# Structural coordination between active sites of a CRISPR reverse transcriptase-integrase complex

Joy Y. Wang[1,2], Christopher M. Hoel[2-4], Basem Al-Shayeb[2,5], Jillian F. Banfield[5-7], Stephen G. Brohawn[2-4] and Jennifer A. Doudna[1-3,8-11*]

[1]Department of Chemistry, University of California, Berkeley, Berkeley, California, USA; [2]California Institute for Quantitative Biosciences, University of California, Berkeley, Berkeley, California, USA; [3]Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, California, USA; [4]Helen Wills Neuroscience Institute, University of California, Berkeley, Berkeley, California, USA; [5]Department of Plant and Microbial Biology, University of California, Berkeley, Berkeley, California, USA; [6]Department of Earth and Planetary Sciences, University of California, Berkeley, Berkeley, California, USA; [7]Department of Environmental Science, Policy, and Management, University of California, Berkeley, Berkeley, California, USA; [8]Innovative Genomics Institute, University of California, Berkeley, Berkeley, California, USA; [9]Howard Hughes Medical Institute, University of California, Berkeley, Berkeley, California, USA; [10]Molecular Biophysics & Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA; [11]Gladstone Institute of Data Science and Biotechnology, Gladstone Institutes, San Francisco, California, USA

*Correspondence to: doudna@berkeley.edu

**ABSTRACT**

CRISPR-Cas systems provide adaptive immunity in bacteria and archaea, beginning with integration of foreign sequences into the host CRISPR genomic locus and followed by transcription and maturation of CRISPR RNAs (crRNAs). In some CRISPR systems, a reverse transcriptase (RT) fusion to the Cas1 integrase and Cas6 maturase creates a single protein that enables concerted sequence integration and crRNA production. To elucidate how the RT-integrase organizes distinct enzymatic activities, we present the cryo-EM structure of a Cas6-RT-Cas1—Cas2 CRISPR integrase complex. The structure reveals a heterohexamer in which the RT directly contacts the integrase and maturase domains, suggesting functional coordination between all three active sites. Together with biochemical experiments, our data support a model of sequential enzymatic activities that enable CRISPR sequence acquisition from RNA and DNA substrates. These findings highlight an expanded capacity of some CRISPR systems to acquire diverse sequences that direct CRISPR-mediated interference.

## INTRODUCTION

CRISPR-Cas (clustered regularly interspaced short palindromic repeats-CRISPR associated) systems provide adaptive immunity against viruses for bacteria and archaea[1,2]. This process begins when an integrase comprising the conserved Cas1 and Cas2 proteins incorporates segments of foreign DNA – protospacers – into the host CRISPR locus, which consists of direct repeats separated by virally-derived spacers[3–13]. Transcription and transcript processing generate CRISPR RNAs (crRNAs), which assemble with Cas proteins to form effector complexes that identify and destroy foreign nucleic acids bearing sequence complementarity to the ~20-nucleotide (nt) spacer segment of the crRNA[14–19].

Although most CRISPR spacer sequences are derived from phage DNA, a minority of CRISPR-Cas systems include a reverse transcriptase (RT), enabling spacer acquisition from RNA sources[20–29]. In some cases, the RT occurs as a fusion protein together with Cas1 on the RT's C-terminus that can facilitate DNA-based storage of transcriptional information and has been used as a tool to record cellular transcriptional activity[29]. RT-Cas1s occur primarily in type III CRISPR-Cas systems that target both RNA and DNA, raising the interesting possibility that these variants provide adaptive immunity against RNA phage or other RNA elements[16,30–33]. Notably, many of these RT-Cas1 fusion proteins include an N-terminal Cas6 domain[24,26,34]. The endoribonuclease Cas6 functions as a maturase that recognizes the CRISPR repeat and processes CRISPR transcripts into mature crRNAs[35-37]. A recent study showed that the Cas6 domain of a Cas6-RT-Cas1 fusion is required for RNA spacer acquisition and co-evolves with the RT domain[34]. However, the molecular basis for functional coordination between Cas1, RT and Cas6 remains unknown.

In this work, we determine the 3.7 Å resolution structure of a naturally-occurring *Thiomicrospira* Cas6-RT-Cas1—Cas2 complex using single particle cryo-electron microscopy (cryo-EM) and investigate its activities *in vitro* to elucidate how RT activity might coordinate with the CRISPR integrase and with the Cas6 maturation nuclease in the context of a Cas6-RT-Cas1

fusion protein. We show that in spite of a difference in arrangement of the Cas1—Cas2 disposition compared to the well-studied *E. coli* integrase, the Cas6-RT-Cas1—Cas2 complex catalyzes site-specific full-site CRISPR spacer integration of double-stranded DNA (dsDNA) substrates. The unique RT conformation observed in the context of the CRISPR integrase fusion protein, in addition to direct interactions between the RT, integrase, and maturase active sites, suggests that structural rearrangements coordinate different activities and substrates during CRISPR sequence acquisition. It is unclear whether RNA serves as the primary source of spacers for these systems, and surprisingly, we find that the RT-integrase favors DNA as an integration substrate and conducts complementary DNA (cDNA) synthesis using both RNA and DNA templates that have minimal homology between primer and template. Together, these findings underscore the versatility of CRISPR RT-integrase proteins to provide for expanded rather than specialized CRISPR spacer acquisition.

## RESULTS

### Architecture of the Cas6-RT-Cas1—Cas2 complex

To investigate the structure of a Cas6-RT-Cas1—Cas2 complex and its activities, we purified a Cas6-RT-Cas1 and its associated Cas2 protein, which are part of a *Thiomicrospira* type III CRISPR locus identified from a groundwater, $CO_2$-driven Geyser dataset (Fig. 1a). To form the complex, the Cas6-RT-Cas1 and Cas2 proteins were assembled with a DNA substrate designed to mimic a genomic integration intermediate[10] (Fig. 1a,b, Supplementary Fig. 1a,b). The structure of the resulting Cas6-RT-Cas1—Cas2 complex was determined by single particle cryo-EM at an overall resolution of 3.7 Å (Fig. 1b-d, Supplementary Fig. 2a,b, 3a-h). Approximately half of the complex displayed relatively lower local resolution, and refinement using a mask that excluded this region generated a reconstruction at 3.4 Å resolution with improved interpretability of high-resolution features. (In the remainder of the text, we refer to these two reconstructions as full complex and masked partial complex, respectively.) A model

corresponding to two Cas2 chains and portions of four Cas6-RT-Cas1 chains was *de novo* built and refined into the higher resolution masked partial complex reconstruction. Portions of three Cas6-RT-Cas1 chains that were excluded by the mask were then rigid body docked and refined in the full complex reconstruction. The final model consists of two entire Cas2 chains, two entire Cas6-RT-Cas1 chains (excluding 5 poorly resolved loops), and the Cas1 domain from two additional Cas6-RT-Cas1 chains.

As observed for previous CRISPR integrase structures[10,12,13], the Cas6-RT-Cas1—Cas2 complex is heterohexameric, consisting of a central Cas2 dimer and two distal Cas6-RT-Cas1 dimers (Fig. 1b-d). The heterohexameric architecture of the core Cas1—Cas2 is a hallmark of CRISPR integrases and is critical to their internal ruler mechanism dictating the length of the spacers that are integrated and their ability to catalyze full-site integration[10,12,13]. Extending from the Cas1—Cas2 core are the Cas6-RT lobes, which are positioned orthogonal to the Cas1 domains relative to the central Cas2 dimer. The Cas6-RT lobe is connected to its associated Cas1 domain via a flexible linker, with the RT domain proximal to the central Cas2 dimer and the Cas6 domain closely abutting the RT domain (Fig. 1c-e). We observed density for the Cas6-RT lobe in two of the four Cas6-RT-Cas1 protomers (Cas1b and b'), presumably due to flexibility of the Cas6-RT domains in the other two Cas6-RT-Cas1 protomers (Cas1a and a'). No density was observed for the DNA substrate despite co-migration during size-exclusion chromatography and indication of binding from electrophoretic mobility shift assays, suggesting that DNA was bound in only some of the complexes collected or that the DNA was lost during grid preparation or is too flexible to see (Supplementary Fig. 1a,b).

The Cas1 and Cas2 domains largely resemble those of the *E. coli* Cas1—Cas2 complex; however, there are some notable differences. When compared to the *E. coli* Cas1—Cas2 complex structure, the C-terminal α-helical domains of a given Cas1 are very similar (2.3 Å RMSD), while the linker and N-terminal β-sheet domains are offset by ~5-7 Å (Supplementary Fig. 4a,b). Each Cas2 protomer has a ferredoxin-like fold, as observed in previous Cas2 structures[38-40]. However,

in contrast to the *E. coli* Cas1—Cas2 complex[6], the β-sheet in Cas2 is composed of three β-strands, and the C-terminal domain following the third β-strand crosses the central axis to pack against the face of the opposing subunit, resulting in an altered dimer interface. Additionally, the C-terminal segment (comprising amino acids 88-96) interacts with Cas1b rather than with Cas1b' as in the *E. coli* Cas1—Cas2 structure.

Previous studies have shown that CRISPR-associated RTs are most closely related to the RTs encoded by group II introns[24-27]. We observe that the RT domain shares the characteristic palm and fingers regions of the canonical right-handed fold in other RT structures but is missing the thumb/X subdomain that is present in retroviral and group II intron RTs[41-43] (Supplementary Fig. 4c). Compared to the group II intron RT structure[41], the finger domains are very similar with an overall RMSD of 1.9 Å, while the palm domains are strikingly different. Interestingly, in the present structure, the palm domain is reoriented such that the FADD motif, containing active site residues conserved across diverse RT subtypes, points away from the substrate binding cleft, suggesting that this conformation may not be competent for reverse transcription.

Finally, the structure of the Cas6 domain closely resembles those of previous Cas6 structures[34,36,44-47] (Supplementary Fig. 4d). When compared directly to the structure of the Cas6 domain from the *Marinomonas mediterranea* (MMB-1) Cas6-RT-Cas1 protein[34], the two structures align with RMSD of 2.4 Å, with the greatest variance in the β-strand of the C-terminal RNA recognition motif (RRM) fold, perhaps due to the subsequent connection to the RT domain, absent in the MMB-1 Cas6 structure (Supplementary Fig. 4e).

**Substrate preferences of Cas6-RT-Cas1—Cas2 for cleavage-ligation**

Previous biochemical and genomic experiments demonstrated both RT and integration activities for the MMB-1 RT-Cas1 fusion protein[27,34]. To test the integrase function of the *Thiomicrospira* Cas6-RT-Cas1—Cas2 complex, we first investigated its substrate preferences for integration into a target DNA molecule and the structural elements that may be involved in

the cleavage-ligation reactions. We conducted integration assays by incubating the purified Cas6-RT-Cas1 and Cas2 proteins with a 5'-fluorophore-labeled 35-base pair (bp) dsDNA or a 35-nt single-stranded DNA (ssDNA) or a single-stranded RNA (ssRNA) substrate and a supercoiled plasmid containing the CRISPR locus that functions as a target for spacer integration. Deoxynucleoside triphosphates (dNTPs) were supplied to reactions with RNA substrates. We observed protospacer ligation by examining the incorporation of the fluorophore-labeled substrate into the CRISPR locus in the target plasmid (Fig. 2a). The results show that the Cas6-RT-Cas1—Cas2 complex can catalyze ligation of all three substrates into the target plasmid (Fig. 2b). Interestingly, the Cas6-RT-Cas1 protein alone is able to catalyze protospacer ligation with DNA substrates. While Cas2 alone shows no activity, its presence increases the cleavage-ligation efficiency of the complex. Time-course experiments that quantified the extent of plasmid-nicking in each reaction showed >99% and 77% conversion from supercoiled plasmid target to open circular product with dsDNA and ssDNA, respectively, 25% with ssRNA, and 13% with no protospacer over two hours (Fig. 2c, Supplementary Fig. 5a). We conducted additional time-course experiments with two DNA/RNA hybrid substrates: one hybrid consisting of a 35-nt RNA oligonucleotide and a shorter 25-nt DNA strand and the second hybrid consisting of the same 35-nt RNA oligonucleotide and its fully complementary 35-nt DNA strand. This resulted in an 89% and 91% conversion from supercoiled plasmid target to open circular product with the first and second hybrid, respectively, after two hours in the absence of dNTPs and a 95% and 97% conversion with the first and second hybrid, respectively, after two hours in the presence of dNTPs (Supplementary Fig. 5b). These results show that this *Thiomicrospira* Cas6-RT-Cas1—Cas2 integrase is significantly more efficient with DNA and DNA/RNA hybrid substrates than ssRNA substrates for cleavage-ligation and that dNTPs do not seem to be required for cleavage-ligation activity.

We wondered if there are elements near the Cas1 active site that enable RT-associated CRISPR integrases to catalyze cleavage-ligation chemistry with RNA substrates, a process not

observed for non-RT-associated Cas1s. We found a cluster of residues (R832, R834, R835, and H879*)* adjacent to the active site that undergoes a large conformational change between Cas1a and Cas1b. In the non-catalytic Cas1b domains, these residues appear to coordinate an unknown density, potentially corresponding to a metal ion or water molecule (Fig. 2d,e). In the catalytic Cas1a domains, this density is not observed and the residues of interest are pushed away from each other, seemingly to accommodate the insertion of an α-helix from the RT domain. Sequence alignments across a diverse sample of RT-associated Cas1 variants show that there is ~100%, ~60%, ~70%, and ~35% conservation for R832, R834, R835, and H879, respectively, across RT-associated Cas1s (Supplementary Fig. 5c). The latter three residues show only a slightly lower sequence conservation across the selection of non-RT-associated Cas1s, though they are not conserved in the classical *E. coli* Cas1 (Supplementary Fig. 4a,b). Given our structural observations, we wondered if this cluster of residues may play an important role for the function of Cas6-RT-Cas1.

To examine the potential role of R835, one of the residues that coordinates the unknown density and undergoes a large conformational change, in protospacer ligation, we conducted integration assays with the Cas1-R835A mutant and the Cas1 active site mutant Cas1-H873A. We quantified the relative amounts of protospacer ligation to parallel wild-type (WT) controls. As expected, the Cas1-H873A mutant abolishes almost all protospacer ligation (Fig. 2f). Interestingly, the Cas1-R835A mutant also affected ligation activity, with a 65% and 55% reduction for dsDNA and ssDNA ligation and >99% reduction for ssRNA. Together, these findings identify R835 as a potential structural element that contributes toward Cas6-RT-Cas1—Cas2's ability to accommodate RNA substrates for protospacer ligation.

To determine whether the *Thiomicrospira* Cas6-RT-Cas1—Cas2 complex can catalyze cleavage-ligation in a site-specific manner (at the ends of the repeat sequence in the CRISPR array) with all three substrates[4,7], we conducted integration reactions using a short linear dsDNA molecule comprising 49 bp of the leader sequence, the 35-bp repeat, and 15 bp of the adjacent

spacer as the target for integration of the substrate (Supplementary Fig. 6a). The results show that the predominant cleavage-ligation product corresponds to nucleophilic attack of the protospacer at the spacer end of the repeat (the expected site for integration) for all three protospacers (Supplementary Fig. 6b-d). This suggests that the integrase complex has some intrinsic target recognition, although the numerous products resulting from off-target nucleophilic attacks indicate that other factors may be required to improve specificity. Curiously, while previous work on the MMB-1 system suggests that the cleavage-ligation products are potential substrates for target-primed reverse transcription[27,48,49], we do not observe any bands that represent extension of the 3' end of the DNA after cleavage-ligation when dNTPs are supplied (Supplementary Fig. 6c). Furthermore, unlike what is observed for the MMB-1 system, dNTPs are not required for RNA ligation. The similar pattern of ligation products for all three substrates suggests that the mode for target recognition is likely the same with both RNA and DNA protospacers, although other factors likely contribute to its difference in efficiency.

**Cas6-RT-Cas1—Cas2 exhibits length selectivity for ssDNA and dsDNA substrates**

At the heart of the Cas6-RT-Cas1—Cas2 structure lies the heterohexameric Cas1—Cas2 core, which has some interesting differences from the *E. coli* integrase structure, raising questions about its functions as a CRISPR integrase. Two positively charged regions of the *E. coli* Cas1—Cas2 structure are critical for protospacer substrate binding: the Arginine Channel, which contacts the DNA substrate where the duplex terminates and the single-stranded overhang enters the active site, and the Arginine Clamp, which stabilizes the middle of the duplex[12]. Although similar charged regions occur in the Cas1 and Cas2 dimers of the Cas6-RT-Cas1—Cas2 structure, the Cas2s and Cas1a'/Cas1b' dimer are rotated further away from the Cas1a/Cas1b such that these charged regions are no longer in the same linear plane as observed in the *E. coli* Cas1—Cas2 structures[10,12,50] (Fig. 3a,b). The differences in the arrangement of the Cas1—Cas2 components relative to their observed positioning in the *E. coli*

Cas1—Cas2 integrase led us to test whether the Cas6-RT-Cas1—Cas2 complex retains the intrinsic ruler mechanism characteristic of CRISPR integrases. Results of integration assays performed with 15- to 115-nt ssDNA and dsDNA substrates showed that the complex has some length selectivity for ssDNA and dsDNA. It is most active with 15- to 55-nt substrates and its activity decreases with substrates longer than 55 nt (Fig. 3c,d). This distribution is larger than the range of spacers observed in the CRISPR array (30-47 nt), though it is not surprising since half-site reactions tend to accept a larger variety of substrate lengths than full-site reactions[4]. What is more surprising is that the length distribution is similar for ssDNA and dsDNA protospacers, suggesting that the complex is able to narrow the selection of protospacer lengths in a manner that does not depend on interaction with a DNA duplex or access to two 3' nucleophilic ends.

**Cas6-RT-Cas1—Cas2 catalyzes full-site integration of dsDNA substrates**

We next set out to determine whether the Cas6-RT-Cas1—Cas2 complex can catalyze full-site CRISPR spacer integration. We devised an assay using chloramphenicol resistance as a selection marker for plasmids with protospacer insertion events near the target region[51,52]. The reporter construct was designed with 163 nt of the leader sequence and the CRISPR repeat sequence immediately upstream of a chloramphenicol resistance gene with a missing ribosomal binding site (RBS) sequence and start codon (Fig. 4a). Insertion of a spacer supplying the missing RBS sequence and start codon to the target region allows translation of the chloramphenicol resistance gene transcript. We performed *in vitro* integration assays with this reporter plasmid and a protospacer with an RBS sequence and start codon, transformed the integration products into *E. coli*, plated the transformants on agar containing chloramphenicol, and sequenced the surviving colonies.

The results show that the Cas6-RT-Cas1—Cas2 complex catalyzes full-site integration with dsDNA substrates. Full-site integration events, characterized by spacer insertion followed

by duplication of the adjacent 35 bp, represent the majority of the insertion events (14 out of 25) (Fig 4b,c). The other insertion events appear to be the result of incomplete or abortive integration, characterized by partial/whole spacer insertions followed by a deletion of variable length of the adjacent sequence. While the incomplete/abortive integration events occurred at many off-target locations, full-site integration events were highly specific, with 12 out of the 14 full-site integration sites located at the leader-repeat junction. This suggests that while the complex might attempt integration at off-target sites, it maintains specificity for full-site integration. We did not observe full-site integration with ssDNA or ssRNA protospacers. It is possible that the Cas6-RT-Cas1—Cas2 complex alone can only conduct half-site reactions with single-stranded protospacers *in vitro* and other host factors may be necessary to support full-site integration with those substrates.

**Unique RT conformation results in close contact with Cas1 active site**

Although the RT domain has been implicated in RNA-derived CRISPR spacer integration[27], its mechanism and possible coordination with Cas1 is not known. Similar to other RT structures[41-43], the RT active site of the Cas6-RT-Cas1—Cas2 complex resides in the palm region and consists of three conserved aspartate residues located on a three-stranded antiparallel β-sheet (Fig. 5a-c). Structural alignment with the group II intron RT reveals that despite close alignment in the fingers region, the palm region of the RT domain is dramatically offset, resulting in a ~90° rotation of the β-sheet containing the active site[41] (Fig. 5b,c). Possibly as a result, the two active site aspartates between β6 and β7 are further removed from the third active site aspartate between β5 and α5. This rotation is surprising given that there is comparatively little variation among the group II intron RT and other retroviral RT active sites[41-43], raising the possibility that the current conformation may represent an inactive conformation.

In solution, however, Cas6-RT-Cas1 alone catalyzes reverse transcription, an activity that is enhanced upon addition of Cas2 (Fig. 5d). Using substrates consisting of a 25-nt labeled

DNA primer annealed to a 35-nt DNA or RNA template, we observe extension of the primer from 25 to 35 nt, corresponding to the full template length, when dNTPs are supplied (Fig. 5e). There are some unexpected dNTP-independent products around 50-55 nt that do not form when the Cas1 active site mutant is used, suggesting that they are cleavage-ligation products (Fig. 5e,f, Supplementary Fig. 7a). Interestingly, in the lane that has only the DNA primer as a substrate, a smeary dNTP-dependent band forms around 45-48 nt, which matches the length of a cDNA product that forms when a second copy of the 25-nt DNA substrate is used as a template for primer extension with only 1-2 bp homology. A follow-up experiment using a 5'-labeled 35-nt DNA or RNA oligonucleotide with the original template sequence also resulted in template-dependent extension of the substrate with only 1 bp homology (Supplementary Fig. 7b). These results suggest that the Cas6-RT-Cas1—Cas2 complex is able to catalyze primer extension with DNA and RNA templates with minimal homology requirements.

To examine RT fidelity, we conducted dNTP drop-out experiments and supplied specific dideoxynucleoside triphosphates (ddNTPs) to terminate primer extension. The results show that Cas6-RT-Cas1—Cas2 complex is faithful to the RNA template, terminating cDNA synthesis at the correct position along the sequence where a necessary dNTP is missing or after a ddNTP is incorporated (Fig. 5f). Comparisons between experiments using the DNA template and those using the RNA template indicate that the complex is more efficient in carrying out primer extension with the DNA template than the RNA template; however, the complex is slightly more error-prone with the DNA template with bands suggesting some misincorporation of additional dNTPs. The results demonstrate that while the visible RT conformation is likely inactive, the complex can carry out primer extension with both RNA and DNA templates, with slightly higher fidelity for RNA than DNA.

**Structural elements contribute to crosstalk between Cas6, RT, and Cas1 active sites**

The structure reveals intriguing elements suggesting potential interplay between the Cas1, RT, and Cas6 domains. The most notable is an RT α-helix that serves as a direct connection between the RT and Cas1 active sites. While the N-terminal end of the RT-helix α9 is attached to the β-sheet containing the RT active site, the C-terminal end points into the catalytic Cas1a active site center (Fig. 6a,b). The helix insertion may be directly related to the conformational difference between Cas1a and Cas1b shown earlier. Structural alignments with other substrate-bound Cas1—Cas2 structures suggest that the helix insertion would result in considerable steric obstruction for the protospacer and target, perhaps even blocking access to the Cas1 active site entirely (Supplementary Fig. 8a,b). The RT-helix would likely need to pull out in order for Cas1 to catalyze integration. We hypothesize that this RT-helix motion may be associated with the rotation of the connected β-sheet, which may bring the RT active site closer to the canonical conformation seen in other RT structures[41-43]. A closer look at the region around the RT-helix reveals additional clues on how this movement could potentially be regulated. The flexible RT-Cas1 linker wraps around the RT-helix, raising the possibility of its involvement in aiding the helix's movement. The RT-helix also interacts with Cas2, suggesting that rearrangement of the Cas2s could also impact the position of the RT-helix and the attached β-sheet containing the RT active site. An examination of continuous heterogeneity within the cryo-EM data with 3D variability analysis in cryoSPARC is consistent with complex motion in this region (Supplementary Movie 1).

To investigate the functional implications of the RT-helix in the crosstalk between the RT and Cas1 domains, we generated a panel of Cas6-RT-Cas1 mutants and tested them in protospacer ligation and RT activity assays. In addition to the Cas1 residues H873 and R835, we also mutated the RT active site residue D540 and the RT-helix lysine (K574), which points into the Cas1a active site. Here, ligation activity is shown as the percent ligation into the target

plasmid by the mutant Cas6-RT-Cas1—Cas2 complex relative to a parallel WT control (Fig. 6c, Supplementary Fig. 8c). Interestingly, while the RT-D540A and RT-K574A mutants result in a 20-33% reduction in DNA ligation, they reduce RNA ligation by more than 80%. It appears that impairing either RT activity or the RT-helix-Cas1a interaction substantially disrupts RNA ligation but not DNA ligation.

The results from the RT activity assays show further evidence of crosstalk between the RT and Cas1 domains. Reverse transcriptase activity levels for the different mutant Cas6-RT-Cas1—Cas2 complexes showed that mutating either D540 or K574 abolishes RT activity. The Cas1 domain mutations also show significant effects: the Cas1-H873A and Cas1-R835A mutants result in a 41% and 53% reduction in dNTP incorporation, respectively, after two hours (Fig. 6d). These results show that despite being far away from the RT active site, the RT-helix lysine as well as the Cas1 residues potentially involved in the RT-helix—Cas1 interaction are significant for RT activity, supporting the hypothesis that the RT-helix plays an essential role in regulating RT activity. Together, these results suggest that there is bidirectional crosstalk between the Cas1 and RT domains.

We next explored the potential crosstalk between the Cas1/RT domains and the Cas6 domain, which lies on the opposite side of the RT. The results show that the Cas6 active site mutant Cas6-R37A reduces RT activity by 57% compared to the WT (Fig. 6d). Interestingly, the Cas6-R37A mutant increases DNA ligation to more than 176% that for the WT while reducing RNA ligation by 74% compared to the WT (Fig. 6c). Though the increase in DNA ligation is puzzling, the Cas6 active site mutant may reduce RNA ligation by way of its effect on RT activity, which is more critical for RNA ligation than for DNA ligation. These results show that Cas6 is not just operating on its own and influences the other domains in the complex.

We finally tested the mutants alongside the WT Cas6-RT-Cas1 in Cas6 processing activity assays. The Cas6-RT-Cas1—Cas2 complex was incubated with a 5’-labeled 35-nt RNA substrate corresponding to the CRISPR repeat sequence. The results show that the WT Cas6-

RT-Cas1—Cas2 processes the 35-nt RNA substrate down to a 26-nt product with 97% efficiency after two hours (Fig. 6e, Supplementary Fig. 8d,e). Mutating the Cas6 active site residue R37 reduces processing efficiency to only 12%. The other mutants show no difference in processing relative to the WT, suggesting that the crosstalk between the Cas6 domain and the other two domains is unidirectional, with the Cas6 active site mutation affecting ligation and RT activity but the Cas1 and RT domain mutations having no effect on Cas6 RNA processing activity.

## DISCUSSION

RT-Cas1 and associated Cas2 proteins represent a unique CRISPR adaptation module shown to be involved in the acquisition of CRISPR spacers directly from RNA in a fascinating example of information flow from RNA to DNA[20,22-29,48]. In this study, we identify a type III *Thiomicrospira* Cas6-RT-Cas1—Cas2 complex containing three active sites for CRISPR spacer integration, reverse transcription, and RNA processing, and we observe all three activities in our reconstituted system. The structure of this Cas6-RT-Cas1—Cas2 complex revealed interesting differences between its Cas1—Cas2 integrase core and that observed for the *E. coli* Cas1—Cas2 integrase. An RT-helix links the RT and Cas1 active sites but also appears to block the position of substrate binding. Functional interactions between the Cas6, RT, and Cas1 domains show bidirectional crosstalk between the RT and Cas1 domains and unidirectional crosstalk from Cas6 to the other two domains.

Although the *Thiomicrospira* Cas6-RT-Cas1—Cas2 complex catalyzes site-specific cleavage-ligation chemistry with dsDNA, ssDNA, and ssRNA substrates, we only observe full-site integration with dsDNA substrates. Its inability to fully integrate single-stranded substrates even in the presence of dNTPs implies that additional host factors may be required to carry out this process *in vivo*. It is still unclear when the RNA substrate is reverse transcribed into cDNA during spacer acquisition. We found that the RT faithfully synthesizes cDNA from

primer/template hybrids with short ~35-nt RNA or DNA templates and with minimal homology requirements; however, the preferred template and primer for these systems *in vivo* remain unknown. The difference in efficiency between the DNA polymerase activity and the RT activity is intriguing, because it suggests the possibility that the RT domain is primarily functioning as a DNA polymerase, although this result may also simply be a reflection of the tradeoff between efficiency and accuracy. Though previous work on the MMB-1 system suggested that the cleavage-ligation products could be substrates for target-primed reverse transcription similar to group II intron retrohoming[27,48,49], we did not see evidence of this in biochemical assays. If such a process were to occur *in vivo*, it is interesting to speculate whether the twisted arrangement of the Cas1—Cas2 components relative to their observed positioning in the *E. coli* Cas1—Cas2 integrase might be advantageous to stall integration at some intermediate step to allow access for the RT.

The structural and functional interactions detected between the Cas1, RT, and Cas6 active sites suggest a model in which the linking RT-helix serves as a regulator of both RT and Cas1 activities through its movement in and out of the Cas1 active site (Fig. 7a,b). Given the RT-helix connection to the β-sheet containing the RT active site, we speculate that the motion of the RT-helix drives the RT to adopt a conformation more closely resembling the active conformation of other RT structures, or vice versa, that the conformational change of the β-sheet drives the motion of the RT-helix (Fig. 7a). We suggest that the Cas6b-RTb and Cas6b'-RTb' domains may be able to adopt two states: one where the Cas6-RT lobe is freely moving, only flexibly attached by a linker to the Cas1 domain in the same chain and another where it docks onto the opposite Cas1a domain, blocking Cas1 from carrying out integration (Fig. 7b). Having on-off switches for the RT and Cas1 active sites may be critical to the process of RNA spacer acquisition where intermediate substrates may pass from one active site to the other. Additionally, the unidirectional crosstalk between the Cas6 to the other two domains also implies

a possible multi-step process in which coordination between the different enzymatic functions creates an ordered process for spacer integration and CRISPR RNA processing.

Our results imply that Cas6-RT-Cas1—Cas2 functions as a coordinated system to enable spacer acquisition from both DNA and RNA sources. Questions remain about the relative efficiency of these reactions in light of our observations that the complex strongly prefers DNA over RNA substrates for cleavage-ligation, suggesting that RNA may not be the primary source of spacers for this system or that additional host factors are required to improve the efficiency of RNA integration. Previous literature examining correlation between the frequency of spacer acquisition and overall gene expression level does indicate that there is variation among these systems *in vivo*[27-29]. It is also possible that the efficiency in integrating RNA substrates varies depending on the substrate as a part of a selection process to distinguish between self and non-self. Future research on RNA spacer acquisition by these RT-Cas1 and Cas6-RT-Cas1 fusions can more broadly shed light on how these systems may potentially generate immunity against RNA invaders, which is still a topic that is not very well understood. Furthermore, deeper insight into these fusions can help make strides toward advancing tools that utilize these proteins to record transcriptional information in cells. By presenting a high-resolution structure of the Cas6-RT-Cas1—Cas2 complex, our work provides key details toward achieving a greater understanding of RT-associated integrases and their functions.

## METHODS

### Cas6-RT-Cas1 Protein Identification.

Hidden Markov Models were built from RT-Cas1 fusion proteins or Cas1 sequences affiliated with RTs obtained from a previous study[26] and used to search for more RT-Cas1 representatives. Metagenomic contigs harboring fusion proteins consisting of RT, Cas1, and Cas6 domains were selected from a groundwater, CO2-driven Geyser dataset. The Cas6-RT-Cas1 and Cas2 proteins and their cognate CRISPR array from a Thiomicrospira genome assembly became the subject of this study following manual genome curation.[53] Sequencing reads from the original dataset were mapped to the *de novo* assembled sequences using Bowtie2 v.2.3.4.15. We retained unplaced mate pairs of mapped reads and mappings were manually checked to identify local misassemblies. Sequence ends prematurely terminated due to the presence of repetitive elements were extended using unplaced or incorrectly placed paired reads. In some cases, extended ends were used to recruit new scaffolds that were then added to the assembly. The extension accuracy and accuracy of local assembly changes were verified by subsequent read mapping. The CRISPR locus is available in Supplementary Data 1.

### Residue Conservation Analysis.

A diverse set of 968 Cas1 proteins and 92 Cas1s fused to or associated with RTs (and in some cases, Cas6) was used for analysis of the conservation of the cluster of residues consisting of R832, R834, R835, and H879. An alignment was generated from the aforementioned set with the Cas6-RT-Cas1 protein of interest using the MAFFT v7.407 LINSI setting with 1000 iterations[54]. The positions of interest were extracted and a sequence logo was generated to display the conservation of residues across the relevant Cas1 proteins.

**Protein Purification.**

The Cas6-RT-Cas1 and Cas2 genes were codon optimized for *E. coli* expression and ordered as G-blocks. Cas6-RT-Cas1 and Cas2 were PCR amplified and cloned separately into a pET-based expression vector with an N-terminal 10xHis-MBP-TEV tag. Plasmids were transformed into chemically competent Rosetta cells. Cells were grown to an $OD_{600}$ of ~0.6 and induced overnight at 16 °C with 0.5 mM isopropyl-β-D-thiogalactopyranoside (IPTG). Cells were harvested and resuspended in lysis buffer (20 mM HEPES, pH 7.5, 500 mM NaCl, 10 mM imidazole, 0.1% Triton X-100, 1 mM Tris (2-carboxyethyl)phosphine (TCEP), Complete EDTA-free protease inhibitor (Roche), 0.5 mM phenylmethylsulfonyl fluoride (PMSF), and 10% glycerol. After cell lysis by sonication, lysate was clarified by centrifugation and the supernatant was incubated on Ni-NTA resin (Qiagen). The resin was washed with wash buffer (20 mM HEPES, pH 7.5, 500 mM NaCl, 10 mM imidazole, 1 mM TCEP, and 5% glycerol), before the protein was eluted with wash buffer supplemented with 300 mM imidazole and then digested with TEV protease overnight. The salt concentration was diluted to 335 mM NaCl using ion-exchange buffer A (20 mM HEPES, pH 7.5, 1 mM TCEP, and 5% glycerol), the cleaved MBP tag was removed with an MBPTrap column (GE Healthcare), and the protein was bound to a HiTrap heparin HP column (GE Healthcare), before elution with a gradient from 335 mM to 1 M KCl. The protein was then concentrated and purified on a Superdex 200 (16/60) column with storage buffer (20 mM HEPES, pH 7.5, 500 mM KCl, 1 mM TCEP, and 5% glycerol). The same purification protocol was used for Cas6-RT-Cas1 (wild-type and mutants) and Cas2. Sequences of proteins are shown in Supplementary Table 1.

**DNA and RNA Substrate Preparation.**

To generate target plasmid pCRISPR, the CRISPR array was reduced and ordered as two fragments, which were amplified by PCR and inserted into a pUC19 backbone by Gibson Assembly. For the full-site selection, pCRISPR_full-site was generated by cloning a *cat*

promoter and chloramphenicol resistance gene into a pUC19 plasmid, removing the RBS sequence and start codon, and then inserting a sequence carrying 163 bp of the leader from pCRISPR and the 35 bp repeat before the chloramphenicol resistance gene. DNA and RNA oligos used in this study were ordered from Integrated DNA Technologies (IDT) and purified on 6% (for >35-nt oligos) or 14% urea-PAGE (for ≤35-nt oligos). Protospacers, dsDNA targets, primer/template hybrids, and the half-site substrate were formed by heating at 95 °C for 5 minutes and slow cooling to room temperature in either HEPES hybridization buffer (20 mM HEPES, pH 7.5, 25 mM KCl, and 10 mM $MgCl_2$) or Tris hybridization buffer (20 mM Tris, pH 7.5, 25 mM KCl, and 10 mM $MgCl_2$), as consistent with the integration buffer in which the substrate will be used. For the half-site substrate, hybridization was carried out with a 1.25-fold excess of the shortest strand before purification on 8% native PAGE. Sequences of cloning primers and DNA and RNA substrates are shown in Supplementary Table 2.

**Integration Assays.**

Integration assays with supercoiled target plasmid were conducted in HEPES integration buffer (20 mM HEPES, pH 7.5, 25 mM KCl, 10 mM $MgCl_2$, 1 mM DTT, 0.01% Nonidet P-40, and 10% DMSO), while integration assays with the short linear dsDNA target were conducted in Tris integration buffer (20 mM Tris, pH 7.5, 25 mM KCl, 10 mM $MgCl_2$, 1 mM DTT, 0.01% Nonidet P-40, and 10% DMSO). Cas6-RT-Cas1 and Cas2 were first pre-complexed at 2 μM at room temperature for 30 min. in the designated integration buffer. The protospacer substrate was then incubated with the Cas6-RT-Cas1 and Cas2 complex for 15 minutes, followed by addition of target plasmid (pCRISPR or pCRISPR_full-site) to 20 ng/mL (~10 nM) or short linear dsDNA target (250 nM). Integration reactions with ssRNA protospacers are supplied with 1 mM dNTPs unless otherwise indicated. The reaction was carried out at 37 °C for 2 hours. A 4 μM and 1 μM concentration was used for all protospacers for the time-course integration assays and integration assays with a short dsDNA linear target, respectively. For all other integration

assays, a 500 nM concentration is used for dsDNA and ssDNA protospacers and a 4 μM concentration is used for ssRNA protospacers. In reactions where fluorescent protospacers are indicated, oligos with a 5' 6-carboxyfluorescein (FAM) attachment were ordered. The fluorescent dsDNA protospacers were generated by hybridizing one 5'-labeled strand with its complementary unlabeled strand. For reactions with a supercoiled target plasmid, the reaction was quenched by the addition of 0.4% SDS and 25 mM EDTA, treated with proteinase K for 15 minutes at room temperature, and then treated with 3.4% SDS, before analysis on a 1.5% agarose gel (unstained for reactions with fluorescent protospacers and pre-stained with SYBR Safe for reactions with unlabeled protospacers). For reactions with a short linear dsDNA target, the reaction was quenched by the addition of 2 vol. of quench buffer (95% formamide, 30 mM EDTA, 0.2% SDS, and 400 μg/mL heparin) and heating at 95 °C for 4 min., before analysis on a 6% urea-PAGE gel. The same amount of reaction product is added to each lane for all gel analyses. When examining integration by open circular product formation, gel bands were visualized by ChemiDoc MP (BioRad) and quantified using Image Lab 6.0 (BioRad). The fraction plasmid nicked was calculated as the ratio of the open circular integration product band intensity to the total intensity of both the open circular band and the supercoiled plasmid band. Fluorescent bands were visualized by Typhoon FLA gel imaging scanner and the band intensities were quantified using ImageQuantTL 8.2. The relative percent ligation activity of the mutant proteins was calculated as the ratio of the product band intensity from the mutant protein complex relative to the product band intensity from the parallel WT control. Statistical analyses were performed using Prism 7 version 7.0c (GraphPad). Data were presented as the mean ± SD (error bars) of three biologically independent experiments. Significance was assessed using unpaired, two-tailed *T* tests ($\alpha = 0.05$).

For full-site integration assays, pCRISPR_full-site is used as the target plasmid. Sixteen 50 uL integration reactions were carried out as described above with the reporter construct for 2 hours. Instead of quenching the reaction, the integration products were purified using the

Qiagen MinElute PCR Purification Kit and electroporated into DH10B cells. Transformants were plated on LB agar containing chloramphenicol and 95 of the surviving colonies were sequenced using Sanger Sequencing and analyzed using SnapGene Version 5.0.8.

**RT and Cas6 Activity Assays.**

RT activity was measured using an ELISA-based colorimetric reverse transcriptase activity assay (Catalog No. 11468120910, Roche Diagnostics, Indianapolis, IN). The HEPES integration buffer was used for the reaction. Cas6-RT-Cas1 and Cas2 were first pre-complexed at 2 μM at room temperature for 30 min. The reactions were conducted in triplicates using the supplied template/primer hybrid poly (A) × oligo $(dT)_{15}$ at 37 °C for 2 hours. After following the assay kit's ELISA procedure, the RT activity was measured in absorption units ($A_{405\ nm}$ - $A_{490\ nm}$) using a fluorescence plate reader (BioTek). Statistical analyses were performed using Prism 7 version 7.0c (GraphPad). Data were presented as the mean ± SD (error bars) of three independent experiments. Significance was assessed by comparing samples to the respective WT controls using unpaired, two-tailed *T* tests ($\alpha = 0.05$). For template-driven cDNA synthesis reactions off a fluorescent DNA primer annealed to a DNA or RNA template, Cas6-RT-Cas1 and Cas2 were first pre-complexed at 2 μM at room temperature for 30 min before incubating with 250 nM primer/template substrate and 1 mM dNTPs, where indicated, at 37 °C for 2 hours. For Cas6 activity assays, Cas6-RT-Cas1 and Cas2 were first pre-complexed at 2 μM at room temperature for 30 min. before incubating with the 5'-labeled 35-nt RNA substrate corresponding to the repeat sequence at 37 °C for 2 hours. The template-driven cDNA synthesis reactions and Cas6 processing reactions were quenched by adding 2 vol. quench buffer and heating at 95 °C for 4 min., before analysis on a 14% urea-PAGE gel. Fluorescent bands were visualized by Typhoon FLA gel imaging scanner and intensities were quantified using ImageQuantTL 8.2. The percent Cas6 activity is quantified as the ratio of the product band intensity to the total intensity of both the product band and the unprocessed RNA substrate band.

**Electrophoretic Mobility Shift Assays.**

Electrophoretic mobility shift assays were performed in Binding buffer (10 mM Tris-HCl, pH 7.5, 5 mM EDTA, 500 mM KCl, 0.5 mM TCEP, 5% glycerol, 50 µg/mL heparin, 50 µg/mL BSA, 0.005% Tween-20). Cas6-RT-Cas1 and Cas2 were first pre-complexed at increasing concentrations (0 µM, 1 µM, 10 µM, 20 µM, 50 µM) for 1 hr before incubating with 250 nM of the half-site substrate for 1 hr at room temperature. The binding reactions were analyzed on a 1% agarose gel containing 1x Tris-acetate-EDTA (TAE) buffer supplemented with 500 mM KCl, using a 1x TAE buffer supplemented with 250 mM KCl as the running buffer at 4 ºC. Fluorescent bands were visualized by Typhoon FLA gel imaging scanner.

**Grid Preparation.**

Cas6-RT-Cas1—Cas2 and the half-site DNA substrate were complexed by mixing 50 µM Cas6-RT-Cas1, 50 µM Cas2, and 12.5 µM DNA half-site substrates in storage buffer and dialyzing in complex buffer (10 mM HEPES, pH 7.5, 5 mM EDTA, 500 mM KCl, 0.5 mM TCEP, and 0.5% glycerol) for 2 hours using the Slide-A-Lyzer MINI Dialysis Devices at room temperature. The complex was concentrated to 100 µM Cas6-RT-Cas1—Cas2 and purified over a Superose 6 Increase 10/300 GL column. For freezing grids, a 3 µl drop of protein was applied to freshly glow discharged Holey Carbon, 300 mesh R 1.2/1.3 gold grids (Quantifoil, Großlöbichau, Germany). A FEI Vitrobot Mark IV (ThermoFisher Scientific) was used with 4°C, 100% humidity, 1 blot force, and a 3 second blot time. Grids were then clipped and used for data collection.

**Cryo-EM Data Acquisition.**

Grids were clipped and transferred to a FEI Titan Krios electron microscope operated at 300 kV. 50 frame movies were recorded on a Gatan K3 Summit direct electron detector in super-resolution counting mode with pixel size of 0.5953 Å. The electron dose was 10.50 $e^{-}$ $Å^{2}$ $s^{-1}$ and

total dose was 49.98 $e^-$ Å$^2$. Initial data collection was performed with a GIF Quantum energy filter inserted, however after 516 micrographs had been collected it became unstable and was removed for the final 2729 micrographs. Nine movies were collected around a central hole position with image shift and defocus was varied from -1 to -3 µm through SerialEM[55]. See Supplementary Table 3 for data collection statistics.

**Cryo-EM Data Processing.**

Motion-correction and dose-weighting were performed on all 3330 movies 2x binned to 1.187 Å per pixel using RELION 3.0's implementation of MotionCor2[56,57]. CTFFIND-4.1 was used to estimate the contrast transfer function (CTF) parameters[58]. Micrographs were then manually examined to remove subjectively bad micrographs, such as empty or contaminated holes, resulting in 2,035 micrographs[59]. Additionally, micrographs with a CTF maximum resolution lower than 4 Å were discarded, yielding 1998 micrographs. Template-free auto-picking of particles was performed with RELION3.1's Laplacian-of-Gaussian (LoG) filter to generate an initial set of particles which were iteratively classified to find a subset of particles that were subsequently used to template-based auto-pick 249,102 particles.

Template picked particles were iteratively 2D-classified in RELION3.0 resulting in a set of 129,175 particles. These particles were imported to cryoSPARC v2. An initial map was generated with an ab-initio job and refined with subsequent homogeneous and non-uniform refinement jobs[60,61]. 'csparc2star.py' from UCSF pyem was used to convert the resulting data to a RELION star file[62]. These particles were then refined in RELION3.0 resulting in a map with ~4.1 Å overall resolution. Subsequent refinement following Bayesian particle polishing, and CTF refinement resulted in a map with ~3.66 Å overall resolution[63]. A mask surrounding the stable side of the complex was generated and supplied to a refinement resulting in a map with ~3.42 Å overall resolution. Further classification in 3D was attempted but did not result in an improved map, nor did the application of C2 symmetry during refinement.

This set of 129,175 particle coordinates were used for training in the Topaz particle-picking pipeline[64]. Training, picking, and extraction in Topaz resulted in 1,305,497 particles. These particles were then processed using a similar pipeline to the above Relion template-picked particles, resulting in 285,342 particles after 2D classification and yielding a ~3.8 Å map before, and a ~3.3 Å map following Bayesian particle polishing. The final maps were not distinguishably improved compared to the template-picked particle set. As with the prior particle set, further classification in 3D could not identify a subset of particles yielding an improved map. A consensus homogeneous refinement of these particles in cryoSPARCv2 was used as input for 3D variability analysis in cryoSPARCv2 (filter resolution 5 Å)[65].

**Modeling, Refinement, and Analysis.**

*De novo* modeling of the best-resolved portion of the complex was performed in Coot using the output map from a masked 3D refinement in Relion sharpened with Phenix.auto_sharpen[66,67]. The *de novo* modeled structure, corresponding to Cas2 chain A (amino acids 1-90), Cas2 chain B (amino acids 1-96), Cas6-RT-Cas1 chain C (amino acids 1-634), Cas6-RT-Cas1 chain D (amino acids 650-981), and Cas6-RT-Cas1 chain E (amino acids 641-981), was refined using Phenix.real_space_refine. Molprobity was used to guide iterative rounds of manual adjustment in Coot and refinement in Phenix[68]. Subsequently, portions of the refined model were copied and rigid-body docked into density corresponding to less-well resolved portions of the complex (corresponding to Cas2 chain A (amino acids 91-96), Cas6-RT-Cas1 chain C (amino acids 641-981), Cas6-RT-Cas1 chain E (amino acids 1-634), and Cas6-RT-Cas1 chain F (amino acids 650-981)) in a postprocessed Relion 3D refinement output map that was masked to include the entire visible complex. A final round of rigid-body refinement was performed in Phenix.real_space_refine. This model consists of two entire Cas2 protomers (amino acids 1-96), two entire Cas6-RT-Cas1 protomers (amino acids 1-981), and two partial Cas6-RT-Cas1 protomers with only the Cas1 domain modeled (amino acids 650-981). Six loops were

unmodeled in Cas6-RT-Cas1 protomers: three in the Cas6 domain (amino acids 95-110, 248-257, and 212-221), two in the RT domain (amino acids 383-388 and 593-616), and a linker between the RT and Cas1 domains (amino acids 635-640). Figures were prepared using Chimera, ChimeraX, Prism 7 version 7.0c (GraphPad), GNU Image Manipulation Program, and Adobe Illustrator 2020 software[69,70].

## DATA AVAILABILITY

The atomic model of the partial Cas6-RT-Cas1—Cas2 complex masked for the stable density is in the Protein Data Bank (PDB) under 7KFT [https://doi.org/10.2210/pdb7KFT/pdb] and the corresponding map is deposited in the Electron Microscopy Data Bank (EMDB) under EMD-22855 [https://www.ebi.ac.uk/pdbe/entry/emdb/EMD-22855]. The atomic model of the full Cas6-RT-Cas1—Cas2 complex is in the Protein Data Bank (PDB) under 7KFU [https://doi.org/10.2210/pdb7KFU/pdb] and the corresponding map is deposited in the Electron Microscopy Data Bank (EMDB) under EMD-22856 [https://www.ebi.ac.uk/pdbe/entry/emdb/EMD-22856]. The original micrograph movies and final particle stack are deposited in the Electron Microscopy Public Image Archive (EMPIAR) under EMPIAR-10642 [https://dx.doi.org/10.6019/EMPIAR-10642]. The uncropped images for the main text or supplementary figures and source data from Fig. 2-6 and Supplementary Fig. 1, 3, 5-8 are provided with this paper as a Source Data file. All relevant data are available from the authors.

## REFERENCES


1. Barrangou, R. et al. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).
2. Marraffini, L. A. CRISPR-Cas immunity in prokaryotes. *Nature* **526**, 55–61 (2015).

3. Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res*. **40**, 5569–5576 (2012).
4. Wright, A. V & Doudna, J. A. Protecting genome integrity during CRISPR immune adaptation. *Nat. Struct. Mol. Biol.* **23**, 876 (2016).
5. McGinn, J. & Marraffini, L. A. Molecular mechanisms of CRISPR–Cas spacer acquisition. *Nature Reviews Microbiology* **17**, 7–12 (2019).
6. Nuñez, J. K. et al. Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nat. Struct. Mol. Biol.* **21**, 528-534 (2014).
7. Nuñez, J. K., Lee, A. S. Y., Engelman, A. & Doudna, J. A. Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature* **519**, 193–198 (2015).
8. Arslan, Z., Hermanns, V., Wurm, R., Wagner, R. & Pul, Ü. Detection and characterization of spacer integration intermediates in type I-E CRISPR-Cas system. *Nucleic Acids Res.* **42**, 7884–7893 (2014).
9. Datsenko, K. A. et al. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat. Commun.* **3**, 1–7 (2012).
10. Wright, A. V. et al. Structures of the CRISPR genome integration complex. *Science* **357**, 1113–1118 (2017).
11. Xiao, Y., Ng, S., Nam, K. H. & Ke, A. How type II CRISPR-Cas establish immunity through Cas1-Cas2-mediated spacer integration. *Nature* **550**, 137–141 (2017).
12. Nuñez, J. K., Harrington, L. B., Kranzusch, P. J., Engelman, A. N. & Doudna, J. A. Foreign DNA capture during CRISPR-Cas adaptive immunity. *Nature* **527**, 535–538 (2015).
13. Wang, J. et al. Structural and mechanistic basis of PAM-dependent spacer acquisition in CRISPR-Cas systems. *Cell* **163**, 840–853 (2015).

14. Pourcel, C., Salvignol, G. & Vergnaud, G. CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151**, 653–663 (2005).
15. Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551–2561 (2005).
16. Hale, C. R. et al. RNA-Guided RNA Cleavage by a CRISPR RNA-Cas protein complex. *Cell* **139**, 945–956 (2009).
17. Garneau, J. E. et al. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67–71 (2010).
18. Brouns, S. J. J. et al. Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960–964 (2008).
19. Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733–740 (2009).
20. Makarova, K. S. et al. Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nature Reviews Microbiology* **18**, 67–83 (2020).
21. Kojima, K. K. & Kanehisa, M. Systematic survey for novel types of prokaryotic retroelements based on gene neighborhood and protein architecture. *Mol. Biol. Evol.* **25**, 1395–1404 (2008).
22. Simon, D. M. & Zimmerly, S. A diversity of uncharacterized reverse transcriptases in bacteria. *Nucleic Acids Res.* **36**, 7219–7229 (2008).
23. Toro, N. & Nisa-Martínez, R. Comprehensive phylogenetic analysis of bacterial reverse transcriptases. *PLoS One* **9**, e114083 (2014).
24. Silas, S. et al. On the origin of reverse transcriptase-using CRISPR-Cas systems and their hyperdiverse, enigmatic spacer repertoires. *MBio* **8**, e00897-17 (2017).

25. Toro, N., Mestre, M. R., Martínez-Abarca, F. & González-Delgado, A. Recruitment of reverse transcriptase-Cas1 fusion proteins by type VI-A CRISPR-Cas systems. *Front. Microbiol.* **10**, 2160 (2019).
26. Toro, N., Martínez-Abarca, F. & González-Delgado, A. The reverse transcriptases associated with CRISPR-Cas systems. *Sci. Rep.* **7**, 1-7 (2017).
27. Silas, S. et al. Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. *Science* **351**, aad4234 (2016).
28. González-Delgado, A., Mestre, M. R., Martínez-Abarca, F. & Toro, N. Spacer acquisition from RNA mediated by a natural reverse transcriptase-Cas1 fusion protein associated with a type III-D CRISPR-Cas system in *Vibrio vulnificus*. *Nucleic Acids Res.* **47**, 10202–10211 (2019).
29. Schmidt, F., Cherepkova, M. Y. & Platt, R. J. Transcriptional recording by CRISPR spacer acquisition from RNA. *Nature* **562**, 380–385 (2018).
30. Hale, C. R. et al. Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. *Mol. Cell* **45**, 292–302 (2012).
31. Tamulaitis, G. et al. Programmable RNA shredding by the type III-A CRISPR-Cas system of *Streptococcus thermophilus*. *Mol. Cell* **56**, 506–517 (2014).
32. Goldberg, G. W., Jiang, W., Bikard, D. & Marraffini, L. A. Conditional tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting. *Nature* **514**, 633–637 (2014).
33. Peng, W., Feng, M., Feng, X., Liang, Y. X. & She, Q. An archaeal CRISPR type III-B system exhibiting distinctive RNA targeting features and mediating dual RNA and DNA interference. *Nucleic Acids Res.* **43**, 406–417 (2015).
34. Mohr, G. et al. A reverse transcriptase-Cas1 fusion protein contains a Cas6 domain required for both CRISPR RNA biogenesis and RNA spacer acquisition. *Mol. Cell* **72**, 700–714 (2018).

35. Hochstrasser, M. L. & Doudna, J. A. Cutting it close: CRISPR-associated endoribonuclease structure and function. *Trends in Biochemical Sciences* **40**, 58–66 (2015).

36. Carte, J., Wang, R., Li, H., Terns, R. M. & Terns, M. P. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes Dev.* **22**, 3489–3496 (2008).

37. Haurwitz, R. E., Jinek, M., Wiedenheft, B., Zhou, K. & Doudna, J. A. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science* **329**, 1355–1358 (2010).

38. Samai, P., Smith, P. & Shuman, S. Structure of a CRISPR-associated protein Cas2 from *Desulfovibrio vulgaris*. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **66**, 1552–1556 (2010).

39. Beloglazova, N. et al. A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. *J. Biol. Chem.* **283**, 20361–20371 (2008).

40. Nam, K. H. et al. Double-stranded endonuclease activity in *Bacillus halodurans* clustered regularly interspaced short palindromic repeats (CRISPR)-associated Cas2 protein. *J. Biol. Chem.* **287**, 35943–35952 (2012).

41. Stamos, J. L., Lentzsch, A. M. & Lambowitz, A. M. Structure of a thermostable group II intron reverse transcriptase with template-primer and its functional and evolutionary implications. *Mol. Cell* **68**, 926-939 (2017).

42. Das, K., Martinez, S. E., Bandwar, R. P. & Arnold, E. Structures of HIV-1 RT-RNA/DNA ternary complexes with dATP and nevirapine reveal conformational flexibility of RNA/DNA: insights into requirements for RNase H cleavage. *Nucleic Acids Res.* **42**, 8125–8137 (2014).

43. Mitchell, M., Gillis, A., Futahashi, M., Fujiwara, H. & Skordalakes, E. Structural basis for telomerase catalytic subunit TERT binding to RNA template and telomeric DNA. *Nat. Struct. Mol. Biol.* **17**, 513–518 (2010).

44. Ebihara, A. et al. Crystal structure of hypothetical protein TTHB192 from *Thermus thermophilus* HB8 reveals a new protein family with an RNA recognition motif-like domain. *Protein Sci.* **15**, 1494–1499 (2006).

45. Niewoehner, O., Jinek, M. & Doudna, J. A. Evolution of CRISPR RNA recognition and processing by Cas6 endonucleases. *Nucleic Acids Res.* **42**, 1341–1353 (2014).

46. Özcan, A. et al. Type IV CRISPR RNA processing and effector complex formation in *Aromatoleum aromaticum*. *Nat. Microbiol.* **4**, 89–96 (2019).

47. Shao, Y. et al. A non-stem-loop CRISPR RNA is processed by dual binding Cas6. *Structure* **24**, 547–554 (2016).

48. Zimmerly, S., Guo, H., Perlman, P. S. & Lambowitz, A. M. Group II intron mobility occurs by target DNA-primed reverse transcription. *Cell* **82**, 545–554 (1995).

49. Lambowitz, A. M. & Zimmerly, S. Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harb. Perspect. Biol.* **3**, 1–19 (2011).

50. Li, J. et al. Structural and mechanistic basis of PAM-dependent spacer acquisition in CRISPR-Cas systems. *Cell* **163**, 840–853 (2015).

51. Díez-Villaseñor, C., Guzmán, N. M., Almendros, C., García-Martínez, J. & Mojica, F. J. M. CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of *Escherichia coli*. *RNA Biol.* **10**, 792–802 (2013).

52. Wright, A. V. et al. A functional mini-integrase in a two-protein type V-C CRISPR system. *Mol. Cell* **73**, 727-737 (2019).

53. Al-Shayeb, B. et al. Clades of huge phages from across Earth's ecosystems. *Nature* **578**, 425–431 (2020).

54. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

55. Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152,** 36–51 (2005).

56. Zivanov, J. *et al.* New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife* **7,** 163 (2018).

57. Zheng, S. Q. *et al.* MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14,** 331–332 (2017).

58. Rohou, A. & Grigorieff, N. CTFFIND4: fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192,** 216–221 (2015).

59. Tang, G. et al. EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157,** 38–46 (2007).

60. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. CryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14,** 290–296 (2017).

61. Punjani, A., Zhang, H. & Fleet, D. J. Non-uniform refinement: adaptive regularization improves single particle cryo-EM reconstruction. *bioRxiv* **179,** 2019.12.15.877092 (2019).

62. Asarnow, D., Palovcak, E. & Cheng, Y. UCSF pyem v0.5. Zenodo https://doi.org/10.5281/zenodo.3576630 (2019).

63. Zivanov, J., Nakane, T. & Scheres, S. H. W. A Bayesian approach to beam-induced motion correction in cryo-EM single-particle analysis. *IUCrJ* **6,** 5–17 (2019).

64. Bepler, T. *et al.* Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. *Nat. Methods* **16,** 1153–1160 (2019).

65. Punjani, A. & Fleet, D. J. 3D variability analysis: directly resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM images. *bioRxiv* 2020.04.08.032466 (2020).

66. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 486–501 (2010).
67. Liebschner, D. *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **75,** 861–877 (2019).
68. Williams, C. J. *et al.* MolProbity: more and better reference data for improved all-atom structure validation. *Protein Sci* **27,** 293–315 (2018).
69. Pettersen, E. F., Goddard, T. D. & Huang, C. C. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).
70. Goddard, T. D. *et al.* UCSF ChimeraX: meeting modern challenges in visualization and analysis. *Protein Sci.* **27,** 14–25 (2018).

**ACKNOWLEDGMENTS**

We thank Paul Tobias for computational resources at the Cal-Cryo EM facility, and Dr. James Hurley and Dr. Eva Nogales for supporting the microscopy work. This material is based upon work supported by the National Science Foundation under award number 1817593. J.Y.W. is supported by the US National Science Foundation Graduate Fellowship and previously by the Berkeley Graduate Fellowship. B.A.-S. is supported by the US National Science Foundation Graduate Fellowship. S.G.B. is a New York Stem Cell Foundation Robertson Neuroscience Investigator. J.A.D. is an investigator of the Howard Hughes Medical Institute. We thank A.V. Wright for input on the manuscript, C. Alza and M. Jain for technical assistance, and G.J. Knott, J.-J. Liu, A. Lapinaite, J. Cofsky, K.M. Soczek, P. Pausch and members of the Doudna laboratory and Brohawn laboratory for comments and discussions.

## AUTHOR CONTRIBUTIONS

J.Y.W. conceived of and designed experiments with input from C.M.H., J.A.D., and S.G.B. J.Y.W. performed protein expression, purification, and biochemical experiments. C.M.H. and J.Y.W. performed cryo-EM data collection, and C.M.H. and S.G.B. processed the cryo-EM data and built and refined the atomic models. B.A.-S. identified the CRISPR-Cas locus of interest and performed bioinformatics experiments and analyses with input from J.F.B. J.Y.W., C.M.H., and J.A.D. wrote the manuscript and all authors edited the manuscript.

## COMPETING INTERESTS

The Regents of the University of California have patents issued and pending for CRISPR technologies on which J.A.D. is an inventor. J.A.D. is a cofounder of Caribou Biosciences, Editas Medicine, Scribe Therapeutics, Intellia Therapeutics and Mammoth Biosciences. J.A.D. is a scientific advisory board member of Caribou Biosciences, Intellia Therapeutics, eFFECTOR Therapeutics, Scribe Therapeutics, Mammoth Biosciences, Synthego, Algen Biotechnologies, Felix Biosciences and Inari. J.A.D. is a Director at Johnson & Johnson and has research projects sponsored by Biogen, Pfizer, AppleTree Partners and Roche. The remaining authors declare no competing interests.

## FIGURE LEGENDS

Figure 1. **A natural Cas6-RT-Cas1 fusion protein within a type III CRISPR-Cas system. a** Illustration of the CRISPR genomic locus encoding the proteins used in this study and schematic of the CRISPR spacer integration reaction. Leader, red; repeat, blue; spacer, yellow; dsDNA protospacer (p-spacer), brown. **b** Cartoon showing Cas6-RT-Cas1 and Cas2 proteins and the DNA substrate, mimicking the half-site intermediate depicted in **a**, used in complex assembly. Proteins colored by domain: Cas6, blue; RT, red; Cas1a, dark green; Cas1b, turquoise; Cas2, lime green. Visible RT-Cas1 linkers are shown as solid turquoise lines. Missing Cas6-RT domains (Cas6a/RTa and Cas6a'/RTa') that are present in the complex but not visible are depicted as semi-transparent, connected to the rest of the structure by dashed dark green lines. Quotation marks indicate presumed mobility. Lengths of the DNA substrate strands are indicated. **c** Overall structure of the complex depicted in surface representation and 90° rotation, with color-coding as shown in **b**. **d** Overall structure of the complex, depicted in ribbon representation. **e** Architecture of a single Cas6-RT-Cas1 protomer and 90° rotation.

Figure 2. **Cas6-RT-Cas1—Cas2 catalyzes cleavage-ligation reactions with dsDNA, ssDNA, and ssRNA substrates *in vitro*. a** Schematic of *in vitro* integration reaction of DNA or RNA protospacer (brown) into a supercoiled pCRISPR plasmid (black) that functions as the target. **b** Integration assay with fluorescent 35-nt dsDNA (0.5 μM), ssDNA (0.5 μM), and ssRNA (4 μM) protospacers. Open circular (OC) and linear ligation products are indicated. Star indicates 6-carboxyfluorescein label. Results are representative of 3 independent experiments. **c** Time-course integration assay comparing extent of plasmid-nicking with dsDNA, ssDNA, and ssRNA protospacers (35 nt; 4 μM). Fraction plasmid nicked is calculated as the fraction of open circular products relative to all plasmid ($n$ = 3 biologically independent experiments). Experimental fits are shown as solid lines. Statistical significance of fraction plasmid nicked after two hours was assessed using unpaired, two-tailed $T$ tests ($\alpha$ = 0.05) ($P$ = 0.000092, dsDNA vs. ssRNA; $P$ = 0.0173, ssDNA vs. ssRNA). Representative gels are shown in Supplementary Fig. 5a. **d** Comparison of Cas1a (dark green) and Cas1b (white), with RT-helix (red) and unknown density (blue). Active site and surrounding residues are shown in stick configuration. Active site residues are labeled with an asterisk *. **e** Cas1a and Cas1b overlay and closeup of active site. Double-headed arrows indicate large conformational differences. Single-headed arrows indicate 90° rotations. Cryo-EM density shown in blue mesh, distances in angstroms. **f** Ligation of 35-nt fluorescent dsDNA (0.5 μM), ssDNA (0.5 μM), and ssRNA (4 μM) protospacers into target pCRISPR by mutant Cas6-RT-Cas1—Cas2 proteins. Percent ligation activity is calculated as the fraction of fluorescent products from the mutant Cas6-RT-Cas1—Cas2 relative to that from the wild-type (error bars, mean ± sd, $n$ = 3 biologically independent experiments). Representative gels are shown in Supplementary Fig. 8c. Uncropped gels and source data for panels **c** and **f** are provided as a Source Data file.

Figure 3. **Cas6-RT-Cas1—Cas2 shows length selectivity for dsDNA and ssDNA substrates for integration. a** Comparison of Cas1—Cas2 disposition within the Cas6-RT-Cas1—Cas2 complex (color-coding from Fig. 1) to substrate-bound *E. coli* Cas1—Cas2 integrase (PDB:5DS5 [https://doi.org/10.2210/pdb5ds5/pdb]) and apo *E. coli* Cas1—Cas2 integrase (PDB:4P6I [https://doi.org/10.2210/pdb4P6I/pdb]) (Cas1a, dark gray; Cas1b, medium gray; Cas2, light gray), shown in surface representation. Two views are shown related by a 90° rotation. Arginine clamp and arginine channel residues are colored in blue. Protospacer substrate is depicted in red outline. **b** Overlay of Cas1—Cas2 domains of Cas6-RT-Cas1—Cas2 complex (color-coding from Fig. 1) with apo *E. coli* Cas1—Cas2 integrase (PDB:4P6I [https://doi.org/10.2210/pdb4P6I/pdb]) (gray), aligned via Cas2 dimer, shown in ribbon representation. Two views are shown related by a 90° rotation. Arrows indicate conformational difference between Cas1—Cas2 domain of Cas6-RT-Cas1—Cas2 complex and that of *E. coli* Cas1—Cas2. **c** Cas6-RT-Cas1—Cas2 integration assays with variable length dsDNA protospacers (15 to 115 bp). Supercoiled pCRISPR plasmid that functions as target for integration (SC) and open circular products (OC) are indicated. Fraction plasmid nicked is calculated as the fraction of open circular products relative to all plasmid (error bars, mean ± sd, $n$ = 3 biologically independent experiments). **d** Cas6-RT-Cas1—Cas2 integration with variable length ssDNA protospacers (15 to 115 nt), quantifying fraction plasmid nicked (error bars, mean ± sd, $n$ = 3 biologically independent experiments). Uncropped gels and source data for panels **c** and **d** are provided as a Source Data file.

Figure 4. **Cas6-RT-Cas1—Cas2 conducts site-specific full-site integration. a** Schematic of chloramphenicol selection screen for full-site integration events near the leader-repeat junction. The selection plasmid contains a CRISPR leader (pink) and repeat (blue) upstream of a chloramphenicol resistance gene (*CmR*; green) with the RBS and start codon removed. Full-site

integration of a protospacer (brown) containing an RBS and start codon allows for translation of *CmR*. Transformants are plated on chloramphenicol plates and clones are sequenced using Sanger sequencing. **b** Representation depicting spacer insertion events near the leader-repeat junction in a selection plasmid with a 163 bp leader. The arrowheads indicate the insertion sites with number label indicating bp away from spacer-end of repeat (color-coding: +0, yellow; -25, brown; -45, magenta, -104, light blue; -106, purple; -110, pink, -138 gray; -144, dark blue; +20, light green; +36, dark green) and the height is scaled to the number of spacer insertion events. The arrows on top indicate full-site integration events, with the corresponding solid colored line adjacent to the arrowhead representing the sequence that is duplicated after spacer insertion. The arrows on the bottom indicate partial/whole spacer insertion events followed by a deletion, with the corresponding dashed colored line adjacent to the arrowhead representing the sequence that is deleted with the spacer insertion. **c** Number of spacer insertion events near the leader-repeat junction. Insertion sites follow color-coding in **b**. The black bars represent full-site integration events with a 35 bp repeat duplication of the adjacent sequence and the gray bars represent partial/whole spacer insertion followed by a deletion of variable length. Source data for panel **c** are provided as a Source Data file.

Figure 5. **Cas6-RT-Cas1 catalyzes cDNA synthesis with both RNA and DNA templates. a** Architecture of RT domain from Cas6-RT-Cas1—Cas2 (red) and 90° rotation, with FADD motif (blue) indicated. **b** Alignment and overlay of RT domain from Cas6-RT-Cas1—Cas2 (red) with group II intron RT (white) bound to primer/template substrate (PDB: 6AR1 [https://doi.org/10.2210/pdb6AR1/pdb]) and 90° rotation, FADD motif from Cas6-RT-Cas1—Cas2 colored blue and YADD motif from group II intron RT colored light blue. Palm, fingers, and thumb/X regions and primer/template are indicated. **c** Closeup of RT active site residues of Cas6-RT-Cas1—Cas2 (blue) and group II intron RT (light blue), shown in stick configuration. Two views are shown related by a 90° rotation. **d** RT activity assays comparing Cas6-RT-Cas1 alone and Cas6-RT-Cas1—Cas2. RT activity is measured using an ELISA-based colorimetric reverse transcriptase activity assay (error bars, mean ± sd, $n$ = 3 biologically independent experiments) (Catalog No. 11468120910, Roche Diagnostics, Indianapolis, IN). Source data are provided as a Source Data file. **e** Template-driven cDNA synthesis reactions off a fluorescent DNA primer annealed to DNA and RNA templates. Substrates are schematized (DNA, blue; RNA, purple). Expected cDNA synthesis reactions are indicated. Star indicates 6-carboxyfluorescein label. Results are representative of 3 independent experiments. **f** Template-driven cDNA synthesis reactions in the absence of different dNTPs and in the presence of added ddNTPs. Results are representative of 3 independent experiments. Uncropped gels are available in a Source Data file.

Figure 6. **Crosstalk between RT and Cas1 and Cas6 active sites. a** Interaction between RT-helix α9 and Cas1 active site (color-coding from Fig. 1). **b** Closeup of RT-helix in Cas1a and 90° rotation. Cas1 active site residues and RT-helix K574 are shown in stick configuration. **c** Ligation of fluorescent 35-nt dsDNA (0.5 μM), ssDNA (0.5 μM), and ssRNA (4 μM) protospacers into target pCRISPR by mutant Cas6-RT-Cas1—Cas2s. The percent ligation activity is calculated as the fraction of fluorescent products from the mutant complex relative to that from the WT complex (error bars, mean ± sd, $n$ = 3 biologically independent experiments). Results are represented by colored bars: WT, black; Cas1 mutants, greens; RT mutants, reds; Cas6 mutant, blue. Cas1 mutant data are the same data shown in Fig. 2e and are depicted in pale greens with gray labels. Representative gels are shown in Supplementary Fig. 8c. **d** RT activity assays comparing the WT and mutant Cas6-RT-Cas1—Cas2s (error bars, mean ± sd, $n$ = 3 biologically independent experiments; Catalog No. 11468120910, Roche Diagnostics, Indianapolis, IN). Statistical significance was assessed by comparing samples to WT control using unpaired, two-tailed $T$ tests ($\alpha$ = 0.05) ($P$ = 0.0048, Cas1-H873A; $P$ = 0.0019, Cas1-

R835A; *P* = 0.0001, RT-D540A; *P* = 0.0001, RT-K574A; *P* = 0.0034, Cas6-R37A). Same color-coding is used from **c** except Cas1 mutant data are shown in darker greens with black labels. **e** Cas6 activity assays comparing the WT and mutant Cas6-RT-Cas1—Cas2 proteins. The percent cleavage is calculated as the fraction of the fluorescent CRISPR repeat RNA that has been cleaved, with color-coding from d (error bars, mean ± sd, *n* = 3 biologically independent experiments). A representative gel is shown in Supplementary Fig. 8e. Source data for panels **c**-**e** are provided as a Source Data file.

Figure 7. **Model for RT activation. a** Cartoon showing hypothesized motion of RT-helix α9 in and out of the Cas1a active site and the hypothesized motion of the β-strands attached to the FADD motif containing the active site residues. Proteins colored by domain: Cas6, blue; RT, red, Cas1a, dark green; Cas1b, turquoise; Cas2, lime green. Visible RT-Cas1 linkers are shown as solid turquoise lines. RT-helix α9 and connecting β-strands containing FADD motif are schematized and shown in pink. Hypothesized motions are indicated by curved arrows. **b** Cartoon showing hypothesized conformational change of the Cas6-RT-Cas1—Cas2 complex to accommodate a protospacer substrate, with color-coding from **a**. Missing Cas6-RT domains (Cas6a/RTa and Cas6a'/RTa') that are present in the complex but not visible are depicted as semi-transparent, connected to the rest of the structure by dashed dark green lines. Quotation marks indicate presumed mobility. Protospacer substrate is schematized in brown. Hypothesized motions are indicated by curved arrows.

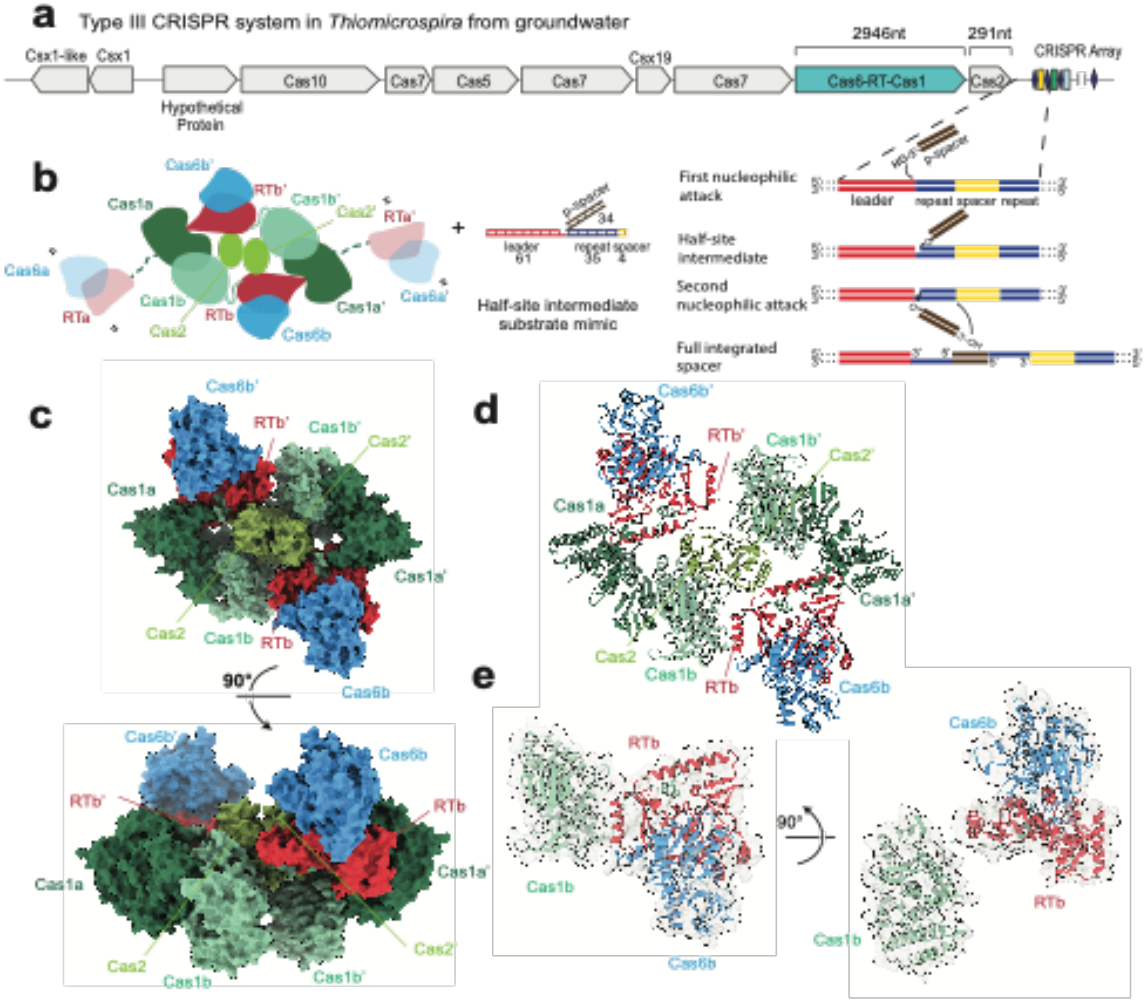


Figure 1. **A natural Cas6-RT-Cas1 fusion protein within a type III CRISPR-Cas system. a** Illustration of the CRISPR genomic locus encoding the proteins used in this study and schematic of the CRISPR spacer integration reaction. Leader, red; repeat, blue; spacer, yellow; dsDNA protospacer (p-spacer), brown. **b** Cartoon showing Cas6-RT-Cas1 and Cas2 proteins and the DNA substrate, mimicking the half-site intermediate depicted in **a**, used in complex assembly. Proteins colored by domain: Cas6, blue; RT, red; Cas1a, dark green; Cas1b, turquoise; Cas2, lime green. Visible RT-Cas1 linkers are shown as solid turquoise lines. Missing Cas6-RT domains (Cas6a/RTa and Cas6a'/RTa') that are present in the complex but not visible are depicted as semi-transparent, connected to the rest of the structure by dashed dark green lines. Quotation marks indicate presumed mobility. Lengths of the DNA substrate strands are indicated. **c** Overall structure of the complex depicted in surface representation and 90° rotation, with color-coding as shown in **b**. **d** Overall structure of the complex, depicted in ribbon representation. **e** Architecture of a single Cas6-RT-Cas1 protomer and 90° rotation.

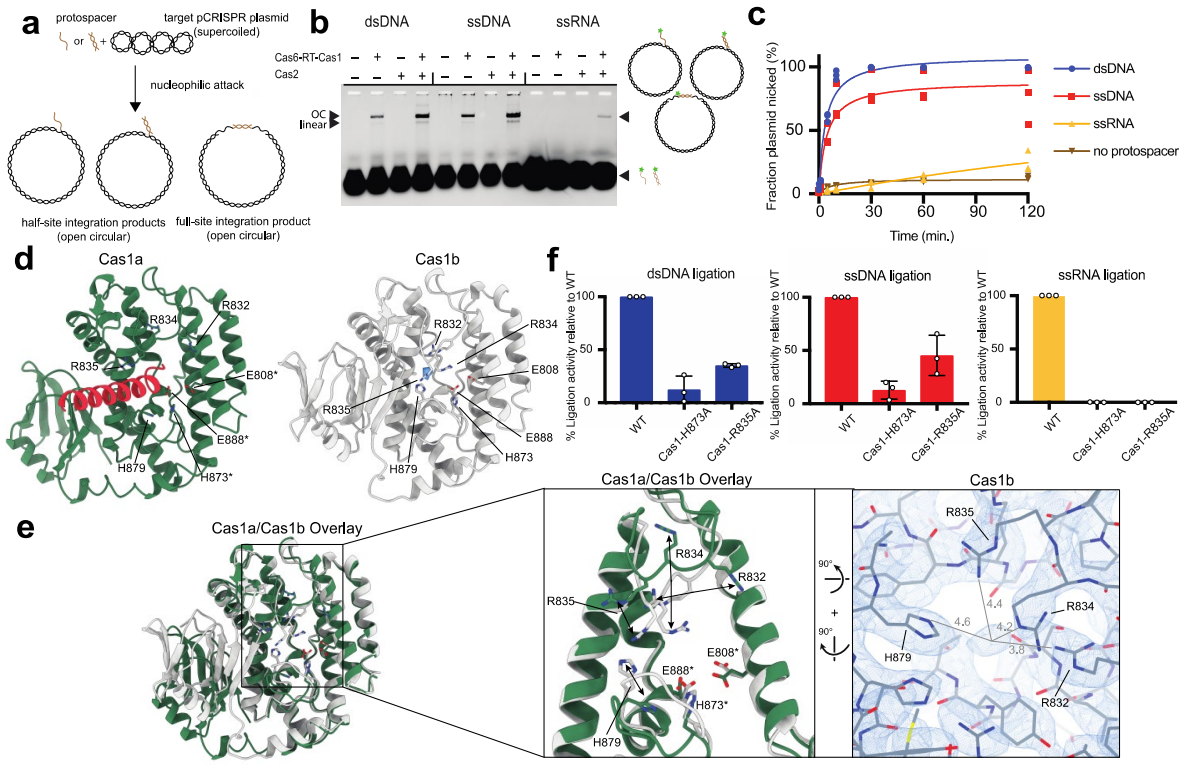


Figure 2. **Cas6-RT-Cas1—Cas2 catalyzes cleavage-ligation reactions with dsDNA, ssDNA, and ssRNA substrates *in vitro*. a** Schematic of *in vitro* integration reaction of DNA or RNA protospacer (brown) into a supercoiled pCRISPR plasmid (black) that functions as the target. **b** Integration assay with fluorescent 35-nt dsDNA (0.5 μM), ssDNA (0.5 μM), and ssRNA (4 μM) protospacers. Open circular (OC) and linear ligation products are indicated. Star indicates 6-carboxyfluorescein label. Results are representative of 3 independent experiments. **c** Time-course integration assay comparing extent of plasmid-nicking with dsDNA, ssDNA, and ssRNA protospacers (35 nt; 4 μM). Fraction plasmid nicked is calculated as the fraction of open circular products relative to all plasmid ($n$ = 3 biologically independent experiments). Experimental fits are shown as solid lines. Statistical significance of fraction plasmid nicked after two hours was assessed using unpaired, two-tailed $T$ tests ($\alpha$ = 0.05) ($P$ = 0.000092, dsDNA vs. ssRNA; $P$ = 0.0173, ssDNA vs. ssRNA). Representative gels are shown in Supplementary Fig. 5a. **d** Comparison of Cas1a (dark green) and Cas1b (white), with RT-helix (red) and unknown density (blue). Active site and surrounding residues are shown in stick configuration. Active site residues are labeled with an asterisk *. **e** Cas1a and Cas1b overlay and closeup of active site. Double-headed arrows indicate large conformational differences. Single-headed arrows indicate 90° rotations. Cryo-EM density shown in blue mesh, distances in angstroms. **f** Ligation of 35-nt fluorescent dsDNA (0.5 μM), ssDNA (0.5 μM), and ssRNA (4 μM) protospacers into target pCRISPR by mutant Cas6-RT-Cas1—Cas2 proteins. Percent ligation activity is calculated as the fraction of fluorescent products from the mutant Cas6-RT-Cas1—Cas2 relative to that from the wild-type (error bars, mean ± sd, $n$ = 3 biologically independent experiments). Representative gels are shown in Supplementary Fig. 8c. Uncropped gels and source data for panels **c** and **f** are provided as a Source Data file.

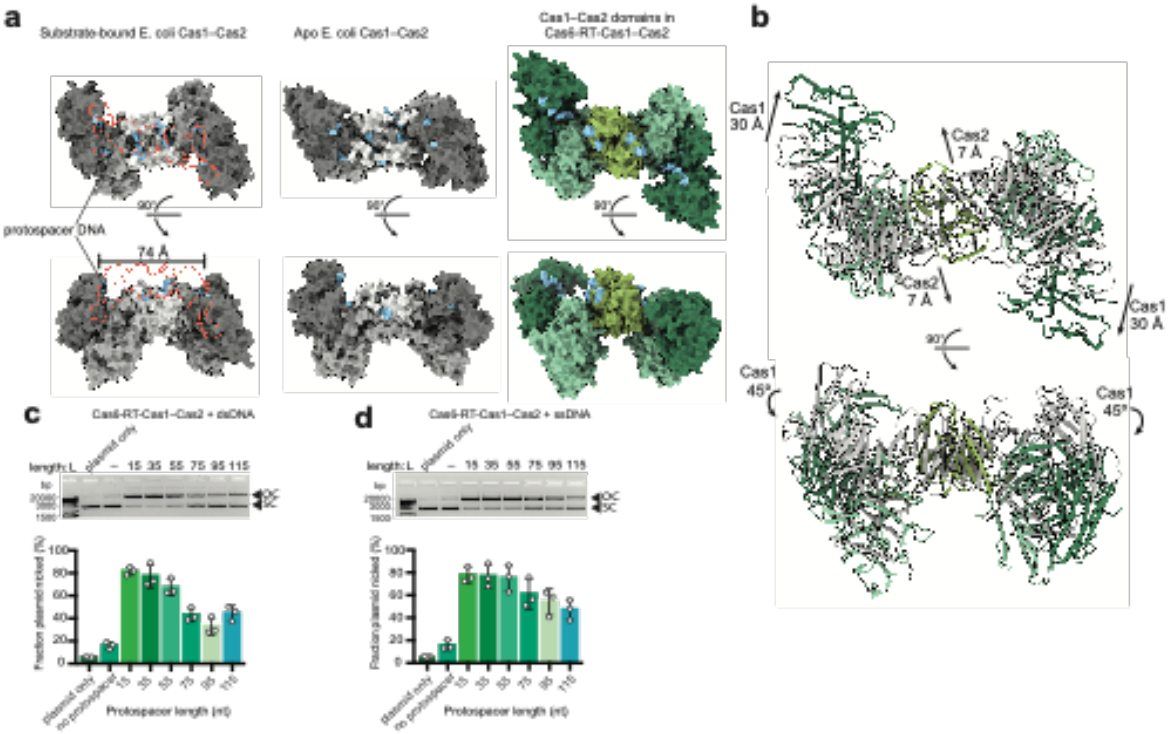


Figure 3. **Cas6-RT-Cas1—Cas2 shows length selectivity for dsDNA and ssDNA substrates for integration. a** Comparison of Cas1—Cas2 disposition within the Cas6-RT-Cas1—Cas2 complex (color-coding from Fig. 1) to substrate-bound *E. coli* Cas1—Cas2 integrase (PDB:5DS5 [https://doi.org/10.2210/pdb5ds5/pdb]) and apo *E. coli* Cas1—Cas2 integrase (PDB:4P6I [https://doi.org/10.2210/pdb4P6I/pdb]) (Cas1a, dark gray; Cas1b, medium gray; Cas2, light gray), shown in surface representation. Two views are shown related by a 90° rotation. Arginine clamp and arginine channel residues are colored in blue. Protospacer substrate is depicted in red outline. **b** Overlay of Cas1—Cas2 domains of Cas6-RT-Cas1—Cas2 complex (color-coding from Fig. 1) with apo *E. coli* Cas1—Cas2 integrase (PDB:4P6I [https://-doi.org/10.2210/pdb4P6I/pdb]) (gray), aligned via Cas2 dimer, shown in ribbon representation. Two views are shown related by a 90° rotation. Arrows indicate conformational difference between Cas1—Cas2 domain of Cas6-RT-Cas1—Cas2 complex and that of *E. coli* Cas1—Cas2. **c** Cas6-RT-Cas1—Cas2 integration assays with variable length dsDNA protospacers (15 to 115 bp). Supercoiled pCRISPR plasmid that functions as target for integration (SC) and open circular products (OC) are indicated. Fraction plasmid nicked is calculated as the fraction of open circular products relative to all plasmid (error bars, mean ± sd, $n = 3$ biologically independent experiments). **d** Cas6-RT-Cas1—Cas2 integration with variable length ssDNA protospacers (15 to 115 nt), quantifying fraction plasmid nicked (error bars, mean ± sd, $n = 3$ biologically independent experiments). Uncropped gels and source data for panels **c** and **d** are provided as a Source Data file.

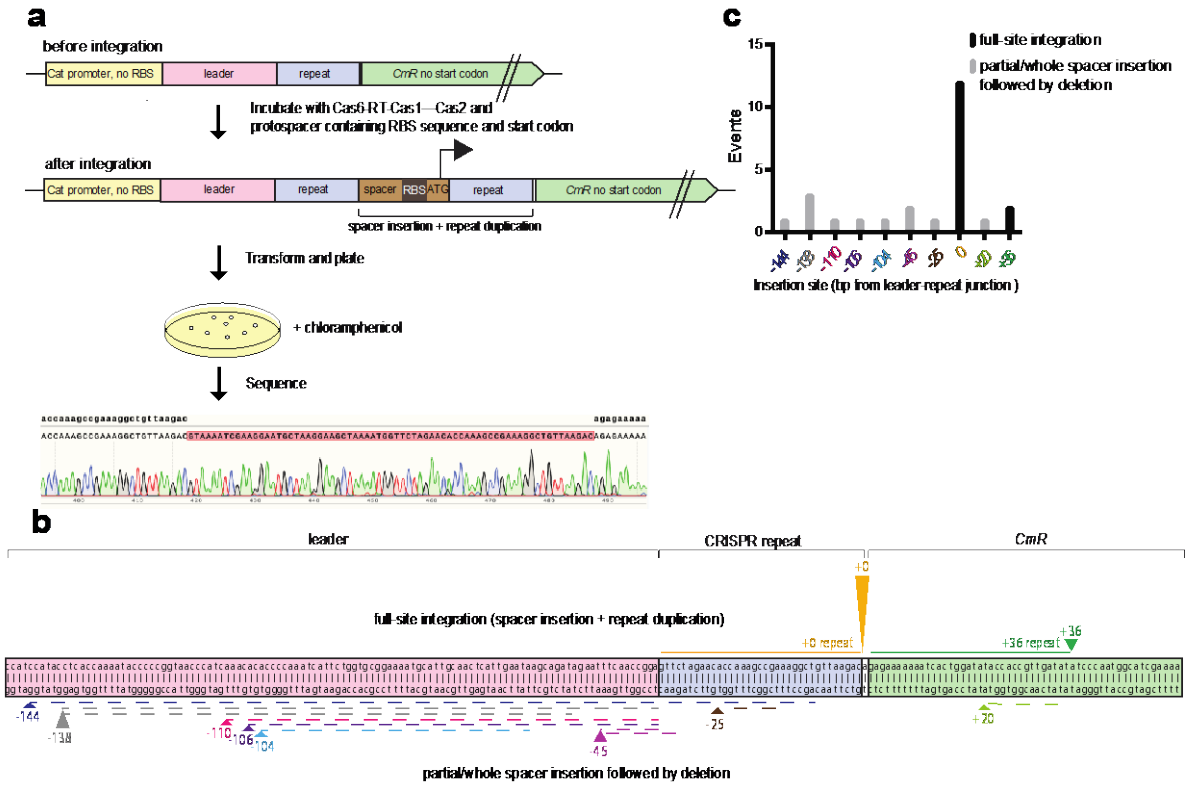


Figure 4. **Cas6-RT-Cas1—Cas2 conducts site-specific full-site integration. a** Schematic of chloramphenicol selection screen for full-site integration events near the leader-repeat junction. The selection plasmid contains a CRISPR leader (pink) and repeat (blue) upstream of a chloramphenicol resistance gene (*CmR*; green) with the RBS and start codon removed. Full-site integration of a protospacer (brown) containing an RBS and start codon allows for translation of the *CmR*. Transformants are plated on chloramphenicol plates and clones are sequenced using Sanger sequencing. **b** Representation depicting spacer insertion events near the leader-repeat junction in a selection plasmid with a 163 bp leader. The arrowheads indicate the insertion sites with number label indicating bp away from spacer-end of repeat (color-coding: +0, yellow; -25, brown; -45, magenta, -104, light blue; -106, purple; -110, pink, -138 gray; -144, dark blue; +20, light green; +36, dark green) and the height is scaled to the number of spacer insertion events. The arrows on top indicate full-site integration events, with the corresponding solid colored line adjacent to the arrowhead representing the sequence that is duplicated after spacer insertion. The arrows on the bottom indicate partial/whole spacer insertion events followed by a deletion, with the corresponding dashed colored line adjacent to the arrowhead representing the sequence that is deleted with the spacer insertion. **c** Number of spacer insertion events near the leader-repeat junction. Insertion sites follow color-coding in **b**. The black bars represent full-site integration events with a 35 bp repeat duplication of the adjacent sequence and the gray bars represent partial/whole spacer insertion followed by a deletion of variable length . Source data for panel **c** are provided as a Source Data file.

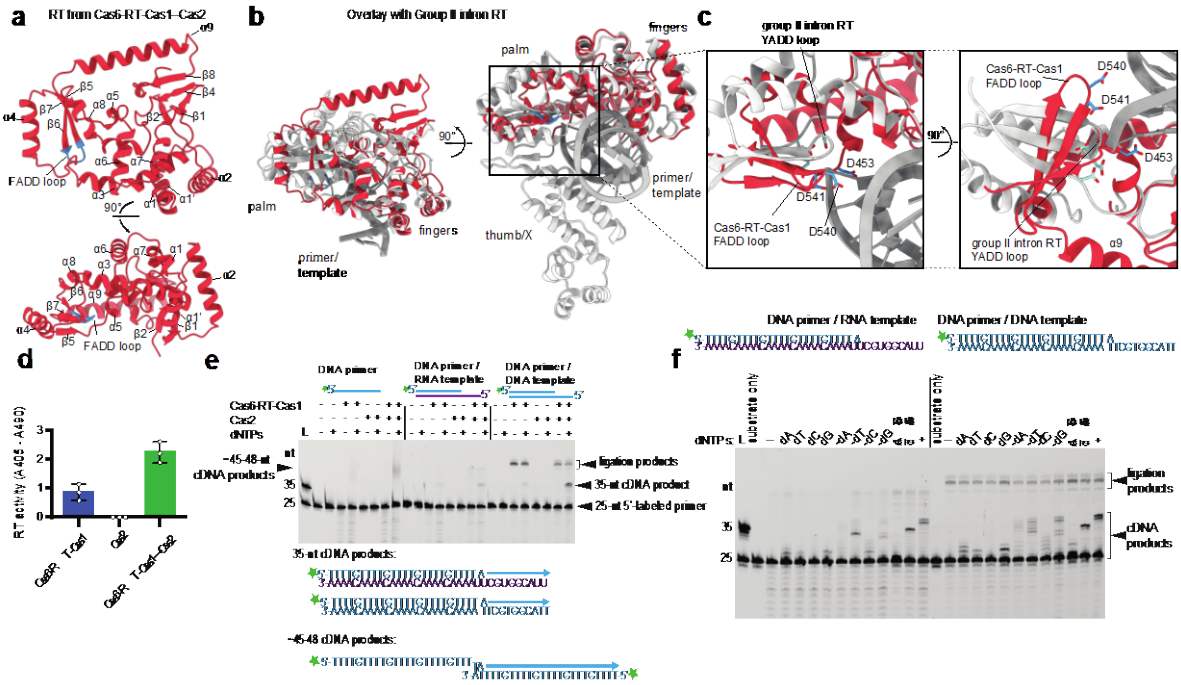


Figure 5. **Cas6-RT-Cas1 catalyzes cDNA synthesis with both RNA and DNA templates. a** Architecture of RT domain from Cas6-RT-Cas1—Cas2 (red) and 90° rotation, with FADD motif (blue) indicated. **b** Alignment and overlay of RT domain from Cas6-RT-Cas1—Cas2 (red) with group II intron RT (white) bound to primer/template substrate (PDB: 6AR1 [https://doi.org/10.2210/pdb6AR1/pdb]) and 90° rotation, FADD motif from Cas6-RT-Cas1—Cas2 colored blue and YADD motif from group II intron RT colored light blue. Palm, fingers, and thumb/X regions and primer/template are indicated. **c** Closeup of RT active site residues of Cas6-RT-Cas1—Cas2 (blue) and group II intron RT (light blue), shown in stick configuration. Two views are shown related by a 90° rotation. **d** RT activity assays comparing Cas6-RT-Cas1 alone and Cas6-RT-Cas1—Cas2. RT activity is measured using an ELISA-based colorimetric reverse transcriptase activity assay (error bars, mean ± sd, $n = 3$ biologically independent experiments) (Catalog No. 11468120910, Roche Diagnostics, Indianapolis, IN). Source data are provided as a Source Data file. **e** Template-driven cDNA synthesis reactions of a fluorescent DNA primer annealed to DNA and RNA templates. Substrates are schematized (DNA, blue; RNA, purple). Expected cDNA synthesis reactions are indicated. Star indicates 6-carboxyfluorescein label. Results are representative of 3 independent experiments. **f** Template-driven cDNA synthesis reactions in the absence of different dNTPs and in the presence of added ddNTPs. Results are representative of 3 independent experiments. Uncropped gels are available in a Source Data file.

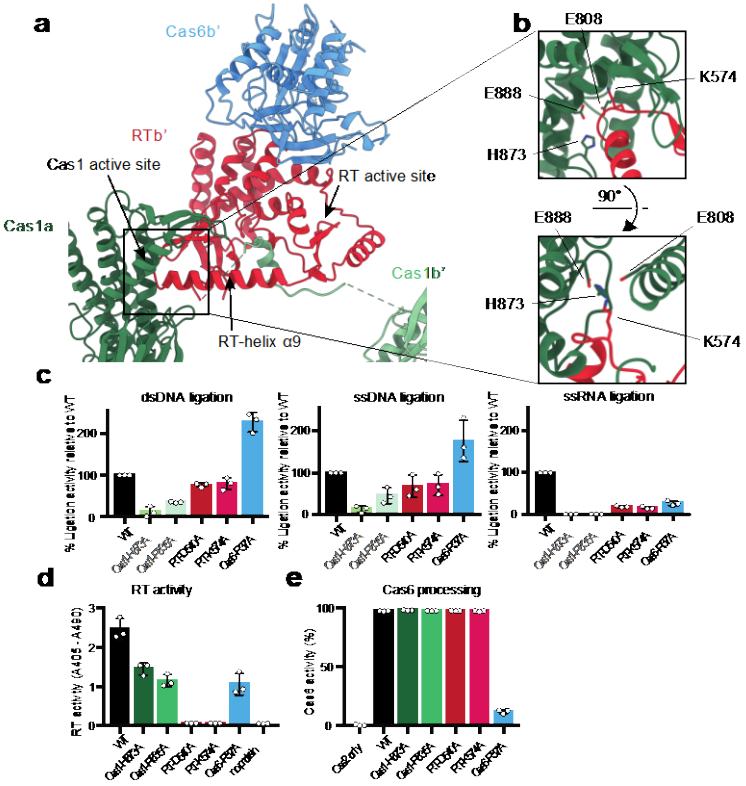


Figure 6. **Crosstalk between RT and Cas1 and Cas6 active sites. a** Interaction between RT-helix α9 and Cas1 active site (color-coding from Fig. 1). **b** Closeup of RT-helix in Cas1a and 90° rotation. Cas1 active site residues and RT-helix K574 are shown in stick configuration. **c** Ligation of fluorescent 35-nt dsDNA (0.5 μM), ssDNA (0.5 μM), and ssRNA (4 μM) protospacers into target pCRISPR by mutant Cas6-RT-Cas1—Cas2s. The percent ligation activity is calculated as the fraction of fluorescent products from the mutant complex relative to that from the WT complex (error bars, mean ± sd, $n = 3$ biologically independent experiments). Results are represented by colored bars: WT, black; Cas1 mutants, greens; RT mutants, reds; Cas6 mutant, blue. Cas1 mutant data are the same data shown in Fig. 2e and are depicted in pale greens with gray labels. Representative gels are shown in Supplementary Fig. 8c. **d** RT activity assays comparing the WT and mutant Cas6-RT-Cas1—Cas2s (error bars, mean ± sd, $n = 3$ biologically independent experiments; Catalog No. 11468120910, Roche Diagnostics, Indianapolis, IN). Statistical significance was assessed by comparing samples to WT control using unpaired, two-tailed $T$ tests ($\alpha = 0.05$) ($P = 0.0048$, Cas1-H873A; $P = 0.0019$, Cas1-R835A; $P = 0.0001$, RT-D540A; $P = 0.0001$, RT-K574A; $P = 0.0034$, Cas6-R37A). Same color-coding is used from **c** except Cas1 mutant data are shown in darker greens with black labels. **e** Cas6 activity assays comparing the WT and mutant Cas6-RT-Cas1—Cas2 proteins. The percent cleavage is calculated as the fraction of the fluorescent CRISPR repeat RNA that has been cleaved, with color-coding from **d** (error bars, mean ± sd, $n = 3$ biologically independent experiments). A representative gel is shown in Supplementary Fig. 8e. Source data for panels **c**-**e** are provided as a Source Data file.

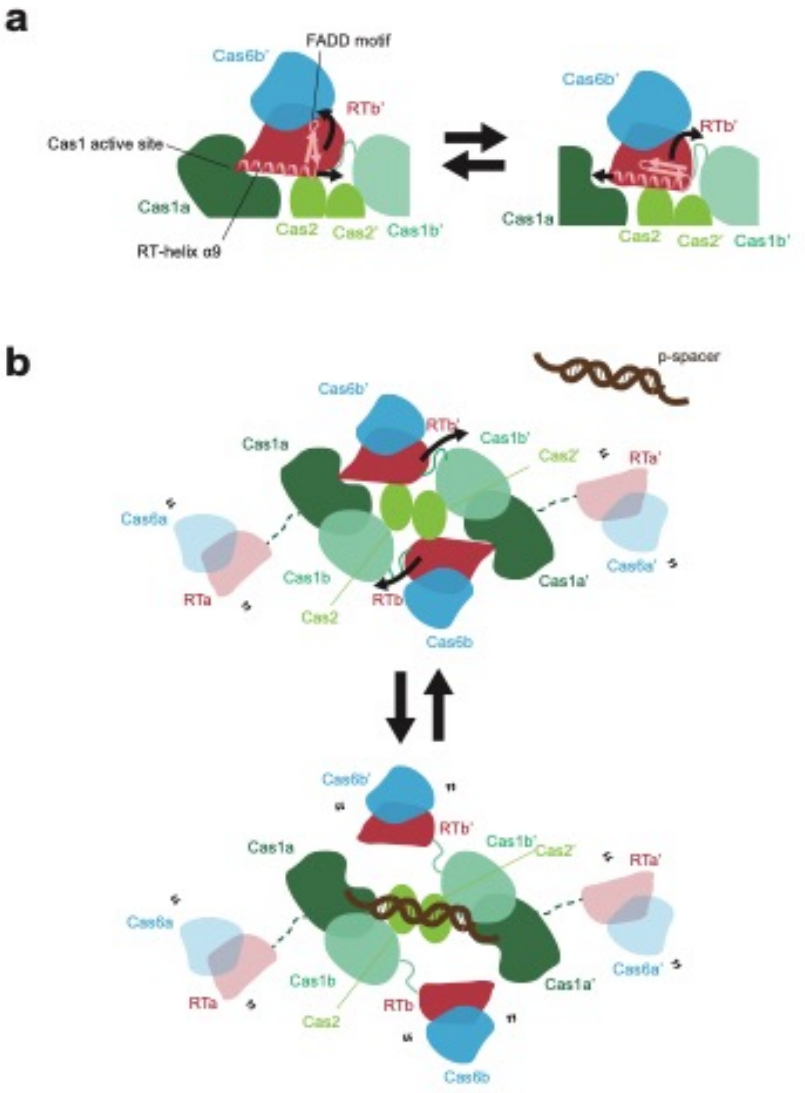


Figure 7. **Model for RT activation. a** Cartoon showing hypothesized motion of RT-helix α9 in and out of the Cas1a active site and the hypothesized motion of the β-strands attached to the FADD motif containing the active site residues. Proteins colored by domain: Cas6, blue; RT, red, Cas1a, dark green; Cas1b, turquoise; Cas2, lime green. Visible RT-Cas1 linkers are shown as solid turquoise lines. RT-helix α9 and connecting β-strands containing FADD motif are schematized and shown in pink. Hypothesized motions are indicated by curved arrows. **b** Cartoon showing hypothesized conformational change of the Cas6-RT-Cas1—Cas2 complex to accommodate a protospacer substrate, with color-coding from **a**. Missing Cas6-RT domains (Cas6a/RTa and Cas6a'/RTa') that are present in the complex but not visible are depicted as semi-transparent, connected to the rest of the structure by dashed dark green lines. Quotation marks indicate presumed mobility. Protospacer substrate is schematized in brown. Hypothesized motions are indicated by curved arrows.