Published in final edited form as:

Nature. 2015 November 26; 527(7579): 535–538. doi:10.1038/nature15760.

Foreign DNA capture during CRISPR–Cas adaptive immunity

James K. Nuñez[1,*], **Lucas B. Harrington**[1,*], **Philip J. Kranzusch**[1,2], **Alan N. Engelman**[3,4], and **Jennifer A. Doudna**[1,2,5,6,7,8]

[1]Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, California 94720, USA.

[2]Howard Hughes Medical Institute, University of California, Berkeley, Berkeley, California 94720, USA.

[3]Department of Cancer Immunology and Virology, Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA.

[4]Department of Medicine, Harvard Medical School, Boston, Massachusetts 02215, USA.

[5]Department of Chemistry, University of California, Berkeley, Berkeley, California 94720, USA.

[6]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA.

[7]Innovative Genomics Initiative, University of California, Berkeley, Berkeley, California 94720, USA.

[8]Center for RNA Systems Biology, University of California, Berkeley, Berkeley, California 94720, USA.

Abstract

Bacteria and archaea generate adaptive immunity against phages and plasmids by integrating foreign DNA of specific 30–40 base pair (bp) lengths into clustered regularly interspaced short palindromic repeats (CRISPR) loci as spacer segments[1–6]. The universally conserved Cas1–Cas2 integrase complex catalyzes spacer acquisition using a direct nucleophilic integration mechanism similar to retroviral integrases and transposases[7–13]. How the Cas1–Cas2 complex selects foreign DNA substrates for integration remains unknown. Here we present X-ray crystal structures of the *Escherichia coli* Cas1–Cas2 complex bound to cognate 33 nucleotide (nt) protospacer DNA substrates. The protein complex creates a curved binding surface spanning the length of the DNA

Correspondence and requests for materials should be addressed to J.A.D. (; Email: doudna@berkeley.edu)
*These authors contributed equally to this work.

Author Contributions

J.K.N. and L.B.H. conducted the crystallography, biochemistry and *in vivo* spacer acquisition assays. J.K.N., L.B.H. and P.J.K. collected the X-ray diffraction data and determined the crystal structures. J.K.N., L.B.H., P.J.K., A.N.E. and J.A.D. designed the study, analyzed all data and wrote the manuscript.

Author Information

Atomic coordinates and structure factors for the reported crystal structures have been deposited at the Protein Data Bank under accession codes 5DS4 (with Mg^{2+}), 5DS5 (no Mg^{2+}) and 5DS6 (splayed DNA).

The authors declare no competing financial interests.

SUPPLEMENTARY INFORMATION. Uncropped gel images with size marker indications.

and splays the ends of the protospacer to allow each terminal nucleophilic 3′–OH to enter a channel leading into the Cas1 active sites. Phosphodiester backbone interactions between the protospacer and the proteins explain the sequence-nonspecific substrate selection observed *in vivo*[2–4]. Our results uncover the structural basis for foreign DNA capture and the mechanism by which Cas1–Cas2 functions as a molecular ruler to dictate the sequence architecture of CRISPR loci.

CRISPR loci are defined by repetitive elements that are separated by similarly-sized spacer sequences acquired from foreign DNA during the adaptation stage of CRISPR–Cas adaptive immunity[6,14]. CRISPR transcripts generated from the loci assemble with Cas proteins to detect and cleave foreign nucleic acids bearing sequence complementarity to the spacer segment[1,5,15–19]. In *E. coli*, expression of the Cas1–Cas2 protein complex triggers acquisition of new 33 bp spacers at the A/T-rich leader end of the CRISPR locus[7–10,20]. How the Cas1–Cas2 complex selects 33 bp protospacers of variable sequences and activates the 3′–OH ends for integration remains unknown. As the Cas1–Cas2 complex is sufficient to initiate spacer acquisition and adaptation of the CRISPR–Cas immune system, we hypothesized that the protein complex alone must provide the structural basis for the unknown mechanism of spacer length determination.

To determine how protospacer variation influences the efficiency of Cas1–Cas2-mediated spacer acquisition, we used an *in vitro* integration assay to test versions of a 33 bp sequence with constant overall length but different 3′ single-stranded overhang lengths[12]. The protospacer sequence is derived from the M13 bacteriophage genome and is highly acquired into the *E. coli* CRISPR locus after infection[8]. Unexpectedly, protospacers with overhanging 3′ nucleotides are strongly preferred by the Cas1–Cas2 complex over a completely double-stranded 33 bp protospacer (Fig. 1a and Extended Data Fig. 1a,b). Single-stranded DNA and substrates with 5′ overhangs are poor substrates for integration, highlighting the ability of Cas1–Cas2 to select specific DNA substrates prior to integration[12]. The most preferred protospacer DNA for *in vitro* integration consists of five overhanging nucleotides on each 3′ end (Extended Data Fig. 1). To determine the molecular basis of Cas1–Cas2 protospacer capture, we assembled Cas1–Cas2 complexes with the preferred protospacer substrate and determined crystal structures of the complex in the presence and absence of Mg^{2+} at 3.0 Å and 3.2 Å resolution, respectively (Extended Data Fig. 2 and Extended Data Table 1).

The structures reveal a hexameric protein architecture comprising four copies of Cas1 and two copies of Cas2, in which the protospacer spans the central Cas2 dimer and terminates within individual Cas1 subunits on each end of the complex (Fig. 1b). Structural superposition of the Cas1–Cas2 complex with and without bound DNA reveals a DNA-induced change in Cas1 subunit orientation in which each Cas1 dimer rotates ~10° in opposing directions against the central Cas2 hub (Extended Data Fig. 3a,b). Cas1–Cas2 protospacer capture positions each single-stranded protospacer 3′ end within a channel leading directly to a Cas1 active site. Simulated annealing omit maps show clear electron density for the double-helical region and the five-nucleotide overhangs on each end of the protospacer (Extended Data Fig. 4a–c). The constrained protein channel guiding each DNA strand from its double-helical region to the single-strand-accommodating Cas1 active site

explains the specificity of Cas1–Cas2 for five-nucleotide 3′ overhang substrates (Fig. 1a and Extended Data Fig. 1). Two of the four Cas1 subunits, colored green in Fig. 1b, are not occupied with the protospacer 3′ ends and are likely non-catalytic since the 3′–OH nucleophile and the scissile phosphodiester bond of the target DNA must be in the same active site for direct nucleophilic integration.

In the active sites, the 3′ terminal base is involved in a stacking interaction with Y217 that positions the nucleophilic 3′–OH ends of the protospacer near the conserved metal-binding residues E141, H208 and D221 (Fig. 1c). Although we cannot assign density for Mg^{2+} in the active sites, these three residues have been shown previously to coordinate a Mn^{2+} ion in the active site of Cas1 from *P. aeruginosa*[21]. Furthermore, alanine mutations at these positions disrupt *in vivo* spacer acquisition[7,8,10]. Thus, the observed positioning of the 3′–OH nucleophiles and catalytic residues likely represents the active configuration of the nucleoprotein complex immediately prior to spacer integration.

All interactions between Cas1–Cas2 and protospacer DNA involve coordination of the phosphate-backbone rather than base-specific contacts, consistent with the variable sequence selection of protospacers that is essential for resistance to diverse foreign sequences[2–4]. Two central regions of the Cas1–Cas2 complex, which we term the Arginine Clamp and the Arginine Channel, stabilize the protospacer (Fig. 2a–d). The Arginine Clamp interacts with the middle of the duplex region where four Arg residues coordinate each DNA strand: Cas1 R41 and Cas2 R16, R77, R78 (Fig. 2c). Reverse charge mutations of Cas1 R41 and Cas2 R16 and R78 drastically reduce spacer acquisition *in vivo*, whereas the Cas2 R77E mutant functions similar to wild-type (WT) Cas2 (Fig. 2e). Thus, Cas1 R41, Cas2 R16 and R78 are the key constituents of the Arg clamp. The contribution of Cas2 to protospacer DNA binding supports the previous hypothesis that the main function of Cas2 is to form a non-catalytic scaffold within the Cas1–Cas2 complex[10].

Cas1 residues R66, R84, R245 and R248 line the Arginine Channel that stabilizes the junction where the duplex region terminates and the ssDNA overhang enters the active site. Reverse charge mutation of each arginine lining the Arginine Channel disrupts spacer acquisition *in vivo* (Fig. 2e). In addition, purified Cas1 R59D or R66D proteins complexed with WT Cas2 are highly defective in integrating 33 bp duplex or five-nucleotide overhang protospacer substrates *in vitro* (Fig. 2f). Fluorescence polarization assays demonstrate that the mutant complexes exhibit dramatically reduced affinity for protospacer DNA, highlighting the critical role of this part of the Cas1–Cas2 complex for protospacer capture and complex stability (Fig. 2g).

The Cas1–Cas2–DNA crystal structures uncover a protein wedge that terminates the protospacer dsDNA region and allows single-stranded DNA overhangs to enter the Arginine Channel. A stacking interaction of the 5′ terminal base (adenine 6 in Fig. 3a,b) with Y22 of Cas1 stabilizes protospacer duplex unwinding, directing each single-stranded 3′ overhang to sharply bend ~90° away from the duplex and into the active site channel (Fig. 3b). A mutation of Y22 to alanine reduces spacer acquisition *in vivo* whereas a phenylalanine mutation has near WT levels of acquisition, consistent with a specific role for Cas1 Y22 base-stacking in protospacer strand splaying (Fig. 3c). Sequence alignment of representative

Cas1 proteins in Type I CRISPR systems reveals that Y22 is not universally conserved in other bacteria, suggesting that additional or different Cas1 residues may stabilize the splayed ends in other CRISPR–Cas systems (Extended Data Fig. 5).

The observed stacking interaction raises the possibility that fully duplexed protospacers are separated by Cas1 Y22, thereby displacing the 5′ end of the duplex, we term the non-nucleophilic strand, from the nucleophilic strand carrying the 3′–OH. DNA transposases and retroviral integrases also utilize end fraying to isolate the reactive DNA strands for chemistry within enzyme active sites[22–24]. To test this potential activity of Cas1–Cas2, we introduced an increasing number of mismatches at the ends of the 33 bp protospacer to disrupt end base pairing and assayed their potential for *in vitro* integration (Fig. 3d and Extended Data Fig. 6a,b). Similar to the 3′ overhang substrates, the 4 and 5 nt frayed ends are highly preferred, presumably due to the lower energy required for capture of these substrates compared to perfectly duplexed ends (Fig. 3d). The complex containing the Cas1 Y22A mutant regains marginal activity with substrates containing 5 or 6 nt splayed ends, suggesting that Y22 steers the non-nucleophilic DNA strand away from the active site (Fig. 3d). Notably, the displaced non-nucleophilic strand is not cleaved into a shorter fragment by Cas1–Cas2, as the protospacer ends are not processed during integration (Extended Data Fig. 6c).

To determine the trajectory of the displaced non-nucleophilic strand after end-splaying, we crystallized Cas1–Cas2 with a protospacer with five-nucleotide frayed ends on both sides (Fig 3a,b). The electron density at the fork is similar to the structures described above, except we observe the first nucleotide of the displaced non-nucleophilic strand pointing in the opposite direction from the nucleophilic ssDNA strand. Clear electron density is not observed for the remaining nucleotides of the displaced strand, indicating that they are not stabilized by the complex.

An alternative crystal form grown in the presence of Mg^{2+} reveals secondary Cas1–DNA interactions that provide additional insight into the mechanism of Cas1–Cas2 genomic DNA target binding and subsequent integration. In addition to the two Cas1 “catalytic” active sites carrying the 3′–OH ends of the protospacer, the “non-catalytic” Cas1 active sites interact with the protospacer DNA from a symmetry mate, revealing a possible coordination of the target DNA during integration (Fig. 4a and Extended Data Fig. 7a). The non-catalytic Cas1 engages the DNA minor groove by contacts with α helix 7, causing a slight kink on the DNA compared to our alternative crystal form lacking Mg^{2+} (Extended Data Fig. 7b). A close-up of the active site shows continuous density for Mg^{2+} with E141, H208, D221 and a phosphate backbone of the presumed target DNA, capturing a snapshot of scissile phosphodiester bond coordination prior to integration (Fig. 4a).

Because integration must occur in the active site that coordinates the 3′–OH of the protospacer DNA, we modeled the protein-DNA interactions from the non-catalytic Cas1 active sites into the catalytic Cas1 active sites. This reveals the positioning of the nucleophilic 3′–OH of the protospacer ends for attacking the scissile phophodiester bond in the modeled DNA (Fig. 4b, c). Further work will be needed to shed light on how the complex specifically recognizes the leader-repeat region of the CRISPR locus for integration, as recently observed *in vitro*[11–13].

Together, these data explain key aspects of Cas1–Cas2 integrase-mediated acquisition of new DNA into bacterial genomes. First, we show that the substrates for integration are double-stranded DNA. Importantly, however, optimal substrates include a central 23 base pair helical region flanked by five single-stranded nucleotides on each 3′ end. If substrates for CRISPR integration come from ssDNA products of RecBCD, as recently suggested, they must somehow anneal or otherwise become double stranded prior to Cas1–Cas2 capture[20]. It remains unclear how the Cas1–Cas2 complex recognizes the AAG protospacer adjacent motif (PAM) during protospacer selection, since the terminal nucleotides containing the 3′–OH nucleophiles are coordinated similarly in the Cas1 active sites (Fig. 1). Second, the Cas1–Cas2 integrase architecture specifies the precise length of integrated DNA, ensuring uniformity of spacer lengths within CRISPR loci. Finally, the structure-based model of DNA target sequence positioning suggests that in addition to catalyzing the integration reaction, Cas1 plays a role in binding the target CRISPR locus. Target binding could possibly disrupt the structural symmetry observed in the crystal structure to coordinate the sequence-specific integration reactions at the leader-end of the CRISPR locus. Insights into target site recognition may offer strategies for altering or enhancing integration site specificity, with implications for use of the Cas1–Cas2 integrase as a genome-modifying technology.

METHODS

Cas1, Cas2 and DNA preparation

The Cas1 and Cas2 proteins from *E. coli* K12 (MG1655) were cloned and separately purified as previously described[10]. Single-stranded DNA oligonucleotides purchased from Integrated DNA Technologies were annealed in 20 mM HEPES-NaOH, pH 7.5, 25 mM KCl, 10 mM $MgCl_2$ by heating at 95 °C for 3 min and slow cooling to room temperature. The pCRISPR DNA target for *in vitro* integration was constructed as previously described[12]. The DNA substrates used for crystallization were gel purified prior to complex formation. The sequences for the five nt overhang substrates used for crystallization are: ssDNA1 (5′-ATTTACTACTCGTTCTGGTGTTTCTCGT-3′) and ssDNA2 (5′-AAACACCAGAACGAGTAGTAAATTGGGC-3′). The sequences for the five-nucleotide splayed substrates are: ssDNA1 (5′-TAAACATTTACTACTCGTTCTGGTGTTTCTCGT-3′) and ssDNA2 (5′-CATCTAAACACCAGAACGAGTAGTAAATTGGGC-3′).

In vivo acquisition and *in vitro* integration assays

The *in vivo* acquisition assays were performed as previously described[7]. The *in vitro* integration reactions were conducted as previously described with slight modifications[12]. After pre-incubation of equimolar Cas1 and Cas2 at 4 °C, 100 nM of the resulting Cas1–Cas2 complex was incubated with 100 nM protospacer DNA for an additional 10–15 min at room temperature. The integration reaction was activated by the addition of 300 ng (~5 nM) pCRISPR, incubated at 37 °C for 1 h and quenched with DNA loading buffer supplemented with EDTA at a final concentration of 20 mM. The reaction products were analyzed on 1.5% agarose gels. Percent integration activity values were determined by quantifying the band intensity of the relaxed pCRISPR product and dividing over the intensity of all bands detected by Image Lab Software (Bio-Rad). We note that the integration activity could be a mixture of half-site and full-site integration products, as described previously[12].

Complex formation, crystallization and structure determination

Purified Cas1 and Cas2 were incubated with protospacer DNA at equimolar concentrations (50 μM) in Buffer A (500 mM KCl, 20 mM HEPES-NaOH, pH 7.5, 1 mM DTT, 10 mM EDTA), followed by overnight dialysis at 4 °C against Buffer B (100 mM KCl, 20 mM HEPES-NaOH, pH 7.5, 1 mM DTT, 5 mM EDTA). The dialyzed sample was applied on a Superdex 75 10/300 column (GE Healthcare) in Buffer B. Peak fractions were pooled and concentrated to ~3 mg ml^{-1} for crystallization. Optimized crystals were grown by hanging-drop vapor diffusion at room temperature in two different conditions, as described in the text. The Mg^{2+}-containing crystals grew as gem-like morphologies in 50 mM MES, pH 6.1, 10% isopropanol and 20 mM $MgCl_2$. The "no Mg^{2+} crystals" grew as rods in 100 mM sodium citrate tribasic pH 5.6, 200 mM sodium acetate and 8% PEG 8000 (w/v). The crystals were briefly transferred into a drop containing either 25% ethylene glycol (with Mg^{2+} crystals) or 30% glycerol (without Mg^{2+} crystals) for cryoprotection and frozen in liquid nitrogen. The Cas1–Cas2 complex with a splayed DNA substrate crystallized in the same conditions as the no Mg^{2+} crystals.

X-ray diffraction data were collected under cryogenic conditions at beamline 8.3.1 at the Lawrence Berkeley National Laboratory Advanced Light Source. Initial phases were obtained by sequential molecular replacement using individual protein components of the Cas1–Cas2 apo structure (PDB 4P6I) as search models. Following initial placement of two Cas1 dimers and a dimer of Cas2, phases were improved by performing one round of rigid body refinement in PHENIX[25]. The resulting maps showed clear unbiased density for protospacer DNA, and subsequent model building was performed through iterative rounds of building in Coot[26] and refinement in PHENIX with NCS restraints on the protein subunits. The asymmetric unit of the three structures contains one copy of the Cas1–Cas2 complex bound to protospacer DNA. Statistics for the final crystal structures are reported in Extended Data Table 1. The final structures are missing clear density for the loop connecting α6 and α7 of Cas1. We assume this loop to be highly disordered as it is also not observed in the apo *E. coli* Cas1 crystal structure (PDB 3NKD) and the apo Cas1–Cas2 complex (PDB 4P6I)[10,27].

Fluorescence polarization

Fluorescence polarization assays were performed in 20 mM HEPES-NaOH, pH 7.5, 25 mM KCl, 5 mM EDTA, 1 μg ml^{-1} BSA and 1 mM DTT. Cas1–Cas2 were complexed and purified over gel filtration for all binding assays. The 3′–fluorescein labeled DNA substrate was added to the protein solution at a final concentration of 5 nM and the DNA–protein mixture was allowed to incubate for 30 min at 22 °C. Measurements were made by excitation at 485 nm and monitoring emission at 535 nm. Data were fit to a binding isotherm to obtain K_d. Each experiment was conducted in triplicates and error bars represent the standard deviation.

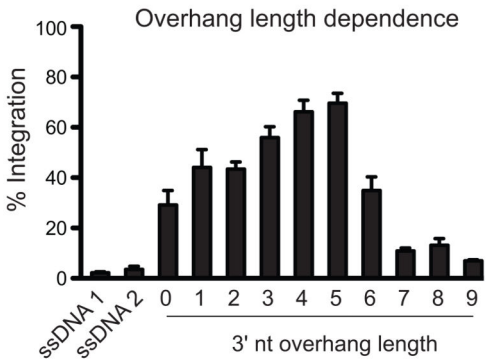
Sequence alignment

The *cas1* sequences were obtained from the National Center for Biotechnology Information (NCBI) Gene Data Bank. A representative *cas1* from each CRISPR Type I subtype were chosen based on previous subtype assignments and the alignment was generated using

MAFFT[28,29]. The organisms chosen for the alignment are: *Escherichia coli K-12, Cronobacter dublinensis 582, Erwinia amylovora, Yersinia pestis biovar Antiqua str. B42003004, Yersinia kristensenii, Hafnia alvei, Sulfolobus solfataricus, Thermotoga maritime, Pseudothermotoga lettingae, Deferribacter desulfuricans, Desulfovibrio vulgaris, Bacillus halodurans, Bacillus cereus, Synechocystis sp. PCC 6803, Cyanothece sp. PCC 8802* and *Limnoraphis robusta*.

Extended Data

a

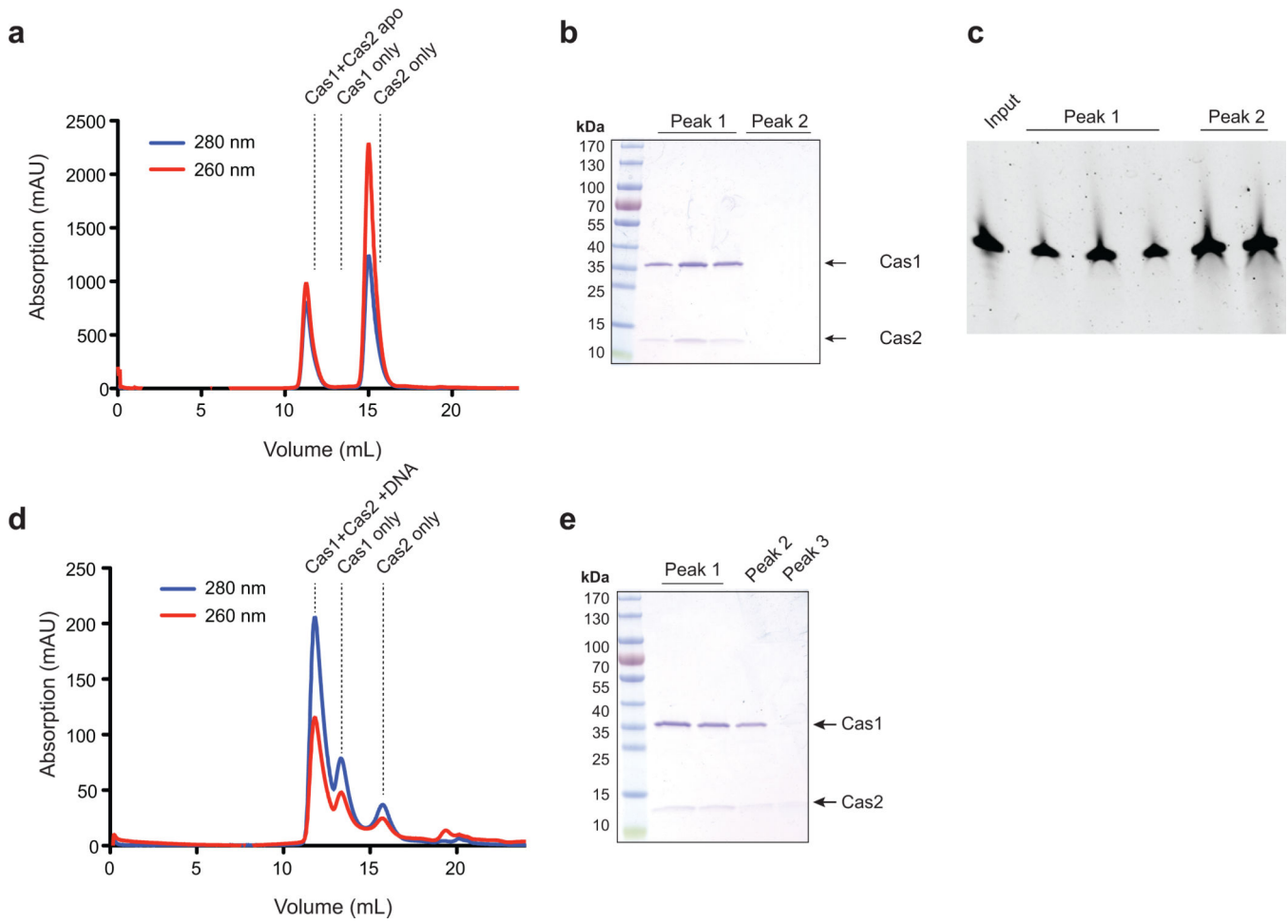


b

Full dsDNA	5' GCCCAATTTACTACTCGTTCTGGTGTTTCTCGT 3' 3' CGGGTTAAATGATGAGCAAGACCACAAAGAGCA 5'	5' ATTTACTACTCGTTCTGGTGTTTCTCGT 3' 3' CGGGTTAAATGATGAGCAAGACCACAAA 5'	**5 nt**
1 nt	5' CCCAATTTACTACTCGTTCTGGTGTTTCTCGT 3' 3' CGGGTTAAATGATGAGCAAGACCACAAAGAGC 5'	5' TTTACTACTCGTTCTGGTGTTTCTCGT 3' 3' CGGGTTAAATGATGAGCAAGACCACAA 5'	**6 nt**
2 nt	5' CCAATTTACTACTCGTTCTGGTGTTTCTCGT 3' 3' CGGGTTAAATGATGAGCAAGACCACAAAGAG 5'	5' TTACTACTCGTTCTGGTGTTTCTCGT 3' 3' CGGGTTAAATGATGAGCAAGACCACA 5'	**7 nt**
3 nt	5' CAATTTACTACTCGTTCTGGTGTTTCTCGT 3' 3' CGGGTTAAATGATGAGCAAGACCACAAAGA 5'	5' TACTACTCGTTCTGGTGTTTCTCGT 3' 3' CGGGTTAAATGATGAGCAAGACCAC 5'	**8 nt**
4 nt	5' AATTTACTACTCGTTCTGGTGTTTCTCGT 3' 3' CGGGTTAAATGATGAGCAAGACCACAAAG 5'	5' ACTACTCGTTCTGGTGTTTCTCGT 3' 3' CGGGTTAAATGATGAGCAAGACCA 5'	**9 nt**

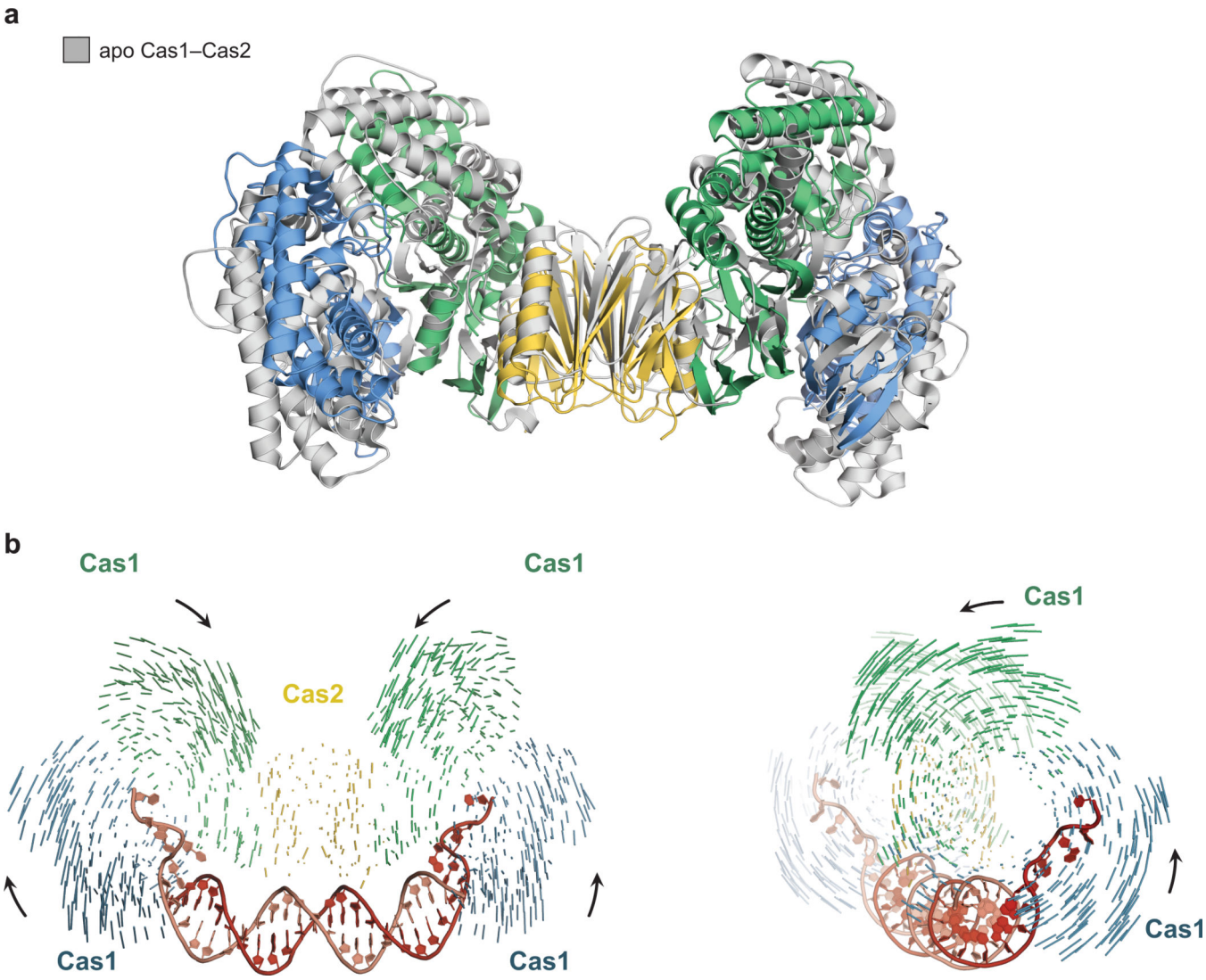
Extended Data Figure 1. Effect of overhang length on integration efficiency

a, A plot of the percent integration of protospacers ± standard deviation with varying 3′ single-stranded DNA extensions. A representative gel is shown in Fig. 1a. **b**, Protospacer sequences used for the assays described in **a** and Fig. 1a, with the red nucleotides indicating the 3′ overhang regions.



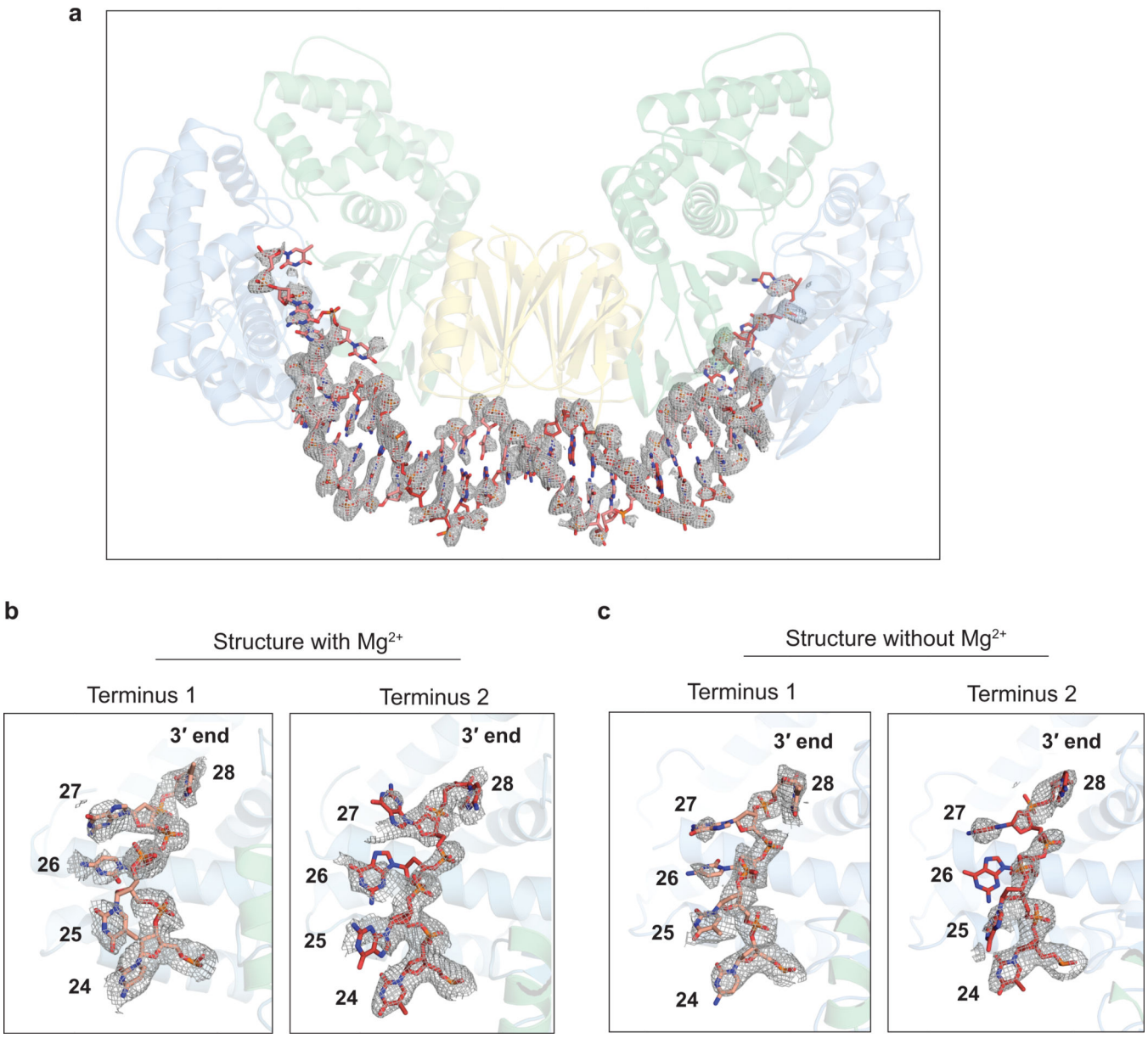
Extended Data Figure 2. Assembly of Cas1–Cas2 complex bound to protospacer DNA

a, Gel filtration chromatogram of pre-assembled Cas1–Cas2 complex with protospacer DNA containing five nt 3′ overhangs. The dotted lines indicate the peak fractions of the Cas1–Cas2 complex without DNA, as shown in **d**. The solid lines indicate the peak fractions of the Cas1–Cas2 complex bound to DNA (first peak) and excess, unbound DNA (second peak). **b, c**, The fractions from Peak 1 (~12 ml) and Peak 2 (~15 ml) were analyzed by Coomassie-stained SDS-PAGE (**b**) and 12% urea-PAGE (**c**) to confirm the presence of Cas1, Cas2 and protospacer DNA. **d**, Gel filtration chromatogram of assembled Cas1–Cas2 without protospacer DNA. **e**, Coomassie-stained SDS-PAGE of the peak fractions from **d**. Supplementary Information contains the full images for **b, c** and **e**.



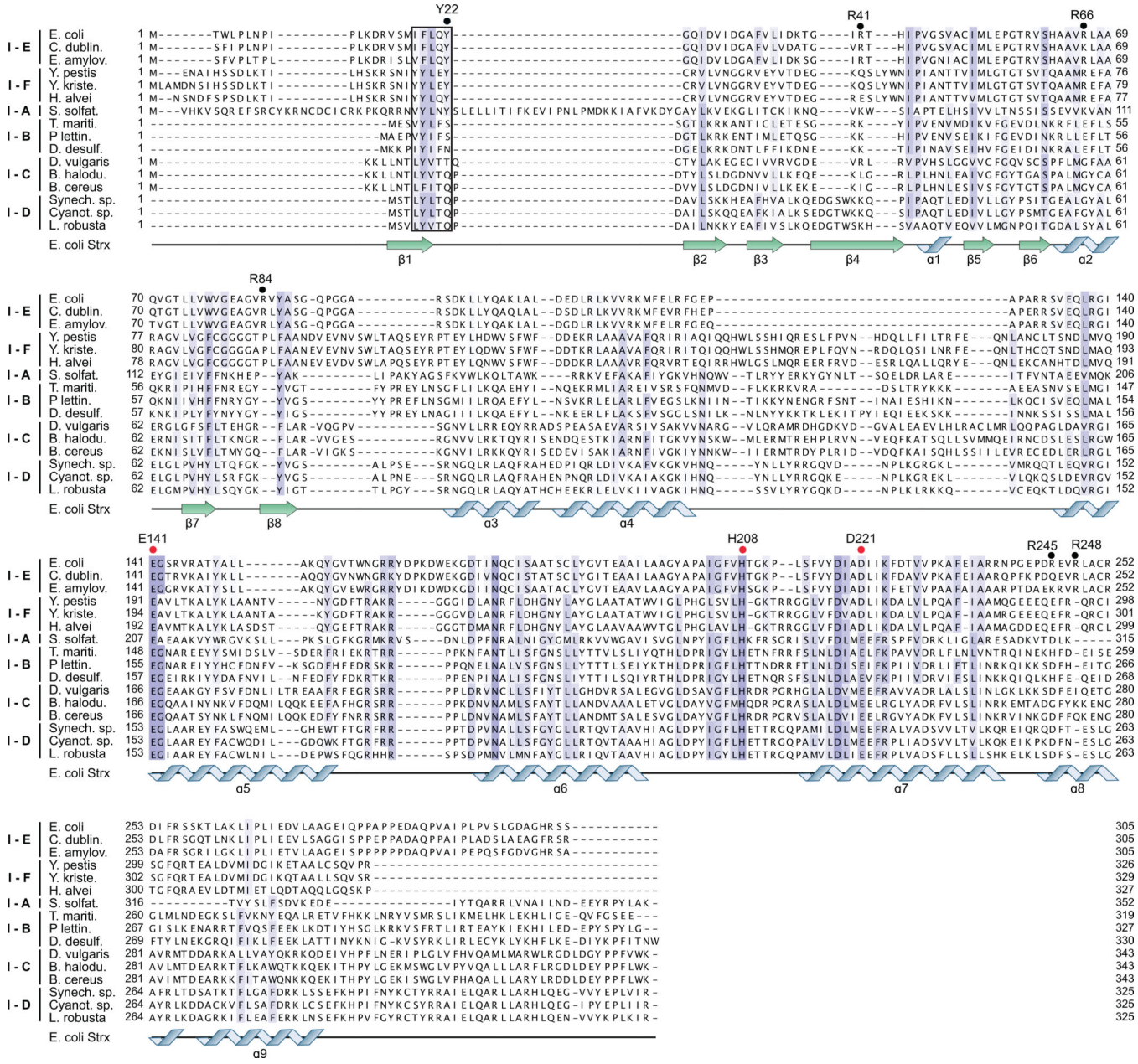
Extended Data Figure 3. Conformational dynamics upon protospacer DNA binding

a, An overlay of the DNA bound Cas1–Cas2 structure with the apo Cas1–Cas2 (gray, PDB 4P6I). **b**, Vector lines depicting the conformational changes the Cas1–Cas2 complex undergoes upon protospacer DNA binding compared to the apo complex (PDB 4P6I). The Cas1 subunits rotate towards the direction of the arrows.



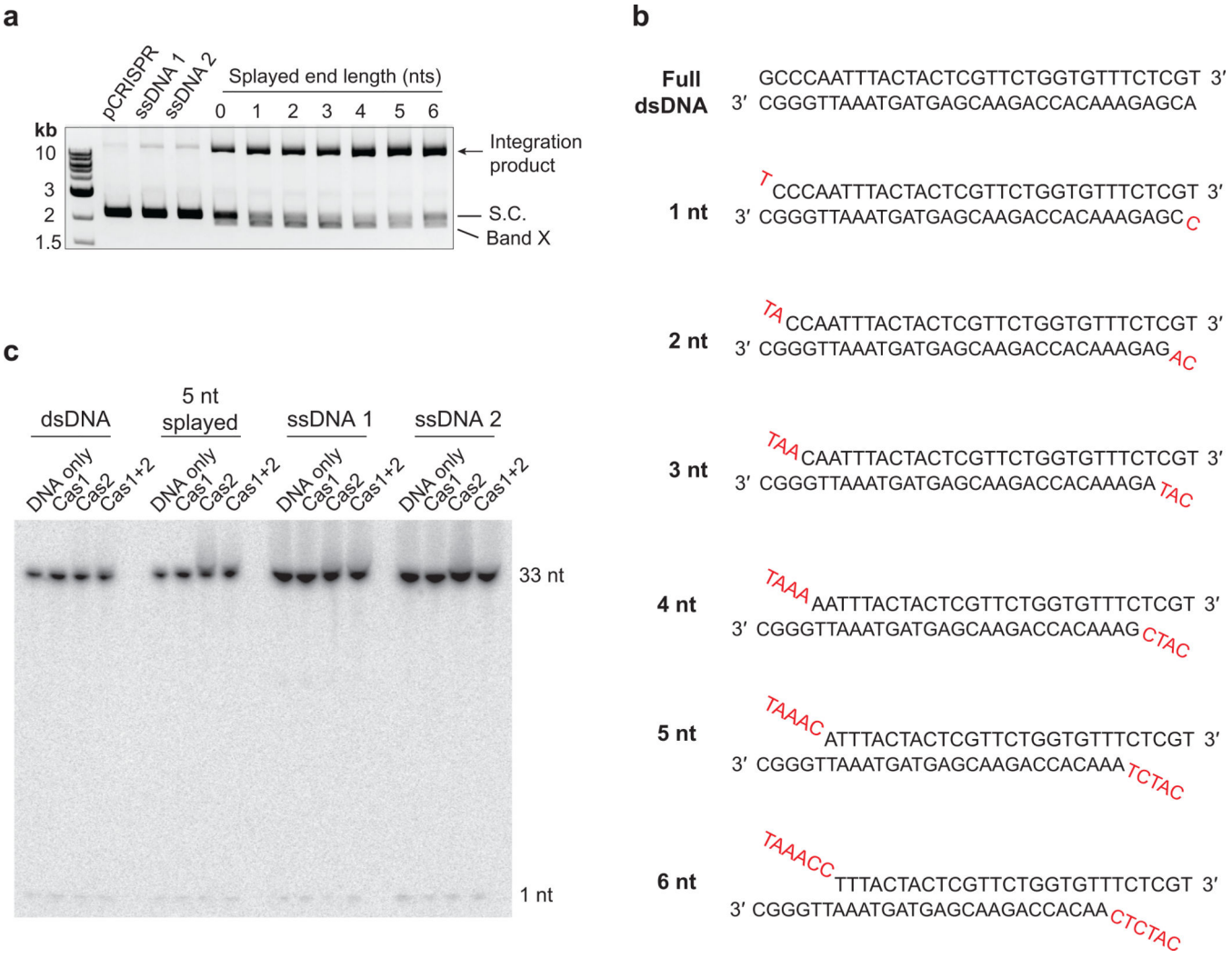
Extended Data Figure 4. Omit maps of the protospacer DNA

a, Simulated annealing F_o–F_c omit electron density map of the entire protospacer DNA using the "no Mg^{2+}" map and model. **b, c**, Simulated annealing F_o–F_c omit electron density maps of the terminal five nucleotides in the active sites of the structures (**a**) with Mg^{2+} or (**b**) without Mg^{2+} in the crystallization condition. The maps are contoured at 2.0 σ.



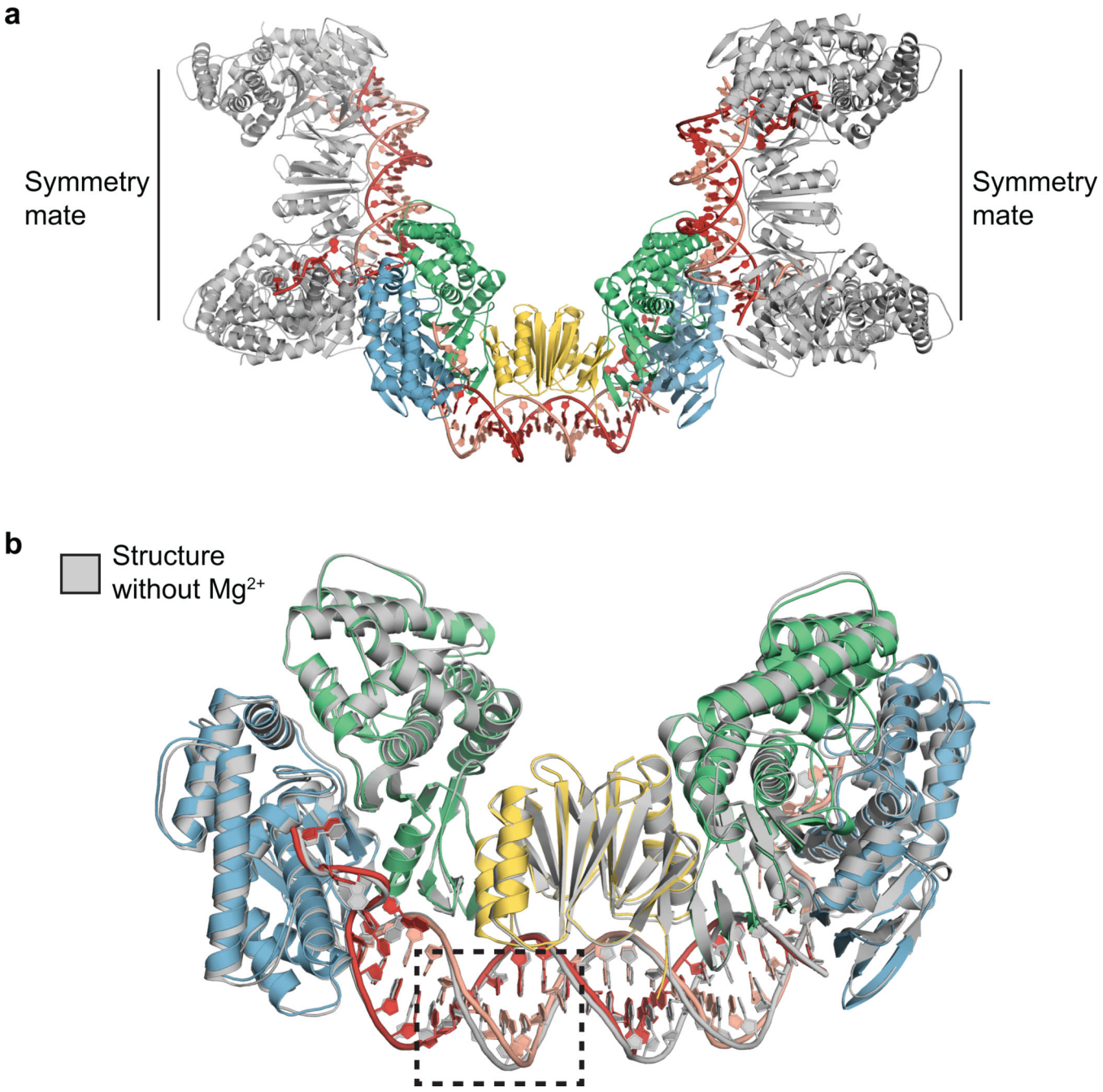
Extended Data Figure 5. Sequence alignment of Cas1 proteins in Type I CRISPR systems

Sequence alignments of Cas1 from representative organisms with Type I CRISPR systems. The *E. coli* sequence is displayed at the top. The dots indicate the residues described in this study, with the red dots indicating the metal-binding residues. The box highlights the non-universal conservation of the *E. coli* Y22 residue in the β1 region of Type I CRISPR systems. The secondary structure representations shown are for the *E. coli* Cas1.



Extended Data Figure 6. Integration of protospacer substrates with splayed ends

a, Representative agarose gel of *in vitro* integration reactions using increasing lengths of splayed ends. The average percent integration of three independent experiments is plotted in Fig. 3d. **b**, Sequences of protospacers used in the integration assays in **a. c**, A 12% denaturing polyacrylamide gel of protospacers after incubation with Cas1–Cas2 for 1 h at 37 °C in integration assay buffer conditions. The indicated DNA substrates are radiolabeled at the 5′ end. Supplementary Information contains the full images for **a** and **c**.



Extended Data Figure 7. Crystallographic packing of the complex bound to Mg^{2+}

a, View of the symmetry mates (gray) contacting the non-catalytic Cas1 subunits (green). **b**, Superposition of our two crystal structures, with or without Mg^{2+}, show a slight DNA kink in the structure bound to Mg^{2+} (dotted box). This region contacts α helix 7 of a symmetry mate, as described in the text.

Extended Data Table 1

Summary of X-ray crystallography data collection and refinement.

	Without Mg^{2+}	With Mg^{2+}	Splayed substrate
Data collection			
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$
Cell dimensions			
a, b, c, (Å)	88.02, 120.01, 196.01	75.66, 165.93, 167.26	88.02, 123.01, 196.01
α, β, γ (°)	90, 90, 90	90, 90, 90	90, 90, 90
Resolution (Å)	49.00–3.20 (3.36 – 3.20)	46.41–2.95 (3.06–2.95)	48.9–3.35 (3.42–3.35)
R_{merge} (%)	30.8 (146)	19.6 (157)	28.5 (126)
R_{pim} (%)	12.8 (61.4)	10.8 (86.3)	21.6 (94.3)
I/σ	6.4 (1.5)	9.8 (1.4)	5.0 (1.3)
$CC_{1/2}$	98.5 (72.4)	99.3 (42.0)	98.3 (72.7)
Completeness (%)	99.8 (99.0)	100 (99.9)	99.6 (97.7)
Redundancy	6.7 (6.6)	7.9 (8.0)	4.1 (4.0)
Wilson B factor ($Å^2$)	63.8	64.0	73.7
Refinement			
Resolution (Å)	49.00–3.20	46.41–2.95	49.00–3.35
No. reflections	35,808 (3,502)	44,960 (4,418)	31,049 (2885)
R_{work}/R_{free}	24.2/27.0	23.0/25.4	23.2/27.4
No. atoms			
Protein	9,375	9,576	9,375
DNA	1,142	1,142	1,165
Metal	0	4	0
Average B-factors ($Å^2$)			
Protein	65.9	66.6	86.6
DNA	76.2	67.2	103.0
Metal		51.6	
R.m.s deviations			
Bond lengths (Å)	0.003	0.003	0.004
Bond angles (°)	0.72	0.75	0.81
Ramachandran statistics (%)			
Favored	96.0	95.0	96.0
Allowed	3.75	4.51	3.58
Outliers	0.25	0.49	0.42

One crystal was used for each structure.
Highest resolution shell is shown in parenthesis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank G. Meigs and the 8.3.1 beamline staff at the Advanced Light Source for assistance with data collection, J. Chen for input on experimental design and members of the Doudna lab for comments and discussions. The 8.3.1 beamline is supported by UC Office of the President, Multicampus Research Programs and Initiatives grant MR-15-328599 and Program for Breakthrough Biomedical Research, which is partially funded by the Sandler Foundation. This project was funded by U.S. National Science Foundation grant No. 1244557 to J.A.D. and by NIH grant AI070042 to A.N.E. J.K.N. and L.B.H. are supported by U.S. National Science Foundation Graduate Research Fellowships and J.K.N. by a UC Berkeley Chancellor′s Graduate Fellowship. P.J.K. is supported as a Howard Hughes Medical Institute Fellow of the Life Sciences Research Foundation. J.A.D. is an Investigator of the Howard Hughes Medical Institute and a member of the Center for RNA Systems Biology.

References

1. Barrangou R, et al. CRISPR provides acquired resistance against viruses in prokaryotes. Science. 2007; 315:1709–1712. [PubMed: 17379808]
2. Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Soria E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. Journal of molecular evolution. 2005; 60:174–182. [PubMed: 15791728]
3. Bolotin A, Quinquis B, Sorokin A, Ehrlich SD. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology. 2005; 151:2551–2561. [PubMed: 16079334]
4. Pourcel C, Salvignol G, Vergnaud G. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, provide additional tools for evolutionary studies. Microbiology. 2005; 151:653–663. [PubMed: 15758212]
5. Garneau JE, et al. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature. 2010; 468:67–71. [PubMed: 21048762]
6. van der Oost J, Westra ER, Jackson RN, Wiedenheft B. Unravelling the structural and mechanistic basis of CRISPR-Cas systems. Nature reviews. Microbiology. 2014; 12:479–492. [PubMed: 24909109]
7. Yosef I, Goren MG, Qimron U. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucleic acids research. 2012; 40:5569–5576. [PubMed: 22402487]
8. Datsenko KA, et al. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. Nature communications. 2012; 3:945.
9. Swarts DC, Mosterd C, van Passel MW, Brouns SJ. CRISPR interference directs strand specific spacer acquisition. PloS one. 2012; 7:e35888. [PubMed: 22558257]
10. Nunez JK, et al. Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. Nature structural & molecular biology. 2014; 21:528–534.
11. Arslan Z, Hermanns V, Wurm R, Wagner R, Pul U. Detection and characterization of spacer integration intermediates in type I-E CRISPR-Cas system. Nucleic acids research. 2014; 42:7884–7893. [PubMed: 24920831]
12. Nunez JK, Lee AS, Engelman A, Doudna JA. Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. Nature. 2015; 519:193–198. [PubMed: 25707795]
13. Rollie C, Schneider S, Brinkmann AS, Bolt EL, White MF. Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition. eLife. 2015; 4
14. Heler R, Marraffini LA, Bikard D. Adapting to new threats: the generation of memory by CRISPR-Cas immune systems. Molecular microbiology. 2014; 93:1–9. [PubMed: 24806524]
15. Brouns SJ, et al. Small CRISPR RNAs guide antiviral defense in prokaryotes. Science. 2008; 321:960–964. [PubMed: 18703739]
16. Carte J, Wang R, Li H, Terns RM, Terns MP. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. Genes & development. 2008; 22:3489–3496. [PubMed: 19141480]
17. Haurwitz RE, Jinek M, Wiedenheft B, Zhou K, Doudna JA. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. Science. 2010; 329:1355–1358. [PubMed: 20829488]

18. Deltcheva E, et al. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature. 2011; 471:602–607. [PubMed: 21455174]
19. Jinek M, et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012; 337:816–821. [PubMed: 22745249]
20. Levy A, et al. CRISPR adaptation biases explain preference for acquisition of foreign DNA. Nature. 2015; 520:505–510. [PubMed: 25874675]
21. Wiedenheft B, et al. Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. Structure. 2009; 17:904–912. [PubMed: 19523907]
22. Savilahti H, Rice PA, Mizuuchi K. The phage Mu transpososome core: DNA requirements for assembly and function. The EMBO journal. 1995; 14:4893–4903. [PubMed: 7588618]
23. Scottoline BP, Chow S, Ellison V, Brown PO. Disruption of the terminal base pairs of retroviral DNA during integration. Genes & development. 1997; 11:371–382. [PubMed: 9030689]
24. Katz RA, Merkel G, Andrake MD, Roder H, Skalka AM. Retroviral integrases promote fraying of viral DNA ends. The Journal of biological chemistry. 2011; 286:25710–25718. [PubMed: 21622554]
25. Adams PD, et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta crystallographica. Section D, Biological crystallography. 2010; 66:213–221. [PubMed: 20124702]
26. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. Acta crystallographica. Section D, Biological crystallography. 2004; 60:2126–2132. [PubMed: 15572765]
27. Babu M, et al. A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. Molecular microbiology. 2011; 79:484–502. [PubMed: 21219465]
28. Makarova KS, et al. Evolution and classification of the CRISPR-Cas systems. Nature reviews. Microbiology. 2011; 9:467–477.
29. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013; 30:772–780. [PubMed: 23329690]

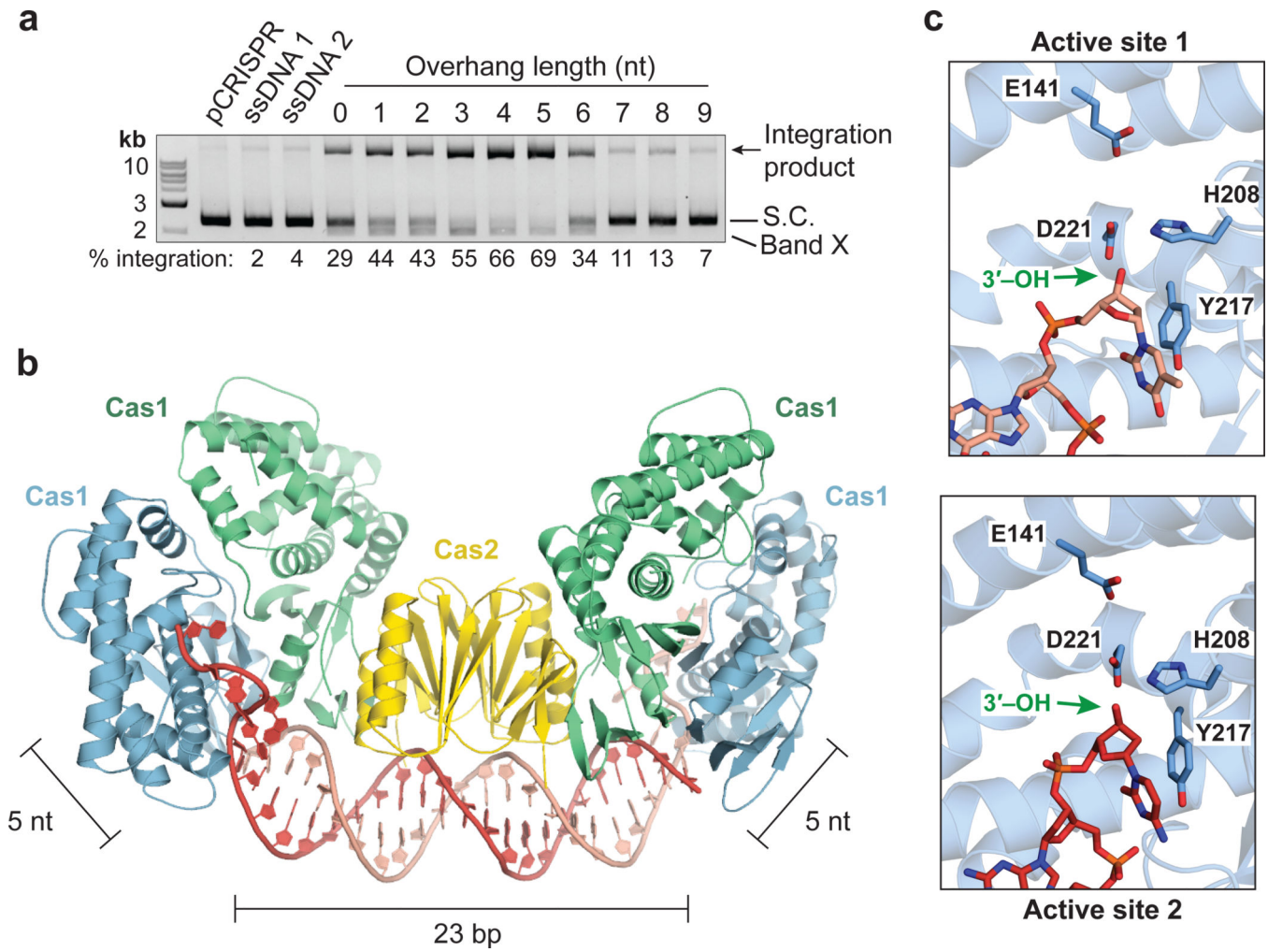


Figure 1. Overall architecture and active site positioning of 3′–OH nucleophile

a, A representative agarose gel of *in vitro* integration reactions using increasing lengths of 3′ single-strand protospacer DNA overhangs. Percent integration values are the average of three independent experiments. **b**, The overall architecture of Cas1–Cas2 bound to protospacer DNA. The line segments indicate DNA regions lengths, spanning a total of 33 nt. **c**, Stick configurations of the two Cas1 active sites (blue subunits in **b**) that coordinate the nucleophilic 3′–OH ends of the protospacer (green arrow). Supplementary Information contains the full image for **a**.

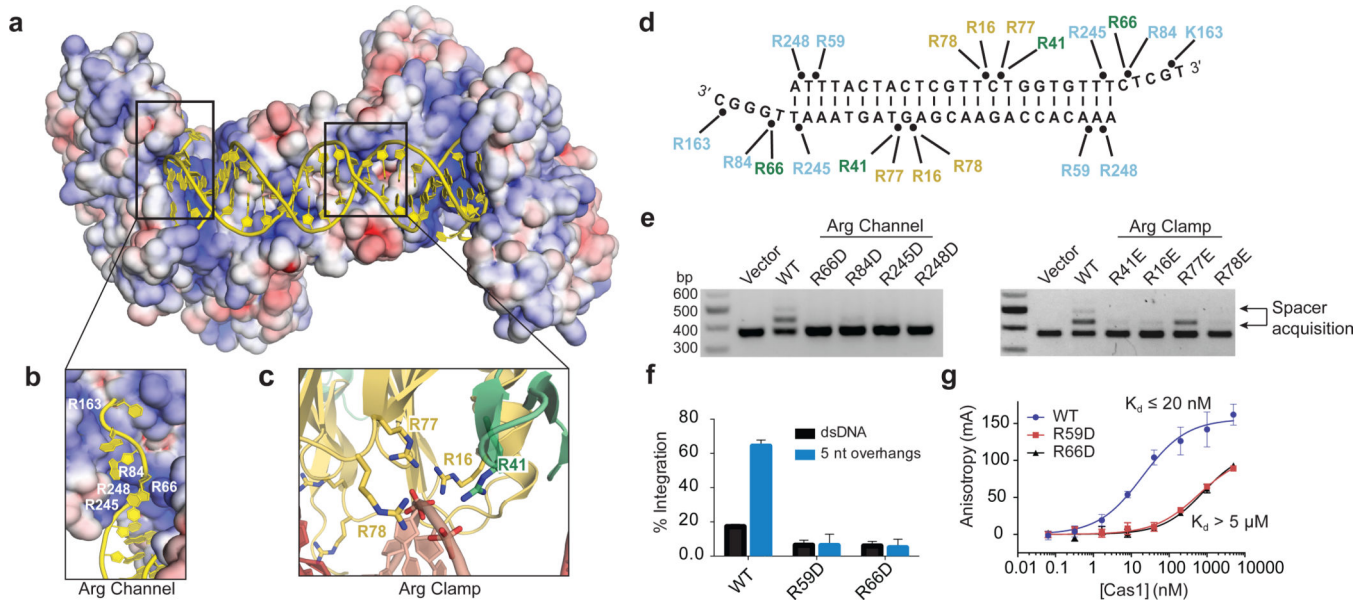


Figure 2. Coordination of protospacer DNA within the complex

a, Electrostatic potential surface representation of the Cas1–Cas2 complex with the protospacer shown in yellow. **b**, Close up of the Arginine Channel that stabilizes the ssDNA overhang. **c**, Stick configuration representation of Arginine Clamp residues that coordinate the protospacer duplex region. **d**, Map of amino acid residues that coordinate the protospacer phosphodiester backbone (black dots). Residue colors indicate Cas1–Cas2 protomers from Fig. 1b. **e**, Agarose gels of *in vivo* spacer acquisition assays of Arginine Channel and Clamp mutant proteins. **f**, Plot of percent *in vitro* integration of either dsDNA (black) or 5 nt overhang (blue) protospacers with Cas1 WT, R59D or R66D complexed with Cas2. **g**, Fluorescence polarization binding assays of a 5 nt overhang protospacer with the same mutants in **f** complexed with Cas2. The calculated relative binding affinities (K_d) are indicated. Error bars represent the standard deviation of three independent experiments. Panel **e–g** data are results of minimally three biological replicates. Supplementary Information contains the full images for **e**.

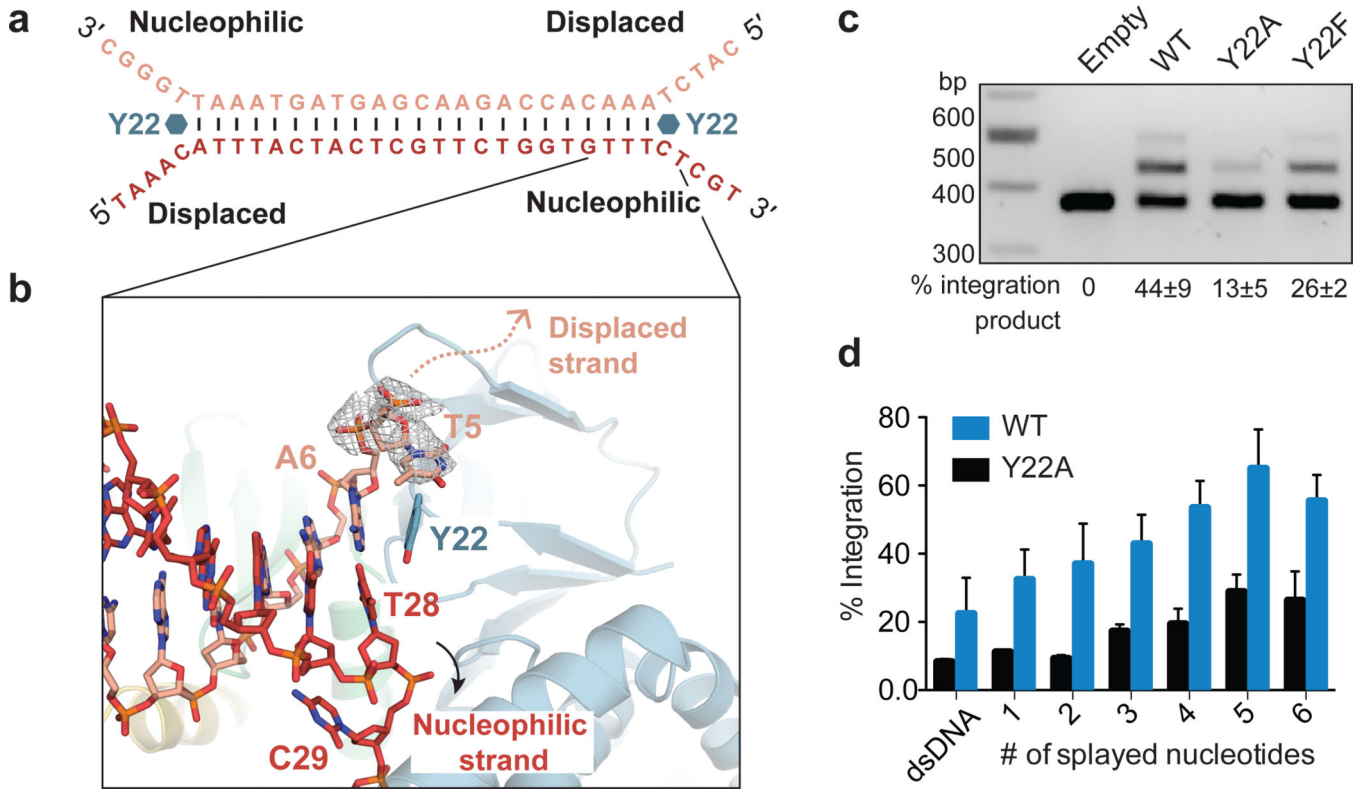


Figure 3. Mechanism of protospacer DNA end separation

a, The 5 nt splayed protospacer sequence used for crystallization to determine the trajectory of the displaced non-nucleophilic strand. Cas1 Y22, involved in base stacking at the fork, is shown in blue. **b**, Close up of the DNA fork showing the base stacking interaction of Y22 with the terminal adenine nucleotide of the non-nucleophilic strand. The nucleotides are numbered from 5′ to 3′ of each DNA strand shown in **a**. The gray mesh shows the $2F_o–F_c$ density contoured at 2.2 σ of the first ejected nucleotide of the displaced strand. The arrows indicate the opposite trajectories of each strand. **c**, Agarose gel of *in vivo* acquisition assay of co-expressed Cas1 WT or the indicated Cas1 mutant with Cas2. Quantification is the mean of three independent experiments ± standard deviation. **d**, Plot of percent integration of increasing splayed nt at the protospacer ends using Cas1 WT (blue) or Y22A (blue) complexed with Cas2. Error bars represent the standard deviation of three independent experiments. Supplementary Information contains the full image for **c**.

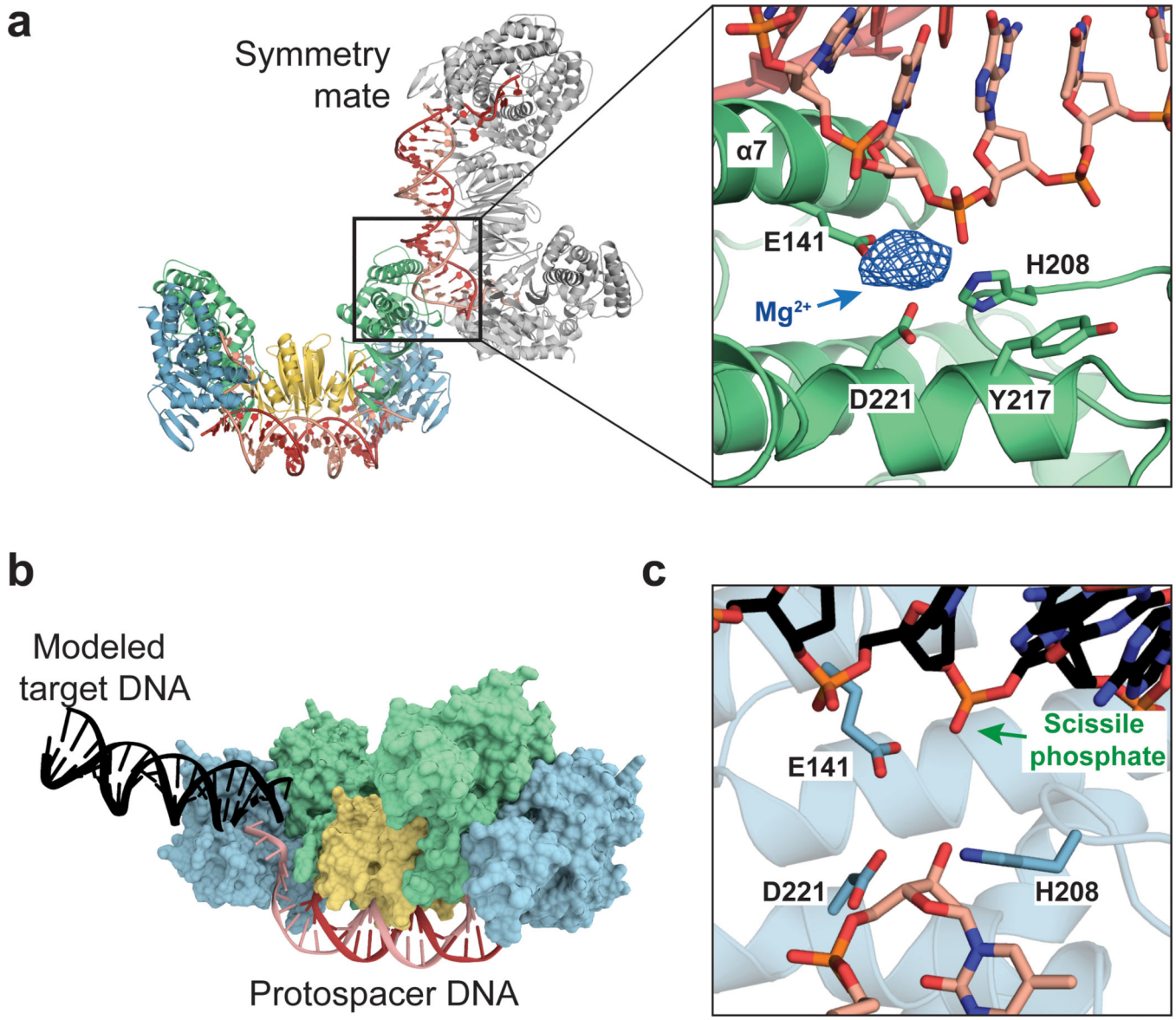


Figure 4. Model of protospacer DNA integration

a, View of crystal packing from a symmetry mate complex (gray) showing coordination of the symmetry DNA along a Cas1 active site. The inset is a close up of the coordination of the phosphodiester backbone with metal-binding residues E141, H208 and D221. The mesh represents a F_o–F_c density for a Mg^{2+} ion, contoured at 2.2 σ. **b, c**, Model of protospacer DNA integration into target DNA (black) and positioning of the scissile phosphate (green arrow) and the 3′–OH nucleophile in the Cas1 active site.