nickel does not support integration.
Agarose gel of integration assay with plasmid target. Integration results in the generation of nicked and toposiomerized plasmids, the latter of which run ahead of the supercoiled plasmid. EDTA or divalent cation were added as indicated.

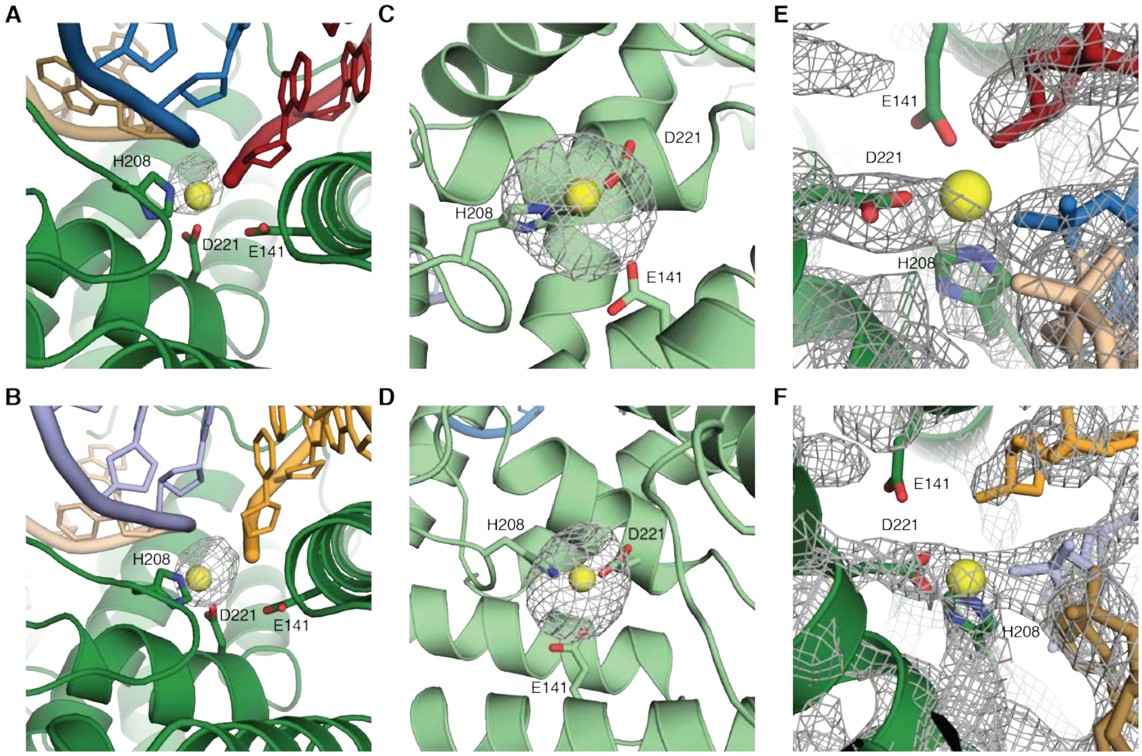

Fig. S7. Anomalous difference maps for active-site nickel atoms. (A-D) Cartoon representation of active-site nickels and anomalous density for leader-side active site (A), spacer-side active site (B), and the two non-catalytic active sites (C,D). Anomalous density is shown as a mesh contoured at 4.0 σ. Peaks in the anomalous difference map were smaller for nickel coordinated in the catalytic active sites, particularly the leader-side active sites. These Ni^{2+} were modelled with lower occupancy than those present in the non-catalytic active sites. Active site residues are shown as sticks. **(E,F)** Leader-side (E) and spacer-side (F) active sites with feature-enhanced map. Density is shown as mesh contoured at 2.0 σ. Active site residues are shown as sticks.

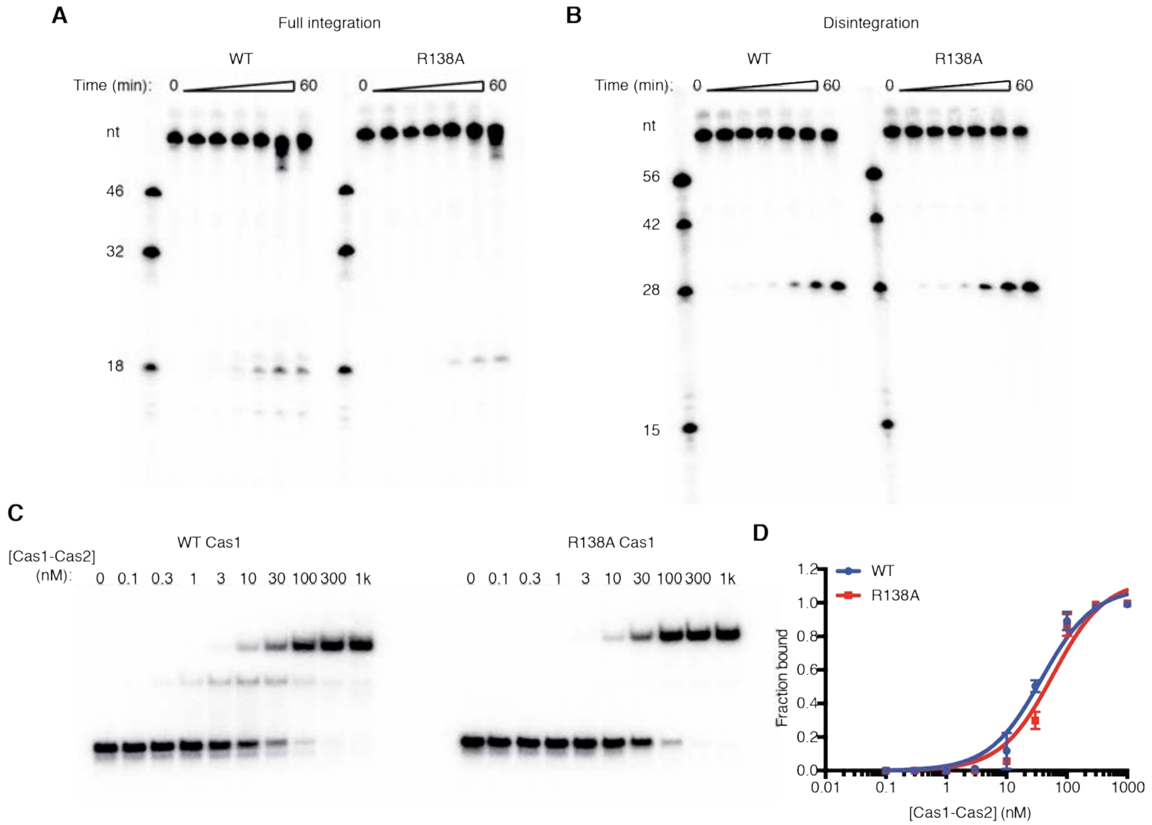

Fig. S8. Raw gels and target binding for R138A Cas1.

(A) Representative urea-PAGE gel of a full-site integration assay with half-site substrate and WT or R138A Cas1. **(B)** Representative urea-PAGE gel of disintegration assay with half-site substrate and WT or R138A Cas1. **(C)** Representative native PAGE gel of EMSA with half-site substrate and WT or R138A Cas1. **(D)** Quantification of half-site substrate binding by WT or R138A Cas1-Cas2. Mean and standard deviation of three independent experiments are shown.

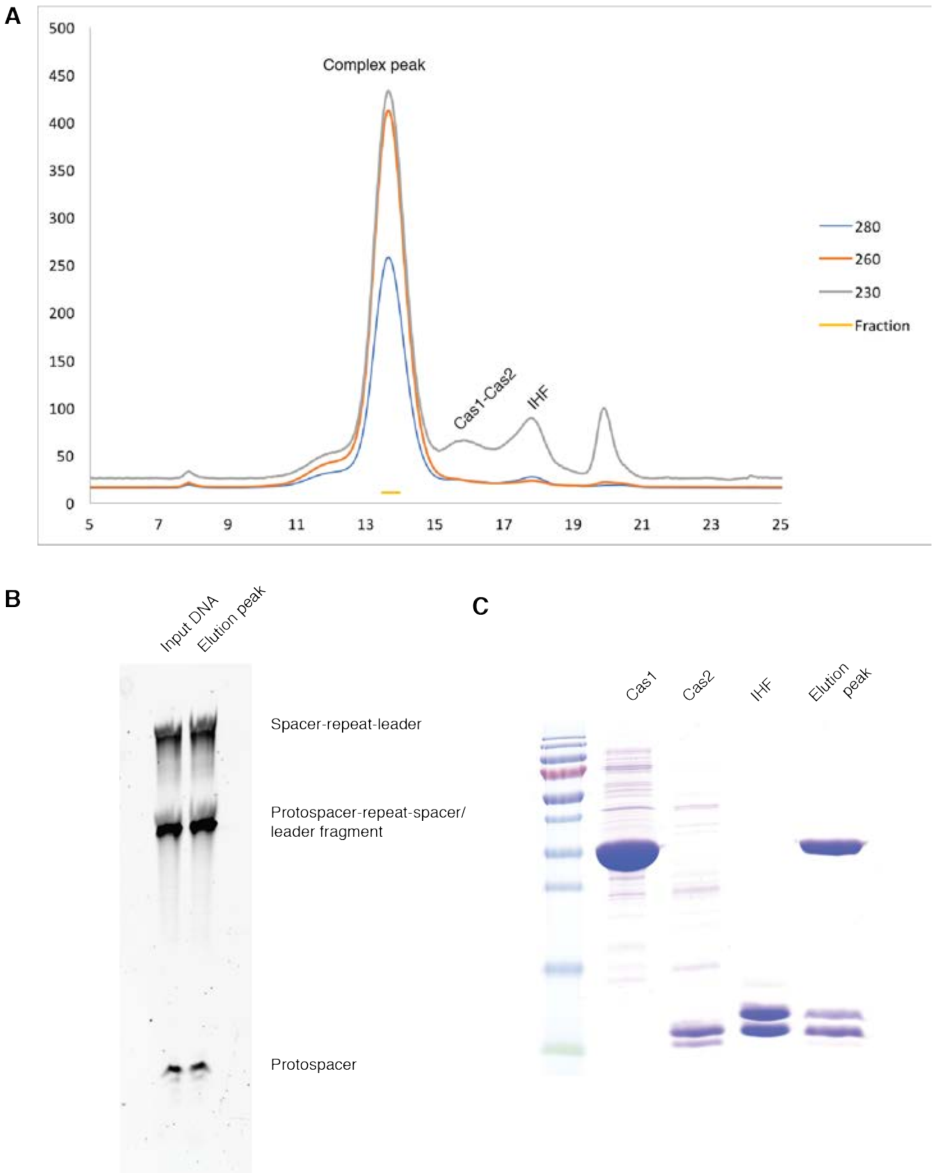

Fig. S9. IHF-Cas1-Cas2-DNA complex formation.
(A) Elution profile of IHF-Cas1-Cas2-DNA complex purified over Superose 6 10/300 column. The collected fractions are indicated with “Fraction.” **(B)** Urea-PAGE of the input DNA and the collected fractions. Strands are annotated. **(C)** SDS-PAGE of input proteins and the elution peak.

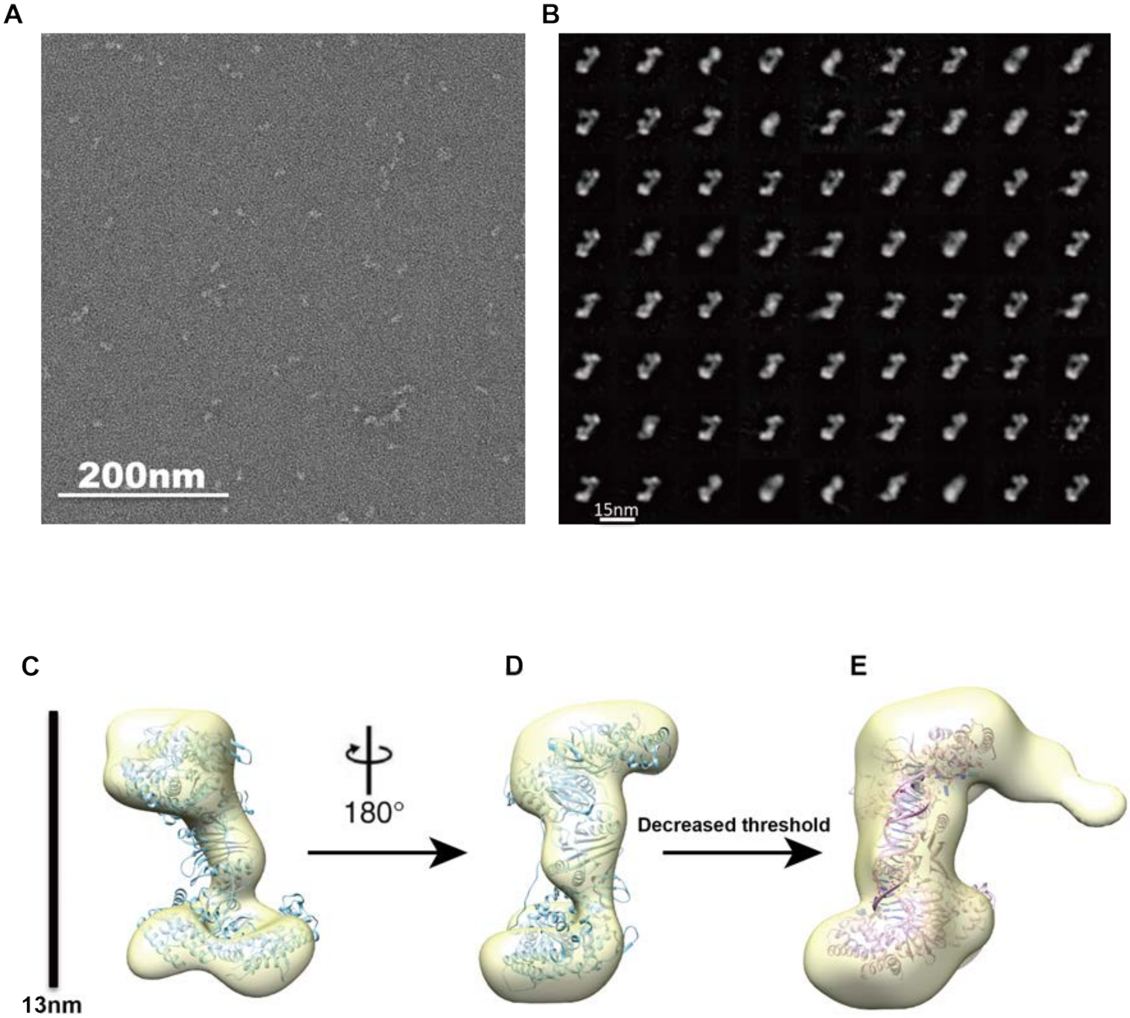

Fig. S10. Negative staining screening of Cas1-Cas2-DNA-IHF complex.
(A) Raw image of Cas1-Cas2-DNA-IHF complex by negative staining. The scale bar is 200nm. **(B)** Reference-free 2D class-averages of Cas1-Cas2-DNA-IHF complex by negative staining. 72 class-averages are shown in the panel. The scale bar is 15nm. **(C-D)** 3D refined model of Cas1-Cas2-DNA-IHF complex by negative staining. Two orientations of the EM map (at the threshold of 5 σ) aligned with atomic model of Cas1-Cas2 complex (PDB code 4p6i) are shown in panel (C) and (D). The EM map in the threshold of 3 σ aligned with Cas1-Cas2-DNA atomic model (PDB code 5ds5) was presented in the third panel (D).

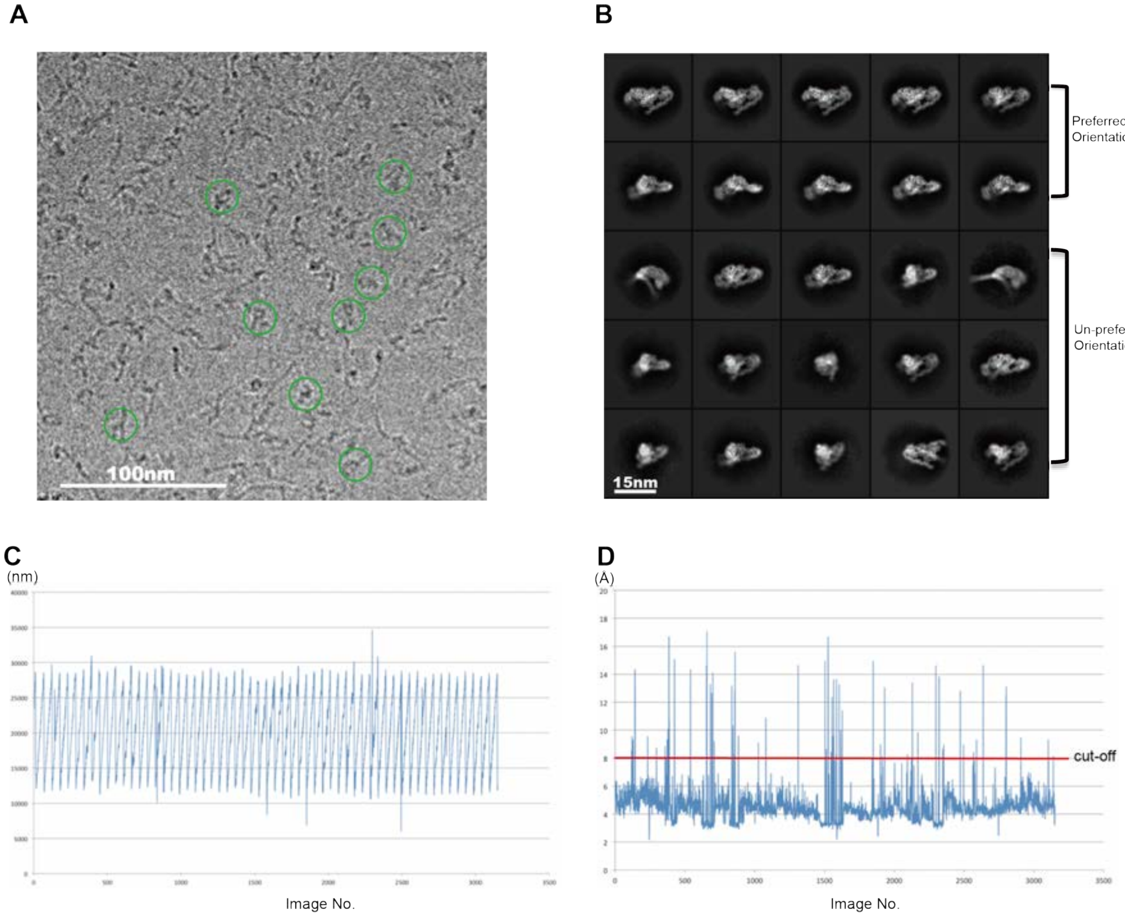

Fig. S11. Cryo-EM data analysis of Cas1-Cas2-DNA-IHF complex.
(A) Drift-corrected image of Cas1-Cas2-DNA-IHF complex by cryo-EM. The scale bar is 100nm. Several particles are marked with green circles. **(B)** Reference-free 2D class-averages of Cas1-Cas2-DNA-IHF complex by cryo-EM. 25 class-averages are shown in the panel. The upper two panels show the averages in preferred orientations. The following 3 panels show the averages in un-preferred orientations. The scale bar is 15nm. **(C)** The defocus value statistic of the whole data set. These values were calculated by CTFFIND4. **(D)** The maximal resolution statistic of the whole data set. These values were calculated in Relion2.0 based on signal intensity in micrograph at different resolutions. The red line indicates the cut-off resolution for micrograph sorting. Only the micrographs with the signal more than 8 Å were kept for 2D and 3D analysis.

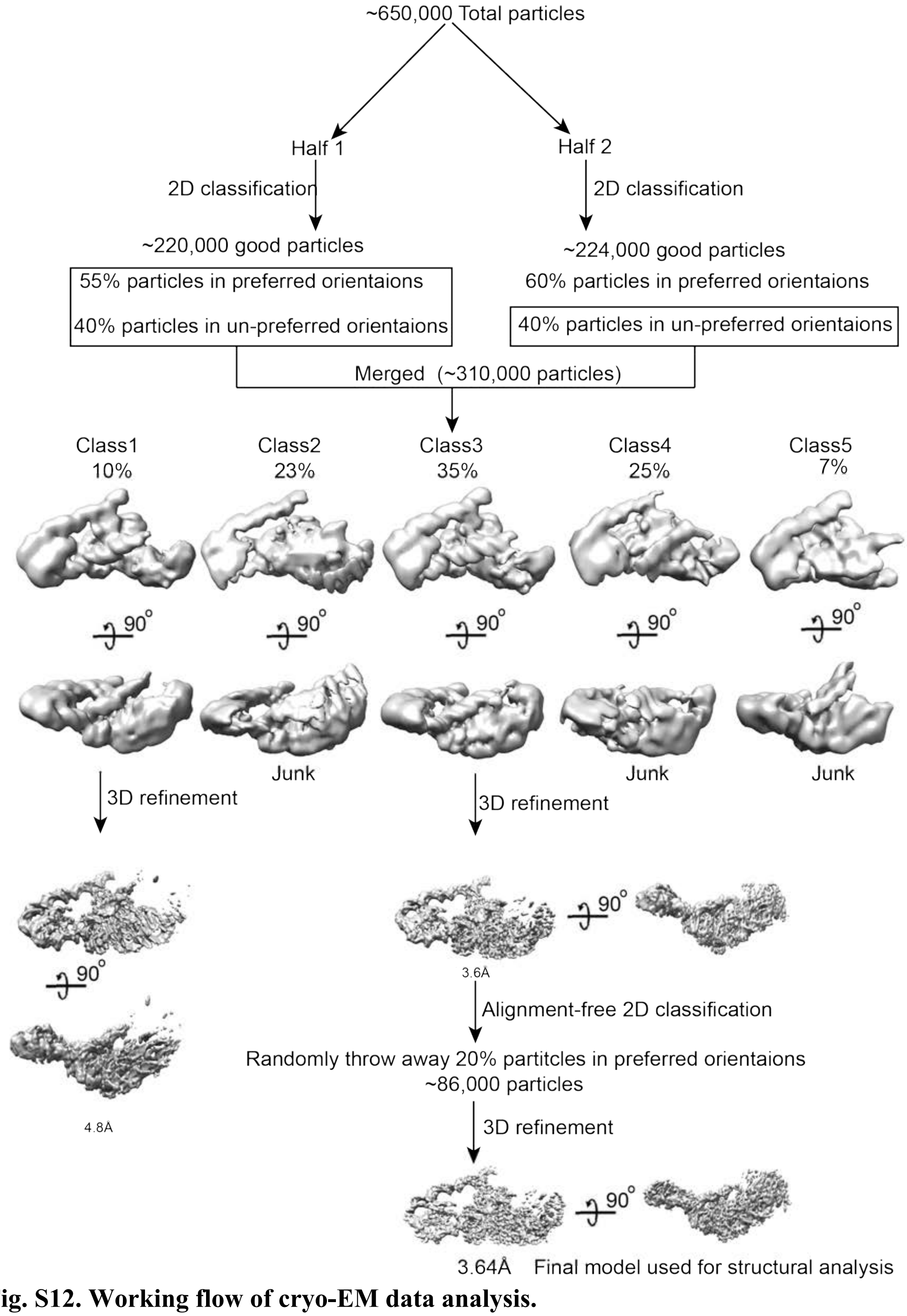
~650,000 Total particles
Half 1
Half 2
2D classification
2D classification
~220,000 good particles
~224,000 good particles
55% particles in preferred orientaions
40% particles in un-preferred orientaions
60% particles in preferred orientaions
40% particles in un-preferred orientaions
Merged (~310,000 particles)
Class1 10%
Class2 23%
Class3 35%
Class4 25%
Class5 7%
90°
90°
90°
90°
90°
Junk
Junk
Junk
3D refinement
3D refinement
90°
90°
4.8A
3.6A
Alignment-free 2D classification
Randomly throw away 20% partitcles in preferred orientaions
~86,000 particles
3D refinement
90°
3.64Å Final model used for structural analysis

Fig. S12. Working flow of cryo-EM data analysis.

About 2,900 micrographs were left after sorting. With the templates generated by manually picked particles, we picked about 650,000 particles in Gautomatch. The total data set was split in to two halves for 2D classification in Relion2.0. The good particles of half 1 and good particles in un-preferred orientations of half 2 were merged and classified into five 3D classes in Cryosparc with the initial model generated by negative staining. Two views of each 3D model are shown. The particle percentage of each class is also presented. Two good classes were further refined in Cryosparc. For class1, the reported resolution of the 3D refined model was 4.8 Å. For class3, the reported resolution of the 3D refined model was 3.6 Å. To further reduce the anisotropic resolution issue introduced by redundant particles in preferred orientations, we performed further alignment-free 2D classification based on the orientation information defined by the previous 3D refinement. 20 percent of particles in preferred orientations were discarded for further 3D refinement, which gave rise to a better isotropic map with the resolution of 3.64 Å.

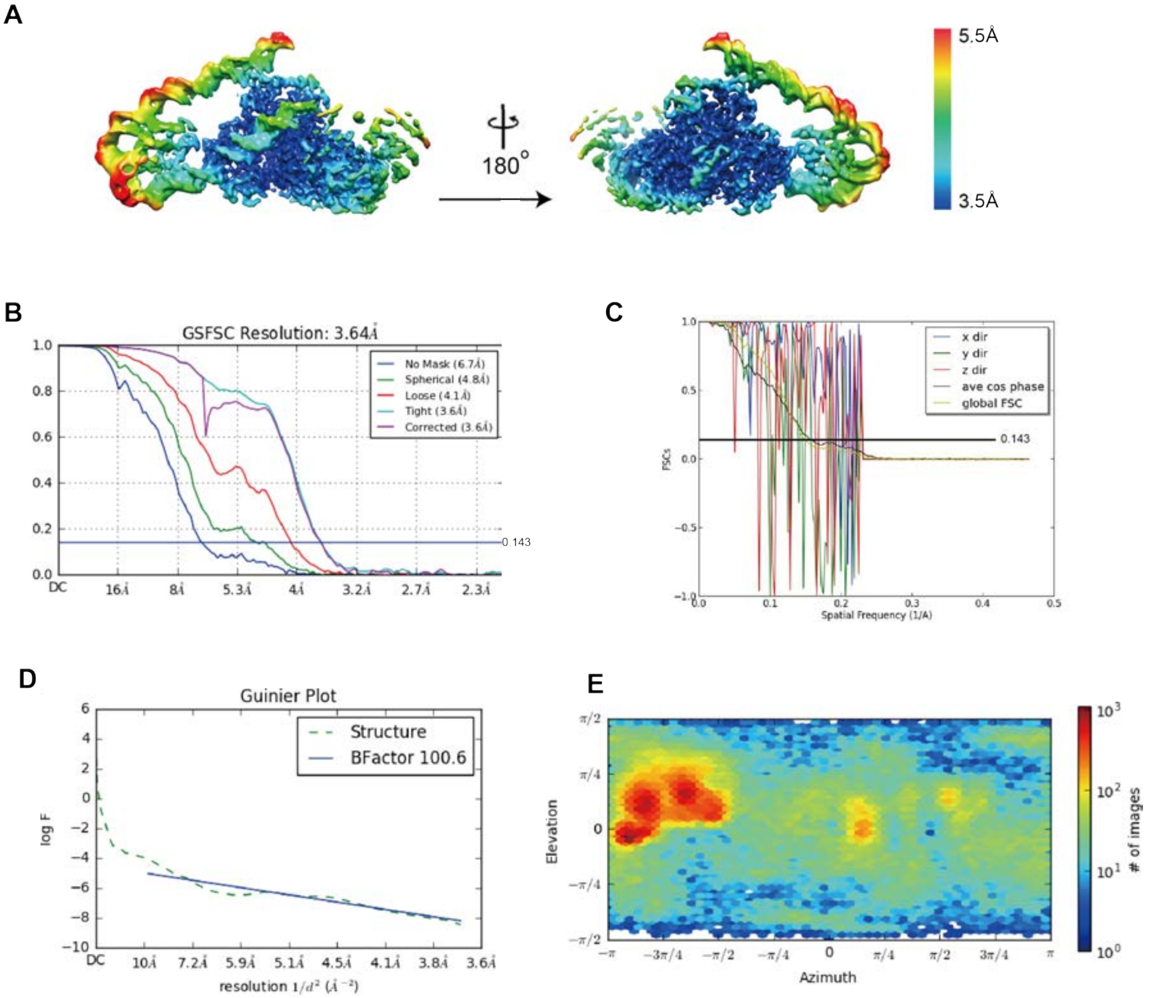

Fig. S13. Validation of EM 3D model.

(A) Cryo-EM structure of the Cas1-Cas2-DNA-IHF used for model building was shown and colored by local resolution calculated in Relion2.0. Resolution ranges from 3.5 Å to 5.5 Å. **(B)** The Fourier shell correlation (FSC) curve calculated using two independent half maps, indicating an overall resolution of 3.64Å. The panel was the standard output from Cryosparc. **(C)** The Fourier shell correlation (FSC) curve along x, y and z directions calculated by ThreeDFSC using two independent half maps. The panel is the standard output of ThreeDFSC. **(D)** The standard output of Guinier Plot for the sharpened model by Cryosparc. The B-factor used for sharpening is 100.6. **(E)** The Euler angle distribution of refined dataset. The panel is the standard output from Cryosparc.

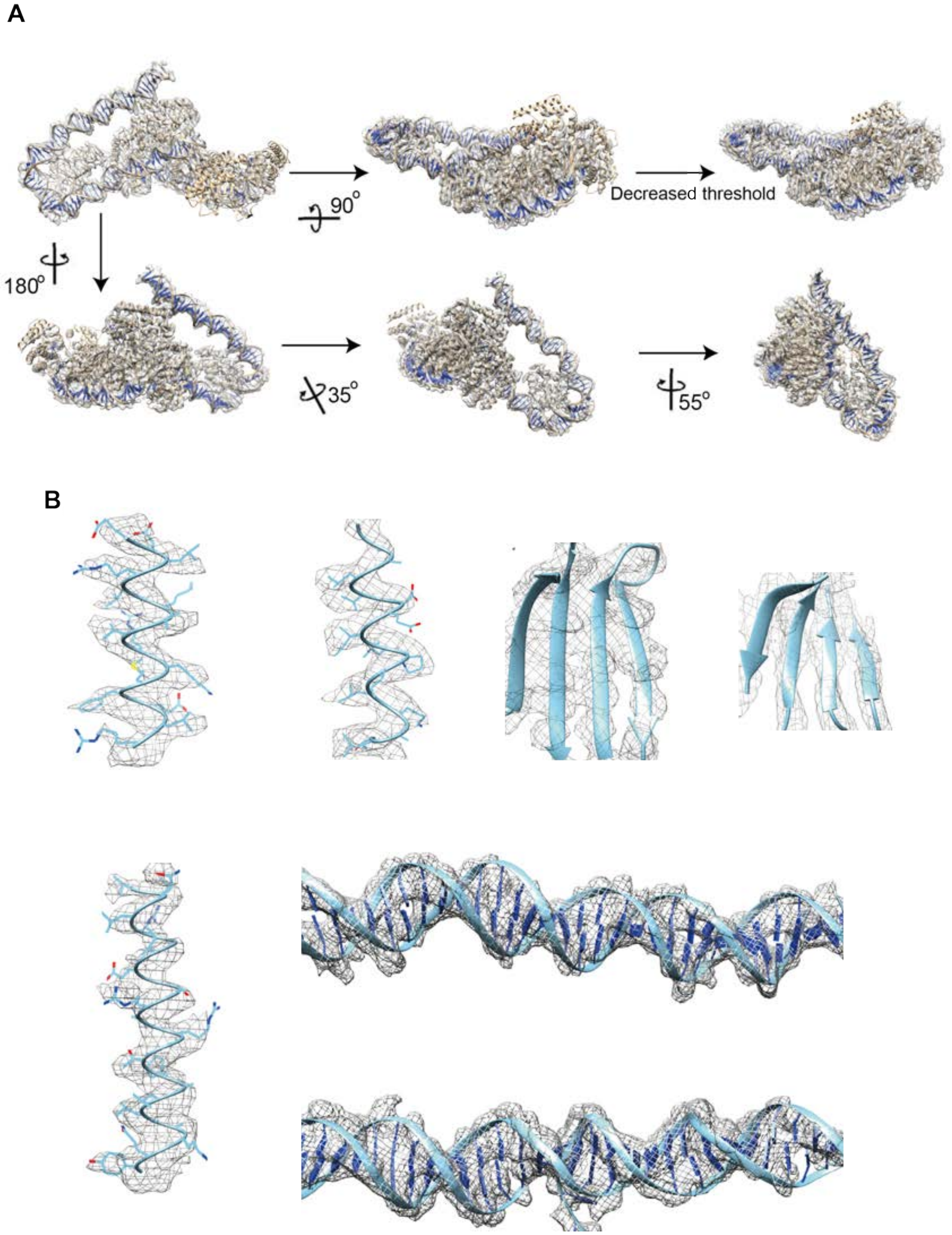

Fig. S14. Atomic model building of Cas1-Cas2-DNA-IHF complex.
(A) The EM density at the threshold of 8.5 σ was aligned with the atomic model and shown in different orientations. The EM density at the threshold of 5.5 σ was shown on the top right panel, which gave more visible density for the flexible Cas1 unit. **(B)** Representative regions of the EM density map of Cas1-Cas2-DNA-IHF complex, into which the atomic model was built.

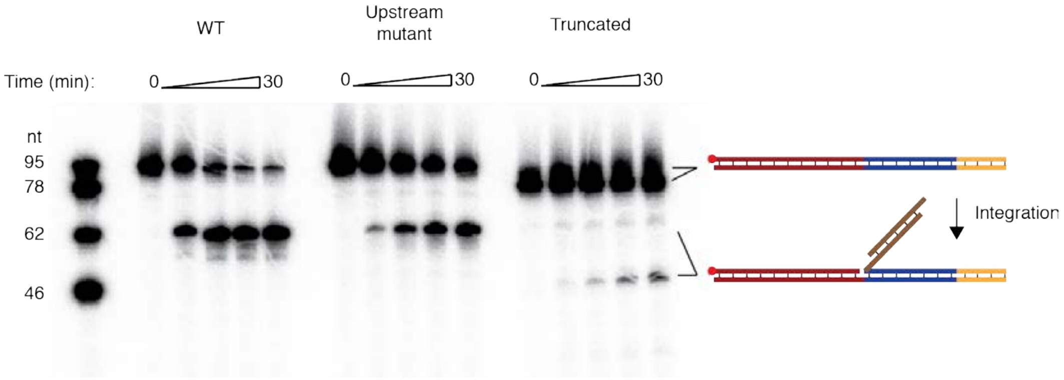

Fig. S15. Integration with limiting leader.
Urea-PAGE of a representative integration assay using unlabeled protospacer and a target with the top strand labeled. The substrate and expected product are shown as cartoons, with the radiolabel indicated with a red circle. The expected positions of the substrate and product bands are shown. Substrates are the same as in Fig. 5e. Time points were taken at 0, 1, 5, 15, and 30 minutes.

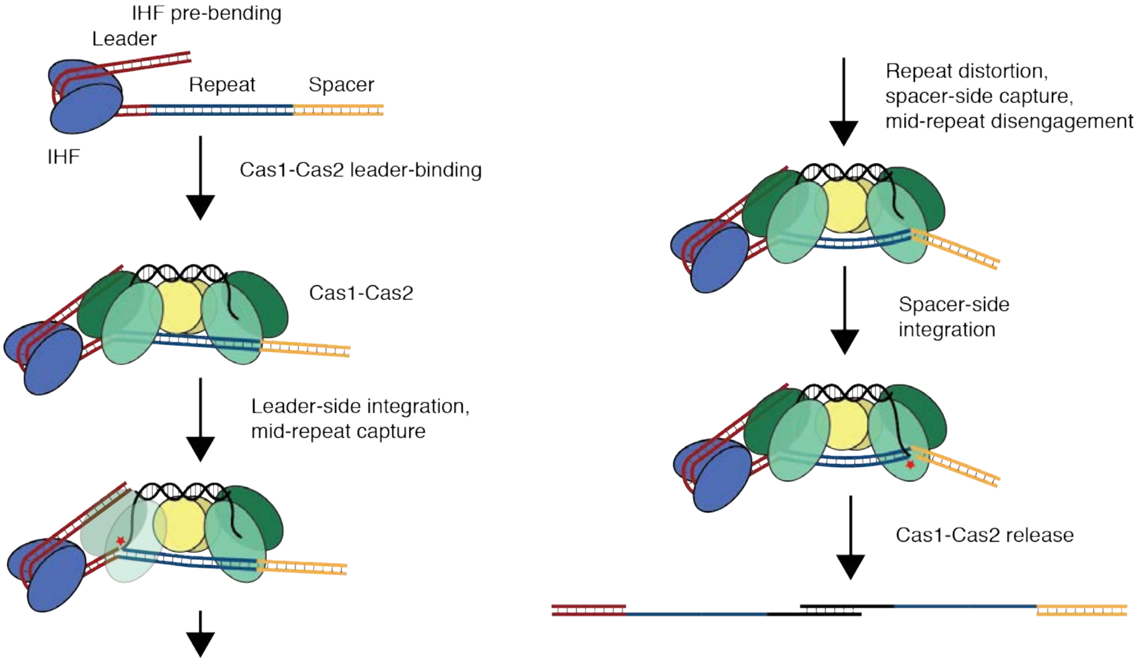

Fig. S16. Model for repeat recognition and integration by Cas1-Cas2.
IHF binding the leader sequence creates a doubled-over DNA structure that allows for simultaneous recognition of the leader sequence and the upstream recognition motif by Cas1-Cas2. Direct sequence readout is largely restricted to this initial recognition. The spacer side of the repeat is flexible at this point, but may be captured by basic residues in the channel formed by Cas1-Cas2. Capture of the repeat-spacer junction requires sequence-dependent distortion of the repeat, allowing off-target integration events to be halted at the half-site step. Integration of the protospacer occurs at both ends of the repeat, though it is unclear whether integration at the leader end precedes stable binding at the spacer end. Following integration, Cas1-Cas2 must release the product to allow for repair of the repeat. The mechanism for product release remains to be discovered.

Table S1. X-ray data collection and refinement statistics

	Half-site 5VVJ	Pseudo-full-site 5VVK	Pseudo-full-site with Ni^{2+} 5VVL
Data collection			
Unique reflections	25669	56364	35241
Space group	$P2_12_12_1$	$P2_1$	$P2_1$
Cell dimensions			
a, b, c (Å)	75.1, 183.1, 196.9	74.9, 187.6, 95.3	74.6, 197.7, 88.8
α, β, γ (°)	90, 90, 90	90, 112.7, 90	90, 111.3, 90
Resolution (Å)	98.46-3.89 (4.03-3.89)	98.81-2.90 (3.00-2.90)	39.5-3.31 (3.43-3.31)
R_{merge}[a] (%)	42.0 (283.6)	13.2 (191.3)	15.5 (146.2)
R_{pim}[b] (%)	12.2 (82.2)	14.3 (206.2)	16.7 (158.1)
$I/\sigma(I)$	6.0 (1.0)	12.8 (1.2)	12.3 (1.5)
$CC_{1/2}$[c]	99.9 (57.9)	99.9 (62.9)	99.7 (71.7)
Completeness (%)	99.8 (99.6)	99.8 (99.6)	99.7 (98.8)
Redundancy	12.9 (12.8)	7.3 (7.2)	6.9 (6.9)
Refinement			
Resolution (Å)	98.46–3.89	98.81–2.90	39.5–3.31
No. reflections	25,719 (2,512)	56,453 (5,610)	35,300 (3,515)
R_{work} / R_{free}[d]	29.1/32.9	21.5/25.2	22.6/26.2
No. atoms	11887	11896	11762
Protein	9534	9688	9757
DNA	2353	2208	1983
Metal			22
B factors (Å^2)			
Protein	149	93	100
DNA	195	138	139
Metal			131
R.m.s. deviations			
Bond lengths (Å)	0.003	0.003	0.003
Bond angles (°)	0.51	0.54	0.50
Ramachandran statistics (%)			
Favored	96.05	97.9	97.04
Allowed	3.79	2.1	2.80
Outliers	0.16	0	0.16

One crystal was used for each structure

Values in parentheses are for highest-resolution shell.

[a] $R_{\mathrm{merge}} = \Sigma_{\mathrm{hkl}}\Sigma_i|I_{\mathrm{hkl},i} – <I_{\mathrm{hkl}}>|/\Sigma_{\mathrm{hkl}}\Sigma_i I_{\mathrm{hkl},i}$, where I_{hkl} is the observed intensity for a given reflection and $<I_{\mathrm{hkl}}>$ is the average intensity of a unique reflection obtained from symmetry-related and redundant measurements.

[b] $R_{\mathrm{pim}} = \Sigma_{\mathrm{hkl}}(1/(n–1))^{1/2}\Sigma_i(|I_{\mathrm{hkl},i} – <I_{\mathrm{hkl}}>|)/\Sigma_{\mathrm{hkl}}\Sigma_i I_{\mathrm{hkl},i}$

[c] $CC_{1/2}$ is the percentage of correlation between intensities from random half-datasets.

[d] R_{work} is $\Sigma_{\mathrm{hkl}}||F_o – F_c||/ \Sigma_{\mathrm{hkl}}|F_o|$, where F_o is the observed amplitude and F_c is the calculated amplitude; R_{free} is the same statistic calculated for a randomly selected subset of the reflections (5% of the total) omitted from the refinement.

Table S2. EM Data collection and model refinement statistics of Cas1-Cas2-IHF-DNA complex

Data Collection	
EM	Titan Krios 300kV, K2 Gatan Summit
Pixel size (Å)	1.07
Defocus range (μm)	−1.2 to −2.8
Reconstruction (Relion)	Cryosparc
Particle Number	86,000
B-factor	100.600
Final resolution (Å)	3.64
Refinement (Phenix)	
Map CC (whole unit cell)	0.801
Map CC (around atoms)	0.735
R.m.s. deviations	
Bond lengths (Å)	0.00
Bond angles (^{o})	0.64
Ramachandran plot	
% favoured	95.44
% allowed	4.48
% outliers	0.07
Molprobity	
Clashscore	13.09

Table S3. DNA substrates used in this study.

Description	Sequence
Pseudo-full-site protospacer-repeat-spacer-hairpin (pseudo-full-site structure, Ni^{2+} pseudo-full-site structure)	GCTACTGGGGCCGAGGGTGTTCCCCGCGCCAGCGGGGATAAACCGAGCAGATATGCTC
Pseudo-full-site protospacer-repeat-leader-hairpin (pseudo-full-site structure, Ni^{2+} pseudo-full-site structure)	CACTGGTGGTCGCCGCGGTTTATCCCCGCTGGCGCGGGGAACACTCTAAGATATTAGA
Pseudo-full-site protospacer fragment (pseudo-full-site structure, Ni^{2+} pseudo-full-site structure, S1)	GCCCCAGTAGC
Pseudo-full-site protospacer fragment (pseudo-full-site structure, Ni^{2+} pseudo-full-site structure, S1)	GACCACCAGTG
Half-site protospacer-repeat-spacer-repeat-leader (half-site structure)	ATTTACTACTCGTTCTGGTGTTTCTCGTGTGTTCCCCGCGCCAGCGGGGATAAACCGAGCAGATATGCTCGGTTTATCCCCGCTGGCGCGGGGAACACTCTAAGATATTAGA
Protospacer strand (half-site structure, EM structure, 1d, 3d-f, 4d, 5e,f, 6c,d, S1, S5, S6, S8, S15)	AAACACCAGAACGAGTAGTAAATTGGGC
Extended leader (EM structure)	ATAAAGTTGGTAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGA
Protospacer-repeat-spacer (EM structure)	ATTTACTACTCGTTCTGGTGTTTCTCGTGTGTTCCCCGCGCCAGCGGGGATAAACCGAGCA
Full-length spacer-repeat-leader (EM structure, 3d,f, 5e,f)	TGCTCGGTTTATCCCCGCTGGCGCGGGGAACACTCTAAACATAACCTATTATTAATTAATGATTTTTTAAGCCAGTCACAATCTACCAACTTTAT
Pseudo-full-site/half-site leader fragment (1d, 3d, 4d, S1, S5, S8)	AATAATAGGTTATGTTTAGA
Half-site protospacer-repeat-spacer (1d, 3d, 4d, S1, S5, S8)	ATTTACTACTCGTTCTGGTGTTTCTCGTGTGTTCCCCGCGCCAGCGGGGATAAACCGAGCACAAATATCATCGC
Half-site spacer-repeat-leader (1d, 3d, 4d, S1, S5, S8)	GCGATGATATTTGTGCTCGGTTTATCCCCGCTGGCGCGGGGAACACTCTAAACATAACCTATTATT
Pseudo-full-site spacer fragment (S1)	GCGATGATATTTGTGCTC
Pseudo-full-site protospacer-repeat-spacer (S1)	CACTGGTGGTCGCCGAGGTTTATCCCCGCTGGCGCGGGGAACACTCTAAACATAACCTATTATT
Pseudo-full-site protospacer-repeat-leader (S1)	GCTACTGGGGCCGAGGGTGTTCCCCGCGCCAGCGGGGATAAACCGAGCACAAATATCATCGC
Protospacer strand (3d,f 4d, 5e,f, 6c,d, S6, S15)	ATTTACTACTCGTTCTGGTGTTTCTCGT
Full-length leader-repeat-spacer (3d,f, 5e,f,)	ATAAAGTTGGTAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGAGTGTTCCCCGCGCCAGCGGGGATAAACCGAGCA
AT mutant leader-repeat-spacer (3d)	ATAAAGTTGGTAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGAGTGTGAAAATAGCCATATTTTCTAAACCGAGCA
AT mutant spacer-repeat-leader (3d)	TGCTCGGTTTAGAAAATATGGCTATTTTCACACTCTAAACATAACCTATTATTAATTAATGATTTTTTAAGCCAGTCACAATCTACCAACTTTAT
GC mutant leader-repeat-spacer (3d)	ATAAAGTTGGTAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGAGTGTAGGGGCGGCCACGCCCCTTAAACCGAGCA
GC mutant spacer-repeat-leader (3d)	TGCTCGGTTTAAGGGGCGTGGCCGCCCCTACACTCTAAACATAACCTATTATTAATTAATGATTTTTTAAGCCAGTCACAATCTACCAACTTTAT
Half-site AT mutant protospacer-repeat-spacer (3e, S5)	ATTTACTACTCGTTCTGGTGTTTCTCGTGTGTGAAAATAGCCATATTTTCTAAACCGAGCACAAATATCATCGC
Half-site AT mutant spacer-repeat-leader (3e, S8)	GCGATGATATTTGTGCTCGGTTTAGAAAATATGGCTATTTTCACACTCTAAACATAACCTATTATT
Half-site GC mutant protospacer-repeat-spacer (3e, S5)	ATTTACTACTCGTTCTGGTGTTTCTCGTGTGTAGGGGCGGCCACGCCCCTTAAACCGAGCACAAATATCATCGC
Half-site GC mutant spacer-repeat-leader (3e, S5)	GCGATGATATTTGTGCTCGGTTTAAGGGGCGTGGCCGCCCCTACACTCTAAACATAACCTATTATT
Mid-repeat mismatch 1 (3f)	ATAAAGTTGGTAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGAGTGTTCCCCGCCCCAGCGGGGATAAACCGAGCA
Mid-repeat mismatch 2 (3f)	ATAAAGTTGGTAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGAGTGTTCCCCGCGGCAGCGGGGATAAACCGAGCA
Mid-repeat mismatch 3 (3f)	ATAAAGTTGGTAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGAGTGTTCCCCGCGCGAGCGGGGATAAACCGAGCA
Mid-repeat mismatch 4 (3f)	ATAAAGTTGGTAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGAGTGTTCCCCGCGCCTGCGGGGATAAACCGAGCA
Mid-repeat double-mismatch (3f)	ATAAAGTTGGTAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATG

	TTTAGAGTGTTCCCCGCGCGTGCGGGGATAAACCGAGCA
Upstream mutant leader-repeat-spacer (5e,f, S15)	ATAAAGTTCCATCATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGAGTGTTCCCCGCGCCAGCGGGGATAAACCGAGCA
Upstream mutant spacer-repeat-leader (5e,f, S15)	TGCTCGGTTTATCCCCGCTGGCGCGGGGAACACTCTAAACATAACCTATTATTAATTAATGATTTTTTAAGCCAGTCACAATGATGGAACTTTAT
Truncated leader-repeat-spacer (5e,f, 6c,d, S15)	GTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGAGTGTTCCCCGCGCCAGCGGGGATAAACCGAGCA
Truncated spacer-repeat-leader (5e,f, 6c,d, S15)	TGCTCGGTTTATCCCCGCTGGCGCGGGGAACACTCTAAACATAACCTATTATTAATTAATGATTTTTTAAGCCAGTCAC
Truncated +1 leader-repeat-spacer (6c,d)	GTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGATTTAGAGTGTTCCCCGCGCCAGCGGGGATAAACCGAGCA
Truncated +1 spacer-repeat-leader (6c,d)	TGCTCGGTTTATCCCCGCTGGCGCGGGGAACACTCTAAATCATAACCTATTATTAATTAATGATTTTTTAAGCCAGTCAC
Truncated +2 leader-repeat-spacer (6c,d)	GTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGATTTTAGAGTGTTCCCCGCGCCAGCGGGGATAAACCGAGCA
Truncated +2 spacer-repeat-leader (6c,d)	TGCTCGGTTTATCCCCGCTGGCGCGGGGAACACTCTAAAATCATAACCTATTATTAATTAATGATTTTTTAAGCCAGTCAC
Truncated +5 leader-repeat-spacer (6c,d)	GTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGATACATTTAGAGTGTTCCCCGCGCCAGCGGGGATAAACCGAGCA
Truncated +5 spacer-repeat-leader (6c,d)	TGCTCGGTTTATCCCCGCTGGCGCGGGGAACACTCTAAATGTATCATAACCTATTATTAATTAATGATTTTTTAAGCCAGTCAC
Leader sequences	
GGTAG->AACGA (5d)	AAGTACTCTTTAACATAATGGATGTGTTGTTTGTGTGATACTATAAAGTTAACGAATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGA
GGTAG->CCATC (5d)	AAGTACTCTTTAACATAATGGATGTGTTGTTTGTGTGATACTATAAAGTTCCATCATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGA
ATGGTAG (5d)	AAGTACTCTTTAACATAATGGATGTGTTGTTTGTGTGATACTATAAAGATGGTAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGA
TAGGTAG (5d)	AAGTACTCTTTAACATAATGGATGTGTTGTTTGTGTGATACTATAAAGTAGGTAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGA
TTCGTAG (5d)	AAGTACTCTTTAACATAATGGATGTGTTGTTTGTGTGATACTATAAAGTTCGTAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGA
TTGCTAG (5d)	AAGTACTCTTTAACATAATGGATGTGTTGTTTGTGTGATACTATAAAGTTGCTAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGA
TTGGAAG (5d)	AAGTACTCTTTAACATAATGGATGTGTTGTTTGTGTGATACTATAAAGTTGGAAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGA
TTGGTTG (5d)	AAGTACTCTTTAACATAATGGATGTGTTGTTTGTGTGATACTATAAAGTTGGTTGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGA
TTGGATC (5d)	AAGTACTCTTTAACATAATGGATGTGTTGTTTGTGTGATACTATAAAGTTGGTACATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGA
IHF flip (5d)	AAGTACTCTTTAACATAATGGATGTGTTGTTTGTGTGATACTATAAAGTTGGTAGATTGTGACTGGCTTCATAACCTATTATTAATTAATGATTTTTTTTTAGA

Sequences of all oligonucleotide DNA substrates used are listed, as well as the mutant leaders used for *in vivo* acquisition assays. Oligo description includes figures panels that the substrates were used to generate. Substrates used for crystallography or electron microscopy are noted with the relevant structure. Mutations from wild-type sequences are highlighted in red.