Title
Mechanisms of CRISPR-Cas Immune Adaptation

Permalink
https://escholarship.org/uc/item/0442858x

Author
Wright, Addison Von

Publication Date
2018

Peer reviewed|Thesis/dissertation

Mechanisms of CRISPR-Cas Immune Adaptation

By

Addison V. Wright

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Molecular and Cell Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Jennifer Doudna, Chair
Professor Donald Rio
Professor Susan Marqusee
Professor Michiko Taga

Summer 2018

ABSTRACT

Mechanisms of CRISPR-Cas Immune Adaptation

by
Addison V. Wright

Doctor of Philosophy in Molecular and Cell Biology
University of California, Berkeley

Professor Jennifer A. Doudna, Chair

Prokaryotes have evolved a diverse array of strategies to prevent or mitigate infection by phage. Among these, CRISPR-Cas systems (clustered regularly interspaced short palindromic repeats - CRISPR-associated) are unique in that they adapt to infections by generating an immunological memory that allows the host cell to mount a robust defense against subsequent infections. These systems are characterized by the presence of a genomic feature called a CRISPR array, which is made up of an AT-rich leader sequence followed by a series of direct repeat sequences of 20-50 base pairs alternating with variable viral-derived spacer sequences of similar length. When a cell is infected by a phage, a small fragment of the phage genome can be captured and inserted into the CRISPR array as a new spacer through a process called acquisition. The CRISPR array can then be transcribed to generate crRNAs (CRISPR RNAs) that assemble with interference Cas proteins to surveil the cell for complementary nucleic acid sequences. If a match is found, the Cas proteins degrade the nucleic acid. While the interference proteins of CRISPR-Cas systems are highly diverse, acquisition is broadly conserved. The proteins Cas1 and Cas2 carry out the integration of new spacers at the CRISPR locus and are found in nearly all identified active CRISPR systems. This work examines the mechanisms of spacer acquisition with a focus on how Cas1 and Cas2 from different CRISPR systems recognize and maintain specificity for the CRISPR array.

Cas1 and Cas2 function as a complex to capture fragments of foreign DNA, called protospacers prior to integration, and insert them at the leader-proximal repeat through an integrase-like mechanism that results in duplication of the repeat. We find that the Cas1-Cas2 from the *Streptococcus pyogenes* type II CRISPR system integrates with high specificity *in vitro* into both plasmid and short linear targets, and we identify sequence motifs in the leader and repeat required for integration. We present the first evidence of full-site integration *in vitro* and show that the sequence requirements for full-site integration are stricter than those for half-site integration. Our biochemical data suggest that full-site integration acts as a checkpoint to ensure specificity, while half-site integration occurs more promiscuously due to the limited potential for it to introduce mutations at off-target sites.

Using x-ray crystallography and cryo-electron microscopy, we identify the structural basis of leader and repeat recognition by Cas1-Cas2 from the *Escherichia coli* type I system. Crystal structures of the proteins bound to substrates mimicking half-site and full-site products, supported with biochemical and bacterial genetic experiments, show that integration requires substantial distortion of the repeat DNA and that the repeat sequence is identified by its deformability. The EM structure of Cas1-Cas2 bound to an extended target and IHF, a host factor required for specificity, reveals that IHF bends the leader DNA 180° to bring an upstream recognition sequence into contact with Cas1 for additional sequence-specific recognition. These structures and assays show that Cas1-Cas2 rely on structural constraints to restrict full-site integration to the CRISPR array.

ACKNOWLEDGEMENTS

It is impossible to sufficiently thank and acknowledge everyone who supported me through my PhD, but I will do my best. Every graduate school journey is different, but none are completed without help. To all the people who encouraged me through the low points and celebrated with me during the highs, and to those who saw things in me I never saw myself, I am forever grateful.

Without Jennifer Doudna, none of this would be possible. Thank you for your support and encouragement. The freedom you gave me and the environment you established in your lab helped me grow into the scientist I am now and instilled in me the confidence to push myself farther and take risks as I move on in my career. I also thank Susan Marqusee and Don Rio for being consistent sources of support, from MCB200 through quals and committee meetings to letter-writing for postdocs, and Michi Taga, whose insight and encouragement have made her a welcome voice on committee as well.

More than anything, my time has been shaped by those I see and work with every day. James Nuñez welcomed me to the lab as my rotation mentor, even if it came as a surprise to him, and made me feel immediately at home. He got me hooked on acquisition from day one, and he has been a valuable friend in addition to an incredible scientist. Sam Sternberg served as my second mentor and managed to drag me kicking and screaming into the world of applications for a time, introducing me to the world of patent lawyers and competitive speed paper-writing while he was at it. I will always remember Stephen Floor, Ross Wilson, Philip Kranzusch, and Yun Bai with the awe I held them in when I first joined the lab, and I hope to one day approach their level of insight. I thank Mitch O'Connell for being no less scientifically impressive but also serving as a drinking and concert buddy in addition to a role model. And just as important has been Kaihong, whose ability to cut to the heart of a situation knows no equal and who you can always count on to tell you the truth.

I am grateful to Megan Hochstrasser, Kevin Doxzen, Akshay Tambe, Alex Seletsky, and Kendall Condon for making the lab an exciting and fun place to work over these years. I will miss karaoke, Tri-Lab, and softball. Those of us who remember Wacky Wednesday miss that too. Thank you to the lunch crew, Gavin Knott, Brady Cress, Kyle Watters, and Christof Fellman, for making forgetting my lunch at home a tempting proposition. And for my baymates in the Younglings Bay, Lucas Harrington and Janice Chen, I can't imagine the last couple years without you. Even though your baseline level of snark made me think that starting a company was just another long-running joke until Lucas starting showing up in collared shirts, it's been great to work alongside such amazing people. Without you, I wouldn't know the first thing about integrating verticals. To Josh and now Joy, it's been amazing to see you get started in the lab, and I can't wait to hear all of your successes. Finally, to Ben Oakes, my coton and friend since the interview days, thank you for your friendship. It's almost enough to make me put aside academia \and work with you.

Outside of lab, my friends and classmates have been critical to keeping me sane. Brian Castellano, Sean Higgins, Katie Herr, Charles Hesser, Elijah Mena, Gavin Schlissel, Kelsey Van Dalfsen, Ross Pederson, and more… Thank you for Friendsgivings, WNDCOTs and WNDCOTOWs, weekend getaways, and roommate hangouts. And to

Derek Ponticelli and Max Ashton, thank you for keeping me grounded and forcing me to have non-science conversations.

Finally, thank you to Sam and to my family. I am grateful to my parents for always supporting me and rooting for me through all the twists and turns. And to Sam, you've been my greatest source of support and comfort these last years. Through the late nights, the stressed-out days, and the wildly inaccurate ETAs you've stood by me. I am so grateful for that, and for your willingness to move to Boston and start a new life with me there. I couldn't do this without you. So to you, and to everyone else who played a role in my life big or small these last five years, thanks.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Biology and applications of CRISPR systems

A portion of the work presented in this chapter has been published previously as part of the following paper: Wright, A.V., Nuñez, J.K. & Doudna, J.A. (2016). Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. *Cell* **164** (1-2), 29-44. Where appropriate, portions of the text and figures have been updated to reflect advancements in the field since the time of publication.

1.1 Overview

CRISPR immunity is an adaptive immune system that allows prokaryotes to defend themselves against repeated challenge from foreign genetic elements such as phage. The initial phase of CRISPR immunity is called acquisition, wherein fragments of foreign DNA are captured and inserted at a unique genomic locus called a CRISPR array. This process is carried out by the proteins Cas1 and Cas2, which form an integrase complex. This dissertation focuses on the mechanism of integration by Cas1-Cas2. In particular, we will explore how Cas1-Cas2 specifically integrate at the CRISPR locus, ensuring successful acquisition and avoiding deleterious mutations from off-target integration. Cas1 evolved from a more traditional transposase, and its evolution from a promiscuous parasitic element to a highly specific molecular recorder represents the critical first step in the evolution of CRISPR immunity. However, the basis of this specificity was previously unknown. The following chapter provides an overview of CRISPR systems generally. In the second chapter, we investigate Cas1-Cas2 from a type II-A CRISPR system and determine its *in vitro* sequence specificity. We show the first evidence of full-site integration *in vitro* and establish the half-site to full-site transition as a critical specificity-determining checkpoint. In the third chapter, we present structural studies of a type I-E Cas1-Cas2 bound to integration intermediates. The structures reveal that recognition of the target DNA is driven largely by the sequence-dependent physical properties of the DNA, with relatively little direct sequence recognition through hydrogen bonding. They also reveal how a host factor, IHF, directs Cas1-Cas2 to the leader-proximal repeat. Chapter Four provides a summary of results and an outlook on the future of the field. In the appendix, we generate a split-Cas9 to probe interactions between the two lobes of the protein and the guide RNA and to serve as a platform for future applications. In addition to shedding light on the evolution of CRISPR systems, this investigation of Cas1-Cas2 specificity is essential for the future development of Cas1-Cas2 as a tool. Use of Cas1-Cas2 outside of its native context requires an understanding of where the complex might integrate in order to predict off-targets, identify potential integration sites in a foreign genome, and to inform engineering of the proteins for altered specificity.

1.2 Abstract

Bacteria and archaea possess a range of defense mechanisms to combat plasmids and viral infections. Unique among these are the CRISPR–Cas (clustered regularly interspaced short palindromic repeats–CRISPR associated) systems, which provide adaptive immunity against foreign nucleic acids. CRISPR systems function by acquiring genetic records of invaders to facilitate robust interference upon reinfection. In this review we discuss recent advances in understanding the diverse mechanisms by which Cas proteins respond to foreign nucleic acids and how these systems have been harnessed for precision genome manipulation in a wide array of organisms.

1.3 Introduction

CRISPR–Cas (clustered regularly interspaced short palindromic repeats–CRISPR associated) adaptive immune systems are found in roughly 50% of bacteria and 90% of archaea (Makarova et al., 2015). These systems function alongside restriction-modification systems, abortive infections, and adsorption blocks to defend prokaryotic populations against phage infection (Labrie et al., 2010). Unlike other mechanisms of cellular defense, which provide generalized protection against any invaders not possessing countermeasures, CRISPR immunity functions analogously to vertebrate adaptive immunity by generating records of previous infections to elicit a rapid and robust response upon reinfection.

CRISPR–Cas systems are generally defined by a genomic locus called the CRISPR array, a series of ~20-50 base-pair (bp) direct repeats separated by unique "spacers" of similar length and preceded by an AT-rich "leader" sequence (Jansen et al., 2002; Kunin et al., 2007). Nearly two decades after CRISPR loci were first identified in *Escherichia coli*, spacers were found to derive from viral genomes and conjugative plasmids, serving as records of previous infection (Bolotin et al., 2005; Ishino et al., 1987; Mojica et al., 2005; Pourcel et al., 2005). Sequences in foreign DNA matching spacers are referred to as "protospacers." In 2007, it was shown that a spacer matching a phage genome immunizes the host microbe against the corresponding phage and that infection by a novel phage leads to the expansion of the CRISPR array by addition of new spacers originating from the phage genome (Barrangou et al., 2007).

CRISPR immunity is divided into three stages: spacer acquisition, CRISPR RNA (crRNA) biogenesis, and interference (Fig. 1.1) (Makarova et al., 2011b; van der Oost et al., 2009). During spacer acquisition, also known as adaptation, foreign DNA is identified, processed, and integrated into the CRISPR locus as a new spacer. The crRNA biogenesis or expression stage involves CRISPR locus transcription, often as a single pre-crRNA, and its subsequent processing into mature crRNAs that each contain a single spacer. In the interference stage, an effector complex uses the crRNA to identify and destroy any phage or plasmid bearing sequence complementarity to the spacer sequence of the crRNA.

These steps are carried out primarily by Cas proteins, which are encoded by *cas* genes flanking the CRISPR arrays. The specific complement of *cas* genes varies widely. CRISPR–Cas systems can be classified based on the presence of "signature genes" into six types, which are additionally grouped into two classes (Fig. 1.2) (Makarova et al., 2011b; 2015; Mohanraju et al., 2016; Shmakov et al., 2015). Types I-III are the best studied, while types IV-VI have been more recently identified (Makarova and Koonin, 2015; Makarova et al., 2015; Shmakov et al., 2015). The signature protein of type I systems is Cas3, a protein with nuclease and helicase domains that functions in foreign DNA degradation to cleave DNA that is recognized by the multi-protein–crRNA complex Cascade (CRISPR-associated complex for antiviral defense). In type II systems, the signature *cas9* gene encodes the sole protein necessary for interference. Type III systems are signified by Cas10, which assembles into a Cascade-like interference complex for target search and destruction. Type IV systems have Csf1, an uncharacterized protein proposed to form part of a Cascade-like complex, though these systems are often found as isolated *cas* genes without an associated CRISPR array (Makarova and Koonin, 2015). Type V systems are highly variable but also contain a single effector bearing a RuvC nuclease domain called Cas12a-e (formerly known as Cpf1, C2c1, C2c3, CasY, and CasX, respectively) (Burstein et al., 2017; Koonin et al., 2017; Shmakov et al., 2015; Zetsche et al., 2015a). Type VI systems have Cas13, previously known as C2c2, a large protein with two HEPN (higher eukaryotes and prokaryotes nucleotide-binding) RNase domains (Shmakov et al., 2015; 2017). Type I, III, and IV systems are considered Class 1 systems based on their multi-subunit effector complexes, while the single-subunit effector type II, V, and VI systems are grouped into Class 2 (Koonin et al., 2017).

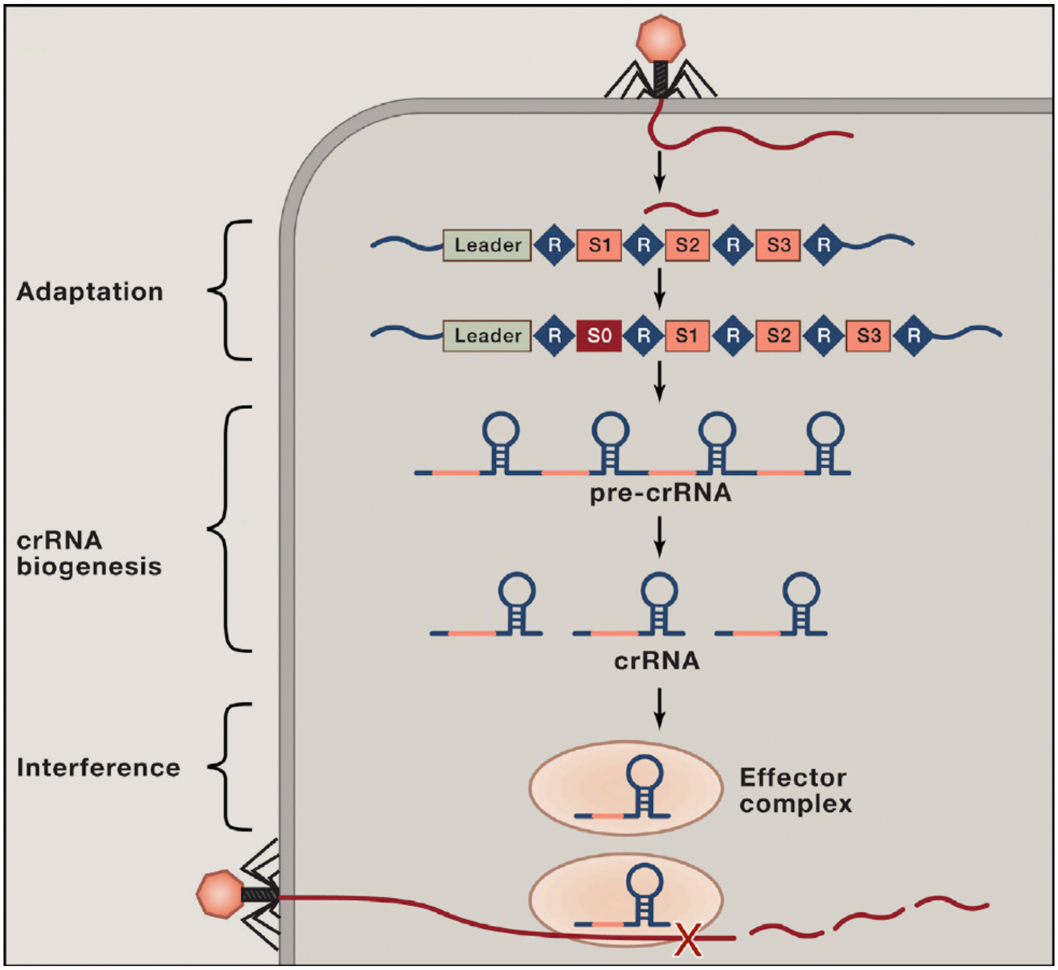


Figure 1.1 | Three stages of CRISPR immunity Upon introduction of foreign DNA, the adaptation machinery selects protospacers and inserts them into the leader end of the CRISPR locus. During crRNA biogenesis, the CRISPR locus is transcribed and sequence elements in the repeats direct processing of the pre-crRNA into crRNAs each with a single spacer. The crRNA then assembles with Cas proteins to form the effector complex, which acts in the interference stage to recognize foreign nucleic acid upon subsequent infection and degrade it.

The study of CRISPR biology has revealed enzyme mechanisms that can be harnessed for precision genome engineering and other applications, leading to an explosion of interest in both native CRISPR pathways and the use of these systems for applications in animals, plants, microbes, and humans. In this review we discuss recent advancements in the field that reveal unexpected divergence as well as unifying themes underlying the three stages of CRISPR immunity. In each case we highlight the ways in which these systems are being harnessed for applications across many areas of biology.

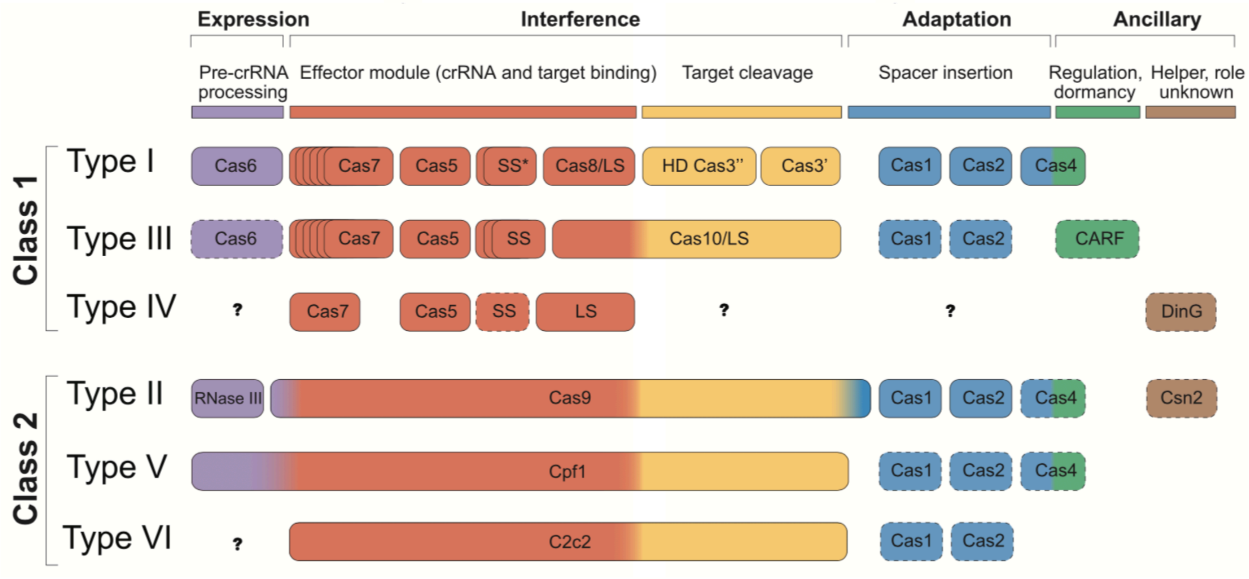


Figure 1.2 | Classification of CRISPR-Cas systems. CRISPR systems are classified as class 1 or class 2 based on whether interference is carried out by a multi-protein or single-protein effector complex. They are further divided into six types based on the specific *cas* genes present. Dashed lines indicate proteins found in only some systems within a given type. Adapted from Mohanraju et al., 2016.

1.4 Acquisition: creating genetic records of past infections

CRISPR immunity begins with the detection and integration of foreign DNA into the host cell's chromosome. In the *Streptococcus thermophilus* type II-A system where acquisition was first detected experimentally, new spacers from bacteriophage DNA are inserted into the leader end of the CRISPR locus, causing duplication of the first repeat to maintain the repeat-spacer architecture (Fig. 1.1) (Barrangou et al., 2007). Subsequent studies using the *E. coli* type I-E system verified that Cas1 and Cas2 mediate spacer acquisition (Datsenko et al., 2012; Swarts et al., 2012; Yosef et al., 2012). The selection of new protospacer sequences is nonrandom and in most systems depends on the presence of a 2-5 nucleotide protospacer adjacent motif (PAM) found next to the protospacer sequence (Deveau et al., 2008; Mojica et al., 2009). PAM-specific selection of protospacers is critical for immunity, as crRNA-guided interference in most systems depends on the PAM sequence for foreign DNA detection and destruction, which avoids self-targeting at the PAM-free CRISPR locus. Interestingly, spacers originating from the host genome are present in almost 20% of CRISPR-containing organisms, suggesting alternative roles of the CRISPR-Cas machinery in directing other processes such endogenous gene regulation and genome evolution (Westra et al., 2014). Spacer acquisition has been observed experimentally in various systems across types I-III. Here we focus on mechanistic studies of acquisition in type I-E and type II-A systems, in which the most comprehensive studies have been done.

1.4.1 Type I acquisition

Acquisition in *E. coli* occurs via two mechanisms – naïve and primed (Fig. 1.3). Naïve acquisition initiates upon infection by previously unencountered DNA and relies on the Cas1-Cas2 integrase complex to recognize and acquire new spacers from foreign DNA. Overexpression of Cas1 and Cas2 in the absence of other Cas proteins leads to the acquisition of 33 bp spacers at the leader-proximal end of the CRISPR array (Datsenko et al., 2012; Yosef et al., 2012). The PAM of the *E. coli* CRISPR–Cas system was identified as 5′-AWG-3′, with the G becoming the first nucleotide of the integrated spacer (Datsenko et al., 2012; Díez-Villaseñor et al., 2013; Levy et al., 2015; Nuñez et al., 2014; Savitskaya et al., 2013; Shmakov et al., 2014; Swarts et al., 2012; Yosef et al., 2012; 2013). In addition to the PAM, a dinucleotide motif, AA, found at the 3′ end of the protospacer was also shown to be present in a disproportionately large number of spacers (Yosef et al., 2013) . A crystal structure of the Cas1-Cas2 complex bound to an unprocessed protospacer revealed sequence-specific contacts with the 5′-CTT-3′ sequence on the PAM-complementary strand, suggesting that Cas1 recognizes PAM sites on potential protospacers before they are processed for integration (Wang et al., 2015).

After a spacer is acquired from a new invader, the resulting crRNA assembles with Cas proteins to form Cascade, the interference complex capable of targeting PAM-adjacent DNA sequences matching the spacer sequence of the crRNA (Brouns et al., 2008; Jore et al., 2011; Lintner et al., 2011). Upon target binding, the helicase/nuclease Cas3 is recruited to the site and processively degrades the foreign DNA (Hochstrasser et al., 2014; Mulepati and Bailey, 2011; Sinkunas et al., 2011; 2013; Westra et al., 2012). Strikingly, when Cascade encounters a mutant PAM or protospacer that prevents Cas3 degradation, hyperactive spacer acquisition from the targeted plasmid or genome is triggered in a process called “priming” (Fig. 1.3) (Datsenko et al., 2012; Li et al., 2014; Richter et al., 2014; Savitskaya et al., 2013; Swarts et al., 2012). Priming increases the host’s repertoire of functional spacers, allowing the host to adapt to invaders that evade the CRISPR–Cas system by mutation. Cascade is capable of binding escape mutant target sites, and single-molecule studies showed that the presence of Cas1 and Cas2 allows for the recruitment of Cas3 to these sites (Blosser et al., 2015; Redding et al., 2015; Richter et al., 2014). The recruited Cas3 can then translocate in either direction, in contrast to the unidirectional movement observed at perfect targets, without degrading the target DNA (Redding et al., 2015). Cas1 and Cas2 may accompany the translocating Cas3 and be activated for protospacer selection, allowing for robust acquisition on either side of the target site. More recent bulk biochemistry has suggested that the degradation products of Cas3 serve directly as protospacers, and *in vivo* work has led to a model that primed acquisition coincides with successful interference rather than acting as a separate process on interference-incompetent phage (Künne et al., 2016; Staals et al., 2016). Further work is required to reconcile these apparently competing models of Cas3 activity in priming.

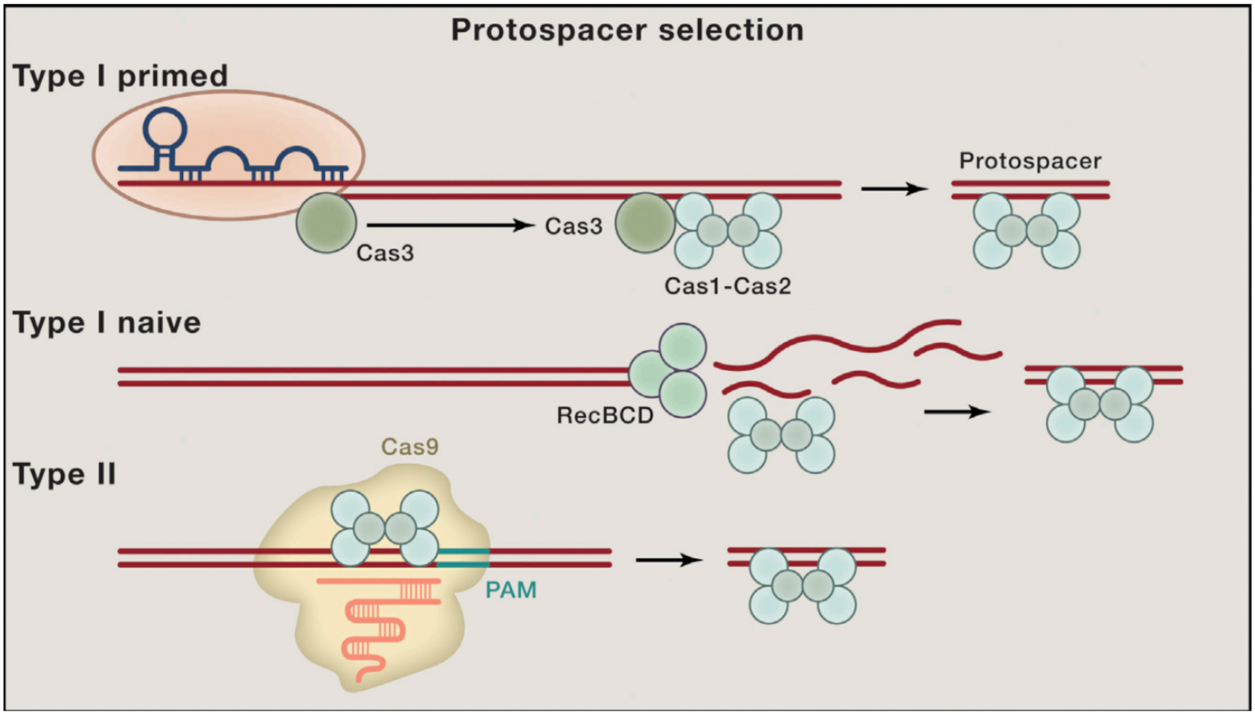


Figure 1.3 | Protospacer selection in adaptation. The selection of protospacers for acquisition is poorly understood, but studies suggest at least three distinct mechanisms for the selection of substrates for integration. In type I systems, primed adaptation occurs when Cascade binds a partially mismatched target. The nuclease/helicase Cas3 is recruited to the target site and then translocates along the target DNA to a new site, possibly degrading the DNA along the way. The DNA is passed to Cas1-Cas2 to be used in integration. In *E. coli*, naïve adaptation involves the nuclease/helicase RecBCD. The degradation products appear to serve as substrates for Cas1-Cas2. In type II systems, Cas9 recognizes PAM sites and likely recruits Cas1-Cas2 to acquire the flanking sequence. The nuclease/helicase AddAB is also involved, possibly in a role similar to that of RecBCD.

Primed acquisition has also been shown experimentally in the *P. atrosepticum* type I–F system, in which Cas2 and Cas3 are naturally fused as a single polypeptide that associates with Cas1, as well as in the *Haloarcula hispanica* type I-B system, where naïve acquisition was not experimentally observed (Li et al., 2014; Richter et al., 2014; 2012). Acquisition in *H. hispanica* also requires Cas4, a 5′→3′ exonuclease found in most type I subtypes as well as type II-B and type V systems that generates integration-competent protospacers (Kieper et al., 2018; Lee et al., 2018; Lemak et al., 2013; Li et al., 2014; Makarova et al., 2015). Although Cas1 and Cas2 may be the minimal proteins required for spacer acquisition in some systems, the association of Cas1, Cas2, and the interference machinery allows the host to coordinate robust adaptive immunity in type I systems.

1.4.2 Self vs non-self recognition

The mechanism underlying the preference for foreign over self DNA during protospacer selection remained poorly understood until a study on spacer acquisition during naïve acquisition. Spacer acquisition in *E. coli* was shown to be highly dependent on DNA replication, and foreign-derived spacers were preferred over self-derived spacers by about 100- to 1,000-fold (Levy et al., 2015). Analysis of the source of self-derived spacers demonstrated that protospacers were acquired largely from genomic loci predicted to frequently generate stalled replication forks and double-stranded DNA. Such harmful dsDNA breaks are repaired by the helicase/nuclease RecBCD complex, which degrades the broken ends until reaching a Chi-site, after which only the 5′ end is degraded (Dillingham and Kowalczykowski, 2008). Due to the lower frequency of Chi sites in foreign DNA, RecBCD is predicted to preferentially degrade plasmids and viral DNA, resulting in the generation of candidate protospacer substrates for Cas1 and Cas2 (Levy et al., 2015) (Fig. 1.3). RecBCD degrades DNA asymmetrically, yielding single-stranded fragments ranging from tens to hundreds of nucleotides long from one strand and kilobases long from the other (Dillingham and Kowalczykowski, 2008). It is unclear how Cas1-Cas2 substrates, which are 33 bp long and partially double-stranded with 3′ overhangs, are generated from RecBCD products (Nuñez et al., 2015a; 2015b; Wang et al., 2015). It is possible that ssDNA products re-anneal to produce partial duplexes, followed by processing to 33 bp by an unknown mechanism prior to integration into the CRISPR locus. Crystal structures of Cas1-Cas2 with bound protospacer reveal that the complex defines the length of the duplex region of the protospacer via a ruler mechanism and may cleave the 3′ overhangs to their final length (Nuñez et al., 2015a; Wang et al., 2015). The involvement of a helicase/nuclease in both type I-E primed and naïve acquisition (Cas3 and RecBCD, respectively) as well as in Cas4-containing subtypes hints at a conserved mechanism for protospacer generation. It is also worth noting that RecBCD is conserved primarily in Gram-negative bacteria, while Gram-positive bacteria and archaea rely on AddAB and HerA-NurA, respectively, to fill a similar role (Blackwood et al., 2013; Dillingham and Kowalczykowski, 2008). Indeed, recent work has shown that AddAB is required for efficient spacer integration by a *S. pyogenes* type II CRISPR system expressed in *Staphylococcus aureus* (Modell et al., 2017).

1.4.3 Mechanism of protospacer integration

Cas1 and Cas2 play central roles in the acquisition of new spacers, where they function as a complex (Nuñez et al., 2014). Crystal structures of Cas1 and Cas2, with or without bound protospacer, revealed two copies of a Cas1 dimer bridged by a central Cas2 dimer (Nuñez et al., 2014; 2015a; Wang et al., 2015). Cas1 functions catalytically, while Cas2 appears to serve a primarily structural role (Arslan et al., 2014; Datsenko et al., 2012; Nuñez et al., 2014; Yosef et al., 2012).

The first insight into the mechanism of protospacer integration was gained by Southern blot analysis of the genomic CRISPR locus of *E. coli* cells overexpressing Cas1 and Cas2 (Arslan et al., 2014). This revealed integration intermediates consistent with two transesterification reactions, where each strand of the protospacer is integrated into opposite sides of the leader-proximal repeat (Fig. 1.4). This integrase-like model was further bolstered by the *in vitro* reconstitution of protospacer integration into a plasmid-encoded CRISPR locus using purified Cas1-Cas2 complex (Nuñez et al., 2015b). The integration reaction required double-stranded DNA protospacers with 3′-OH ends that are integrated into plasmid DNA via a direct nucleophilic transesterification reaction, reminiscent of retroviral integrases and DNA transposases (Engelman et al., 1991; Mizuuchi and Adzuma, 1991). Cas1 was discovered to have an evolutionary link with transposases as well, as *cas1* genes have been found in bacterial transposons, where they are functional for mobilization of the transposon (Béguin et al., 2016; Hickman and Dyda, 2015; Krupovic et al., 2014).

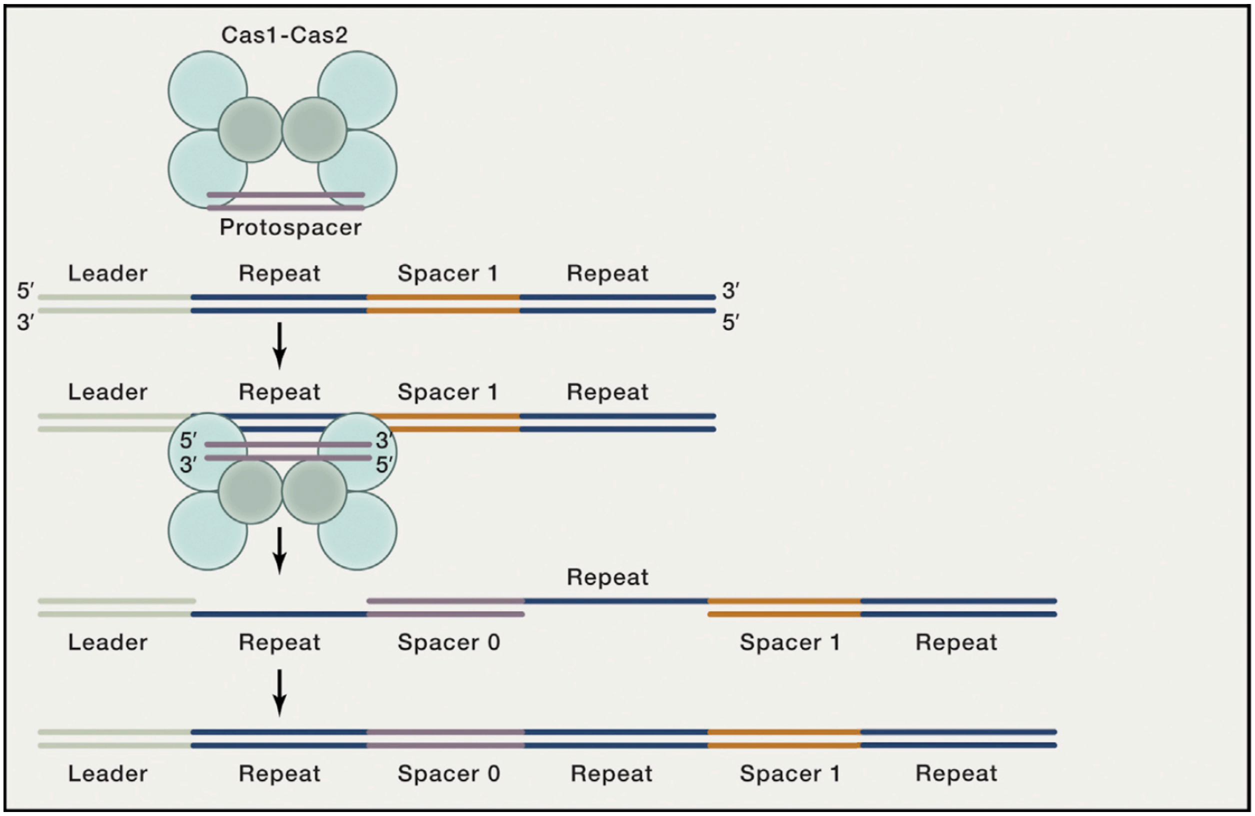


Figure 1.4 | Protospacer integration. Cas1-Cas2 act as an integrase to insert protospacers into the CRISPR locus as new spacers. The complex with protospacer bound recognizes the leader-adjacent repeat and catalyzes a pair of transesterification reactions. The 3′ OH of each protospacer makes a nucleophilic attack on the repeat backbone, one at the leader side and one at the spacer side. The resulting gapped product is then repaired, causing duplication of the first repeat.

Although deep sequencing of *in vitro* integration products revealed preferential protospacer integration adjacent to the first repeat, confirming that Cas1-Cas2 directly recognize the CRISPR locus, integration also occurred at the borders of every repeat at varying levels (Nuñez et al., 2015b). This contrasts with spacer acquisition only occurring at the first repeat in *E. coli in vivo* (Datsenko et al., 2012; Swarts et al., 2012; Yosef et al., 2012). To determine if the Cas1-Cas2 complex has sequence specificity for the leader-repeat sequence, a recent study took advantage of the Cas1-catalyzed disintegration reaction, a reversal of the integration reaction also observed with retroviral integrases and transposases (Chow et al., 1992; Rollie et al., 2015). Disintegration activity was stimulated when using the correct leader-repeat border sequences, highlighting intrinsic sequence-specific recognition by Cas1. Furthermore, disintegration was faster at the leader-repeat junction compared to the repeat distal end (Rollie et al., 2015). Taken together, protospacer integration likely begins at the leader-repeat junction via sequence-specific recognition by Cas1, followed by a second nucleophilic attack at the repeat distal end. This ensures precise duplication of the first repeat, as observed *in vivo*, after DNA repair by host proteins. Subsequent studies revealed that the *E. coli* system requires an additional factor, IHF (Integration Host Factor), to bind in the leader for specific integration to occur (Nuñez et al., 2016; Yoganand et al., 2017). The integration mechanism is hypothesized to be highly specific, as almost all acquired spacers with a corresponding AAG PAM are oriented with the 5′-G at the leader-proximal end, leading to functional crRNA-dependent targeting by Cascade and Cas3 (Shmakov et al., 2014). A preference for integration in the proper orientation was observed *in vitro* when protospacers with a 5′-G were used (Nuñez et al., 2015b); however, inclusion of part of the PAM in spacers has only been observed in *E. coli*, raising the question of how Cas1-Cas2 in other systems properly orient the integration reaction.

1.4.4 Type II acquisition

While most mechanistic work on acquisition has been performed in type I systems, recent studies in type II systems have also shed light on key aspects of spacer acquisition. One generalizable finding in type II systems is the dependence of acquisition on infection by defective phage (Hynes et al., 2014). A significant problem with CRISPR immunity is the time required for foreign DNA to be identified, integrated into the CRISPR locus, transcribed, processed and assembled into an interference complex that must then begin the search for appropriate targets. Since lytic phage can kill cells within 20 minutes, providing insufficient time for this multi-step process, Hynes and colleagues tested the hypothesis that initial immunization takes place from infection by a defective phage. Supplementation of active phage with UV-irradiated phage or phage susceptible to a restriction-modification system stimulated spacer acquisition compared to that observed with active phage alone (Hynes et al., 2014). The authors speculate that acquisition from compromised phage might also represent the dominant mode of acquisition in wild populations, allowing for a small subset of the population to acquire resistance and escape without needing to outpace a rapidly reproducing phage.

1.4.5 Type II acquisition machinery

Type II systems are subdivided into II-A, II-B, and II-C based on the presence or absence of an additional *cas* gene alongside the minimal complement of *cas1*, *cas2*, and *cas9*. Type II-A systems contain *csn2* while type II-B systems, which are least commonly found, contain *cas4* (Chylinski et al., 2014; Makarova et al., 2011b). Type II-C systems comprise only the minimal three genes. Csn2 has been shown to be essential for acquisition in several type II-A systems (Barrangou et al., 2007; Heler et al., 2015; Wei et al., 2015b). It forms a tetramer with a torroidal architecture that binds and slides along free DNA ends, though its function in CRISPR systems is unclear (Arslan et al., 2013; Ellinger et al., 2012; Koo et al., 2012; Lee et al., 2012). Cas4, discussed above, is likely involved in acquisition in type II-B systems. type II-C systems, which constitute the majority of identified type II systems (Chylinski et al., 2014; Makarova et al., 2015), are possibly functional for acquisition in the absence of auxiliary acquisition factors, though in the case of the *Campylobacter jejuni* system acquisition was only observed following infection by phage encoding a Cas4 homolog (Hooton and Connerton, 2014).

Two studies demonstrated that, in addition to Cas1, Cas2, and Csn2, Cas9 plays a necessary role in the acquisition of new spacers in type II systems (Heler et al., 2015; Wei et al., 2015b). Both groups, one working with the CRISPR1 type II-A system of *S. thermophilus*, the other with the type II-A system of *Streptococcus pyogenes* and the CRISPR3 system of *S. thermophilus*, also type II-A, showed that wild-type or catalytically inactive Cas9 (dCas9) supported robust spacer acquisition whereas deletion of Cas9 abolished spacer acquisition. It is proposed that Cas9 serves to recognize PAM sites in potential protospacers and mark them for recognition by Cas1 and Cas2 (Fig. 1.3). This hypothesis was confirmed by mutating the PAM-interacting residues of Cas9, resulting in complete loss in PAM-specificity in the newly acquired spacers (Heler et al., 2015). This presents a striking contrast to the *E. coli* type I-E system, where Cas1-Cas2 recognizes PAM sequences independently.

Intriguingly, expression of dCas9 results in the acquisition of primarily self-targeting spacers, suggesting that many acquisition events lead to self-targeting and suicide (Wei et al., 2015b). Microbial populations may rely on a few individuals to acquire phage resistance while the rest succumb to infection or CRISPR-mediated suicide. Some systems, such as that found in *E. coli*, may evolve to use host processes to bias acquisition away from self-targeting. Alternatively, *S. thermophilus* might have mechanisms of self-non-self discrimination that were masked in the strain overexpressing CRISPR proteins. Phage challenge experiments with wild-type *S. thermophilus* revealed that some sequences were acquired as spacers disproportionately often across multiple experiments, suggesting that the type II acquisition machinery has preferences in addition to Cas9-dependent PAM selection, though no clear pattern emerged with respect to the genomic location or sequence of protospacers that indicated a basis for the preferences (Paez-Espino et al., 2013).

Additionally, it was demonstrated that the four proteins of the *S. pyogenes* CRISPR system (Cas1, Cas2, Csn2, and Cas9) form a complex, suggesting that Cas9 directly recruits the acquisition proteins to potential targets (Heler et al., 2015). While drawing comparisons between the involvement of Cas9 in acquisition and primed acquisition in type I systems is tempting, neither group saw evidence that acquisition was affected by the presence of existing spacers matching or closely matching the infecting phage or plasmid (Heler et al., 2015; Wei et al., 2015b). In addition, while the trans-activating crRNA (tracrRNA) that forms a complex with Cas9 and the crRNA is necessary for acquisition, it is unclear whether a corresponding crRNA is also required (Heler et al., 2015; Wei et al., 2015b). Future mechanistic work will be required to shed light on the similarities between Cas9-mediated spacer acquisition and the primed acquisition in type I systems.

1.4.6 Type II protospacer integration

The sequence requirements for protospacer integration in type II-A systems were recently demonstrated in *S. thermophilus* (Wei et al., 2015a). Similar to *E. coli*, the leader and a single repeat were sufficient to direct integration. Furthermore, only the ten nucleotides of the leader proximal to the first repeat are required to license the integration of new spacers, in contrast to the 60 nt minimal requirement in *E. coli* (Wei et al., 2015a; Yosef et al., 2012). A limited mutational study of the repeat showed that the first two nucleotides are necessary for acquisition, while the final two nucleotides can be mutated without consequence (Wei et al., 2015a). Thus, Cas1-Cas2-catalyzed integration at the leader-repeat junction is sequence-specific while the attack at the repeat-spacer junction is determined by a ruler mechanism, in agreement with observations from experiments in the *E. coli* system (Díez-Villaseñor et al., 2013). Together, these findings support the functional conservation of the Cas1-Cas2 integrase complex despite divergent mechanisms of protospacer selection between types I and II CRISPR–Cas systems.

1.4.7 CRISPR integrases as genome modifying tools

As with many other Cas proteins, the Cas1-Cas2 integrase complex shows promise for use in modifying genomes. While Cas1-Cas2 catalyze a reaction similar to that of many integrases and transposases, they exhibit several fundamental differences that make them uniquely suited to certain applications. Cas1-Cas2 complexes lack sequence specificity for the DNA substrate to be integrated, a property that could make the system ideal for barcoding genomes. Genome barcoding allows for tracking lineages originating from individual cells, facilitating studies of population evolution, cancer, development, and infection (Blundell and Levy, 2014). Cas1-Cas2 complexes integrate short DNA sequences, in contrast with current techniques based on recombinases that integrate entire plasmids, resulting in potential fitness costs and unwanted negative selection (Blundell and Levy, 2014). Interestingly, *in vitro* integration of DNA substrates into plasmid targets revealed integration into non-CRISPR sites (Nuñez et al., 2015b), suggesting that Cas1-Cas2 may be harnessed to integrate into a wide array of target sequences. Recent work has demonstrated the ability for Cas1-Cas2 to encode information, delivered as electroporated oligonucleotides, in a time-resolved manner, allowing for reconstruction of a brief movie from the CRISPR arrays of a bacterial population (Shipman et al., 2017). An extension of this principle might allow for Cas1-Cas2 to be used to record information about cellular state or transcriptional activity as well. A greater understanding of the minimal functional recognition motif for various Cas1-Cas2 integrases will facilitate the development of these technologies.

1.5 crRNP biogenesis: Generating molecular sentinels for the cell

CRISPR immune systems use RNA-programmed proteins to patrol the cell in search of DNA molecules bearing sequences complementary to the crRNA. Assembly of these molecular sentinels begins with transcription of the CRISPR locus to generate long, precursor CRISPR RNAs (pre-crRNAs), followed by processing into short crRNA guides (Brouns et al., 2008; Carte et al., 2008). The promoter is embedded within the AT-rich leader sequence upstream of the repeat-spacer array, or sometimes within the repeat sequences (Zhang et al., 2013). Here we briefly review the processing of pre-crRNAs catalyzed by the Cas6 endoribonuclease family in type I and III systems and a distinct processing pathway in type II systems that involves endogenous RNase III, Cas9, and a transactivating crRNA (tracrRNA). The crRNA biogenesis pathway has been extensively reviewed elsewhere (Charpentier et al., 2015; Hochstrasser and Doudna, 2015). More recently, Cas12a and Cas13 have been shown to directly process their own pre-crRNA based on recognition of the repeat hairpin, while Cas12b and Cas12e appear to rely on tracrRNA-mediated processing (Burstein et al., 2017; East-Seletsky et al., 2016; Fonfara et al., 2016; Shmakov et al., 2015).

1.5.1 Processing by Cas6 endoribonucleases

Type I and type III systems employ Cas6 endoribonucleases to cleave pre-crRNAs sequence-specifically within each repeat (Brouns et al., 2008; Carte et al., 2008; Haurwitz et al., 2010). Although Cas6 homologs are variable in sequence, they share a conserved cleavage mechanism that results in crRNA guides comprising an entire spacer sequence flanked by portions of the repeat sequence on the 5′ and 3′ ends. Mature crRNA guides consist of an 8 nt 5′ handle derived from the repeat sequence and variable lengths of the repeat at the 3′ handle, which is further trimmed by as yet unidentified cellular nuclease(s) in type III systems (Hale et al., 2008). A notable exception is in type I-C systems, which utilize a Cas5 variant for crRNA processing, leaving an 11 nt 5′ handle and 21-26 nt at the 3′ end (Garside et al., 2012; Nam et al., 2012). In other type I systems, Cas5 subunits serve a non-catalytic role capping the 5′ end of the crRNA in Cascade complexes.

In type I-C, I-D, I-E and I-F systems, the repeats form stable hairpin structures that allow for structure- and sequence-specific cleavage by Cas6 at the base of the hairpin (Gesner et al., 2011; Haurwitz et al., 2010; Sashital et al., 2011). After cleavage, the hairpin constitutes the 3′ handle of the crRNA. The Cas6 proteins in *Haloferax volcanii* (Cas6b), *E. coli* and *T. thermophilus* (Cas6e), and *Pseudomonas aeruginosa* (Cas6f) remain stably bound to the 3′ handle and eventually becomes part of the Cascade complex (Brendel et al., 2014; Brouns et al., 2008; Gesner et al., 2011; Haurwitz et al., 2010; Sashital et al., 2011).

Type I-A, I-B, III-A and III-B repeat sequences are non-palindromic and predicted to be unstructured in solution (Kunin et al., 2007). Thus, the respective Cas6 is thought to rely on sequence for specificity rather than structure. Interestingly, a crystal structure of the type I-A Cas6 bound to its cognate RNA structure reveals Cas6 inducing a 3 bp hairpin in the RNA that positions the scissile phosphate in the enzyme active site (Shao and Li, 2013). It remains unknown whether other Cas6s that recognize non-palindromic repeats have a similar mechanism of RNA stabilization. Following or concurrent with the maturation of the crRNAs, the Cas proteins involved in interference assemble into the final effector complex that functions to recognize and destroy targets bearing sequence complementarity to the crRNA. In systems where Cas6 remains bound to the crRNA, it may serve to nucleate the assembly of the subunits that constitute the effector complex backbone along the crRNA. In type III systems, the number of backbone subunits defining the complex length is variable, and any unprotected crRNA remaining is degraded (Hale et al., 2008; Staals et al., 2014).

1.5.2 Processing in type II systems

Type II systems rely on a different mechanism to process pre-crRNAs. In types II-A and II-B, pre-crRNA cleavage specificity is aided by a tracrRNA that has sequence complementarity to the CRISPR repeat sequence (Deltcheva et al., 2011). The gene encoding the tracrRNA is typically located either proximal to or within the CRISPR–*cas* locus (Chylinski et al., 2014). Upon crRNA:tracrRNA base pairing, which is stabilized by

Cas9, endogenous RNase III cleaves the pre-crRNA at the repeat. The reliance on RNAse III, which is not found in archaea, may explain why type II systems are limited to bacteria (Garrett et al., 2015). An unknown nuclease trims the 5′ end of the crRNA to remove the flanking repeat sequence and portions of the spacer. In *S. pyogenes*, the 30 nt spacer sequence is trimmed to the 20 nt that base-pairs with complementary foreign sequences during interference (Deltcheva et al., 2011; Jinek et al., 2012).

In the *Neisseria meningitidis* and *C. jejuni* Type II-C systems, each repeat sequence encodes a promoter, resulting in varying lengths of pre-crRNAs depending on the transcription start site (Dugar et al., 2013; Zhang et al., 2013). Although RNase III-mediated pre-crRNA processing can still occur, RNase III is dispensable for interference in these systems (Zhang et al., 2013). Thus, Cas9 is able to complex with the pre-crRNA and unprocessed tracrRNA for functional target interference without further processing of the pre-crRNAs.

1.5.3 Cas6 as a biotechnology tool

The Cas6 homolog from type I-F systems, Cas6f (also known as Csy4), was the first Cas protein to be repurposed as a tool. Following demonstration of the sequence-specificity of Cas6f binding and cleavage, the protein has been used for the purification of tagged RNA transcripts from cells (Haurwitz et al., 2010; Lee et al., 2013; Salvail-Lacoste et al., 2013; Sternberg et al., 2012). Subsequent studies showed that Cas6f could be used to alter the translation and stability of tagged mRNAs, allowing for post-transcriptional regulation of protein expression (Borchardt et al., 2015; Du et al., 2016; Nissim et al., 2014). Cas6f has also been used alongside Cas9 to process multiple guide RNAs from a single transcript, greatly facilitating multiplexed editing (Tsai et al., 2014).

1.6 Interference: Precise, programmable DNA binding and cleavage

Implementation of CRISPR systems to provide immunity involves RNA-guided recognition and precision cutting of DNA molecules, a property that makes them useful for genome engineering and control of gene expression. The extreme diversity of the crRNP targeting complexes is largely responsible for the variability observed in different CRISPR types. Whereas types I and III use multi-protein complexes, types II, V, and VI rely on a single protein for interference. Extensive studies have elucidated the mechanisms and structures of several complexes from most CRISPR types, revealing the commonality of target binding through crRNA base-pairing and high divergence in the machineries and modes of target cleavage. In-depth reviews focused solely on interference have been presented elsewhere (Garcia-Doval and Jinek, 2017; Plagens et al., 2015; Pyenson and Marraffini, 2017; Tsui and Li, 2015).

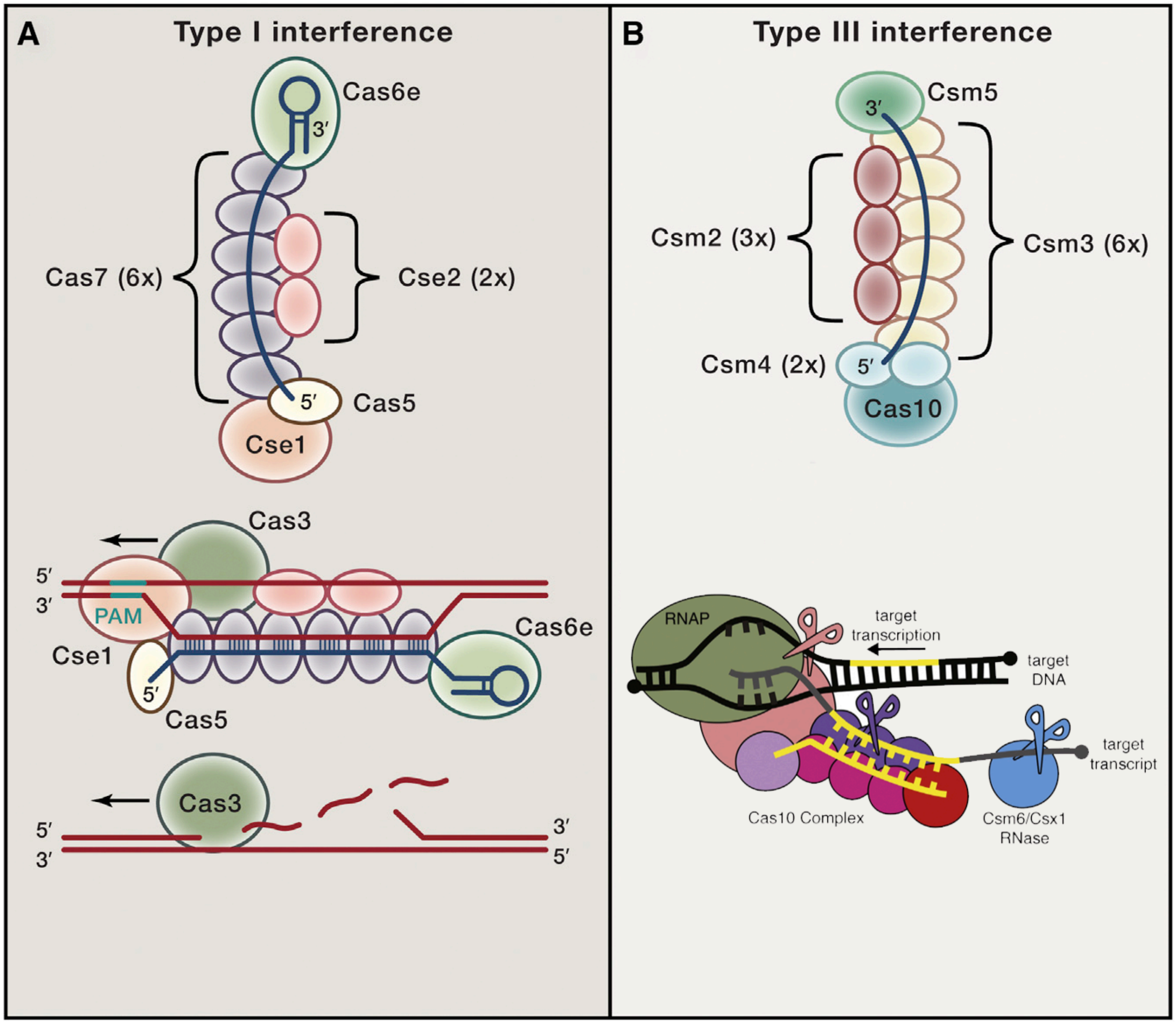


Figure 1.5 | Class 1 interference mechanisms. (A) Interference in type I systems is carried out by Cascade, which recognizes the target DNA, and Cas3, which degrades it. A type I-E complex is schematized here. (B) Interference in type III systems is carried out by the Csm or Cmr complex and involves cleavage of the target RNA by the Csm3 or Cmr4 subunits as well as nonspecific ssDNA and RNA cleavage by Cas10 and Csm6/Csx1. A type III-A complex is schematized here. Adapted in part from Pyenson and Marraffini, 2017.

1.6.1 Type I interference

In type I systems, the roles of target DNA recognition and degradation are segregated into two distinct components. The crRNA-guided Cascade complex binds and unwinds the DNA target sequence (Brouns et al., 2008), then recruits Cas3 to degrade the target in a processive manner through the combined action of its HD nuclease and helicase domains (Fig. 1.5a) (Makarova et al., 2011b; Mulepati and Bailey, 2013; Sinkunas et al., 2013; Westra et al., 2012). Each type I subtype (I-A through I-F) has a distinct complement of Cascade components and, in some cases, significant variation of the *cas3* gene (Makarova et al., 2011b).

The *E. coli* Cascade complex has served as the model system for understanding the mechanism of type I interference. In addition to the central 61 nt crRNA bearing the 32 nt spacer, the complex comprises five proteins in different stoichiometries: $(Cse1)_1$, $(Cse2)_2$, $(Cas5)_1$, $(Cas7)_6$, $(Cas6)_1$. The Cas7 subunits form the "backbone" that polymerizes along the crRNA and determines the crescent-shaped, semi-helical architecture seen in all structurally characterized Cascade complexes (Hochstrasser et al., 2014; Jackson et al., 2014; Jore et al., 2011; Mulepati et al., 2014; Wiedenheft et al., 2011a; Zhao et al., 2014). Cas6 (Cas6e in type I-E systems) remains bound to the 3′ hairpin following CRISPR maturation, while Cas5 binds the 5′ handle (Brouns et al., 2008; Jore et al., 2011). A "small subunit" (Cse2 in type I-E) is often found in two copies forming the "belly" of the structure and helps stabilize the crRNA and target DNA (Jackson et al., 2014; Mulepati et al., 2014; Zhao et al., 2014). A "large subunit" (Cse1 in type I-E, Cas8 in most other subtypes) binds at the 5′ end of the crRNA and recognizes the PAM sequences and recruits Cas3 to an authenticated target (Fig. 1.5a) (Hochstrasser et al., 2014; Sashital et al., 2012). While Cas6 does not always remain with the complex and the small subunit is often found as a fusion with the large subunit, the overall architecture of Cascade complexes in generally conserved (Makarova et al., 2011b; Plagens et al., 2012; Sokolowski et al., 2014).

Cascade pre-arranges the spacer segment of the crRNA in six five-base segments of pseudo A-form conformation, with the sixth base flipped out and bound by a Cas7 subunit (Jackson et al., 2014; Mulepati et al., 2014; Zhao et al., 2014). To initiate interference, Cascade first recognizes trinucleotide PAM sites in the target strand of foreign DNA through specific interactions with Cse1 (Sashital et al., 2012). Upon PAM binding, the DNA target is unwound starting at the PAM-proximal end of the protospacer to form an R loop structure (Hochstrasser et al., 2014; Rollins et al., 2015; Rutkauskas et al., 2015; Sashital et al., 2012; Szczelkun et al., 2014; van Erp et al., 2015). Each stretch of five exposed bases in the crRNA is free to bind the target DNA, leading to a stable but highly distorted and discontinuous crRNA:target strand duplex (Mulepati et al., 2014; Szczelkun et al., 2014). Cascade undergoes a conformational change upon target binding that enables recruitment of Cas3 to the Cse1 subunit (Hochstrasser et al., 2014; Mulepati et al., 2014). Cas3 binds and nicks the displaced strand using its catalytic center of the HD nuclease domain (Gong et al., 2014; Huo et al., 2014; Mulepati and Bailey, 2013; Sinkunas et al., 2013; Westra et al., 2012). The ATP-dependent helicase activity of Cas3 is then activated, causing metal- and ATP-dependent 3′→5′ translocation and processive degradation of the non-target strand (Gong et al., 2014; Huo et al., 2014; Westra et al., 2012). Cas3 initially degrades only 200-300 nt of the nontarget strand, though it continues translocating for many kilobases (Redding et al., 2015). Exposed ssDNA on the target strand may then become a substrate for other ssDNA nucleases or an additional Cas3 molecule to complete the degradation of foreign DNA (Mulepati and Bailey, 2013; Redding et al., 2015; Sinkunas et al., 2013). In addition to the PAM, target interference also relies on a seed region at the 3′ end of the spacer segment of the crRNA (Semenova et al., 2011; Wiedenheft et al., 2011b). Single point mutations of the seed region of the *E. coli* Cascade complex, at the 1 to 5 and 7 to 8 position of the spacer, is enough to decrease target DNA binding and subsequent interference (Semenova et al., 2011).

Differences in the *cas3* gene among type I subtypes suggests some variability in interference mechanism. In some type I-E species, Cas3 is fused to Cse1 by a linker that allows it to stably associate with the Cascade complex (Westra et al., 2012). In type I-A systems, the Cas3 helicase and nuclease domains exist as separate polypeptides that both associate with the Cascade complex (Plagens et al., 2014). In type I-F systems, Cas3 is fused to Cas2, lending further genetic support for the interaction between the interference and acquisition machinery during primed acquisition (Makarova et al., 2015; Richter and Fineran, 2013; Richter et al., 2012). How these fusions and domain separations affect the processive degradation observed in type I-E systems requires further study.

1.6.2 Type II interference

In contrast to the multi-subunit effector complexes seen in type I and type III systems (but similar to Cpf1 of type V systems), the type II signature protein Cas9 functions as an individual protein, along with a crRNA and tracrRNA, to interrogate DNA targets and destroy matching sequences by cleaving both strands of the target (Fig. 1.6) (Gasiunas et al., 2012; Jinek et al., 2012). Extensive studies on Cas9 have yielded a range of structures of *S. pyogenes* Cas9 in different substrate-bound states as well as structures of several orthologs (Anders et al., 2014; Jiang et al., 2015; Jinek et al., 2014; Nishimasu et al., 2015; 2014). Many of these structures, as well as the mechanism of Cas9 target search and recognition, are reviewed elsewhere (Jiang and Doudna, 2015; van der Oost et al., 2014); here we focus on recent advances.

Structures of Cas9 have revealed two distinct lobes, the nuclease lobe and the α-helical or REC lobe (Anders et al., 2014; Jinek et al., 2014; Nishimasu et al., 2014; 2015). The nuclease lobe is composed of the HNH nuclease domain, which cleaves the target strand, a RuvC-like nuclease domain, which cleaves the non-target strand and is separated into three distinct regions in the primary sequence by the intervening α-helical lobe and the HNH domain, and a C-terminal PAM-interacting domain (Anders et al., 2014; Jinek et al., 2014; Nishimasu et al., 2014; 2015). The α-helical lobe contains an arginine-rich “bridge helix”, which connects the two lobes and interacts with the guide RNA, and is the most variable region of Cas9, with insertions or deletions accounting for much of the wide variation in size seen in Cas9 orthologs (Chylinski et al., 2014; Jinek et al., 2014; Nishimasu et al., 2014).

Cas9 initiates its target search by probing duplexed DNA for an appropriate PAM before initiating target unwinding (Sternberg et al., 2014). The target unwinds from the seed region, the first 10-12 nucleotides following the PAM, toward the PAM-distal end (Szczelkun et al., 2014). A perfect or near-perfect match leads to cleavage of both DNA strands, with mismatches being more tolerated outside of the seed region (Cong et al., 2013; Jiang et al., 2013; Jinek et al., 2012; Sternberg et al., 2014). The mechanism by which mismatches distant from the cleavage site prevent cleavage appears to rely on the structural flexibility of the HNH domain, which has yet to be crystallized in proximity to the scissile phosphate (Anders et al., 2014; Nishimasu et al., 2014; 2015). FRET assays show that the HNH domain swings into a catalytically competent position only upon binding to a cognate double-stranded DNA substrate, underscoring the multiple steps of conformational control of Cas9-catalyzed DNA cleavage (Sternberg et al., 2015). The RuvC domain is in turn allosterically regulated by the HNH domain. Cleavage of the non-target strand requires movement of the HNH domain into an active position, even when mismatched substrates allow full unwinding of the non-target strand (Sternberg et al., 2015).

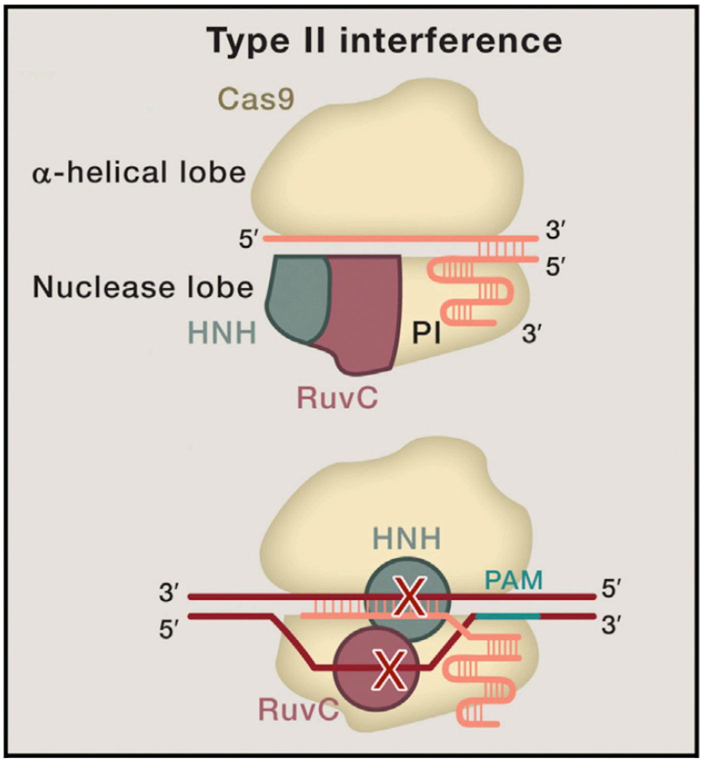


Figure 1.6 | Type II interference. In type II systems, Cas9 forms the effector complex with a crRNA and a tracrRNA. Cas9 is composed of a nuclease lobe and an α-helical lobe. The nuclease lobe contains both the HNH and RuvC-like nuclease domains as well as the PAM-interacting domain. The 3′ hairpins of the tracrRNA bind the nuclease lobe, while the stemloop and spacer line the channel between the two lobes. Binding to a matching PAM-adjacent target causes the HNH domain to move into position to cleave the annealed strand, while the displaced strand is fed into the RuvC active site for cleavage.

Recent crystal structures of *S. pyogenes* Cas9-sgRNA surveillance complex and of the smaller *Staphylococcus aureus* Cas9 in a target-bound state provided new insights into Cas9 function (Jiang et al., 2015; Nishimasu et al., 2015). The sgRNA-bound structure revealed how binding of sgRNA shifts Cas9 from the auto-inhibited state observed in the apo form to a conformation competent for target search (Jiang et al., 2015; Jinek et al., 2014). As previously observed in low-resolution electron microscopy structures, a nucleic acid binding cleft is formed between the two lobes upon sgRNA binding (Jinek et al., 2014). Furthermore, two PAM-interacting arginine residues are pre-positioned to allow for scanning of potential target DNA, a finding that may explain the necessity of tracrRNA in directing PAM-dependent spacer acquisition. Surprisingly, while the 3′ hairpins of the tracrRNA have been shown to provide nearly all of the binding energy and specificity for Cas9, the repeat-anti-repeat region of the sgRNA as well as the seed sequence were required to induce the conformational rearrangement (Briner et al., 2014; Jiang et al., 2015; Wright et al., 2015). The seed sequence of the sgRNA was also found to be pre-ordered in an A-form helix, analogous to the pre-ordered seed region of guide RNA observed in eukaryotic Argonaute structures and the type I and type III effector complexes, where the entire crRNA is pre-arranged in a target-binding-competent state (Jackson et al., 2014; Kuhn and Joshua-Tor, 2013; Mulepati et al., 2014; Osawa et al., 2015; Taylor et al., 2015; Zhao et al., 2014). The observed pre-ordering of the guide RNA provides an energetic compensation for the unwinding of the target duplex to facilitate binding.

Cas9 from the type II-C CRISPR system of *S. aureus* was crystallized in complex with sgRNA and a single-stranded DNA target sequence, providing insight into the structural variation between more distantly-related Cas9 (Nishimasu et al., 2015). *S. aureus* Cas9 is significantly smaller than the Cas9 of *S. pyogenes* (1,053 vs. 1,368 amino acids) and recognizes a significantly different guide RNA and PAM site. The *S. aureus* Cas9 structure revealed a smaller α-helical lobe, with domains in the middle and PAM-proximal side notably absent, while the nuclease lobe is largely conserved (Nishimasu et al., 2015). The authors proposed a new domain designation, the wedge domain, which diverges significantly between the two proteins and appears integral to determining guide RNA orthogonality. Another small Cas9, that from *Actinomyces naeslundi*, was previously crystallized in the apo form, but the absence of bound substrate and significant disordered regions limited detailed exploration of the differences between the orthologs (Jinek et al., 2014). Other recent work with type II-C Cas9 proteins from *N. meningitidis* and *Corynebacterium diphtheriae*, among other type II-C orthologs, revealed that these enzymes have a reduced ability to unwind dsDNA compared to *S. pyogenes* Cas9 and exhibit efficient PAM-independent and in some cases tracrRNA-independent cleavage of ssDNA (Ma et al., 2015; Zhang et al., 2015). This activity may allow for more efficient interference with ssDNA plasmid or phage or represent a more ancestral activity that predates the expansion of the α-helical lobe to facilitate more robust DNA unwinding.

1.6.3 Type III interference

Type III systems are classified into types III-A to III-D based on their effector complexes, with III-A and III-B being the best characterized (Makarova et al., 2015). The former is constituted by the Csm complex, and the latter by the Cmr complex (Makarova et al., 2011b). Phylogenetic studies suggested that some *csm* and *cmr* genes are distant homologs of *cas* genes that compose the Cascade complex of type I systems, and subsequent structural studies have revealed a striking structural conservation between Cascade and the Csm and Cmr complexes (Hochstrasser et al., 2014; Jackson et al., 2014; Makarova et al., 2013; Mulepati et al., 2014; Osawa et al., 2015; Staals et al., 2014; Taylor et al., 2015; Zhao et al., 2014). For a detailed discussion of the structural similarities between these complexes, refer to Jackson and Wiedenheft, 2015. Briefly, Csm3 (in III-A systems) or Cmr4 (in III-B) polymerizes along the crRNA as a helical backbone, analogously to Cas7, while Csm2 or Cmr5 take the role of Cse2 as the small subunit (Fig. 1.5b) (Jackson and Wiedenheft, 2015). Similar to Cascade, the crRNA is pre-arranged for binding with kinks every six nucleotides. The target RNA binds in a distorted manner, forming five-nucleotide helical stretches with the sixth base flipped out to allow for the extreme deviation from helical nucleic acid observed in all structures (Osawa et al., 2015; Taylor et al., 2015). Cmr3 and Csm4 bind the 5′ crRNA handle, while Cas10 (also referred to as Csm1 and Cmr2) serves as the large subunit (Makarova et al., 2011a; Osawa et al., 2015; Staals et al., 2014; Taylor et al., 2015). Csm5, Cmr6, and Cmr1 also share homology with Cas7 and cap the helical backbone at the 3′ end of the crRNA. In type III-B systems, two major crRNA species are generally observed, differing by six nucleotides (Juranek et al., 2012; Staals et al., 2014). Cryo-electron microscopy captured two Cmr complexes of different sizes, with one complex having one fewer Cmr4 and Cmr5 subunit, suggesting that the different crRNA lengths are the result of different complex sizes, or vice versa (Taylor et al., 2015).

Despite the structural similarities, the type III interference complexes function quite distinctly from Cascade. The substrate specificity of Csm and Cmr complexes has only recently been clarified. Early *in vivo* genetic experiments suggested Csm targeted DNA, while *in vitro* studies of Cmr showed binding and cleavage activity against RNA only (Hale et al., 2009; Marraffini and Sontheimer, 2008), leading to a model wherein the two subtypes had evolved distinct and complementary substrate preferences. This simple model was soon complicated by the observation that Csm complexes *in vitro* also bind and cleave RNA while exhibiting no activity against DNA (Staals et al., 2014; Tamulaitis et al., 2014). Meanwhile, the *in vivo* DNA-targeting activity of III-A systems was shown to depend on transcription at the target site, in contrast to the transcription-independent targeting seen in type I and type II systems, and a similar activity was observed for a III-B system *in vivo* (Deng et al., 2013; Goldberg et al., 2014). These observations were reconciled by the discovery that the Csm complex from *Staphylococcus epidermidis* exhibits both RNA cleavage and DNA cleavage when directed against the non-template strand of actively-transcribed DNA (Samai et al., 2015). Subsequent studies with additional III-A and III-B systems revealed a conserved mechanism whereby binding of a target RNA activates a nonspecific ssDNase activity until the target RNA is cleaved (Elmore et al., 2016; Estrella et al., 2016; Han et al., 2016; Kazlauskiene et al., 2016; Liu et al., 2017c).

DNA and RNA interference are carried out by distinct subunits of the type III complexes. RNA interference is mediated by the backbone subunit Csm3 (or Cmr4 in III-B systems), which cleaves the target every 6 nucleotides in the active site of a separate subunit by activating the ribose 2′ OH for nucleophilic attack in a manner typical of metal-independent RNases (Osawa et al., 2015; Staals et al., 2014; Tamulaitis et al., 2014; Taylor et al., 2015). Cas10 cleaves ssDNA, possibly in transcription bubbles, and different groups have observed a dependence on either the protein's HD nuclease domain or palm polymerase domain for cleavage (Elmore et al., 2016; Estrella et al., 2016; Han et al., 2016; Kazlauskiene et al., 2017; Liu et al., 2017c; Samai et al., 2015). More recent work has revealed that type III systems have a nonspecific RNase activity that is also activated by target RNA binding (Han et al., 2017; Kazlauskiene et al., 2017; Niewoehner et al., 1AD). Upon target binding, Cas10 synthesizes cyclic oligoadenylates from ATP, which then serve as second messengers to activate Csm6 or Csx1, a nonspecific RNase that does not physically associate with the Csm or Cmr complex. This activity allows the host cell to degrade accumulated viral transcripts and is essential when the CRISPR array only targets late-expressed phage genes (Jiang et al., 2016).

The distinct behavior of type III systems provides the host microbe with the ability to tolerate temperate phages (Goldberg et al., 2014). While type I and type II systems target and degrade any protospacer-containing DNA type III systems ignore foreign DNA until transcription begins that poses a threat to the cell. This has the advantage of allowing cells to acquire advantageous genes contained in prophages, such as antibiotic resistance genes, and causing cell suicide in the event that a lysogenic phage becomes lytic and begins transcribing genes with matching spacers (Goldberg et al., 2014). However, the strand-specific nature of both the RNA targeting and transcription-dependent DNA targeting imposes an additional restriction on the integration step of acquisition, as only one direction of integration will yield productive interference. The means by which this apparent limitation is overcome are unclear. Type III systems are also frequently found coexisting with type I systems, and they have been demonstrated to function in interfering against viruses that have escaped from type I interference (Makarova et al., 2011b; Silas et al., 2017).

Type III systems are also unusual in their lack of a PAM. Rather than recognizing a distinct motif to avoid auto-immunity at the CRISPR locus, the Csm and Cmr complexes instead check for complementarity between the repeat-derived region of the crRNA with the target and do not cleave if a full match is detected (Marraffini and Sontheimer, 2010; Samai et al., 2015; Staals et al., 2014; Tamulaitis et al., 2014). The specificity of type III effector complexes for single-stranded targets might provide a rationale for their distinct mode of target authentication. For type I and type II effector complexes, which target dsDNA, PAM recognition allows for an initial binding event to facilitate subsequent unwinding of the target to probe for complementarity to the crRNA (Hochstrasser et al., 2014; Rollins et al., 2015; Sternberg et al., 2014; Szczelkun et al., 2014; Westra et al., 2012). Type III complexes can immediately probe a potential single-stranded target for complementarity to their bound crRNA without a need to license initial unwinding, and the exposed nature of a single-stranded target facilitates the check for complementarity to the repeat-derived region of the guide.

1.6.4 Type V interference

Type V systems have been identified more recently, but initial work demonstrated that these systems are functional for interference (Makarova et al., 2015; Zetsche et al., 2015a). The systems appear most similar to type II systems, possessing only the acquisition machinery and a single additional protein (Makarova et al., 2015; Vestergaard et al., 2014). Five subtypes of Class V systems have been identified with widely varying interference proteins (Burstein et al., 2017; Koonin et al., 2017; Shmakov et al., 2015; 2017). Type V-A through V-E are characterized by the presence of Cas12a to Cas12e, formerly known as Cpf1, C2c1, C2c3, CasY, and CasX respectively (Koonin et al., 2017). All five proteins are evolved from the same family of transposon-associated TpnB proteins as Cas9 and have a C-terminal RuvC (Koonin et al., 2017; Shmakov et al., 2015). However, the proteins show little similarity to each other, and the phylogenetic grouping of the associated *cas1* genes with various branches of type I and type III *cas1* genes suggests that each of these subtypes originated from distinct recombination events between CRISPR systems and *tpnB* genes (Koonin et al., 2017; Shmakov et al., 2015).

Cas12a is the best-studied of the type V interference proteins, but due to the diversity of these proteins it is unclear whether its mechanism of action is broadly conserved. The studied Cas12 proteins, Cas12a included, target DNA and recognize a T-rich 5′ PAM sequence (Burstein et al., 2017; Zetsche et al., 2015a). The single RuvC domain of Cas12a cleaves both strands of the target dsDNA, resulting in a several-nucleotide 5′ nucleotide distal to the PAM site (Swarts et al., 2017; Zetsche et al., 2015a). The RuvC can also carry out nonspecific ssDNA cleavage when Cas12a is activated by a complementary target DNA (Chen et al., 2018; Gootenberg et al., 2018). Structural studies of Cas12a and Cas12b revealed that the proteins, like Cas9, adopt a bilobed structure with the guide RNA and target DNA binding across the central channel

(De Dong et al., 2016; Gao et al., 2016; Liu et al., 2016; Stella et al., 2017; Swarts et al., 2017; Wu et al., 2017; Yamano et al., 2016; 2017; Yang et al., 2016). How the single RuvC nuclease domain of Cas12a cuts both strands of the target DNA and ssDNA in *trans*, as well as whether this activity is conserved across the divergent Cas12 proteins, remains to be seen.

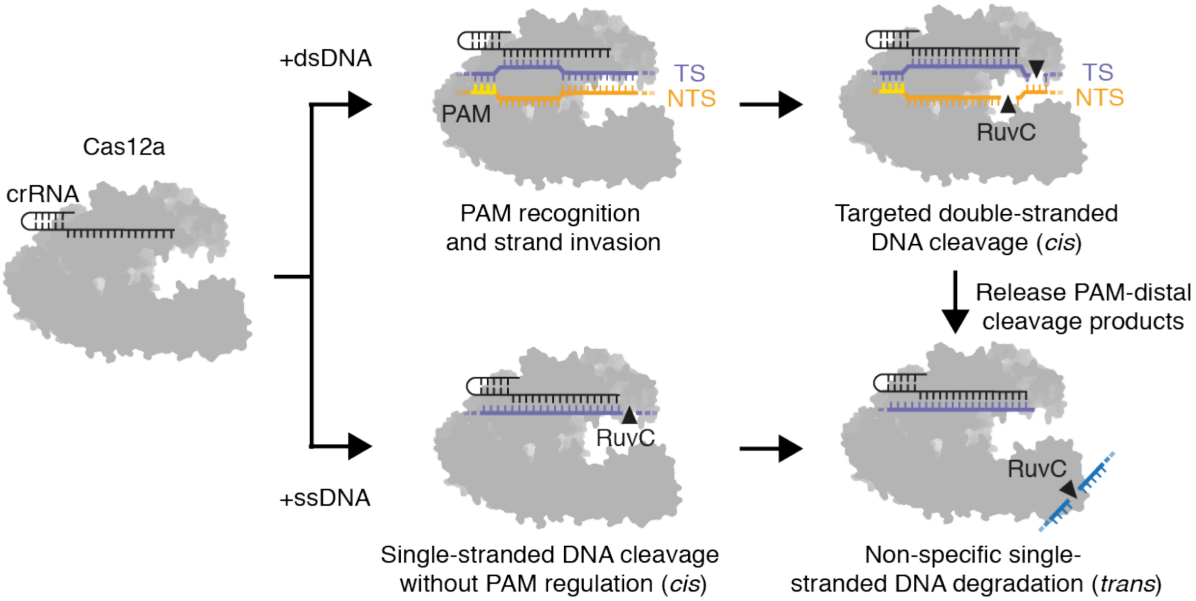

Cas12a
crRNA
+dsDNA
+ssDNA
TS
NTS
PAM
PAM recognition and strand invasion
TS
NTS
RuvC
Targeted double-stranded DNA cleavage (*cis*)
Release PAM-distal cleavage products
RuvC
Single-stranded DNA cleavage without PAM regulation (*cis*)
RuvC
Non-specific single-stranded DNA degradation (*trans*)

Figure 1.7 | Type V interference. Cas12a forms an effector complex with a crRNA. Binding of either a double-stranded or single-stranded complementary target activates a single RuvC domain for cleavage, resulting in cleavage of both the target DNA and non-specific single-stranded DNA. Adopted from Chen et al., 2018.

1.6.5 Type VI interference

Type VI systems are characterized by the presence of Cas13 (formerly C2c2) and are unique among CRISPR types in that they appear to target RNA exclusively (Koonin et al., 2017). Type VI systems are grouped into four subtypes based on the Cas13 sequence and the presence of auxiliary factors with apparent roles in regulating Cas13 activity (Koonin et al., 2017; Yan et al., 2018). Here we will discuss the conserved aspects of the systems. Cas13 has two HEPN (higher eukaryotes and prokaryotes nucleotide-binding domain) domains that together form a single RNase active site, as well as a separate active site responsible for processing the pre-crRNA (East-Seletsky et al., 2016; Knott et al., 2017; Liu et al., 2017a; 2017b; Shmakov et al., 2015). When a complementary RNA target is bound, the compound HEPN active site, which is distal from the target RNA binding cleft, is activated as a nonspecific RNase, resulting in cleavage of both the target RNA (*cis* cleavage) and any other RNA as well (*trans* cleavage) (Fig. 1.8) (Abudayyeh et al., 2016; East-Seletsky et al., 2016). In some circumstances, Cas13 has been used to accomplish targeted knockdown of a specific transcript *in vivo* without evidence of *trans* cleavage (Abudayyeh et al., 2016; 2017). The mechanism by which *trans* cleavage is suppressed *in vivo* and what extent the differential activity is ortholog-dependent requires further characterization.

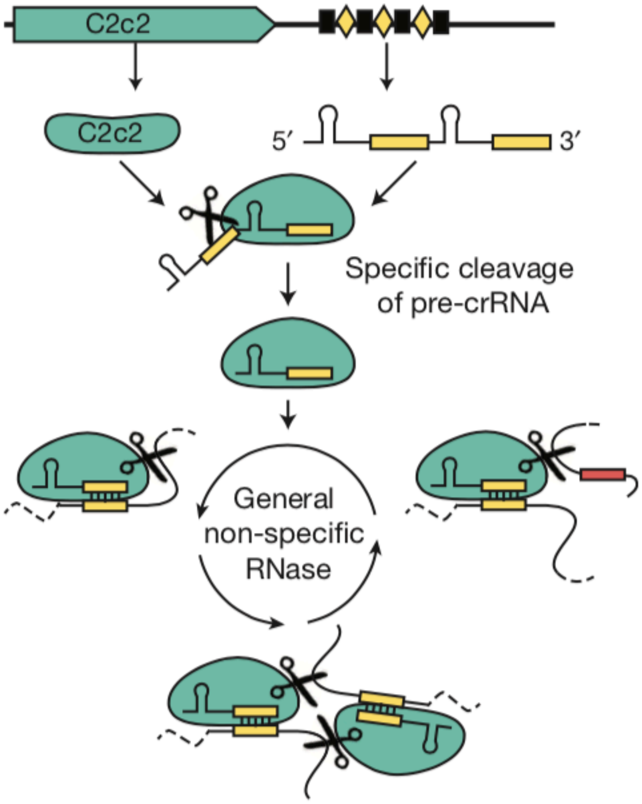


Figure 1.8 | Type VI interference. Cas13 (C2c2) processes its own pre-crRNA into a functional crRNA. Binding a complementary RNA activates a non-specific RNase activity that can cleave the bound target RNA as well as other RNA molecules in *trans*. Adapted from East-Seletsky et al., 2016.

1.6.6 Interference complexes as genome editing tools

Most tool development of Cas proteins has focused on exploiting the programmable sequence-specific DNA recognition of interference complexes. Cas9 from *S. pyogenes* in particular has proven enormously useful for genome engineering. The ability to render Cas9 a two-component system by fusing the crRNA and tracrRNA into a single guide RNA (sgRNA) has allowed for its easy use for genome editing, transcriptional control, RNA targeting, and imaging (Jiang and Marraffini, 2015; Sternberg and Doudna, 2015). Cas9 has been used in various cell types and organisms ranging from mice and monkeys to primary human T cells and stem cells as well as plants, bacteria, and fungi. Recent work has focused on developing various chemical- and light-inducible Cas9 constructs to allow for greater spatiotemporal control and on employing Cas9 orthologs with different PAM sequences and smaller sizes, allowing for easier packaging in adeno-associated virus vectors (Davis et al., 2015; Nihongaki et al., 2015; Polstein and Gersbach, 2015; Ran et al., 2015; Zetsche et al., 2015b).

Other interference complexes have already been used or have the potential to be useful for genome manipulation as well. Although the multi-subunit composition of Cascade make it less tractable for genome engineering compared to Cas9, its large size and stable binding has been used for transcriptional silencing in *E. coli* (Rath et al., 2015). No published work has shown the application of Csm or Cmr complexes, but either could likely be used for various RNA modulation applications in cells. Cas12a has been adopted for genome editing alongside Cas9, where its AT-rich PAM and ready ability to utilize multiplexed guides due to its intrinsic guide-processing ability provide situational advantages (Swarts and Jinek, 2018). Cas13 has been developed for a separate set of applications based on its RNA-targeting abilities, including gene knockdown, RNA editing, and regulation of splicing (Abudayyeh et al., 2017; Cox et al., 2017; Konermann et al., 2018). The *trans* activity of both Cas12a and Cas13 has been exploited *in vitro* for highly sensitive detection of specific DNA or RNA sequences, using cleavage of a reporter nucleic acid with both a fluorescent dye and a quencher to generate signal (Chen et al., 2018; East-Seletsky et al., 2016; Gootenberg et al., 2018; 2017).

While Cas9 has already seen extensive use in the research setting, challenges remain for its application in the clinic. While making programmed cuts has become largely trivial, biasing DNA repair toward homology-directed repair rather than non-homologous end joining remains a challenge (Chu et al., 2015; Maruyama et al., 2015). Delivery of Cas9, either as an RNP or on a plasmid or viral vector, to particular tissues in whole organisms is another challenge that must be addressed to enable clinical applications (D'Astolfo et al., 2015; Gori et al., 2015; Howes and Schofield, 2015; Lin et al., 2014; Zuris et al., 2015). As the field continues to advance rapidly, clinical trials may occur within a few years, with therapies possibly following within a decade. Engineering of crop plants with Cas9 is already underway; regulatory rulings have so far considered knockout plants not to be genetically modified organisms, but the regulatory fate of other modifications is currently being considered (Servick, 2015).

1.7 Concluding remarks

Despite the rapid progress of the field since the first demonstration of CRISPR immunity in 2007, many mechanistic questions remain unanswered. Fundamental aspects of acquisition, such as how substrates for Cas1-Cas2-mediated integration are generated and the mechanism and extent of self vs. non-self discrimination in different CRISPR subtypes, are still a mystery. While crRNA biogenesis and interference are reasonably well understood for certain model subtypes (type I-E, type II-A), the sheer diversity of CRISPR systems means that many subtypes with potentially distinct mechanisms remain unexamined or undiscovered. The diversity of type V and type VI are still being explored, and type IV systems, bearing some familiar *cas* genes but no identifiable CRISPR locus, have yet to be characterized experimentally and almost certainly rely on mechanisms distinct from those of traditional CRISPR systems (Makarova and Koonin, 2015).

Other aspects of CRISPR-Cas systems lie beyond the scope of this work. We have not discussed the non-immune functions of CRISPR-Cas systems, some of which appear to have evolved to serve regulatory rather than defense roles (for reviews, see Westra et al., 2014, and Ratner et al., 2015). Phage evasion of CRISPR immunity is another active area of research, with identified mechanisms including DNA modification, specialized anti-CRISPR proteins, and mutational escape (Bondy-Denomy et al., 2013; Bondy-Denomy et al., 2015; Bryson et al., 2015; Deveau et al., 2008; Paez-Espino et al., 2015; Pawluk et al., 2014). The context-dependent regulation of CRISPR-Cas systems in response to phage infection and stress signals has also been explored but requires further study (Bondy-Denomy and Davidson, 2014; Garrett et al., 2015; Kenchappa et al., 2013; Patterson et al., 2015; Pul et al., 2010). The rapid development of technology derived from CRISPR-Cas systems, most notably Cas9 but also Cas6f/Csy4, Cascade and Cpf1, has fueled intense interest in the field. The arms race between bacteria and bacteriophage has generated powerful molecular biology tools, from restriction enzymes that enabled recombinant DNA technology to Cas9, which started the "CRISPR revolution" in modern genome engineering. CRISPR systems haven proven to be both fascinating and enormously useful. Further study of bacterial immune systems, both CRISPR systems and those yet undiscovered, promises to yield further unforeseen discoveries and exciting new technologies.

1.8 Acknowledgments

This work was funded by US National Science Foundation grant number 1244557 to J.A.D. A.V.W. and J.K.N. are NSF Graduate Research Fellows. Megan Hochstrasser provided valuable input on the manuscript.

CHAPTER 2

Protecting genome integrity during CRISPR immune adaptation

A portion of the work presented in this chapter has been published previously as part of the following paper: Wright, A.V. & Doudna, J.A. (2016). Protecting genome integrity during CRISPR immune adaptation. *Nat. Struct. Mol. Biol.* **23** (10) 876-883.

Addison Wright performed all experiments. Addison Wright and Jennifer Doudna designed experiments, analyzed data, and wrote the manuscript.

2.1 Abstract

Bacterial CRISPR-Cas systems include genomic arrays of short repeats flanking foreign DNA sequences that provide adaptive immunity against viruses. Integration of foreign DNA must occur specifically to avoid damaging the genome or the CRISPR array, but surprisingly promiscuous activity occurs *in vitro*. Here we reconstitute full-site DNA integration and show that the *Streptococcus pyogenes* type II-A Cas1-Cas2 integrase maintains specificity in part through limitations on the second integration step. At non-CRISPR sites, integration stalls at the half-site intermediate, enabling reaction reversal. *S. pyogenes* Cas1-Cas2 is highly specific for the leader-proximal repeat and recognizes the repeat's palindromic ends, fitting a model of independent recognition by distal Cas1 active sites. These findings show how CRISPR systems maintain host genome integrity and suggest that DNA insertion sites will be less common than previous work suggests, preventing toxicity during CRISPR immune adaptation.

2.2 Introduction

CRISPR-Cas (Clustered Regularly Interspaced Short Palindromic Repeats-CRISPR-associated) systems provide bacteria and archaea with adaptive immunity against foreign genetic elements (Barrangou et al., 2007). These systems capture small pieces of foreign DNA, called protospacers, and integrate them at a CRISPR locus in a process called acquisition (Sternberg et al., 2016; Wright et al., 2016). The CRISPR locus is composed of a series of direct repeats, 20-50 nt long, separated by viral- and plasmid-derived spacers of similar length, along with an upstream leader sequence that contains the promoter for the locus (Bolotin et al., 2005; Mojica et al., 2005; Pourcel et al., 2005). The locus is transcribed and processed into CRISPR RNAs (crRNAs), which assemble with Cas proteins to identify and degrade complementary nucleic acids (Brouns et al., 2008).

Cas proteins and the mechanisms by which they carry out immunity vary widely, leading to the classification of CRISPR systems into six types (Makarova et al., 2015; Shmakov et al., 2015). Cas1 and Cas2 are the only proteins conserved across all identified active CRISPR systems, and are necessary and in some systems sufficient for acquisition (Makarova et al., 2015; Yosef et al., 2012). They catalyze integration of new protospacers at the leader-proximal repeat via an integrase-like mechanism (Nuñez et al., 2014; 2015b). Full integration requires cleavage-ligation reactions to take place at either end of the repeat, with each 3′ end of the protospacer making a nucleophilic attack on the backbone of the target DNA (Arslan et al., 2014; Nuñez et al., 2015b). The resulting gapped product is repaired by DNA repair factors to yield a new spacer and a duplicated repeat (Ivančić-Baće et al., 2015).

The Cas1-Cas2 integrase, in contrast to other transposases and integrases, has high specificity for the target (genomic) DNA sequence but little preference for the sequence of the integrated DNA sequence fragment (Nuñez et al., 2015b). Precision in target recognition is required for maintenance of the CRISPR locus as well as to avoid disruption of other genes by off-target integration. Traditional transposases avoid genome instability by maintaining transposition at a low level, targeting untranscribed regions, or using additional factors to direct integration to a defined attachment site (Bainton et al., 1993; Craigie and Bushman, 2012; Goryshin and Reznikoff, 1998; Wang and Higgins, 1994). Cas1-Cas2, in contrast, largely rely on intrinsic sequence specificity while maintaining sufficient activity to respond robustly to invasion. *In vivo* work with several CRISPR systems has revealed that mutations in the leader or repeat effectively abolish acquisition (Wang et al., 2016; Wei et al., 2015a; Yosef et al., 2012). However, while *in vitro* studies with Cas1-Cas2 from type I systems have demonstrated a preference for the leader-repeat sequence, other target sites also appeared to be tolerated, raising the question of how fidelity is maintained *in vivo* (Nuñez et al., 2015b; Rollie et al., 2015). Recent work has shown that IHF binding in the leader region increases specificity for *E. coli* Cas1-Cas2, but nonspecific integration still occurs in the absence of IHF, and it is unclear whether similar mechanisms exist in other CRISPR systems (Nuñez et al., 2016).

We set out to determine how Cas1-Cas2 from a type II system recognize the CRISPR locus and avoid off-target integration. Given that *in vitro* assays have so far only investigated half-site products, where only one end of the protospacer is integrated, we hypothesized that the full-site reaction might have greater substrate specificity. We show that the *S. pyogenes* type II Cas1-Cas2 integrase has high sequence specificity and catalyzes full-site integration into short targets. A broad range of substrates supported half-site integration, but only proper targets and protospacers allowed for full-site integration. We also identified a set of sequence requirements distinct from those observed in type I systems, suggesting different recognition modes for different Cas1 families. These findings imply that Cas1-Cas2 might sample potential sites with partial integration *in vivo*, but only catalyzes the complete reaction at proper integration sites.

2.3 Results

2.3.1 Sequence specificity of a type II CRISPR integrase

CRISPR systems contain variable sets of protein-coding genes that enable adaptive RNA-guided DNA binding and cleavage. Of the different types of CRISPR systems, type II pathways have among the smallest number of required proteins, hinting at possible functional and mechanistic differences with other CRISPR types (Makarova et al., 2015). To test the integration activity of a type II Cas1-Cas2, we expressed and purified Cas1 and Cas2 from the *Streptococcus pyogenes* type II-A CRISPR system. We tested the proteins for integration activity with plasmid substrates containing either no CRISPR locus (pUC19), the CRISPR locus from the *Escherichia coli* type I system (pEcoCR), or the native *S. pyogenes* CRISPR locus (pSpyCR).

We monitored protospacer integration by plasmid topology – successful integration generates open-circle plasmid, while integration followed by disintegration is believed to generate partially relaxed topoisomers, which migrate ahead of the input plasmid on ethidium bromide-stained gels (Fig. 2.1a) (Nuñez et al., 2015b). As previously observed, *E. coli* Cas1-Cas2 (EcoCas1-Cas2) exhibited robust activity with plasmids regardless of the presence of its CRISPR locus, converting most of the supercoiled plasmid into open-circle or topoisomerized plasmid (Fig. 2.1b) (Nuñez et al., 2015b). *S. pyogenes* Cas1-Cas2 (SpyCas1-Cas2), however, only integrated when the target contained the *S. pyogenes* CRISPR locus. Incubation of SpyCas1-Cas2 with non-target plasmids generated only topoisomers, visible as a faint band ahead of the supercoiled plasmid, indicating abortive integration. This suggests that SpyCas1-Cas2 integrates into potential targets but quickly reverses the reaction if the target sequence is incorrect. We also observed that Csn2, another Cas protein in the *S. pyogenes* CRISPR system known to be involved in acquisition, was not required for integration and actually reduced the generation of integration products for pSpyCR and topoisomers for the off-target plasmids, virtually eliminating activity against pUC19 (Heler et al., 2015; Wei et al., 2015b). The reduction in activity may stem from the sequestration of protospacers by Csn2, which binds linear DNA (Arslan et al., 2013). Cas9 is also required for acquisition *in vivo*, but the dispensability of both Cas9 and Csn2 for *in vitro* integration suggests that they are involved in a different step of acquisition, such as the generation of protospacers (Heler et al., 2015; Wei et al., 2015b).

We next tested integration activity using radiolabeled protospacers, which allow for more direct visualization of stable integration products (Nuñez et al., 2015b). While we observed some integration into off-target plasmids, integration was much more efficient into pSpyCR (Fig. 2.1c). SpyCas1-Cas2 also integrated DNA into a linearized plasmid at levels comparable to supercoiled plasmid, in contrast to previous observations that the *E. coli* proteins require bent or supercoiled target DNA (Nuñez et al., 2016). These results suggest that, within the minimal reconstitution system presented here, the SpyCas1-Cas2 integrase is more efficient at avoiding off-target integration than EcoCas1-Cas2 (Nuñez et al., 2015b; Rollie et al., 2015).

2.3.2 SpyCas1-Cas2 integrates at the leader-proximal repeat

To test whether the integration observed with SpyCas1-Cas2 into pSpyCR occurred specifically at the CRISPR locus, we performed high-throughput sequencing of the integration products of SpyCas1-Cas2 and pUC19, pEcoCR, and pSpyCR and mapped integration sites (Nuñez et al., 2015b; 2016). When we incubated SpyCas1-Cas2 with protospacers and either of the off-target plasmids, integration occurred at low levels across the plasmid sequence (Fig. 2.2a,b). Protospacer-containing reads exhibited a similar profile to the protospacer-free reads, with the exception of a peak downstream of the origin on both plasmids (Fig. 2.2a,b, 2.3a-d). When we used pSpyCR as the target, the integration sites mapped predominantly to the plus strand of the leader-repeat junction and the minus strand of the first repeat-spacer junction (Fig. 2.2c, 2.3e,f). These integration sites correspond precisely to the sites expected for a leader-proximal integration event and made up 60% of all mapped integration events. Many of the remaining integration peaks occurred at either end of the downstream repeats, suggesting a low level of leader-independent integration (Fig. 2.2d). Previous sequencing experiments with EcoCas1-Cas2 under the same conditions showed a weaker preference for the cognate leader-proximal repeat as well as notable integration hotspots in pUC19 (Nuñez et al., 2015b). The SpyCas1-Cas2 integration profile more closely resembles the activity of EcoCas1-Cas2 in the presence of IHF (Nuñez et al., 2016). These results support the hypothesis that SpyCas1-Cas2 relies on intrinsic sequence specificity to ensure accurate integration of new spacers, in contrast to the *E. coli* proteins, which require an additional host factor for faithful acquisition.

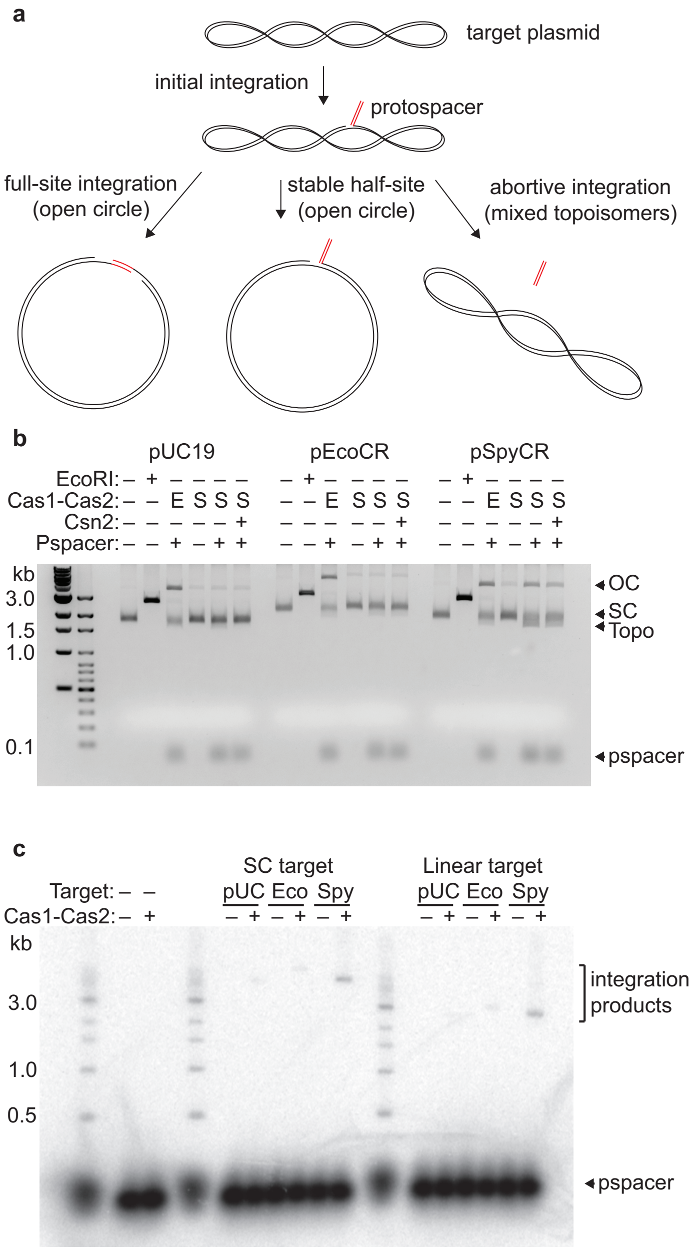
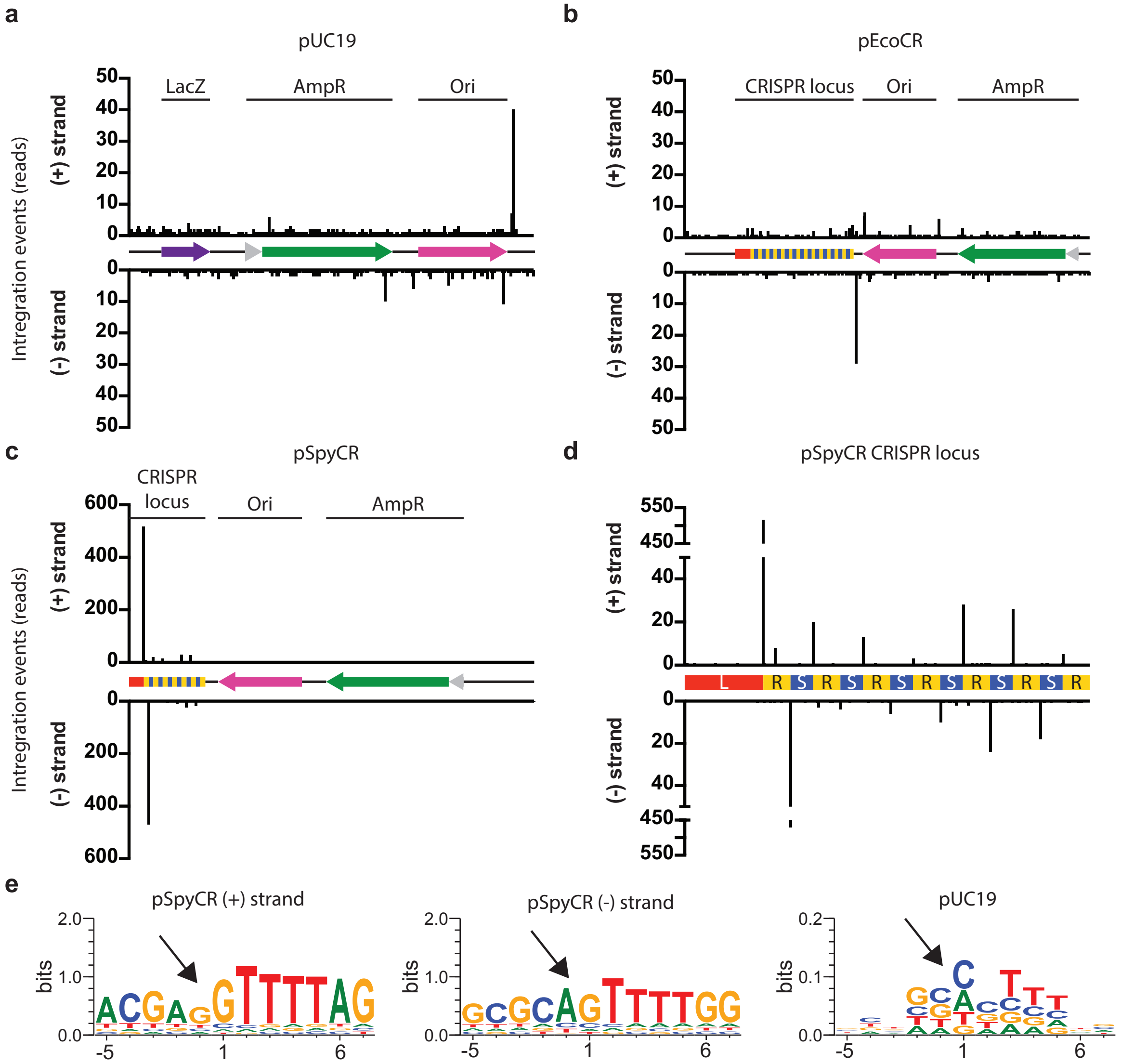


Figure 2.1 | *S. pyogenes* Cas1-Cas2 integrates only into plasmids with the target locus. (**a**) Schematic of Cas1-Cas2 integration into a supercoiled plasmid. Integration of a protospacer can yield three distinct products. (**b**) Integration assay with unlabeled protospacer. Cas1-Cas2 from *E. coli* (E) or *S. pyogenes* (S) were used. EcoRI-digested plasmid was used as a linear standard. Open-circle (OC), supercoiled, and topoisomer (Topo) products, as well as free protospacer (pspacer) are indicated. (**c**) Integration assay with radiolabeled protospacer and supercoiled (SC) or NdeI-digested (linear) targets.

Alignments of the sequences flanking the integration sites on either the plus or minus strand of pSpyCR suggest a strong preference for the ends of the repeat sequence (Fig. 2.2e). Unsurprisingly, the leader sequence is over-represented in the plus strand reads, but the consensus sequence shows a stronger bias for the repeat itself. The ends of the *S. pyogenes* repeat form a palindrome, which is readily apparent in the strong similarity between the consensus integration sites on the two strands, suggesting that Cas1 might recognize the same sequence at either end of the repeat. The integration sites in pUC19 show a much weaker sequence bias, with a slight preference for Ts in the +3 to +5 position and for C in the +1 position. The most over-represented integration site in both pUC19 and pEcoCR has the sequence 5′-CTTTTG-3′, and the +2 to +5 positions of the other highly-represented sites in pUC19 and pEcoCR are also unusually T-rich, suggesting that this motif may play a role in directing SpyCas1-Cas2 to the CRISPR repeat (Table 2.1). The significance of the preference for a C at the integration site is unclear, and it might suggest that the repeat is not necessarily the most strongly-preferred recognition sequence.



5’-ACGAGGTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAACTGCGC-3’
3’-TGCTCCAAAATCTCGATACGACAAAACTTACCAGGGTTTTGACGCG-5’

Figure 2.2 | SpyCas1-Cas2 integration is highly specific for the leader-proximal repeat. (**a-c**) Integration sites in pUC19 (**a**), pEcoCR (**b**), or pSpyCR (**c**). Note difference in scale for (**c**). Source data is deposited online. (**d**) Magnified view of the pSpyCR CRISPR locus. (**e**) WebLogos for integration sites in pSpyCR and pUC19. Arrows indicate the site and direction of integration. The sequence of the repeat and part of the leader and first spacer are shown below. The leader is red, repeat is yellow, and spacer is blue. Integration sites are indicated with arrows.

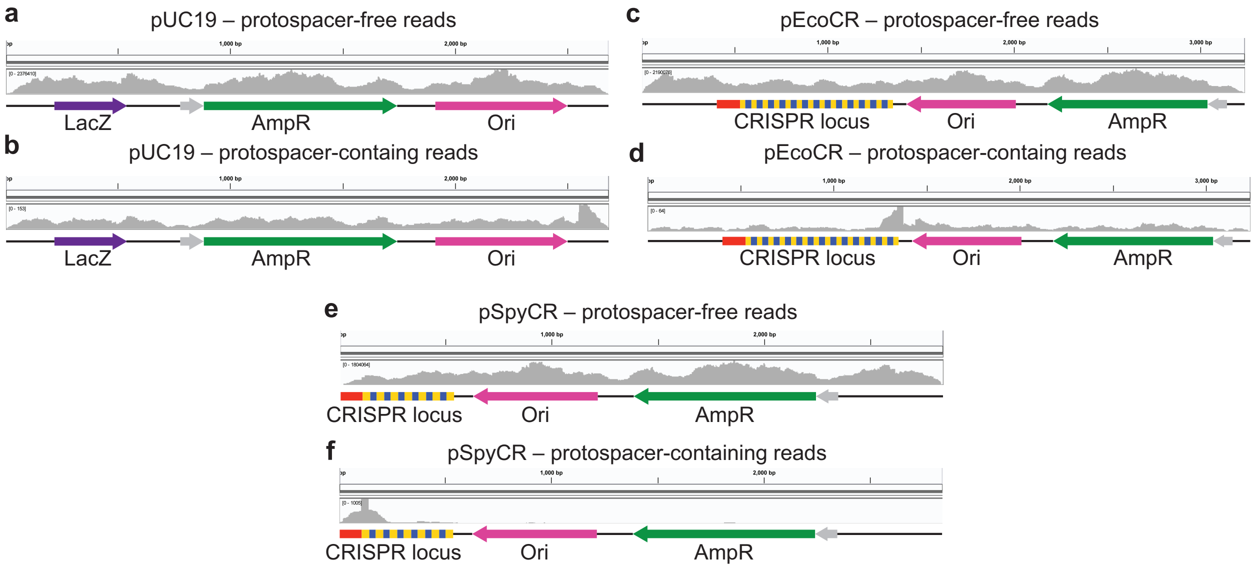


Figure 2.3 | High-throughput sequencing of SpyCas1-Cas2 integration shows preference for leader-proximal repeat. (**a**-**f**) Mapped reads from the integration reaction into pUC19 (**a**,**b**), pEcoCR (**c**,**d**), and pSpyCR (**e**,**f**). Plasmid features are shown below pile-up.

pUC19

Position	Strand	Reads	Sequence
2551	+	40	5′-CTTTTGC-3′
2552	+	29	5′-TTTTGCT-3′
2487	–	11	5′-CCCCTGA-3′
1702	–	10	5′-CTATTTC-3′
2540	+	7	5′-CTTTTGC-3′
932	+	6	5′-ATTTTGC-3′
1890	+	6	5′-CCCGTAG-3′

pEcoCR

Position	Strand	Reads	Sequence
1375	–	29	5′-CTTTTGC-3′
1374	–	16	5′-TTTTGCT-3′
1440	+	8	5′-CCTGACG-3′
1439	+	7	5’-CCCTGAC-3′
2036	+	6	5′-TCTGACG-3′

Table 2.1 | Highly represented sequences in non-target plasmid integration sites. Integration sites with more than 5 reads are shown for pUC19 and pEcoCR. The sequence of the ligated stand from the +1 to +7 position is shown.

2.3.3 SpyCas1-Cas2 integrates into short linear targets

Taking advantage of the increased sequence-specificity of SpyCas1-Cas2, we investigated whether the proteins could integrate site-specifically into short oligonucleotide targets. Previous work showed that EcoCas1-Cas2 can accurately integrate into such targets only in the presence of IHF (Nuñez et al., 2016). Based on the minimal leader sequence identified *in vivo* for the *Streptococcus thermophilus* type II system, we designed targets with 11 nucleotides of the leader, the full 36-nt repeat, and a 27-nt spacer (Fig. 2.4a) (Wei et al., 2015a). Incubation of SpyCas1-Cas2 with radiolabeled protospacers generated ligation products corresponding precisely to integration at either end of the repeat, confirming that the intrinsic sequence specificity of Cas1-Cas2 is sufficient for faithful integration (Fig. 2.4b,c). Integration required both Cas1 and Cas2, and a mutation in the predicted active site of Cas1 (H205A) ablated activity (Nuñez et al., 2014; 2015b). Integration also required a divalent cation, with magnesium providing the most robust activity, and a 3′ OH on the strand being integrated (Fig 2.5). The apparent preference for magnesium contrasts with EcoCas1-Cas2, which exhibits greater activity with manganese (Nuñez et al., 2015b). The crystal structure of *S. pyogenes* Cas1 revealed that the expected metal-binding residues were positioned farther apart than in other Cas1 structures, which may affect the metal preference (Ka et al., 2016).

To confirm that the integration sites had the expected strand specificity, we incubated protospacer-bound Cas1-Cas2 with targets with only one strand labeled. Integration into the plus-strand-labeled target produced a fragment matching the length of the leader, while the minus-strand-labeled target produced a spacer-sized fragment (Fig. 2.4d). Low levels of cleavage also occurred in the absence of protospacer, likely indicating that Cas1-Cas2 can use either water or a 3′ OH from another target as a nucleophile. This cleavage was restricted to the expected integration sites, confirming it results from sequence-specific Cas1-Cas2 activity. Based on the observation that EcoCas1-Cas2 prefers protospacers with 5-nt 3′ overhangs, we tested protospacers with 3′ overhangs ranging from 0 to 7 nt and found that 4-nt overhangs allowed for a modest increase in the integration rate at the leader end and a 2-fold increase in integration at the spacer end after 10 minutes relative to blunt ends (Fig. 2.6) (Nuñez et al., 2015a; Wang et al., 2015). Interestingly, single-nucleotide overhangs allowed for nearly as efficient integration as 4-nt overhangs.

These results confirm that SpyCas1-Cas2 does not require additional factors for specificity *in vitro* and that integration relies on direct sequence recognition rather than any structural or topological features of the target. Integration reactions carried out at 4° C revealed faster integration at the leader-side attack site, suggesting that the sequence at this site is more favorable for binding or catalysis (Fig. 2.4c). Integration reactions performed with protospacers with one strand ending with a 3′ deoxy nucleotide, which can only make a single nucleophilic attack, still yielded both leader-side and spacer-side integration products, indicating that the two attacks are not strictly coordinated or ordered (Fig. 2.5b).

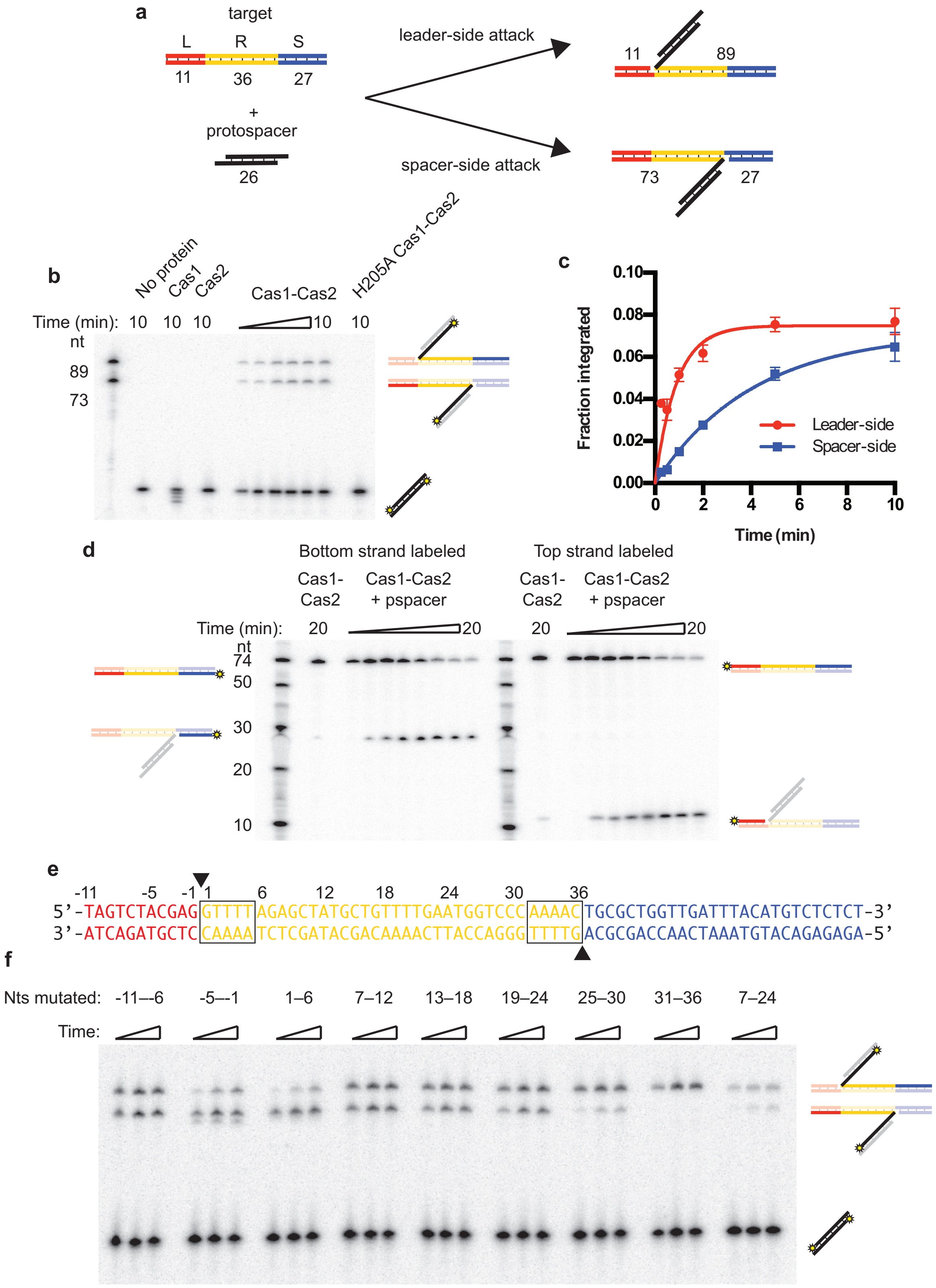


Figure 2.4 | SpyCas1-Cas2 recognizes sequences at the repeat ends for integration.

Figure 2.4 | SpyCas1-Cas2 recognizes sequences at the repeat ends for integration. (**a**) Schematic of integration into short linear targets. Leader (L) is shown in red, repeat (R) in yellow, spacer (S) in blue, and protospacer in black. The lengths of the substrate and expected products are indicated. (**b**) Integration reaction with short target and radiolabeled protospacer. Expected products are indicated Radiolabel is indicated with a star. (**c**) Quantification of (**b**). Points represents the mean of three experiments, with error bars representing the standard deviation. (**d**) Integration with radiolabeled target. Expected products are indicated. (**e**) Sequence of the WT target. Leader is shown in red, repeat in yellow, spacer in blue. Integration sites are indicated with arrowheads. Palindromic ends are boxed, and nucleotide numbering is indicated above the sequence. (**f**) Integration with mutant substrates. Transversion mutants (A→T, G→C) were made for all nucleotides in the indicated regions.

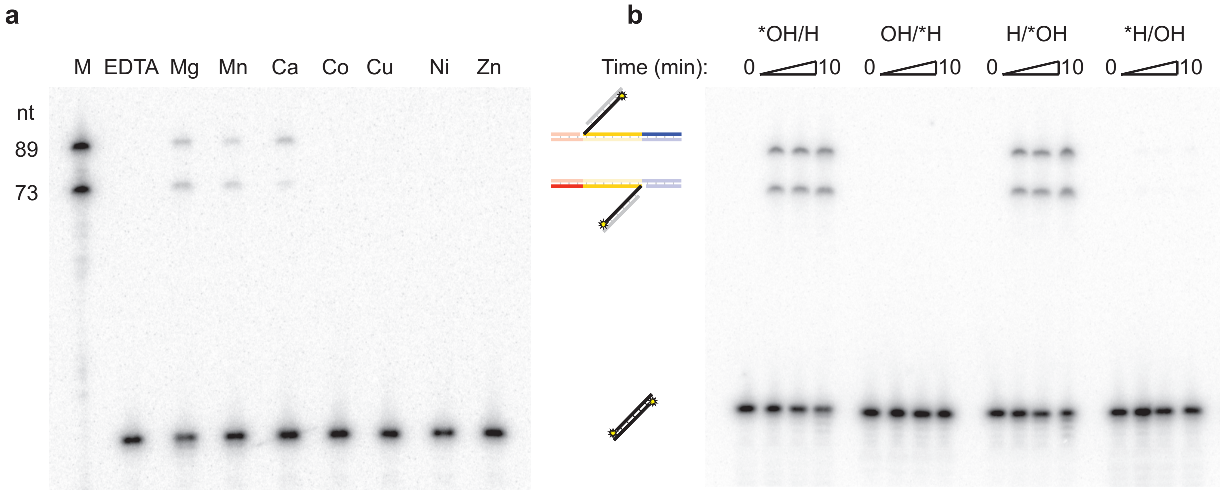


Figure 2.5 | SpyCas1-Cas2 requires divalent cations and 3' OH for integration. (**a**) Integration with radiolabeled protospacer and divalent cations. (**b**) Integration with protospacers lacking a 3′ nucleophile. 3′ deoxy strands are noted as “H”, and the labeled strand is indicated with an asterisk.

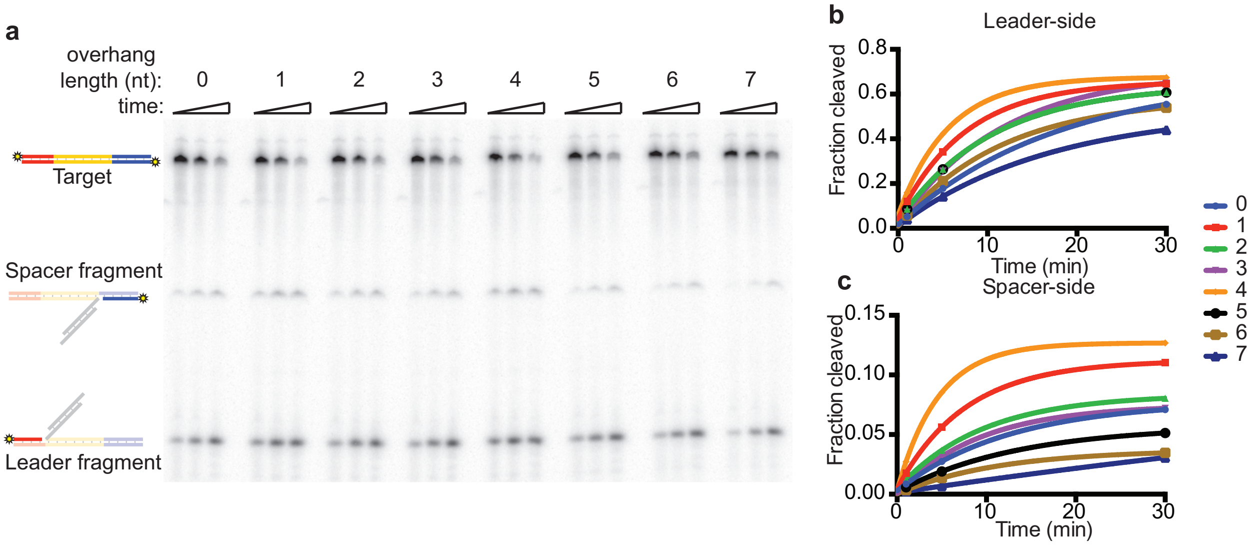


Figure 2.6 | 4-nucleotide overhangs allow for more rapid integration. (**a**) Integration assay with labeled target and protospacers with 3′ overhangs. Uncropped gel is shown in Supplementary Data Set 1. (**b**,**c**) Quantification of leader-side and spacer-side integration in (**a**).

2.3.4 Mutated targets support partial integration

We next investigated the sequence determinants of integration at either site by testing targets with five- or six-nucleotide blocks mutated across the leader and repeat sequence (Fig. 2.4e). Strikingly, mutations immediately adjacent to each integration site severely affected integration at that site without reducing integration at the distal integration site (Fig. 2.4f). Mutating the final five nucleotides of the leader or the first six nucleotides of the repeat substantially reduced the rate of integration at the leader-repeat junction (by 65% and 85% after two minutes, respectively), suggesting that sequence recognition occurs across the integration site, consistent with *in vivo* experiments (Wei et al., 2015a). Mutating the spacer-proximal end of the repeat eliminated all detectable integration at that site, and mutating nucleotides 25-30 also caused an observable reduction in integration rates. The internal sequence of the repeat appeared less critical for recognition. Mutating any block of six nucleotides was insufficient to produce an obvious difference in integration efficiency, while mutating nucleotides 7-24 reduced but did not eliminate integration. Cas1-Cas2 can tolerate degeneracy even in the leader and repeat ends, as single-nucleotide mutations in these regions also failed to noticeably impact integration (Fig. 2.7).

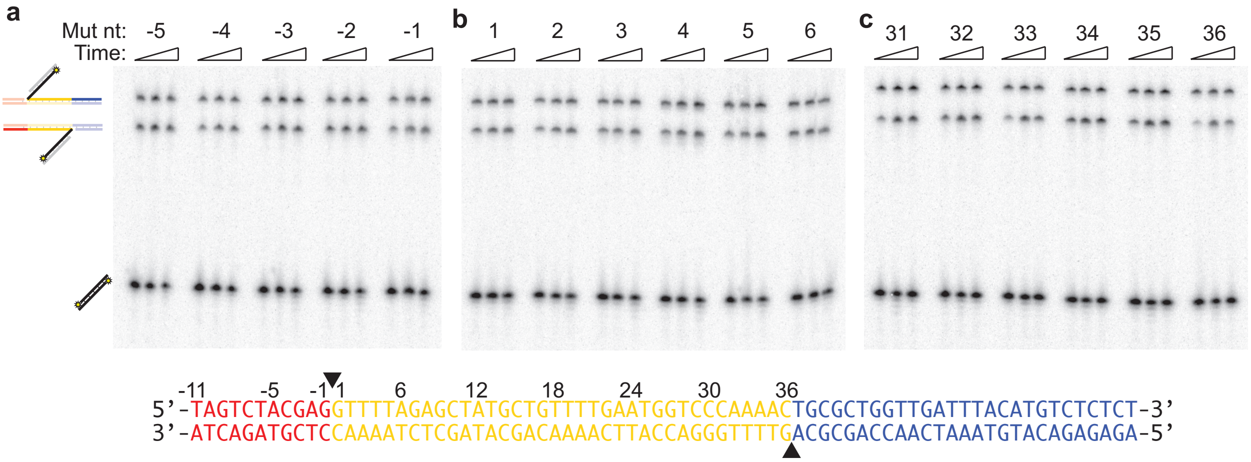

a
Mut nt:
-5
-4
-3
-2
-1
Time:
b
1
2
3
4
5
6
c
31
32
33
34
35
36
-11
-5
-1
1
6
12
18
24
30
36
5'-TAGTCTACGAGGTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAACTGCGCTGGTTGATTTACATGTCTCTCT-3'
3'-ATCAGATGCTCCAAAATCTCGATACGACAAAACTTACCAGGGTTTTGACGCGACCAACTAAATGTACAGAGAGA-5'

Figure 2.7 | Single nucleotide mutations have minor effects on integration. (**a-c**) Integration assays using targets with single-nucleotide transversions (C→G, A→T) at the indicated position in the leader (**a**), leader-proximal repeat (**b**), and spacer-proximal repeat (**c**). The WT target sequence is shown below for reference.

Together with the consensus sequences from the sequencing experiment, these data demonstrate the importance of the palindromic ends of the CRISPR repeat for recognition and integration (Fig. 2.4e). Structures of Cas1-Cas2 from *E. coli* show the complex to be a symmetrical hexamer, suggesting that, if the overall architecture is conserved, the catalytic Cas1 positioned at each integration site recognizes the inverted repeats by the same protein-nucleic acid contacts (Nuñez et al., 2014; 2015a; Wang et al., 2015). The leader sequence appears to provide additional sequence-specific contacts, explaining the more rapid integration observed at the leader-proximal site and the ability for Cas1-Cas2 to integrate, albeit weakly, when the leader-proximal portion of the repeat palindrome is mutated. The importance of the spacer-adjacent sequence for integration and the comparative dispensability of internal sequences represent a difference from type I systems, where the ends of repeats are often less palindromic and internal sequences dictate leader-distal integration (Nuñez et al., 2016; Wang et al., 2016).

2.3.5 Cas1-Cas2 catalyzes full-site integration *in vitro*

The ability for Cas1-Cas2 to integrate efficiently at only one end of a mutated target suggests that off-target integration events might be halted as half-site intermediates. To further explore the relevance of full-site integration in avoiding off-target integration, we designed hairpin targets that would yield a product of distinct length when a radiolabeled protospacer is fully integrated (Fig. 2.8a). By using targets with hairpins at either the leader or spacer end we were able to follow formation of both full-site products and half-site intermediates.

Incubation of Cas1-Cas2 with labeled protospacer and either hairpin target led to formation of the expected full-site product band (Fig. 2.8b,c). When the reaction was carried out with protospacers with one strand terminating in a dideoxy nucleotide, which should only be capable of making a half-site product, very low levels of the putative full-site band were produced. The apparent full-site products in these reactions likely stem from either uncoordinated half-site attacks by two protospacers on the same target or from protospacer-independent cleavage at the leader-repeat junction, as discussed above. Use of a protospacer containing both 3′ OH nucleophiles led to a ~50-fold increase in full-site product formation, suggesting that the large majority of these products are from full-site integration events. The spacer-side hairpin target, which had no detectable background in the single dideoxy control, was used for subsequent experiments.

Quantification of the three products formed, which, from largest to smallest, correspond to half-site products, full-site products, and a combination of full-site and half-site products, supported the model that leader-side integration occurs more quickly. Leader-side half-sites are most abundant product at the first time-point and disappear as full-site products are formed, suggesting that this half-site represents an intermediate that is converted to a full-site product (Fig. 2.8d). Spacer-side half-sites are produced more slowly, accounting for only 12% of integration products after 15 seconds, as opposed to the 56% made up of leader-side half-sites, and accumulate gradually rather than exhibiting a burst phase and decay (Fig. 2.8e). While spacer-side half-sites may also progress to full-site products, the faster rate of leader-side attack means that leader-side half-sites likely constitute the most common intermediate. Full-site products constitute ~60% of total integration events, confirming that SpyCas1-Cas2 efficiently catalyzes the full integration reaction without requiring the participation of Csn2 or additional factors.

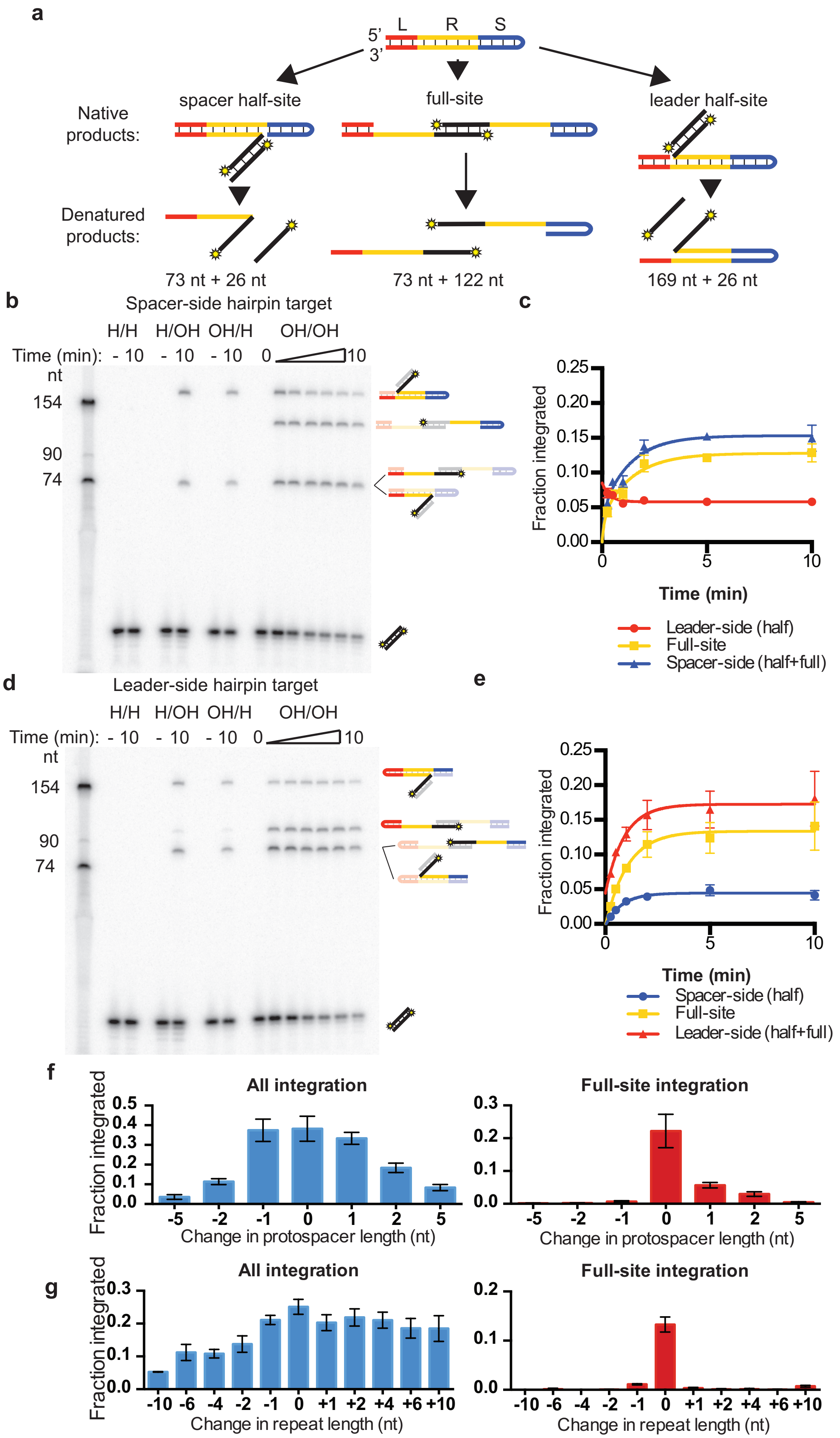


Figure 2.8 | Improper substrates are arrested as half-site intermediates.

Figure 2.8 | Improper substrates are arrested as half-site intermediates. (**a**) Schematic of hairpin target assay for detection of full-site products. The spacer-side hairpin target is shown, with leader in red, repeat in yellow, and spacer in blue. For the leader-side hairpin target, expected product sizes are 158 nt for the spacer-side half-site, 100 nt for the full-site product, and 84 nt for the leader-side integration products. Radiolabel is indicated with a star. (**b,d**) Full-site integration assays. Protospacers with one (H/OH, OH/H) or two (H/H) 3′ dideoxy termini were used as controls. Products are indicated next to each band. (**c**,**e**) Quantification of (**b**,**d**), respectively. (**f**) Quantification of integration with variable-length protospacers. A representative gel is shown in Fig. 2.9a (**g**) Quantification of integration with variable-length repeats. A representative gel is shown in Fig. 2.9b. For all quantifications, experiments were carried out in triplicate. Mean values were plotted and error bars represent standard deviation.

2.3.6 Improper substrates support only half-site integration

We next tested whether full-site integration can serve as a mechanism for the rejection of improper substrates other than partial recognition sequences. Acquired spacers are frequently of a defined length within an array, but integration assays with EcoCas1-Cas2 showed efficient integration of protospacers across a much wider length range (Nuñez et al., 2015b; Yosef et al., 2013). We tested whether protospacers that deviated from the 30-nt length observed in the *S. pyogenes* locus could support either half-site or full-site integration. We observed that lengthening or shortening the protospacer by a single nucleotide had minimal effects on overall integration, while adding or removing two or more nucleotides led to greater reductions (Fig. 2.8f). The full-site reaction, however, was far more sensitive to protospacer length. A 30-nt protospacer facilitated roughly 4-fold more full-site integration than a 31-nt protospacer and 30-fold greater full-site integration than a 29-nt protospacer (Fig. 2.8f). This suggests that protospacers of varying lengths can be bound by Cas1-Cas2 with one end positioned for cleavage, but only a correctly-sized protospacer can be positioned correctly at both active sites. We observed a strong preference for leader-side integration for mis-sized protospacers, suggesting that Cas1-Cas2 preferentially positions the properly coordinated protospacer end at the leader-repeat junction (Fig. 2.9a).

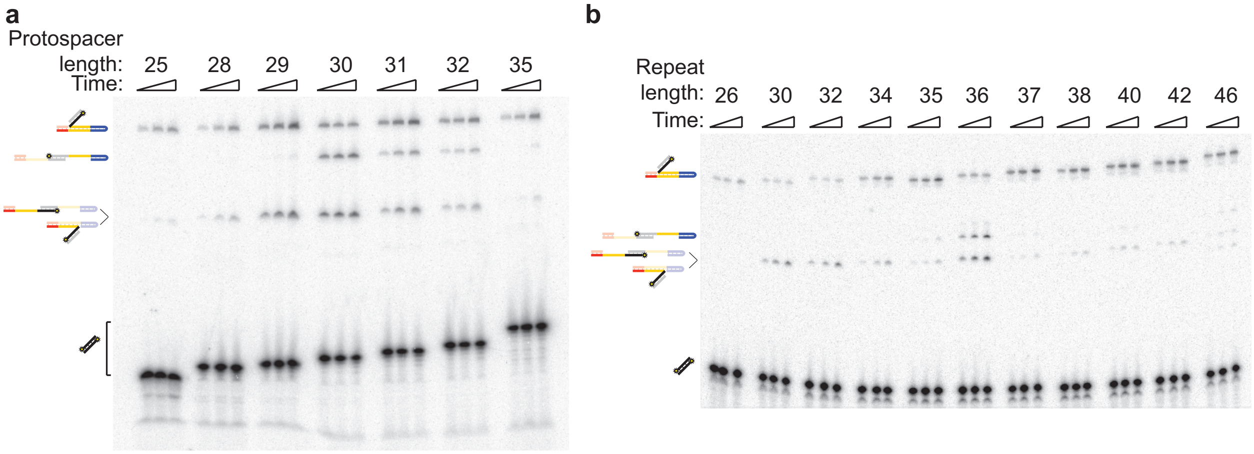

a
Protospacer
length: 25 28 29 30 31 32 35
Time:
b
Repeat
length: 26 30 32 34 35 36 37 38 40 42 46
Time:

Figure 2.9 | Full-site integration requires proper-length substrates. (**a**) Integration assay with hairpin target and radiolabeled variable-length protospacer. Product bands are indicated. (**b**) Integration assay with variable-length hairpin target and radiolabeled protospacer. Product bands are indicated.

The ability of Cas1-Cas2 to independently recognize repeat-terminal integration sequences raised the possibility that two recognition sequences with arbitrary spacing could support full-site integration. We added or removed sequences in the middle of the repeat of hairpin targets to test this possibility. Longer repeats had little effect on overall integration and shortened repeats supported ~50% of wild-type activity, but strikingly, a single nucleotide added or removed almost entirely abrogated full-site integration (Fig. 2.8g). For most mutant targets a preference for the leader-side was observed, again suggesting that Cas1-Cas2 preferentially forms the pre-reaction complex at the leader-repeat junction if both sites cannot be properly coordinated (Fig. 2.9b). Together with the data from mutated targets, this suggests that repeat-specific integration is maintained by a combination of sequence-recognition and ruler mechanisms, where appropriate sequences must be present with precise spacing to allow for full-site integration.

2.3.7 Off-target half-sites are disintegrated by Cas1-Cas2

To test whether half-site intermediates are disintegrated or converted to full-site products at different rates depending on the target sequence, we incubated Cas1-Cas2 with Y-DNA substrates mimicking a leader-side half-site intermediate (Fig. 2.10a). Substrates had either a perfect target sequence or an “off-target” sequence containing a correct leader-repeat junction (from -28 to +8) followed by a scrambled sequence. We observed disintegration as ligation of the plus-strand leader fragment and cleavage of the integrated protospacer strand, while progression to a full-site product causes cleavage of the minus-strand spacer fragment and ligation of the unintegrated protospacer strand. Cas1-Cas2 catalyzed disintegration of both half-site substrates, but roughly 50% more of the off-target half-site was disintegrated after 10 minutes (Fig. 2.10b,c, Fig. 2.11a). Only the reaction with a perfect target half-site yielded minus-strand cleavage products consistent with full-site integration, and the unintegrated strand of the protospacer was readily ligated into a larger product consistent with integration into the minus strand (Fig. 2.10d,e). Low levels of ligation products were also observed for the off-target substrate, but the lack of minus-strand cleavage for this substrate indicates that these products likely result from disintegration of the protospacer followed by re-integration into the plus strand (Fig. 2.10d,e).

To confirm that the wild-type ligation products resulted from full-site integration, rather than disintegration followed by re-integration, we substituted a leader fragment lacking a 3′ OH to prevent disintegration. No disintegration occurred, but ligation of the other strand occurred at levels comparable to the original substrate (Fig. 2.11b). We also tested wild-type and off-target spacer-side half-sites. These substrates supported activity similar to that of leader-side half-sites, with disintegration occurring at both wild-type and off-target substrates and secondary attack occurring only for the wild-type substrate (Fig. 2.11c-g). These data support the model that off-target integration events are arrested at the half-site stage or disintegrated, while perfect targets support full-site integration.

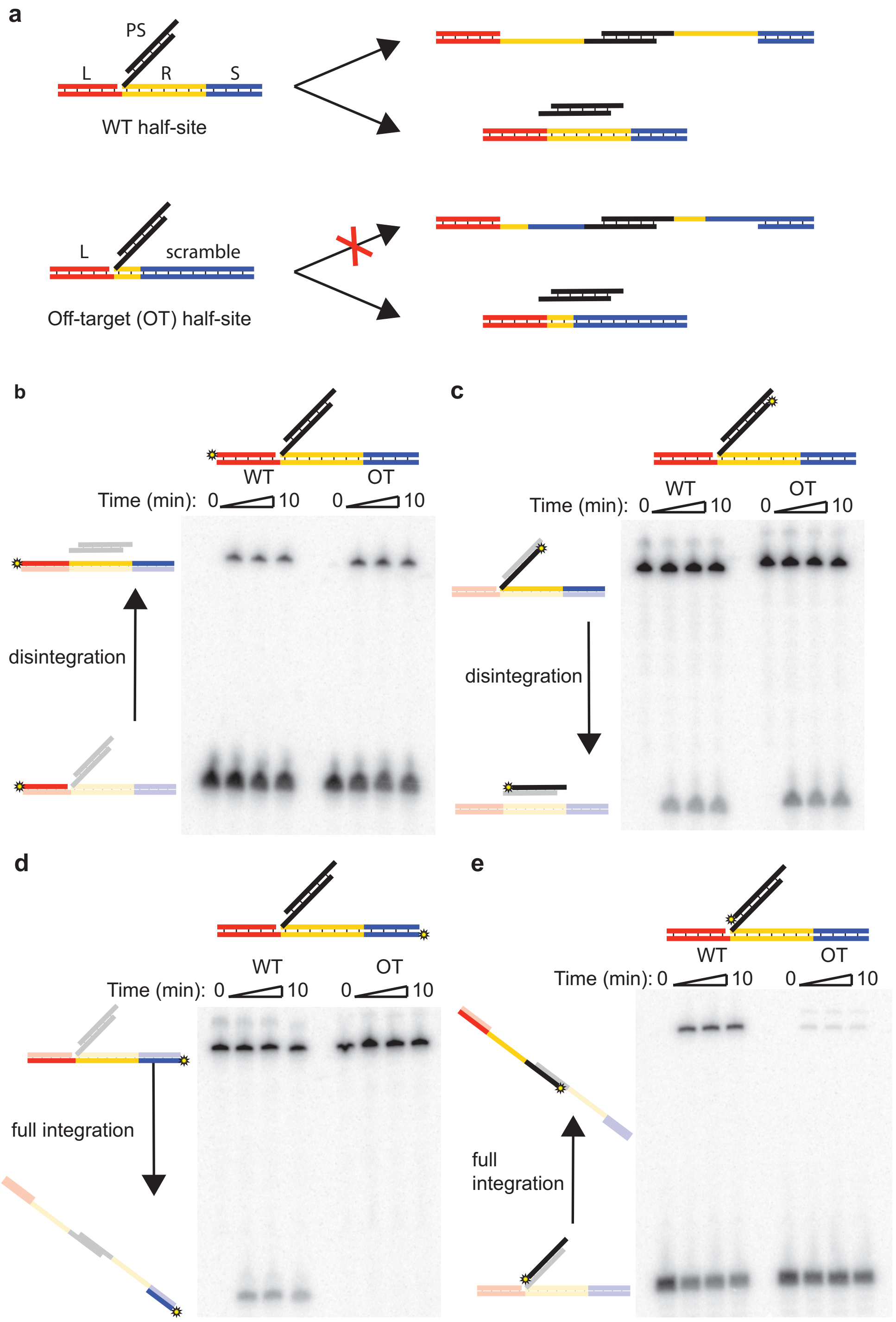


Figure 2.10 | Off-target half-sites are only resolved by disintegration. (**a**) Schematic of wild-type (WT) and off-target (OT) half-site substrates and full-site integration and disintegration products. (**b**,**c**) Disintegration assays with labeled leader (**b**) or labeled integrated protospacer (**c**). Radiolabel is indicated with a star. Quantification is shown in Supplementary Fig. 6a. (**d**,**e**) Full-site integration assays with labeled minus strand (**d**) or labeled unintegrated protospacer strand (**e**).

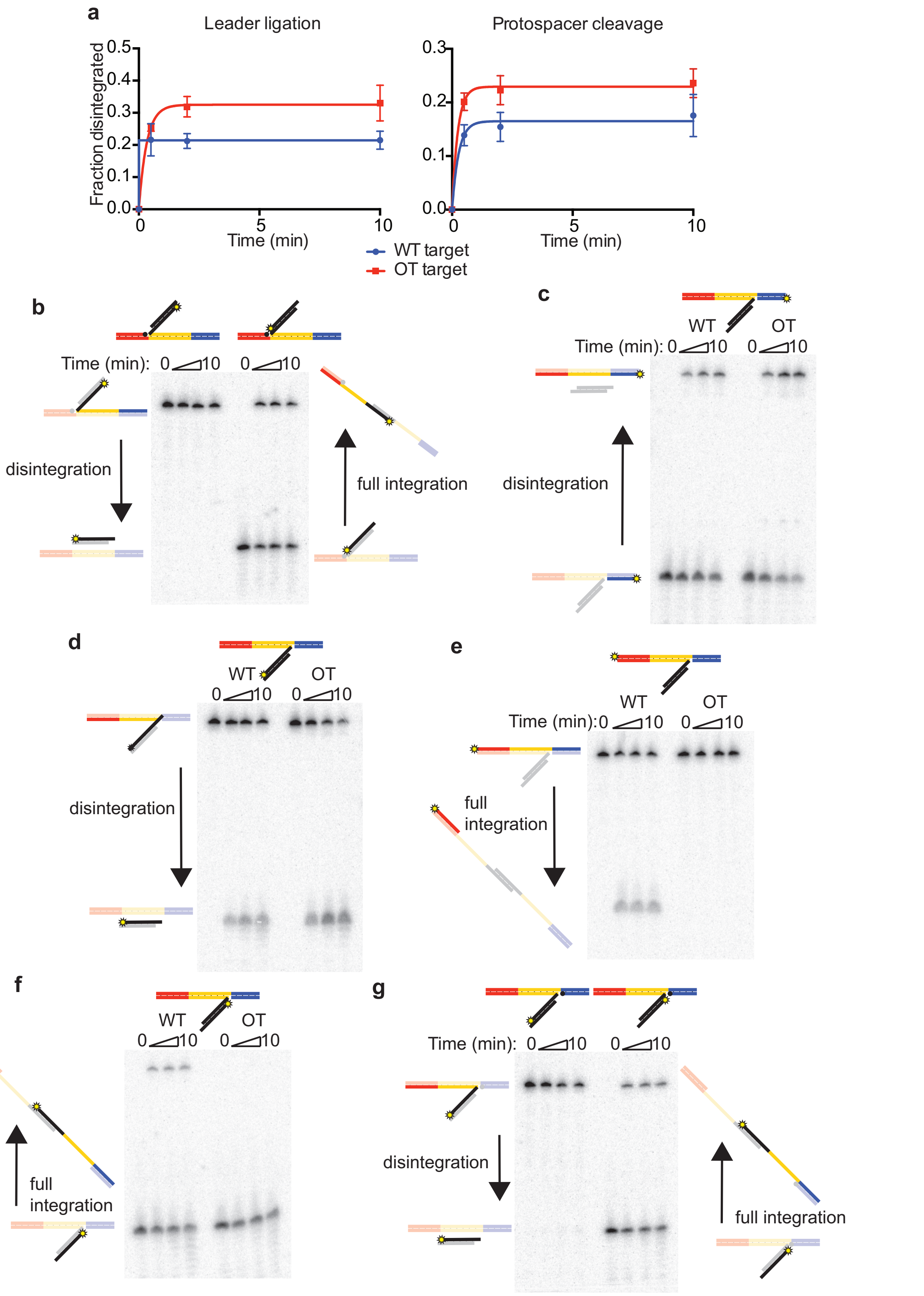


Figure 2.11 | Wild-type leader-side and spacer-side half-sites both support full-site integration. (**a**) Quantification of disintegration reaction shown in Figure 2.10b,c. Reaction was performed in triplicate. Means are plotted, with error bars representing standard deviation. (**b**) Disintegration and full-site integration assay with 3′ deoxy leader fragment and labeled protospacer strands on leader-side WT half-

site. (**c,d**) Disintegration assays with wild-type (WT) and off-target (OT) spacer-side half-site substrates with labeled spacer fragment (**c**) and labeled integrated protospacer strand (**d**). (**e,f**) Full-site integration assays with spacer-side half site with labeled plus strand (**e**) and labeled unintegrated protospacer (**f**). (**g**) Disintegration and full-site integration assay with 3′ deoxy spacer fragment and labeled protospacer strands on spacer-side half-site.

2.4 Discussion

Faithful recognition of the CRISPR locus by Cas1-Cas2 is essential for both the generation of new immunity and the maintenance of genome stability. The results presented here suggest a model for how target recognition and protospacer integration are coordinated by type II Cas1-Cas2. At proper targets, Cas1 active sites recognize the palindromic ends of the repeat. Initial integration can occur at either end of the repeat, but additional sequence contacts with the leader favor faster integration at the leader-proximal end. Integration at both sites yields a full-site product that can be repaired and propagated through subsequent cell divisions (Fig. 2.12). Partial recognition events can also occur, resulting in integration of one end of the protospacer. However, if another recognition site is not present 36 nucleotides downstream, the reaction arrests at the half-site step, allowing for disintegration by Cas1-Cas2 *in vitro* and, if the same partial reaction occurs *in vivo*, either disintegration by Cas1-Cas2 or excision by DNA repair proteins. Additional mechanisms may prevent off-target half-sites *in vivo*, but the disintegration activity of Cas1-Cas2 and the relative ease of repairing a nick following flap excision might render these unnecessary. Regardless of the physiological relevance of the half-site reaction, it appears that Cas1-Cas2 maintain high specificity for the full-site reaction by combining bipartite recognition of the inverted motifs with a strict ruler mechanism.

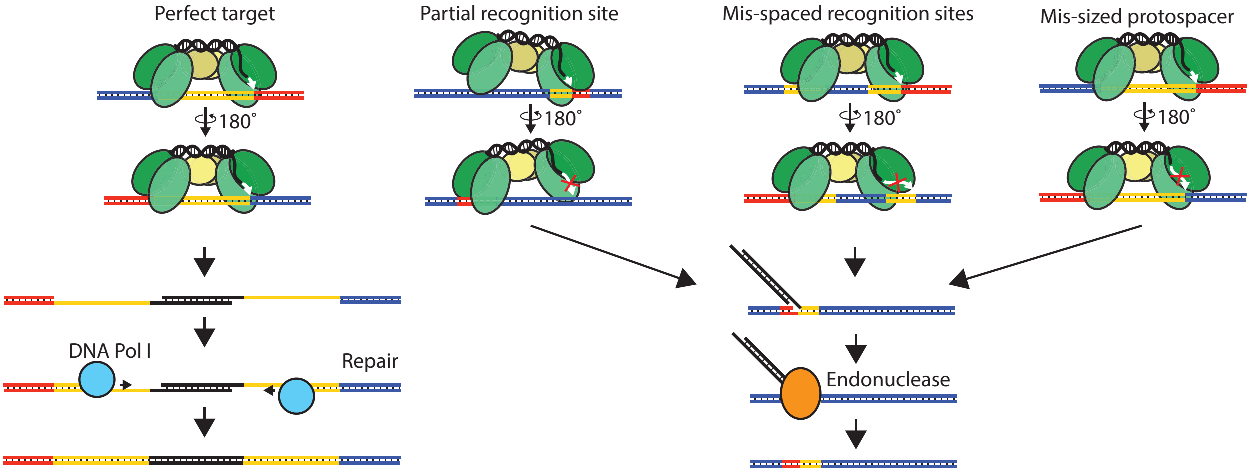


Figure 2.12 | Model for maintenance of genome stability by SpyCas1-Cas2. Integration by Cas1-Cas2 requires both a proper-length protospacer and a target containing appropriately-spaced recognition sequences for both Cas1 active sites. If all criteria are met, full-site integration yields a product that can be repaired and propagated in the genome. At off-target integration sites or with mis-sized protospacers, half-site integration can occur, but the resulting intermediate can be repaired by Cas1-Cas2 disintegration or possibly by the action of cellular DNA repair factors, preventing lasting damage to the genome. The schematic assumes that the Cas1-Cas2 complex architecture from *E. coli* is conserved. Leader sequences are represented in red, repeat sequences are yellow, and nonspecific sequences are blue. Nucleophilic attacks are shown as white arrows.

Our experiments revealed sequence requirements for a type II Cas1-Cas2 that differ from those of previously-studied type I Cas1-Cas2 integrases. Both types rely on sequences at the leader-repeat junction to direct integration, but type I systems recognize internal sequences to direct the spacer-side attack while type II systems directly recognize the end of the repeat (Nuñez et al., 2016; Wang et al., 2016; Wei et al., 2015a). While the structure of the putative *S. pyogenes* Cas1-Cas2 complex is unknown, the ruler mechanism and symmetrical sequence recognition are consistent with a symmetrical structure like that of the *E. coli* complex. Under this model, it seems the *S. pyogenes* proteins utilize a recognition mode reminiscent of type II restriction enzymes and other homodimeric DNA-binding proteins (Pingoud and Jeltsch, 2001). Type II Cas1 genes comprise a distinct phylogenetic branch from those of type I and type III (Makarova et al., 2015). One possible explanation for the divergence of the two families might be the differing roles of the 3′ end of the repeat in the interference complexes. In type I and type III systems, the 3′ end of the repeat is retained as the 5′ tag of the crRNA, where it is recognized by Cas5 in type I systems and Cmr3 or Csm4 in type III systems (Plagens et al., 2015). In type II systems, the 3′ end is cleaved by RNase III after annealing with the tracrRNA (Deltcheva et al., 2011). Evolutionary pressure may have caused type I and III repeats to lose their palindromic ends, forcing Cas1-Cas2 to recognize internal sequences, while type II systems could maintain palindromic recognition.

Cas1 is believed to have evolved from a transposase, and Cas1 homologs have been identified in transposons (or Casposons) and confirmed as active (Hickman and Dyda, 2015; Krupovic et al., 2014; Makarova et al., 2015). The domestication of Cas1 would have required dramatic changes in substrate specificity, most strikingly a shift from integrating essentially at random to integrating only at a defined site. Our results show how this specificity is maintained without allosteric regulation or long stretches of sequence recognition and reveal how the proposed structure of the Cas1-Cas2 complex and the spacing of Cas1 active sites dictate recognition and specificity. By understanding how the Cas1-Cas2 integrase recognizes its target site, we can also make strides toward exploiting this site-specific integrase activity for biological applications. Both half-site and full-site integration by Cas1-Cas2 have potential uses, such tagging genomic sites, introducing barcodes, and other applications where the specific integration and short integration fragment of Cas1-Cas2 provide advantages over transposases. A greater knowledge of the diverse sequence requirements of Cas1-Cas2 integrases will allow for the prediction of genomic targets and possibly the generation of new integrases with desired specifities.

2.5 Methods

2.5.1 Protein purification

The *S. pyogenes* Cas1, Cas2, and Csn2 genes were PCR amplified from the *S. pyogenes* M1 GAS genome and separately cloned into pET16b (Novagen) with an N-terminal His_6-MBP tag (Nuñez et al., 2014). Each construct was expressed separately in BL21 (DE3) 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 KCl, 10 mM imidazole, 0.1%

Triton X-100, 2 mM *tris*(2-carboxyethyl)phosphine (TCEP), 0.5 mM phenylmethylsulfonyl fluoride (PMSF), Complete EDTA-free protease inhibitor (Roche), and 10% glycerol). Cells were lysed by sonication, and lysate was cleared by centrifugation. The supernatant was incubated on Ni-NTA resin (Qiagen). The resin was washed with wash buffer (20 mM HEPES, pH 7.5, 500 mM KCl, 10 mM imidazole, 1 mM TCEP, 5% glycerol), and protein was eluted with wash buffer supplemented with 300 mM imidazole. The proteins were dialyzed against wash buffer without imidazole and incubated with TEV protease overnight to remove the affinity tags. Tags were separated by binding to Ni-NTA resin. Cas2 was bound to a HiTrap heparin HP column (GE Healthcare) and eluted with a gradient from 500 mM to 1 M KCl. Cas1 and Csn2 were dialyzed against buffer with 150 mM KCl before binding to a HiTrap heparin HP column or HiTrap Q HP column (GE Healthcare), respectively, and eluted with a gradient from 150 mM to 1 M KCl. All proteins were further purified on a Superdex 75 (16/60) column with gel filtration buffer (20 mM HEPES, pH 7.5, 150 mM KCl, 10 mM imidazole, 1 mM TCEP, 5% glycerol), except for Cas2, which was purified with gel filtration buffer supplemented to 500 mM KCl. H205A Cas1 was generated by ‘Round the Horn mutagenesis using the primers 5′-TTTAAGCCCAAACTGAGTCATACAT-3′ and 5′-GCTGCTAATCAGTTTAATCAGTTCAATTTTGC-3′ and purified by the same procedures (Hemsley et al., 1989). *E. coli* Cas1 and Cas2 were purified as previously described (Nuñez et al., 2014).

2.5.2 DNA substrate preparation

pEcoCR was generated as described for pCRISPR in Nuñez et al., 2015a. pSpyCR was cloned by PCR-amplifying the CRISPR array and leader sequence from *S. pyogenes* M1 GAS using the primers 5′-AGAGAGGAATTCTACTCTTAATAAATGCAGTAATACAGGGGC-3′ and 5′-AGAGAGACATGTCTCTTTCTCAAGTTATCATCGGCAATG-3′ and ligating into PciI-EcoRI-digested pUC19. Plasmids were linearized by digestion with NdeI (New England Biosciences) and purified by phenol-chloroform extraction followed by ethanol precipitation.

All oligonucleotides were synthesized by Integrated DNA Technologies. Protospacers and dsDNA targets were hybridized by heating to 95° C and slow-cooling to room temperature in hybridization buffer (20 mM Tris, pH 7.5, 100 mM KCl, 5 mM $MgCl_2$) and purified on 8% native PAGE. Protospacers and targets were labeled using with [γ-^{32}P]-ATP (Perkin Elmer) and T4 polynucleotide kinase (New England Biosciences). For dsDNA substrates with a single strand labeled, hybridization was carried out with 2-fold excess of the unlabeled strand and followed by gel purification. Mutant dsDNA targets were generated by overlap extension PCR. Hairpin targets used in the variable repeat experiment, including the unaltered repeat, were made by Klenow fragment fill-in of partial hairpins, and the final products were purified on 6% urea-PAGE. Half-site intermediate substrates were made by hybridizing the labeled strand with 5-fold excess of all other strands and purifying on 8% native PAGE. Sequences of all substrates are shown in Table 2.2.

Description	Sequence
30mer protospacer – 2.1b,c, 2.2, 2.3, 2.5	GACAGAGTTACTACTCGTTCTGGCTCTGTC
RC	GACAGAGCCAGAACGAGTAGTAACTCTGTC
33mer protospacer – 2.1b	GCGAGAATTACTACTCGTTCTGGTGTTTCTCGC
RC	GCGAGAAACACCAGAACGAGTAGTAATTCTCGC
1-nt overhang protospacer – 2.5	ACAGAGTTACTACTCGTTCTGGCTCTGTC
RC	ACAGAGCCAGAACGAGTAGTAACTCTGTC
2-nt overhang protospacer – 2.5	CAGAGTTACTACTCGTTCTGGCTCTGTC
RC	CAGAGCCAGAACGAGTAGTAACTCTGTC
3-nt overhang protospacer – 2.5	AGAGTTACTACTCGTTCTGGCTCTGTC
RC	AGAGCCAGAACGAGTAGTAACTCTGTC
4-nt overhang protospacer – 2.4, 2.6, 2.7 2.8, 2.9, 2.10	GAGTTACTACTCGTTCTGGCTCTGTC
RC	GAGCCAGAACGAGTAGTAACTCTGTC
5-nt overhang protospacer – 2.5	AGTTACTACTCGTTCTGGCTCTGTC
RC	AGCCAGAACGAGTAGTAACTCTGTC
6-nt overhang protospacer – 2.5	GTTACTACTCGTTCTGGCTCTGTC
RC	GCCAGAACGAGTAGTAACTCTGTC
7-nt overhang protospacer – 2.5	TTACTACTCGTTCTGGCTCTGTC
RC	CCAGAACGAGTAGTAACTCTGTC
4-nt overhang 3' ddO protospacer – 2.5b, 2.8b,d	GAGTTACTACTCGTTCTGGCTCTGTddC
RC	GAGCCAGAACGAGTAGTAACTCTGTddC
WT target – 2.4b,c,d, 2.5, 2.6	TAGTCTACGAG**GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC**TGCGCTGGTTGATTTACATGTCTCTCT
RC	AGAGAGACATGTAAATCAACCAGCGCA**GTTTTGGGACCATTCAAAACAGCATAGCTCTAAAAC**CTCGTAGACTA
-11–-6 mut target fwd – 2.4f	ATCAGAACGAG**GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC**TGC
-5–-1 mut target fwd – 2.4f	TAGTCTTGCTC**GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC**TGC
1–6 mut target fwd – 2.4f	TAGTCTACGAG**CAAAATGAGCTATGCTGTTTTGAATGGTCCCAAAAC**TGC
7-12 mut target fwd – 2.4f	TAGTCTACGAG**GTTTTACTCGATTGCTGTTTTGAATGGTCCCAAAAC**TGC
-11–12 mut target rev – 2.4f	AGAGAGACATGTAAATCAACCAGCGCA**GTTTTGGGACCATTCAAAACAGC**
13-18 mut target fwd – 2.4f	TAGTCTACGAG**GTTTTAGAGCTAACGACATTTGAATGGTCCCAAAAC**TGC
13-18 mut target rev – 2.4f	AGAGAGACATGTAAATCAACCAGCGCA**GTTTTGGGACCATTCAAATGTCG**
19-24 mut target fwd – 2.4f	TAGTCTACGAG**GTTTTAGAGCTATGCTGTAAACTTTGGTCCCAAAAC**TGC
19-24 mut target rev – 2.4f	AGAGAGACATGTAAATCAACCAGCGCA**GTTTTGGGACCAAAGTTTACAGC**
25-30 mut target fwd – 2.4f	TAGTCTACGAG**GTTTTAGAGCTATGCTGTTTTGAAACCAGGCAAAAC**TGC
25-30 mut target rev – 2.4f	AGAGAGACATGTAAATCAACCAGCGCA**GTTTTGCCTGGTTTCAAAACAGC**
31-36 mut target fwd – 2.4f	TAGTCTACGAG**GTTTTAGAGCTATGCTGTTTTGAATGGTCCGTTTTG**TGC
31-36 mut target rev – 2.4f	AGAGAGACATGTAAATCAACCAGCGCA**CAAAACGGACCATTCAAAACAGC**
7–24 mut target fwd – 2.4f	TAGTCTACGAG**GTTTTACTCGATACGACAAAACTTTGGTCCCAAAAC**TGC
7–24 mut target rev – 2.4f	AGAGAGACATGTAAATCAACCAGCGCA**GTTTTGGGACCAAAGTTTTGTCG**
Spacer-side hairpin target – 2.8b,f, 2.9a	TAGTCTACGAG**GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC**TGCGCTGGTTGATTTACATGTCTCTCTcgatagAGAGAGACATGTAAATCAACCAGCGCA**GTTTTGGGACCATTCAAAACAGCATAGCTCTAAAAC**CTCGTAGACTA
Leader-side hairpin target – 2.8c,e	GACATGTAAATCAACCAGCGCA**GTTTTGGGACCATTCAAAACAGCATAGCTCTAAAAC**CTCGTAGACTATTTTTcgatagAAAAATAGTCTACGAG**GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC**TGCGCTGGTTGATTTACATGTC
Partial hairpin – 2.8g, 2.9b	TAGTCTACGAG**GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC**TGCGCTGGTTGCTCGCTcgatagAGCGAGCAAC
-1 partial hairpin – 2.8g, 2.9b	TAGTCTACGAG**GTTTTAGAGCTATGCGTTTTGAATGGTCCCAAAAC**TGCGCTGGTTGCTCGCTcgatagAGCGAGCAAC
-2 partial hairpin – 2.8g, 2.9b	TAGTCTACGAG**GTTTTAGAGCTATGGTTTTGAATGGTCCCAAAAC**TGCGCTGGTTGCTCGCTcgatagAGCGAGCAAC
-4 partial hairpin – 2.8g, 2.9b	TAGTCTACGAG**GTTTTAGAGCTAGTTTTGAATGGTCCCAAAAC**TGCGCTGGTTGCTCGCTcgatagAGCGAGCAAC
-6 partial hairpin – 2.8g, 2.9b	TAGTCTACGAG**GTTTTAGAGCGTTTTGAATGGTCCCAAAAC**TGCGCTGGTTGCTCGCTcgatagAGCGAGCAAC
-10 partial hairpin – 2.8g, 2.9b	TAGTCTACGAG**GTTTTAGAGTTTTGAATGCCCAAAAC**TGCGCTGGTTGCTCGCTcgatagAGCGAGCAAC
+1 partial hairpin – 2.8g, 2.9b	TAGTCTACGAG**GTTTTAGAGCTATGCTAGTTTTGAATGGTCCCAAAAC**TGCGCTGGTTGCTCGCTcgatagAGCGAGCAAC
+2 partial hairpin – 2.8g, 2.9b	TAGTCTACGAG**GTTTTAGAGCTATGCTGAGTTTTGAATGGTCCCAAAAC**TGCGCTGGTTGCTCGCTcgatagAGCGAGCAAC
+4 partial hairpin – 2.8g, 2.9b	TAGTCTACGAG**GTTTTAGAGCTATGCTGACTGTTTTGAATGGTCCCAAAAC**TGCGCTGGTTGCTCGCTcgatagAGCGAGCAAC

+6 partial hairpin – 2.8g, 2.9b	TAGTCTACGAG**GTTTTAGAGCTATGCT**GACTGA**GTTTTGAATGGTCCCAAAAC**TGCGCTGGTTGCTCGCTcgatagAGCGAGCAAC
+10 partial hairpin – 2.8g, 2.9b	TAGTCTACGAG**GTTTTAGAGCTATGCT**GACTGACTGA**GTTTTGAATGGTCCCAAAAC**TGCGCTGGTTGCTCGCTcgatagAGCGAGCAAC
Half-site leader – 2.10, 2.11a	AATTTTTTAGACAAAAATAGTCTACGAG
WT half-site protospacer-repeat-spacer – 2.10, 2.11a,b	GAGTTACTACTCGTTCTGGCTCTGTC**GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC**GGCGCTGGTTGATTTCTTCTTGCGAG
WT leader half-site minus strand – 2.10, 2.11a	CTCGCAAGAAGAAATCAACCAGCGCC**GTTTTGGGACCATTCAAAACAGCATAGCTCTAAAAC**CTCGTAGACTATTTTTGTCTAAAAAATT
Mut half-site protospacer-repeat-spacer – 2.10, 2.11a	GAGTTACTACTCGTTCTGGCTCTGTC**GTTTTAGA**TAAGGTTCCGACGTATACTCGTTGTCCGTTCTTAATGGCTGTGCGGCAGTATGT
Mut half-site minus strand – 2.10, 2.11a	ACATACTGCCGCACAGCCATTAAGAACGGACAACGAGTATACGTCGGAACCTTA**TCTAAAAC**CTCGTAGACTATTTTTGTCTAAAAAATT
3' ddO half-site leader – 2.11b	AATTTTTTAGACAAAAATAGTCTACGAddC
ddO leader half-site minus strand – 2.11b	CTCGCAAGAAGAAATCAACCAGCGCCGTTTTGGGACCATTCAAAACAGCATAGCTCTAAAACGTCGTAGACTATTTTTGTCTAAAAAATT
Half-site spacer – 2.11c-f	CTCGCAAGAAGAAATCAACCAGCGCC
WT half-site protospacer-repeat-leader – 2.11c-g	GAGCCAGAACGAGTAGTAACTCTGTG**GTTTTGGGACCATTCAAAACAGCATAGCTCTAAAAC**CTCGTAGACTATTTTTGTCTAAAAAATT
WT spacer half-site plus strand – 2.11c-g	AATTTTTTAGACAAAAATAGTCTACGAG**GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAACG**GCGCTGGTTGATTTCTTCTTGCGAG
Mut half-site protospacer-repeat-leader – 2.11c-f	GAGCCAGAACGAGTAGTAACTCTGTG**GTTTTGGG**GTACAATTATCTATTTCTACACATCCGCTTAGACTGTACTCCAATAAAAGATAAAA
Mut half-site plus strand – 2.11c-f	TTTTATCTTTTATTGGAGTACAGTCTAAGCGGATGTGTAGAAATAGATAATTGTAC**CCCAAAAC**GGCGCTGGTTGATTTCTTCTTGCGAG
3' ddO half-site spacer – 2.11g	CTCGCAAGAAGAAATCAACCAGCGCddC

Table 2.2 | DNA substrates used in this study. Relevant figures are indicated after the description. RC indicates the complementary strand of the previous oligonucleotide. All sequences are written 5′ to 3′. Red indicates mutated or added nucleotides, bold indicates repeat sequences. Single nucleotide mutations used in Figure 2.7 are omitted.

2.5.3 Integration assays

Integration assays using plasmid target and unlabeled protospacer were carried out as previously described (Nuñez et al., 2015b). Briefly, 1 μM Cas1 and Cas2 were incubated together on ice in integration buffer (20 mM HEPES, pH 7.5, 25 mM KCl, 10 mM $MgCl_2$, 1 mM DTT, 10% DMSO) for 30 minutes. Cas1-Cas2 was diluted to a final concentration of 75 nM and incubated with 200 nM protospacer for 10 minutes. *E. coli* Cas1-Cas2 were provided with a 33-nt protospacer, and *S. pyogenes* Cas1-Cas2 were provided with a 30-nt protospacer. Where indicated, 75 nM Csn2 was added before the addition of protospacer. Plasmid was added to a final concentration of 7.5 nM and the reaction was carried out at 37° C for 1 hour before quenching with 0.4% SDS and 25 mM EDTA, extracting with phenol-chloroform-isoamyl alcohol, and analyzing products on a 1.5% ethidium bromide-stained agarose gel. Assays with radiolabeled protospacer and plasmid target were carried out with ~1 nM protospacer, 100 nM Cas1-Cas2, and 75 nM plasmid, an unstained agarose gel was used, and the gel was dried and visualized using phosphorimaging.

Integration assays with linear dsDNA targets and radiolabeled protospacers were carried out with 100 nM Cas1-Cas2, ~1 nM protospacer, and 100 nM target in integration buffer supplemented with 0.01% nonidet P-40. Assays to test for metal-dependence were carried out with integration buffer lacking $MgCl_2$ and supplemented with 10 mM of EDTA or the chloride salt of the indicated metal. Reactions were incubated at 16° C, except where otherwise noted, and timepoints were taken at 0.25, 0.5, 1, 2, 5, and 10 minutes and quenched by the addition of an equal volume of 95% formamide and 50 mM EDTA. Samples were run on 8% urea-PAGE. Assays with radiolabeled targets contained 100 nM Cas1-Cas2, 25 nM protospacer, and ~1 nM target and were carried out at room temperature, with the final timepoint taken at 20 rather than 10 minutes. Samples were run on 10% urea-PAGE. Reactions with hairpin substrates were performed as described for linear dsDNA substrates, except that samples were analyzed with 6% urea-PAGE. Reactions with half-site intermediate substrates were performed with 100 nM Cas1-Cas2 and ~1 nM substrate at room temperature. Time points were taken at 0.5, 2 and 10 minutes and analyzed with 10% urea-PAGE. All gels were visualized with phosphorimaging and quantified with ImageQuant (GE Healthcare). For quantification of hairpin integration experiments, "all integration" is calculated by the equation (2×top + 2×bottom)/(total), where "top" is the intensity of the top integration band (half-site only), "bottom" is the bottom integration band (full-site and half-site), and "total" is the sum of all bands in the lane. "Full-site integration" is calculated by the equation (2×middle)/(total), where "middle" is the intensity of the middle integration band (full-site only). In the first equation, the integration band intensities are doubled to account for the presence of the unintegrated protospacer strand in the free protospacer band. The middle band intensities are doubled in the second equation to account for the presence of the second protospacer strand in the bottom integration band. All non-linear regression was performed in Prism (Graphpad).

2.5.4 High-throughput sequencing

Sequencing of integration products was performed as previously described (Nuñez et al., 2015b). The reaction was carried out with 75 nM Cas1-Cas2, 200 nM blunt-ended protospacer, and 7.5 nM of pUC19, pEcoCR, or pSpyCR in integration buffer at 37° C for one hour. Integration products were fragmented to ~100 bp using dsDNA Fragmentase (New England Biolabs), end-repaired, A-tailed, and ligated with the NEBNext adapter for Illumina (New England Biolabs). Libraries were amplified using Q5 polymerase and NEBNext universal and index primers for Illumina (New England Biolabs) and sequenced on an Illumina HiSeq2500 in rapid run mode with 150 nt single reads.

3′ adapter sequences were removed from reads using Cutadapt (Marcel Martin, 2015). Reads containing at least 10 nucleotides of protospacer sequence with no errors were identified and trimmed using Cutadapt, and the resulting reads were mapped to the respective plasmids using Bowtie, allowing 2 mismatches and requiring unique mapping (Langmead et al., 2009). Reads without protospacer sequences were also mapped with the same criteria to establish background. The consensus integration sequences were generated with WebLogo using all integration events within a given data set (Crooks et al., 2004). Raw read alignments were visualized using the Integrative Genomics Viewer (Robinson et al., 2011; Thorvaldsdóttir et al., 2013).

2.6 Accession codes

Sequencing reads have been deposited in the Sequence Read Archive under accession code SRP079023.

2.7 Acknowledgments

We are grateful to J.K. Nuñez for technical assistance and members of the Doudna lab for discussion and input on the manuscript. This project was funded by US National

Science Foundation grant No. 1244557 (J.A.D.) and National Institute of General Medicine Sciences grant No. 1P50GM102706-01 (J.H. Cate). A.V.W. is supported by a US National Science Foundation Graduate Fellowship. J.A.D. is an Investigator of the Howard Hughes Medical Institute and a Paul Allen Distinguished Investigator. This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 Instrumentation Grants S10RR029668 and S10RR027303.

CHAPTER 3

Structures of the CRISPR genome integration complex

A portion of the work presented in this chapter has been published previously as part of the following paper: Wright, A.V., Liu, J.-J., Knott, G.J., Doxzen, K.W., Nogales, E., and Doudna, J.A. (2017). Structures of the CRISPR genome integration complex. *Science*. **357** (6356) 1113-1118.

Addison Wright performed biochemical and bacterial genetics experiments, generated samples for electron microscopy, solved the crystal structures with assistance from Gavin Knott and Kevin Doxzen. Jun-Jie Liu performed electron microscopy and solved the structure with assistance from Addison Wright. Addison Wright, Jun-Jie Liu, Eva Nogales, and Jennifer Doudna designed the study and wrote the manuscript.

3.1 Abstract

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

3.2 Introduction

CRISPR-Cas (Clustered Regularly Interspaced Short Palindromic Repeats-CRISPR associated) bacterial adaptive immune systems store fragments of viral DNA in the CRISPR array, a genomic locus comprising direct sequence repeats separated by virally-derived spacer sequences, both of approximately 20-50 base pairs in length (Barrangou et al., 2007; Bolotin et al., 2005; Mojica et al., 2005; Pourcel et al., 2005). In most systems, a transcriptional promoter located in an AT-rich leader sequence preceding the first CRISPR repeat gives rise to precursor CRISPR transcripts that are processed and used to recognize viral nucleic acids by base pairing with complementary sequences. Bacteria acquire immunity to new viruses when the CRISPR integrase, a heterohexameric complex of four Cas1 and two Cas2 proteins, inserts new viral DNA at the first CRISPR repeat following the leader sequence (Nuñez et al., 2014; 2015b; Yosef et al., 2012). Integration involves nucleophilic attack by the 3’ ends of the viral DNA fragment, called a protospacer, at each end of the repeat (Fig. 3.1a) (Nuñez et al., 2015b). Half-site intermediates form when one of the two protospacer DNA ends attacks the CRISPR locus integration site, and can either progress to full-site integration products or be disintegrated, leaving the target sequence intact (Nuñez et al., 2015b; Rollie et al., 2015).

To ensure effective acquisition of new immunity and avoid deleterious insertions into the genome, integration by Cas1-Cas2 must be highly specific for the CRISPR locus. In the type I CRISPR system from *E. coli*, acquisition requires sequences spanning the leader-repeat junction as well as an inverted repeat motif in the repeat (Goren et al., 2016; Moch et al., 2017; Nuñez et al., 2016; Rollie et al., 2015). IHF (Integration Host Factor), a histone-like protein, binds in the leader and assists in recruiting Cas1-Cas2 to the leader-proximal repeat, possibly involving a secondary upstream binding site (Fagerlund et al., 2017; Nuñez et al., 2016; Yoganand et al., 2017). The mechanism by which Cas1-Cas2 recognizes these sequences is not yet known.

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

3.3 Results

3.3.1 Target binding in the half-site intermediate

To determine the mechanism by which Cas1-Cas2 recognizes its target sequence, we crystallized the integrase bound to DNA substrates representing a half-site integration intermediate as well as the full-site integration product (Fig. 3.1a). The full-site product mimic, which we term the pseudo-full-site substrate, was designed with a break in the middle of the protospacer to allow Cas1-Cas2 to access the repeat (Fig. 3.1a). Both substrates bound to Cas1-Cas2 with high affinity (Fig. 3.2). The half-site-bound structure, refined at 3.9 Å resolution, revealed an overall complex architecture similar to that of the previously-solved protospacer-bound structures (Fig. 3.1b, 3.3, Table 3.1) (Nuñez et al., 2015a; Wang et al., 2015). A Cas2 dimer sits at the center of two Cas1 dimers, with the protospacer DNA stretching across the flat back of the complex. The first 18 base pairs of the repeat sequence bind across a central channel formed by Cas2 and the non-catalytic Cas1 monomers, with the leader-repeat junction positioned across a Cas1 active site (Fig. 3.1b, 3.4a,b). Seven nucleotides of the spacer-proximal repeat are unresolved, while the repeat-spacer junction binds at the distal Cas1 active site. Basic residues on both Cas2 (K38, R40) and the non-catalytic Cas1 monomers (K12, K259) are positioned to contact the phosphate backbone of the mid-repeat DNA (Fig. 3.1b,c) (Wang et al., 2015). Charge-swap mutations of these residues reduce or eliminate acquisition of new spacers *in vivo*, confirming their importance for the CRISPR integration reaction (Fig. 3.5a).

Although earlier work suggested that inverted sequence motifs in the repeat might form a cruciform structure during target recognition, our structure shows that the center of the repeat remains a canonical duplex at this intermediate stage of integration (Arslan et al., 2014; Babu et al., 2010; Nuñez et al., 2015b). Although the inverted repeat sequences are critical for spacer acquisition, we found no evidence of sequence-specific contacts in these motifs (Fig. 3.1c) (Goren et al., 2016; Moch et al., 2017; Wang et al., 2016). Contacts between the mid-repeat DNA and the integrase proteins are limited to nonspecific backbone interactions, with no regions of Cas1 or Cas2 positioned to interrogate either the major or minor groove. To test for contacts in solution, we performed hydroxyl radical footprinting of the half-site substrate bound by the complex (Fig. 3.1d). Protection of the backbone is clearly seen in the protospacer, including in the single-stranded end where the DNA binds in a channel of Cas1. Only weak protection occurs near the ends of the repeat on the non-integrated target strand and largely does not overlap with the inverted repeats. Several hypersensitive nucleotides are apparent at the beginning of the second inverted repeat even in the absence of protein, suggesting that these nucleotides exhibit increased flexibility or a distorted conformation in solution. Although direct sequence readout could involve a distinct but transient binding mode prior to half-site integration, our data suggest that integrase recognition of the repeat sequence likely relies on a mechanism other than base-specific hydrogen-bonding.

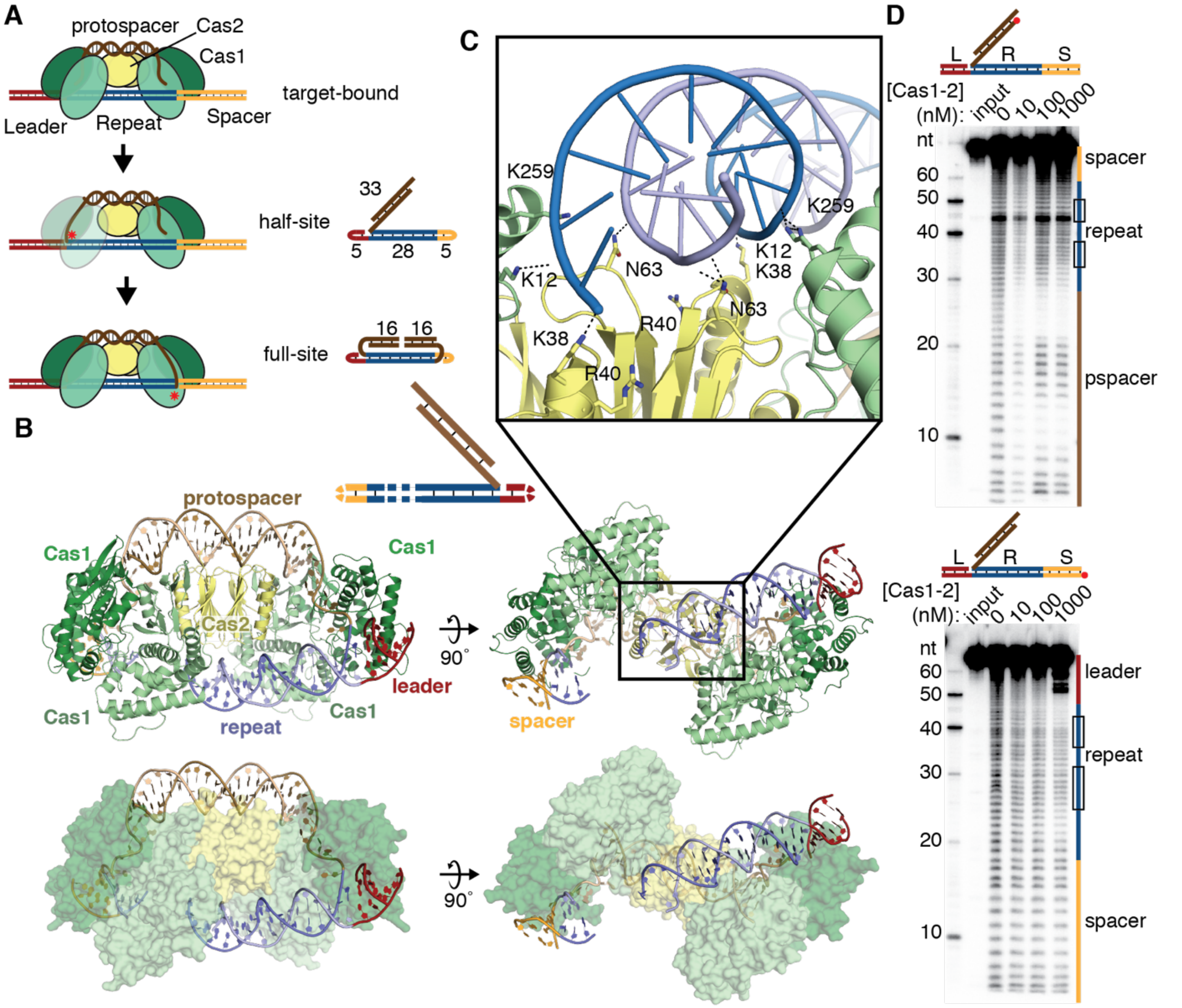


Figure 3.1 | Half-site binding by Cas1-Cas2. (**a**) Cartoon of steps of integration by Cas1-Cas2. Crystallography substrates are shown next to the corresponding reaction intermediate, with nucleotide lengths indicated. Red stars represent integration events. (**b**) Cartoon and surface representations of half-site substrate bound by Cas1-Cas2. DNA is colored as in (a). A substrate schematic is shown above, with disordered regions shown as dashed lines. (**c**) Close-up of backbone interactions between Cas1-Cas2 and half-site repeat DNA. Polar contacts are shown as dotted lines. (**d**) Hydroxyl radical footprinting of radiolabeled half-site DNA. Input is untreated DNA. The substrates are shown above the gel, with the radiolabel indicated with a red circle. Regions of the gel corresponding to the leader, repeat, spacer, and protospacer (pspacer) are indicated alongside the gel. The inverted repeat regions of the repeat are shown as boxes.

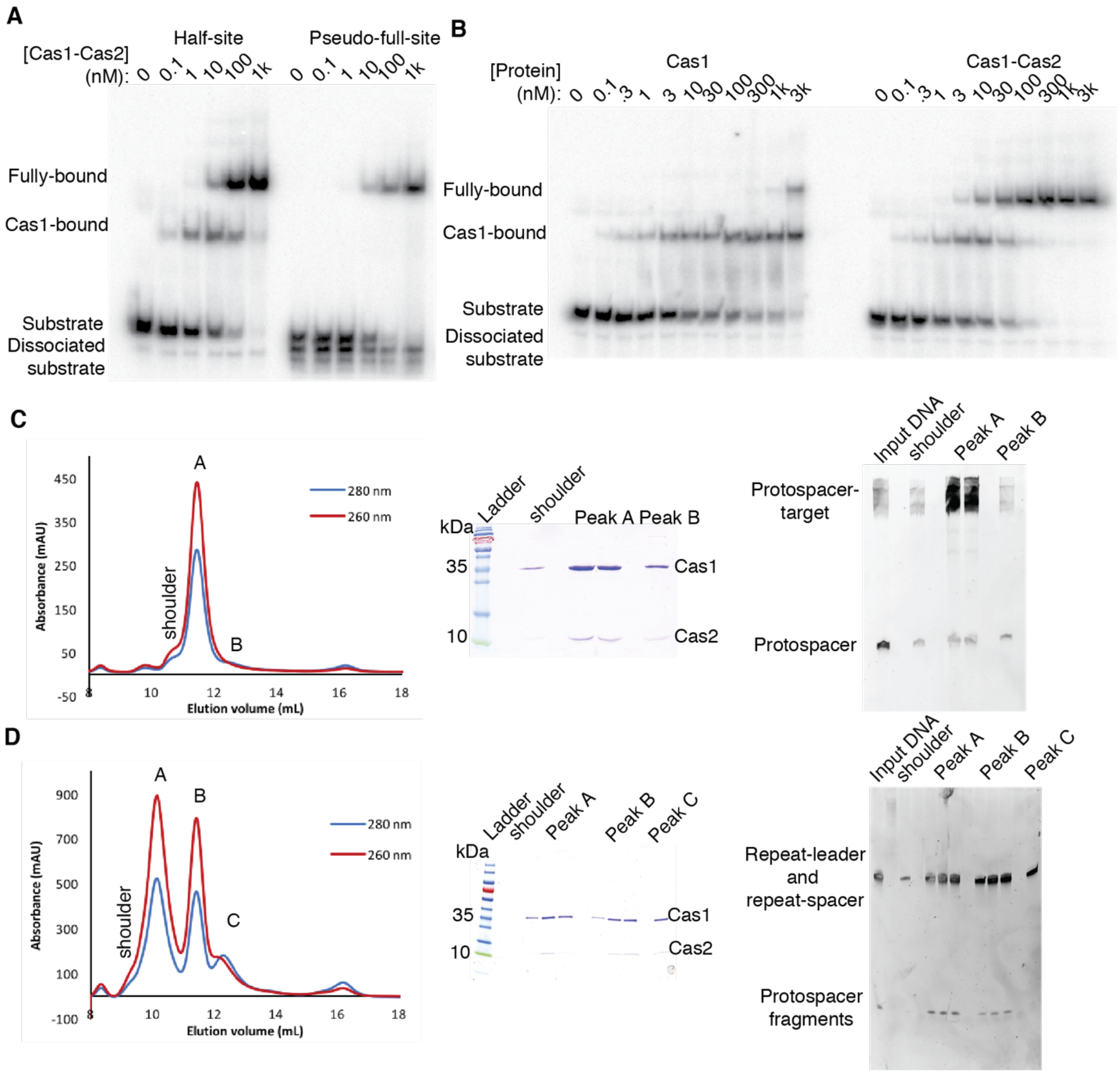


Figure 3.2 | Half-site and pseudo-full-site binding by Cas1 and Cas1-Cas2. (**a**) EMSA of radiolabeled half-site and pseudo-full-site substrates with Cas1-Cas2. Protein concentrations are indicated, and bands are labeled. (**b**) EMSA of half-site substrate by either Cas1 alone or Cas1-Cas2. The intermediate band observed in half-site binding appears to result from binding of free Cas1 dimer. (**c**) Purification of half-site-bound complex by size exclusion chromatography. A representative S200 size exclusion trace is shown. Samples were taken from the labeled peaks and analyzed on SDS-PAGE with Coomassie brilliant blue and urea-PAGE with SybrGold. The smear for the protospacer-target DNA strand results from partial renaturation of the hairpin structure. Peak A was used for crystallography. (**d**) Purification of pseudo-full-site-bound complex by size exclusion chromatography, with gels as described for (c). Peak B was used for crystallography.

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

One crystal was used for each structure.

Values in parentheses are for highest-resolution shell.

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

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

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

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

Table 3.1 | X-ray data collection and refinement statistics.

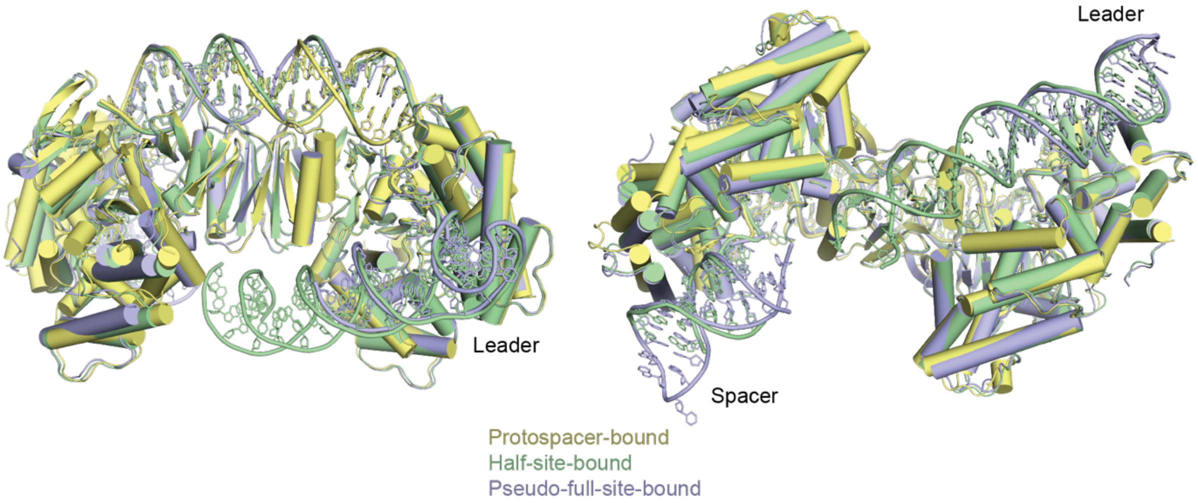


Figure 3.3 | Superposition of protospacer-bound, half-site-bound, and pseudo-full-site-bound Cas1-Cas2. Structural alignment of our target-bound structures with a previously-solved protospacer-bound structure (PDB code 5DS5). Alignments were made using the Cas2 dimer as the reference. The protospacer-bound structure is shown in yellow, half-site bound in green, and pseudo-full-site-bound in blue. Only modest structural rearrangements occur upon target binding, with the spacer-side Cas1 dimer (left side) rotating slightly to position the active site closer to the central channel.

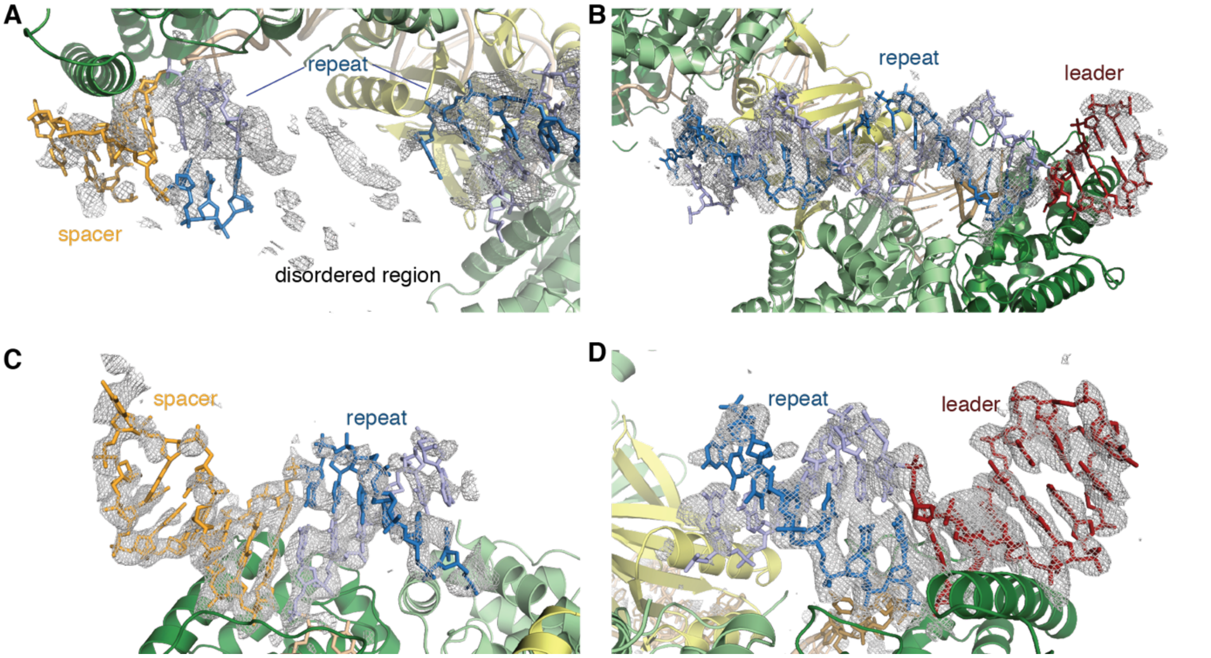


Figure 3.4 | Simulated annealing omit maps for Cas1-Cas2 bound to half-site and pseudo-full-site DNA. (**a**) F_o-F_c omit map for the entire target DNA using half-site map and model, showing the leader and early to mid-repeat DNA. (**b**) F_o-F_c omit map showing the repeat-spacer junction and the unresolved region. (**c**) F_o-F_c omit map of the target DNA using the pseudo-full-site map and model. Leader-repeat region is shown. (**d**) F_o-F_c omit map showing the spacer-repeat region. Maps are contoured at 2.0 σ.

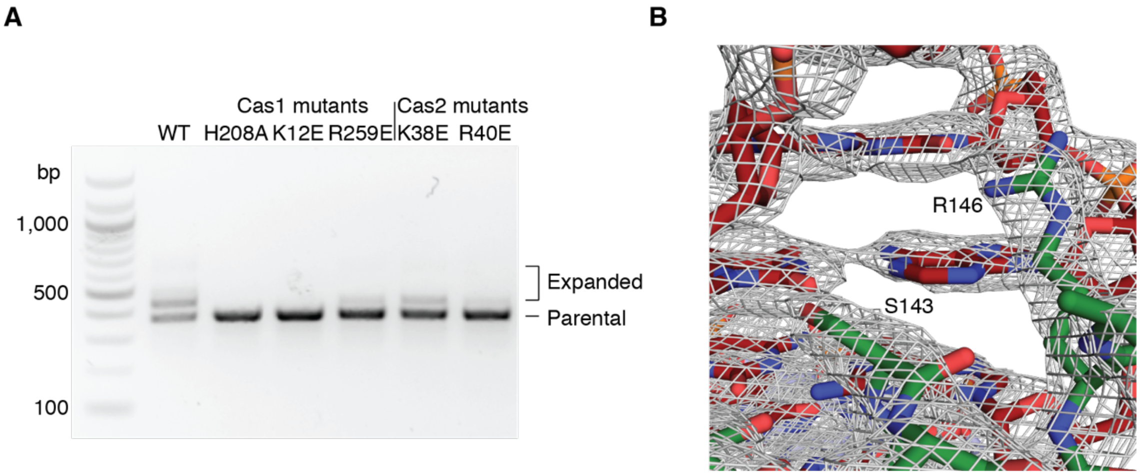


Figure 3.5 | Residues involved in mid-repeat and leader interactions. (**a**) Agarose gel of *in vivo* acquisition assays performed with the indicated Cas1 or Cas2 mutants. Cas1 H208A is used as a negative control. (**b**) Feature-enhanced map of the leader and interacting residues. Map is shown as mesh at 2.0 σ.

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

The pseudo-full-site-bound structure was solved at 2.9 Å and reveals more details of the interaction between Cas1 and the target DNA (Table 3.1). The nucleotides at both the leader-adjacent and spacer-adjacent integration sites are clearly resolved, while the middle of the repeat was disordered, suggesting that the repeat disengages from Cas2 following full integration (Fig 3.4c,d, 3.6a). Previous crystal structures suggested that the Cas1 α-helix 7 might interact with target DNA, and we indeed observe insertion of this helix into the minor groove of both the leader and spacer regions of the target DNA (Fig. 3.6b) (Nuñez et al., 2015a). The terminal residues of the leader sequence contribute to integration efficiency, and our structure reveals that several residues make hydrogen bonds with the minor-groove face of leader bases (McGinn and Marraffini, 2016; Rollie et al., 2015; Wang et al., 2016; Wright and Doudna, 2016). Cas1 R146 hydrogen bonds with A-3 and T-4 and is essential for integration *in vivo*, suggesting that it may also stabilize binding through interactions with the phosphate backbone (Fig. 3.5b, 3.6b,c). Cas1 S143 interacts with T-3 of the non-integrated target strand, though it is dispensable for *in vivo* activity (Fig. 3.5b, 3.6b,c).

3.3.3 Integration requires DNA distortion

Notably, both the half-site and the pseudo-full-site structures reveal significant distortion of the target DNA. The DNA exhibits a sharp kink at both integration sites, with the bases on either side of the leader-repeat and repeat-spacer junction forming a nearly 30° angle (Fig. 3.7a). The repeat-spacer junction of the half-site substrate exhibits a similar kink, which indicates that the distortion occurs not as a result of integration but instead upon Cas1-Cas2 binding to the target. Binding across the Cas2 dimer surface also forces a bend in the repeat, mostly localized to the region directly over Cas2 (Fig. 3.7b).

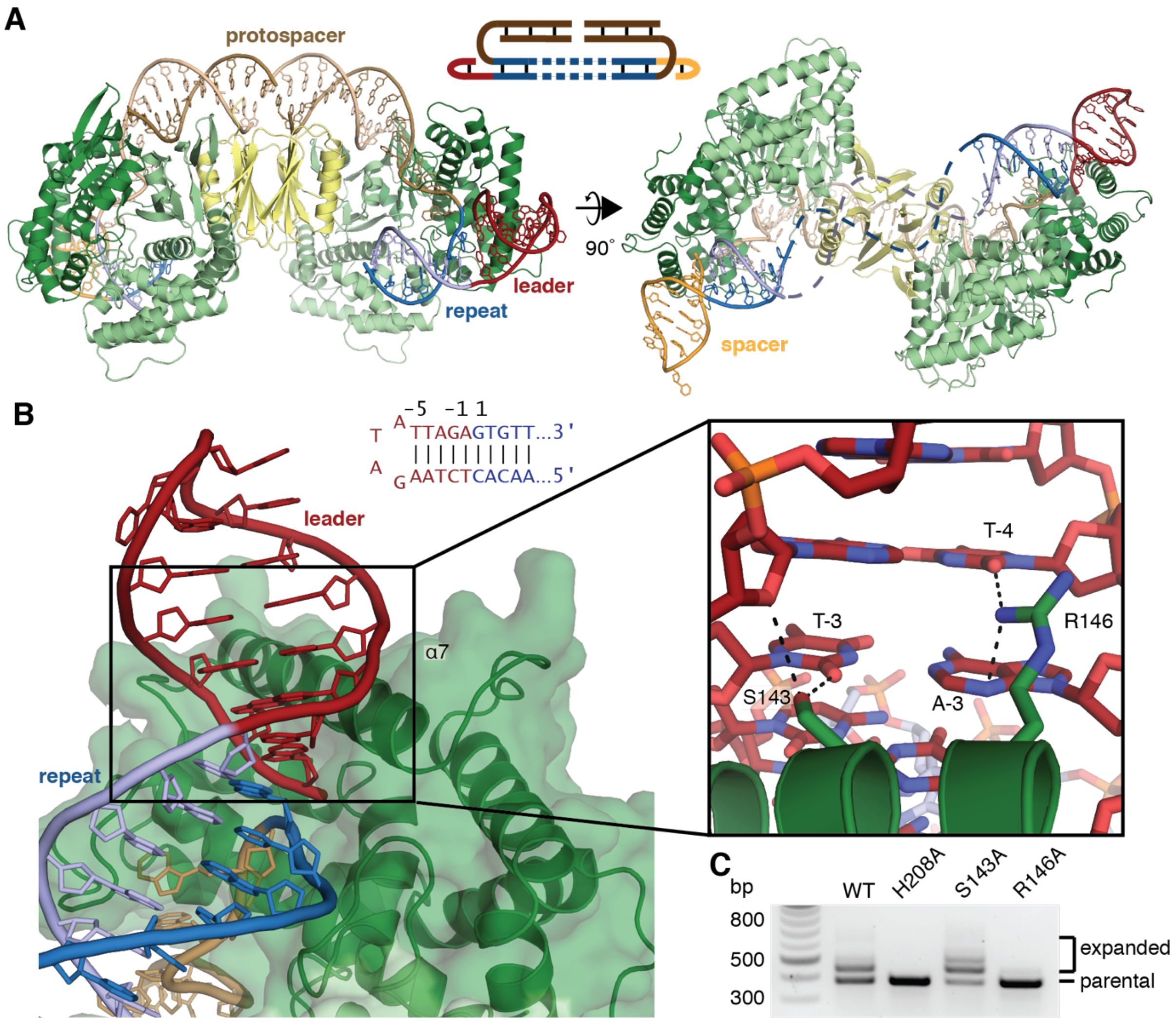


Figure 3.6 | Pseudo-full-site binding by Cas1-Cas2. (**a**) Overview of pseudo-full-site substrate binding by Cas1-Cas2. In the second view, the expected path of the disordered DNA is shown as dashed lines. A schematic of the substrate is shown, with the disordered region as dashed lines. (**b**) A view of minor groove insertion by α-helix 7. Dotted lines in close-up show polar contacts. The sequence of the leader-repeat junction and residue numbering are shown. (**c**) Agarose gel of a representative *in vivo* acquisition assay with indicated Cas1 mutants and wild-type Cas2. Acquisition results in expansion of the CRISPR array, which is visible as larger bands above the parental locus. The H208A active-site mutant is used as a negative control.

Both structures show that the repeat must also undergo twist deformation to be properly positioned in both active sites. Modeling B-form DNA into the disordered regions of the repeat results in the incorrect backbone being positioned in the spacer-side active site (Fig. 3.7c). Connecting the resolved regions of DNA requires that the missing region be under-wound by approximately one third of a turn relative to canonical B-form DNA. It is unclear how this distortion is distributed across the disordered region, and the lack of order might indicate that the DNA adopts a range of conformations to accommodate the strain. The required bending and under-winding of the repeat, together with the lack of sequence-specific contacts in the repeat, suggests that Cas1-Cas2 recognizes the target through indirect readout based on the repeat's sequence-dependent deformability. The poly-G stretches in the inverted repeat motifs in particular may facilitate the adoption of strained conformations to allow binding across both active sites (Gardiner et al., 2003; Olson et al., 1998).

To investigate whether these motifs are required for the DNA to be coordinated at opposing active sites, we performed *in vitro* integration assays using repeats with mutations known to prevent acquisition *in vivo* (Fig. 3.7d) (Goren et al., 2016). The mutations did not significantly affect leader-side integration, but they prevented integration at the repeat-spacer junction. Half-site substrates bearing the same mutations were unable to be converted to full-site products, despite supporting binding and disintegration, while wild-type half-sites were readily converted to full-site products (Fig. 3.7e, 3.8). These results confirm that the repeat sequence is important not for binding and recruitment of Cas1-Cas2 but instead for determining the ability of the target to reach the spacer-side active site.

To further investigate the importance of DNA deformation for spacer-side integration, we performed integration assays using targets with single- or double-base mismatches between the inverted repeats (Fig. 3.7f). The introduction of a mismatch is expected to disrupt the DNA duplex and generate a flexible hinge in the middle of the repeat. Mismatches immediately before the second inverted repeat increased the rate of spacer-side integration, indicating that increasing the deformability of the repeat at specific sites enhances full-site integration. These data support the model that sequence-dependent distortion is necessary for recognition and integration at the repeat. Notably, both G→C and G→A transitions in the inverted repeats prevented full-site integration, suggesting that the necessary deformation of the repeat depends on factors other than or in addition to GC content, such as specific purine-pyrimidine steps in the region where mismatches favor integration.

3.3.4 Active site geometry

To better understand Cas1 active site geometry, we grew pseudo-full-site-bound crystals in the presence of Ni^{2+}, which does not support catalysis but should allow for Mg^{2+}-like coordination geometry, and solved the structure to 3.3 Å resolution (Fig. 3.9, Table 3.1). We observed density and peaks in the anomalous difference map for a single Ni^{2+} located at each of the four Cas1 active sites, though the metals are at lower occupancy in the substrate-engaged active sites, potentially due to lower solvent accessibility at these sites (Fig. 3.10a, 3.11a-d). At the non-catalytic active sites, the metal is coordinated by H208 and D221, as previously described (Nuñez et al., 2015a; Wiedenheft et al., 2009). In the post-integration active sites, the phosphate of the newly-formed phosphodiester bond bridging the protospacer and the repeat coordinates the metal, and the free 3' OH of the cleaved leader or spacer is in close proximity. E141, which has been annotated as a metal-coordinating residue, had poor side-chain density in all monomers and appeared to be outside the range of a favorable interaction with the metal (Fig. 3.11e,f). The absolute requirement of E141 for activity suggests that it may play another role in catalysis, perhaps acting as a proton donor for the leaving 3' hydroxyl (Nuñez et al., 2014; Wiedenheft et al., 2009).

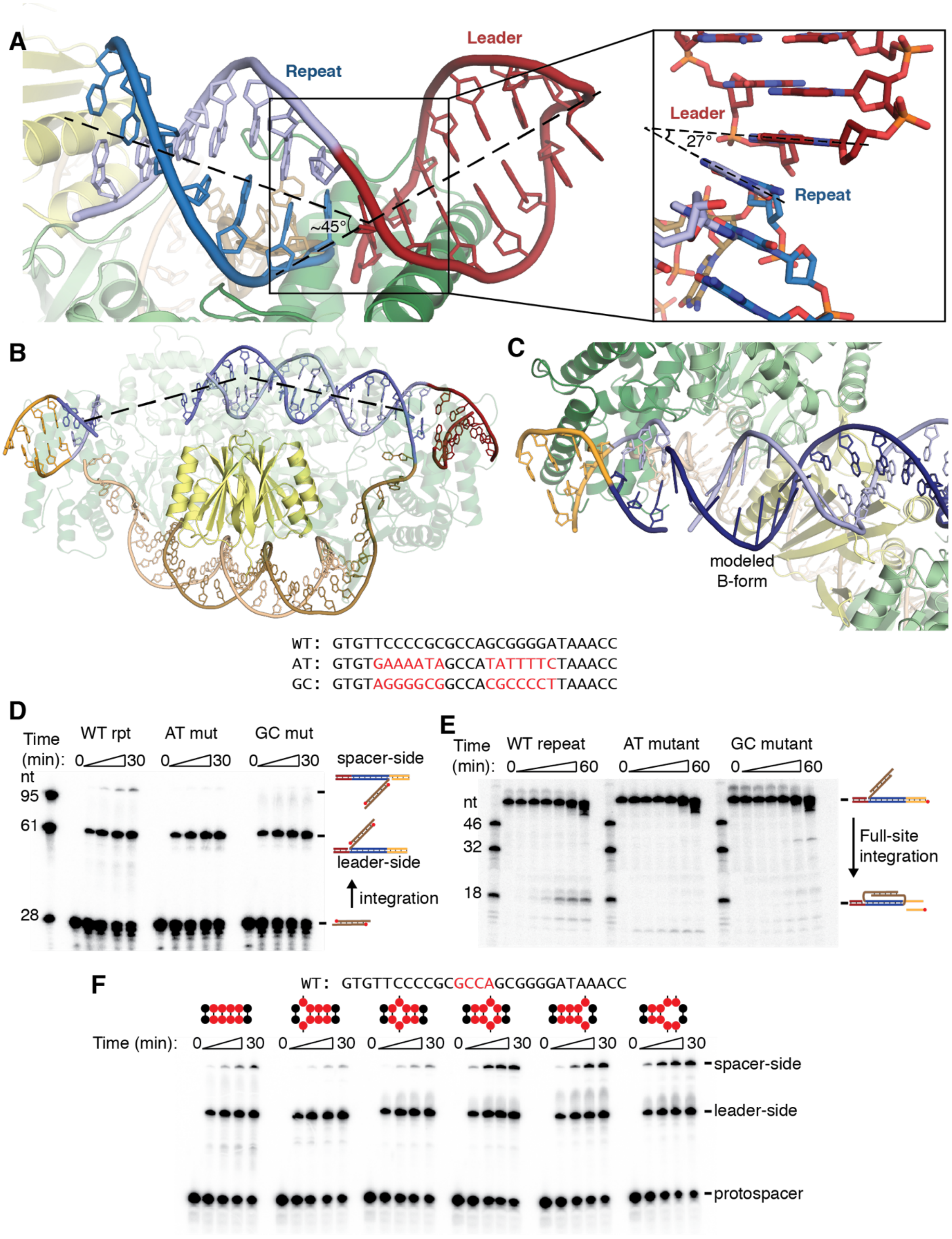


Figure 3.7 | Integration involves DNA distortion. (**a**) View of kink introduced at leader-repeat junction. The kink in the pseudo-full-site structure is highlighted with a dashed line showing the central axis of the DNA. The inset shows the bases before and after the integration site. Part of the backbone is omitted for clarity, and the angle formed by adjacent bases is shown with dashed lines. (**b**) Representation of the half-

site repeat bending over the Cas2 dimer. The DNA trajectory is fit with a dashed line to show the localized bending. (**c**) Modeled B-form DNA fails to connect resolved regions of the half-site repeat. Modeled bases are shown with bases as sticks rather than rings. The (+) strand and (–) strand are shown in dark and light blue, respectively, to show that the modeled DNA does not properly join with the spacer-proximal DNA. (**d**) Urea-PAGE gel of integration assay with radiolabeled protospacer. The substrate and expected products are shown as cartoons with the radiolabel represented with a red circle. Their expected positions are indicated. The repeat sequences are shown above, with the mutated regions highlighted in red. Timepoints were taken at 0, 1, 5, 15, and 30 minutes. (**e**) Urea-PAGE gel of second-site integration assay using mutant repeat sequences. The substrate and expected product are schematized with the radiolabel indicated with a red circle, and their expected positions are indicated on the gel. The mutant repeats are the same as in (d). Timepoints were taken at 0, .5, 1, 2, 10, 30, and 60 mintues. (**f**) Integration assay with radiolabeled protospacer and mismatched repeats. Mismatches were introduced in the region of the repeat highlighted in red in the wild-type sequence above the gel. The positions of the mismatches are schematized above each time course, with the red circles representing the highlighted mid-repeat nucleotides. Timepoints were taken at 0, 1, 5, 15, and 30 minutes.

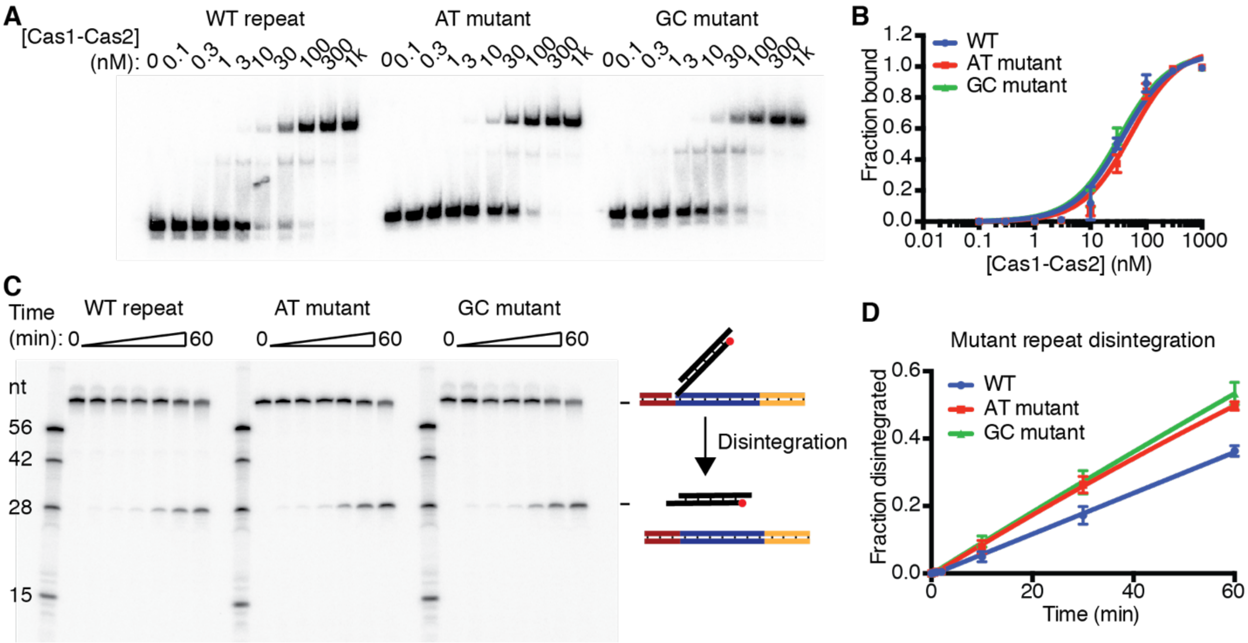


Figure 3.8 | Half-site substrates with mutant inverted repeats support robust binding and disintegration. (**a**) Representative gel showing EMSA of radiolabeled half-site substrates with wild-type repeats or mutant repeats as described in the text. (**b**) Quantification of EMSAs of half-site substrates. Mean and standard deviation of three independent experiments are plotted. (**c**) Representative urea-PAGE gel of disintegration assays performed with radiolabeled wild-type or mutant repeat half-site substrates. Substrate and expected product are schematized, with the radiolabel indicated as a red circle. (**d**) Quantification of disintegration assays of half-site substrates. Mean and standard deviation of three independent experiments are plotted, and rates were fitted as a pseudo-first-order reaction.

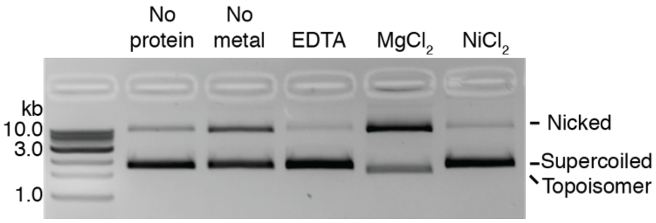


Figure 3.9 | Nickel does not support integration. Agarose gel of integration assay with plasmid target. Integration results in the generation of nicked and toposiomerized plasmids, the latter of which run ahead of the supercoiled plasmid. EDTA or divalent cation were added as indicated.

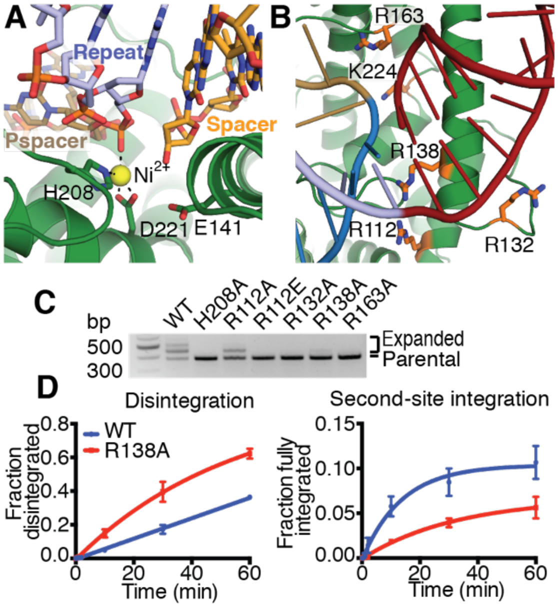


Figure 3.10 | Full-site integration requires a basic clamp around the active site. (**a**) Metal coordination in the spacer-side active site. Active site residues, repeat, spacer, and protospacer (pspacer) are labeled, and coordination is shown as dotted lines. (**b**) View of basic residues surrounding leader-repeat junction. Basic residues in close proximity to the target DNA backbone on either side of the integration site are shown as sticks and colored orange. (**c**) Agarose gel of *in vivo* acquisition assay with indicated Cas1 mutants. H208A Cas1 is used as a negative control. (**d**) Quantification of disintegration and second-site integration time-course assays by wild-type and R138A Cas1. Mean and standard deviation of three independent experiments are plotted. Representative gels are shown in Figure 3.12.

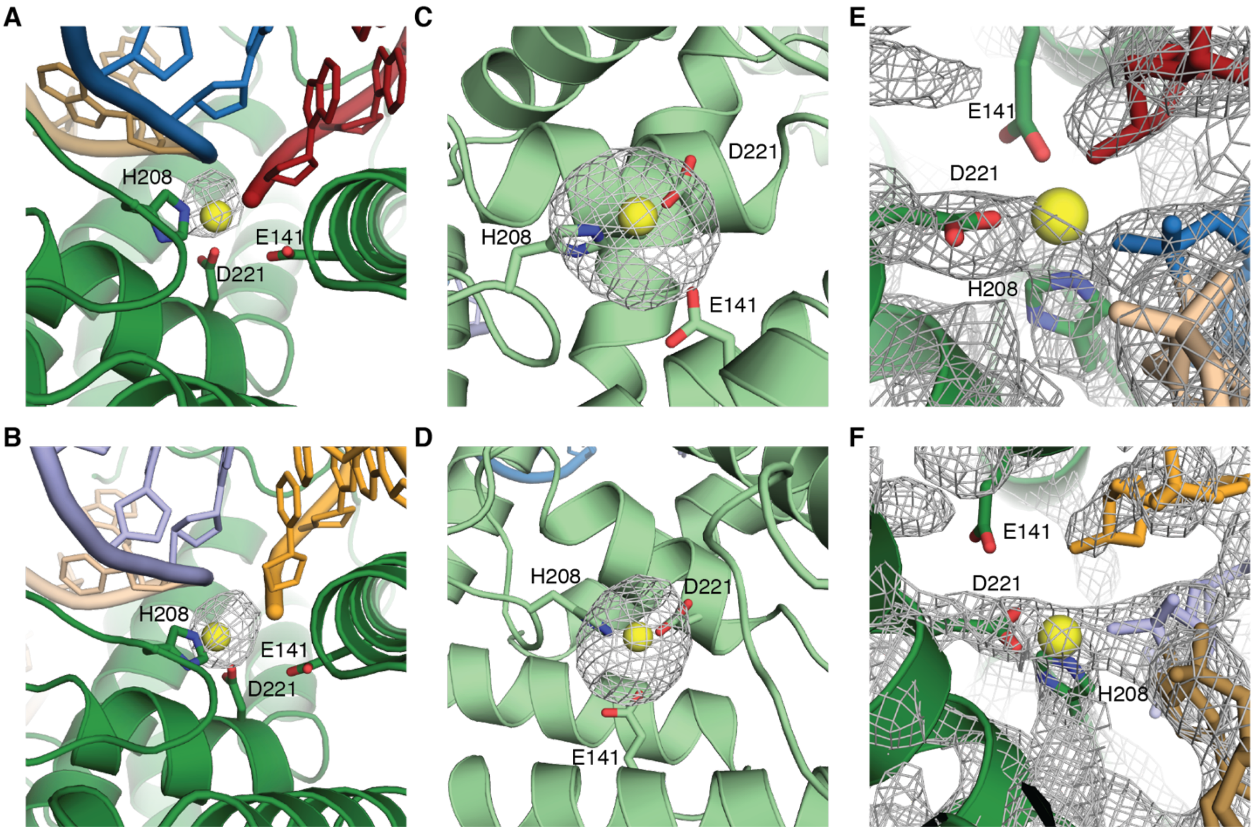


Figure 3.11 | Anomalous difference maps for active-site nickel atoms. (**a-d**) Cartoon representation of active-site nickels and anomalous density for leader-side active site (a), spacer-side active site (b), and the two non-catalytic active sites (c,d). Anomalous density is shown as a mesh contoured at 4.0 σ. Peaks in the anomalous difference map were smaller for nickel coordinated in the catalytic active sites, particularly the leader-side active sites. These Ni^{2+} were modelled with lower occupancy than those present in the non-catalytic active sites. Active site residues are shown as sticks. (**e,f**) Leader-side (e) and spacer-side (f)

active sites with feature-enhanced map. Density is shown as mesh contoured at 2.0 σ. Active site residues are shown as sticks.

In vivo CRISPR integration assays to test the role of basic residues in the integrase that might contact either side of the DNA integration site showed that alanine mutants of Cas1 R132, R138, and R163 eliminate or nearly eliminate acquisition (Fig. 3.10b,c). The R112A Cas1 mutant maintained some activity, but the R112E mutation prevented acquisition. The importance of all of these residues may reflect the need for a strong network of favorable contacts to capture the DNA in a strained conformation. To test this hypothesis, we performed disintegration and second-site integration assays with an R138A Cas1 mutant. This mutation reduced the rate of second-site integration by 50%, but R138A Cas1 exhibited wild-type-like binding and enhanced disintegration activity, likely due to faster product release or the reduced rate of the competing forward reaction (Fig. 3.10d, 3.12). These data confirm that R138 is dispensable for catalysis but is important for trapping the DNA at the distal active site.

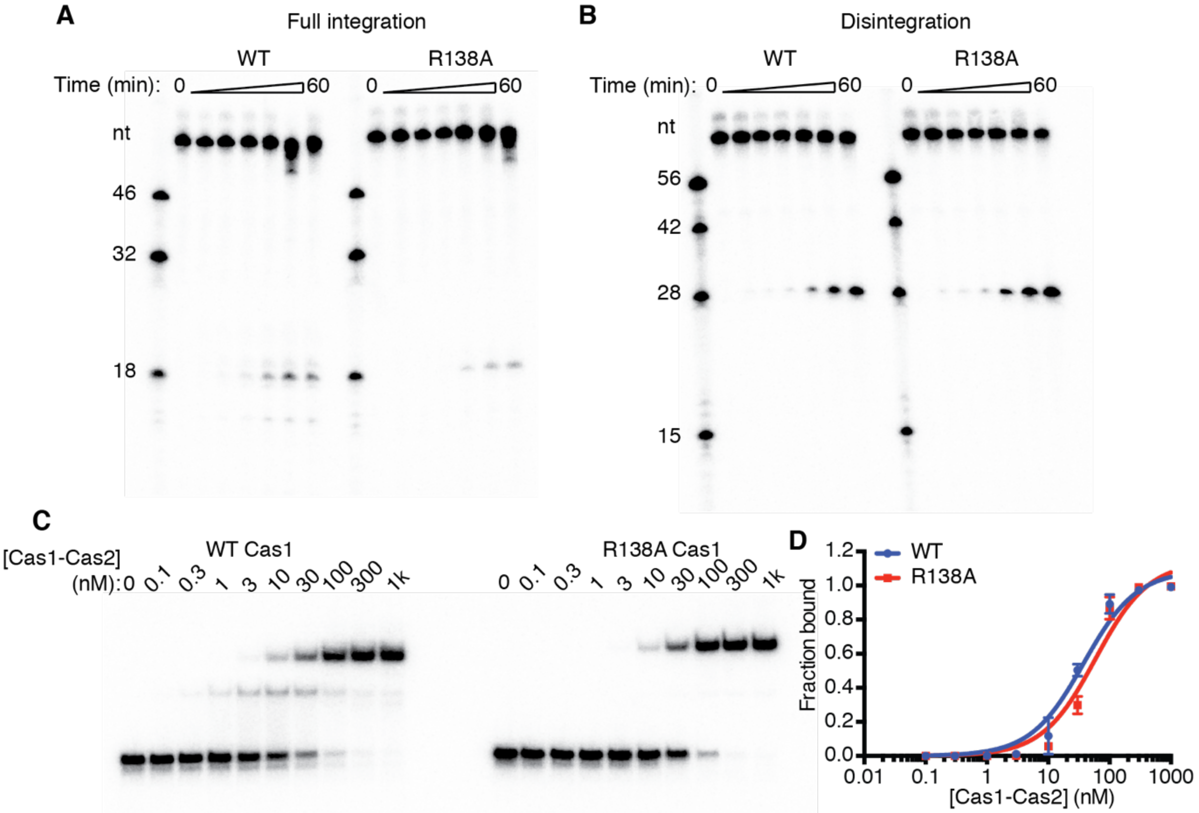


Figure 3.12 | Raw gels and target binding for R138A Cas1. (**a**) Representative urea-PAGE gel of a full-site integration assay with half-site substrate and WT or R138A Cas1. (**b**) Representative urea-PAGE gel of disintegration assay with half-site substrate and WT or R138A Cas1. (**c**) Representative native PAGE gel of EMSA with half-site substrate and WT or R138A Cas1. (**d**) Quantification of half-site substrate binding by WT or R138A Cas1-Cas2. Mean and standard deviation of three independent experiments are shown.

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

To investigate the mechanism by which IHF recruits Cas1-Cas2 to the leader-proximal repeat, we purified the Cas1-Cas2 and IHF bound to a half-site substrate with an extended leader sequence (Fig. 3.13). Negatively stained samples were used to generate an initial low-resolution reconstruction that showed additional density attached to Cas1-Cas2 module that we could assigned to IHF (Fig. 3.14). We then used cryo-EM to solve the structure at a final resolution of 3.6 Å (Fig. 3.15-3.17, Table 3.2). We generated a complete model of the Cas1-Cas2-IHF-DNA holo-complex by first fitting the crystal structure of half-site-bound Cas1-Cas2 solved in this work and the published atomic model of the IHF module (PDB:1IHF) into the cryo-EM map, followed by manually rebuilding the models to fit the density. The DNA substrates were manually built ab initio and the resulting complete model was improved by real-space refinement (Fig. 3.18).

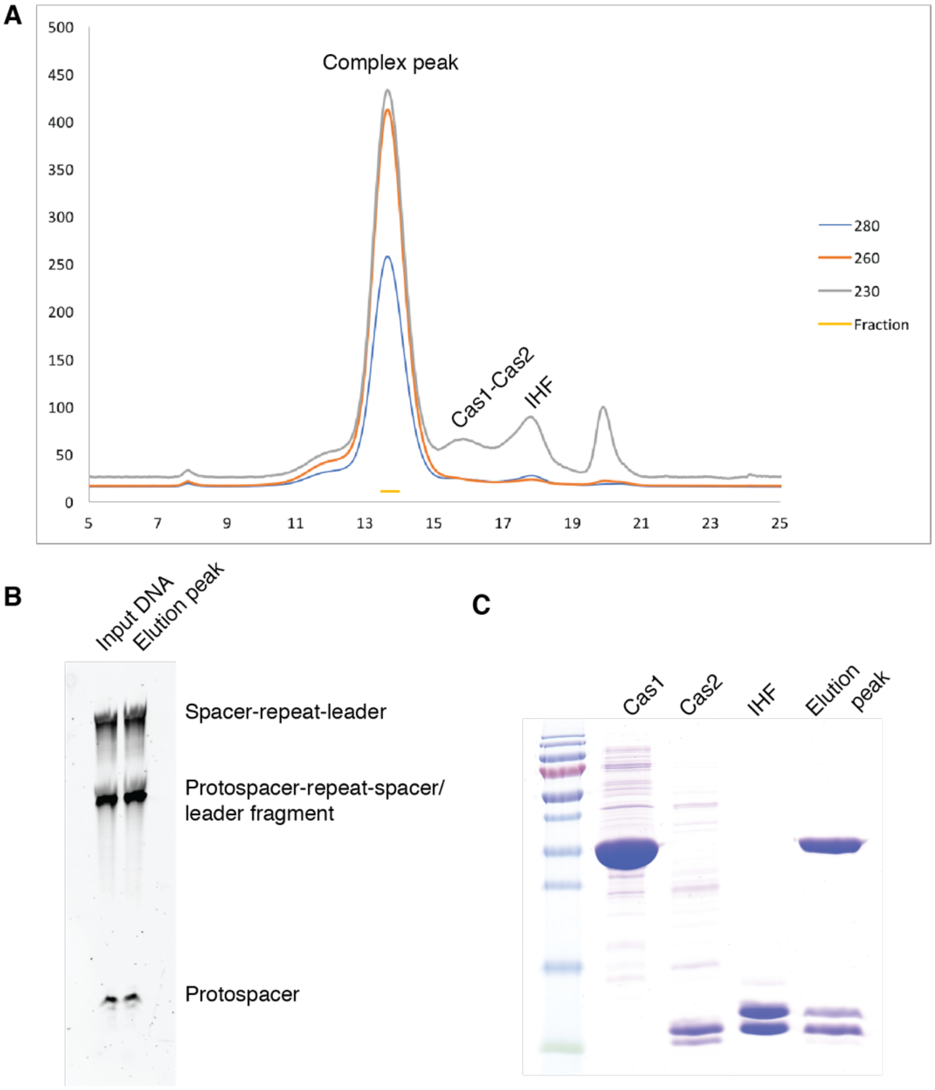


Figure 3.13 | IHF-Cas1-Cas2-DNA complex formation. (**a**) Elution profile of IHF-Cas1-Cas2-DNA complex purified over Superose 6 10/300 column. The collected fractions are indicated with "Fraction." (**b**) Urea-PAGE of the input DNA and the collected fractions. Strands are annotated. (**c**) SDS-PAGE of input proteins and the elution peak.

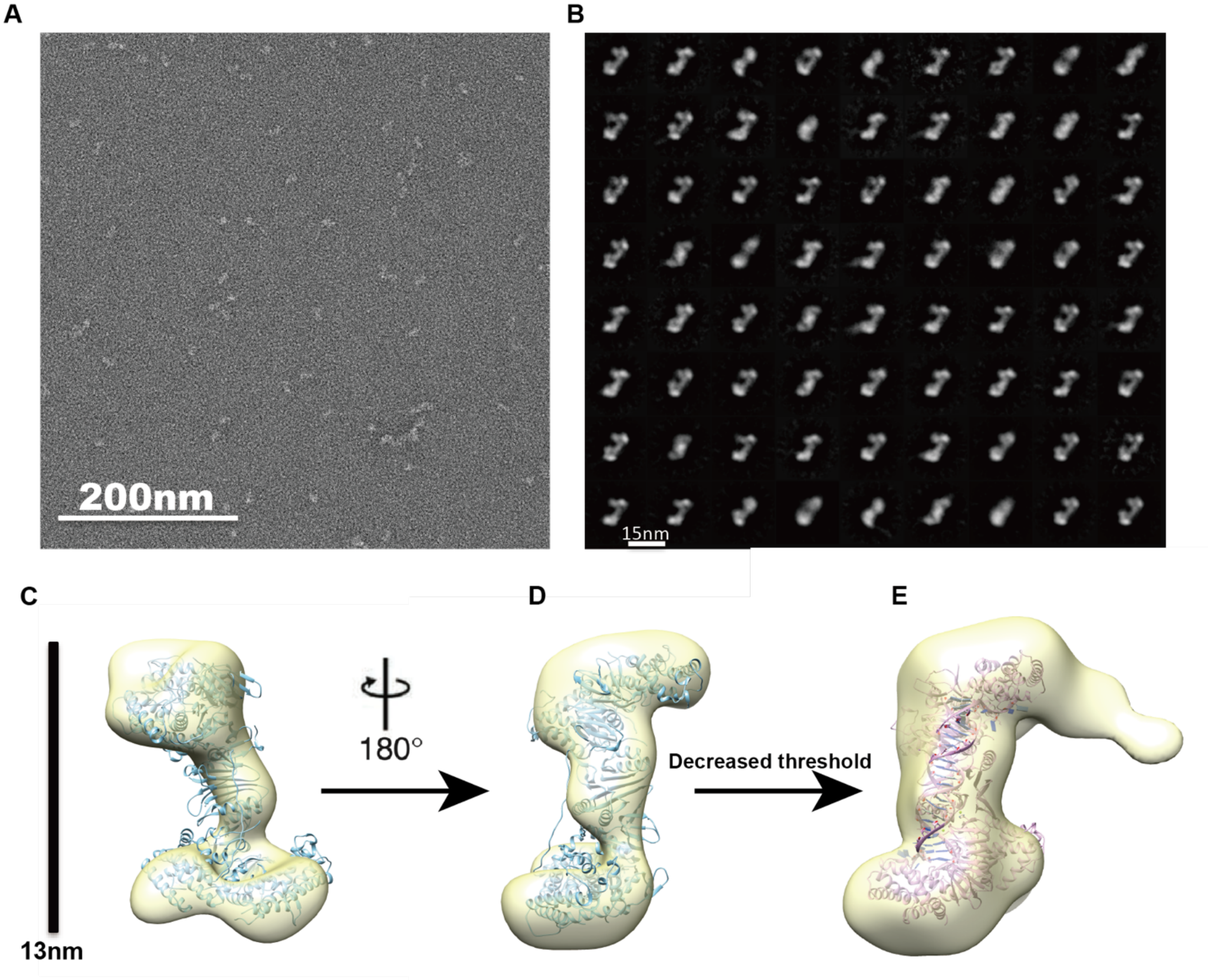


Figure 3.14 | Negative staining screening of Cas1-Cas2-DNA-IHF complex. (**a**) Raw image of Cas1-Cas2-DNA-IHF complex by negative staining. The scale bar is 200nm. (**b**) Reference-free 2D class-averages of Cas1-Cas2-DNA-IHF complex by negative staining. 72 class-averages are shown in the panel. The scale bar is 15nm. (**c-d**) 3D refined model of Cas1-Cas2-DNA-IHF complex by negative staining. Two orientations of the EM map (at the threshold of 5 σ) aligned with atomic model of Cas1-Cas2 complex (PDB code 4p6i) are shown in panel (c) and (d). The EM map in the threshold of 3 σ aligned with Cas1-Cas2-DNA atomic model (PDB code 5ds5) was presented in the third panel (d).

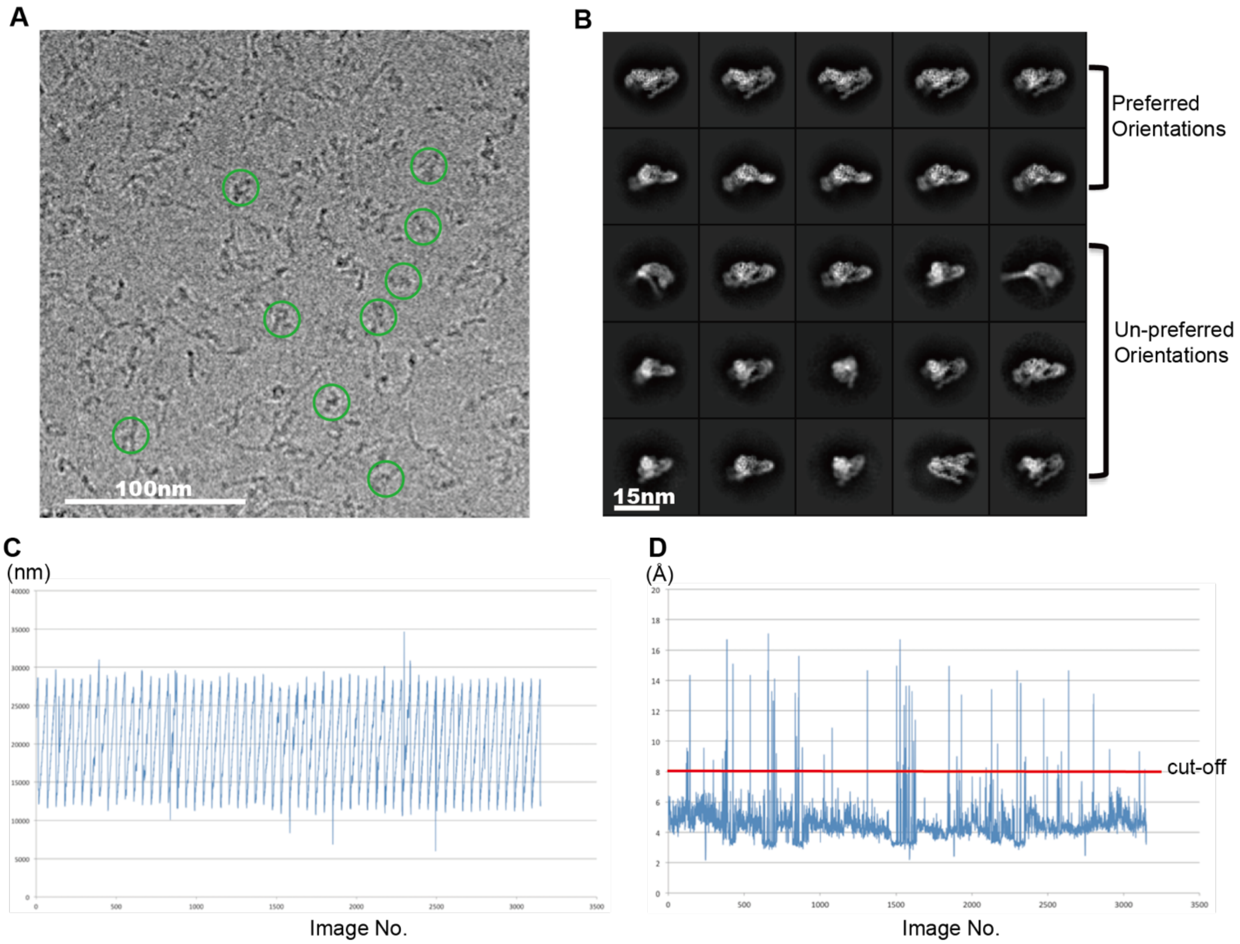


Figure 3.15 | Cryo-EM data analysis of Cas1-Cas2-DNA-IHF complex. (**a**) Drift-corrected image of Cas1-Cas2-DNA-IHF complex by cryo-EM. The scale bar is 100nm. Several particles are marked with green circles. (**b**) Reference-free 2D class-averages of Cas1-Cas2-DNA-IHF complex by cryo-EM. 25 class-averages are shown in the panel. The upper two panels show the averages in preferred orientations. The following 3 panels show the averages in un-preferred orientations. The scale bar is 15nm. (**c**) The defocus value statistic of the whole data set. These values were calculated by CTFFIND4. (**d**) The maximal resolution statistic of the whole data set. These values were calculated in Relion2.0 based on signal intensity in micrograph at different resolutions. The red line indicates the cut-off resolution for micrograph sorting. Only the micrographs with signal more than 8 Å were kept for 2D and 3D analysis.

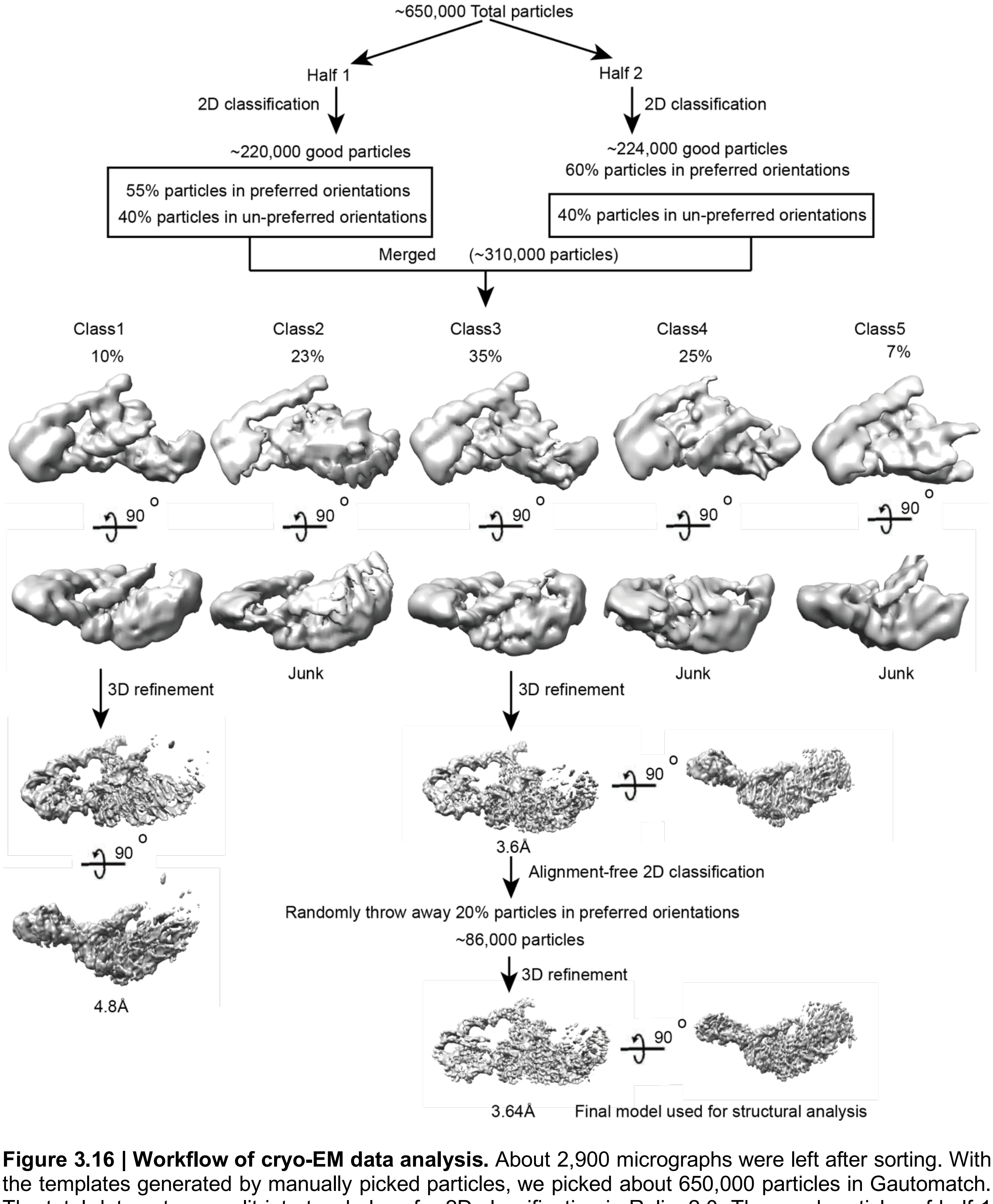

~650,000 Total particles
Half 1
Half 2
2D classification
2D classification
~220,000 good particles
~224,000 good particles
60% particles in preferred orientations
55% particles in preferred orientations
40% particles in un-preferred orientations
40% particles in un-preferred orientations
Merged (~310,000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

Figure 3.16 | Workflow of cryo-EM data analysis. About 2,900 micrographs were left after sorting. With the templates generated by manually picked particles, we picked about 650,000 particles in Gautomatch. The total data set was split into two halves for 2D classification in Relion2.0. The good particles of half 1 and good particles in un-preferred orientations of half 2 were merged and classified into five 3D classes in Cryosparc with the initial model generated by negative staining. Two views of each 3D model are shown. The particle percentage of each class is also presented. Two good classes were further refined in Cryosparc. For class1, the reported resolution of the 3D refined model was 4.8 Å. For class3, the reported resolution of the 3D refined model was 3.6 Å. To further reduce the anisotropic resolution issue introduced by redundant particles in preferred orientations, we performed further alignment-free 2D classification based on the orientation information defined by the previous 3D refinement. 20 percent of particles in preferred orientations were discarded for further 3D refinement, which gave rise to a better isotropic map with the resolution of 3.64 Å.

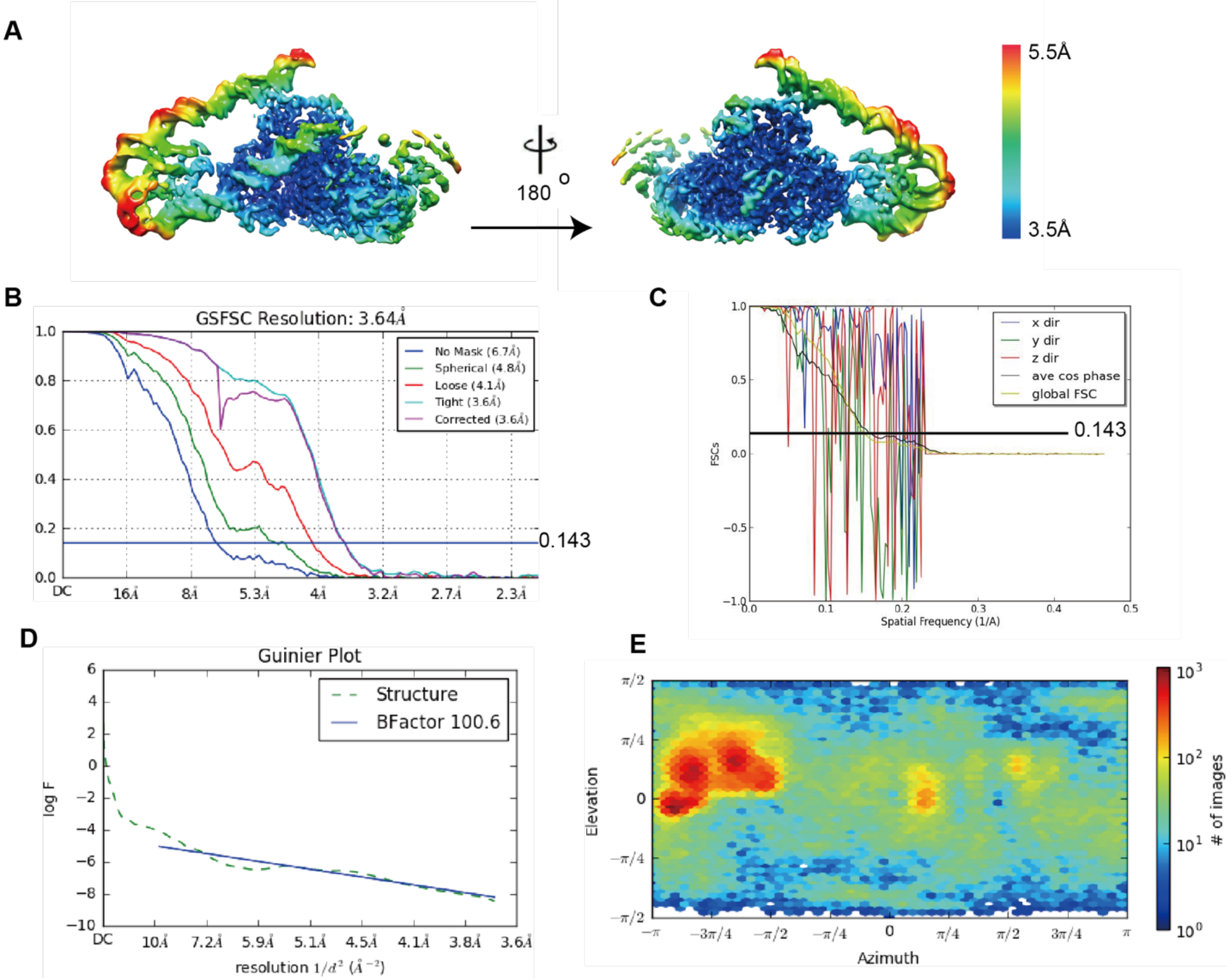


Figure 3.17 | Validation of EM 3D model. (**a**) Cryo-EM structure of the Cas1-Cas2-DNA-IHF used for model building was shown and colored by local resolution calculated in Relion2.0. Resolution ranges from 3.5 Å to 5.5 Å. (**b**) The Fourier shell correlation (FSC) curve calculated using two independent half maps, indicating an overall resolution of 3.64Å. The panel was the standard output from Cryosparc. (**c**) The Fourier shell correlation (FSC) curve along x, y and z directions calculated by ThreeDFSC using two independent half maps. The panel is the standard output of ThreeDFSC. (**d**) The standard output of Guinier Plot for the sharpened model by Cryosparc. The B-factor used for sharpening is 100.6. (**e**) The Euler angle distribution of refined dataset. The panel is the standard output from Cryosparc.

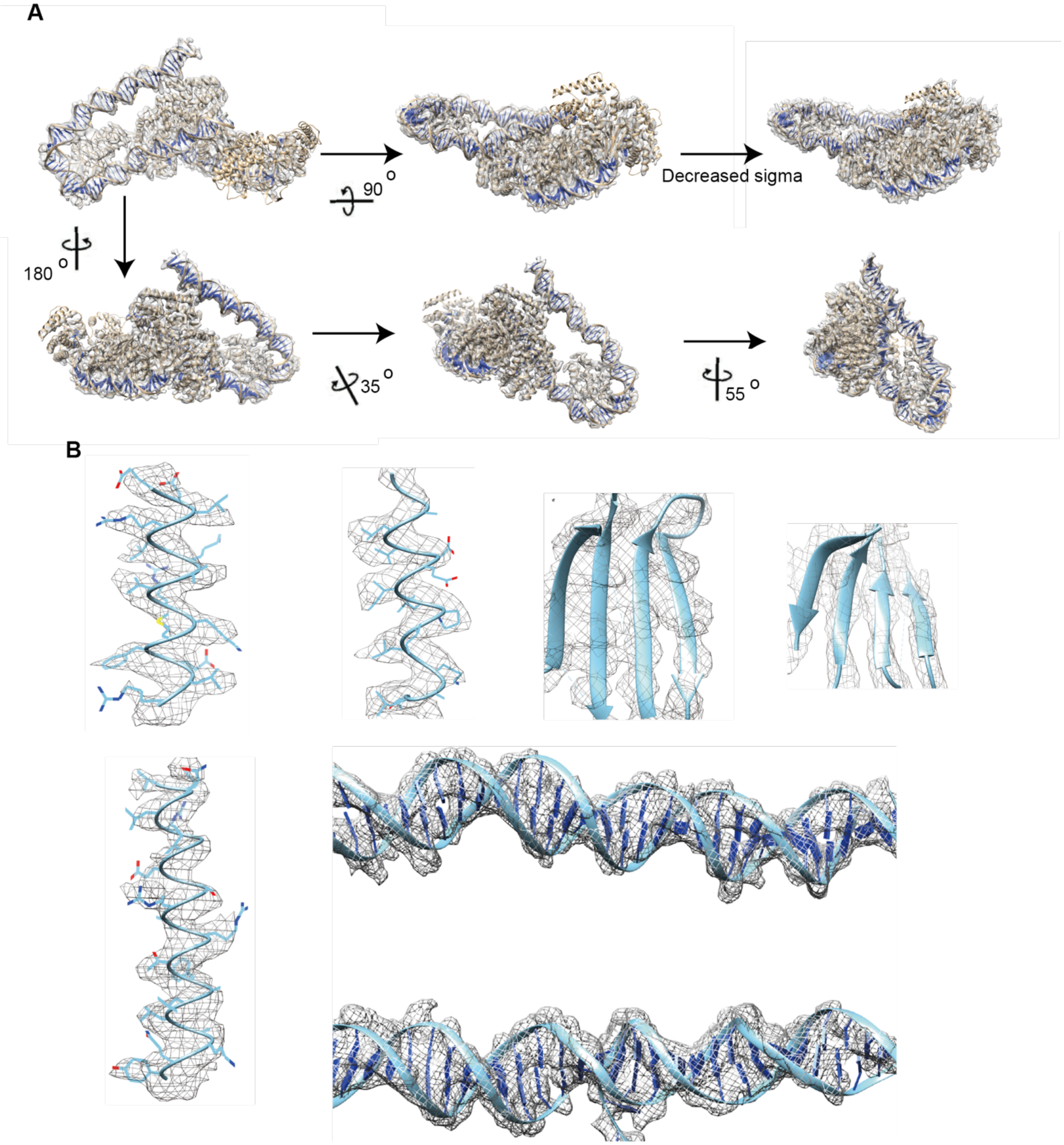


Figure 3.18 | Atomic model building of Cas1-Cas2-DNA-IHF complex. (**a**) The EM density at the threshold of 8.5 σ was aligned with the atomic model and shown in different orientations. The EM density at the threshold of 5.5 σ was shown on the top right panel, which gave more visible density for the flexible Cas1 unit. (**B**) Representative regions of the EM density map of Cas1-Cas2-DNA-IHF complex, into which the atomic model was built.

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

Table 3.2 | EM data collection and model refinement statistics of Cas1-Cas2-IHF-DNA complex

Compared to the holo-complex, Cas1-Cas2 and the repeat are overall in the same conformation as in the half-site crystal structure, and disorder of the spacer end of the complex again prevented building the DNA across to the distal active site. The structure shows how IHF binds the leader immediately upstream of Cas1-Cas2 and induces a 180° turn in the DNA, directing it back toward the Cas1-Cas2 complex (Fig. 3.19a) (Rice et al., 1996). The upstream binding motif interacts with one of the non-catalytic Cas1 protomers, with the loop between $\alpha6$ and $\alpha7$ inserting into the minor groove. R117 and Q136 interact with the phosphate backbone, and R131 and R132 are positioned to hydrogen bond with the minor groove face of bases in the conserved recognition region (Fig. 3.19b). R132 is essential for integration *in vivo*, but it is difficult to assess the importance of its role in upstream readout given that R132 on the catalytic Cas1 protomer is implicated in the basic clamp described above (Fig. 3.10b, 3.19c). R131 and Q136 also contribute significantly to DNA binding, as alanine mutations of either reduce acquisition. Mutation of the conserved upstream sequence as a block eliminated acquisition, as previously noted, and single nucleotide mutations revealed G-53, which is recognized by R131, as particularly important for recognition (Fig. 3.19d) (Yoganand et al., 2017).

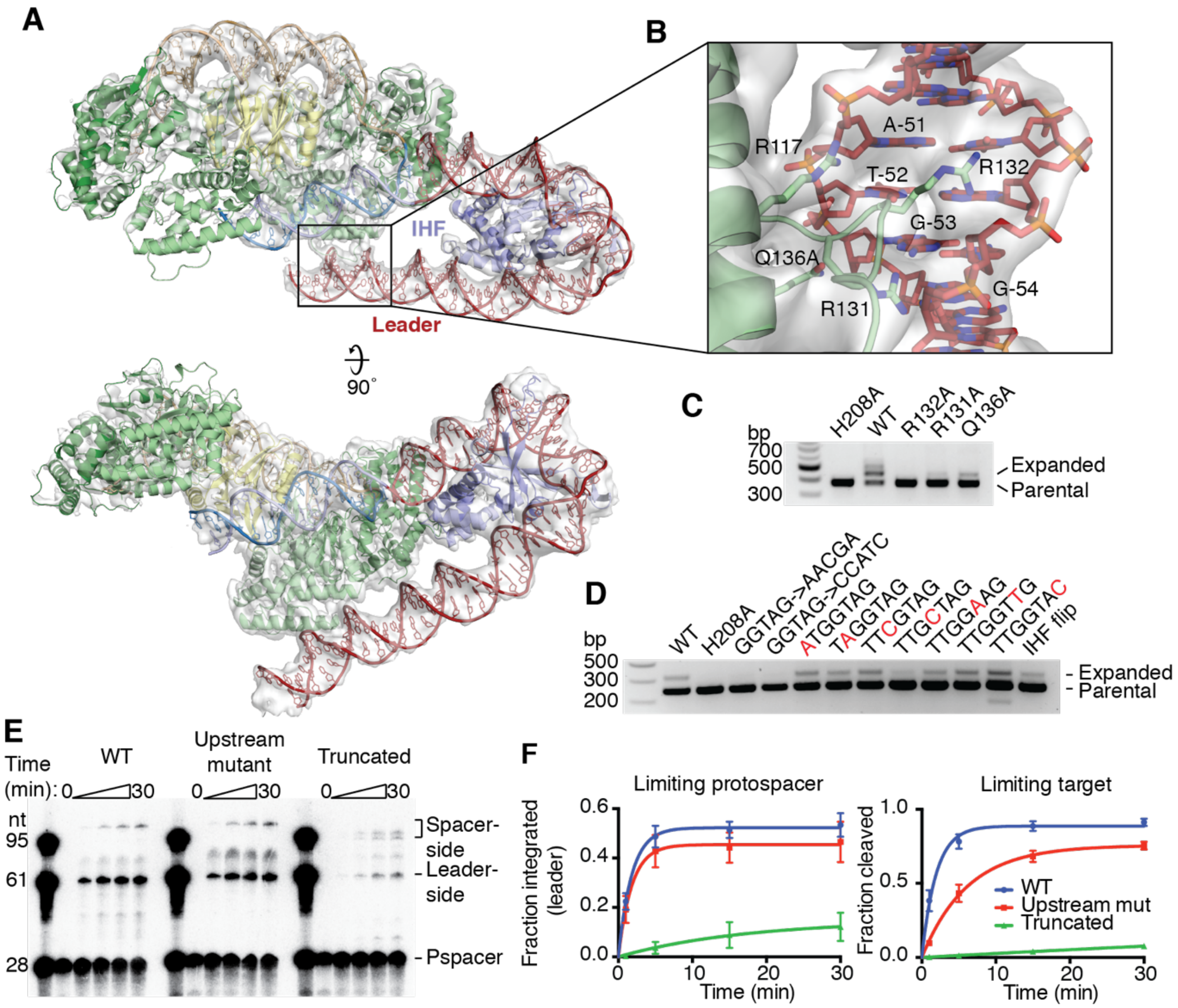


Figure 3.19 | Upstream sequence recognition by Cas1. (**a**) Cryo-EM structure of Cas1-Cas2 with IHF and extended leader. The atomic model is shown as a cartoon, and the electron density is shown as a transparent surface. Density is shown using an 8 σ threshold. (**b**) View of upstream sequence readout by Cas1. Electron density is shown as a transparent surface using an 8 σ threshold. Relevant Cas1 residues are labeled. Bases in the conserved recognition sequence are labeled, with numbering such that the final residue of the leader is -1. (**c**) Acquisition assay with wild-type Cas2 and the indicated Cas1 mutants. H208A Cas1 is used as a negative control. (**d**) Acquisition assay with wild-type proteins and the noted mutations in the leader sequence. Single-nucleotide mutations in the conserved recognition region are highlighted in red. “IHF flip” denotes the leader sequence with the IHF binding sequence reversed in place. H208A Cas1 is used as a negative control. (**e**) Integration assay with radiolabeled protospacer and targets with variable leaders. “Upstream mutant” substrate has the “GGTAG→CCATC” mutation in the conserved recognition motif, while the “Truncated” substrate begins at residue -46, after the recognition motif. Time points were taken at 0, 1, 5, 15, and 30 minutes. (**f**) Quantification of integration assays with limiting protospacer and limiting target. Mean and standard deviation of three independent experiments are shown. A representative gel of the limiting target experiment is shown in Figure 3.20.

To determine how much the IHF-dependent recruitment of Cas1-Cas2 depends on upstream sequence recognition as opposed to nonspecific stabilizing interactions, we performed *in vitro* integration assays with targets containing leaders with mutations in the upstream binding region or leaders truncated prior to the upstream interaction region (Fig.

3.19e). Mutations in the binding site reduced the rate of leader-side integration three-fold when target is limiting (Fig. 3.19f, 3.20). The rate effect is masked when the target is in excess over protospacer-bound complex, but a higher level of off-target integration is observed (Fig. 3.20). The increased importance of the upstream sequence for *in vivo* acquisition suggests that it may be important for initial identification of the target in context of genomic DNA, while it is dispensable when the correct target is saturating and no competitor is present. Truncation of the leader had a much more significant effect, with the rate of leader-side integration reduced ~100-fold when target was limiting (Fig. 3.19f). Spacer-side integration was also affected by the truncation, as indicated by the appearance of a second band consistent with misplaced integration within the repeat (Fig. 3.19e). These results show that nonspecific interactions with the leader DNA are critical for robust Cas1-Cas2 activity and specificity, while the sequence-specific interactions aid in efficient recognition.

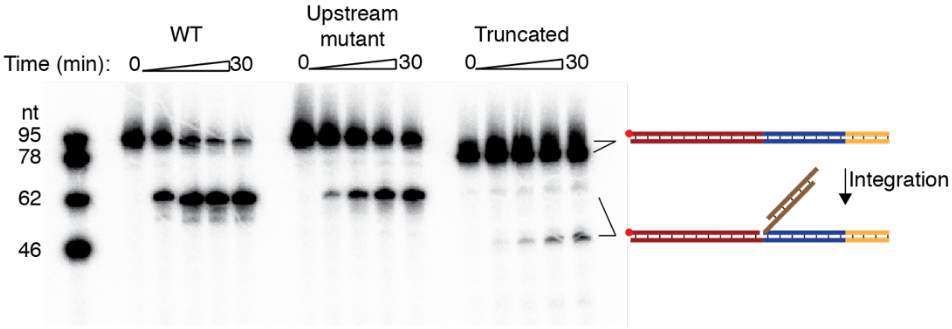


Figure 3.20 | Integration with limiting target. Urea-PAGE of a representative integration assay using unlabeled protospacer and a target with the top strand labeled. The substrate and expected product are shown as cartoons, with the radiolabel indicated with a red circle. The expected positions of the substrate and product bands are shown. Substrates are the same as in Fig. 3.19e. Time points were taken at 0, 1, 5, 15, and 30 minutes.

3.3.6 Suppression of off-target integration by IHF

We also investigated whether IHF contributes to Cas1-Cas2 recruitment by mechanisms other than juxtaposition of the upstream binding site. Our structure reveals that Cas1 and the alpha protomer of IHF (IHF-α) are in close proximity, with a solvent-inaccessible surface of 200 Å^2 between the two proteins (Fig. 3.21a). However, there is no significant continuous electron density between the proteins. Mutations of IHF-α residues near the interface with Cas1 identified E10 and D14 as important for acquisition (Fig. 6b). These residues might interact favorably with Cas1 R131 or R132 to aid in Cas1 recruitment. However, reversing the orientation of the IHF binding site in the leader, which should position IHF-β rather than IHF-α to interact with Cas1, did not severely impact acquisition, suggesting that any interaction that occurs is not highly specific (Fig. 3.19d).

To further investigate the role of IHF, we performed integration assays with and without IHF, using a truncated leader to prevent contribution from upstream interactions (Fig. 3.21c). In the absence of IHF, off-target integration occurs in the leader, demonstrating a role for IHF in limiting spurious integration events. Shifting the IHF binding site one to five nucleotides farther away from the leader-repeat junction led to a modest decrease in the efficiency of leader-side integration, though the site of integration was unaltered (Fig. 3.21c,d). This supports the model that contacts between IHF and Cas1 contribute to specific and efficient CRISPR locus expansion, though recruitment of the upstream binding site appears to be the more important contribution.

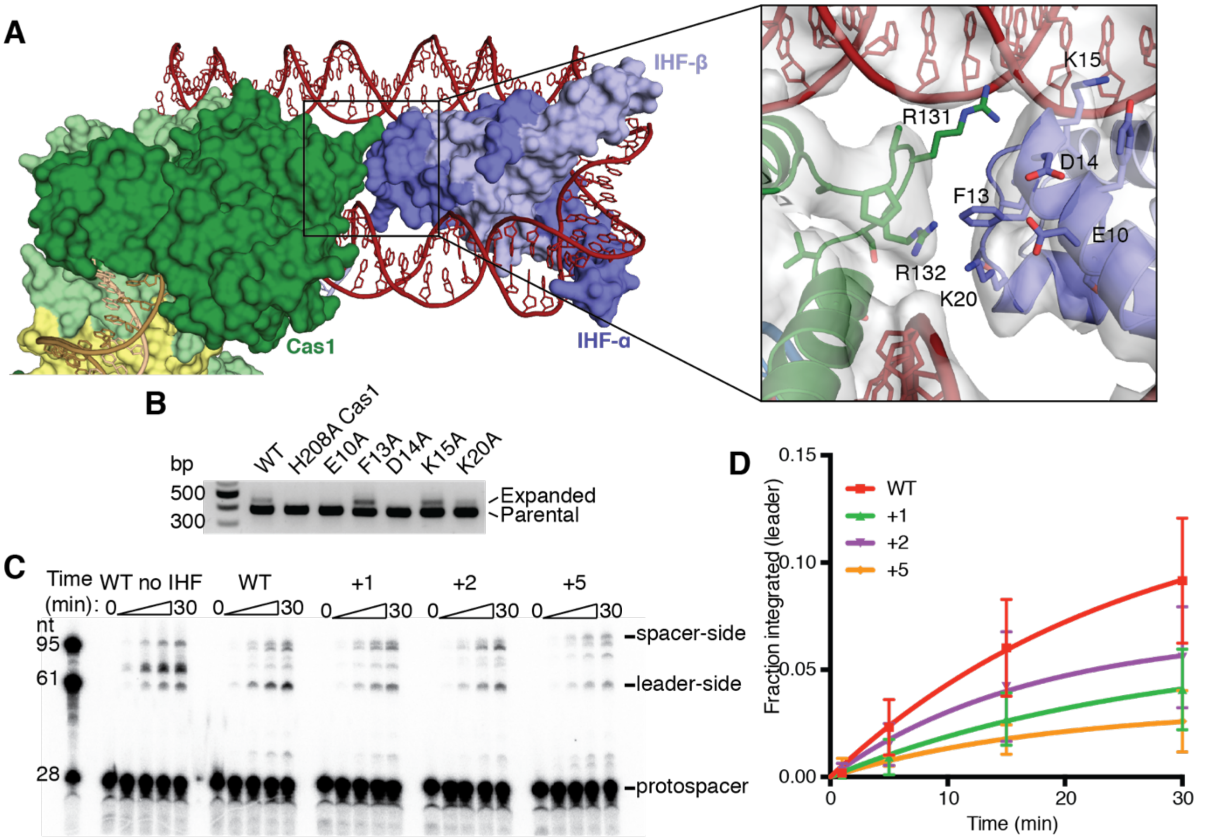


Figure 3.21 | Interactions between Cas1 and IHF. (**a**) Surface and cartoon representations of the interface between Cas1 and IHF-α. In the inset, residues at the interaction surface are shown as sticks, and residues of interest are labeled. Electron density is shown as a surface with an 8 σ threshold. (**b**) Acquisition assay with wild-type Cas1 and Cas2 and the indicated IHF-α mutants. H208A Cas1 is used as a negative control. (**c**) Integration assays with radiolabeled protospacer and targets with truncated leaders. IHF is included unless otherwise noted. Mutant substrates have 1, 2, or 5 base pairs inserted between the IHF recognition sequence and the Cas1 recognition sequence of the leader. Time points were taken at 0, 1, 5, 15, and 30 minutes. (**d**) Quantification of leader-side integration with radiolabeled protospacer and truncated targets. Mean and standard deviation of three independent replicates are shown.

3.4 Conclusions

These data show that the type I Cas1-Cas2 from *E. coli* relies heavily on active site positioning and structural features of the DNA, rather than direct sequence recognition, to localize DNA integration to the CRISPR locus (Fig. 3.22). The ability of the DNA substrate duplex to access both Cas1 active sites regulates recognition of the CRISPR repeat, with the GC-rich inverted repeats allowing for twist deformation while the mid-repeat sequence acts as a hinge, and IHF aids in recruitment at the leader by providing a secondary binding surface for the complex. The lack of direct sequence recognition might reflect the evolutionary origins of Cas1 as a more promiscuous transposase (Béguin et al., 2016; Hickman and Dyda, 2015; Krupovic et al., 2014). DNA target site bending is a common feature in transposases and integrases, where it disfavors the disintegration reaction by ejecting DNA from the integrase active sites once integration is achieved (Maertens et al., 2010; Montaño et al., 2012). While Cas1-Cas2 may use a similar mechanism, as suggested by the displacement of the mid-repeat upon full-site integration, CRISPR systems appear to have exploited the requirement for DNA bending to provide sequence specificity for the integration reaction. The role played by IHF also represents a surprising variation on a feature sometimes seen in transposases. In both λ and μ phage mobilization pathways, IHF, or the related protein HU, are involved in bringing recognition sequences on the viral DNA into contact with the integrase (Laxmikanthan et al., 2016; Montaño et al., 2012). Notably, in the phage pathways, IHF aids in the recognition of donor DNA, while in CRISPR acquisition it is important for recognition of the target DNA, highlighting the shift in substrate selectivity from donor to target that was essential for the "domestication" of Cas1 for use in immunity (Béguin et al., 2016; Krupovic et al., 2014).

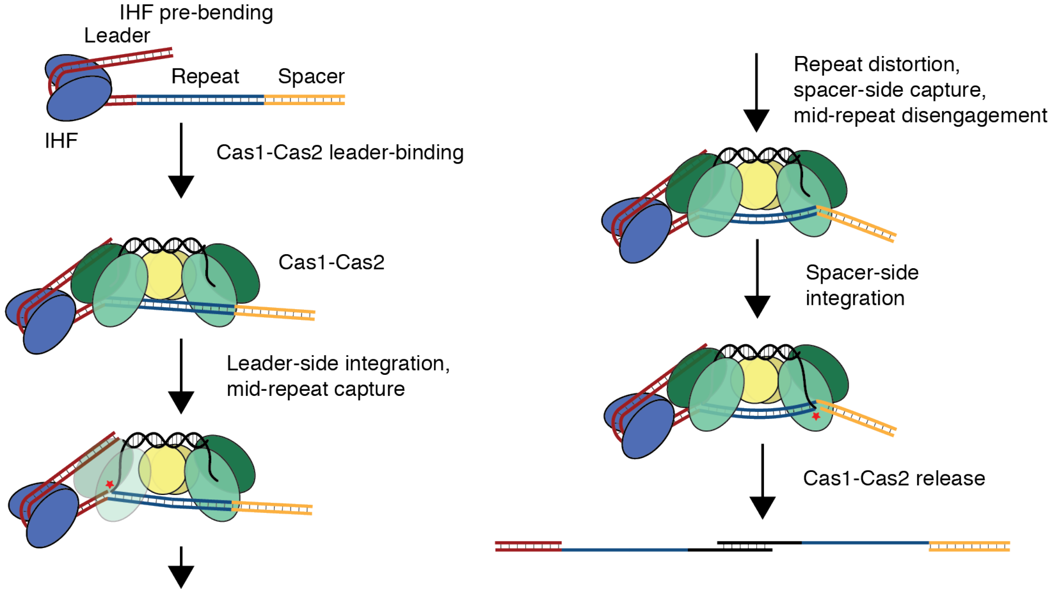


Figure 3.22 | Model for repeat recognition and integration by Cas1-Cas2. IHF binding the leader sequence creates a doubled-over DNA structure that allows for simultaneous recognition of the leader sequence and the upstream recognition motif by Cas1-Cas2. Direct sequence readout is largely restricted to this initial recognition. The spacer side of the repeat is flexible at this point, but may be captured by basic residues in the channel formed by Cas1-Cas2. Capture of the repeat-spacer junction requires sequence-dependent distortion of the repeat, allowing off-target integration events to be halted at the half-site step. Integration of the protospacer occurs at both ends of the repeat, though it is unclear whether integration at the leader end precedes stable binding at the spacer end. Following integration, Cas1-Cas2 must release the product to allow for repair of the repeat. The mechanism for product release remains to be discovered.

The unique substrate preferences of the CRISPR integrase could make it useful as a molecular recording device for barcoding genomes or generating locus-specific sequence insertions (Shipman et al., 2016). Bacterial transposases including Tn5 and MuA provide robust tools for DNA tagging, insertion and deletion, but they are promiscuous in their target selection and require sequence-specific interactions with the donor DNA that limit their use in some systems (Adey and Shendure, 2012; Goryshin et al., 1999; Nadler et al., 2016). While the CRISPR integrase shares the reaction chemistry of other transposases, its unique substrate sequence independence coupled with its selectivity for target DNA sequences may enable a complementary set of applications. The architecture of the CRISPR integration complexes presented here suggests that subtle adjustment of the distance between Cas1 active sites could reprogram the CRISPR integrase to recognize different integration target sites. Changes in integrase architecture could thereby be exploited for genome tagging applications and may also explain natural divergence of CRISPR arrays in bacteria.

3.5 Materials and methods

3.5.1 Protein and DNA preparation

Cas1 and Cas2 proteins from *E. coli* K12 (MG1655) were individually purified as previously described (Nuñez et al., 2014). IHF from *E. coli* K12 (MG1655) was purified as a heterodimer as previously described (Nuñez et al., 2016). DNA oligonucleotides were purchased from Integrated DNA Technologies or Dharmacon and were purified using urea-PAGE. DNA substrates for crystallography were prepared by mixing the appropriate ssDNA oligonucleotides in 20 mM HEPES-NaOH, pH 7.5, 25 mM KCl, 10 mM $MgCl_2$, incubating at 95°C for 5 minutes, slow-cooling to room temperature, and purifying over an 8% native polyacrylamide gel. Radiolabeled substrates were prepared by labeling with T4 polynucleotide kinase (New England Biolabs) and [γ-^{32}P]-ATP (Perkin Elmer) and annealing with a two-fold excess of the unlabeled strands, with the exception of radiolabeled protospacers, which were annealed and purified using native PAGE prior to radiolabeling the duplex. Substrates used for hydroxyl radical footprinting were further purified on an 8% native polyacrylamide gel. Sequences for all substrates are shown in Table 3.3.

3.5.2 Complex formation, crystallization, and data collection

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

X-ray diffraction data for the half-site and pseudo-full-site were collected under cryogenic conditions at beamline 8.3.1 at the Lawrence Berkeley National Laboratory Advanced Light Source with a wavelength of 1.1158 Å. Native X-ray diffraction data for the pseudo-full-site with nickel was collected under cryogenic conditions at beamline 9-2 at the Stanford Synchrotron Radiation Lightsource with a wavelength of 0.9795 Å. All data were collected with a Pilatus3 S 6M detector (Dectris). Anomalous data were collected from crystals grown in the presence of Ni^{2+} at 8345.7 eV, an energy between the inflection point and peak anomalous energies. All data were indexed in *XDS* and scaled in *XSCALE* before merging in *AIMLESS* (Evans and Murshudov, 2013; Kabsch, 2010). Resolution cut-offs were determined using correlation-coefficient threshold of 0.5 (Diederichs and Karplus, 2013).

3.5.3 Negative staining EM microscopy and image processing

Cas1-Cas2-DNA-IHF complexes were assembled by co-incubating Cas1 and Cas2 at 50 μM each in buffer containing 20 mM HEPES, pH 7.5, 150 mM KCl, 5 mM EDTA, and 1 mM DTT. IHF and half-site DNA were incubated in the same buffer at 20 μM and 10 μM, respectively. After an hour, equal volumes Cas1-Cas2 and IHF-DNA were combined, such that Cas1-Cas2 complex was in 1.25-fold excess over DNA, and allowed to complex for 30 minutes before purifying over a Superose 6 10/300 column (GE Healthcare). The complexes were diluted to a final concentration of 50~80 nM and negatively stained in a 2% (w/v) solution of uranyl acetate (Electron Microscopy Sciences) following the standard deep-staining procedure on glow-discharged holey carbon-coated EM copper grids covered with a thin layer of continuous carbon. The negatively stained specimen was then mounted onto a transmission electron microscope holder and examined by an FEI Tecnai Spirit electron microscope operated at 120-kV. Magnified digital micrographs of the specimen were automatically taken at a nominal magnification of 80,000 on a Gatan Ultrascan 4000 CCD camera with a pixel size of 1.5 Angstroms at the specimen level within Leginon. The defocus values used were about -1.0 to -1.8 μm, and the total accumulated dose at the specimen was about 60 electrons per $Å^2$. The particles were automatically picked, CTF corrected and then 2D-classified without reference in Appion (Lander et al., 2009). 10 good 2D class averages were imported into EMAN2 for generating the initial 3D model based on common line method (Tang et al., 2007). Good particles sorted by the 2D classification were further refined against the initial model with SPIDER (Shaikh et al., 2008).

Description	Sequence
Pseudo-full-site protospacer-repeat-spacer-hairpin (pseudo-full-site structure, Ni^{2+} pseudo-full-site structure)	GCTACTGGGGCCGAGGGTGTTCCCCGCGCCAGCGGGGATAAACCGAGCAGATATGCTC
Pseudo-full-site protospacer-repeat-leader-hairpin (pseudo-full-site structure, Ni^{2+} pseudo-full-site structure)	CACTGGTGGTCGCCGCGGTTTATCCCCGCTGGCGCGGGGAACACTCTAAGATATTAGA
Pseudo-full-site protospacer fragment (pseudo-full-site structure, Ni^{2+} pseudo-full-site structure, 3.2)	GCCCCAGTAGC
Pseudo-full-site protospacer fragment (pseudo-full-site structure, Ni^{2+} pseudo-full-site structure, 3.2)	GACCACCAGTG
Half-site protospacer-repeat-spacer-repeat-leader (half-site structure)	ATTTACTACTCGTTCTGGTGTTTCTCGTGTGTTCCCCGCGCCAGCGGGGATAAACCGAGCAGATATGCTCGGTTTATCCCCGCTGGCGCGGGGAACACTCTAAGATATTAGA
Protospacer strand (half-site structure, EM structure, 3.1d, 3.2, 3.7d-f, 3.8, 3.9, 3.10d, 3.12, 3.19e,f, 3.20, 3.21c,d)	AAACACCAGAACGAGTAGTAAATTGGGC
Extended leader (EM structure)	ATAAAGTTGGTAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGA
Protospacer-repeat-spacer (EM structure)	ATTTACTACTCGTTCTGGTGTTTCTCGTGTGTTCCCCGCGCCAGCGGGGATAAACCGAGCA
Full-length spacer-repeat-leader (EM structure, 3.7d,f, 3.19e,f)	TGCTCGGTTTATCCCCGCTGGCGCGGGGAACACTCTAAACATAACCTATTATTAATTAATGATTTTTTAAGCCAGTCACAATCTACCAACTTTAT
Pseudo-full-site/half-site leader fragment (3.1d, 3.2, 3.7d, 3.8, 3.10d, 3.12)	AATAATAGGTTATGTTTAGA
Half-site protospacer-repeat-spacer (3.1d, 3.2, 3.7d, 3.8, 3.10d, 3.12)	ATTTACTACTCGTTCTGGTGTTTCTCGTGTGTTCCCCGCGCCAGCGGGGATAAACCGAGCACAAATATCATCGC
Half-site spacer-repeat-leader (3.1d, 3.2, 3.7d, 3.8, 3.10d, 3.12)	GCGATGATATTTGTGCTCGGTTTATCCCCGCTGGCGCGGGGAACACTCTAAACATAACCTATTATT
Pseudo-full-site spacer fragment (3.2)	GCGATGATATTTGTGCTC
Pseudo-full-site protospacer-repeat-spacer (3.2)	CACTGGTGGTCGCCGAGGTTTATCCCCGCTGGCGCGGGGAACACTCTAAACATAACCTATTATT
Pseudo-full-site protospacer-repeat-leader (3.2)	GCTACTGGGGCCGAGGGTGTTCCCCGCGCCAGCGGGGATAAACCGAGCACAAATATCATCGC
Protospacer strand (3.7d,f, 3.9, 3.10d, 3.19e,f, 3.20, 3.21c,d)	ATTTACTACTCGTTCTGGTGTTTCTCGT
Full-length leader-repeat-spacer (3.7d,f, 3.19e,f,)	ATAAAGTTGGTAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGAGTGTTCCCCGCGCCAGCGGGGATAAACCGAGCA
AT mutant leader-repeat-spacer (3.7d)	ATAAAGTTGGTAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGAGTGTGAAAATAGCCATATTTTCTAAACCGAGCA
AT mutant spacer-repeat-leader (3.7d)	TGCTCGGTTTAGAAAATATGGCTATTTTCACACTCTAAACATAACCTATTATTAATTAATGATTTTTTAAGCCAGTCACAATCTACCAACTTTAT
GC mutant leader-repeat-spacer (3.7d)	ATAAAGTTGGTAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGAGTGTAGGGGCGGCCACGCCCCTTAAACCGAGCA
GC mutant spacer-repeat-leader (3.7d)	TGCTCGGTTTAAGGGGCGTGGCCGCCCCTACACTCTAAACATAACCTATTATTAATTAATGATTTTTTAAGCCAGTCACAATCTACCAACTTTAT
Half-site AT mutant protospacer-repeat-spacer (3.7e, 3.8)	ATTTACTACTCGTTCTGGTGTTTCTCGTGTGTGAAAATAGCCATATTTTCTAAACCGAGCACAAATATCATCGC
Half-site AT mutant spacer-repeat-leader (3.7e, 3.12)	GCGATGATATTTGTGCTCGGTTTAGAAAATATGGCTATTTTCACACTCTAAACATAACCTATTATT
Half-site GC mutant protospacer-repeat-spacer (3.7e, 3.8)	ATTTACTACTCGTTCTGGTGTTTCTCGTGTGTAGGGGCGGCCACGCCCCTTAAACCGAGCACAAATATCATCGC
Half-site GC mutant spacer-repeat-leader (3.7e, 3.8)	GCGATGATATTTGTGCTCGGTTTAAGGGGCGTGGCCGCCCCTACACTCTAAACATAACCTATTATT
Mid-repeat mismatch 1 (3.7f)	ATAAAGTTGGTAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGAGTGTTCCCCGCCCCAGCGGGGATAAACCGAGCA
Mid-repeat mismatch 2 (3.7f)	ATAAAGTTGGTAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGAGTGTTCCCCGCGGCAGCGGGGATAAACCGAGCA
Mid-repeat mismatch 3 (3.7f)	ATAAAGTTGGTAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGAGTGTTCCCCGCGCGAGCGGGGATAAACCGAGCA
Mid-repeat mismatch 4 (3.7f)	ATAAAGTTGGTAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGAGTGTTCCCCGCGCCTGCGGGGATAAACCGAGCA
Mid-repeat double-mismatch (3.7f)	ATAAAGTTGGTAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGAGTGTTCCCCGCGCGTGCGGGGATAAACCGAGCA
Upstream mutant leader-repeat-spacer (3.19e,f, 3.20)	ATAAAGTTCCATCATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGAGTGTTCCCCGCGCCAGCGGGGATAAACCGAGCA
Upstream mutant spacer-repeat-leader (3.19e,f, 3.20)	TGCTCGGTTTATCCCCGCTGGCGCGGGGAACACTCTAAACATAACCTATTATTAATTAATGATTTTTTAAGCCAGTCACAATGATGGAACTTTAT
Truncated leader-repeat-spacer (3.19e,f, 3.20, 3.21c,d)	GTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGAGTGTTCCCCGCGCCAGCGGGGATAAACCGAGCA

Truncated spacer-repeat-leader (3.19e,f, 3.20, 3.21c,d)	TGCTCGGTTTATCCCCGCTGGCGCGGGGAACACTCTAAACATAACCTATTATTAATTAATGATTTTTTAAGCCAGTCAC
Truncated +1 leader-repeat-spacer (3.21c,d)	GTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGATTTAGAGTGTTCCCCGCGCCAGCGGGGATAAACCGAGCA
Truncated +1 spacer-repeat-leader (3.21c,d)	TGCTCGGTTTATCCCCGCTGGCGCGGGGAACACTCTAAATCATAACCTATTATTAATTAATGATTTTTTAAGCCAGTCAC
Truncated +2 leader-repeat-spacer (3.21c,d)	GTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGATTTTAGAGTGTTCCCCGCGCCAGCGGGGATAAACCGAGCA
Truncated +2 spacer-repeat-leader (3.21c,d)	TGCTCGGTTTATCCCCGCTGGCGCGGGGAACACTCTAAAATCATAACCTATTATTAATTAATGATTTTTTAAGCCAGTCAC
Truncated +5 leader-repeat-spacer (3.21c,d)	GTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGATACATTTAGAGTGTTCCCCGCGCCAGCGGGGATAAACCGAGCA
Truncated +5 spacer-repeat-leader (3.21c,d)	TGCTCGGTTTATCCCCGCTGGCGCGGGGAACACTCTAAATGTATCATAACCTATTATTAATTAATGATTTTTTAAGCCAGTCAC
Leader sequences	
GGTAG->AACGA (3.19d)	AAGTACTCTTTAACATAATGGATGTGTTGTTTGTGTGATACTATAAAGTTAACGAATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGA
GGTAG->CCATC (3.19d)	AAGTACTCTTTAACATAATGGATGTGTTGTTTGTGTGATACTATAAAGTTCCATCATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGA
ATGGTAG (3.19d)	AAGTACTCTTTAACATAATGGATGTGTTGTTTGTGTGATACTATAAAGATGGTAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGA
TAGGTAG (3.19d)	AAGTACTCTTTAACATAATGGATGTGTTGTTTGTGTGATACTATAAAGTAGGTAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGA
TTCGTAG (3.19d)	AAGTACTCTTTAACATAATGGATGTGTTGTTTGTGTGATACTATAAAGTTCGTAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGA
TTGCTAG (3.19d)	AAGTACTCTTTAACATAATGGATGTGTTGTTTGTGTGATACTATAAAGTTGCTAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGA
TTGGAAG (3.19d)	AAGTACTCTTTAACATAATGGATGTGTTGTTTGTGTGATACTATAAAGTTGGAAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGA
TTGGTTG (3.19d)	AAGTACTCTTTAACATAATGGATGTGTTGTTTGTGTGATACTATAAAGTTGGTTGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGA
TTGGATC (3.19d)	AAGTACTCTTTAACATAATGGATGTGTTGTTTGTGTGATACTATAAAGTTGGTACATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGA
IHF flip (3.19d)	AAGTACTCTTTAACATAATGGATGTGTTGTTTGTGTGATACTATAAAGTTGGTAGATTGTGACTGGCTTCATAACCTATTATTAATTAATGATTTTTTTTTAGA

Table 3.3 | DNA substrates used in this study. Sequences of all oligonucleotide DNA substrates used are listed, as well as the mutant leaders used for *in vivo* acquisition assays. Oligo description includes figures panels that the substrates were used to generate. Substrates used for crystallography or electron microscopy are noted with the relevant structure. Mutations from wild-type sequences are highlighted in red.

3.5.4 Cryo-EM microscopy

Cas1-Cas2-DNA-IHF complexes in a buffer containing 20 mM HEPES, pH 7.5, 150 mM KCl, 5 mM EDTA, 1 mM DTT, and 0.1% glycerol were used for cryo-EM sample preparation. Immediately after glow-discharging the grid for 14 seconds using a Solaris plasma cleaner, 3.6 µl droplets of the sample (~1µM) were placed onto C-flat grids with 2 µm holes and 2 µm spacing between holes (Protochips Inc.). The grids were rapidly plunged into liquid ethane using an FEI Vitrobot MarkIV maintained at 8 °C and 100% humidity, after being blotted for 4.5 seconds with a blot force of 12. Data were acquired using an FEI Titan Krios transmission electron microscope operated at 300 keV, at a nominal magnification of ×24,500 (1.07 Å pixel size), and with defocus ranging from −1.2 to −2.8 µm. A total of ~3,000 micrographs were recorded using SerialEM on a Gatan K2 Summit direct electron detector operated in super-resolution mode (Mastronarde, 2003). We collected a 6.0s exposure fractionated into 30, 200 ms frames with a dose of 6.8 e-Å^{-2}s^{-1}.

3.5.5 Image processing and reconstruction for cryo-EM

The 28 frames (we skipped the first 2 frames) of each image stack in super-resolution model were aligned, decimated, and summed and dose-weighted using Motioncor2 (Zheng et al., 2017). CTF values of the summed-micrographs were determined using CTFFIND4 and then applied to dose-weighted summed-micrographs for further processing (Rohou and Grigorieff, 2015). Initial particle picking to generate template images was performed using EMAN2. About 20,000 particles were stacked and then imported into Relion2.0 for reference-free 2D classification (Kimanius et al., 2016). Particle picking for the complete dataset was carried out using Gautomatch (http://www.mrc-lmb.cam.ac.uk/kzhang/) with templates generated in previous 2D classification. About 650,000 good particles were selected in total. Due to the preferred orientation issue, random half of the particles in preferred orientations were thrown away, and then only 410,000 particles were left for further processing. Using the 3D model got from negative staining and low-pass filtered to 60Å as a reference, we performed 3D classification using RELION2.0. 3D refinements of the 2 best classes were performed in Cryosparc by importing the 3d models generated from 3D classification (Punjani et al., 2017). The local resolution was calculated by Relion2.0. The reported resolution was based on the gold standard FSC criterion using two independent half-maps. The model resolution in different orientations was calculated using two independent half-maps with the ThreeDFSC script shared by Philip Baldwin (https://github.com/nysbc/Anisotropy).

3.5.6 Model building and refinement

Initial phases for the half-site and pseudo-full-site crystal structures were calculated by molecular replacement with the protospacer-bound Cas1-Cas2 complex (Protein Data Bank accession number 5DS5) in *PHASER* (McCoy et al., 2007). The low resolution of the half-site-bound structure and the disorder of the DNA in particular precluded confident placement of individual nucleotides. DNA from the leader-repeat and repeat-spacer junction generated from the pseudo-full-site structure were used to generate initial models at the corresponding density in the half-site structure. Regular B-form DNA was used as an initial model for the early and mid-repeat regions of the half-site DNA and modified to fit the trajectory and helical pitch of the visible density. The structures were completed through iterative model-building in *COOT* and refinement in *PHENIX* (Afonine et al., 2012; Emsley et al., 2010). The pseudo-full-site structure was refined using NCS and reference-model restraints until the final rounds of refinement. For the lower resolution half-site structure, refinement was carried out using reference-model, NCS, and secondary structure restraints. Anomalous difference maps to identify Ni^{2+} sites were generated using data truncated to 6.2 Å. Web 3DNA was used to analyze structural parameters of the DNA (Zheng et al., 2009).

To generate a complete model for the cryo-EM map, the crystal structure of half-site-bound Cas1-Cas2 solved in this work and published atomic model of IHF module (PDB accession code 1IHF) were first fitted into the refined 3D-reconstruction map using UCSF Chimera (*18*) and then manually rebuilt in Coot to fit the density. The DNA substrates were manually built ab initio in Coot based on the EM density. To improve backbone geometry, the atomic model of Cas1-Cas2-IHF-DNA model was subjected to

PHENIX real space refinement (global minimization and ADP refinement) with Ramachandran, rotamer, and nucleic-acid restraints. The final model was validated using Molprobity (Chen et al., 2010). Structural analysis was performed in Coot and figures were prepared using PyMOL (Schrodinger LLC) and UCSF Chimera. Data collection and refinement statistics are in Table 3.2.

3.5.7 Electrophoretic mobility shift assays

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

3.5.8 Hydroxyl radical footprinting

Cas1 and Cas2 were coincubated in buffer containing 20 mM HEPES pH 7.5, 50 mM KCl, 10 mM EDTA, and 100 μg/mL BSA on ice for 30 minutes. Cas1-Cas2 was added to 1 nM DNA at a final concentration of 0, 10, 100, or 1000 nM and allowed to bind at room temperature for one hour. Hydroxyl radical cleavage was carried out as previously described, except that additional EDTA was not added (Carey and Smale, 2007). The DNA was ethanol precipitated and resuspended in loading buffer containing 95% formamide and 10 mM EDTA, incubated at 95° for 5 minutes, and resolved on a 12% denaturing polyacrylamide gel. The gel was dried and visualized with phosphorimaging.

3.5.9 *In vivo* acquisition assays

In vivo acquisition assays using Cas1 or Cas2 mutants were performed as previously described (Nuñez et al., 2014). Assays involving mutations in the leader region were performed using our pCDF-Cas1-Cas2 expression plasmid with the leader and single repeat from the BL21 CRISPR-I locus cloned into the XbaI site. Amplification was performed with primers specific to the plasmid-based locus. Complete leader sequences are shown in Table 3.3. Assays using IHF-α point mutants were performed in a IHF-α knockout strain with the mutant IHF-α and wild-type IHF-β expressed off a plasmid, as previously described (Nuñez et al., 2016). Both IHF mutant assays and leader mutant assays were grown under induction for two 24-hour cycles before analysis to allow for higher levels of integration (Nuñez et al., 2016).

3.5.10 Integration, second-site integration, and disintegration assays

Integration assays with plasmid target were performed largely as previously described (Nuñez et al., 2015b). pCRISPR was used as a target, Cas1-Cas2 were at 100 nM, protospacer at 100 nM, and plasmid at 7.5 nM, and the reaction was carried out for one hour. Metal was omitted from the reaction buffer or added at 10 mM where indicated. Integration assays with radiolabeled protospacer were performed as previously described, except that protospacer concentration was changed to 10 nM and target concentration was 100 nM, and reactions were carried out at room temperature (Nuñez et al., 2016). IHF was included at 200 nM unless otherwise noted. Integration assays with radiolabeled leaders were carried out with 200 nM Cas1-Cas2, 100 nM unlabeled protospacer, 50 nM IHF, and 10 nM labeled target. Reactions were carried out at room temperature. Samples were run on a 12% denaturing polyacrylamide gel. Progression of half-site substrates to full-site and disintegration assays were both performed as previously described for Cas1-Cas2 from *Streptococcus pyogenes*, with 100 nM protein and 1 nM radiolabeled DNA substrate (Wright and Doudna, 2016). Reactions were performed at room temperature, time points were taken at 0.5, 1, 2, 10, 30, and 60 minutes, and samples were run on a 12% denaturing polyacrylamide gel. Gels was dried and visualized with phosphorimaging. Bands were quantified with ImageQuant (GE Healthcare) and data were analyzed with Prism and fit with a one-phase association model (GraphPad).

3.6 Acknowledgments

We thank G. Meigs and the Advanced Light Source 8.3.1 beamline staff and D. Tzanko, A. Gonzalez and the Stanford Synchrotron Radiation Light Source 9-2 beamline staff for assistance with data collection, and A. East-Seletsky for input on the manuscript. Beamline 8.3.1 at the Advanced Light Source is operated by the University of California Office of the President, Multicampus Research Programs and Initiatives grant MR-15-328599 and Program for Breakthrough Biomedical Research, which is partially funded by the Sandler Foundation. Use of the Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, is supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Contract No. DE-AC02-76SF00515. The EM data was collected in the HHMI EM facility located in UC Berkeley. We thank D. B. Toso and P. Grob for expert electron microscopy assistance, A. Chintangal for computational support, and members of Nogales lab for helpful discussions on EM data processing. The SSRL Structural Molecular Biology Program is supported by the DOE Office of Biological and Environmental Research, and by the National Institutes of Health, National Institute of General Medical Sciences (including P41GM103393). This project was funded by US National Science Foundation grant no. 1244557 (J.A.D.) and National Institute of General Medicine Sciences grant no. 1P50GM102706-01 (J.H. Cate). A.V.W. and K.W.D. are supported by a US National Science Foundation Graduate Research Fellowship, and G.J.K. is funded by Howard Hughes Medical Institute. J.A.D. and E.N. are investigators of the Howard Hughes Medical Institute and a member of the Center for RNA Systems Biology. Atomic coordinates and structure factors for the reported crystal structures have been deposited at the Protein Data Bank under accession codes 5VVJ (half-site-bound) 5VVK (pseudo-full-site-bound) and 5VVL (pseudo-full-site-bound with

Ni^{2+}). The cryo-EM structure and map have been deposited at the Protein Data Bank under accession code 5WFE and the Electron Microscopy Data Bank under accession code EMD-8827. A patent was filed by the University of California for the use of Cas1-Cas2 for integrating DNA into genomes. J.A.D. is a cofounder and Scientific Advisory Board member of Caribou Biosciences and Intellia Therapeutics and a cofounder of Editas Medicine, all of which develop CRISPR-based technologies. Correspondence and requests for materials should be addressed to J.A.D. (doudna@berkeley.edu).

CHAPTER 4

Conclusion and future directions

4.1 Specificity of Cas1-Cas2

The sequence specificity of Cas1-Cas2 is unusual among transposases and is essential for the functionality of the complex as part of an immune system. Without a high degree of specificity, acquisition would be inefficient and the risk of mutation from off-target integration would impose costs that would likely outweigh the benefits of a CRISPR system. The conservation of Cas1-Cas2 across virtually all CRISPR types suggests that the domestication of Cas1 from a promiscuous transposase was the essential first step in the evolution of these systems. While novel interference proteins have been coopted into CRISPR systems multiple times across evolutionary time, the evolution of the adaptation module appears to have occurred only once (Mohanraju et al., 2016).

This work provides valuable insight into how the Cas1-Cas2 integrase has evolved specificity for the CRISPR array. The proteins have not acquired new structural motifs, such as a helix-turn-helix or zinc-finger domain, to read out the target sequence. Instead, the structural constraints imposed by a relatively rigid complex with distant active sites provide specificity. We observe that the requirements for integration by a single active site in both the type I-E and type II-A systems are relatively lax, resulting in half-site integration at sequences not expected to support integration *in vivo*. The requirement for full-site integration, however, imposes additional restrictions such that successful integration *in vivo* occurs primarily at the CRISPR locus even under conditions favoring aggressive integration, with most off-target events occurring at sites with sequence similarity to the CRISPR array (Nivala et al., 2018).

Despite the apparently conserved role of full-site integration as the specificity-determining step of integration, our study of type I and type II systems reveals different strategies for how the Cas1-Cas2 complex maintains specificity. For the type I-E system, DNA distortion predominates over direct sequence recognition. Initial recognition is driven by IHF, which creates a DNA structure that Cas1-Cas2 recognizes through both base-specific contacts and non-specific backbone interactions. Specificity for the repeat, however, is driven by the sequence-dependent deformability of the mid-repeat region. Contacts between the protein complex and the repeat DNA are almost entirely peripheral. The type II-A system relies more strongly on direct sequence recognition. Our work identified the inverted repeat motifs on either end of repeat as critical for integration at that end of the repeat, and subsequent structural work identified additional sequence-specific contacts between a loop of Cas1 and these motifs (Xiao et al., 2017). While the repeat DNA is still bent as it traverses the complex, there is minimal twist deformation, explaining the reduced dependence on mid-repeat sequences. The structural parameters of the protein complex remain essential for dictating a strict ruler mechanism during repeat recognition.

Overall, our research suggests that the evolution of Cas1-Cas2 has been essentially conservative. The properties of the complex that provide specificity, such as its requirement for DNA bending or the involvement of IHF, are often seen in transposases as well. However, for transposases, these properties are important for recognition of the donor DNA or for driving the reaction forward, not for recognition of a specific target. Cas1-Cas2 is unusual in exploiting these properties to generate sequence specificity.

4.2 Diversity of acquisition systems

While the studies presented here address two distinct branches of Cas1 diversity, they cover only a tiny fraction of the range of acquisition systems. Recent work has increased our understanding of other subtypes, particularly type I variants, but a great deal remain entirely unexplored. While investigation of a type I-F system identified that IHF plays a similar role as in the *E. coli* I-E system, a study of a I-A system revealed a reliance on a currently unidentified ATP-dependent host factor for leader recognition as well as a requirement for 500 nucleotides of the leader, in contrast to the 60 nucleotides required by the *E. coli* system (Fagerlund et al., 2017; Rollie et al., 2018). While the principle of interacting with host proteins to gain specificity for the leader is conserved, the nature of the interaction is clearly very different. A broader investigation is needed to uncover the diversity of these interactions and establish how CRISPR systems have evolved to cooperate with different sets of proteins in diverse bacteria and archaea. An important question is whether these interactions are conserved within a given subtype and therefore restrict subtypes to a range of compatible hosts, or if these interactions are plastic and readily evolve when, for example, a type I-E system is transferred to a host lacking IHF.

Acquisition outside of type I and II systems remains almost entirely unstudied, but there is evidence that these other types have evolved unique mechanisms to address the challenges of protospacer generation and target selection. Some type III and type VI systems, which primarily target RNA, have Cas1 proteins fused to a reverse transcriptase, and in one instance have been shown to acquire directly from RNA (Silas et al., 2017b; 2016; Toro et al., 2017). While some aspects of the reverse transcription and integration mechanism have been revealed, generating a coherent picture of how single-stranded RNA fragment is converted into a fully-integrated double-stranded DNA spacer requires further study (Silas et al., 2016). Just as interesting is the question of how RNA-targeting systems without reverse transcriptase-Cas1 proteins acquire spacers with the necessary strand and orientation specificity necessary to target an RNA transcript, as well as whether these systems can defend against RNA viruses or merely interfere with the transcription of DNA viruses.

The expanding array of type V systems also provides hints of novel variations on acquisition systems. Type V-C and V-D systems have *cas1* genes but no *cas2* as well as CRISPR arrays with unusually short spacers (Burstein et al., 2017). We are currently undertaking work to establish whether the Cas1 proteins from these systems act as a minimal integration complex with a shortened protospacer ruler due to the lack of Cas2. These systems may shed light on the origins of CRISPR systems, as it is possible that Cas2 was incorporated into the systems later and that the ancestral Cas1 was sufficient for protospacer integration.

A final area of ongoing research is the cooperation between Cas1-Cas2 and components of the interference pathway. The close links between acquisition and interference were first seen in the type I priming pathway, and subsequent work has only confirmed that the interference proteins are an important and sometimes obligate part of adaptation. Elements of the priming pathway have been reconstituted biochemically, and there is some indication that Cas1-Cas2 is capable of modulating the activity of Cas3, but the nature and degree of their cooperation has yet to be fully established (Künne et al., 2016; Redding et al., 2015). Type II systems also require Cas9 and the tracrRNA in addition to Cas1, Cas2, and Csn2 (Heler et al., 2015; Wei et al., 2015b). The mechanisms involved in this interaction remain unknown and deserve further study. Other CRISPR types likely have their own unique interactions between the acquisition and interference machinery, and it will be fascinating to observe as these interactions are discovered and characterized.

6.3 Applications

The activity of Cas1-Cas2 makes the complex uniquely suited to certain specialized applications. From a technology perspective, the proteins act fundamentally as a recording device. One potential application of this activity is the barcoding of individual cells in a population. By supplying Cas1-Cas2 with randomized sequences as protospacers, individual cells could be marked with a unique barcode at a known genomic locus. This would be useful in tracking cell fate in studies of development, oncogenesis, or other fields where lineage tracing is critical. Cas9 has already been used for similar purposes, either by targeting a tandem array to result in a random assortment of cleavage and repair events or by targeting the guide itself to produce a continually evolving sequence through rounds of cleavage and repair (Kalhor et al., 2017; McKenna et al., 2016). A Cas1-Cas2-based method would have the advantage of introducing a standardized, predictable, and highly diverse and customizable set of barcodes in a single event.

Another application currently being developed is using Cas1-Cas2 to report on cellular state. Multiple groups have shown that Cas1-Cas2 can be used to record information, provided either via electroporated protospacers or via plasmids with inducible copy-number variations, into bacterial genomes where it can be read out later by sequencing (Sheth et al., 2017; Shipman et al., 2017). In principle, any input that results in the generation of a DNA substrate could be recorded. If this technique could be coupled to transcription, perhaps using a reverse transcriptase-Cas1, it could provide valuable time-resolved information about the transcriptional programs of cell types in an organism. Again, Cas9 has also been used for similar applications, in this case by putting Cas9 or the guide RNA under inducible promoters and using cleavage or base editing at a targeted site as a means of recording (Tang and Liu, 2018). Here Cas1-Cas2 have the advantage of being able to encode much richer information. Whereas the Cas9-based method is limited to recording the duration and intensity of a stimulus, the ability of Cas1-Cas2 to capture diverse sequences allows the integrated sequence to act as another layer of information in addition to the amount of integration that has occurred.

Both of these applications take advantage of the distinct characteristics of Cas1-Cas2 not shared with other transposases, namely their ability to integrate short, random pieces of DNA and their specificity for a single target site. As development of these tools moves forward, a continued exploration of the basic mechanism of Cas1-Cas2 will be required to inform it. The results presented in this work provide information about specificity that is essential for predicting where in a foreign genome Cas1-Cas2 might integrate, allowing for identification of likely off-target sites and designing a desired integration site. Currently, using Cas1-Cas2 in a new organism will likely require knocking in a CRISPR locus to serve as an integration target, but targeting the proteins to a pre-existing locus has clear advantages. The structural restraints on integration make the prospect of engineering Cas1-Cas2 to recognize new targets an interesting problem, as it will likely require mutations that subtly alter the relative orientation of the active sites rather than simply directed mutations of base-interacting residues. It will be fascinating to observe as these applications move beyond the proof-of-concept phase and potentially provide Cas1-Cas2 a place as a research tool alongside the other Cas proteins already in use.

References

Abudayyeh, O.O., Gootenberg, J.S., Essletzbichler, P., Han, S., Joung, J., Belanto, J.J., Verdine, V., Cox, D.B.T., Kellner, M.J., Regev, A., et al. (2017). RNA targeting with CRISPR-Cas13. Nature *550*, 280–284.

Abudayyeh, O.O., Gootenberg, J.S., Konermann, S., Joung, J., Slaymaker, I.M., Cox, D.B.T., Shmakov, S., Makarova, K.S., Semenova, E., Minakhin, L., et al. (2016). C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. Science *353*, aaf5573.

Adey, A., and Shendure, J. (2012). Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing. Genome Res. *22*, 1139–1143.

Afonine, P.V., Grosse-Kunstleve, R.W., Echols, N., Headd, J.J., Moriarty, N.W., Mustyakimov, M., Terwilliger, T.C., Urzhumtsev, A., Zwart, P.H., and Adams, P.D. (2012). Towards automated crystallographic structure refinement with phenix.refine. Acta Crystallogr. D Biol. Crystallogr. *68*, 352–367.

Anders, C., Niewoehner, O., Duerst, A., and Jinek, M. (2014). Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. Nature *513*, 569–573.

Arslan, Z., Hermanns, V., Wurm, R., Wagner, R., and Pul, U. (2014). Detection and characterization of spacer integration intermediates in type I-E CRISPR-Cas system. Nucleic Acids Research *42*, 7884–7893.

Arslan, Z., Wurm, R., Brener, O., Ellinger, P., Nagel-Steger, L., Oesterhelt, F., Schmitt, L., Willbold, D., Wagner, R., Gohlke, H., et al. (2013). Double-strand DNA end-binding and sliding of the toroidal CRISPR-associated protein Csn2. Nucleic Acids Research *41*, 6347–6359.

Babu, M., Beloglazova, N., Flick, R., Graham, C., Skarina, T., Nocek, B., Gagarinova, A., Pogoutse, O., Brown, G., Binkowski, A., et al. (2010). A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. Molecular Microbiology *79*, 484–502.

Bainton, R.J., Kubo, K.M., Feng, J.N., and Craig, N.L. (1993). Tn7 transposition: target DNA recognition is mediated by multiple Tn7-encoded proteins in a purified in vitro system. Cell *72*, 931–943.

Barrangou, R., and Marraffini, L.A. (2014). CRISPR-Cas Systems: Prokaryotes Upgrade to Adaptive Immunity. Molecular Cell *54*, 234–244.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., and Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. Science *315*, 1709–1712.

Béguin, P., Charpin, N., Koonin, E.V., Forterre, P., and Krupovic, M. (2016). Casposon integration shows strong target site preference and recapitulates protospacer integration by CRISPR-Cas systems. Nucleic Acids Research *44*, 10367–10376.

Blackwood, J.K., Rzechorzek, N.J., Bray, S.M., Maman, J.D., Pellegrini, L., and Robinson, N.P. (2013). End-resection at DNA double-strand breaks in the three domains of life: Figure 1. Biochem. Soc. Trans. *41*, 314–320.

Blosser, T.R., Loeff, L., Westra, E.R., Vlot, M., Künne, T., Sobota, M., Dekker, C., Brouns, S.J.J., and Joo, C. (2015). Two distinct DNA binding modes guide dual roles of a CRISPR-Cas protein complex. Molecular Cell *58*, 60–70.

Blundell, J.R., and Levy, S.F. (2014). Beyond genome sequencing: Lineage tracking with barcodes to study the dynamics of evolution, infection, and cancer. Genomics *104*, 417–430.

Bolotin, A., Quinquis, B., Sorokin, A., and Ehrlich, S.D. (2005). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology (Reading, Engl.) *151*, 2551–2561.

Borchardt, E.K., Vandoros, L.A., Huang, M., Lackey, P.E., Marzluff, W.F., and Asokan, A. (2015). Controlling mRNA stability and translation with the CRISPR endoribonuclease Csy4. Rna *21*, 1921–1930.

Brendel, J., Stoll, B., Lange, S.J., Sharma, K., Lenz, C., Stachler, A.-E., Maier, L.-K., Richter, H., Nickel, L., Schmitz, R.A., et al. (2014). A complex of Cas proteins 5, 6, and 7 is required for the biogenesis and stability of clustered regularly interspaced short palindromic repeats (crispr)-derived rnas (crrnas) in Haloferax volcanii. J. Biol. Chem. *289*, 7164–7177.

Briner, A.E., Donohoue, P.D., Gomaa, A.A., Selle, K., Slorach, E.M., Nye, C.H., Haurwitz, R.E., Beisel, C.L., May, A.P., and Barrangou, R. (2014). Guide RNA functional modules direct Cas9 activity and orthogonality. Molecular Cell *56*, 333–339.

Brouns, S.J.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J.H., Snijders, A.P.L., Dickman, M.J., Makarova, K.S., Koonin, E.V., and van der Oost, J. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. Science *321*, 960–964.

Burstein, D., Harrington, L.B., Strutt, S.C., Probst, A.J., Anantharaman, K., Thomas, B.C., Doudna, J.A., and Banfield, J.F. (2017). New CRISPR-Cas systems from uncultivated microbes. Nature *542*, 237–241.

Carey, M., and Smale, S.T. (2007). Hydroxyl-Radical Footprinting. Cold Spring Harbor Protocols *2007*, pdb.prot4810.

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

Charpentier, E., Richter, H., van der Oost, J., and White, M.F. (2015). Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. FEMS Microbiology Reviews *39*, 428–441.

Chen, J.S., Ma, E., Harrington, L.B., Da Costa, M., Tian, X., Palefsky, J.M., and Doudna, J.A. (2018). CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. Science *360*, 436–439.

Chen, V.B., Arendall, W.B., Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., and Richardson, D.C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr. D Biol. Crystallogr. *66*, 12–21.

Chow, S.A., Vincent, K.A., Ellison, V., and Brown, P.O. (1992). Reversal of integration and DNA splicing mediated by integrase of human immunodeficiency virus. Science *255*, 723–726.

Chu, V.T., Weber, T., Wefers, B., Wurst, W., Sander, S., Rajewsky, K., and Kühn, R. (2015). Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. Nat. Biotechnol. *33*, 543–548.

Chylinski, K., Makarova, K.S., Charpentier, E., and Koonin, E.V. (2014). Classification and evolution of type II CRISPR-Cas systems. Nucleic Acids Research *42*, 6091–6105.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., et al. (2013). Multiplex Genome Engineering Using CRISPR/Cas Systems. Science *339*, 819–823.

Cox, D.B.T., Gootenberg, J.S., Abudayyeh, O.O., Franklin, B., Kellner, M.J., Joung, J., and Zhang, F. (2017). RNA editing with CRISPR-Cas13. Science *358*, 1019–1027.

Craigie, R., and Bushman, F.D. (2012). HIV DNA Integration. Cold Spring Harbor Perspectives in Medicine *2*, a006890–a006890.

Crooks, G.E., Hon, G., Chandonia, J.-M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res. *14*, 1188–1190.

D'Astolfo, D.S., Pagliero, R.J., Pras, A., Karthaus, W.R., Clevers, H., Prasad, V., Lebbink, R.J., Rehmann, H., and Geijsen, N. (2015). Efficient intracellular delivery of native proteins. Cell *161*, 674–690.

Datsenko, K.A., Pougach, K., Tikhonov, A., Wanner, B.L., Severinov, K., and Semenova, E. (2012). Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. Nature Communications *3*, 945.

Davis, K.M., Pattanayak, V., Thompson, D.B., Zuris, J.A., and Liu, D.R. (2015). Small molecule-triggered Cas9 protein with improved genome-editing specificity. Nat. Chem. Biol. *11*, 316–318.

Deltcheva, E., Chylinski, K., Sharma, C.M., Gonzales, K., Chao, Y., Pirzada, Z.A., Eckert, M.R., Vogel, J., and Charpentier, E. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature *471*, 602–607.

Deng, L., Garrett, R.A., Shah, S.A., Peng, X., and She, Q. (2013). A novel interference mechanism by a type IIIB CRISPR-Cmr module in Sulfolobus. Molecular Microbiology *87*, 1088–1099.

Deveau, H., Barrangou, R., Garneau, J.E., Labonté, J., Fremaux, C., Boyaval, P., Romero, D.A., Horvath, P., and Moineau, S. (2008). Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. Journal of Bacteriology *190*, 1390–1400.

Diederichs, K., and Karplus, P.A. (2013). Better models by discarding data? Acta Crystallogr. D Biol. Crystallogr. *69*, 1215–1222.

Dillingham, M.S., and Kowalczykowski, S.C. (2008). RecBCD enzyme and the repair of double-stranded DNA breaks. Microbiol. Mol. Biol. Rev. *72*, 642–71–TableofContents.

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

Du, P., Miao, C., Lou, Q., Wang, Z., and Lou, C. (2016). Engineering Translational Activators with CRISPR-Cas System. ACS Synth Biol *5*, 74–80.

Dugar, G., Herbig, A., Förstner, K.U., Heidrich, N., Reinhardt, R., Nieselt, K., and Sharma, C.M. (2013). High-resolution transcriptome maps reveal strain-specific regulatory features of multiple Campylobacter jejuni isolates. PLoS Genet *9*, e1003495.

East-Seletsky, A., O'Connell, M.R., Knight, S.C., Burstein, D., Cate, J.H.D., Tjian, R., and Doudna, J.A. (2016). Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. Nature *538*, 270–273.

Ellinger, P., Arslan, Z., Wurm, R., Tschapek, B., MacKenzie, C., Pfeffer, K., Panjikar, S., Wagner, R., Schmitt, L., Gohlke, H., et al. (2012). The crystal structure of the CRISPR-associated protein Csn2 from Streptococcus agalactiae. J. Struct. Biol. *178*, 350–362.

Elmore, J.R., Sheppard, N.F., Ramia, N., Deighan, T., Li, H., Terns, R.M., and Terns, M.P. (2016). Bipartite recognition of target RNAs activates DNA cleavage by the Type III-B CRISPR-Cas system. Genes Dev. *30*, 447–459.

Emsley, P., Lohkamp, B., Scott, W.G., and Cowtan, K. (2010). Features and development of Coot. Acta Crystallogr. D Biol. Crystallogr. *66*, 486–501.

Engelman, A., Mizuuchi, K., and Craigie, R. (1991). HIV-1 DNA integration: mechanism of viral DNA cleavage and DNA strand transfer. Cell *67*, 1211–1221.

Estrella, M.A., Kuo, F.-T., and Bailey, S. (2016). RNA-activated DNA cleavage by the Type III-B CRISPR-Cas effector complex. Genes Dev. *30*, 460–470.

Evans, P.R., and Murshudov, G.N. (2013). How good are my data and what is the resolution? Acta Crystallogr. D Biol. Crystallogr. *69*, 1204–1214.

Fagerlund, R.D., Wilkinson, M.E., Klykov, O., Barendregt, A., Pearce, F.G., Kieper, S.N., Maxwell, H.W.R., Capolupo, A., Heck, A.J.R., Krause, K.L., et al. (2017). Spacer capture and integration by a type I-F Cas1?Cas2-3 CRISPR adaptation complex. Proceedings of the National Academy of Sciences *23*, 201618421–17.

Fonfara, I., Richter, H., Bratovič, M., Le Rhun, A., and Charpentier, E. (2016). The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. Nature *532*, 517–521.

Garcia-Doval, C., and Jinek, M. (2017). ScienceDirect Molecular architectures and mechanisms of Class 2 CRISPR-associated nucleases. Curr. Opin. Struct. Biol. *47*, 157–166.

Gardiner, E.J., Hunter, C.A., Packer, M.J., Palmer, D.S., and Willett, P. (2003). Sequence-dependent DNA Structure: A Database of Octamer Structural Parameters. J. Mol. Biol. *332*, 1025–1035.

Garrett, R., Shah, S., Erdmann, S., Liu, G., Mousaei, M., León-Sobrino, C., Peng, W., Gudbergsdottir, S., Deng, L., Vestergaard, G., et al. (2015). CRISPR-Cas Adaptive Immune Systems of the Sulfolobales: Unravelling Their Complexity and Diversity. Life *5*, 783–817.

Garside, E.L., Schellenberg, M.J., Gesner, E.M., Bonanno, J.B., Sauder, J.M., Burley, S.K., Almo, S.C., Mehta, G., and MacMillan, A.M. (2012). Cas5d processes pre-crRNA and is a member of a larger family of CRISPR RNA endonucleases. Rna *18*, 2020–2028.

Gasiunas, G., Barrangou, R., Horvath, P., and Siksnys, V. (2012). Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc. Natl. Acad. Sci. U.S.a. *109*, E2579–E2586.

Gesner, E.M., Schellenberg, M.J., Garside, E.L., George, M.M., and MacMillan, A.M. (2011). Recognition and maturation of effector RNAs in a CRISPR interference pathway. Nat. Struct. Mol. Biol. *18*, 688–692.

Goldberg, G.W., Jiang, W., Bikard, D., and Marraffini, L.A. (2014). Conditional tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting. Nature *514*, 633–637.

Gong, B., Shin, M., Sun, J., Jung, C.-H., Bolt, E.L., van der Oost, J., and Kim, J.-S. (2014). Molecular insights into DNA interference by CRISPR-associated nuclease-helicase Cas3. Proc. Natl. Acad. Sci. U.S.a. *111*, 16359–16364.

Gootenberg, J.S., Abudayyeh, O.O., Kellner, M.J., Joung, J., Collins, J.J., and Zhang, F. (2018). Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a, and Csm6. Science *360*, 439–444.

Gootenberg, J.S., Abudayyeh, O.O., Lee, J.W., Essletzbichler, P., Dy, A.J., Joung, J., Verdine, V., Donghia, N., Daringer, N.M., Freije, C.A., et al. (2017). Nucleic acid detection with CRISPR-Cas13a/C2c2. Science *356*, 438–442.

Goren, M.G., Doron, S., Globus, R., Amitai, G., Sorek, R., and Qimron, U. (2016). Repeat Size Determination by Two Molecular Rulers in the Type I-E CRISPR Array. CellReports *16*, 2811–2818.

Gori, J.L., Hsu, P.D., Maeder, M.L., Shen, S., Welstead, G.G., and Bumcrot, D. (2015). Delivery and Specificity of CRISPR-Cas9 Genome Editing Technologies for Human Gene Therapy. Hum. Gene Ther. *26*, 443–451.

Goryshin, I.Y., and Reznikoff, W.S. (1998). Tn5 in vitro transposition. Journal of Biological Chemistry *273*, 7367–7374.

Goryshin, I.Y., Jendrisak, J., Hoffman, L.M., Meis, R., and Reznikoff, W.S. (1999). Insertional transposon mutagenesis by electroporation of released Tn5 transposition complexes. Nat. Biotechnol. *18*, 97–100.

Hale, C.R., Zhao, P., Olson, S., Duff, M.O., Graveley, B.R., Wells, L., Terns, R.M., and Terns, M.P. (2009). RNA-Guided RNA Cleavage by a CRISPR RNA-Cas Protein Complex. Cell *139*, 945–956.

Hale, C., Kleppe, K., Terns, R.M., and Terns, M.P. (2008). Prokaryotic silencing (psi)RNAs in Pyrococcus furiosus. Rna *14*, 2572–2579.

Han, W., Li, Y., Deng, L., Feng, M., Peng, W., Hallstrøm, S., Zhang, J., Peng, N., Liang, Y.X., White, M.F., et al. (2017a). A type III-B CRISPR-Cas effector complex mediating massive target DNA destruction. Nucleic Acids Research *45*, 1983–1993.

Han, W., Pan, S., López-Méndez, B., Montoya, G., and She, Q. (2017b). Allosteric regulation of Csx1, a type IIIB-associated CARF domain ribonuclease by RNAs carrying a tetraadenylate tail. Nucleic Acids Research 1–11.

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

Heler, R., Samai, P., Modell, J.W., Weiner, C., Goldberg, G.W., Bikard, D., and Marraffini, L.A. (2015). Cas9 specifies functional viral targets during CRISPR-Cas adaptation. Nature *519*, 199–202.

Hemsley, A., Arnheim, N., Toney, M.D., Cortopassi, G., and Galas, D.J. (1989). A simple method for site-directed mutagenesis using the polymerase chain reaction. Nucleic Acids Research *17*, 6545–6551.

Hickman, A.B., and Dyda, F. (2015). The casposon-encoded Cas1 protein from Aciduliprofundum boonei is a DNA integrase that generates target site duplications. Nucleic Acids Research *43*, 10576–10587.

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

Hochstrasser, M.L., Taylor, D.W., Bhat, P., Guegler, C.K., Sternberg, S.H., Nogales, E., and Doudna, J.A. (2014). CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. Proc. Natl. Acad. Sci. U.S.a. *111*, 6618–6623.

Hooton, S.P.T., and Connerton, I.F. (2014). Campylobacter jejuni acquire new host-derived CRISPR spacers when in association with bacteriophages harboring a CRISPR-like Cas4 protein. Front Microbiol *5*, 744.

Howes, R., and Schofield, C. (2015). Genome engineering using Adeno-Associated Virus (AAV). Methods Mol. Biol. *1239*, 75–103.

Hsu, P.D., Lander, E.S., and Zhang, F. (2014). Development and Applications of CRISPR-Cas9 for Genome Engineering. Cell *157*, 1262–1278.

Hsu, P.D., Scott, D.A., Weinstein, J.A., Ran, F.A., Konermann, S., Agarwala, V., Li, Y., Fine, E.J., Wu, X., Shalem, O., et al. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. Nat. Biotechnol. *31*, 827–832.

Huo, Y., Nam, K.H., Ding, F., Lee, H., Wu, L., Xiao, Y., Farchione, M.D., Zhou, S., Rajashankar, K., Kurinov, I., et al. (2014). Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. Nat. Struct. Mol. Biol. *21*, 771–777.

Hynes, A.P., Villion, M., and Moineau, S. (2014). Adaptation in bacterial CRISPR-Cas immunity can be driven by defective phages. Nature Communications *5*, 4399.

Ishino, Y., Shinagawa, H., Makino, K., Amemura, M., and Nakata, A. (1987). Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. Journal of Bacteriology *169*, 5429–5433.

Ivančić-Baće, I., Cass, S.D., Wearne, S.J., and Bolt, E.L. (2015). Different genome stability proteins underpin primed and naïve adaptation in E. coliCRISPR-Cas immunity. Nucleic Acids Research *43*, 10821–10830.

Jackson, R.N., and Wiedenheft, B. (2015). A Conserved Structural Chassis for Mounting Versatile CRISPR RNA-Guided Immune Responses. Molecular Cell *58*, 722–728.

Jackson, R.N., Golden, S.M., van Erp, P.B.G., Carter, J., Westra, E.R., Brouns, S.J.J., van der Oost, J., Terwilliger, T.C., Read, R.J., and Wiedenheft, B. (2014). Structural biology. Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli. Science *345*, 1473–1479.

Jansen, R., Embden, J.D.A.V., Gaastra, W., and Schouls, L.M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. Molecular Microbiology *43*, 1565–1575.

Jiang, F., and Doudna, J.A. (2015). The structural biology of CRISPR-Cas systems. Curr. Opin. Struct. Biol. *30C*, 100–111.

Jiang, F., Zhou, K., Ma, L., Gressel, S., and Doudna, J.A. (2015). STRUCTURAL BIOLOGY. A Cas9-guide RNA complex preorganized for target DNA recognition. Science *348*, 1477–1481.

Jiang, W., and Marraffini, L.A. (2015). CRISPR-Cas: New Tools for Genetic Manipulations from Bacterial Immunity Systems. Annu. Rev. Microbiol. *69*, 209–228.

Jiang, W., Bikard, D., Cox, D., Zhang, F., and Marraffini, L.A. (2013). RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nat. Biotechnol. *31*, 233–239.

Jiang, W., Samai, P., and Marraffini, L.A. (2016). Degradation of Phage Transcripts by CRISPR-Associated RNases Enables Type III CRISPR-Cas Immunity. Cell *164*, 710–721.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. Science *337*, 816–821.

Jinek, M., Jiang, F., Taylor, D.W., Sternberg, S.H., Kaya, E., Ma, E., Anders, C., Hauer, M., Zhou, K., Lin, S., et al. (2014). Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. Science *343*, 1247997.

Jore, M.M., Lundgren, M., van Duijn, E., Bultema, J.B., Westra, E.R., Waghmare, S.P., Wiedenheft, B., Pul, Ü., Wurm, R., Wagner, R., et al. (2011). Structural basis for CRISPR RNA-guided DNA recognition by Cascade. Nat. Struct. Mol. Biol. *18*, 529–536.

Juranek, S., Eban, T., Altuvia, Y., Brown, M., Morozov, P., Tuschl, T., and Margalit, H. (2012). A genome-wide view of the expression and processing patterns of Thermus thermophilus HB8 CRISPR RNAs. Rna *18*, 783–794.

Ka, D., Lee, H., Jung, Y.-D., Kim, K., Seok, C., Suh, N., and Bae, E. (2016). Crystal Structure of Streptococcus pyogenes Cas1 and Its Interaction with Csn2 in the Type II CRISPR-Cas System. Structure *24*, 70–79.

Kabsch, W. (2010). XDS. Acta Crystallogr. D Biol. Crystallogr. *66*, 125–132.

Kalhor, R., Mali, P., and Church, G.M. (2017). Rapidly evolving homing CRISPR barcodes. Nat. Methods *14*, 195–200.

Kazlauskiene, M., Kostiuk, G., Venclovas, Č., Tamulaitis, G., and Siksnys, V. (2017). A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems. Science *357*, 605–609.

Kazlauskiene, M., Tamulaitis, G., Kostiuk, G., Venclovas, Č., and Siksnys, V. (2016). Spatiotemporal Control of Type III-A CRISPR-Cas Immunity: Coupling DNA Degradation with the Target RNA Recognition. Molecular Cell *62*, 295–306.

Kennedy, M.J., Hughes, R.M., Peteya, L.A., Schwartz, J.W., Ehlers, M.D., and Tucker, C.L. (2010). Rapid blue-light-mediated induction of protein interactions in living cells. Nat. Methods *7*, 973–975.

Kieper, S.N., Almendros, C., Behler, J., McKenzie, R.E., Nobrega, F.L., Haagsma, A.C., Vink, J.N.A., Hess, W.R., and Brouns, S.J.J. (2018). Cas4 Facilitates PAM-Compatible Spacer Selection during CRISPR Adaptation. CellReports *22*, 3377–3384.

Kim, H.J., Lee, H.J., Kim, H., Cho, S.W., and Kim, J.-S. (2009). Targeted genome editing in human cells with zinc finger nucleases constructed via modular assembly. Genome Res. *19*, 1279–1288.

Kimanius, D., Forsberg, B.O., Scheres, S.H., and Lindahl, E. (2016). Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. eLife *5*, 19.

Knott, G.J., East-Seletsky, A., Cofsky, J.C., Holton, J.M., Charles, E., O'Connell, M.R., and Doudna, J.A. (2017). Guide-bound structures of an RNA-targeting A-cleaving CRISPR-Cas13a enzyme. Nat. Struct. Mol. Biol. *24*, 825–833.

Konermann, S., Brigham, M.D., Trevino, A., Hsu, P.D., Heidenreich, M., Cong, L., Platt, R.J., Scott, D.A., Church, G.M., and Zhang, F. (2013). Optical control of mammalian endogenous transcription and epigenetic states. Nature *500*, 472–476.

Konermann, S., Lotfy, P., Brideau, N.J., Oki, J., Shokhirev, M.N., and Hsu, P.D. (2018). Transcriptome Engineering with RNA-Targeting Type VI-D CRISPR Effectors. Cell *173*, 665–668.e14.

Koo, Y., Jung, D.-K., and Bae, E. (2012). Crystal structure of Streptococcus pyogenes Csn2 reveals calcium-dependent conformational changes in its tertiary and quaternary structure. PLoS ONE *7*, e33401.

Koonin, E.V., Makarova, K.S., and Zhang, F. (2017). ScienceDirect Diversity, classification and evolution of CRISPR-Cas systems. Current Opinion in Microbiology *37*, 67–78.

Krupovic, M., Makarova, K.S., Forterre, P., Prangishvili, D., and Koonin, E.V. (2014). Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. BMC Biol. *12*, 36.

Kuhn, C.-D., and Joshua-Tor, L. (2013). Eukaryotic Argonautes come into focus. Trends Biochem. Sci. *38*, 263–271.

Kunin, V., Sorek, R., and Hugenholtz, P. (2007). Evolutionary conservation of sequence and secondary structures in CRISPR repeats. Genome Biol. *8*, R61.

Künne, T., Kieper, S.N., Bannenberg, J.W., Vogel, A.I.M., Miellet, W.R., Klein, M., Depken, M., Suarez-Diez, M., and Brouns, S.J.J. (2016). Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation. Molecular Cell *63*, 852–864.

Labrie, S.J., Samson, J.E., and Moineau, S. (2010). Bacteriophage resistance mechanisms. Nat. Rev. Microbiol. *8*, 317–327.

Lander, G.C., Estrin, E., Matyskiela, M.E., Bashore, C., Nogales, E., and Martin, A. (2012). Complete subunit architecture of the proteasome regulatory particle. Nature *482*, 186–191.

Lander, G.C., Stagg, S.M., Voss, N.R., Cheng, A., Fellmann, D., Pulokas, J., Yoshioka, C., Irving, C., Mulder, A., Lau, P.-W., et al. (2009). Appion: an integrated, database-driven pipeline to facilitate EM image processing. J. Struct. Biol. *166*, 95–102.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Laxmikanthan, G., Xu, C., Brilot, A.F., Warren, D., Steele, L., Seah, N., Tong, W., Grigorieff, N., Landy, A., and Van Duyne, G.D. (2016). Structure of a Holliday junction complex reveals mechanisms governing a highly regulated DNA transaction. eLife *5*, 1–23.

Lee, H., Zhou, Y., Taylor, D.W., and Sashital, D.G. (2018). Cas4-Dependent Prespacer Processing Ensures High-Fidelity Programming of CRISPR Arrays. Molecular Cell *70*, 48–59.e5.

Lee, H.Y., Haurwitz, R.E., Apffel, A., Zhou, K., Smart, B., Wenger, C.D., Laderman, S., Bruhn, L., and Doudna, J.A. (2013). RNA-protein analysis using a conditional CRISPR nuclease. Proc. Natl. Acad. Sci. U.S.a. *110*, 5416–5421.

Lee, K.-H., Lee, S.-G., Eun Lee, K., Jeon, H., Robinson, H., and Oh, B.-H. (2012). Identification, structural, and biochemical characterization of a group of large Csn2 proteins involved in CRISPR-mediated bacterial immunity. Proteins *80*, 2573–2582.

Lemak, S., Beloglazova, N., Nocek, B., Skarina, T., Flick, R., Brown, G., Popovic, A., Joachimiak, A., Savchenko, A., and Yakunin, A.F. (2013). Toroidal Structure and DNA Cleavage by the CRISPR-Associated [4Fe-4S] Cluster Containing Cas4 Nuclease SSO0001 from Sulfolobus solfataricus. J. Am. Chem. Soc. *135*, 17476–17487.

Levy, A., Goren, M.G., Yosef, I., Auster, O., Manor, M., Amitai, G., Edgar, R., Qimron, U., and Sorek, R. (2015). CRISPR adaptation biases explain preference for acquisition of foreign DNA. Nature *520*, 505–510.

Li, M., Wang, R., Zhao, D., and Xiang, H. (2014). Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process. Nucleic Acids Research *42*, 2483–2492.

Liang, F.-S., Ho, W.Q., and Crabtree, G.R. (2011). Engineering the ABA plant stress pathway for regulation of induced proximity. Sci Signal *4*, rs2–rs2.

Lienert, F., Torella, J.P., Chen, J.-H., Norsworthy, M., Richardson, R.R., and Silver, P.A. (2013). Two- and three-input TALE-based AND logic computation in embryonic stem cells. Nucleic Acids Research *41*, 9967–9975.

Lin, S., Staahl, B.T., Alla, R.K., and Doudna, J.A. (2014). Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. eLife *3*, e04766.

Lintner, N.G., Kerou, M., Brumfield, S.K., Graham, S., Liu, H., Naismith, J.H., Sdano, M., Peng, N., She, Q., Copié, V., et al. (2011). Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE). J. Biol. Chem. *286*, 21643–21656.

Liu, L., Li, X., Ma, J., Li, Z., You, L., Wang, J., Wang, M., Zhang, X., and Wang, Y. (2017a). The Molecular Architecture for RNA-Guided RNA Cleavage by Cas13a. Cell *170*, 714–726.e10.

Liu, L., Li, X., Wang, J., Wang, M., Chen, P., Yin, M., Li, J., Sheng, G., and Wang, Y. (2017b). Two Distant Catalytic Sites Are Responsible for C2c2 RNase Activities. Cell *168*, 121–134.e12.

Liu, T.Y., Iavarone, A.T., and Doudna, J.A. (2017c). RNA and DNA Targeting by a Reconstituted Thermus thermophilus Type III-A CRISPR-Cas System. PLoS ONE *12*, e0170552–20.

Ma, E., Harrington, L.B., O’Connell, M.R., Zhou, K., and Doudna, J.A. (2015). Single-Stranded DNA Cleavage by Divergent CRISPR-Cas9 Enzymes. Molecular Cell *60*, 398–407.

Maertens, G.N., Hare, S., and Cherepanov, P. (2010). The mechanism of retroviral integration from X-ray structures of its key intermediates. Nature *468*, 326–329.

Makarova, K.S., and Koonin, E.V. (2015). Annotation and Classification of CRISPR-Cas Systems. Methods Mol. Biol. *1311*, 47–75.

Makarova, K.S., Aravind, L., Wolf, Y.I., and Koonin, E.V. (2011a). Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. Biology Direct *6*, 38.

Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J.M., Wolf, Y.I., Yakunin, A.F., et al. (2011b). Evolution and classification of the CRISPR–Cas systems. Nat. Rev. Microbiol. *9*, 467–477.

Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J., Barrangou, R., Brouns, S.J.J., Charpentier, E., Haft, D.H., et al. (2015). An updated evolutionary classification of CRISPR-Cas systems. Nat. Rev. Microbiol. *13*, 722–736.

Makarova, K.S., Wolf, Y.I., and Koonin, E.V. (2013). The basic building blocks and evolution of CRISPR-CAS systems. Biochem. Soc. Trans. *41*, 1392–1400.

Marcel Martin (2015). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal *17*, 10–12.

Marraffini, L.A., and Sontheimer, E.J. (2008). CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science *322*, 1843–1845.

Marraffini, L.A., and Sontheimer, E.J. (2010). Self versus non-self discrimination during CRISPR RNA-directed immunity. Nature *463*, 568–571.

Maruyama, T., Dougan, S.K., Truttmann, M.C., Bilate, A.M., Ingram, J.R., and Ploegh, H.L. (2015). Increasing the efficiency of precise genome editing with CRISPR-Cas9 by inhibition of nonhomologous end joining. Nat. Biotechnol. *33*, 538–542.

Mastronarde, D.N. (2003). SerialEM: A Program for Automated Tilt Series Acquisition on Tecnai Microscopes Using Prediction of Specimen Position. Https://Doi.org/10.1017/S1431927603445911 1–2.

McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C., and Read, R.J. (2007). Phaser crystallographic software. J Appl Crystallogr *40*, 658–674.

McGinn, J., and Marraffini, L.A. (2016). CRISPR-Cas Systems Optimize Their Immune Response by Specifying the Site of Spacer Integration. Molecular Cell *64*, 616–623.

McKenna, A., Findlay, G.M., Gagnon, J.A., Horwitz, M.S., Schier, A.F., and Shendure, J. (2016). Whole-organism lineage tracing by combinatorial and cumulative genome editing. Science *353*, aaf7907.

Mizuuchi, K., and Adzuma, K. (1991). Inversion of the phosphate chirality at the target site of Mu DNA strand transfer: evidence for a one-step transesterification mechanism. Cell *66*, 129–140.

Moch, C., Fromant, M., Blanquet, S., and Plateau, P. (2017). DNA binding specificities of Escherichia coli Cas1-Cas2 integrase drive its recruitment at the CRISPR locus. Nucleic Acids Research *45*, 2714–2723.

Modell, J.W., Jiang, W., and Marraffini, L.A. (2017). CRISPR–Cas systems exploit viral DNA injection to establish and maintain adaptive immunity. Nature *544*, 101–104.

Mohanraju, P., Makarova, K.S., Zetsche, B., Zhang, F., Koonin, E.V., and van der Oost, J. (2016). Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. Science *353*, aad5147.

Mojica, F.J.M., Díez-Villaseñor, C., García-Martínez, J., and Almendros, C. (2009). Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology (Reading, Engl.) *155*, 733–740.

Mojica, F.J.M., Díez-Villaseñor, C., García-Martínez, J., and Soria, E. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J Mol Evol *60*, 174–182.

Montaño, S.P., Pigli, Y.Z., and Rice, P.A. (2012). The Mu transpososome structure sheds light on DDE recombinase evolution. Nature *491*, 413–417.

Mulepati, S., and Bailey, S. (2011). Structural and biochemical analysis of nuclease domain of clustered regularly interspaced short palindromic repeat (CRISPR)-associated protein 3 (Cas3). J. Biol. Chem. *286*, 31896–31903.

Mulepati, S., and Bailey, S. (2013). In vitro reconstitution of an Escherichia coli RNA-guided immune system reveals unidirectional, ATP-dependent degradation of DNA target. J. Biol. Chem. *288*, 22184–22192.

Mulepati, S., Héroux, A., and Bailey, S. (2014). Structural biology. Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. Science *345*, 1479–1484.

Nadler, D.C., Morgan, S.-A., Flamholz, A., Kortright, K.E., and Savage, D.F. (2016). Rapid construction of metabolite biosensors using domain-insertion profiling. Nature Communications *7*, 12266.

Nam, K.H., Haitjema, C., Liu, X., Ding, F., Wang, H., DeLisa, M.P., and Ke, A. (2012). Cas5d protein processes pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvulg CRISPR-Cas system. Structure *20*, 1574–1584.

Niewoehner, O., Garcia-Doval, C., Rostøl, J.T., Berk, C., Schwede, F., Bigler, L., Hall, J., Marraffini, L.A., and Jinek, M. (1AD). Type III CRISPR–Cas systems produce cyclic oligoadenylate second messengers. Nature 1–25.

Nihongaki, Y., Kawano, F., Nakajima, T., and Sato, M. (2015). Photoactivatable CRISPR-Cas9 for optogenetic genome editing. Nat. Biotechnol. *33*, 755–760.

Nishimasu, H., Le Cong, Yan, W.X., Ran, F.A., Zetsche, B., Li, Y., Kurabayashi, A., Ishitani, R., Zhang, F., and Nureki, O. (2015). Crystal Structure of Staphylococcus aureus Cas9. Cell *162*, 1113–1126.

Nishimasu, H., Ran, F.A., Hsu, P.D., Konermann, S., Shehata, S.I., Dohmae, N., Ishitani, R., Zhang, F., and Nureki, O. (2014). Crystal Structure of Cas9 in Complex with Guide RNA and Target DNA. Cell *156*, 935–949.

Nissim, L., Perli, S.D., Fridkin, A., Perez-Pinera, P., and Lu, T.K. (2014). Multiplexed and programmable regulation of gene networks with an integrated RNA and CRISPR/Cas toolkit in human cells. Molecular Cell *54*, 698–710.

Nivala, J., Shipman, S.L., and Church, G.M. (2018). Spontaneous CRISPR loci generation in vivo by non-canonical spacer integration. Nature Microbiology *3*, 310–318.

Nuñez, J.K., Bai, L., Harrington, L.B., Hinder, T.L., and Doudna, J.A. (2016). CRISPR Immunological Memory Requires a Host Factor for Specificity. Molecular Cell *62*, 824–833.

Nuñez, J.K., Harrington, L.B., Kranzusch, P.J., Engelman, A.N., and Doudna, J.A. (2015a). Foreign DNA capture during CRISPR-Cas adaptive immunity. Nature *527*, 535–538.

Nuñez, J.K., Kranzusch, P.J., Noeske, J., Wright, A.V., Davies, C.W., and Doudna, J.A. (2014). Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. Nat. Struct. Mol. Biol. *21*, 528–534.

Nuñez, J.K., Lee, A.S.Y., Engelman, A., and Doudna, J.A. (2015b). Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. Nature *519*, 193–198.

Olson, W.K., Gorin, A.A., Lu, X., Hock, L.M., and Zhurkin, V.B. (1998). DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. Proceedings of the National Academy of Sciences *95*, 11163–11168.

Osawa, T., Inanaga, H., Sato, C., and Numata, T. (2015). Crystal Structure of the CRISPR-Cas RNA Silencing Cmr Complex Bound to a Target Analog. Molecular Cell *58*, 418–430.

Paez-Espino, D., Morovic, W., Sun, C.L., Thomas, B.C., Ueda, K.-I., Stahl, B., Barrangou, R., and Banfield, J.F. (2013). Strong bias in the bacterial CRISPR elements that confer immunity to phage. Nature Communications *4*, 1430.

Pingoud, A., and Jeltsch, A. (2001). Structure and function of type II restriction endonucleases. Nucleic Acids Research *29*, 3705–3727.

Plagens, A., Richter, H., Charpentier, E., and Randau, L. (2015). DNA and RNA interference mechanisms by CRISPR-Cas surveillance complexes. FEMS Microbiology Reviews *39*, 442–463.

Plagens, A., Tjaden, B., Hagemann, A., Randau, L., and Hensel, R. (2012). Characterization of the CRISPR/Cas subtype I-A system of the hyperthermophilic crenarchaeon Thermoproteus tenax. Journal of Bacteriology *194*, 2491–2500.

Plagens, A., Tripp, V., Daume, M., Sharma, K., Klingl, A., Hrle, A., Conti, E., Urlaub, H., and Randau, L. (2014). In vitro assembly and activity of an archaeal CRISPR-Cas type I-A Cascade interference complex. Nucleic Acids Research *42*, 5125–5138.

Polstein, L.R., and Gersbach, C.A. (2015). A light-inducible CRISPR-Cas9 system for control of endogenous gene activation. Nat. Chem. Biol. *11*, 198–200.

Pourcel, C., Salvignol, G., and Vergnaud, G. (2005). CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology (Reading, Engl.) *151*, 653–663.

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

Pyenson, N.C., and Marraffini, L.A. (2017). Type III CRISPR-Cas systems: when DNA cleavage just isn't enough. Current Opinion in Microbiology *37*, 150–154.

Ran, F.A., Cong, L., Yan, W.X., Scott, D.A., Gootenberg, J.S., Kriz, A.J., Zetsche, B., Shalem, O., Wu, X., Makarova, K.S., et al. (2015). In vivo genome editing using Staphylococcus aureus Cas9. Nature *520*, 186–191.

Ran, F.A., Hsu, P.D., Lin, C.-Y., Gootenberg, J.S., Konermann, S., Trevino, A.E., Scott, D.A., Inoue, A., Matoba, S., Zhang, Y., et al. (2013). Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. Cell *154*, 1380–1389.

Rath, D., Amlinger, L., Hoekzema, M., Devulapally, P.R., and Lundgren, M. (2015). Efficient programmable gene silencing by Cascade. Nucleic Acids Research *43*, 237–246.

Redding, S., Sternberg, S.H., Marshall, M., Gibb, B., Bhat, P., Guegler, C.K., Wiedenheft, B., Doudna, J.A., and Greene, E.C. (2015). Surveillance and Processing of Foreign DNA by the Escherichia coli CRISPR-Cas System. Cell *163*, 854–865.

Rice, P.A., Yang, S.-W., Mizuuchi, K., and Nash, H.A. (1996). Crystal Structure of an IHF-DNA Complex: A Protein-Induced DNA U-Turn. Cell *87*, 1295–1306.

Richter, C., and Fineran, P.C. (2013). The subtype I-F CRISPR-Cas system influences pathogenicity island retention in Pectobacterium atrosepticum via crRNA generation and Csy complex formation. Biochem. Soc. Trans. *41*, 1468–1474.

Richter, C., Dy, R.L., McKenzie, R.E., Watson, B.N.J., Taylor, C., Chang, J.T., McNeil, M.B., Staals, R.H.J., and Fineran, P.C. (2014). Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. Nucleic Acids Research *42*, 8516–8526.

Richter, C., Gristwood, T., Clulow, J.S., and Fineran, P.C. (2012). In vivo protein interactions and complex formation in the Pectobacterium atrosepticum subtype I-F CRISPR/Cas System. PLoS ONE *7*, e49549.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26.

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

Rollie, C., Graham, S., Rouillon, C., and White, M.F. (2018). Prespacer processing and specific integration in a Type I-A CRISPR system. Nucleic Acids Research *46*, 1007–1020.

Rollie, C., Schneider, S., Brinkmann, A.S., Bolt, E.L., and White, M.F. (2015). Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition. eLife *4*, e08716.

Rollins, M.F., Schuman, J.T., Paulus, K., Bukhari, H.S.T., and Wiedenheft, B. (2015). Mechanism of foreign DNA recognition by a CRISPR RNA-guided surveillance complex from Pseudomonas aeruginosa. Nucleic Acids Research *43*, 2216–2222.

Rutkauskas, M., Sinkunas, T., Songailiene, I., Tikhomirova, M.S., Siksnys, V., and Seidel, R. (2015). Directional R-Loop Formation by the CRISPR-Cas Surveillance Complex Cascade Provides Efficient Off-Target Site Rejection. CellReports *10*, 1534–1543.

Salvail-Lacoste, A., Di Tomasso, G., Piette, B.L., and Legault, P. (2013). Affinity purification of T7 RNA transcripts with homogeneous ends using ARiBo and CRISPR tags. Rna *19*, 1003–1014.

Samai, P., Pyenson, N., Jiang, W., Goldberg, G.W., Hatoum-Aslan, A., and Marraffini, L.A. (2015). Co-transcriptional DNA and RNA Cleavage during Type III CRISPR-Cas Immunity. Cell *161*, 1164–1174.

Sapranauskas, R., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P., and Siksnys, V. (2011). The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. Nucleic Acids Research *39*, 9275–9282.

Sashital, D.G., Jinek, M., and Doudna, J.A. (2011). An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3. Nat. Struct. Mol. Biol. *18*, 680–687.

Sashital, D.G., Wiedenheft, B., and Doudna, J.A. (2012). Mechanism of foreign DNA selection in a bacterial adaptive immune system. Molecular Cell *46*, 606–615.

Savitskaya, E., Semenova, E., Dedkov, V., Metlitskaya, A., and Severinov, K. (2013). High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in E. coli. RNA Biology *10*, 716–725.

Semenova, E., Jore, M.M., Datsenko, K.A., Semenova, A., Westra, E.R., Wanner, B., van der Oost, J., Brouns, S.J.J., and Severinov, K. (2011). Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc. Natl. Acad. Sci. U.S.a. *108*, 10098–10103.

Servick, K. (2015). SCIENCE POLICY. U.S. to review agricultural biotech regulations. Science *349*, 131–131.

Shaikh, T.R., Gao, H., Baxter, W.T., Asturias, F.J., Boisset, N., Leith, A., and Frank, J. (2008). SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs. Nat Protoc *3*, 1941–1974.

Shao, Y., and Li, H. (2013). Recognition and cleavage of a nonstructured CRISPR RNA by its processing endoribonuclease Cas6. Structure *21*, 385–393.

Shekhawat, S.S., and Ghosh, I. (2011). Split-protein systems: beyond binary protein-protein interactions. Curr Opin Chem Biol *15*, 789–797.

Sheth, R.U., Yim, S.S., Wu, F.L., and Wang, H.H. (2017). Multiplex recording of cellular events over time on CRISPR biological tape. Science *358*, 1457–1461.

Shipman, S.L., Nivala, J., Macklis, J.D., and Church, G.M. (2016). Molecular recordings by directed CRISPR spacer acquisition. Science *353*, aaf1175.

Shipman, S.L., Nivala, J., Macklis, J.D., and Church, G.M. (2017). CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria. Nature *547*, 345–349.

Shmakov, S., Savitskaya, E., Semenova, E., Logacheva, M.D., Datsenko, K.A., and Severinov, K. (2014). Pervasive generation of oppositely oriented spacers during CRISPR adaptation. Nucleic Acids Research *42*, 5907–5916.

Shmakov, S., Abudayyeh, O.O., Makarova, K.S., Wolf, Y.I., Gootenberg, J.S., Semenova, E., Minakhin, L., Joung, J., Konermann, S., Severinov, K., et al. (2015). Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. Molecular Cell *60*, 385–397.

Shmakov, S., Smargon, A., Scott, D., Cox, D., Pyzocha, N., Yan, W., Abudayyeh, O.O., Gootenberg, J.S., Makarova, K.S., Wolf, Y.I., et al. (2017). Diversity and evolution of class 2 CRISPR-Cas systems. Nat. Rev. Microbiol. *15*, 169–182.

Silas, S., Lucas-Elio, P., Jackson, S.A., Aroca-Crevillén, A., Hansen, L.L., Fineran, P.C., Fire, A.Z., and Sánchez-Amat, A. (2017a). Type III CRISPR-Cas systems can provide redundancy to counteract viral escape from type I systems. eLife *6*, aaf5573.

Silas, S., Makarova, K.S., Shmakov, S., Paez-Espino, D., Mohr, G., Liu, Y., Davison, M., Roux, S., Krishnamurthy, S.R., Fu, B.X.H., et al. (2017b). On the Origin of Reverse Transcriptase-Using CRISPR-Cas Systems and Their Hyperdiverse, Enigmatic Spacer Repertoires. MBio *8*, e00897–17.

Silas, S., Mohr, G., Sidote, D.J., Markham, L.M., Sánchez-Amat, A., Bhaya, D., Lambowitz, A.M., and Fire, A.Z. (2016). Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. Science *351*, aad4234.

Sinkunas, T., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P., and Siksnys, V. (2011). Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. Embo J. *30*, 1335–1342.

Sinkunas, T., Gasiunas, G., Waghmare, S.P., Dickman, M.J., Barrangou, R., Horvath, P., and Siksnys, V. (2013). In vitro reconstitution of Cascade-mediated CRISPR immunity in Streptococcus thermophilus. Embo J. *32*, 385–394.

Sokolowski, R.D., Graham, S., and White, M.F. (2014). Cas6 specificity and CRISPR RNA loading in a complex CRISPR-Cas system. Nucleic Acids Research *42*, 6532–6541.

Staals, R.H.J., Jackson, S.A., Biswas, A., Brouns, S.J.J., Brown, C.M., and Fineran, P.C. (2016). Interference-driven spacer acquisition is dominant over naive and primed adaptation in a native CRISPR-Cas system. Nature Communications *7*, 12853.

Staals, R.H.J., Zhu, Y., Taylor, D.W., Kornfeld, J.E., Sharma, K., Barendregt, A., Koehorst, J.J., Vlot, M., Neupane, N., Varossieau, K., et al. (2014). RNA targeting by the type III-A CRISPR-Cas Csm complex of Thermus thermophilus. Molecular Cell *56*, 518–530.

Sternberg, S.H., and Doudna, J.A. (2015). Expanding the Biologist's Toolkit with CRISPR-Cas9. Molecular Cell *58*, 568–574.

Sternberg, S.H., Haurwitz, R.E., and Doudna, J.A. (2012). Mechanism of substrate selection by a highly specific CRISPR endoribonuclease. Rna *18*, 661–672.

Sternberg, S.H., LaFrance, B., Kaplan, M., and Doudna, J.A. (2015). Conformational control of DNA target cleavage by CRISPR-Cas9. Nature *527*, 110–113.

Sternberg, S.H., Redding, S., Jinek, M., Greene, E.C., and Doudna, J.A. (2014). DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature *507*, 62–67.

Sternberg, S.H., Richter, H., Charpentier, E., and Qimron, U. (2016). Adaptation in CRISPR-Cas Systems. Molecular Cell *61*, 797–808.

Suloway, C., Pulokas, J., Fellmann, D., Cheng, A., Guerra, F., Quispe, J., Stagg, S., Potter, C.S., and Carragher, B. (2005). Automated molecular microscopy: the new Leginon system. J. Struct. Biol. *151*, 41–60.

Swarts, D.C., and Jinek, M. (2018). Cas9 versus Cas12a/Cpf1: Structure-function comparisons and implications for genome editing. WIREs RNA *61*, e1481.

Swarts, D.C., Mosterd, C., van Passel, M.W.J., and Brouns, S.J.J. (2012). CRISPR interference directs strand specific spacer acquisition. PLoS ONE *7*, e35888.

Swarts, D.C., van der Oost, J., and Jinek, M. (2017). Structural Basis for Guide RNA Processing and Seed- Dependent DNA Targeting by CRISPR-Cas12a. Molecular Cell *66*, 221–233.e224.

Szczelkun, M.D., Tikhomirova, M.S., Sinkunas, T., Gasiunas, G., Karvelis, T., Pschera, P., Siksnys, V., and Seidel, R. (2014). Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. Proc. Natl. Acad. Sci. U.S.a. *111*, 9798–9803.

Tamulaitis, G., Kazlauskiene, M., Manakova, E., Venclovas, Č., Nwokeoji, A.O., Dickman, M.J., Horvath, P., and Siksnys, V. (2014). Programmable RNA shredding by the type III-A CRISPR-Cas system of Streptococcus thermophilus. Molecular Cell *56*, 506–517.

Tang, G., Peng, L., Baldwin, P.R., Mann, D.S., Jiang, W., Rees, I., and Ludtke, S.J. (2007). EMAN2: An extensible image processing suite for electron microscopy. J. Struct. Biol. *157*, 38–46.

Tang, W., and Liu, D.R. (2018). Rewritable multi-event analog recording in bacterial and mammalian cells. Science *360*.

Taylor, D.W., Zhu, Y., Staals, R.H.J., Kornfeld, J.E., Shinkai, A., van der Oost, J., Nogales, E., and Doudna, J.A. (2015). Structural biology. Structures of the CRISPR-Cmr complex reveal mode of RNA target positioning. Science *348*, 581–585.

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinformatics *14*, 178–192.

Toro, N., Martínez-Abarca, F., and González-Delgado, A. (2017). The Reverse Transcriptases Associated with CRISPR-Cas Systems. Scientific Reports *7*, 7089.

Tsai, S.Q., Wyvekens, N., Khayter, C., Foden, J.A., Thapar, V., Reyon, D., Goodwin, M.J., Aryee, M.J., and Joung, J.K. (2014). Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. Nat. Biotechnol. *32*, 569–576.

Tsui, T.K.M., and Li, H. (2015). Structure Principles of CRISPR-Cas Surveillance and Effector Complexes. Annu. Rev. Biophys. *44*, 229–255.

van der Oost, J., Jore, M.M., Westra, E.R., Lundgren, M., and Brouns, S.J.J. (2009). CRISPR-based adaptive and heritable immunity in prokaryotes. Trends Biochem. Sci. *34*, 401–407.

van der Oost, J., Westra, E.R., Jackson, R.N., and Wiedenheft, B. (2014). Unravelling the structural and mechanistic basis of CRISPR–Cas systems. Nat. Rev. Microbiol. *12*, 479–492.

van Erp, P.B.G., Jackson, R.N., Carter, J., Golden, S.M., Bailey, S., and Wiedenheft, B. (2015). Mechanism of CRISPR-RNA guided recognition of DNA targets in Escherichia coli. Nucleic Acids Research *43*, 8381–8391.

van Heel, M., Harauz, G., Orlova, E.V., Schmidt, R., and Schatz, M. (1996). A new generation of the IMAGIC image processing system. J. Struct. Biol. *116*, 17–24.

Vestergaard, G., Garrett, R.A., and Shah, S.A. (2014). CRISPR adaptive immune systems of Archaea. RNA Biology *11*, 156–167.

Wang, J., Li, J., Zhao, H., Sheng, G., Wang, M., Yin, M., and Wang, Y. (2015). Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems. Cell *163*, 840–853.

Wang, R., Li, M., Gong, L., Hu, S., and Xiang, H. (2016). DNA motifs determining the accuracy of repeat duplication during CRISPR adaptation in Haloarcula hispanica. Nucleic Acids Research *44*, 4266–4277.

Wang, X., and Higgins, N.P. (1994). “Muprints” of the lac operon demonstrate physiological control over the randomness of in vivo transposition. Molecular Microbiology *12*, 665–677.

Wei, Y., Chesne, M.T., Terns, R.M., and Terns, M.P. (2015a). Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in Streptococcus thermophilus. Nucleic Acids Research gku1407.

Wei, Y., Terns, R.M., and Terns, M.P. (2015b). Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. Genes Dev. *29*, 356–361.

Westra, E.R., Buckling, A., and Fineran, P.C. (2014). CRISPR–Cas systems: beyond adaptive immunity. Nat. Rev. Microbiol. *12*, 317–326.

Westra, E.R., Nilges, B., van Erp, P.B.G., van der Oost, J., Dame, R.T., and Brouns, S.J.J. (2012). Cascade-mediated binding and bending of negatively supercoiled DNA. RNA Biology *9*, 1134–1138.

Wiedenheft, B., Lander, G.C., Zhou, K., Jore, M.M., Brouns, S.J.J., van der Oost, J., Doudna, J.A., and Nogales, E. (2011a). Structures of the RNA-guided surveillance complex from a bacterial immune system. Nature *477*, 486–489.

Wiedenheft, B., van Duijn, E., Bultema, J.B., Bultema, J., Waghmare, S.P., Waghmare, S., Zhou, K., Barendregt, A., Westphal, W., Heck, A.J.R., et al. (2011b). RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. Proc. Natl. Acad. Sci. U.S.a. *108*, 10092–10097.

Wiedenheft, B., Zhou, K., Jinek, M., Coyle, S.M., Ma, W., and Doudna, J.A. (2009). Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. Structure *17*, 904–912.

Wright, A.V., and Doudna, J.A. (2016). Protecting genome integrity during CRISPR immune adaptation. Nat. Struct. Mol. Biol. *23*, 876–883.

Wright, A.V., Nuñez, J.K., and Doudna, J.A. (2016). Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. Cell *164*, 29–44.

Wright, A.V., Sternberg, S.H., Taylor, D.W., Staahl, B.T., Bardales, J.A., Kornfeld, J.E., and Doudna, J.A. (2015). Rational design of a split-Cas9 enzyme complex. Proc. Natl. Acad. Sci. U.S.a. *112*, 2984–2989.

Xiao, Y., Ng, S., Nam, K.H., and Ke, A. (2017). How type II CRISPR-Cas establish immunity through Cas1-Cas2-mediated spacer integration. Nature *550*, 137–141.

Yan, W.X., Chong, S., Zhang, H., Makarova, K.S., Koonin, E.V., Cheng, D.R., and Scott, D.A. (2018). Cas13d Is a Compact RNA-Targeting Type VI CRISPR Effector Positively Modulated by a WYL-Domain-Containing Accessory Protein. Molecular Cell *70*, 327–339.e5.

Yoganand, K.N.R., Sivathanu, R., Nimkar, S., and Anand, B. (2017). Asymmetric positioning of Cas1–2 complex and Integration Host Factor induced DNA bending guide the unidirectional homing of protospacer in CRISPR-Cas type I-E system. Nucleic Acids Research *45*, 367–381.

Yosef, I., Goren, M.G., and Qimron, U. (2012). Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucleic Acids Research *40*, 5569–5576.

Yosef, I., Shitrit, D., Goren, M.G., Burstein, D., Pupko, T., and Qimron, U. (2013). DNA motifs determining the efficiency of adaptation into the Escherichia coli CRISPR array. Proc. Natl. Acad. Sci. U.S.a. *110*, 14396–14401.

Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., Essletzbichler, P., Volz, S.E., Joung, J., van der Oost, J., Regev, A., et al. (2015a). Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell *163*, 759–771.

Zetsche, B., Volz, S.E., and Zhang, F. (2015b). A split-Cas9 architecture for inducible genome editing and transcription modulation. Nat. Biotechnol. *33*, 139–142.

Zhang, Y., Heidrich, N., Ampattu, B.J., Gunderson, C.W., Seifert, H.S., Schoen, C., Vogel, J., and Sontheimer, E.J. (2013). Processing-independent CRISPR RNAs limit natural transformation in Neisseria meningitidis. Molecular Cell *50*, 488–503.

Zhang, Y., Rajan, R., Seifert, H.S., Mondragón, A., and Sontheimer, E.J. (2015). DNase H Activity of Neisseria meningitidis Cas9. Molecular Cell *60*, 242–255.

Zhao, H., Sheng, G., Wang, J., Wang, M., Bunkoczi, G., Gong, W., Wei, Z., and Wang, Y. (2014). Crystal structure of the RNA-guided immune surveillance Cascade complex in Escherichia coli. Nature *515*, 147–150.

Zheng, G., Lu, X.-J., and Olson, W.K. (2009). Web 3DNA--a web server for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures. Nucleic Acids Research *37*, W240–W246.

Zheng, S.Q., Palovcak, E., Armache, J.-P., Verba, K.A., Cheng, Y., and Agard, D.A. (2017). MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. Nat. Methods *14*, 331–332.

Zuris, J.A., Thompson, D.B., Shu, Y., Guilinger, J.P., Bessen, J.L., Hu, J.H., Maeder, M.L., Joung, J.K., Chen, Z.-Y., and Liu, D.R. (2015). Cationic lipid-mediated delivery of proteins enables efficient protein-based genome editing in vitro and in vivo. Nat. Biotechnol. *33*, 73–80.

APPENDIX I

Rational design of a split-Cas9 enzyme complex

A portion of the work presented in this chapter has been published previously as part of the following paper: Wright, A.V., Sternberg, S.H., Taylor, D.W., Staahl, B.T., Bardales, J.A., Kornfeld, J.E., and Doudna, J.A. (2015). Rational design of a split-Cas9 enzyme complex. *Proc. Natl. Acad. Sci. USA* **112** (10), 2984-2989.

Samuel Sternberg designed the split-Cas9 constructs. Addison Wright designed and performed biochemical experiments, with input from Samuel Sternberg. David Taylor and Jack Kornberg performed single particle electron microscopy experiments. Brett Staahl, Jorge Bardales, and Addison Wright performed genome editing experiments. Addison Wright, Samuel Sternberg, and Jennifer Doudna interpreted the data and wrote the manuscript.

I.1 Abstract

Cas9, an RNA-guided DNA endonuclease found in clustered regularly interspaced short palindromic repeats (CRISPR) bacterial immune systems, is a versatile tool for genome editing, transcriptional regulation and cellular imaging applications. Structures of *Steptococcus pyogenes* Cas9 alone or bound to single-guide RNA (sgRNA) and target DNA revealed a bi-lobed protein architecture that undergoes major conformational changes upon guide RNA and DNA binding. To investigate the molecular determinants and relevance of the inter-lobe rearrangement for target recognition and cleavage, we designed a split Cas9 enzyme in which the nuclease lobe and α-helical lobe are expressed as separate polypeptides. Although the lobes do not interact on their own, the sgRNA recruits them into a ternary complex that recapitulates the activity of full-length Cas9 and catalyzes site-specific DNA cleavage. The use of a modified sgRNA abrogates split-Cas9 activity by preventing dimerization, allowing for the development of an inducible dimerization system. We propose that split-Cas9 can act as a highly regulatable platform for genome engineering applications.

I.2 Significance statement

Bacteria have evolved clustered regularly interspaced short palindromic repeats (CRISPR) together with CRISPR-associated (Cas) proteins to defend themselves against viral infection. RNAs derived from the CRISPR locus assemble with Cas proteins into programmable DNA-targeting complexes that destroy DNA molecules complementary to the guide RNA. In type II CRISPR-Cas systems, the Cas9 protein binds and cleaves target DNA sequences at sites complementary to a 20-nucleotide (nt) guide RNA sequence. This activity has been harnessed for a wide range of genome engineering applications. This study explores the structural features that enable Cas9 to bind and cleave target DNAs, and the results suggest a new way of regulating Cas9 by physical separation of the catalytic domains from the rest of the protein scaffold.

I.3 Introduction

Bacteria use RNA-guided adaptive immune systems encoded by CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats)-Cas (CRISPR-associated) genomic loci to defend against invasive DNA (Barrangou and Marraffini, 2014; van der Oost et al., 2014). In type II CRISPR-Cas systems, a single enzyme called Cas9 is responsible for targeting and cleavage of foreign DNA (Sapranauskas et al., 2011). The ability to program Cas9 for DNA cleavage at sites defined by engineered single-guide RNAs (sgRNAs) (Jinek et al., 2012) has led to its adoption as a robust and versatile platform for genome engineering (for recent reviews, see (Doudna and Charpentier, 2014; Hsu et al., 2014; Mali et al., 2013)).

Cas9 contains two nuclease active sites that function together to generate DNA double-strand breaks (DSBs) at sites complementary to the 20-nt guide RNA sequence and adjacent to a PAM (protospacer adjacent motif). Structural studies of the *Streptococcus pyogenes* Cas9 showed that the protein exhibits a bi-lobed architecture comprising the catalytic nuclease lobe and the α-helical lobe of the enzyme (Jinek et al., 2014). Electron microscopy (EM) studies and comparisons to X-ray crystal structures with and without a bound guide RNA and target DNA revealed a large-scale conformational rearrangement of the two lobes relative to each other upon nucleic acid binding (Jinek et al., 2014; Nishimasu et al., 2014). Strikingly, RNA binding induces the nuclease lobe to rotate approximately 100° relative to the α-helical lobe, generating a nucleic-acid binding cleft that can accommodate DNA, and interactions between the two lobes appear to be mediated primarily through contacts with the bound nucleic acid rather than direct protein-protein contacts (Jinek et al., 2014; Nishimasu et al., 2014). These observations suggested that the two structural lobes of Cas9 might be separable into independent polypeptides that retain the ability to assemble into an active enzyme complex. Such a system would enable analysis of the functionally distinct properties of each Cas9 structural region, and might offer a unique mechanism for controlling active protein assembly.

Here we show that two distinct Cas9 polypeptides encompassing the α-helical and nuclease lobes can be stably expressed and purified. Filter binding and negative-stain EM experiments demonstrate that the split-Cas9 assembles with sgRNA into a ternary complex resembling that of full-length Cas9–RNA. Furthermore, DNA cleavage assays reveal that the enzymatic activity of split-Cas9 closely mimics that of WT Cas9. Split-Cas9 is functional for genome editing in human cells with full-length sgRNAs, but can be inactivated with shortened sgRNAs that give rise to destabilized complexes. Together these data show how the Cas9 protein can be re-engineered as a split enzyme whose assembly and function is regulatable by the sgRNA, providing a new platform for controlled use of Cas9 for genome engineering applications in cells.

I.4 Results

I.4.1 Design and functional validation of split-Cas9

The nuclease lobe of Cas9 includes the RuvC and HNH nuclease domains, as well as a C-terminal domain that is involved in PAM recognition (Fig. I.1a) (Anders et al., 2014; Jinek et al., 2014; Nishimasu et al., 2014). The RuvC domain comprises three distinct motifs: motifs II and III are interrupted by the HNH domain, and motifs I and II are interrupted by a large lobe composed entirely of α-helices. This α-helical lobe, also referred to as the recognition (REC) lobe (Nishimasu et al., 2014), forms a broad cleft that makes extensive contacts with the sgRNA and target DNA. We previously showed that the α-helical lobe undergoes a large rotation relative to the nuclease lobe upon guide RNA binding to create a central channel where target DNA is bound (Jinek et al., 2014).

Using available crystal structures as a guide, we designed a split-Cas9 in which the native structure of both lobes was kept as intact as possible (Fig. I.1a). In particular, rather than simply split the full-length Cas9 sequence internally at a single junction, we constructed the nuclease lobe by directly linking the N-terminal RuvCI motif to the remainder of the nuclease lobe located ~650 amino acids away in primary sequence, with the intervening polypeptide comprising the α-helical lobe. Two crossover points between the lobes occur at residues ~56 and ~720 (Fig. I.1b): the C-terminal connection is disordered in both apo-Cas9 and sgRNA/DNA-bound structures, and the N-terminal connection occurs between the RuvCI motif and the bridge helix. We connected residue

E57 from RuvCI with residue G729 from RuvCII using a three-amino acid linker, and removed a short, poorly conserved α-helix from the RuvCII motif that does not appear to play an important structural role in the sgRNA/DNA-bound state (Fig. I.1b). The α-helical lobe spans residues G56–S714, with the N-terminus encompassing the entirety of the bridge helix.

To determine whether the lobes could function as separate polypeptides, we separately over-expressed both lobes in *Escherichia coli* and purified them by affinity and size-exclusion chromatography (Fig. I.1c, I.2). We investigated whether split-Cas9 (α-helical lobe plus nuclease lobe) would recapitulate the activity of WT Cas9 using a standard cleavage assay with sgRNA and a radiolabeled double-stranded DNA (dsDNA) target (Fig. I.1d). No cleavage was observed with either lobe individually, but the reconstituted split-Cas9 enzyme complex exhibited robust target DNA cleavage (Fig. I.1d, I.2). Split-Cas9 maintained the same site and pattern of cleavage as WT Cas9, including the “trimming” of the non-target strand that we observed previously (Jinek et al., 2012), and functioned equally well with a dual guide RNA composed of crRNA and tracrRNA (Fig. I.3). In addition, we confirmed that split-Cas9 activity was dependent on complementarity between the sgRNA and target DNA as well as the presence of a 5’-NGG-3’ PAM (Fig. I.3).

When we investigated the kinetics of DNA cleavage under pseudo-first order conditions using excess enzyme, we found that split-Cas9 was ~10-fold slower than WT, though it reached the same endpoint after 5 minutes (Fig. I.3 and Table I.1). This may result from slower kinetics of protein-RNA complex formation, a reduced rate of dsDNA recognition and unwinding, or a minor defect in nuclease domain activation. DNA binding experiments using nuclease-inactive split-dCas9 (D10A/H840A mutations) revealed a significantly weaker affinity of split-Cas9 for target DNA than WT Cas9 (Fig. I.4), suggesting that slower kinetics of dsDNA binding likely limit the observed rate of cleavage. Collectively, these results demonstrate that the enzymatic activity of WT Cas9 does not require a direct linkage between the α-helical and nuclease lobes, though their physical connection within RNA-protein complexes increases the affinity for the target DNA substrate. Remarkably, while previous work shows that the RNA-induced large-scale rearrangement of both lobes is necessary for WT Cas9 to achieve an active conformation (Jinek et al., 2014), our experiments reveal that the sgRNA is entirely sufficient to recruit and dimerize the separate lobes into an active enzyme complex. Furthermore, communication through the sgRNA enables PAM recognition, dsDNA unwinding, and DNA cleavage, despite the absence of extensive protein-protein interactions between the lobes.

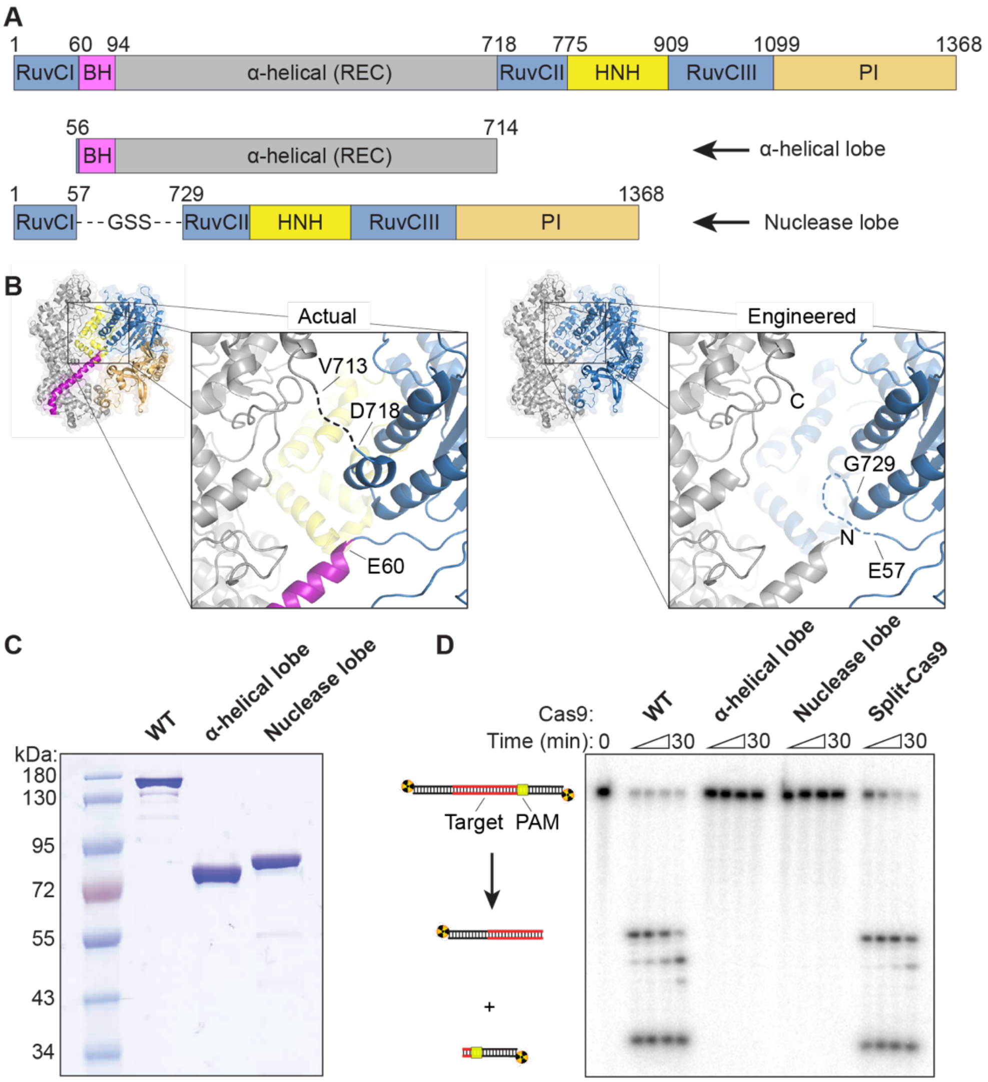


Figure I.1 | Cas9 can be split into two separate polypeptides that retain the ability to catalyze RNA-guided dsDNA cleavage. (**a**) Domain organization of WT Cas9 (top) and split-Cas9 (bottom), composed of the α-helical lobe and nuclease lobe. Domain junctions are numbered according to Nishimasu *et al.* (Nishimasu et al., 2014). BH, bridge helix; REC, recognition lobe; PI, PAM-interacting. The PI domain can be further subdivided into Topo-homology and C-terminal domains (Jinek et al., 2014). (**b**) Crystal structures of sgRNA/DNA-bound Cas9 (PDB ID: 4OO8) (Nishimasu et al., 2014) colored according to domain (left) or by lobe (right), with the α-helical and nuclease lobes depicted in grey and blue, respectively. Nucleic acids are omitted for clarity. In the observed interface between the lobes (inset, left), the dashed line represents a disordered linker spanning residues V713–D718. In the engineered interface (inset, right), the dashed line represents a GGS linker connecting E57 to G729, and new N- and C-termini of the α-helical lobe are shown. (**c**) Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) analysis of purified WT Cas9 (159 kDa), the α-helical lobe (77 kDa), and the nuclease lobe (81 kDa). The gel was stained with Coomassie Brilliant Blue. (**d**) DNA cleavage assay with the indicated Cas9 construct, analyzed by denaturing PAGE. Reactions contained ~1 nM radiolabeled dsDNA and 100 nM protein–sgRNA complex; split-Cas9 contained a two-fold molar excess of the α-helical lobe. Quantified data and kinetic analysis can be found in Fig. I.3 and Table I.1.

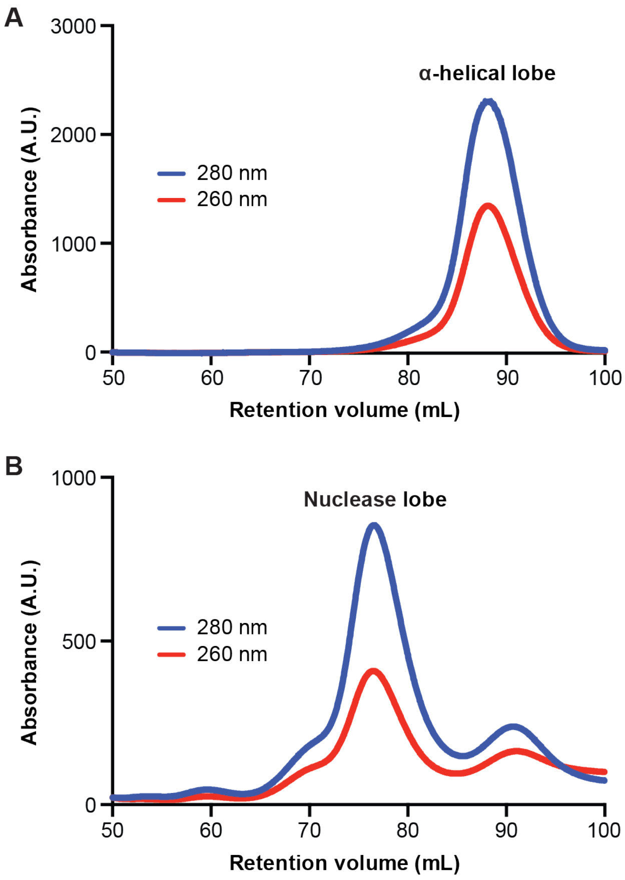


Figure I.2 | Size exclusion chromatograms of purified α-helical and nuclease lobes. Following cleavage of the affinity tag by TEV protease and further clean-up using ortho-Ni-NTA and ion exchange columns (α-helical lobe) or an ortho-Ni-NTA column alone (nuclease lobe), the polypeptides were concentrated and injected onto a HiLoad 16/60 Superdex 200 gel filtration column. The α-helical (77 kDa) and nuclease lobes (81 kDa) eluted at 88.4 mL and 76.8 mL, respectively. Both polypeptides were soluble and exhibited consistent activity across multiple rounds of freeze-thawing.

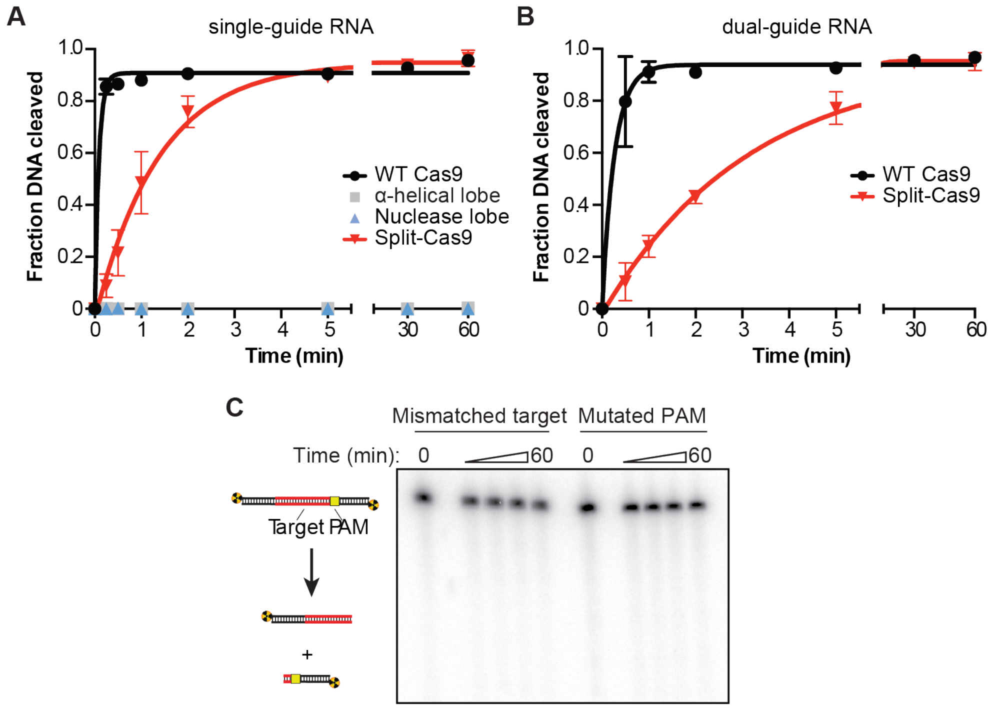


Figure I.3 | Split-Cas9 activity is mediated by single-guide and dual-guide RNAs, and requires RNA:DNA complementarity and a PAM. (**a**) DNA cleavage time courses using a single-guide RNA and

WT Cas9, individual α-helical and nuclease lobes, or split-Cas9. Values for WT and split-Cas9 were averaged from three independent experiments, and error bars represent the standard deviation. Rate constants can be found in Table I.1. (**b**) DNA cleavage time courses using a dual-guide RNA (crRNA:tracrRNA hybrid) and WT Cas9 or split-Cas9. Data are presented as in *A*. (**c**) DNA cleavage assay with split-Cas9 and DNA substrates containing a mismatched target or mutated PAM (Table I.2), analyzed by denaturing PAGE. Reactions contained ~1 nM radiolabeled dsDNA and 100 nM Cas9–sgRNA complex.

	Rate constant (k_{obs}) for indicated sgRNA		
Protein	**Full length**	**Δhairpin1**	**Δhairpins1-2**
WT Cas9	11.3 ± 0.9 min^{-1}	9.2 ± 0.4 min^{-1}	6.0 ± 0.6 min^{-1}
α-helical lobe	N.D.	–	–
Nuclease lobe	N.D.	–	–
Split-Cas9	1.0 ± 0.2 min^{-1}	(5.5 ± 0.8) $\times 10^{-4}$ min^{-1}	N.D.

Table I.1 | Cleavage rate constants for WT and split-Cas9 using different sgRNA constructs. Three independent experiments were performed for each conditions, and the values represent the mean ± S.E.M. N.D., cleavage not detected. –, experiment not performed.

I.4.2 sgRNA motifs recruit both Cas9 lobes to form a ternary complex

We next wanted to investigate RNA molecular determinants that promote heterodimerization of the α-helical and nuclease lobes. Crystal structures of sgRNA/DNA-bound Cas9 show that the spacer (guide) and stem-loop motifs at the 5' end of the sgRNA primarily contact the α-helical lobe, whereas two hairpins at the 3' end bind the outside face of the nuclease lobe (Fig. I.5a). The nexus motif, recently shown to be critical for activity (Briner et al., 2014), occupies a central position between the lobes and forms extensive interactions with the bridge helix. Based on this interaction profile, we generated a full-length sgRNA and two shorter sgRNA constructs that were selectively truncated from either the 5' or 3' end (Fig. I.5b), and determined their affinities for WT Cas9, the individual α-helical and nuclease lobes, and split-Cas9 using a filter binding assay.

The full-length sgRNA is bound by WT Cas9 with an equilibrium dissociation constant (K_d) of 10 ± 2 pM, whereas the lobes individually and together have K_d values in the range of 0.2–0.8 nM (Fig. I.5c and Table I.2). The difference between WT and split-Cas9 likely reflects the increased entropic cost required to assemble a ternary versus binary complex. Interestingly, WT Cas9 bound a truncated sgRNA comprising only the 3' hairpins (Δspacer–nexus) with an affinity that was indistinguishable from the full-length sgRNA (Fig. I.5d and Table I.2), indicating that these hairpins provide the major source of binding energy for the WT protein-RNA complex (Hsu et al., 2013). Consistent with the crystal structure, the nuclease lobe still bound the 5'-truncated sgRNA as tightly as the full-length sgRNA, while the affinity of the α-helical lobe was reduced by over three orders of magnitude (K_d > 100 nM).

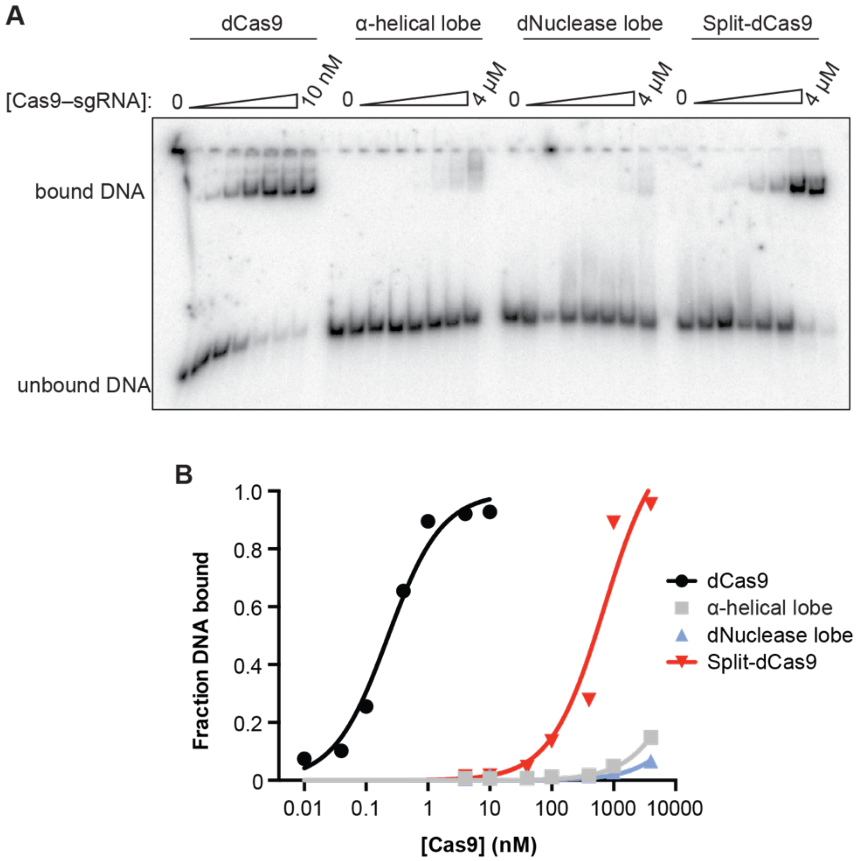


Figure I.4 | Split-Cas9 exhibits substantially weaker binding affinity for target DNA than WT Cas9. (**a**) Radiolabeled target dsDNA was incubated with increasing concentrations of Cas9–sgRNA complexes using catalytically inactive mutants of WT Cas9 and the nuclease lobe, and reaction products were resolved by native PAGE. The distinct Cas9 constructs in each titration are indicated (top). (**b**) Quantified binding data from (a). Split-dCas9–RNA binds dsDNA with an apparent equilibrium dissociation constant of ~700 nM, which is more than 3 orders of magnitude greater than that determined for dCas9–RNA (K_d ≈ 0.2 nM). However, the apparent affinity measured here is likely to be much weaker than the actual affinity, since the low split-dCas9–sgRNA concentrations that were tested will also favor dissociation of the ternary complex formed between the sgRNA, α-helical lobe, and nuclease lobe. Thus, the observed binding curve is likely a convolution of equilibria between the protein and sgRNA, and between the protein–sgRNA complex and dsDNA. Individual lobes together with sgRNA do not appreciably bind dsDNA at the tested concentrations.

We similarly reasoned that removing the two hairpins from the 3’ end of the sgRNA (Δhairpins) would selectively perturb interactions with the nuclease lobe. Indeed, the affinity of the 3’-truncated sgRNA for the nuclease lobe decreased by over three orders of magnitude relative to full-length sgRNA (K_d >100 nM), whereas the affinity of the α-helical lobe was unchanged (Fig. I.5d and Table I.2). Our results demonstrate that sgRNA truncations specifically destabilize binding to only one of the two lobes, and that the affinity of split-Cas9 is limited by the highest affinity interaction with either lobe. These findings highlight the multiple, independent molecular contacts formed between the sgRNA and the two lobes of Cas9.

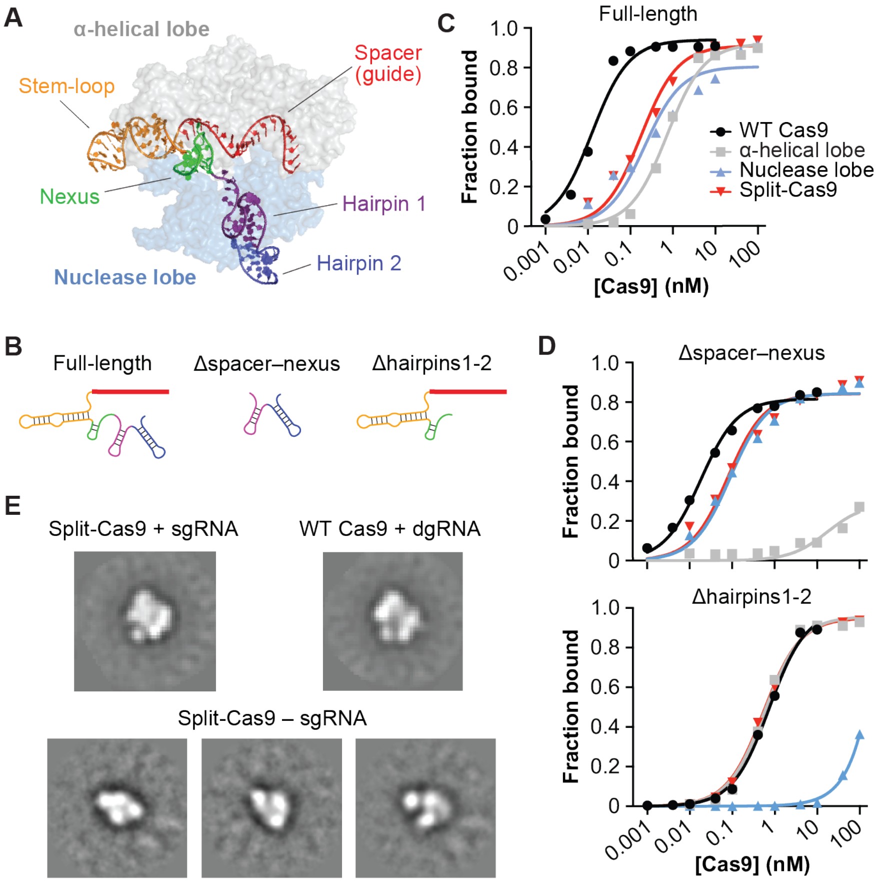


Figure I.5 | Split-Cas9 assembly requires the sgRNA. (**a**) Crystal structure of sgRNA/DNA-bound Cas9 (PDB ID: 4OO8) (Nishimasu et al., 2014): Cas9 is colored by lobe and shown as a transparent surface, the sgRNA is colored by motif according to Briner *et al.*, 2014 , and the DNA is omitted for clarity. (**b**) Cartoon representations of full-length and truncated sgRNA variants used in binding experiments; specific motifs of the sgRNA are colored as in (a). (**c-d**) Results from binding experiments using full-length sgRNA (c), and Δhairpins1-2 and Δspacer-nexus sgRNA truncations (d). Radiolabeled RNAs were incubated with increasing concentrations of WT Cas9, individual α-helical and nuclease lobes, or split-Cas9, and the fraction of protein-bound RNA was determined by nitrocellulose filter binding. Equilibrium dissociation constants (K_d) determined from three independent experiments are shown in Table I.2. (**e**) Reference-free class averages from negative-stain EM images of split-Cas9 reconstituted with single-guide RNA (top left), WT Cas9 reconstituted with dual-guide RNA (top right), and split-Cas9 in the absence of guide RNA (bottom). For split-Cas9 without sgRNA, several class averages are shown. The width of the boxes corresponds to ~336 Å. Data with WT Cas9 is adapted from Jinek *et al.*, 2014.

Based on our binding data and on the minimal contacts observed between the two Cas9 lobes in available structures, we hypothesized that the sgRNA would be required for heterodimerization of the α-helical and nuclease lobes. To test this, we performed analytical negative-stain electron microscopy with the polypeptides corresponding to each lobe alone and together in the presence and absence of sgRNA. Raw micrographs of a sample containing both polypeptides and the sgRNA revealed bi-lobed densities that had dimensions consistent with our earlier reconstructions of the Cas9–RNA complex (Fig. I.6) (Jinek et al., 2014), and the resulting class averages were indistinguishable from those obtained using WT Cas9 (Fig. 1.5e). In contrast, we observed smaller particles that had dimensions more consistent with single lobes when the polypeptides were mixed together in the absence of sgRNA (Fig. I.5e, Fig. I.6). These results indicate that a full-length sgRNA acts as a molecular scaffold in dimerizing the two lobes.

	Equilibrium dissociation constant (K_d) for indicated sgRNA*		
Protein	**Full length**	**Δhairpins1–2**	**Δspacer–nexus**
WT Cas9	10 ± 2 pM	0.86 ± 0.12 nM	16 ± 2 pM
α-helical lobe	0.75 ± 0.12 nM	0.70 ± 0.13 nM	> 100 nM
Nuclease lobe	0.30 ± 0.07 nM	> 100 nM	0.17 ± 0.06 nM
Split-Cas9	0.23 ± 0.04 nM	1.05 ± 0.05 nM	0.17 ± 0.07 nM

* Three independent experiments were performed for each condition, and the values represent the mean ± S.E.M.

Table I.2 | Equilibrium dissociation constants for protein–sgRNA interactions.

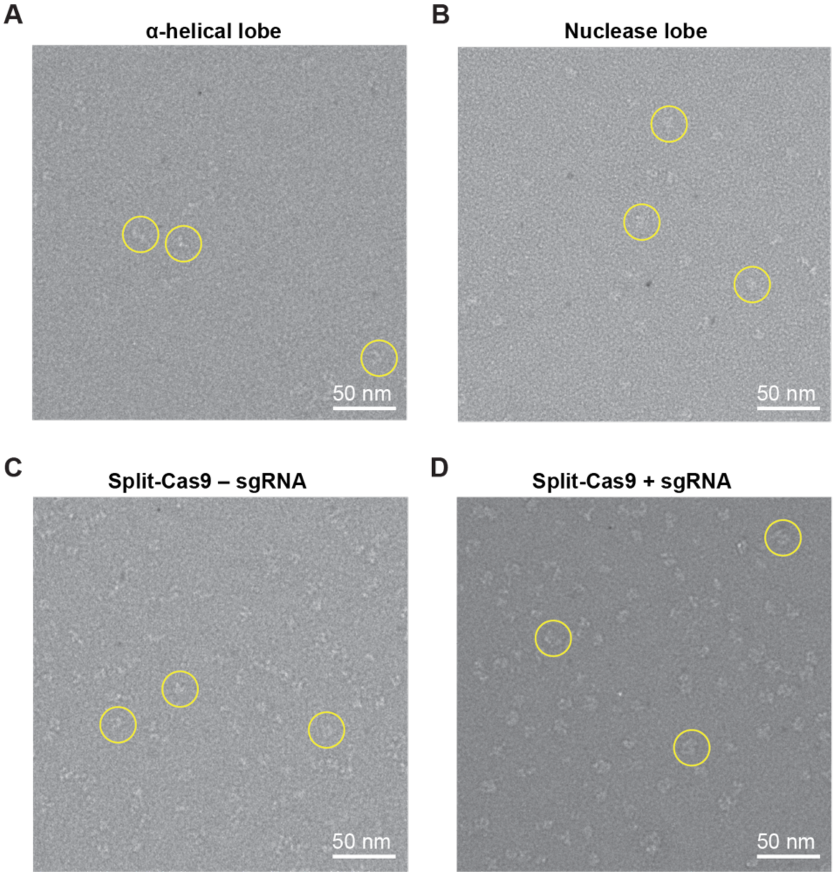


Fig. I.6 | Split-Cas9 heterodimerization requires the sgRNA. (**a-d**) Raw electron micrographs of negatively-stained α-helical and nuclease lobes alone (a,b), together (c), or together with sgRNA (d). Particles having dimensions consistent with WT Cas9–RNA complexes, and thus indicative of heterodimer formation, are only observed in the presence of sgRNA. Representative particles are circled (yellow), and the scale bar indicates 50 nm.

I.4.3 Split-Cas9 functions in mammalian cells for genome editing

To determine whether split-Cas9 would retain the ability to generate site-specific genomic edits *in vivo*, we targeted the *EMX1* locus in HEK293T cells by nucleofection using reconstituted Cas9–sgRNA ribonucleoprotein (RNP) complexes (Fig. I.7a) (Lin et al., 2014). Split-Cas9 generated indels with efficiencies of up to 0.6% and 2% in cells that were unsynchronized or nocodazole synchronized, respectively, compared to 22% and 34% with WT Cas9 (Fig. I.7b). The reduced levels of editing may be due in part to disruption of the ternary complex during dilution and nucleofection, as the complex is limited by the affinity of the α-helical lobe for sgRNA, or to slower kinetics of DNA cleavage in cells. Additionally, because each copy of sgRNA must recruit both lobes to form an active complex, we suspect that the activity in cells may be particularly sensitive to the stoichiometry between the sgRNA and either lobe. In agreement with this, we found that the *in vitro* DNA cleavage activity of split-Cas9 decreased as the sgRNA concentration was increased above that of both lobes (Fig. I.8), suggesting that excess sgRNA may titrate the lobes apart from each other. While our results leave room for optimization of split-Cas9 activity in cells, they demonstrate that the intrinsic genome editing capabilities are retained when Cas9 comprises two individual polypeptides.

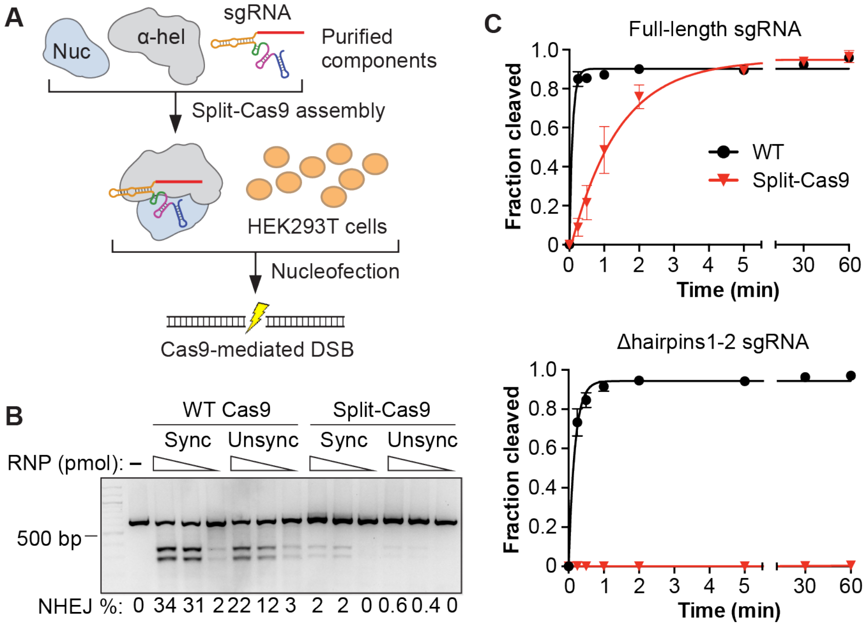


Figure I.7 | Genomic editing function and selective inactivation of split-Cas9. (**a**). Schematic of the split-Cas9 RNP nucleofection assay using a full-length *EMX1*-targeting sgRNA. Illustration and protocol adapted from Lin *et al*., 2014. (**b**) Analysis of editing efficiencies by nonhomologous end joining (NHEJ) using a T7 endonuclease I assay and agarose gel electrophoresis. Cells were nucleofected with 100, 30, or 10 pmol of WT or split-Cas9 ribonucleoprotein (RNP) complexes after arrest at mitosis with nocodazole (Sync) or during normal growth (Unsync). Editing efficiencies are shown at the bottom. (**c**) DNA cleavage time courses using WT and split-Cas9 with either a full-length sgRNA (top) or the Δhairpins1–2 sgRNA (bottom). Values were averaged from three independent experiments, and error bars represent the standard deviation. Rate constants can be found in Table I.1.

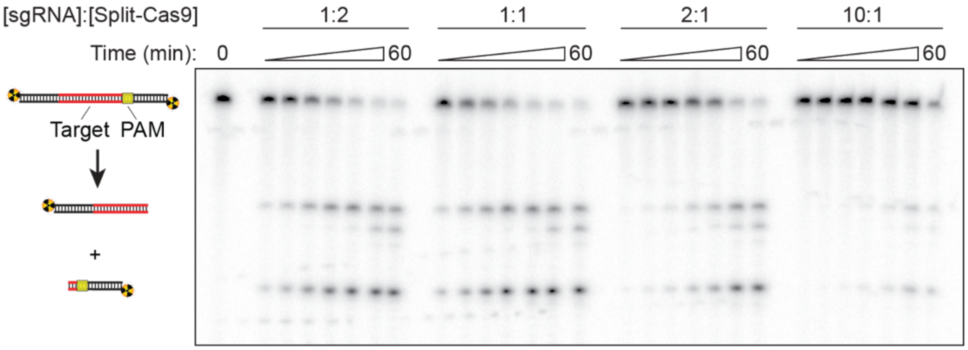


Figure I.8 | Excess sgRNA reduces the DNA cleavage activity of split-Cas9. DNA cleavage assay with varying molar ratios of protein to sgRNA, analyzed by denaturing PAGE. Reactions contained ~1 nM radiolabeled dsDNA, 100 nM α-helical and nuclease lobes, and 50–1000 nM sgRNA. The extent of product formation decreases substantially as the sgRNA concentration surpasses the lobe concentration. This observation suggests that stoichiometric excesses of sgRNA titrate the individual lobes away from each and onto independent sgRNA molecules, a hypothesis supported by the finding that distinct sgRNA motifs interact with either lobe.

I.4.4 Engineered sgRNAs selectively preclude split-Cas9 but not WT Cas9 activity

The potential for enhanced spatiotemporal control of genome engineering events with split-Cas9 prompted us to investigate ways in which sgRNA-mediated dimerization of the α-helical and nuclease lobes could be perturbed. In particular, we reasoned that certain 3'-truncated or modified sgRNAs, which have weak affinity for the nuclease lobe (Fig. I.5e) but still support robust DNA cleavage activity of WT Cas9 (Jinek et al., 2012), would selectively inactivate split-Cas9 activity through their inability to effectively recruit and dimerize both lobes into a functional enzyme complex. Thus, the activity of split-Cas9 in cells could be made dependent upon inducible protein-protein dimerization domains (Fig. I.9).

When we tested sgRNA variants that lacked one or both hairpins at the 3' end for their ability to support *in vitro* cleavage, split-Cas9 activity was either severely compromised or completely abolished whereas WT Cas9 activity was slightly reduced relative to a full-length sgRNA (Fig. I.7c, I.10, Table I.1). A recent report found that sgRNAs in which only the first hairpin is deleted function robustly in cells (Briner et al., 2014), and we found that similar designs supported DNA cleavage activity of WT Cas9 but not split-Cas9 *in vitro* (Fig. I.10). Thus, rationally designed variants of the sgRNA scaffold can be used to prevent RNA-mediated heterodimerization of the two lobes without compromising the intrinsic RNA-guided DNA cleaving capabilities of Cas9.

I.5 Discussion

Here we have successfully designed a split version of Cas9 that maintains the cleavage activity of the native enzyme. We demonstrated that the sgRNA is necessary and sufficient to dimerize the nuclease and α-helical lobes into an active complex, and furthermore showed that multiple distinct sgRNA motifs interact with the lobes independently. Split-Cas9 is active for genome editing in cells, albeit at a reduced level relative to WT Cas9, and can be rendered nonfunctional through the removal of one or more of the hairpins at the 3’ end of sgRNA. Although optimization will help split-Cas9 to function effectively in cells, we have shown the potential to enable a variety of interesting and useful applications.

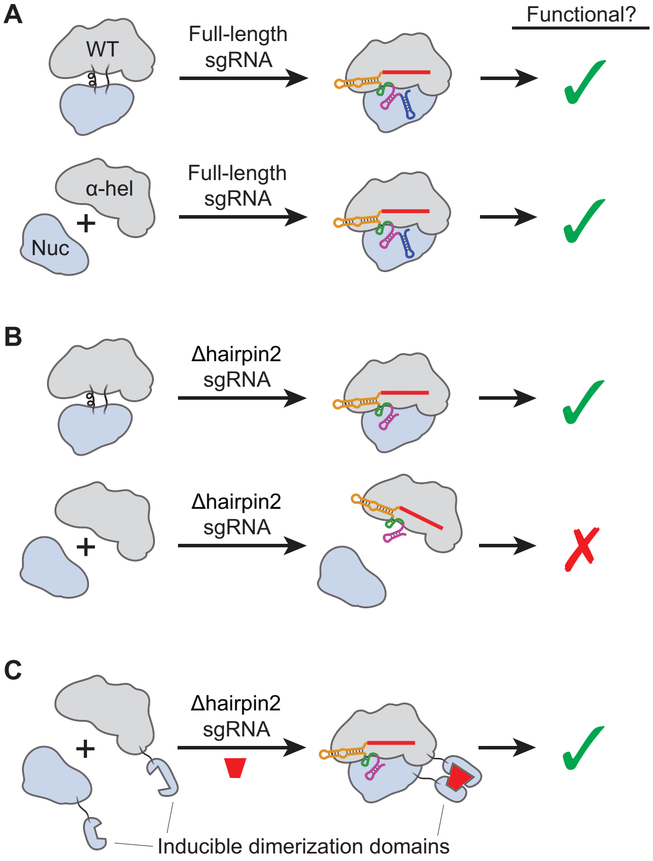


Figure I.9 | Strategy for inducible control of genome engineering by a split-Cas9 enzyme complex. (**a**) Because the α-helical and nuclease lobes dimerize in the presence of sgRNA, both WT and split-Cas9 are functional genome editing tools in cells using full-length sgRNA. (**b**) sgRNA variants with 3’-hairpin truncations have substantially weaker affinity for the nuclease lobe and thus do not efficiently assemble a functional split-Cas9 complex, leading to an inactive enzyme. In contrast, in vitro DNA cleavage by WT Cas9 is minimally affected by these truncations, indicating that the intrinsic activity of the Cas9–sgRNA enzyme complex does not require hairpins at the 3’ end. (**c**) We propose an inducible split-Cas9 system, in which exogenous dimerization domains control the assembly of a functional ternary complex between a 3’-truncated sgRNA and the α-helical and nuclease lobes. By fusing both lobes to domains that dimerize only upon some external stimulus (e.g. a small molecule; red trapezoid), split-Cas9 can be specifically activated for a desired genome engineering outcome.

Split-protein systems have often been designed such that the functional unit is restored through the interactions of an exogenous pair of proteins (Shekhawat and Ghosh, 2011). For example, DNA-binding and effector domains of modified TALEs have been fused to CRY2 and CIB1 to enable light-inducible control of gene expression (Konermann et al., 2013). Alternatively, genes may be split such that a single functional polypeptide is the final product of mRNA trans-splicing or intein-based protein splicing (Lienert et al., 2013). Our strategy with split-Cas9 differs in that it faithfully recapitulates the functionality of full-length Cas9 using the very same RNA ligand that WT Cas9 requires. In this sense, sgRNA-mediated dimerization and activation of split-Cas9 may be viewed analogously to the sgRNA-mediated rearrangement and structural activation of WT Cas9.

Optimizing expression levels of the components of the split-Cas9 system could increase its effectiveness. In particular, while the sgRNA should be kept limiting to avoid titrating the lobes away from the ternary complex, overall expression must be high enough to overcome the reduced affinity of both protein components for the sgRNA scaffold. Split-Cas9 could be regulated by the combinatorial use of promoters, restricting activity to highly specific subsets of tissues or creating a “coincidence detector” with two inducible promoters. Split-Cas9 could also be developed for use with adeno-associated viral vectors, where the smaller coding regions of each lobe would enable the use of effector or reporter domains that are currently prohibited by limited packaging capacity. We also suggest that split-Cas9 can be converted into a regulatable system using exogenous dimerization domains (Fig. I.9). Fusing both lobes to domains that selectively dimerize upon chemical or optical induction, such as the abscisic acid-inducible PYL-ABI dimer (Liang et al., 2011) or the blue light-inducible CRY2-CIB1 dimer (Kennedy et al., 2010) would allow for enhanced spatiotemporal control of genome engineering events (Hsu et al., 2014). Dimerization domains may also increase the efficiency of complex formation by making lobe assembly independent of the sgRNA. We propose that the combined use of inducible dimerization domains with compromised sgRNA variants that enable DNA targeting but not split-Cas9 assembly would eliminate leaky activity in the absence of inducer while still allowing for robust activation, creating an extremely sensitive inducible system.

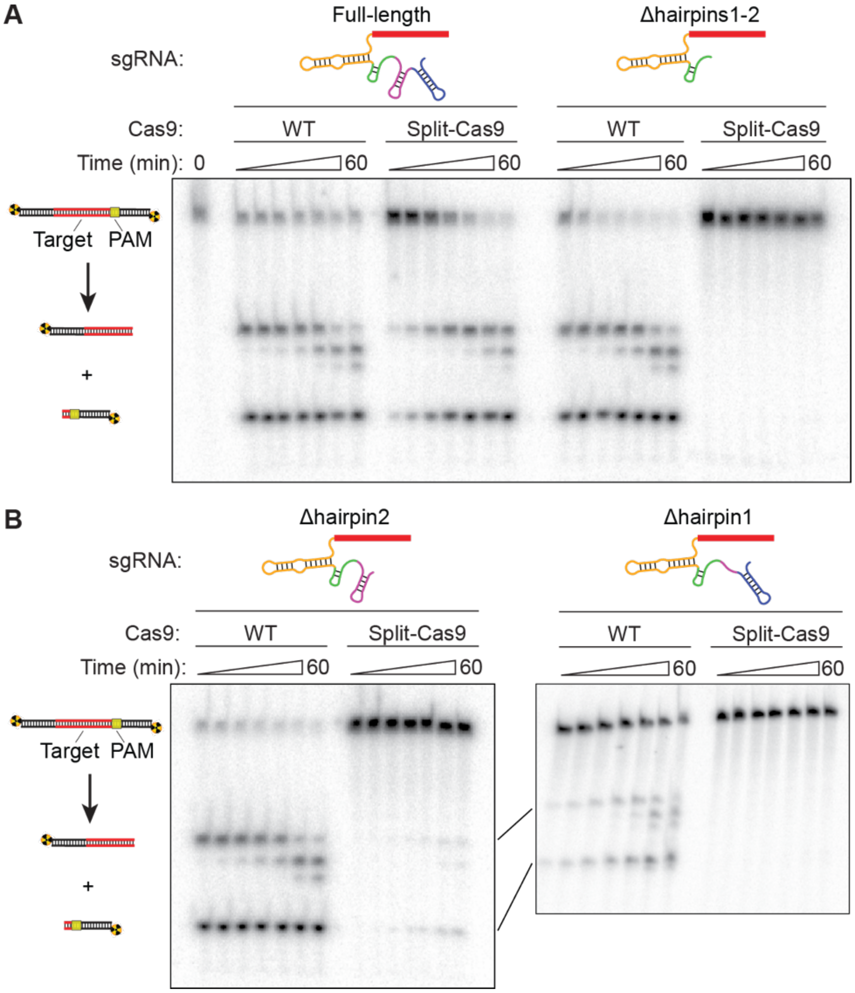


Figure I.10 | 3'-truncated sgRNA variants selectively inactivate split-Cas9. (**a-b**) DNA cleavage assays with WT and split-Cas9 and a panel of four different sgRNAs, analyzed by denaturing PAGE. (**a**) Full-length sgRNAs promote DNA cleavage activity of both WT and split-Cas9, whereas split-Cas9 activity is completely lost with an sgRNA lacking both hairpins at the 3' end (Δhairpins1–2). (**b**) sgRNA variants where only one hairpin is removed show minimal effects on WT Cas9 activity but severely (Δhairpin2) or completely (Δhairpin1) inactivate split-Cas9.

Finally, our study provides important insights into the structure-function relationship of the native Cas9 enzyme. The ability of sgRNA to act as a molecular scaffold in assembling two separate polypeptides highlights the crucial role that the sgRNA plays in orchestrating conformational rearrangements of WT Cas9. The separation of recognition and catalytic functions into two separate lobes may have evolved to eliminate non-specific nuclease activity and control licensing of Cas9 for DNA interrogation. Our results also invite comparisons to the mechanisms of RNP assembly and DNA targeting in other CRISPR-Cas systems, particularly the type I-E Cascade interference complex. While Cascade is composed of 11 distinct subunits, none of which possess nuclease activity, the guide RNA (crRNA) plays a similarly critical role in scaffolding the assembly of distinct domains into a structure that is primed to engage DNA targets (Jackson et al., 2014; Mulepati et al., 2014; Wiedenheft et al., 2011a). Similar principles are likely to govern the assembly and activity of other CRISPR RNA-guided DNA targeting complexes (van der Oost et al., 2014). Thus, distinct CRISPR–Cas systems may have evolved similar organizational strategies in parallel that utilize the guide RNA for structural assembly and conformational activation.

I.6 Materials and methods

I.6.1 Cloning and protein purification

The expression vector for purification of the nuclease lobe was generated by Around the Horn (ATH) polymerase chain reaction (PCR) using a pre-existing pET-based expression vector for *S. pyogenes* Cas9. The final construct encodes an N-terminal decahistidine-maltose binding protein (His_{10}-MBP) tag, a tobacco etch virus (TEV) protease cleavage site, residues 1–57, a glycine-serine-serine linker, and residues 729–1368. The vector for the catalytically inactive dNuclease lobe was generated by ATH PCR of a similar dCas9 (D10A/H840A) vector. The vector for expression of the α-helical lobe was generated by PCR amplification of *S. pyogenes* Cas9 residues 56-714 and assembly of the resulting fragment into a His_{10}-MBP expression vector via ligation-independent cloning.

Each protein was over-expressed in *E. coli* BL21 Rosetta 2(DE3) (EMD Biosciences) by growing in 2xYT medium at 37 °C to an optical density of 0.5, inducing with 0.5 mM IPTG, and growing an additional 16 hours at 18°C. Cells were lysed by sonication in a buffer containing 50 mM Tris pH 7.5, 500 mM NaCl, 1 mM TCEP, 5% glycerol, and a protease inhibitor cocktail (Roche). The clarified lysate was bound in batch to Ni-NTA agarose (Qiagen). The resin was washed extensively with 20 mM Tris pH 7.5, 500 mM NaCl, 1 mM TCEP, 10 mM imidazole, 5% glycerol; and the bound protein was eluted in 20 mM Tris pH 7.5, 500 mM NaCl, 1 mM TCEP, 300 mM imidazole, 5% glycerol. The His_{10}-MBP affinity tag was removed with His_6-tagged TEV protease during overnight dialysis against 20 mM Tris pH 7.5, 500 mM NaCl, 1 mM TCEP, 5% glycerol. The protein was then flowed over Ni-NTA agarose to remove TEV protease and the cleaved affinity tag.

The α-helical lobe was dialyzed for 2 h against 20 mM Tris pH 7.5, 125 mM KCl, 1 mM TCEP, 5% glycerol, and purified on a 5 mL HiTrap SP Sepharose column (GE Healthcare), with elution over a linear gradient from 125 mM – 1 M KCl. The α-helical lobe was further purified by size exclusion chromatography on a Superdex 200 16/60 column (GE Healthcare) in 20 mM Tris pH 7.5, 200 mM KCl, 1 mM TCEP, 5% glycerol. The nuclease lobe was purified via size exclusion chromatography immediately following the ortho-Ni-NTA step. The dNuclease lobe was purified as described for the nuclease lobe, except the size exclusion chromatography was performed with 20 mM Tris pH 7.5, 500 mM KCl, 1 mM TCEP, 5% glycerol. Full-length Cas9 was purified as previously described (Jinek et al., 2014).

I.6.2 *In vitro* transcription of sgRNA

Linearized plasmid DNA was used as a template for *in vitro* transcription of full-length, Δhairpins1-2, and Δhairpin2 λ1 sgRNA. The appropriate region of the sgRNA, along with a T7 RNA polymerase promoter sequence, was PCR-amplified and restriction-cloned into EcoRI/BamHI sites of a pUC19 vector, and the resulting vector was digested with BamHI to enable run-off transcription. The Δhairpin1 and Δspacer-nexus λ1 sgRNA, as well as λ1 crRNA and tracrRNA, were transcribed from a single-stranded DNA template with an annealed T7 promoter oligonucleotide. The DNA template for EMX1 sgRNA was produced by overlapping PCR as previously described (Lin et al., 2014).

Transcription reactions (1 mL) were conducted in buffer containing 50 mM Tris, pH 8.1, 25 mM $MgCl_2$, 0.01% Triton X-100, 2 mM spermidine, and 10 mM DTT, along with 5 mM each of ATP, GTP, CTP, and UTP, 100 μg/mL T7 polymerase, and approximately 1 μM DNA template. Reactions were incubated at 37 °C overnight and subsequently treated with 5 units DNase (Promega) for 1 hour. Reactions were then quenched with 800 μL of 95% formamide, 0.05% bromophenol blue, and 20 mM EDTA, and loaded onto a 7M urea 10% polyacrylamide gel. The appropriate band was excised, and the RNA was eluted from the gel overnight in DEPC-treated water. The sgRNA was ethanol-precipitated and resuspended in DEPC-treated water. Concentrations were determined by A_{260nm} using a NanoDrop (Thermo Scientific). For filter binding assays, the sgRNA was dephosphorylated with calf intestinal phosphatase (New England Biolabs) prior to radiolabeling with T4 polynucleotide kinase (New England Biolabs) and γ-[^{32}P]-ATP (Perkin Elmer). The radiolabeled sgRNA was gel purified as described above. Sequences of transcribed RNAs as well as all other RNA and DNA substrates used in the study are in Table I.3

I.6.3 Split-Cas9 complex reconstitution

Split-Cas9 complexes were reconstituted prior to cleavage and binding assays at 37 °C for 10 minutes in a buffer containing 20 mM Tris pH 7.5, 100 mM KCl, 5 mM $MgCl_2$, 1 mM DTT, and 5% glycerol. For binding assays, reactions containing equimolar amounts of the dNuclease lobe, the α-helical lobe, and *in vitro* transcribed sgRNA. For cleavage assays, reactions contained equimolar amounts of the nuclease lobe and sgRNA (or crRNA:tracrRNA),with a two-fold molar excess of the α-helical lobe.

I.6.4 DNA cleavage assays

All cleavage assays were performed in 1X Cleavage Buffer, which contained 20 mM Tris pH 7.5, 100 mM KCl, 5 mM $MgCl_2$, 1 mM DTT, and 5% glycerol. Preformed complexes were diluted in Cleavage Buffer, and reactions were initiated with the addition of radiolabeled dsDNA substrates. Final reaction concentrations were 100 nM protein:RNA complex and ~1 nM radiolabeled DNA target. The concentration of Cas9 was chosen to be sufficiently above the K_d for the sgRNA such that complex assembly is unlikely to be rate-limiting, except in the case of split-Cas9 and the Δhairpins1-2 sgRNA (K_d >100 nM). Reactions proceeded at room temperature, and aliquots were removed at selected time points and quenched with an equal volume of buffer containing 50 mM EDTA, 0.02% bromophenol blue, and 90% formamide. Reaction products were resolved by 7M urea-PAGE, gels were dried, and DNA was visualized by phosphorimaging and quantified using ImageQuant software (GE Healthcare). The percentage of DNA cleaved was determined by dividing the amount of cleaved DNA by the sum of uncleaved and cleaved DNA. Kinetic analysis was performed using Prism (GraphPad Software). Reported observed rate constants (k_{obs}) are the average of three independent experiments ± standard error of the mean. Graphed values are the averaged timepoints of three independent experiments with error bars representing the standard deviation.

Description	Sequence	Used in
λ1 target dsDNA	5′–AGCAGAAATCTCTGCTGACGCATAAAGATGAGACGC**TGG**AGTACAAACGTCAGCT–3′ 3′–TCGTCTTTAGAGACGACTGCGTATTTCTACTCTGCGACCTCATGTTTGCAGTCGA–5′	I.1d, I.3a,b, I.4, I.7c, I.8, I.9
λ1 target dsDNA, mutated PAM	5′–AGCAGAAATCTCTGCTGACGCATAAAGATGAGACGCTCGAGTACAAACGTCAGCT–3′ 3′–TCGTCTTTAGAGACGACTGCGTATTTCTACTCTGCGAGCTCATGTTTGCAGTCGA–5′	I.3c
λ2 target dsDNA, (λ1 mismatch)	5′–GAGTGGAAGGATGCCAGTGATAAGTGGAATGCCATG**TGG**GCTGTCAAAATTGAGC–3′ 3′–CTCACCTTCCTACGGTCACTATTCACCTTACGGTACACCCGACAGTTTTAACTCG–5′	I.3c
λ1 sgRNA, full-length	5′–GACGCAUAAAGAUGAGACGCGUUUUAGAGCUAUGCUGUUUUGGAAACAAAACAGCAUAG CAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUUU UUGGAUC–3′	I.1d, I.3a,c, I.5c,e, I.6d, I.7c, I.8, I.10a
λ1 crRNA	5′–GACGCAUAAAGAUGAGACGCGUUUUAGAGCUAUGCUGUUUUG–3′	I.3b, I.5e
tracrRNA	5′–GGACAGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCG AGUCGGUGCUUUUU–3′	I.3b, I.5e
λ1 sgRNA, Δhairpins1–2	5′–GACGCAUAAAGAUGAGACGCGUUUUAGAGCUAUGCUGUUUUGGAAACAAAACAGCAUAG CAAGUUAAAAUAAGGCUAGUCCGUGGAUC–3′	I.5d, I.7c, I.10a
λ1 sgRNA, Δhairpin2	5′–GACGCAUAAAGAUGAGACGCGUUUUAGAGCUAUGCUGUUUUGGAAACAAAACAGCAUAG CAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGAUC–3′	I.10b
λ1 sgRNA, Δhairpin1	5′–GACGCAUAAAGAUGAGACGCGUUUUAGAGCUAUGCUGUUUUGGAAACAAAACAGCAUAG CAAGUUAAAAUAAGGCUAGUCCGUUGGCACCGAGUCGGUGCUUUUUUU–3′	I.10b
λ1 sgRNA, Δspacer–nexus	5′–GGUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUUUUU–3′	I.5d
***EMX1* sgRNA, full-length**	5′–GGUCACCUCCAAUGACUAGGGGUUUAAGAGCUAUGCUGGAAACAGCAUAGCAAGUUUAA AUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUUUU–3′	I.7b

Table I.3 | DNA and RNA substrates used in this study. RNA guide sequences and complementary DNA target strand sequences are shown in red; PAM sites are highlighted in yellow.

I.6.5 Electrophoretic mobility shift assays (EMSAs)

All binding assays were performed in 1X Binding Buffer, which contains 20 mM Tris pH 7.5, 100 mM KCl, 5 mM $MgCl_2$, 1 mM DTT, 5% glycerol, 50 μg/mL heparin, 0.01% Tween-20, 100 μg/mL BSA. Preformed complexes were diluted into 1X Binding Buffer, after which radiolabeled dsDNA substrates were added to a final concentration of <0.2

nM. Reactions were incubated at room temperature for 60 min and then resolved at 4 ºC on a native 8% polyacrylamide gel containing 0.5X TBE and 5 mM $MgCl_2$. Gels were dried and DNA was visualized by phosphorimaging and quantified using ImageQuant software (GE Healthcare). The fraction of DNA bound (amount of bound DNA divided by the sum of free and bound DNA) was plotted versus concentration of protein and fit to a binding isotherm using Prism (GraphPad Software).

I.6.6 Filter-binding assays

All filter-binding assays were performed in 1X RNA Binding Buffer, which contains 20 mM Tris pH7.5, 100 mM KCl, 5 mM $MgCl_2$, 1 mM DTT, 5% glycerol, 0.01% Igepal CA-630, 10 μg/mL yeast tRNA, and 10 μg/mL BSA. WT Cas9, α-helical lobe, nuclease lobe, or an equimolar mix of both lobes (split-Cas9) were incubated with <0.02 nM radiolabeled sgRNA for 60 min at room temperature. Tufryn (Pall Corporation), Protran (Whatman), and Hybond-N+ (GE Healthcare) membranes were soaked in buffer containing 20 mM Tris pH7.5, 100 mM KCl, 5 mM $MgCl_2$, 1 mM DTT, 5% glycerol, and arranged on a dot blot apparatus. Binding reactions were separated through the membranes by the application of vacuum, and after drying, the membranes were visualized by phosphorimaging and quantified using ImageQuant Software (GE Healthcare). The fraction of sgRNA bound was plotted versus the concentration of protein and fit to a binding isotherm using Prism (GraphPad Software). Reported K_d values are the average of three independent experiments, and errors represent the standard error of the mean.

I.6.7 Negative-stain electron microscopy and image processing

We prepared and imaged negatively-stained samples of the α-helical lobe, nuclease lobe, and split-Cas9 (α-helical lobe and nuclease lobe) with and without sgRNA as described previously (Jinek et al., 2012). Data were acquired using a Tecnai F20 Twin transmission electron microscope operated at 120 keV at a nominal magnification of either 80,000× (1.45 Å at the specimen level) using low-dose exposures (~20 $e^{-}Å^{-2}$) with a randomly set defocus ranging from −0.7 to −1.6 μm. A total of 150–200 images of each Cas9 sample was automatically recorded on a Gatan 4k × 4k CCD camera using the MSI-Raster application within LEGINON (Suloway et al., 2005). Low-resolution negative stain class averages of Lid particles from the yeast 26S proteasome (Lander et al., 2012) were used as references for template-based particle picking. The Lid complex was used as a template to avoid selection bias because it bears minimal to no structural resemblance to Cas9. Cas9 complexes were extracted using a 224 × 224-pixel box size. These particles were subjected to 2D reference-free alignment and classification using multivariate statistical analysis and multi-reference alignment in IMAGIC (van Heel et al., 1996).

I.6.8 Cas9 and split-Cas9 RNP assembly and nucleofection

The split-Cas9 RNP was prepared immediately before the experiment by incubating both lobes with sgRNA at molar ratios of 1.2:1:1.2 (α-helical lobe:nuclease lobe:sgRNA) for 10 min at 37 °C in 20 mM HEPES pH 7.5, 150 mM KCl, 1 mM $MgCl_2$, 10% glycerol, and 1 mM TCEP. The nucleofections were carried out as previously described for Cas9, using 10, 30, and 100 pmol of RNP complex for approximately 2x10^{5} cells (Lin et al., 2014). Where indicated, cells were synchronized with 200 ng/mL nocodazole for 17 hr prior to nucleofection. Neither WT Cas9 nor the split-Cas9 lobes had nuclear localization signals, which may have led to reduced editing levels, particularly for the unsynchronized cells.

I.6.9 Analysis of in-cell genome editing efficiency

Determination of the percentage of indels induced at the target region was performed as previously described (Lin et al., 2014). In brief, 640-nt regions of the EMX1 locus containing the target sites were PCR amplified, and the resulting products were denatured, re-annealed, and digested with T7 endonuclease I (New England Biolabs), which cleaves mismatched heteroduplex DNA (Kim et al., 2009). The products were resolved on a 2% agarose gel containing SYBR Gold (Life Technologies), and band intensities were determined using Image Lab (Bio-Rad Laboratories). Editing efficiencies was determined using the formula (1 – (1 – (b + c / a + b +c))1/2) x 100, , where “a” is the band intensity of DNA substrate and “b” and “c” are the cleavage products (Ran et al., 2013).

I.7 Acknowledgments

We thank B. Castellano and M. Kaplan for assisting with cloning and protein purification; E. Nogales for use of EM resources and expert assistance in EM analysis; and members of the Doudna lab for helpful discussions. S.H.S. is a National Science Foundation Graduate Research Fellow and a National Defense Science & Engineering Graduate Research Fellow. A.V.W. is a National Science Foundation Graduate Research Fellow. B.T.S. is a Roche Postdoctoral Fellow, RPF 311. J.A.D is a Howard Hughes Medical Institute Investigator.