

MOLECULAR MECHANISMS OF DNA UPTAKE INTO CRISPR-CAS LOCI

by

JULIE MARIE GRAINY

(Under the Direction of Michael P. Terns)

ABSTRACT

An essential primary step of the prokaryotic CRISPR-Cas defense pathway includes acquiring foreign spacer DNA into the CRISPR locus. This first step, termed adaptation, provides the necessary immunity against the foreign mobile genetic element from which the DNA was captured. In this dissertation, a number of approaches were used to investigate the adaptation stage of CRISPR-Cas prokaryotic defense systems. The mechanisms of spacer integration into the CRISPR locus was investigated *in vitro* using two different model organisms, *Pyrococcus furiosus* and *Streptococcus thermophilus*. For *P. furiosus*, it was determined that Cas1 and Cas2 are sufficient to accurately integrate spacers into a minimal CRISPR locus, while Cas1 from *S. thermophilus* seems to be capable of independent spacer integration. The sequences and their relative spacing in the *P. furiosus* CRISPR repeat were found to contribute to accurate spacer integration at the sites required for maintaining the proper repeat size as the CRISPR locus grows. The precursor step of spacer generation was also considered here, to investigate how foreign DNA is selected and processed to the appropriate size for the integration of functional spacers. A complex consisting of Cas1, Cas2, Cas4-1, and Cas4-2 from *P. furiosus* was observed and demonstrated to be active in spacer integration into the CRISPR repeat *in*

vitro. Furthermore, Cas4-2 exhibited *in vitro* nuclease activity, supporting previous *in vivo* evidence of Cas4-2 involvement in pre-spacer generation. Together these *in vitro* studies of Cas protein molecular mechanisms add to the constantly growing model of CRISPR-Cas adaptation.

INDEX WORDS: CRISPR; Cas; Cas1; Cas2; Cas4; adaptation; *Pyrococcus furiosus*;
Streptococcus thermophilus

MOLECULAR MECHANISMS OF DNA UPTAKE INTO CRISPR-CAS LOCI

by

JULIE MARIE GRAINY

B.S., Waynesburg University, 2013

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial
Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2019

© 2019

Julie Marie Grainy

All Rights Reserved

MOLECULAR MECHANISMS OF DNA UPTAKE INTO CRISPR-CAS LOCI

by

JULIE MARIE GRAINY

Major Professor:
Committee:

Michael Terns
Anna Karls
Robert Maier
Claiborne Glover

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2019

DEDICATION

I dedicate this dissertation to my husband, Brandon Grainy, and daughter, Penelope Grainy. I cannot imagine a world without you. This work is the result of your loyal support at every step of the process. Thank you for the endless love and encouragement to pursue my dreams.

ACKNOWLEDGEMENTS

None of this work would have been possible without help and support from a number of people. First and foremost, I would like to thank Dr. Michael Terns for fostering my development as a research scientist. Your careful critiques and feedback throughout the process of experimental design all the way to publication have been my key to success. I would also like to thank my committee members, Dr. Anna Karls, Dr. Robert Maier, and Dr. Claiborne Glover for their helpful discussions and advice as I navigated my research projects over the past six years.

I am fortunate to have been surrounded by amazing labmates, past and present, including Dr. Rebecca Terns, Dr. Joshua Elmore, Dr. Sonali Majumdar, Dr. Morgan Teachey, Michael Ellis, Nolan Sheppard, Walter Woodside, Kawanda Foster, Ralph Zhang, Jenny Kim, Elizabeth Watts, Clare Edwards, and a constant flux of talented undergraduate students. Specifically, I need to acknowledge the undergraduate students who have worked directly with me, Jesse Hu, Alyssia Mitchell, Xavia Taylor, Ela Mitchel, and James Steward. Each of you aided the progression of my projects, and allowed me the opportunity to practice and improve my mentoring abilities. Thank you to all of the Terns lab members for making lab an enjoyable and stimulating environment. I value your direct and indirect involvement in my research, and more importantly your friendships.

Finally, I owe a substantial acknowledgement to my family and friends. Thank you for being an endless source of support in the good times and the bad. Thank you to

my husband, Brandon Grainy, for listening to countless periods of doubt and fear, and always reminding me that I am strong and capable of anything. Thank you for committing your life to me and for never limiting my dreams in any way. Thank you to Penelope Grainy for making me a mother, and for being a constant beam of joy and energy. Your presence during this past year has provided me with extra love, laughs, and motivation to be the best version of myself. I still can't believe that I have the privilege of being your mom. Thank you for letting me sleep (usually). Thank you to my family members, Jonathan Tischer, Carrie Tischer, Amy Guerrero, and Laura Brown. It was comforting to know that I could call each of you for encouragement, advice, or just to vent. Thank you for making it a priority to gather together as much as possible, even though we all live in different parts of the country now. Your love and support shines through, despite the distances between us. Thank you for listening to all of my science fair presentations, and fostering my passion for biology. To my hobbits, Angela Coole, Britta Rea, Jennelle Brigham, and Kelsey Mahonski, I acknowledge decades of valuable friendship. You have each, indubitably, been a steadfast source of entertainment, love, support, and weirdness. Thank you for keeping me young, while at the same time helping me navigate becoming an adult. Thank you for brutal honesty and unconditional friendship. Last, but not least, thank you to Pepper for being the best dog I could ever ask for. Your unspoken affection and emotional support has been one of the best parts of coming home each day.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	v
CHAPTER	
1 INTRODUCTION AND LITERATURE REVIEW	1
2 CRISPR REPEAT SEQUENCES AND RELATIVE SPACING SPECIFY DNA INTEGRATION BY <i>PYROCOCCUS FURIOSUS</i> CAS1 AND CAS2	47
3 CHARACTERIZING CAS4-1 AND CAS4-2 <i>IN VITRO</i> TO DETERMINE THEIR FUNCTION IN CRISPR-CAS ADAPTATION IN <i>PYROCOCCUS FURIOSUS</i>	127
4 DEVELOPING TOOLS TO STUDY ADAPTATION OF A RNA TARGETING TYPE III-A CRISPR-CAS SYSTEM <i>IN VITRO</i>	163
5 DISCUSSION	188

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Introduction to CRISPR-Cas systems

In the never-ending arms race between prokaryotes and invading phages, transposons, and other mobile genetic elements, various defense mechanisms have evolved to give an advantage to host cells that can overcome an attack from an invader. CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats)-Cas (CRISPR-associated proteins) systems allow prokaryotes to gain immunity towards viruses (1) and plasmids (2), and are found in about 50% of bacteria and 90% of archaea (3). CRISPR-Cas is a nucleic acid-guided defense mechanism grouped with other similar defense systems such as RNA interference and argonaute-centered defense (4-8). However, CRISPR is unique in that it is a heritable immune memory (9-12).

The key to the heritable nature of these systems are the CRISPR loci found on the host genome, which includes an array of DNA “spacers” that are 30-40 bp sequences of DNA derived from the invader, separated by 25-35 bp direct repeats (1). Preceding the repeat-spacer array is an AT-rich “leader” sequence and CRISPR-associated genes required for immunity, which are clustered into one or more operons (3,13,14). CRISPR arrays were first implicated as a genetic memory bank of previously encountered mobile genetic elements when the spacer sequences were found to match sequences found in

viral genomes and conjugative plasmids (15-18). Two years later, in 2007, spacers contained in CRISPR arrays were discovered to provide immunity to the host prokaryote upon infection with a phage containing a sequence matching the spacer (1). Additionally it was shown that infection with a novel phage resulted in a new spacer being added to the CRISPR array from the invading DNA (1). The decade to follow uncovered the foundational mechanisms of CRISPR-Cas immunity, and revealed diverse systems that utilize unique pathways to achieve the same core goal of capturing sequences of nucleic acid from the invader, and using that as a guide to target degradation of the invading nucleic acid if infection ever occurred again.

CRISPR-Cas immunity can be boiled down to three distinct stages (11,19): adaptation, CRISPR RNA (crRNA) biogenesis, and invader silencing (Figure 1.1). During adaptation, new spacers are acquired through mechanisms that capture fragments of the invading nucleic acid and insert it into the CRISPR array on the host genome. During crRNA biogenesis, the CRISPR array is transcribed and processed into crRNAs that will serve as guides to bind via complementary base pairing to the invader DNA or RNA (20-22). During invader silencing, effector Cas proteins are guided to the invader by the crRNA and proceed to degrade the nucleic acid in a targeted manner (23-26).

The precise targeted cleavage of DNA by CRISPR-Cas systems has been repurposed as a method of genome editing, most notably with the Cas9 nuclease. The success of CRISPR derived genome editing tools has been one of the driving factors to investigate the basic biology of CRISPR systems (27-36), and as more and more CRISPR-Cas systems are discovered and characterized we continue to add to our

growing knowledge of the arms race between prokaryotes and mobile genetic elements and new biochemical tools for research, healthcare, and industry.

Classification of CRISPR-Cas systems

As CRISPR-Cas systems with unique features began to be discovered, a classification scheme was established to group the systems into defined categories. The diversity of these systems derives from the variety of *cas* genes found flanking CRISPR arrays. The Cas proteins coordinate all three stages of CRISPR immunity, and the wide range of systems have been classified by signature proteins (19,32). While all systems use RNA-guided ribonucleoprotein complexes to recognize and degrade foreign genetic elements, the arms race between host and mobile genetic elements has resulted in rapid evolution of *cas* genes, particularly the nuclease component of the effector module (37). This accounts for diversification of molecular mechanisms of CRISPR immunity (3,32,38).

First, all systems fall into two classes. Class 1 includes multi-subunit effector complexes (38), while Class 2 includes systems with a single, large, multidomain Cas protein functioning independently in invader silencing (38) (such as the extensively studied and utilized Cas9). The classes are further broken down into three Types for each class. Class 1 includes Type I, III, and IV, while Class 2 includes Type II, V, and VI (39). Each Type contains distinct effector Cas nucleases (40). At their narrowest classification of subtype, CRISPR-Cas systems are further distinguished by subtler differences in the organization of CRISPR loci and usually have subtype-specific Cas proteins (41). Twenty-four of the major subtypes are represented in Figure 1.2, but I will only be elaborating on the subtypes relevant for the studies included in this dissertation.

Each Type has a unique signature protein that is responsible for target cleavage. Type I systems each contain a signature Cas3 nuclease/helicase recruited by the multi-protein effector complex to cleave the target DNA. Type III systems each contain a Cas10 nuclease as part of the multi-protein effector complex to cleave target DNA in a transcription-dependent manner (42). Examples of both a multi-subunit effector complex and a single effector protein are represented in Figure 1.2. The detailed mechanisms of CRISPR-Cas adaptation, crRNA biogenesis, and invader silencing will be discussed below, with an emphasis on Type I and Type III systems.

Mechanisms of CRISPR-Cas adaptation

The investigation of the adaptation stage of CRISPR-Cas immunity was pioneered by studies of the Type I-E system in *Escherichia coli* and the II-A systems in *Streptococcus thermophilus* and *Streptococcus pyogenes* (1,43-47). Adaptation was first demonstrated in *S. thermophilus* through experiments that showed acquisition of immunity. New spacers were added at the leader proximal repeat junction— leading to a historical order of previous infections (1). After that key study, adaptation was examined in more detail with the I-E system of *E. coli* to determine requirements for *cis*- and *trans*-acting factors (47,48). Cas1 and Cas2 were determined to be required for adaptation *in vivo* (47-49). Since that time many studies using *in vivo*, *in vitro*, and structural approaches to understand the mechanism of adaptation have all contributed to a growing model of adaptation represented in Figure 1.3A and discussed in further detail below.

The adaptation process starts with Cas proteins binding to “protospacer” DNA (the sequence that will eventually become a spacer in the CRISPR array) and introducing

a double-strand break or exonucleolytic DNA trimming at sites harboring protospacer adjacent motifs (PAMs). The excised protospacer is integrated into the CRISPR array, almost invariably at the first repeat, next to the leader sequence (50). Adaptation is often broken down into two stages, pre-spacer generation and spacer integration, which will be discussed in further detail below. It is also important to note that there is mechanistic diversity between systems, and key differences will be mentioned throughout.

Pre-spacer generation

Distinguishing self vs. non-self is essential for avoiding acquiring spacers from the host genome, which results in auto-immunity. However, acquiring spacers from self is actually very frequent (47,51,52). Some organisms have ways to skew spacer acquisition towards mobile genetic elements. A study of the Type I-E system in *E. coli* showed that foreign spacers were preferred over self-derived spacers by about 100- to 1,000-fold and that adaptation was dependent on DNA replication (51). Stalled DNA replication, commonly seen at locations such as termination sites, are prone to dsDNA breaks, and thus dsDNA breaks could be a major source for protospacers. The study found that self-spacers were acquired in a pattern that dropped off at Chi sequences (5'-GCTGGTGG), implicating the dsDNA break repair activity of RecBCD as a mechanism contributing to spacer selection. In organisms that contain RecBCD, dsDNA break repair machinery relies on Chi sites to prevent the Rec nucleases from continuing to chew back the DNA (53). However, foreign DNA contains very few Chi sites, which would result in non-stalled nuclease activity by RecBCD, providing a large pool of potential protospacers for adaptation. Similarly AddAB, a homologous dsDNA break repair system, was

required for robust spacer integration in *S. pyogenes* Type II system (54). Moreover, a detailed analysis of spacers acquired from plasmid or genome sources in *Pyrococcus furiosus* revealed preferential uptake from regions prone to DNA nicking and/or double stranded breaks, including active transposons, ribosomal RNA genes, origins of replication, and sites of recombination (55). While the sources of preferred protospacers are becoming evident, there does not appear to be a defined mechanisms for avoiding acquiring spacers from host DNA. From a population perspective, as long as some (or one) of the prokaryotic cells survive the infection, the CRISPR-Cas immunity can be inherited vertically.

Regardless of self or foreign origin, protospacer sources usually contain a PAM sequence. The PAM is a 2-7 nt element commonly found neighboring the sequence that will ultimately become a new spacer (56,57) and is recognized by proteins involved in adaptation. Diverse PAM sequences have been discovered in many system by mapping acquired protospacer to plasmid and viral genomes and examining the flanking sequences for conserved motifs (47-49,56,58). The PAM is important in order for the effector proteins to be targeted specifically to the invader and not target degradation at the spacer located in the host CRISPR array (56,59).

The diversity of adaptation mechanisms is illustrated by a range of modes for PAM recognition and pre-spacer processing. For Class 2 systems the single multi-domain effector protein responsible for invader silencing contains a PAM recognition domain to confirm it is targeting the foreign DNA rather than the spacer DNA at the site of the CRISPR array on the host's genome. Therefore it seemed logical that the PAM recognition domain of Cas9 (the effector protein of Type II systems) was shown to be

essential for acquiring new spacers with PAM specificity (43,46). Cas9 forms a complex with Cas1, Cas2, and Csn2 (another protein involved in Type II-A adaptation), suggesting cross-talk between adaptation and invader-silencing proteins. This practical use of one PAM recognizing protein for both adaptation and invader silencing has not held true in other organisms.

In Type I systems, the effector proteins (including the ones responsible for PAM recognition during invader silencing) are not required for acquiring spacers with PAM specificity. Type I systems often have Cas4 proteins, some of which have demonstrated PAM recognition and cleavage (60-66). In addition to PAM specificity, Cas4 proteins also maintain the proper spacer length and spacer orientation, two critical elements for functional invader silencing (63). Cas4 proteins are also found in II-B and V systems (67-71), but their involvement in pre-spacer generation has not been studied in those systems yet. In a system lacking Cas4 (Type I-E), Cas1 is demonstrated to recognize the PAM and process the pre-spacer into the form needed for successful integration into the CRISPR array (72). Structural evidence shows Cas1 recognizing the PAM through sequence-specific contacts with the unprocessed protospacer (72).

It appears that Type III systems may lack PAMs altogether. In order to prevent self-targeting during invader silencing, the effector complex instead uses the repeat-derived region of the crRNA to determine self (spacer at a CRISPR array) versus non-self (the target nucleic acid) (23,73-76). Type III systems are also unique from Type I and II because they can acquire immunity from RNA. There has been some exploration into how dsDNA spacers in the CRISPR array are produced from an RNA species. In some cases, a reverse transcriptase is present in the CRISPR-Cas locus or actually fused to

Cas1 (77,78), which allows the invader RNA to serve as the template to make the dsDNA spacer for integration into the CRISPR. However, it is still unclear how systems lacking reverse transcriptase are able to achieve this.

Ultimately, the pre-spacer generation stage of adaptation leads to a DNA product that is ready to be integrated into the CRISPR array. In order for the integrated spacer to be transcribed into a crRNA that is functional in the invader silencing stage of CRISPR-Cas immunity, the generated pre-spacer needs to be a discrete size and have originated adjacent to a PAM sequence (Figure 1.3B). The final orientation of the spacer in the CRISPR array will also be vital for functional targeting. A crRNA containing a correctly oriented spacer will bind to the target strand of the invader nucleic acid, positioning the PAM recognizing domain of the effector protein(s) at the PAM sequence of the target strand. Once the PAM is recognized, the effector protein(s) degrade the invader. In the case of an incorrectly oriented spacer, the strand that the crRNA is targeted to will not contain the PAM-sequence, and invader degradation will not occur. Therefore, it is essential that the subsequent integration step of adaptation occur in a specific manner.

Spacer integration

Cas1 and Cas2 are universally found in all systems with very few exceptions, which suggests a conserved mechanism across CRISPR systems for spacer integration (3,19). Cas1 is the enzymatically active subunit that performs the integration of spacers, while Cas2 plays a structural role (44,79). Spacer integration intermediates, detected by Southern blot analysis of genomic CRISPR locus, demonstrated products consistent with two transesterification reactions occurring at both borders of the repeat on opposite

strands (79). This implicated an integrase-like mechanism of inserting spacer DNA into the CRISPR in a manner that would duplicate the repeat sequence. This model was supported by *in vitro* reconstitution of integrating pre-spacers into a plasmid containing a CRISPR locus by Cas1 and Cas2 from *E. coli* (80). The pre-spacers required 3' hydroxyl ends just like DNA transposases and retroviral integrases (81,82). The 3' hydroxyl ends of each strand of the incoming spacer make two nucleophilic attacks, one at the first base of the repeat, on the top strand, and the other at the last base of the repeat, on the bottom strand (Figure 1.3C). As the two strands of the repeat dissociate, with the new spacer between them, the ssDNA repeats are filled in by host DNA polymerase ensuring the repeat is properly duplicated each time a new spacer is added (83).

In order to avoid spacers being integrated into random locations on the genome, Cas1 and Cas2 need to be directed to the CRISPR repeat. Sequencing *in vitro* spacer integration products revealed that Cas1 and Cas2 have a preference for CRISPR repeats (80). More specifically, Cas1 was shown to have sequence specificity at the leader-repeat junction for spacer disintegration activity (the reverse reaction of spacer integration) (84). The mid-repeat sequence in the Type I-E system seems to be essential for Cas1 and Cas2 to bind to the repeat (85), which is where Cas2 and the non-catalytic Cas1 are predicated to be situated during spacer integration. The sequence of the repeat has also been implicated in regulating the order of the two-step transesterification event of spacer integration. Some studies have demonstrated the sequence at the leader-proximal end of the repeat to be essential for integration, while the leader-distal end of the repeat is dispensable (50). This supports a model in which the first step of integration occurs at the

leader-repeat junction in a sequence-specific manner, followed by the second site attack at the repeat-spacer junction in a sequence-independent manner.

In vivo, new spacers are directed to the first repeat, but in Type I-E and I-F systems that have been studied *in vitro* Cas1 and Cas2 readily integrate spacers at every repeat in the CRISPR array (80,86). Both of these studied systems found that the addition of integration host factor (IHF) caused integration to be focused at the leader-adjacent repeat (87-89). In contrast, the Cas1 and Cas2 proteins in the Type II-A system seem to have an intrinsic ability to direct spacer integration to the first repeat (90). Further confirming this difference between systems is the observation that the Type II-A leader could be reduced to 10 bp *in vivo* (45), while the Type I-E system requires the full 60 bp leader (47).

Ultimately, the spacer integration step of adaptation leads to a new fragment of invader DNA added to the CRISPR array and the duplication of the repeat sequence. It is essential that the two transesterification reactions occur precisely at the first and last nucleotide of the repeat in order to properly duplicate the repeat. It is clear that both *cis*- and *trans*-acting factors are involved in regulating the important specificity of spacer integration to ensure that the spacer is integrated at the correct location and in the proper orientation.

Primed adaptation

The first time encountering a mobile genetic element and acquiring a spacer is termed naïve adaptation. This is a somewhat “dumb” and inefficient defense mechanism. But if the invading DNA is encountered again, the invader silencing step of CRISPR-Cas

immunity results in a robust uptake of more spacers surrounding the target site that was degraded by the effector proteins (48,49,68). This is called primed adaptation, as it is primed by a spacer already present on the CRISPR. In some instances, the invader has mutated since the previous encounter, and if there is a mismatch in the PAM or target sequence, the efficiency of primed adaptation actually increases (91). It is thought that the mismatch causes the effector proteins to slow down their nuclease activity, allowing time for the adaptation proteins to be recruited to the site to grab new spacers to add to the CRISPR array (92,93). While primed adaptation has obvious benefits to the host, it would not be possible without first undergoing basic naïve adaptation.

Mechanisms of invader silencing

While adaptation is the critical first stage in CRISPR-Cas immunity, it would serve no purpose without the subsequent steps that result in invader silencing. The mechanisms of crRNA biogenesis and invader silencing are elaborated on in this section, thus completing the full cycle of CRISPR-Cas defense.

CRISPR RNA biogenesis: Processing the CRISPR array into RNAs to guide the effector complex

The CRISPR array is typically transcribed into a single long pre-crRNA and processed into mature crRNA by the Cas6 nuclease or an external RNase (21,94,95). In Class 1 systems the effector complex binds the transcript and Cas6 is recruited to process the crRNA (20,96-99).

In Type II systems of Class 2 this processing is achieved by RNase III and trans-acting CRISPR (tracr)RNA (22,24,100-103). In other Class 2 systems, the single effector protein Cas12a (Type V) or Cas13 (Type VI) processes their own crRNAs (103-106). This important step of generating functional crRNA that will accurately guide the effector nucleases to the invading nucleic acid is reviewed in great detail here: (94,95).

Invader silencing: Interference by target binding and cleavage

The crRNA acts as a guide that directs the effector complex of Cas proteins to the protospacer sequence of the invading nucleic acid (107,108). The vast diversity of effector complexes across subtypes accounts for the mechanistic variability. For Class 2 systems the nuclease domain of the single large effector protein is responsible for cleavage (22,24,100,102,109). In Type I systems the multi-subunit effector complex binds to the protospacer and subsequently recruits the Cas3 helicase/nuclease protein to facilitate cleavage of the target. The helicase domain unwinds the dsDNA while the nuclease domain cleaves the DNA (110-112). In contrast, for Type III systems the Cas10 nuclease responsible for target cleavage is a stable subunit of the effector complex (113). Besides the nuclease being a separate protein (Cas3) or part of the complex (Cas10), the overall structure of the multi-subunit effector complexes of Type I and Type III systems are very similar (113). Despite this similarity, Type III systems are functionally very distinct from Type I in that DNA cleavage requires primary binding of the RNA transcript at that location (114,115). The mechanism of this mode of invader silencing involves the binding of the target RNA activating ssDNAse activity (42,116-119). Furthermore, Cas10 (the subunit responsible for ssDNAse activity at the target site),

synthesizes cyclic oligoadenylates from ATP to signal the activation of Csm6 (Type III-A) or Csx1 (Type III-B) (proteins not part of the effector complex) which proceed to cleave RNA in a nonspecific manner. The detailed mechanisms of these systems are summarized in great detail in these reviews: (108,120-122).

Type III systems typically coexist with Type I systems (19,123). In certain cases they share adaptation machinery, and in other cases they have their own set of adaptation proteins. Two prime examples of these two scenarios include the CRISPR-Cas systems found in *Pyrococcus furiosus* and *Streptococcus thermophilus*, which I elaborate on below.

CRISPR-Cas systems of *Pyrococcus furiosus*

The model archaeal organism, *P. furiosus*, was first isolated from marine thermal sediment and optimally grows at 100°C (124). This often renders thermostable recombinant proteins, allowing for detailed mechanistic and structural studies (125). The COM1 strain has seven active CRISPR loci, each with a common 30 bp repeat shared in all arrays (125). The genome contains three Class 1 multi-subunit effector complexes: Type I-A (a Csa system), I-G (a Cst system), and III-B (a Cmr system). The organization of these CRISPR-Cas systems, and their locations on the genome, are displayed in Figure 1.4. Evolutionarily acquiring multiple immune systems was likely a result of the arms race against viruses (39). In this particular organism, the systems work simultaneously (126), but can also each function independently if the other two systems are knocked out (127). Each CRISPR array is transcribed and processed into crRNA (125,128,129) by the metal-independent endoribonuclease Cas6 (21,125,130,131). Each mature crRNA

contains one spacer, that will act as a guide RNA for any of the three effector complexes of *P. furiosus*, and target the foreign DNA or RNA (only by the Type III-B system) (42,126,127,132-134). The Type I-A and I-G systems each have the signature Cas3 nuclease, while the Type III-B system has the signature Cas10 nuclease. The crRNA guiding the III-B effector complex will bind to the target RNA, which activates cleavage of the target DNA (42), as discussed above.

While effector module genes can be shuffled and passed around, the adaptation module remains fixed in most cases (135). This allows *P. furiosus* to have multiple immune systems that share one set of adaptation proteins, something also observed with other organisms (38,41,135). The mechanisms of *Pyrococcus furiosus* adaptation have begun to be elucidated by *in vivo* studies. Cas1, Cas2, and two similar but distinct Cas4 proteins have been shown to be essential for proper adaptation *in vivo* (63). One of the *cas4* genes (*cas4-1*) is located next to the *cas1* and *cas2* genes, while the other *cas4* gene (*cas4-2*) is an isolated gene not near any other Cas genes or CRISPR arrays (Figure 1.4). Despite not being part of the core adaptation locus, Cas4-2 still contributes significantly to the accuracy of spacer integration (63). Cas4-1 and Cas4-2 were demonstrated to play a role in accurate pre-spacer generation (63), while the Cas1 and Cas2 proteins are assumed to be directly involved in the spacer integration step of adaptation.

Pyrococcus furiosus is a useful model organism for studying multiple CRISPR-Cas immune systems that all share one core set of adaptation machinery that continually add new spacers to all seven CRISPR arrays.

CRISPR-Cas system of *Streptococcus thermophilus*

The model organism *Streptococcus thermophilus* is an easily cultured, non-pathogenic, gram-positive bacterium with attractive applications in the dairy industry. The JIM8232 strain contains a Type III-A system with 17 unique spacers in the associated CRISPR array (Figure 1.5) (136). The crRNA biogenesis protein (Cas6) and the effector proteins (Csm1-Csm5 and Csm6-1) from *S. thermophilus* JIM8232 can achieve invader silencing when expressed in *E. coli* (137). The Csm1 subunit is the Cas10 family protein responsible for the DNase activity (23), while Csm3 provides the RNase activity of the complex (23,74,75). The accessory protein, Csm6, is activated by cyclic oligoadenylate secondary messengers produced by Csm1 and results in non-specific RNase activity to accelerate the destruction of the invader (138). This is the only active CRISPR-Cas system in the JIM8232 strain, indicating that it should be capable of acquiring new spacers without borrowing the adaptation machinery from a separate system.

Very little is known about the adaptation stage of Type III systems. While Type I and II systems have reported the requirement for dsDNA pre-spacers, it will be interesting to see if this RNA targeting Type III-A systems will have a mechanism to incorporate ssRNA or ssDNA pre-spacers into the CRISPR array. Adaptation has been observed *in vivo* for a Type III-B system, which uses a reverse transcriptase fused to Cas1 to incorporate new spacers from RNA sources (77). Since the JIM8232 strain Type III-A system lacks a reverse transcriptase, it will be an important model system for studying if/how RNA can be used as a source of new spacer uptake in the absence of the reverse transcriptase.

S. thermophilus JIM8232 is a useful model organism for studying Type III-A CRISPR-Cas immunity, in that it has its own set of adaptation proteins that are predicted to be capable of spacer integration. While the mechanisms of adaptation have primarily been studied with Type I and Type II systems, this provides a unique opportunity to investigate a diverse system that has RNA targeting capabilities.

Overview of dissertation

The next three chapters describe the approaches I have used to investigate the adaptation stage of CRISPR-Cas prokaryotic defense systems. Chapter 2 narrows in on the spacer integration step of adaptation. The study analyzes the *in vitro* activity of recombinant Cas1 and Cas2 from *Pyrococcus furiosus* to identify sequences that direct DNA integration to the CRISPR repeat. Chapter 3 shifts the attention to the pre-spacer generation step of adaptation. The study identifies physical interactions between Cas1/Cas2/Cas4-1/Cas4-2 from *Pyrococcus furiosus*, and seeks to elucidate the individual functions of Cas4-1 and Cas4-2 *in vitro* in order to better understand the function of the adaptation machinery as a whole. I report *in vitro* nuclease activity by Cas4-2 and build the characterized function into the ever-growing model of adaptation. Chapter 4 diverges from studying the CRISPR-Cas systems of *Pyrococcus furiosus* to explore the RNA targeting Type III-A system of *Streptococcus thermophilus*. The report defines essential components for studying the adaptation stage of the system *in vitro*, and provides evidence for spacer integration by Cas1. In summary, these three chapters combined provide a glimpse into the fascinating molecular mechanisms of CRISPR-Cas adaptation.

References

1. 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.
2. Marraffini, L.A. and Sontheimer, E.J. (2008) CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science*, **322**, 1843-1845.
3. Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J., Barrangou, R., Brouns, S.J., Charpentier, E., Haft, D.H. *et al.* (2015) An updated evolutionary classification of CRISPR-Cas systems. *Nature reviews. Microbiology*, **13**, 722-736.
4. Hur, J.K., Olovnikov, I. and Aravin, A.A. (2014) Prokaryotic Argonautes defend genomes against invasive DNA. *Trends in biochemical sciences*, **39**, 257-259.
5. Hutvagner, G. and Simard, M.J. (2008) Argonaute proteins: key players in RNA silencing. *Nat Rev Mol Cell Biol*, **9**, 22-32.
6. Koonin, E.V. (2017) Evolution of RNA- and DNA-guided antiviral defense systems in prokaryotes and eukaryotes: common ancestry vs convergence. *Biology direct*, **12**, 5.
7. Shabalina, S.A. and Koonin, E.V. (2008) Origins and evolution of eukaryotic RNA interference. *Trends Ecol Evol*, **23**, 578-587.

8. Swarts, D.C., Makarova, K., Wang, Y., Nakanishi, K., Ketting, R.F., Koonin, E.V., Patel, D.J. and van der Oost, J. (2014) The evolutionary journey of Argonaute proteins. *Nature structural & molecular biology*, **21**, 743-753.
9. Makarova, K.S., Grishin, N.V., Shabalina, S.A., Wolf, Y.I. and Koonin, E.V. (2006) A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biology direct*, **1**, 7.
10. Marraffini, L.A. and Sontheimer, E.J. (2010) CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nature reviews. Genetics*, **11**, 181-190.
11. van der Oost, J., Jore, M.M., Westra, E.R., Lundgren, M. and Brouns, S.J. (2009) CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends in biochemical sciences*, **34**, 401-407.
12. Koonin, E.V. and Makarova, K.S. (2009) CRISPR-Cas: an adaptive immunity system in prokaryotes. *Fl000 Biol Rep*, **1**, 95.
13. Kunin, V., Sorek, R. and Hugenholtz, P. (2007) Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome biology*, **8**, R61.
14. Makarova, K.S., Wolf, Y.I. and Koonin, E.V. (2013) The basic building blocks and evolution of CRISPR-CAS systems. *Biochemical Society transactions*, **41**, 1392-1400.

15. Bolotin, A., Quinquis, B., Sorokin, A. and Ehrlich, S.D. (2005) Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology*, **151**, 2551-2561.
16. 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*, **151**, 653-663.
17. 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.
18. Mojica, F.J., Diez-Villasenor, C., Garcia-Martinez, J. and Soria, E. (2005) Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol*, **60**, 174-182.
19. Makarova, K.S., Aravind, L., Wolf, Y.I. and Koonin, E.V. (2011) Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biology direct*, **6**, 38.
20. Brouns, S.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J., Snijders, A.P., 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.
21. 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.

22. 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.
23. 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.
24. 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.
25. Jore, M.M., Lundgren, M., van Duijn, E., Bultema, J.B., Westra, E.R., Waghmare, S.P., Wiedenheft, B., Pul, U., Wurm, R., Wagner, R. *et al.* (2011) Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nature structural & molecular biology*, **18**, 529-536.
26. 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**, E2579-2586.
27. Barrangou, R. and Horvath, P. (2017) A decade of discovery: CRISPR functions and applications. *Nat Microbiol*, **2**, 17092.
28. Barrangou, R. and Marraffini, L.A. (2014) CRISPR-Cas systems: Prokaryotes upgrade to adaptive immunity. *Mol Cell*, **54**, 234-244.
29. Doudna, J.A. and Charpentier, E. (2014) Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science*, **346**, 1258096.

30. Hsu, P.D., Lander, E.S. and Zhang, F. (2014) Development and applications of CRISPR-Cas9 for genome engineering. *Cell*, **157**, 1262-1278.
31. Komor, A.C., Badran, A.H. and Liu, D.R. (2017) CRISPR-Based Technologies for the Manipulation of Eukaryotic Genomes. *Cell*, **168**, 20-36.
32. Koonin, E.V., Makarova, K.S. and Zhang, F. (2017) Diversity, classification and evolution of CRISPR-Cas systems. *Current opinion in microbiology*, **37**, 67-78.
33. 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.
34. Sorek, R., Lawrence, C.M. and Wiedenheft, B. (2013) CRISPR-mediated adaptive immune systems in bacteria and archaea. *Annual review of biochemistry*, **82**, 237-266.
35. Wright, A.V., Nunez, J.K. and Doudna, J.A. (2016) Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. *Cell*, **164**, 29-44.
36. Wu, W.Y., Lebbink, J.H.G., Kanaar, R., Geijsen, N. and van der Oost, J. (2018) Genome editing by natural and engineered CRISPR-associated nucleases. *Nat Chem Biol*, **14**, 642-651.
37. Takeuchi, N., Wolf, Y.I., Makarova, K.S. and Koonin, E.V. (2012) Nature and intensity of selection pressure on CRISPR-associated genes. *Journal of bacteriology*, **194**, 1216-1225.
38. Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J., Wolf, Y.I., Yakunin, A.F. *et al.* (2011)

- Evolution and classification of the CRISPR-Cas systems. *Nature reviews. Microbiology*, **9**, 467-477.
39. Koonin, E.V. and Makarova, K.S. (2019) Origins and evolution of CRISPR-Cas systems. *Philos Trans R Soc Lond B Biol Sci*, **374**, 20180087.
 40. Garneau, J.E., Dupuis, M.E., Villion, M., Romero, D.A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadan, A.H. and Moineau, S. (2010) The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature*, **468**, 67-71.
 41. 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. *Nature reviews. Microbiology*, **15**, 169-182.
 42. 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 & development*, **30**, 447-459.
 43. 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.
 44. Nunez, 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. *Nature structural & molecular biology*, **21**, 528-534.

45. Wei, Y., Chesne, M.T., Terns, R.M. and Terns, M.P. (2015) Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in *Streptococcus thermophilus*. *Nucleic acids research*, **43**, 1749-1758.
46. Wei, Y., Terns, R.M. and Terns, M.P. (2015) Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. *Genes & development*, **29**, 356-361.
47. 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.
48. 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. *Nat Commun*, **3**, 945.
49. Swarts, D.C., Mosterd, C., van Passel, M.W. and Brouns, S.J. (2012) CRISPR interference directs strand specific spacer acquisition. *PloS one*, **7**, e35888.
50. Jackson, S.A., McKenzie, R.E., Fagerlund, R.D., Kieper, S.N., Fineran, P.C. and Brouns, S.J. (2017) CRISPR-Cas: Adapting to change. *Science*, **356**, eaal5056.
51. 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.
52. Paez-Espino, D., Morovic, W., Sun, C.L., Thomas, B.C., Ueda, K., Stahl, B., Barrangou, R. and Banfield, J.F. (2013) Strong bias in the bacterial CRISPR elements that confer immunity to phage. *Nat Commun*, **4**, 1430.

53. Dillingham, M.S. and Kowalczykowski, S.C. (2008) RecBCD enzyme and the repair of double-stranded DNA breaks. *Microbiol Mol Biol Rev*, **72**, 642-671, Table of Contents.
54. 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.
55. Shiimori, M., Garrett, S.C., Chambers, D.P., Glover, C.V.C., 3rd, Graveley, B.R. and Terns, M.P. (2017) Role of free DNA ends and protospacer adjacent motifs for CRISPR DNA uptake in *Pyrococcus furiosus*. *Nucleic acids research*, **45**, 11281-11294.
56. Deveau, H., Barrangou, R., Garneau, J.E., Labonte, 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.
57. Mojica, F.J., Diez-Villasenor, C., Garcia-Martinez, J. and Almendros, C. (2009) Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology*, **155**, 733-740.
58. Erdmann, S. and Garrett, R.A. (2012) Selective and hyperactive uptake of foreign DNA by adaptive immune systems of an archaeon via two distinct mechanisms. *Molecular microbiology*, **85**, 1044-1056.
59. Shah, S.A., Erdmann, S., Mojica, F.J. and Garrett, R.A. (2013) Protospacer recognition motifs: mixed identities and functional diversity. *RNA biology*, **10**, 891-899.

60. 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.
61. 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. *Cell Reports*, **22**, 3377-3384.
62. 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.e45.
63. Shiimori, M., Garrett, S.C., Graveley, B.R. and Terns, M.P. (2018) Cas4 Nucleases Define the PAM, Length, and Orientation of DNA Fragments Integrated at CRISPR Loci. *Mol Cell*, **70**, 814-824 e816.
64. Lee, H., Dhingra, Y. and Sashital, D.G. (2019) The Cas4-Cas1-Cas2 complex mediates precise prespacer processing during CRISPR adaptation. *eLife*, **8**, e44248.
65. Zhang, Z., Pan, S., Liu, T., Li, Y. and Peng, N. (2019) Cas4 nucleases can effect specific integration of CRISPR spacers. *Journal of bacteriology*, **201**, e00747-00718.
66. Almendros, C., Nobrega, F.L., McKenzie, R.E. and Brouns, S.J.J. (2019) Cas4-Cas1 fusions drive efficient PAM selection and control CRISPR adaptation. *Nucleic acids research*, **47**, 5223-5230.

67. 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*. *Journal of the American Chemical Society*, **135**, 17476-17487.
68. 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.
69. Lee, H., Zhou, Y., Taylor, D.W. and Sashital, D.G. (2018) Cas4-Dependent Prespacer Processing Ensures High-Fidelity Programming of CRISPR Arrays. *Mol Cell*, **70**, 48-59 e45.
70. 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. *Cell Rep*, **22**, 3377-3384.
71. Kaushik, V., Verma, V.V. and Goel, M. (2018) Functional divergence and comparative in-silico study of Cas4 proteins of DUF83 class. *J Mol Recognit*, **31**, e2694.
72. 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.
73. Marraffini, L.A. and Sontheimer, E.J. (2010) Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature*, **463**, 568-571.

74. Staals, R.H., 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*. *Mol Cell*, **56**, 518-530.
75. Tamulaitis, G., Kazlauskienė, M., Manakova, E., Venclovas, C., 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*. *Mol Cell*, **56**, 506-517.
76. Johnson, K., Learn, B.A., Estrella, M.A. and Bailey, S. (2019) -Target sequence requirements of a type III-B CRISPR-Cas immune system. *The Journal of biological chemistry*, 1-19.
77. Silas, S., Mohr, G., Sidote, D.J., Markham, L.M., Sanchez-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.
78. 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.* (2017) On the Origin of Reverse Transcriptase-Using CRISPR-Cas Systems and Their Hyperdiverse, Enigmatic Spacer Repertoires. *mBio*, **8**, e00897-00817.
79. 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.

80. Nunez, J.K., Lee, A.S., Engelman, A. and Doudna, J.A. (2015) Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature*, **519**, 193-198.
81. 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.
82. 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.
83. Liu, T., Liu, Z., Ye, Q., Pan, S., Wang, X., Li, Y., Peng, W., Liang, Y., She, Q. and Peng, N. (2017) Coupling transcriptional activation of CRISPR-Cas system and DNA repair genes by Csa3a in *Sulfolobus islandicus*. *Nucleic acids research*, **45**, 8978-8992.
84. 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.
85. 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.
86. 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. *Proc Natl Acad Sci U S A*, **114**, E5122-E5128.

87. Nunez, J.K., Bai, L., Harrington, L.B., Hinder, T.L. and Doudna, J.A. (2016) CRISPR Immunological Memory Requires a Host Factor for Specificity. *Mol Cell*, **62**, 824-833.
88. Yoganand, K.N., 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.
89. 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**, 1113-1118.
90. Wright, A.V. and Doudna, J.A. (2016) Protecting genome integrity during CRISPR immune adaptation. *Nature structural & molecular biology*, **23**, 876-883.
91. Xue, C., Seetharam, A.S., Musharova, O., Severinov, K., Brouns, S.J., Severin, A.J. and Sashital, D.G. (2015) CRISPR interference and priming varies with individual spacer sequences. *Nucleic acids research*, **43**, 10831-10847.
92. Kunne, T., Kieper, S.N., Bannenberg, J.W., Vogel, A.I., Miellet, W.R., Klein, M., Depken, M., Suarez-Diez, M. and Brouns, S.J. (2016) Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation. *Mol Cell*, **63**, 852-864.
93. Staals, R.H., Jackson, S.A., Biswas, A., Brouns, S.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.

94. 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 Microbiol Rev*, **39**, 428-441.
95. Hochstrasser, M.L. and Doudna, J.A. (2015) Cutting it close: CRISPR-associated endoribonuclease structure and function. *Trends in biochemical sciences*, **40**, 58-66.
96. Wiedenheft, B., Lander, G.C., Zhou, K., Jore, M.M., Brouns, S.J.J., van der Oost, J., Doudna, J.A. and Nogales, E. (2011) Structures of the RNA-guided surveillance complex from a bacterial immune system. *Nature*, **477**, 486-489.
97. Rouillon, C., Zhou, M., Zhang, J., Politis, A., Beilsten-Edmands, V., Cannone, G., Graham, S., Robinson, C.V., Spagnolo, L. and White, M.F. (2013) Structure of the CRISPR interference complex CSM reveals key similarities with cascade. *Mol Cell*, **52**, 124-134.
98. Spilman, M., Cocozaki, A., Hale, C., Shao, Y., Ramia, N., Terns, R., Terns, M., Li, H. and Stagg, S. (2013) Structure of an RNA silencing complex of the CRISPR-Cas immune system. *Mol Cell*, **52**, 146-152.
99. Staals, R.H.J., Agari, Y., Maki-Yonekura, S., Zhu, Y., Taylor, D.W., van Duijn, E., Barendregt, A., Vlot, M., Koehorst, J.J., Sakamoto, K. *et al.* (2013) Structure and activity of the RNA-targeting Type III-B CRISPR-Cas complex of *Thermus thermophilus*. *Mol Cell*, **52**, 135-145.
100. Jiang, F. and Doudna, J.A. (2017) CRISPR-Cas9 Structures and Mechanisms. *Annu Rev Biophys*, **46**, 505-529.

101. 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.
102. Liu, L., Chen, P., Wang, M., Li, X., Wang, J., Yin, M. and Wang, Y. (2017) C2c1-sgRNA Complex Structure Reveals RNA-Guided DNA Cleavage Mechanism. *Mol Cell*, **65**, 310-322.
103. 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. *Mol Cell*, **60**, 385-397.
104. 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.
105. East-Seletsky, A., O'Connell, M.R., Knight, S.C., Burstein, D., Cate, J.H., 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.
106. Fonfara, I., Richter, H., Bratovic, M., Le Rhun, A. and Charpentier, E. (2016) The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. *Nature*, **532**, 517-521.
107. Nishimasu, H. and Nureki, O. (2017) Structures and mechanisms of CRISPR RNA-guided effector nucleases. *Curr Opin Struct Biol*, **43**, 68-78.

108. Plagens, A., Richter, H., Charpentier, E. and Randau, L. (2015) DNA and RNA interference mechanisms by CRISPR-Cas surveillance complexes. *FEMS Microbiol Rev*, **39**, 442-463.
109. Liu, L., Li, X., Wang, J., Wang, M., Chen, P., Yin, M., Li, J., Sheng, G. and Wang, Y. (2017) Two Distant Catalytic Sites Are Responsible for C2c2 RNase Activities. *Cell*, **168**, 121-134 e112.
110. 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.
111. 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.
112. 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. *The EMBO journal*, **30**, 1335-1342.
113. Jackson, R.N. and Wiedenheft, B. (2015) A Conserved Structural Chassis for Mounting Versatile CRISPR RNA-Guided Immune Responses. *Mol Cell*, **58**, 722-728.
114. 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.

115. 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.
116. Estrella, M.A., Kuo, F.T. and Bailey, S. (2016) RNA-activated DNA cleavage by the Type III-B CRISPR-Cas effector complex. *Genes & development*, **30**, 460-470.
117. Han, W., Li, Y., Deng, L., Feng, M., Peng, W., Hallstrom, S., Zhang, J., Peng, N., Liang, Y.X., White, M.F. *et al.* (2017) A type III-B CRISPR-Cas effector complex mediating massive target DNA destruction. *Nucleic acids research*, **45**, 1983-1993.
118. Kazlauskienė, M., Tamulaitis, G., Kostiuk, G., Venclovas, C. and Siksnys, V. (2016) Spatiotemporal Control of Type III-A CRISPR-Cas Immunity: Coupling DNA Degradation with the Target RNA Recognition. *Mol Cell*, **62**, 295-306.
119. Liu, T.Y., Iavarone, A.T. and Doudna, J.A. (2017) RNA and DNA Targeting by a Reconstituted *Thermus thermophilus* Type III-A CRISPR-Cas System. *PloS one*, **12**, e0170552.
120. Garcia-Doval, C. and Jinek, M. (2017) Molecular architectures and mechanisms of Class 2 CRISPR-associated nucleases. *Curr Opin Struct Biol*, **47**, 157-166.
121. 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.
122. Tsui, T.K. and Li, H. (2015) Structure Principles of CRISPR-Cas Surveillance and Effector Complexes. *Annu Rev Biophys*, **44**, 229-255.

123. Silas, S., Lucas-Elio, P., Jackson, S.A., Aroca-Crevillen, A., Hansen, L.L., Fineran, P.C., Fire, A.Z. and Sanchez-Amat, A. (2017) Type III CRISPR-Cas systems can provide redundancy to counteract viral escape from type I systems. *eLife*, **6**, e27601.
124. Fiala, G. and Stetter, K.O. (1986) *Pyrococcus-Furiosus* Sp-Nov Represents a Novel Genus of Marine Heterotrophic Archaeobacteria Growing Optimally at 100-Degrees C. *Arch Microbiol*, **145**, 56-61.
125. Terns, R.M. and Terns, M.P. (2013) The RNA- and DNA-targeting CRISPR-Cas immune systems of *Pyrococcus furiosus*. *Biochemical Society transactions*, **41**, 1416-1421.
126. Majumdar, S., Zhao, P., Pfister, N.T., Compton, M., Olson, S., Glover, C.V., 3rd, Wells, L., Graveley, B.R., Terns, R.M. and Terns, M.P. (2015) Three CRISPR-Cas immune effector complexes coexist in *Pyrococcus furiosus*. *RNA*, **21**, 1147-1158.
127. Elmore, J., Deighan, T., Westpheling, J., Terns, R.M. and Terns, M.P. (2015) DNA targeting by the type I-G and type I-A CRISPR-Cas systems of *Pyrococcus furiosus*. *Nucleic acids research*, **43**, 10353-10363.
128. 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.
129. Hale, C., Kleppe, K., Terns, R.M. and Terns, M.P. (2008) Prokaryotic silencing (psi)RNAs in *Pyrococcus furiosus*. *RNA*, **14**, 2572-2579.

130. Wang, R., Preamplume, G., Terns, M.P., Terns, R.M. and Li, H. (2011) Interaction of the Cas6 ribonuclease with CRISPR RNAs: recognition and cleavage. *Structure*, **19**, 257-264.
131. Carte, J., Pfister, N.T., Compton, M.M., Terns, R.M. and Terns, M.P. (2010) Binding and cleavage of CRISPR RNA by Cas6. *RNA*, **16**, 2181-2188.
132. Majumdar, S. and Terns, M.P. (2019) CRISPR RNA-guided DNA cleavage by reconstituted Type I-A immune effector complexes. *Extremophiles*, **23**, 19-33.
133. Majumdar, S., Ligon, M., Skinner, W.C., Terns, R.M. and Terns, M.P. (2017) Target DNA recognition and cleavage by a reconstituted Type I-G CRISPR-Cas immune effector complex. *Extremophiles*, **21**, 95-107.
134. Hale, C.R., Majumdar, S., Elmore, J., Pfister, N., Compton, M., Olson, S., Resch, A.M., Glover, C.V., 3rd, Graveley, B.R., Terns, R.M. *et al.* (2012) Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. *Mol Cell*, **45**, 292-302.
135. Hudaiberdiev, S., Shmakov, S., Wolf, Y.I., Terns, M.P., Makarova, K.S. and Koonin, E.V. (2017) Phylogenomics of Cas4 family nucleases. *BMC Evol Biol*, **17**, 232.
136. Delorme, C., Bartholini, C., Luraschi, M., Pons, N., Loux, V., Almeida, M., Guedon, E., Gibrat, J.F. and Renault, P. (2011) Complete genome sequence of the pigmented *Streptococcus thermophilus* strain JIM8232. *Journal of bacteriology*, **193**, 5581-5582.

137. Ichikawa, H.T., Cooper, J.C., Lo, L., Potter, J., Terns, R.M. and Terns, M.P.
(2017) Programmable type III-A CRISPR-Cas DNA targeting modules. *PloS one*, **12**, e0176221.
138. Kazlauskienė, M., Kostiuk, G., Venclovas, C., Tamulaitis, G. and Siksnys, V.
(2017) A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems. *Science*, **357**, 605-609.
139. Terns, R.M. and Terns, M.P. (2014) CRISPR-based technologies: prokaryotic defense weapons repurposed. *Trends in genetics : TIG*, **30**, 111-118.

Figure 1.1. Overview of CRISPR-Cas defense.

During adaptation, foreign DNA is recognized by the host prokaryote. A short DNA fragment directly next to a “protospacer adjacent motif” (PAM) is captured and integrated into the CRISPR locus on the host’s genome. On the foreign DNA this fragment is called a protospacer, and after integration into the CRISPR locus this sequence is called a spacer. During crRNA biogenesis, the CRISPR locus is transcribed and processed into CRISPR RNAs (crRNAs). The crRNAs assemble into an effector complex with CRISPR-associated (Cas) proteins and guide the complex to the foreign nucleic acid through base complementarity. During invader silencing, if the same PAM sequence is recognized, the effector complex is prompted to degrade the invading nucleic acid, and the threat is silenced. Figure adapted from Terns and Terns 2014 (139).

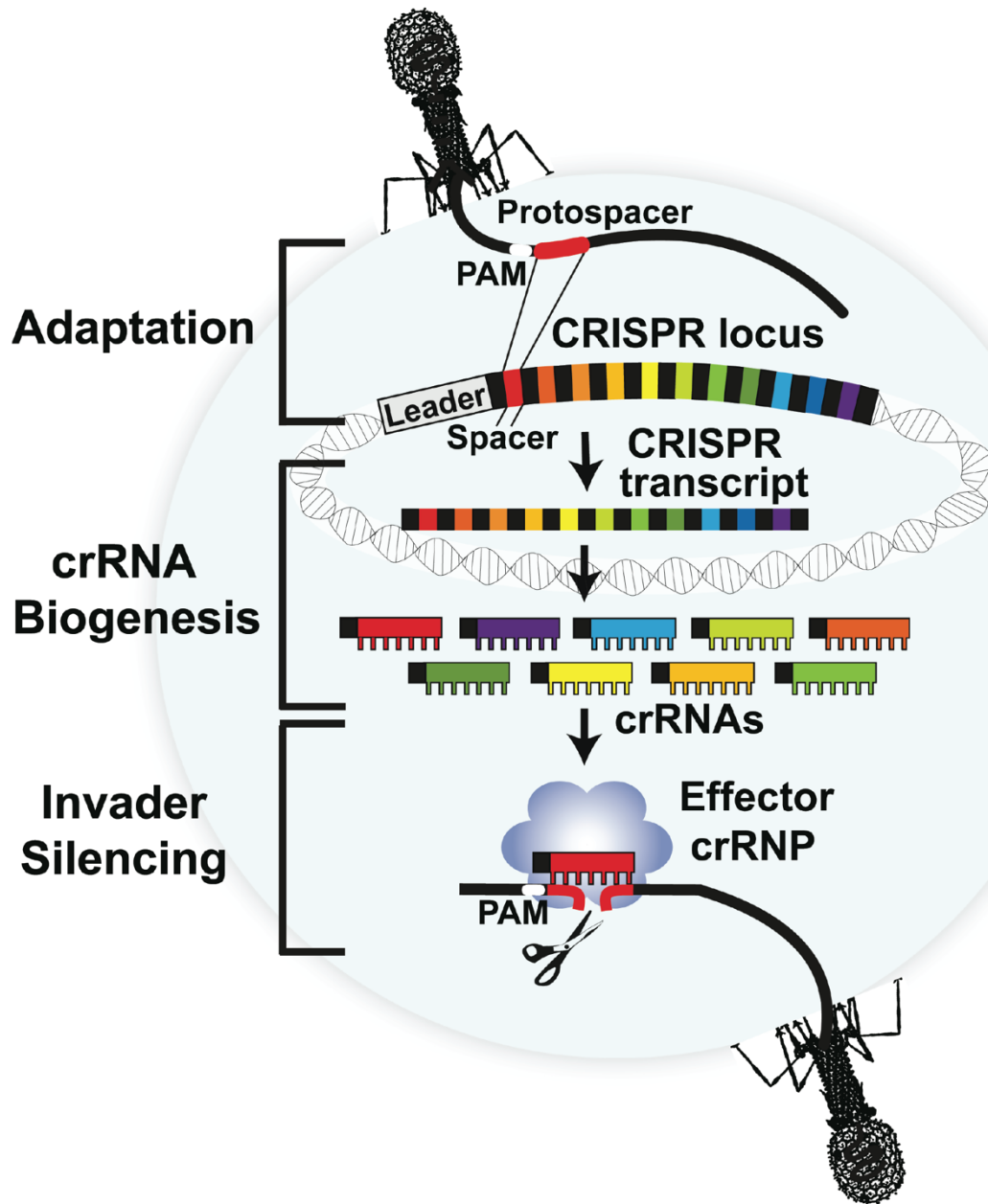


Figure 1.2. Classification of CRISPR-Cas systems.

CRISPR-Cas systems are classified into two classes, six main types, and several subtypes based on the *cas* genes contained in their respective systems. Class 1 systems all contain multi-subunit effector complexes, while Class 2 systems contain single effector proteins. The principal Cas proteins of each of the six CRISPR-Cas system types are illustrated. Dispensable genes are indicated by a dashed outline. SS is the small subunit and LS is the large subunit of the effector complexes. Example architectures of the multi-subunit (Class 1) and single protein (Class 2) effector complexes are illustrated for reference. Figure adapted and modified from Koonin and Makarova 2019 (39).

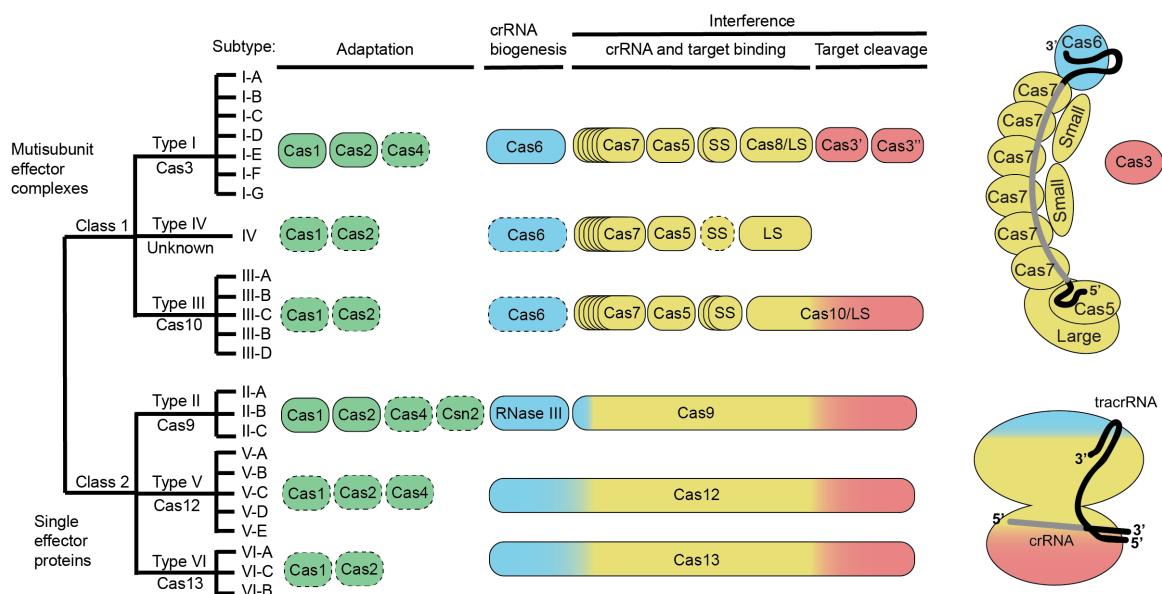


Figure 1.3. Model of CRISPR-Cas adaptation.

(A) Adaptation occurs in two main steps: pre-spacer generation and spacer integration.

(B) During pre-spacer generation, the PAM is cleaved from the protospacer, and the pre-spacer is processed to the proper size necessary for spacer integration. (C) During spacer integration, the 3' hydroxyl groups of the spacer attack the borders of the CRISPR repeat on the top and bottom strands. The resulting single-stranded repeats are repaired by host polymerase and ligase, allowing duplication of the repeat each time a new spacer is added to the array.

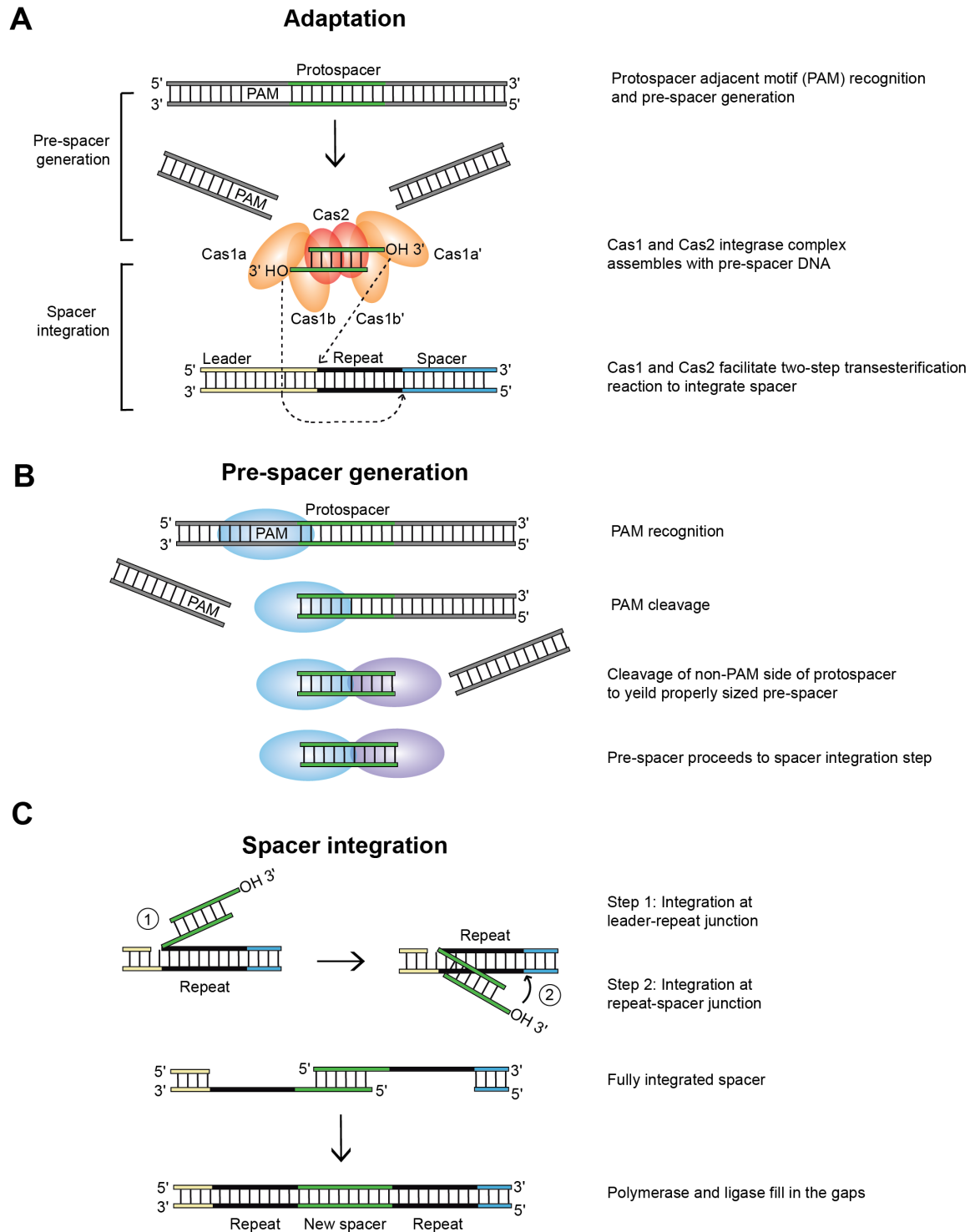


Figure 1.4. CRISPR-Cas systems of *Pyrococcus furiosus*.

(A) The arrangement *P. furiosus* *cas* genes is illustrated. *P. furiosus* has three CRISPR-Cas systems present on its genome: a Type III-B Cmr system (blue), a Type I-G Cst system (yellow), and a Type I-A Csa system (red). One set of adaptation genes (green) is shared by all three systems and services all seven active CRISPR arrays on the genome.

(B) The locations of each active CRISPR locus and the *cas* genes are illustrated. Figure adapted and modified from Shiimori 2018 (63).

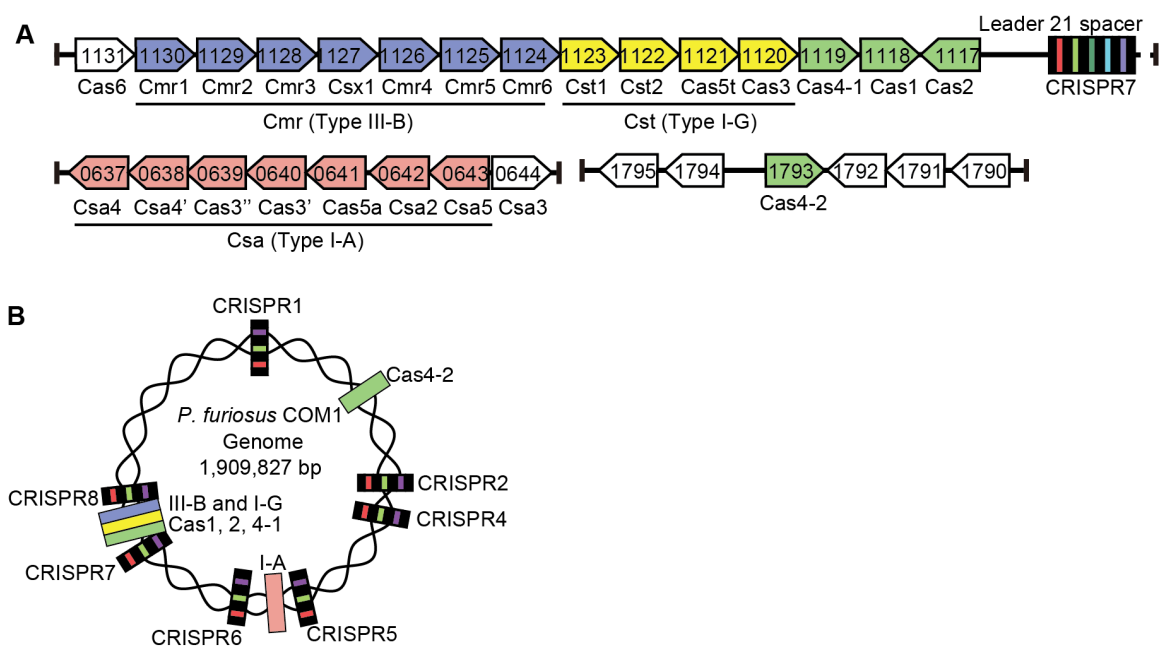
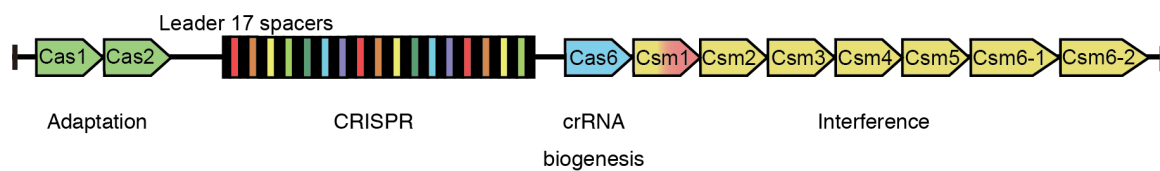


Figure 1.5. Type III-A CRISPR-Cas system of *Streptococcus thermophilus*.

The arrangement of *S. thermophilus* Type III-A CRISPR-*cas* genes is illustrated. The system includes Cas1 and Cas2 proteins predicted to be involved in adaptation (green), Cas6 for crRNA biogenesis (blue), and Csm1-Csm6-2 for invader silencing (yellow).



CHAPTER 2

CRISPR REPEAT SEQUENCES AND RELATIVE SPACING SPECIFY DNA
INTEGRATION BY *PYROCOCCUS FURIOSUS* CAS1 AND CAS2¹

¹Julie Grainy, Sandra Garrett, Brenton R. Graveley, Michael P. Terns. Accepted by Nucleic Acids Research. Reprinted here with permission of publisher.

Abstract

Acquiring foreign spacer DNA into the CRISPR locus is an essential primary step of the CRISPR-Cas pathway in prokaryotes for developing host immunity to mobile genetic elements. Here, we investigate spacer integration *in vitro* using proteins from *Pyrococcus furiosus* and demonstrate that Cas1 and Cas2 are sufficient to accurately integrate spacers into a minimal CRISPR locus. Using high-throughput sequencing, we identified high frequency spacer integration occurring at the same CRISPR repeat border sites utilized *in vivo*, as well as at several non-CRISPR plasmid sequences which share features with repeats. Analysis of non-CRISPR integration sites revealed that Cas1 and Cas2 are directed to catalyze full-site spacer integration at specific DNA stretches where guanines and/or cytosines are 30 base pairs apart and the intervening sequence harbors several positionally conserved bases. Moreover, assaying a series of CRISPR repeat mutations, followed by sequencing of the integration products, revealed that the specificity of integration is primarily directed by sequences at the leader-repeat junction as well as an adenine-rich sequence block in the mid-repeat. Together, our results indicate that *P. furiosus* Cas1 and Cas2 recognize multiple sequence features distributed over a 30 base pair DNA region for accurate spacer integration at the CRISPR repeat.

Introduction

CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) arrays, found on the genomes of roughly half of bacteria and the majority of archaea, harbor a genetic memory of previously encountered nucleic acid invaders (1,2). Cas (CRISPR-associated) proteins recurrently add short, ~30-40 bp DNA sequences (called spacers) to CRISPR

arrays, and coordinate the defense against the nucleic acid invaders upon subsequent infections (1,3,4). CRISPR uptake of spacer DNA fragments occurs through the process of adaptation and leads to heritable genetic memories of the infecting virus or other mobile genetic element (5-8). During adaptation, segments of the invader are cleaved and typically processed at protospacer adjacent motifs (PAMs) before being directionally integrated into the CRISPR genetic memory bank (9-12). The CRISPR locus includes a leader sequence followed by an array of short repeating DNA sequences, separated by unique spacers derived from past infections (13). In order to defend against the invading mobile genetic element, the CRISPR locus is transcribed and processed into CRISPR RNAs (crRNAs) that guide effector Cas proteins to the invading nucleic acid through base complementarity (3,14,15). When bound to a complementary target nucleic acid, crRNA-associated Cas nucleases degrade the invader and the threat is silenced (16-19). Diverse CRISPR-Cas systems have evolved that each employ distinct Cas protein family members and mechanisms for defense and have been categorized into 6 types and more than 30 subtypes (20,21). In contrast, the proteins that perform the integration of new spacers into CRISPR arrays, Cas1 and Cas2, are common components found in the various types of active CRISPR-Cas systems, and sequences of CRISPR leaders and repeats tend to coevolve with Cas1 sequences (20,22).

The model organism *Pyrococcus furiosus*, an extreme hyperthermophile, has proven to be useful for investigating the understudied archaeal adaptation mechanisms (12,23). The *P. furiosus* genome contains type I-A, I-G, and III-B CRISPR-Cas systems that use crRNA guides generated from seven active CRISPR arrays (4,24-28). New spacers are added to each of these CRISPR arrays by a single, shared set of Cas1 and

Cas2 proteins (23), after proper processing of the protospacer by Cas4-1 and Cas4-2 (12). Cas1 has been shown to functionally and physically interact with Cas4 proteins in several studied systems (12,29-31). *P. furiosus* Cas4-1 and Cas4-2 nucleases are necessary for PAM recognition and removal during spacer generation, for setting spacer length, and for ensuring that new spacers are efficiently integrated in the correct orientation that supports immunity (12). However, in the absence of both Cas4-1 and Cas4-2 proteins, Cas1 and Cas2 proteins retain the ability to accurately integrate spacers at CRISPR repeats *in vivo* (12) raising the key question of what specific *cis*-acting determinants guide Cas1 and Cas2 proteins to localize at CRISPR loci.

The *P. furiosus* Cas1 and Cas2 genes are encoded immediately adjacent to genes of the type I-G Cas proteins (Cst module) (23) suggesting that they co-evolved to function as the adaptation module associated with the type I-G system (rather than III-B or I-A system). Figure S1 displays a model of *P. furiosus* adaptation based on *in vivo* experiments conducted with *P. furiosus* (12,23) and *in vitro* experiments with type I-E and I-A systems (11,30,32-36). The spacer integration step of this process is achieved through a two-step transesterification reaction, where the 3' hydroxyl groups of the incoming, predominately double-stranded DNA spacer attack both borders of the first repeat sequence, on opposite strands (34,37). The majority of studies suggest that type I and II systems first attack the top strand at the leader-repeat junction (LR), followed by a second attack of the repeat-spacer junction (RS) on the bottom strand (8,38). Direct, *in vitro* kinetic tests on the order of attacks have only been done for I-E and II-A systems (6,32,35), and these tests should be expanded to include more systems, given their diversity and ability to achieve immunity through unique mechanisms.

Details of the CRISPR spacer integration reaction have begun to be elucidated through *in vivo* and *in vitro* studies across a variety of CRISPR-Cas systems. In both type I and II systems, a Cas1₄-Cas2₂ complex has been found consisting of two Cas1 dimers bridged by a central Cas2 dimer, with Cas1 serving as the integrase and Cas2 likely playing a structural role (11,31,34,35,39,40). The Cas1-Cas2 integrase complex is directed to integrate new spacers at the leader-proximal repeat (1,6,7,39,41,42), but the mechanisms guiding this targeted integration are proving to be quite diverse across different subtypes. Some systems require non-Cas host factors, such as the leader-binding integration host factor (IHF), to guide Cas1-Cas2 to the first repeat in the array (32,35,43,44). In contrast, the type II-A systems lack a requirement for IHF and instead have an intrinsic ability to direct spacer integration, likely due to protein-DNA interactions between Cas1 and sequences surrounding the leader-repeat junction (6,40,41). It remains to be determined how spacer integration is directed in adaptation systems of archaea, which lack IHF, but preliminary findings implicate both sequence and protein components. The large (531 bp) archaeal leader in *Sulfolobus solfataricus* is important for type I-A adaptation, and *in vitro* spacer integration required unidentified host factor(s) from cell lysate and ATP hydrolysis (36). As this has been the only report on *in vitro* spacer integration in archaeal systems, there is still much unknown about how spacers are directed for integration at CRISPR repeats in archaeal organisms.

Sequences in the leader and repeat have been shown to influence proper spacer integration. Two *in vivo* studies investigating I-E and I-B spacer addition and repeat duplication identified mid-repeat motifs that orchestrate accurate spacer integration at LR and RS sites (45,46). In the I-E study, the essential motifs were situated in palindromic

inverted repeats (IRs) within the CRISPR repeat (45), while in the I-B study, accurate integration sites relied on a motif between the IRs. While IRs are a common feature among many groups of CRISPR repeats, both studies found that the ability for IRs to form cruciform structures was not necessary for adaptation (45,46). Recent high resolution cryo-EM and X-ray crystallography structures showed Cas1-Cas2 bound to pre-spacers being integrated into minimal CRISPR loci (11,35,40), and suggested that these important mid-repeat motifs likely serve as binding sites for the integrase machinery. Several *in vitro* studies have investigated *cis*-acting elements of the repeat and leader through a variety of sequence mutations (6,8,32,35,36,40). However, results have varied depending on the system, again highlighting the mechanistic diversity of CRISPR systems. It is important to note that these *in vitro* mutational studies have been limited by drawing conclusions solely on apparent integration efficiency rather than determining the sites of integration by direct sequencing methods.

Here we established an *in vitro* system capable of accurately integrating spacer DNA at the proper junctions of *P. furiosus* CRISPR repeats. Interestingly, spacers also became integrated into plasmid DNA at a subset of non-CRISPR (“off-target”) DNA regions that share specific sequence features with CRISPR repeats. We found *P. furiosus* Cas1 and Cas2 to be necessary and sufficient for complete (full-site) spacer integration. Specific DNA determinants important for efficient and accurate integration of spacer DNA were identified both by evaluating reaction efficiency and monitoring the precise location of spacer integration by sequencing for an extensive panel of CRISPR repeat and leader mutant target DNAs. Together, our results indicate that *P. furiosus* Cas1 and Cas2 recognize multiple sequence features distributed over an approximately 30 base pair

DNA stretch to enable accurate integration of spacer DNA into the repeats of target CRISPR loci. Our results implicate specific sequence elements and their relative spacing to each other in properly positioning the *P. furiosus* Cas1 and Cas2 integrase proteins for accurate spacer integration at the CRISPR repeat.

Materials and Methods

Cloning, expression and purification

The *cas1* (PF1117) and *cas2* (PF1118) genes were amplified from *Pyrococcus furiosus* COM1 genomic DNA, individually cloned into pET21d expression vectors, and transformed into *E. coli* BL21 RIPL strain. Cultures were grown at 37°C in 200 mL of Luria broth to an OD₆₀₀ of 0.4-0.6, and expression of the C-terminal 6x histidine tagged proteins was induced with 1 mM IPTG during overnight growth at room temperature. Harvested cells were lysed in 40 mM Tris, 500 mM KCl, 10% glycerol, 5 mM imidazole, pH 7.5. Following thermal precipitation at 70°C for 30 minutes, the cell lysate was centrifuged at 14,000 rpm at 4°C for 30 minutes and the soluble fraction was collected and filtered (0.8 µm filter pore size Millex filter unit, Millipore). His-tagged proteins were purified by Ni²⁺ affinity column chromatography, using a stepwise increase of imidazole (10, 20, 50, 100, 250, 500 mM) in 40 mM Tris, 500 mM KCl, 10% glycerol, pH 7.5. Peak elution fractions were dialyzed using Slide-a-lyzer mini dialysis cassettes (Thermo Fisher) into 40 mM Tris, 200 mM KCl, 10% glycerol, pH 7.5, and stored at 4°C prior to use for functional assays. C-terminal 6x histidine tagged proteins were

confirmed to be functional for adaptation *in vivo* (data not shown), ruling out potential detrimental effects of the tag.

DNA substrate preparation

DNA oligonucleotides were from Eurofins Genomics (for minimal CRISPR substrates and PCR primers) and Integrated DNA Technologies (for pre-spacer DNA and hairpin CRISPR substrates). Oligonucleotides used to make pre-spacers and CRISPR substrates were separated by 15% denaturing polyacrylamide gel electrophoresis in 1x TBE, detected by ethidium bromide staining, and the band of the expected oligonucleotide size was excised. Oligonucleotides were eluted from the gel slices overnight at 4°C in 500 μ L of elution buffer (0.5 M ammonium acetate, 1 mM EDTA (pH 8.0), and 0.1% SDS), extracted with phenol/chloroform/isoamyl alcohol (pH 8.0), ethanol precipitated, and resuspended in 20 mM Tris, 100 mM KCl, 5% glycerol, pH 7.5. Corresponding oligonucleotides were annealed by incubating at 95°C for 5 minutes followed by slow cooling until 23°C. Annealing was confirmed by 10% non-denaturing polyacrylamide gel electrophoresis in 1x TBE. Pre-spacers and half-site CRISPR substrates were 5' radiolabeled with T4 polynucleotide kinase (PNK) and [γ -³²P] ATP. In the case of the half-site substrates, a second gel extraction and precipitation was performed after annealing. Oligonucleotide sequences can be found in Supplementary Table S1.

Spacer integration assay

Unless stated otherwise, a final concentration of 1 μ M Cas1, 1 μ M Cas2, 1 mM DTT, and 10 mM MgCl₂ was incubated in reaction buffer (20 mM Tris, 100 mM KCl, 5%

glycerol, pH 7.5) at 4°C for one hour. 20 nM of radiolabeled pre-spacer or 100 nM of unlabeled pre-spacer was added to the reactions and incubated at 4°C for an additional 15 minutes. Finally, plasmid or linear DNA was added to a final concentration of 5 nM for pCR7, or 100 nM for minimal CRISPR substrate, and the reaction was incubated at 70°C for one hour. Reactions with pCR7 were quenched with EDTA and proteinase K (Life Technologies). Products were mixed with gel loading dye (purple, NEB) and separated on 1% agarose gels in 1x TAE. Reactions with minimal CRISPR substrates were quenched with an equal volume of Gel Loading Buffer II (Thermofisher) and 25 mM EDTA. Samples were boiled for 5 minutes before separation by 12% denaturing 7M urea-containing polyacrylamide gel electrophoresis in 1x TBE. Gels were dried and radioactivity was detected with phosphorimaging (Storm 840 Scanner GE Healthcare).

Analysis of integration by high-throughput sequencing

Library preparation. To analyze integrations by high-throughput sequencing, the spacer integration assay was performed as described above using unlabeled pre-spacer. Following incubation, DNA was isolated using the DNA Clean and Concentrator Kit (Zymo Research, Irvine CA). For the plasmid integration samples, excess un-integrated pre-spacer was removed using Agencourt AMPure XP beads (Beckman Coulter, Indianapolis, IN). Next, Illumina adapter sequence with an N10 random primer was annealed to the plasmid DNA and extended (thermocycler conditions: 98°C for 30s, 25°C for 30s, 35°C for 30s, 45°C for 30S, and 72°C for 5 min). Following extension, excess adapter was removed using AMPure beads, and PCR was done to amplify plasmid DNA that contained integrated pre-spacer: forward primers were specific for the pre-spacer,

while reverse primers targeted the Illumina adapter introduced with the random anneal and extension step. This amplified both full-site and half-site integration events with no discrimination. Illumina barcodes and additional adapter sequences were added with a final PCR and the resulting library was separated on a 1% agarose gel to select for DNA in a 400 to 700 bp size range. DNA was isolated using the Zymo Gel DNA Recovery Kit (Zymo Research, Irvine CA) and sequenced on an Illumina MiSeq in a 100 by 50 cycle run. Only the 100 bp Read 1 data was used in this analysis.

For the minimal linear CRISPR substrate products, 1 μ l of eluted DNA was used as a PCR template. Primers to add Illumina adaptor sequences were annealed to the newly integrated spacer and the 3' end of either the plus or minus strand of the CRISPR substrate. DNA Clean and Concentrator Kit (Zymo Research, Irvine CA) was used to isolate the PCR product, and 1 μ l of this product was used as the template for a second PCR using primers to add Illumina barcodes. These products were purified on a 1% agarose gel and extracted with a Gel Purification Kit (Zymo Research, Irvine CA).

Mapping integration events. After sequencing, samples were de-multiplexed by barcode and analyzed to determine sites of integration. For plasmid data, the complete pre-spacer sequence was located in each read and 50 bp of sequence immediately downstream from the end of the pre-spacer was extracted. These 50 bp sequences were aligned to the appropriate plasmid reference using Bowtie (47). To visualize the distribution of integration events, alignment output files were converted into coverage files using bedtools (48) and displayed on a custom UCSC genome browser track hub

(<https://www.genome.ucsc.edu>). An initial inspection of the integration tracks revealed that large peaks occurred outside of the CRISPR arrays and data suggested that these peaks were due to particular sequence and spatial features. These trends were both analyzed in an unbiased plasmid-wide manner. To determine sequence preferences at the sites of integration, the base at the integration point, along with upstream and downstream context sequence, was extracted from the reference sequence (bedtools) and used to make sequence logos (49). For spacing trends, we took the browser track files and assessed the distances between two large peaks occurring on the same strand or on opposite strands. To do this, 500 random 50 bp intervals (“windows”) were selected for each plasmid. Within each of these windows, the two highest peaks on the plus strand and the minus strand were identified and the bp distance between these peaks was determined (highest to second highest on the plus strand, highest to second highest on the minus strand, highest plus strand to highest minus strand). Distance values from all 500 windows were then binned and counted. For the minimal linear CRISPR integration data, the spacer-target junction was determined from each read and counts for each potential integration point were totaled.

Results

P. furiosus* Cas1 and Cas2 facilitate spacer integration *in vitro

To determine the requirements for the integration of new spacers into the CRISPR array, Cas1 and Cas2 recombinant proteins cloned from *P. furiosus* were expressed and purified from *Escherichia coli* and tested in spacer integration assays with a synthetic pre-spacer

and a plasmid containing a minimal CRISPR array (pCR7) (Figure 1). The CRISPR array had a complete 508 bp leader followed by three 30 bp repeats, separated by two 37 bp spacers derived from *P. furiosus* CRISPR locus 7 (Figure 1A). The 37 bp sequence of the pre-spacer DNA was selected due to its high frequency of acquisition from a plasmid during transformation into *P. furiosus* (23). When pCR7 was incubated *in vitro* with Cas1, Cas2, and radiolabeled pre-spacers in the presence of MgCl₂, full-site or half-site integration products (which cause plasmid nicking detectable by agarose gel electrophoresis and DNA staining) were detected by autoradiography (Figure 1B, compare lanes 1-5). Spacer integration was also observed using plasmids that lacked the leader sequence (pCR7 no leader) or control plasmids devoid of a CRISPR locus (pControl) (Figure 1B). Radiolabeled integration products were detected in the nicked and linear conformation of the plasmid, and required the presence of Cas1, Cas2 and MgCl₂ (Figure 1B). Cas1 alone was able to produce extremely low levels of integration, while Cas2 alone could not (Figure S2A) indicating the importance of both Cas1 and Cas2 in catalyzing spacer integration into the target DNA molecules. Pre-spacers with blunt-ends or 3' overhangs (five nucleotides in length) were both integrated into pCR7 (Figure S2B), but subsequent experiments were carried out using pre-spacers with 3' overhangs. Moreover, integration occurred efficiently into linear plasmid DNA, indicating that integration by Cas1 and Cas2 is not strictly dependent on the supercoiled topology of circular plasmid DNA (Figure S2B).

We next mapped the precise sites of spacer integration in each of the tested plasmids by high-throughput sequencing of the integration products (Figure 1C-E). Integration into pCR7 occurred at very low levels at sites distributed throughout the

entire plasmid with a marked increased preference at each of the three CRISPR repeats (occurring precisely at the same top strand and bottom strand repeat junctions as is observed *in vivo*) (Figure 1C and 1F and (23)), as well as several non-CRISPR locations (Figure 1C). The high frequency of integration at the first repeat was lost when the adjacent leader was deleted, but the levels of integration at the second and third repeat remained high (Figure 1D and 1G). The non-CRISPR integration sites remained when the leader or entire CRISPR locus was absent from the plasmid (Figure 1D and 1E). One such highly preferred non-CRISPR site of integration (bracketed in Figure 1C) was investigated further in subsequent experiments. This exact site was also highly preferred in an alternative plasmid backbone containing the same non-CRISPR sequence (pControl2) (Figure S3).

Non-CRISPR integration sites resemble CRISPR repeats

A closer look at the most highly targeted non-CRISPR plasmid DNA integration sites revealed similarities between these sites and the CRISPR repeat, both in size and sequence identity (Figure 2). Pairs of high frequency integration sites on opposite strands, similar to the pattern observed at the repeats, were observed scattered along the plasmid and the distances between the paired non-CRISPR integration sites were consistent with the size of a CRISPR repeat (30 bp), as shown with the non-CRISPR example (Figure 2A). We quantified this spacing trend by randomly selecting windows across the plasmid and determining the distance between the largest peaks on opposite strands. There was a significant trend wherein the highest plus strand peak was 30-31 bp from the highest minus strand peak for both plasmid DNA that contained (pCR7) or

lacked (pControl) the CRISPR array (Figure 2B and 2C). In contrast, when the same analysis was performed for peaks that were both on the plus strand or both on the minus strand, no such 30 bp spacing preference was observed, suggesting the paired peaks on opposite strands are due to the Cas1 and Cas2-catalyzed, full-site integration (two-step transesterification reaction) (Figure S4).

In addition to trends in peak spacing, we also noted that certain bases were conserved in the highly integrated non-CRISPR sites. We therefore analyzed the sequences surrounding these non-CRISPR integration sites to determine if there was any similarity to the leader-repeat elements of the CRISPR locus. The nucleotide identity at the site of integration across the entire plasmid exhibited a strong preference for guanine, with the second most preferred base being cytosine, which is in agreement with the nucleotide identity of the natural *P. furiosus* CRISPR repeat borders (Figure 2D and 2E). Sequences for the top ten most highly integrated non-CRISPR sites of 30 bp spacing are listed in Figure 2F, and their locations are indicated on the plasmid map in Figure S5. A WebLogo created from these top ten sequences revealed similarities to the repeat, particularly at the borders and the center of the repeat (Figure 2G). Specifically, nucleotides G1, T2, A6, A12, T14, A17, A18, A20, A23, T24, T25, A28, and A29 of the repeat, and C-3, C-4, G-7, A-8, A-9, and A-10 of the leader were often enriched in the non-CRISPR sites (Figure 2G). When all integration sites were considered (not just ones with a paired peak pattern), repeat-like features were conserved; most notably, G1, T2, T3, A6, T14, A17, A18, A29, as well as C-3 in the leader (Figure 2H and 2I). The same spacing and sequence trends are seen when analyzing plasmids lacking the leader (pCR7 no leader) and missing the entire leader-CRISPR array (pControl), demonstrating that this

phenomenon is not due to the presence of the CRISPR locus but rather reflects an intrinsic integration site preference of Cas1 and Cas2 (Figure S4 and S6). Taken together, this analysis of integration sites across the plasmid suggests that spacer integration by *P. furiosus* Cas1 and Cas2 proteins is being directed by specific sequences organized in a defined spatial configuration.

Full-site integration

To further elucidate the mechanism of integration at the repeat, an assay using a minimal linear CRISPR array was established (Figure 3A and 3B). Unlike plasmid integration data, the results of these assays allowed us to distinguish half-site and full-site spacer DNA integrations. When using a linear CRISPR array with only 10 bp of the leader, a single repeat, and a single spacer (Figure 3A), products representing integration at both leader-repeat (LR) and repeat-spacer (RS) borders of the repeat were observed (Figure 3C), and confirmed through high-throughput sequencing (Figure 3D). The same high degree of specificity was observed when using the sequence of the highly preferred non-CRISPR site described above (Figure 1, 2, S3), even though the sequence is only 43% identical to the wildtype (WT) repeat (Figures 3A, 3C, and 3D). This orientation was selected because of the higher repeat sequence conservation, but it is important to note that the alternative orientation also has conserved elements identified to influence integration location (G1, T3, A6, T14, A18, A20, T24, and A29 as well as the C-3). When the sequences of the non-CRISPR substrate that are conserved with the WT repeat were mutated (NC Mut), efficiency was reduced (Figure 3C) and the specificity of

integration at the sites 30 bp apart was abolished to less than 1% of the total reads (Figure 3D).

Next, to determine if an integrated spacer at one of the two repeat junctions would progress to a full-site integration, CRISPR DNA substrates modeled to be half-site DNA intermediates were assayed (Figure 3E-H). Half-site substrates were rapidly transformed into full-site products regardless of which junction of the repeat the spacer started at (Figure 3E and 3F). Disintegration of the half-site spacer was also separately examined since it is expected to reflect the reverse reaction of spacer integration that should be occurring at some level in the reactions (50). Half-site spacer disintegration was observed at lower levels (Figure 3G and 3H) than full-site conversion (Figure 3E and 3F) showing that most, if not all, products observed with the half-site substrates resulted in full-site integration rather than disintegration followed by reintegration at the second site. Additionally, disintegration did not require the presence of Cas2 (Figure S7C and S7D), which was needed for the progression to full-site (Figure S7A and S7B). The results show that *P. furiosus* Cas1 and Cas2 execute accurate second-site integrations by attacking the proper top or bottom strand leader-repeat or repeat-spacer junctions.

We next addressed if *P. furiosus* Cas1 and Cas2 catalyze sequential, two-step transesterification reactions and if there is a preference for first-site integration at the leader-repeat junction vs. the repeat-spacer junction (Figure 4). This was investigated using CRISPR substrates with a DNA hairpin structure at either the leader end or the spacer end, which allowed full-site products to be distinguished from the two possible half-site products on the basis of the size of fragments separated by gel electrophoresis. Both hairpin substrates resulted in rapid, full-site integration but half-site products at both

the leader-repeat and repeat-spacer junctions were also observed as expected (Figure 4A and 4B). In order to confirm that the full-site product was indeed a result of the two-step integration event from a single spacer rather than two separate half-site integration events by two independent spacers, a dideoxy group was substituted for the 3' hydroxyl on one strand or the other of the pre-spacer. With the lack of second 3' hydroxyl group required for the sequential second-site integration, reactions were halted after the first-site integration event, verifying that the full-site product was not produced by independent half-site integrations by two separate spacers. Additionally, with the dideoxy pre-spacers, the LR and RS half-site products were observed at approximately equal levels, suggesting no preference for one border of the repeat over the other for the first transesterification attack (Figure 4C). When the dideoxy was present on both strands (no 3' hydroxyl groups), no integration was observed, as expected. Furthermore, full-site integration was also observed with the highly preferred, non-CRISPR plasmid sequence (Figure 4D) described above (Figures 1, 2, S3, 3). In addition to versatility for the first-site of integration, no preference for orientation was observed for the newly integrated spacer, suggesting that pre-spacer internal sequences do not guide orientation to ensure functional crRNA-mediated invader defense (Figure S8 and (12)).

Together these results indicate that full-site spacer integration reactions facilitated by *P. furiosus* Cas1 and Cas2 occur approximately equally well from either direction. Additionally, the abundant full-site integration products observed with the CRISPR and non-CRISPR sequences (Figure 4) suggests that many or all of the paired integration sites on the plasmids (Figure 1C-E) are full-site integration events. Finally, the large leader and region upstream of the non-CRISPR sequence can be shortened to 10 bp while

maintaining accuracy of spacer integration (Figure 3A-D). However, a mutation to repeat-conserved nucleotides of the non-CRISPR substrate abolished this specificity (Figures 3A-D), supporting the notion that Cas1 and Cas2 are guided by specific sequence elements described above (Figure 2).

Sequences within the repeat guide integration to occur at a distance of 30 bp

The observation that spacers were selectively integrated at CRISPR repeat and repeat-like sequences alike led us to further investigate the characteristics of the repeat sequence that are directing Cas1 and Cas2-facilitated spacer integration. We introduced a series of mutations to the minimal CRISPR DNA substrates (Figure 5A-B) and examined the effects of each mutation on spacer integration efficiency (level of integration observed on a gel with autoradiography relative to the WT repeat) and specificity (location of integration determined and quantified through sequencing). The first (G1 of the top strand) and last (C30 of the bottom strand) nucleotides of the wildtype repeat are the sites for transesterification attacks during integration, and when both of these nucleotides were mutated to adenine or thymine (G1A, C30T and G1T, C30A mutants), the efficiency and specificity were both reduced severely, while a mutation to cytosine and guanine (G1C, C30G mutant) did not significantly alter the reaction (Figure 5C and 5F). Individual mutations of the first or last nucleotide did not have the same severe effect as the simultaneous mutations, except in the case of G1T mutation which was detrimental to both integration efficiency and specificity (Figure S9). Thus, the identity of the base at the sites of transesterification attack on the CRISPR repeat is an important component for specifying integration at a CRISPR repeat by the Cas1 and Cas2 proteins.

DNA base substitution block mutations throughout the minimal CRISPR DNA substrates were used to detect additional elements that coordinate accurate and specific spacer integration (Figure 5). Mutation of the 10 bp leader (B1 mutant) resulted in a moderate reduction in integration efficiency and specificity, while a mutation to the eight nucleotides spanning the leader-repeat junction (B2 mutant) had a detrimental effect on both efficiency and specificity of integration (Figure 5D and 5G). Even with an intact leader sequence, a mutation to the first four nucleotides of the repeat (B3 mutant) still caused major disruption, especially to the leader-repeat integration site which was reduced to less than 2% of total integration reads (Figure 5G). In contrast, if the last four nucleotides of the repeat were mutated (B8 mutant), there was no significant effect on either the efficiency or specificity of integration relative to wild type (Figure 5D and 5G). However, if both the first and last four nucleotides were mutated simultaneously (B10 mutant), the specificity at both junctions was drastically impacted, more so than B3 mutation alone (Figure 5D and 5G). Mutations to the inverted repeats (B4 and B7 mutants) did not have a significant effect, and if they were mutated together (B9 mutant), efficiency was somewhat reduced but specificity remained largely intact excluding a critical role for any potential DNA cruciform structure in directing integration (Figure 5D and 5G). In the case of a mid-repeat mutation (B5 mutant), integration at the repeat-spacer junction was inefficient but remained highly specific to the correct sites of integration. However, a more 3' mid-repeat mutation (B6 mutant), which appeared to have integration efficiency similar to wild type levels, severely disrupted the specificity of integration (Figure 5D and 5G). This demonstrated that integration products that appear to be the expected length on the denaturing polyacrylamide gel could actually be

off-target integration products. To confirm this, we changed the length of the minimal CRISPR DNA substrate so that off-target products would be a length that was distinct from the leader-repeat and repeat-spacer products (Figure S10). We noticed that the sequence substituted in the B6 mutant (5'-ACCCCTC-3') had partial sequence overlap with a sequence (5'-CCCCT-3') found in the leader immediately upstream (-1 to -5) of the first repeat that we considered might be responsible for the altered integration specificity. However, this possibility was ruled out when the same dramatic loss on spacer integration fidelity with a B6 mutant was observed by replacement with a different sequence with no similarity to sequences within the leader (B6b; 5'-GTTTTCT-3') (Figure S10). It is clear that this adenine-rich section of the internal CRISPR repeat (5'-CAAAAGA-3' at position 16-22 of the repeat) is contributing significantly to proper spacer integration.

The finding that the vast majority of integration events on pCR7 were at a distance of 30 bases (Figure 2B) from one another suggested that the size of the repeat is a consequence of the proper spacing of the important sequence elements specifying integration sites. To test this, three nucleotides were either inserted or deleted from the center of the repeat in a region that can withstand base substitutions (B5 mutant). Expanding the spacing between sequence elements within the repeat through base insertion (Ins mutant) caused an apparent reduction in efficiency with a clear disruption in integration specificity at both borders of the repeat (Figure 5E and 5H). Similarly, reducing the spacing between sequence elements within the repeat by base deletion (Del mutant) did not dramatically reduce overall efficiency of integration but did significantly impair the specificity of integration at both borders of the repeat (Figure 5E and 5H).

The drastic effects of these distance-altering mutations on integration fidelity suggest that proper spacing of the important repeat elements is crucial for accurate spacer integration.

Taken together, the results of the block mutations and insertions and deletions (Figure 5D, E, G, H) revealed that sequences immediately adjacent to the leader-repeat junction (5'-CCCT|GTTA-3' at positions -4 through +4) as well as in an adenine-rich internal-repeat region (5'-CAAAAGA-3' at positions 16-22 of the repeat) are particularly critical elements for accurate spacer integration at the LR and RS borders of the repeat. Paired integration sites on opposite strands of the CRISPR DNA, at a set distance of 30 bp, relies heavily on these two motifs. These results are in agreement with the size and sequence preferences of the non-CRISPR plasmid integration sites (Figure 2 and 3).

Discussion

The ability to accurately integrate new spacer DNA sequences into a CRISPR repeat in a site-specific manner is critical for generating functional CRISPR RNAs that endow prokaryotic organisms with heritable immunity against viruses and other potentially detrimental mobile genetic elements. Our work provides the first *in vitro* characterization of spacer integration for the hyperthermophilic archaeon *Pyrococcus furiosus*, and importantly sheds light on the mechanism by which Cas1 and Cas2 are directed to integrate spacers into CRISPR repeats. Our results are consistent with a model (Figure 6) in which *P. furiosus* Cas1 and Cas2 are guided to perform properly spaced two-site integration at the leader-proximal repeat by certain sequences, most notably motifs at the leader-repeat junction and central region of the repeat. Proper ~30 bp spacing of integration sites (which spans three helical turns of double-stranded DNA) is likely

maintained by the architecture of the presumed *P. furiosus* Cas1-Cas2 integrase complex, while the fidelity of integration is governed by sequence preferences and ultimately, by the overall structural properties of the repeat. Furthermore, integration at both junctions of the CRISPR repeat relies on the proper spacing of sequences throughout the DNA, allowing the two-site, full spacer integration to occur without a strict order, potentially even simultaneously.

Our study yielded valuable insight into CRISPR elements governing spacer integration by analyzing effects on spacer integration efficiency and specificity for a wide array of leader and repeat mutations (Figure 5) as well as thorough careful inspection of off-target DNA integration sites (Figure 2). By sequencing integration products to confirm spacer positioning, we were able to more fully interpret the effects of DNA mutations. Some products that were the expected size actually represented off-target integration sites (Blocks 2, 3 and 6 of Figure 5), and some mutations that caused reduced efficiency still maintained a preference for LR and RS sites (G1A-C30T, Blocks 1, 5, and 9 of Figure 5). Thus, analyzing *in vitro* integration products through gel electrophoresis to address efficiency, combined with sequencing to address specificity, provided a more comprehensive approach to studying integration than gel analysis alone. In addition, given the rich information gleaned from inspection of off-target DNA integration sites, we conclude that off-target events can provide insight into key, context-independent elements that are recognized by proteins during adaptation. Although off-target integration poses potential deleterious effects *in vivo* (6), evidence suggests that non-canonical spacer integration might serve in the spontaneous generation of CRISPR loci (51). The specific DNA sequences we identified from these approaches (Figure 2 and 5;

represented in our model, Figure 6), suggest that points distributed throughout the repeat collectively contribute to accurate spacer integration. Block DNA mutations (Figure 5) determined that some of these elements hold a greater importance than others, including two essential motifs at the leader-repeat junction and central region of the repeat, which are discussed in greater detail below.

Leader-repeat motif

Based on our findings (Figure 2 and 5), the motif at the leader-repeat junction is defined by GTT at positions 1-3 of the repeat and a C at the -3 position of the leader (Figure 6). The preference for guanine at sites of integration has been previously observed for Cas1 and Cas2 in I-A (36,50), I-B (46), I-E (34,50), II-A (6,40), and V-C systems (52), further implicating G1 as a key feature for directing integration to the leader-repeat junction. We predict that Cas1 directly interacts with this motif to guide integration at G1, which agrees with X-ray crystallographic evidence revealing base-specific DNA contacts of Cas1 at positions T1 and T2 of the repeat and -1 through -4 of the leader in a type II-A system (40). The importance of the C at position -3 is further supported by its conservation across all group 1 leader sequences of the Pyrococcales, with the exception of CRISPR locus 8 of *P. furiosus*, which has low activity compared to the other active loci (23,53). Similarly, a WebLogo of integration sites for the archaeal type I-A system of *S. solfataricus* revealed a preference for a C at the -2 position (36). Our identification of an important leader-repeat motif for the *P. furiosus* CRISPR-Cas system is in agreement with several studies that report the sequence spanning the LR junction to be important for spacer integration examined *in vivo* (7,37,41,42,46,54) and *in vitro*

(32,36,40). This important motif is unique to the first repeat, and it may play a role in making the leader-proximal repeat the preferred repeat in the CRISPR array *in vivo*.

Adenine-rich central motif

The centrally located repeat motif identified in this study is defined as an adenine-rich sequence located at positions 16-22 of the repeat (Figure 6). The adenine bases at positions 17-18 likely contribute significantly to this motif, as they were particularly conserved in the non-CRISPR sites of integration (Figure 2). We predict this motif plays an upstream binding role to orient Cas1 and Cas2 in the correct position for integration at the two borders of the repeat, since we found that mutations to this motif (Block 6) significantly disrupted both LR and RS integration sites, as opposed to just one or the other (Figure 5 and S10). Our model is supported by other *in vitro* studies, which demonstrated motifs in the mid-repeat to be important for Cas1+Cas2 binding to the repeat (33) and completion from half-site to full-site integration (35). *In vivo* studies have also implicated internal motifs as Cas1 and Cas2 docking sites that coordinate the proper spacing of the paired integration sites (45,46). While the internal motifs identified across different systems vary in location and attributes, they could all be playing a common role to recruit and properly orient Cas1 and Cas2 for accurately spaced integration.

Unique properties of *P. furiosus* Cas1 and Cas2

Prior work characterizing the integration properties of Cas1 and Cas2 from different organisms and distinct CRISPR types, have revealed variability in mechanisms

governing: 1) the order of spacer integration at LR and RS sites, and 2) how integration is directed to the leader-proximal repeat vs. downstream repeats. There is *in vitro* evidence suggesting that Cas1 and Cas2 from type I-A (36,50), type I-E (32) and type II-A (6,40) systems attack at the leader-repeat junction first followed by integration at the repeat-spacer junction. The affinity of Cas1 and Cas2 for the sequence spanning the leader-repeat junction reported *in vitro* (32,36,40,50) is presumably what causes this to be the first integration site. In contrast, our results with *P. furiosus* Cas1 and Cas2 show that full-site integration occurs by an unordered or simultaneous attack at both borders of the repeat (Figure 3 and 4). This difference could be explained by the relative importance of different motifs in the repeat: in systems where Cas1 and Cas2 primarily recognize the leader-repeat sequence, there is sequential integration. In *P. furiosus*, Cas1 and Cas2 have a much wider recognition zone, spanning sequence elements spread throughout the entire repeat (Figure 2), leading to simultaneous or unordered integration events. Additionally, some systems that do demonstrate sequential integration also report a stable Cas1-Cas2 complex (6,32). While our work shows a clear requirement for both Cas1 and Cas2 for efficient, full-site integration (Figure 3, 4, S7), *P. furiosus* Cas1 and Cas2 do not form a stable complex *in vitro* (our unpublished results) and instead could be brought together by interacting with the spacer and/or CRISPR repeat DNA elements.

In vivo, new spacers are almost exclusively integrated at the leader-proximal (first) repeat as opposed to downstream repeats of a CRISPR array (1,41,42). Two distinct mechanisms (factor-dependent and factor-independent) have been proposed to explain how various Cas1 and Cas2 proteins are normally guided to integrate selectively at the leader-proximal (first) repeat (6,32,35,36,40,43). For type I-E and I-F systems, this

is only achieved in the presence of Integration Host Factor (IHF), which binds to specific elements in the leader and interacts with Cas1 and Cas2 to recruit the proteins to the first repeat (32,35,43). Similarly, the I-A system in *S. sulfataricus* required the presence of an unidentified host factor from the cell lysate to achieve specific integration at the first repeat (36). In contrast, Cas1 and Cas2 of II-A systems display an intrinsic preference for the first repeat (without additional factors) due to strong affinity for, and reliance on, the sequence surrounding the LR junction (6,40). We found that *P. furiosus* Cas1 and Cas2 integrate at each repeat of a CRISPR array *in vitro* (Figure 1) similar to what is observed for I-E and I-F systems, which integrate at every repeat in the absence of IHF (32,34,43). This is an indication that a yet unidentified factor in *P. furiosus* cells (there are no known close IHF homologs) may be responsible for directing the proteins to the leader-proximal repeat.

Functional integration

In vivo, *P. furiosus* integrates new spacers at the leader-proximal repeat with an orientation that results in crRNA that is complementary to the PAM-containing strand of the invading DNA and is functional for immunity (12). Since we did not observe orientation biases in our *in vitro* results, we predict that additional factors beyond Cas1 and Cas2 are involved in regulating a strict order to integration and in recognizing the leader to direct spacer integration to the first repeat. It is likely that a larger portion of the leader is important for maintaining specific integration, as was the case for the large leader of the archaeal thermophile, *S. sulfataricus* *in vitro* (36). Given that Cas4-1 and Cas4-2 are required for proper spacer orientation *in vivo* (12), it is possible that these two

proteins play a role in directing the order of Cas1-directed transesterifications, perhaps via interactions between Cas4 proteins and Cas1 and/ or the leader-repeat motif.

Considerable effort was directed at testing these hypotheses *in vitro*, but we were unable to identify the proper conditions for Cas4-1 and Cas4-2 activity (our unpublished data).

Further genetic and biochemical analysis will be required to elucidate the factor(s) involved in targeting integration to the first repeat and to determine if the order of integration sites is regulated.

We have established the importance of distinct elements within the CRISPR DNA array for spacer integration, but how do these elements give rise to function? We suggest in our model that the properties of the sequence elements and their relative spacing collectively influence direct interaction with the Cas1-Cas2 proteins, as well as enable the flexibility of the repeat. Together these factors complement the architecture of the bound Cas1-Cas2 proteins to achieve properly spaced integration sites. This model is supported by the apparently functionally important bending of the repeat observed in the structures of Cas1-Cas2 in half-site and full-site integration conformations for I-E (35) and II-A systems (40). Adenine-tracts are known to confer intrinsic DNA bending properties (55,56). The adenine-richness of the identified key functional elements within the *P. furiosus* repeat (summarized in Figure 6), may provide the needed repeat DNA malleability for serving as target DNA for spacer integration. In conclusion, we demonstrated that *P. furiosus* Cas1 and Cas2 have an intrinsic specificity for the sequences and relative spacing of the CRISPR repeats in order to facilitate accurate and full-site integration. Future studies will address how these intrinsic activities of Cas1 and

Cas2 are refined to ensure that integration is correctly oriented and localized to the leader-proximal repeat, as is observed *in vivo*.

References

1. 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.
2. Mojica, F.J., Diez-Villasenor, C., Garcia-Martinez, J. and Soria, E. (2005) Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol*, **60**, 174-182.
3. Brouns, S.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J., Snijders, A.P., 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.
4. 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.
5. Sternberg, S.H., Richter, H., Charpentier, E. and Qimron, U. (2016) Adaptation in CRISPR-Cas Systems. *Mol Cell*, **61**, 797-808.
6. Wright, A.V. and Doudna, J.A. (2016) Protecting genome integrity during CRISPR immune adaptation. *Nat Struct Mol Biol*, **23**, 876-883.

7. Yosef, I., Goren, M.G. and Qimron, U. (2012) Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res*, **40**, 5569-5576.
8. McGinn, J. and Marraffini, L.A. (2019) Molecular mechanisms of CRISPR-Cas spacer acquisition. *Nat Rev Microbiol*, **17**, 7-12.
9. Deveau, H., Barrangou, R., Garneau, J.E., Labonte, J., Fremaux, C., Boyaval, P., Romero, D.A., Horvath, P. and Moineau, S. (2008) Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J Bacteriol*, **190**, 1390-1400.
10. Shah, S.A., Erdmann, S., Mojica, F.J. and Garrett, R.A. (2013) Protospacer recognition motifs: mixed identities and functional diversity. *RNA biology*, **10**, 891-899.
11. 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.
12. Shiimori, M., Garrett, S.C., Graveley, B.R. and Terns, M.P. (2018) Cas4 Nucleases Define the PAM, Length, and Orientation of DNA Fragments Integrated at CRISPR Loci. *Mol Cell*, **70**, 814-824 e816.
13. Grissa, I., Vergnaud, G. and Pourcel, C. (2007) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics*, **8**, 172.
14. 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.

15. 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 Microbiol Rev*, **39**, 428-441.
16. Jackson, R.N., van Erp, P.B., Sternberg, S.H. and Wiedenheft, B. (2017) Conformational regulation of CRISPR-associated nucleases. *Curr Opin Microbiol*, **37**, 110-119.
17. Marraffini, L.A. (2015) CRISPR-Cas immunity in prokaryotes. *Nature*, **526**, 55-61.
18. 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.
19. Hille, F., Richter, H., Wong, S.P., Bratovic, M., Ressel, S. and Charpentier, E. (2018) The Biology of CRISPR-Cas: Backward and Forward. *Cell*, **172**, 1239-1259.
20. Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J., Barrangou, R., Brouns, S.J., Charpentier, E., Haft, D.H. *et al.* (2015) An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol*, **13**, 722-736.
21. Koonin, E.V., Makarova, K.S. and Zhang, F. (2017) Diversity, classification and evolution of CRISPR-Cas systems. *Curr Opin Microbiol*, **37**, 67-78.
22. Alkhnbashi, O.S., Shah, S.A., Garrett, R.A., Saunders, S.J., Costa, F. and Backofen, R. (2016) Characterizing leader sequences of CRISPR loci. *Bioinformatics*, **32**, i576-i585.

23. Shiimori, M., Garrett, S.C., Chambers, D.P., Glover, C.V.C., 3rd, Graveley, B.R. and Terns, M.P. (2017) Role of free DNA ends and protospacer adjacent motifs for CRISPR DNA uptake in *Pyrococcus furiosus*. *Nucleic Acids Res*, **45**, 11281-11294.
24. Majumdar, S., Ligon, M., Skinner, W.C., Terns, R.M. and Terns, M.P. (2017) Target DNA recognition and cleavage by a reconstituted Type I-G CRISPR-Cas immune effector complex. *Extremophiles*, **21**, 95-107.
25. 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.
26. Hale, C.R., Cocozaki, A., Li, H., Terns, R.M. and Terns, M.P. (2014) Target RNA capture and cleavage by the Cmr type III-B CRISPR-Cas effector complex. *Genes Dev*, **28**, 2432-2443.
27. Elmore, J., Deighan, T., Westpheling, J., Terns, R.M. and Terns, M.P. (2015) DNA targeting by the type I-G and type I-A CRISPR-Cas systems of *Pyrococcus furiosus*. *Nucleic Acids Res*, **43**, 10353-10363.
28. Majumdar, S. and Terns, M.P. (2019) CRISPR RNA-guided DNA cleavage by reconstituted Type I-A immune effector complexes. *Extremophiles*, **23**, 19-33.
29. 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*. *J Bacteriol*, **194**, 2491-2500.

30. Lee, H., Zhou, Y., Taylor, D.W. and Sashital, D.G. (2018) Cas4-Dependent Prespacer Processing Ensures High-Fidelity Programming of CRISPR Arrays. *Mol Cell*, **70**, 48-59 e45.
31. Lee, H., Dhingra, Y. and Sashital, D.G. (2019) The Cas4-Cas1-Cas2 complex mediates precise prespacer processing during CRISPR adaptation. *eLife*, **8**, e44248.
32. Nunez, J.K., Bai, L., Harrington, L.B., Hinder, T.L. and Doudna, J.A. (2016) CRISPR Immunological Memory Requires a Host Factor for Specificity. *Mol Cell*, **62**, 824-833.
33. 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 Res*, **45**, 2714-2723.
34. Nunez, J.K., Lee, A.S., Engelman, A. and Doudna, J.A. (2015) Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature*, **519**, 193-198.
35. 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**, 1113-1118.
36. 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 Res*, **46**, 1007-1020.

37. 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 Res*, **42**, 7884-7893.
38. Jackson, S.A., McKenzie, R.E., Fagerlund, R.D., Kieper, S.N., Fineran, P.C. and Brouns, S.J. (2017) CRISPR-Cas: Adapting to change. *Science*, **356**, eaal5056.
39. Nunez, 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.
40. 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.
41. Wei, Y., Chesne, M.T., Terns, R.M. and Terns, M.P. (2015) Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in *Streptococcus thermophilus*. *Nucleic Acids Res*, **43**, 1749-1758.
42. McGinn, J. and Marraffini, L.A. (2016) CRISPR-Cas Systems Optimize Their Immune Response by Specifying the Site of Spacer Integration. *Mol Cell*, **64**, 616-623.
43. 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. *Proc Natl Acad Sci U S A*, **114**, E5122-E5128.
44. Yoganand, K.N., 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 Res, **45**, 367-381.

45. 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. *Cell Rep*, **16**, 2811-2818.
46. 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 Res*, **44**, 4266-4277.
47. 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.
48. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841-842.
49. Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) WebLogo: a sequence logo generator. *Genome Res*, **14**, 1188-1190.
50. 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.
51. Nivala, J., Shipman, S.L. and Church, G.M. (2018) Spontaneous CRISPR loci generation in vivo by non-canonical spacer integration. *Nat Microbiol*, **3**, 310-318.
52. Wright, A.V., Wang, J.Y., Burstein, D., Harrington, L.B., Paez-Espino, D., Kyrpides, N.C., Iavarone, A.T., Banfield, J.F. and Doudna, J.A. (2019) A

- Functional Mini-Integrase in a Two-Protein-type V-C CRISPR System. *Mol Cell*, **73**, 727-737 e723.
53. Norais, C., Moisan, A., Gaspin, C. and Clouet-d'Orval, B. (2013) Diversity of CRISPR systems in the euryarchaeal Pyrococcales. *RNA biology*, **10**, 659-670.
54. Diez-Villasenor, C., Guzman, N.M., Almendros, C., Garcia-Martinez, 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.
55. Haran, T.E. and Mohanty, U. (2009) The unique structure of A-tracts and intrinsic DNA bending. *Q Rev Biophys*, **42**, 41-81.
56. Rettig, M., Germann, M.W., Wang, S. and Wilson, W.D. (2013) Molecular basis for sequence-dependent induced DNA bending. *Chembiochem*, **14**, 323-331.

Figure 2.1. *P. furiosus* Cas1 and Cas2 facilitate spacer integration *in vitro*.

(A) Schematic of the *in vitro* spacer integration reaction. pCR7 containing a minimal CRISPR array was incubated with *Pyrococcus furiosus* (Pfu) Cas1 and Cas2 purified from *E. coli* and radiolabeled pre-spacers (PS). Integration products include half-site or full-site integration (nicked form of the plasmid). (B) Integration assay comparing pCR7, pCR7 with a deletion of the leader, and pControl lacking the entire leader and CRISPR array. Products separated by agarose gel electrophoresis and detected by EtBr followed by autoradiography. Noncontiguous portions of the same gel are indicated by dashed lines. (C, D, E) Sites of spacer integration were identified by high-throughput sequencing; peak heights in tracks C-G directly reflect the number of reads that showed an integration event at that position in the plasmid. Integration sites along pCR7, pCR7 no leader, and pControl, respectively. Brackets above the tracks highlight sites of integration at the CRISPR and a highly preferred non-CRISPR integration site, which was investigated further in later experiments. (F, G) Increased resolution of integration at CRISPR repeats of pCR7 and pCR7 no leader.

Figure 2.2. Non-CRISPR integration sites resemble CRISPR repeats.

(A) Spacing between maximum integration peaks on plasmid tracks shown in Figure 2.1 were measured within 500 random 50 bp windows. (B, C) The number of times each bp spacing between maximum peaks was counted on pCR7 (B) and pControl (C). (D, E) Percent of total integration sites located at guanine, cytosine, adenine, and thymine on pCR7 (D) and pControl (E). (F) The top ten 30 bp non-CRISPR sites on pCR7. List includes any sequences that had more than 40 reads at either the “LR” or “RS” location of the sequence. Homology is highlighted for each nucleotide that matches the CRISPR repeat. (G) WebLogo of non-CRISPR sequences listed in F. (H, I) WebLogo of all integration events on pCR7 (H) and pControl (I).

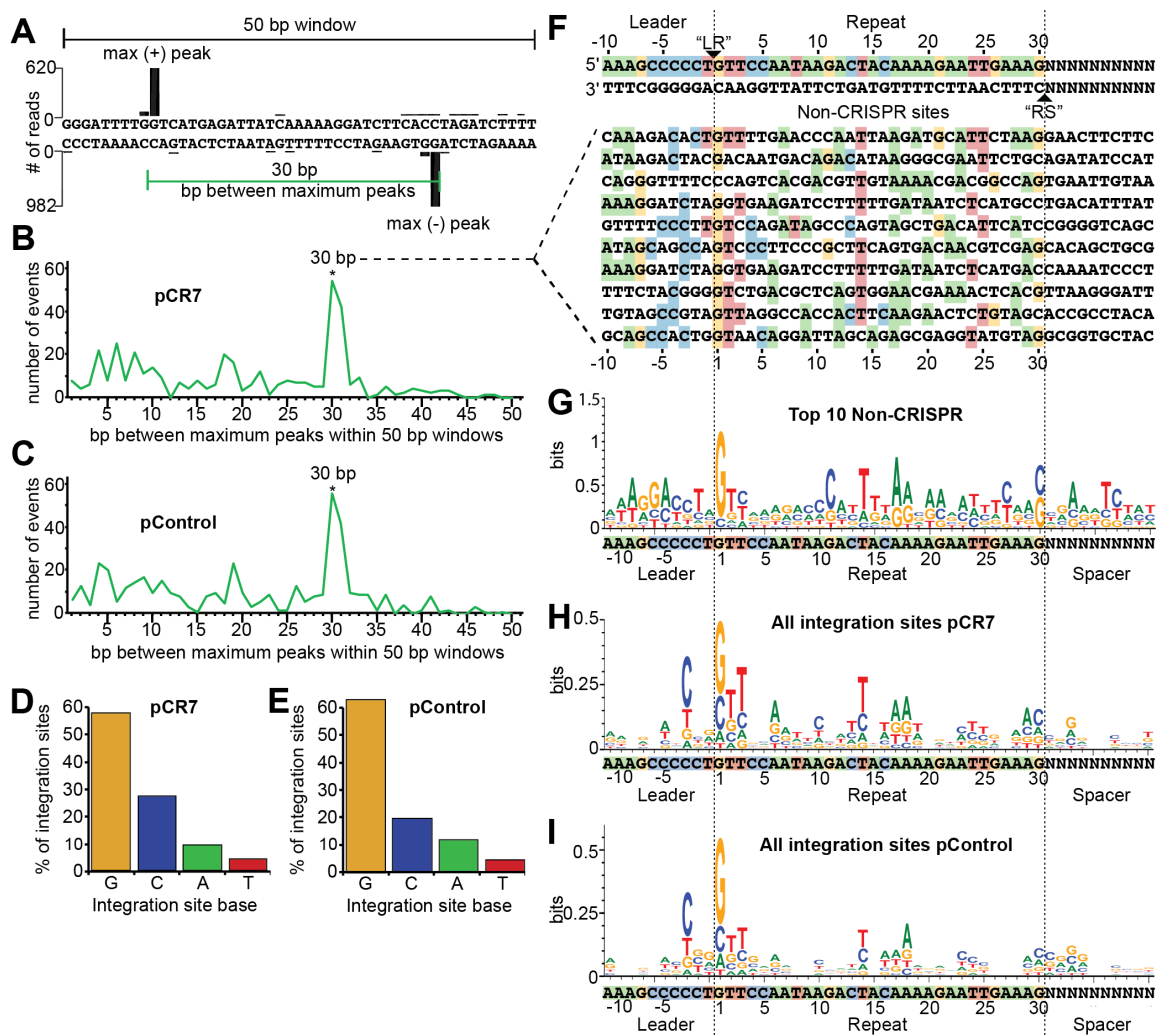


Figure 2.3. Integration into a minimal CRISPR reveals full-site integration.

(A) Annotated sequence of minimal linear CRISPR. LR marks leader-repeat integration site. RS marks repeat-spacer integration site. IR1 and IR2 mark inverted repeats. Below the WT sequence, non-CRISPR (NC) and non-CRISPR mutant (NC Mut) sequences are displayed. Dots above NC sequence indicate homology with WT repeat. (B) Schematic of *in vitro* spacer integration assay with a minimal linear CRISPR containing a single repeat. Potential reaction products include full-site or half-site integration at either the leader-repeat junction (LR) or the repeat-spacer junction (RS). (C) Integration products separated by denaturing PAGE and detected with autoradiography. Noncontiguous lanes from the same gel are indicated by dashed lines. (D) Sequenced integration products displayed as % of reads mapped to sites of integration. (E, F) Continuation of LR half-site (E) or RS half-site (F) substrates to full-site integration detected with radiolabel on bottom strand of the spacer (E) or top strand of leader (F). (G, H) Spacer disintegration from LR (G) or RS (H) half-site substrates detected with radiolabel on spacer.

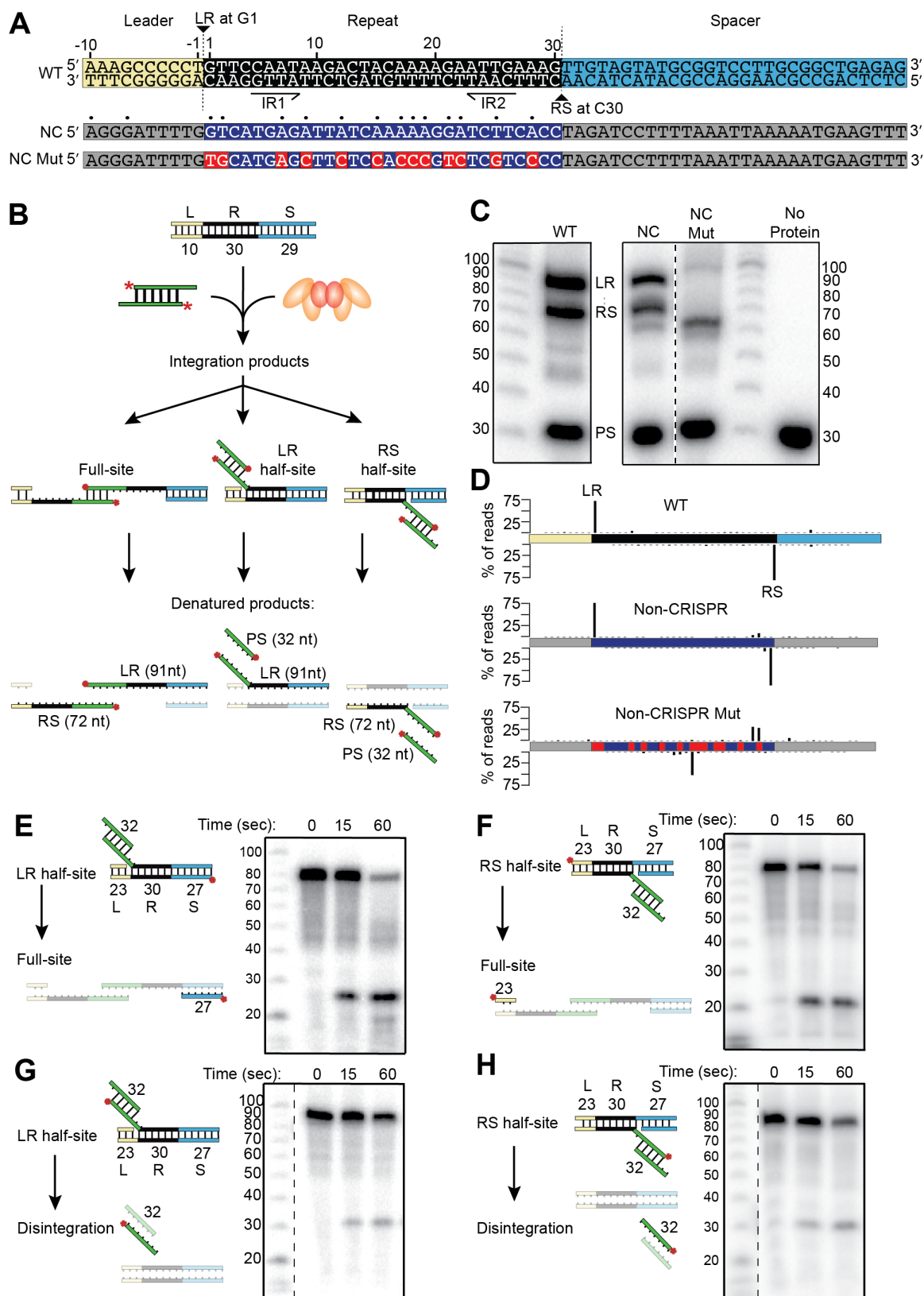


Figure 2.4. Integration into hairpin CRISPRs suggests no preferred order to the two-step transesterification attacks.

(A, B) Full-site integration detected using a hairpin turn in the (A) spacer or (B) leader ends of the minimal CRISPR substrates. (C) Site of first transesterification attacks identified with pre-spacers with 3' dideoxy groups on top, bottom, or both strands. (D) Full-site integration into hairpin substrate containing the non-CRISPR sequence.

Noncontiguous lanes from the same gel are indicated by dashed lines.

Figure 2.5. Integration efficiency and specificity is altered by mutations to the CRISPR repeat.

(A) Annotated sequence of minimal linear CRISPR. LR marks leader-repeat integration site. RS marks repeat-spacer integration site. IR1 and IR2 mark inverted repeats. (B) List of substitution, insertion, and deletion mutations made to the minimal CRISPR. (C, D, E) Integration assays comparing mutations to the first and last nucleotide of the repeat (C), block mutations of the leader and repeat (D), and 3 bp insertions or deletions at the center of the repeat (E). (F, G, H) Integration sites along the minimal CRISPRs. Integration sites mapping to the nucleotide level resolution for all mutations can be found in Figure 2.S11, and a summary of integration specificity at LR and RS junctions of all mutants can be found in Table S2.

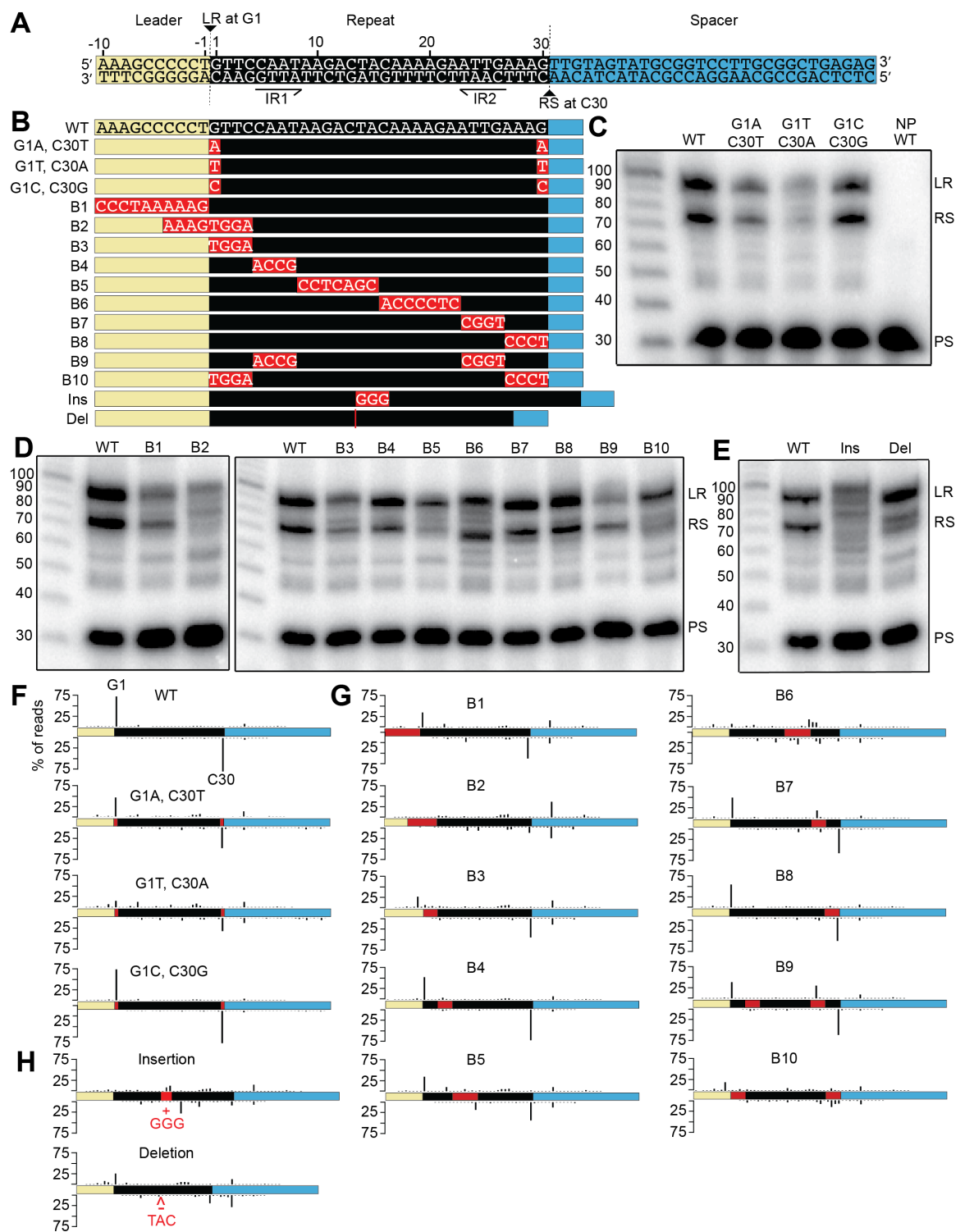


Figure 2.6. *P. furiosus* spacer integration model.

Cas1 (orange) and Cas2 (red) have an affinity for several elements within the repeat (black) and leader (tan) in order to guide accurate spacer (green) integration. The architecture of the protein-spacer-CRISPR ternary complex positions the two-step transesterification attacks 30 bp apart at the preferred G and C bases of the repeat borders. Boxed in purple are the most essential motifs including the sequence spanning the leader-repeat junction and the internal adenine-rich region of nucleotides 16-22 (Block 6). Other nucleotides highlighted throughout the entire repeat collectively contribute to the additive specificity.

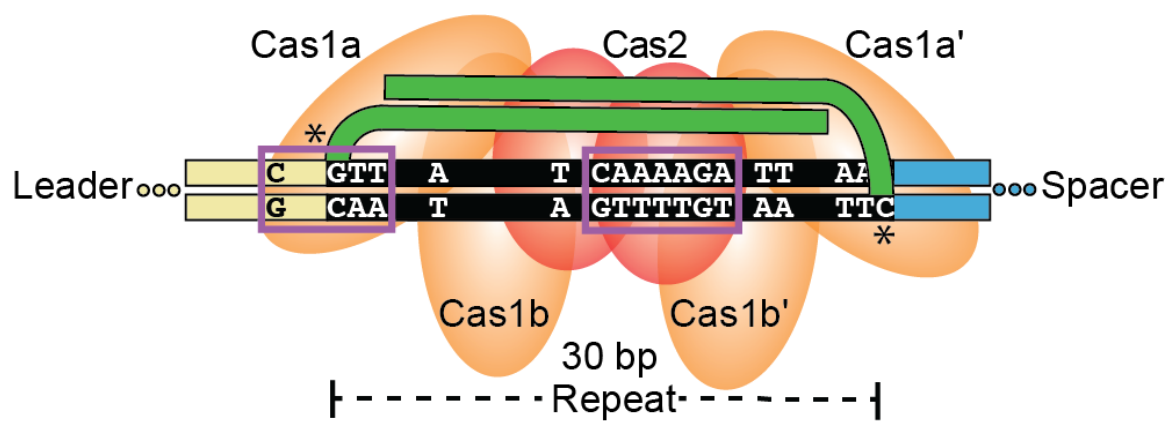


Figure 2.S1. CRISPR adaptation model for *Pyrococcus furiosus*.

Cas4-1 and Cas4-2 generate pre-spacers, which Cas1 and Cas2 then integrate into the CRISPR array by a two-step transesterification reaction at the borders of the leader-proximal repeat. The order of the two integration steps is unknown for this system.

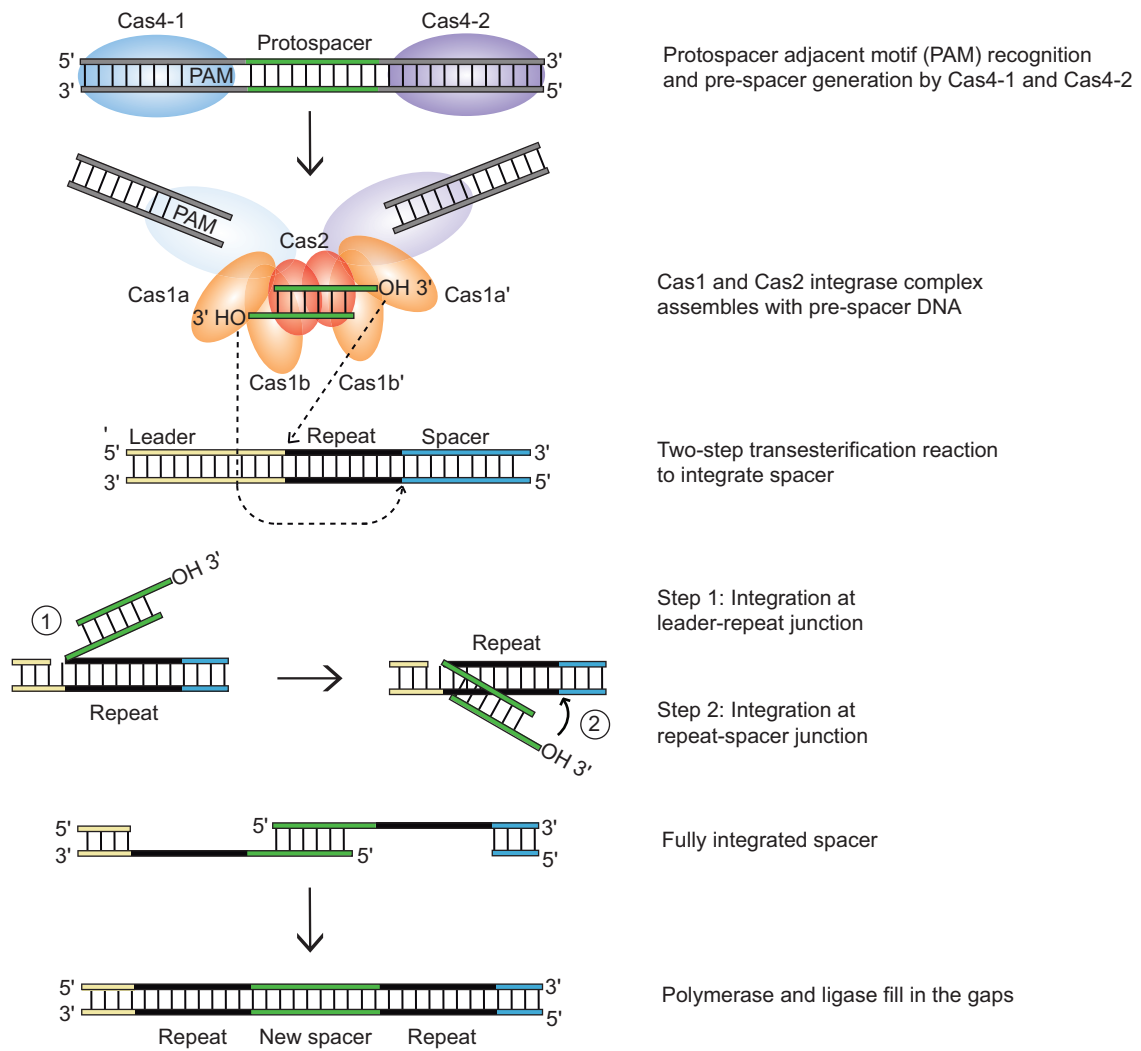


Figure 2.S2. Cas2 is required for robust spacer integration

(A) Integration assay comparing Cas1+2, Cas2 alone, and Cas1 alone. Experimental conditions were the same as described for Figure 2.1. (B) Integration assay comparing pre-spacers with five nucleotide long 3' overhangs and pre-spacers with blunt ends. Linear and circular pCR7 substrates were both tested as the target DNA. Reaction products were separated on 10% denaturing PAGE, followed by autoradiography.

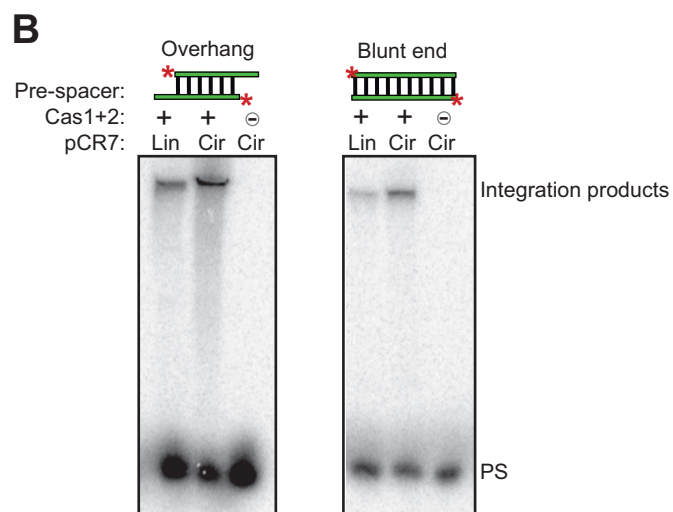
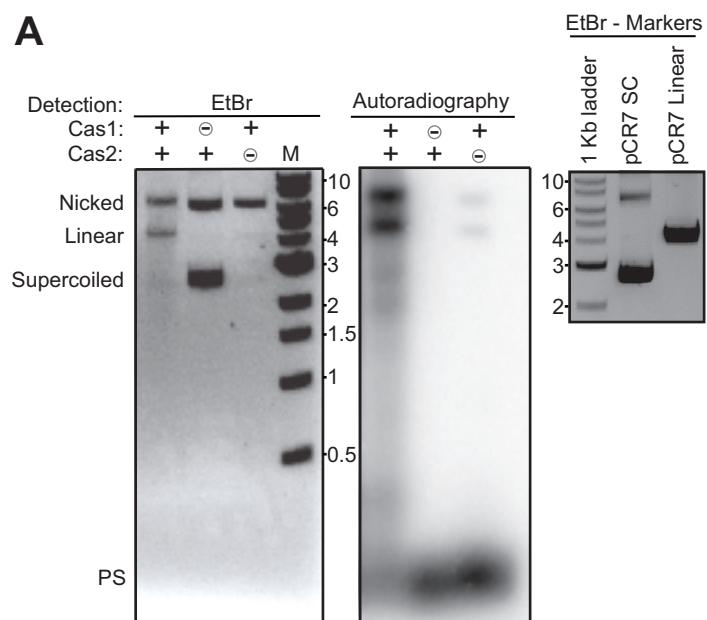


Figure 2.S3. Non-CRISPR integration site

(A) Integration locations on pCR7. Increased resolution of the highly preferred non-CRISPR integration site (boxed in green) displayed below. (B) Integration locations on pControl2, an alternative plasmid with a different backbone from pCR7. Increased resolution of the same highly preferred non-CRISPR integration site (boxed in green) displayed below. pCR7 and pControl2 share a 120 bp conserved region, in which this preferred site is located, while the rest of the plasmid backbones are not conserved.

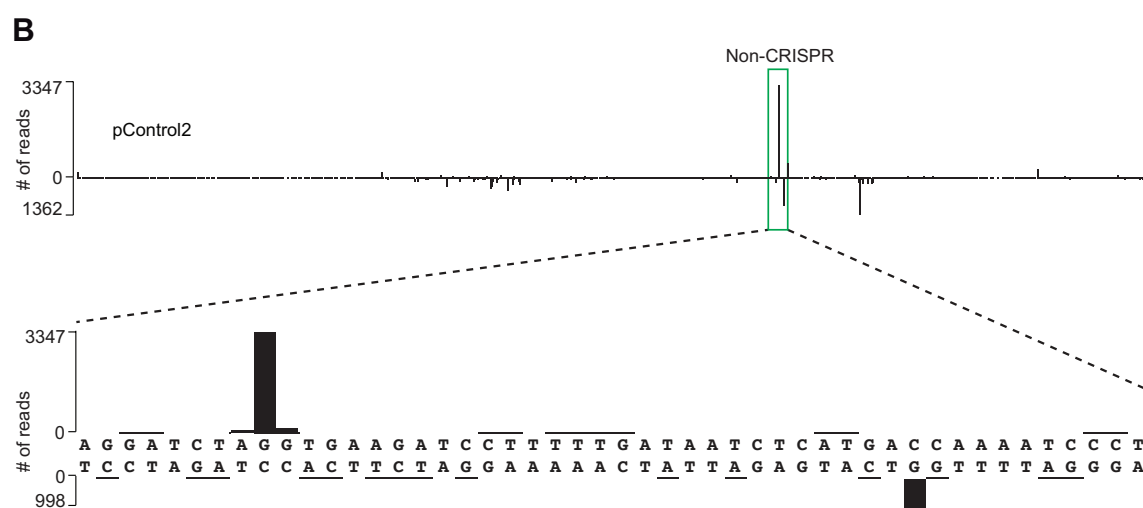
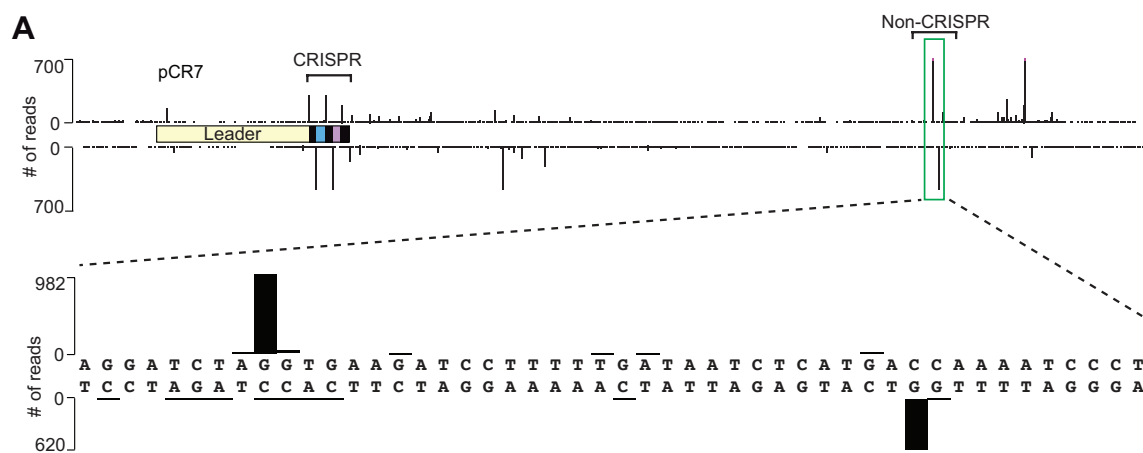


Figure 2.S4. Spacing of integration peaks

(A) Spacing between maximum integration peaks within a 50 bp window was measured across the entire plasmid by making 500 random windows. Distance was measured three ways: Highest plus strand peak to highest minus strand peak (green), two highest plus strand peaks (red), and two highest minus strand peaks (blue). (B) Peaks on opposite strands (green), plus strand (red), and minus strand (blue) were compared by counting the number of times each bp spacing between highest peaks was found in the 500 random windows.

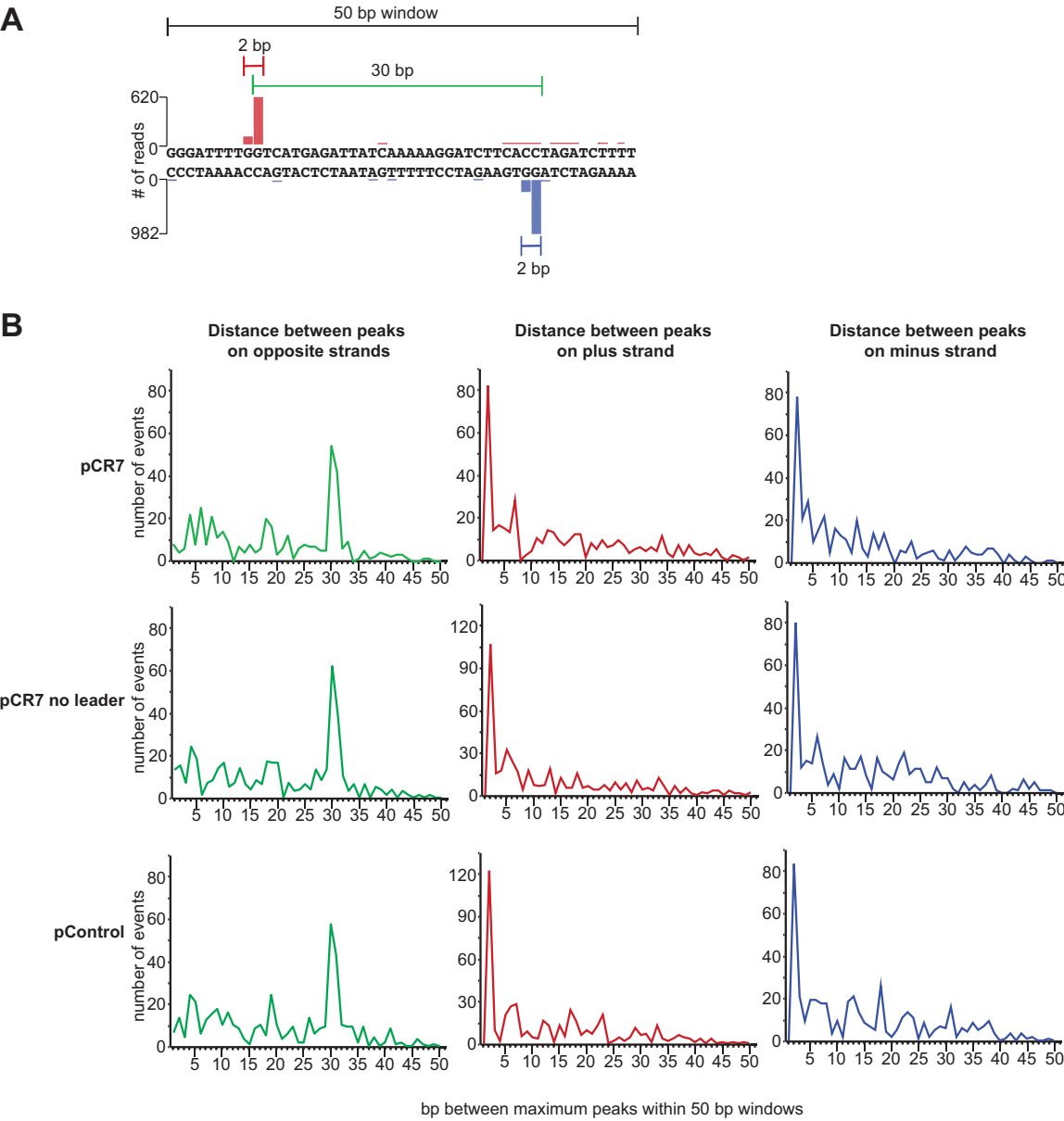


Figure 2.S5. Non-CRISPR integration sites on pCR7

(A) Integration sites along pCR7, with top ten 30 bp non-CRISPR sites marked 1 through 10. Highly integrated sites that were 31 bp are marked A through E. (B, C) Sequences of 30 bp (B) and 31 bp (C) non-CRISPR sites with corresponding numbers and letters mapping to the track above. Homology with the WT repeat is highlighted for each sequence. The 31 bp sequences are displayed in both orientations, in order to show homology with WT repeat from both sites of integration in the +1 position.

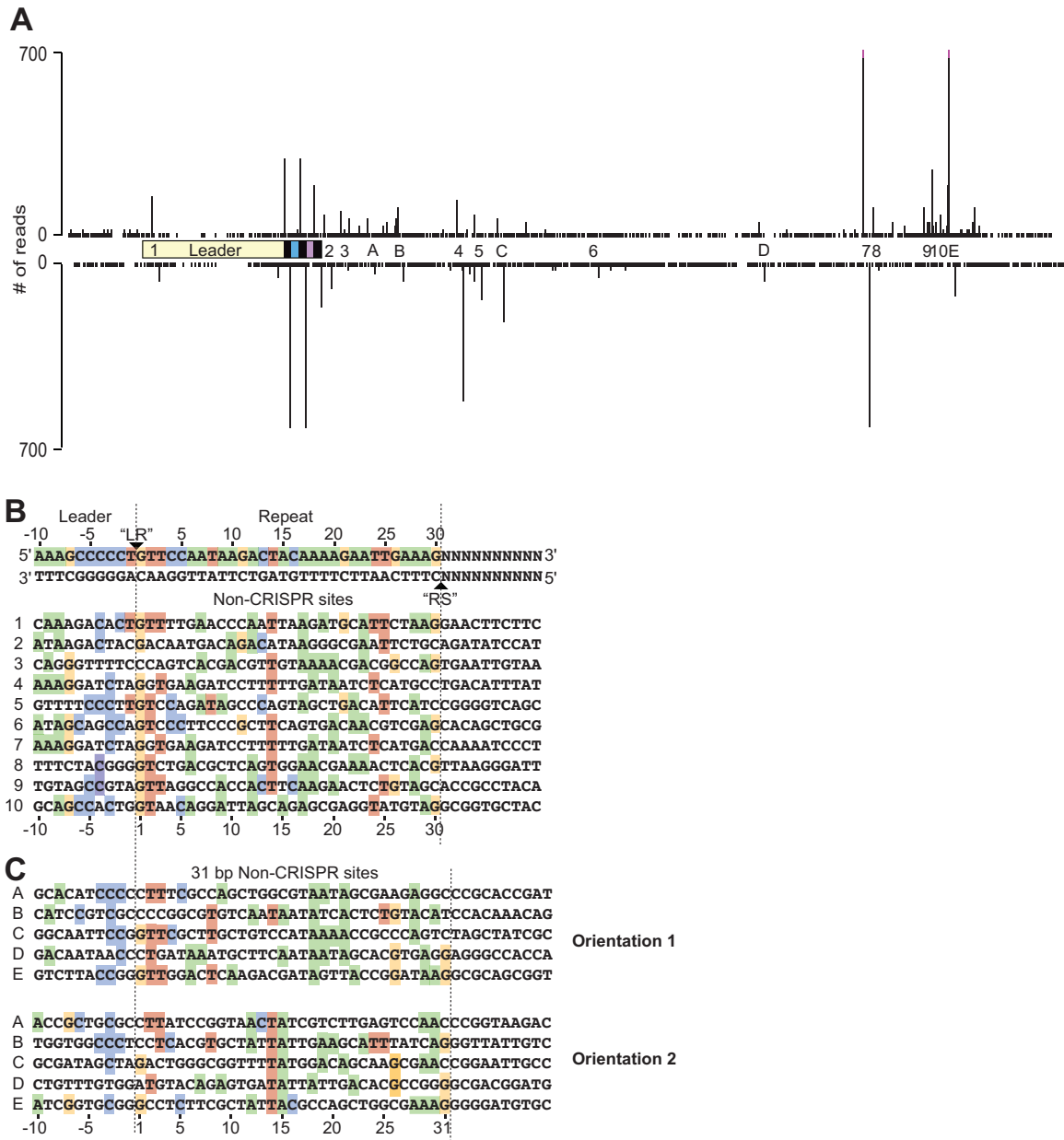


Figure 2.S6. WebLogos and base preferences of all integration sites

(A) WebLogos of all integration locations. (B) Percent of total or unique integration sites that were located at guanine, cytosine, adenine, and thymine. Many positions in the plasmid had multiple reads showing spacer integration at that point; for “total reads” the base at the integration point was counted once for each supporting read. For “unique reads” the base at a distinct position was counted only once if one or more reads showed integration at that point.

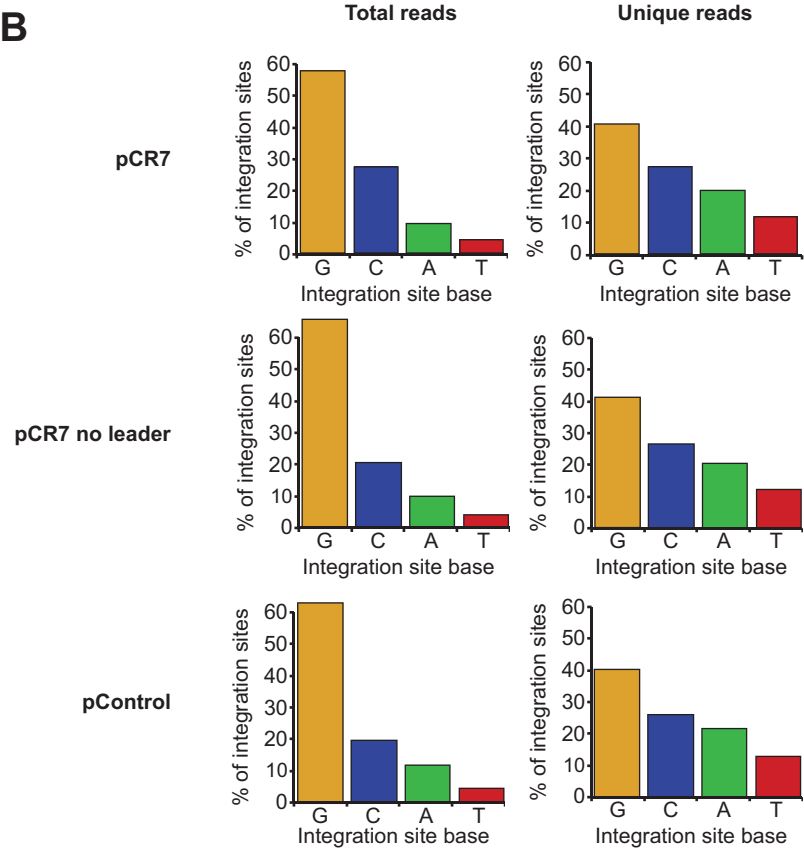
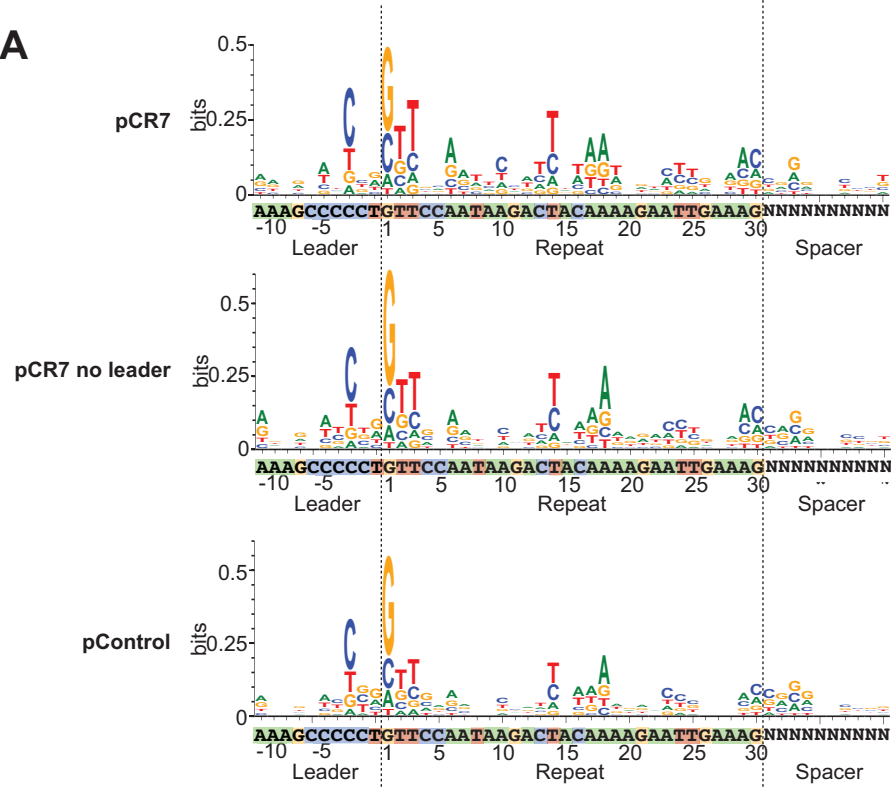


Figure 2.S7. Full-site integration occurs with Cas1 and Cas2, while disintegration is achieved by Cas1 alone.

(A, B) Continuation of LR half-site (A) or RS half-site (B) substrates to full-site integration detected with radiolabel on bottom strand of the spacer (A) or top strand of leader (B). Reactions contain no protein, Cas1, or Cas1 and Cas2. **(C, D)** Spacer disintegration from LR (C) or RS (D) half-site substrates detected with radiolabel on spacer. Reactions contain no protein, Cas1, or Cas1 and Cas2.

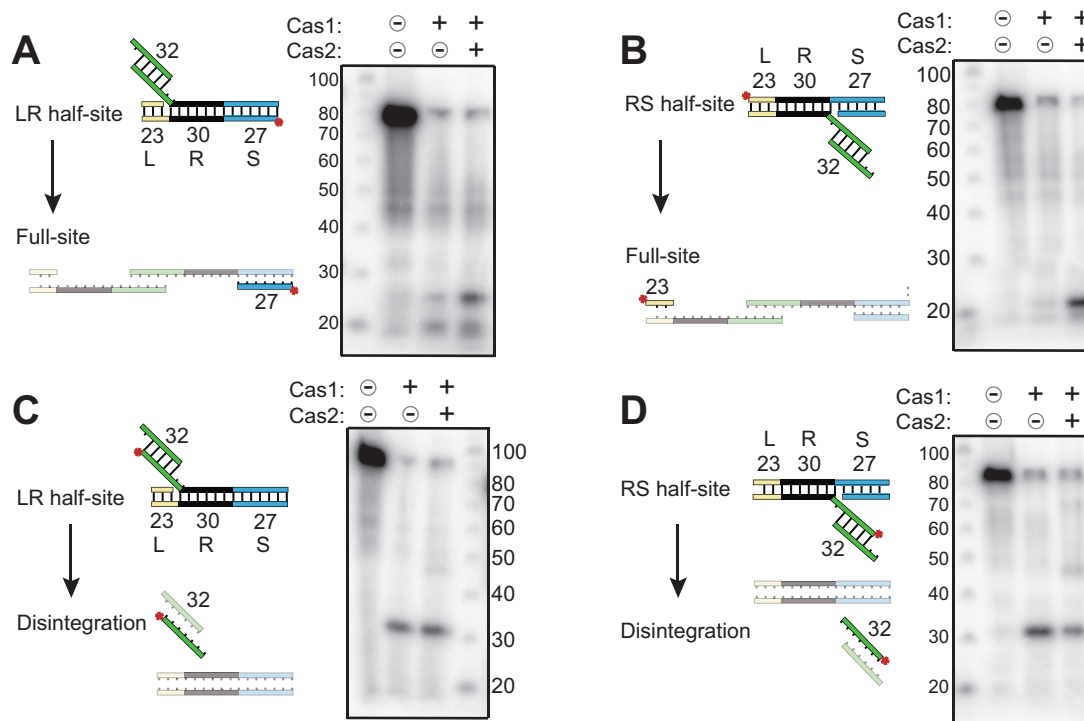


Figure 2.S8. PCR of integration sites determine pre-spacers are integrated in both correct and inverse orientations.

(A) Spacer integration assays were performed with either CRISPR plasmid (pCR) or minimal linear CRISPRs (LRS). Integration products were used as templates for the following PCR reactions: 1) Spacer with correct orientation integrated at LR, 2) Spacer with correct orientation integrated at RS, 3) Spacer with inverse orientation integrated at LR, 4) Spacer with inverse orientation integrated at RS. Plasmids containing correct (C1) or inverse (C2) orientation spacers were used as positive PCR controls, and pCR with no new spacer (C3) was used in negative PCR controls. (B) The PCR amplified integration products were separated on a 1% agarose gel in 1x TAE. Expected PCR products are 125 bp for LR integration and 162 bp for RS integration.

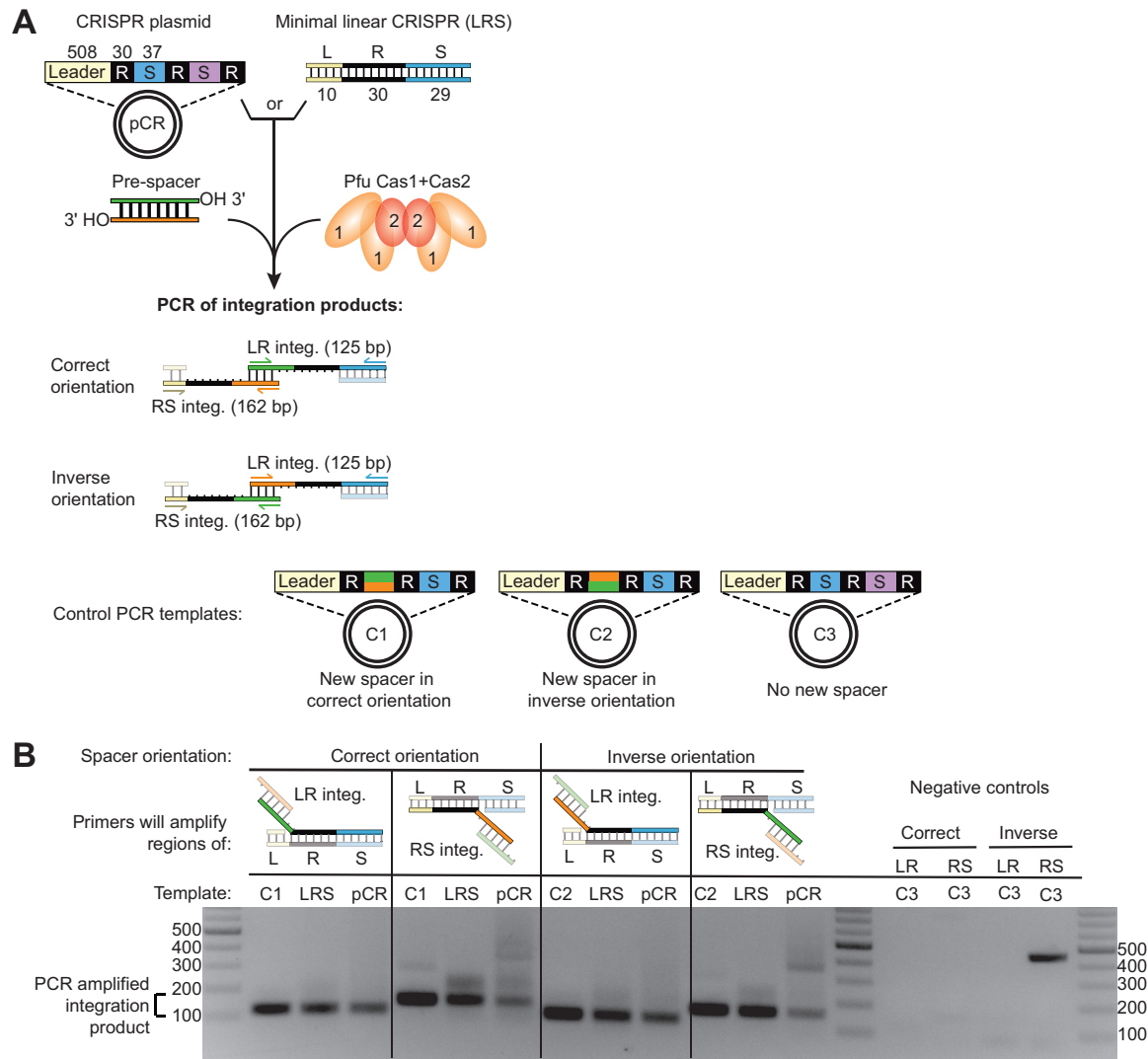


Figure 2.S9. Mutations to the first or last nucleotide of the repeat

(A) Annotated sequence of minimal linear CRISPR. (B) List of substitution mutations made to the minimal CRISPR. (C, D) Integration assay comparing mutations to the (C) first and (D) last nucleotide of the repeat. Noncontiguous lanes from the same gel are indicated by dashed lines. (E) Integration sites along the minimal CRISPRs.

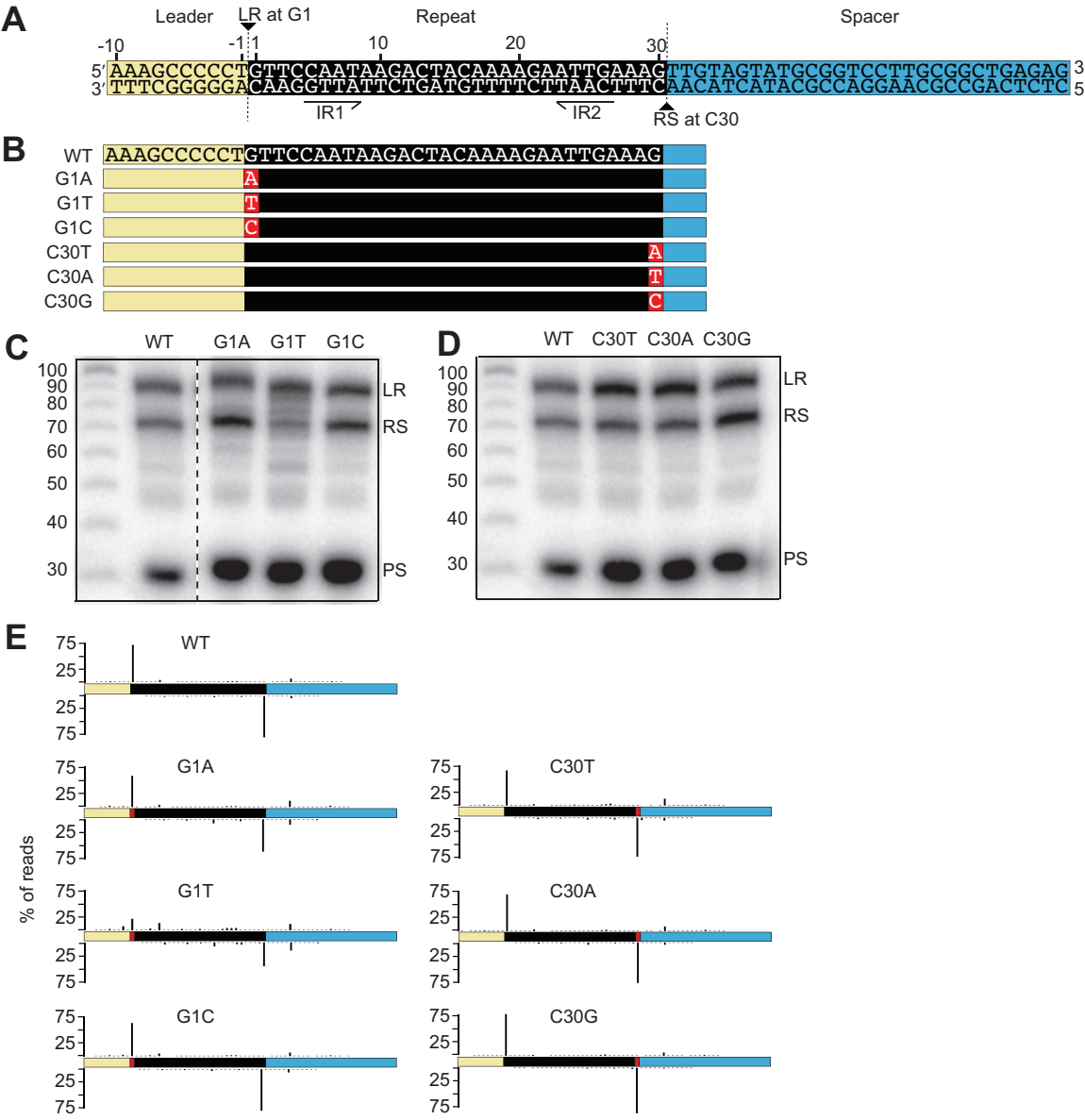


Figure 2.S10. Mutations made to block 6 of the repeat

(A) Annotated sequence of minimal linear CRISPR. (B) List of substitution mutations made to the minimal CRISPR. (C) Integration assay comparing original minimal CRISPR length to a shortened minimal CRISPR to distinguish the integration products of the block 6 mutation. LR integration product of short minimal CRISPR is 64 bp. (D) Integration assay comparing alternative mutations to block 6. (E) Integration sites along the minimal CRISPRs.

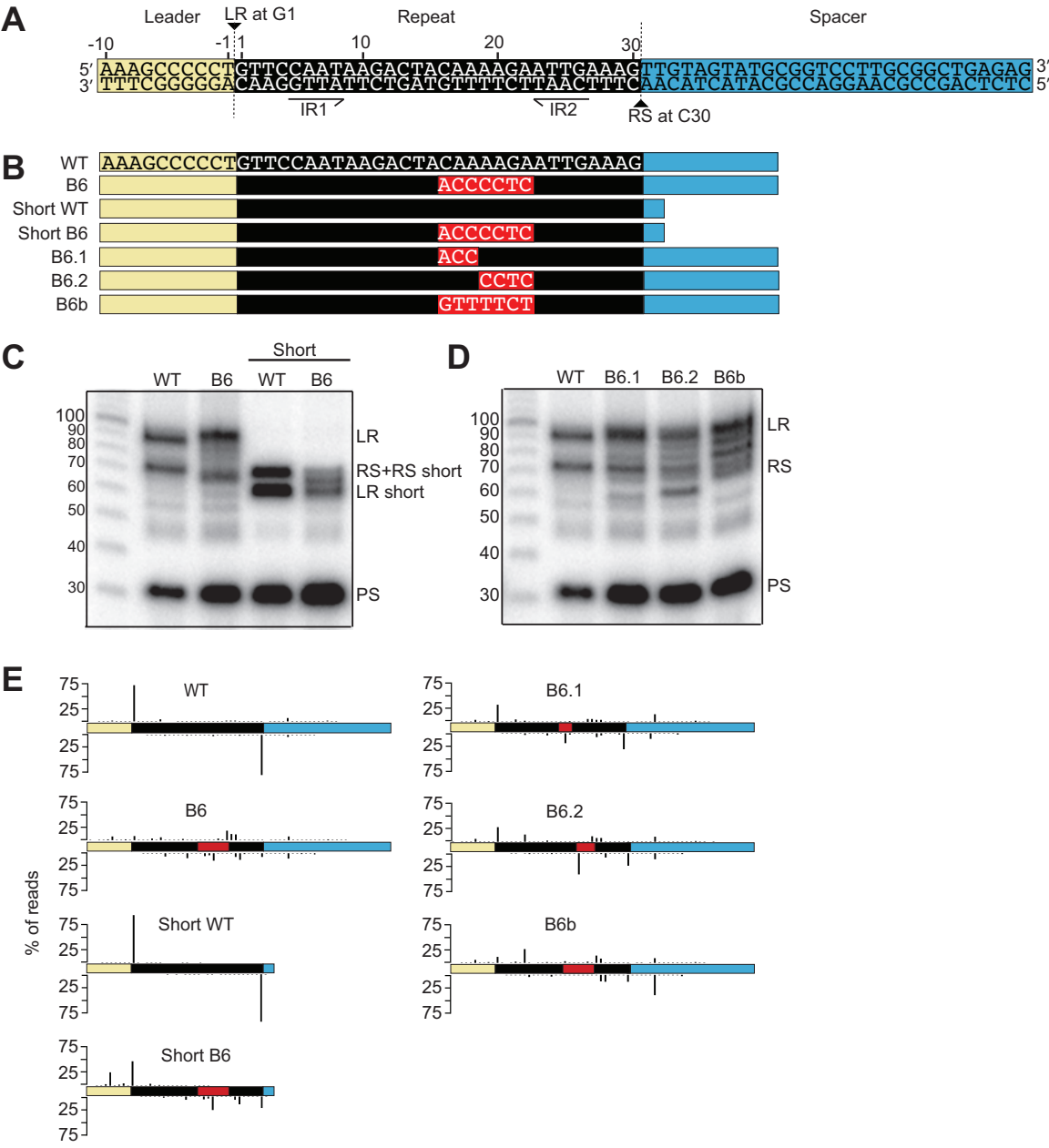
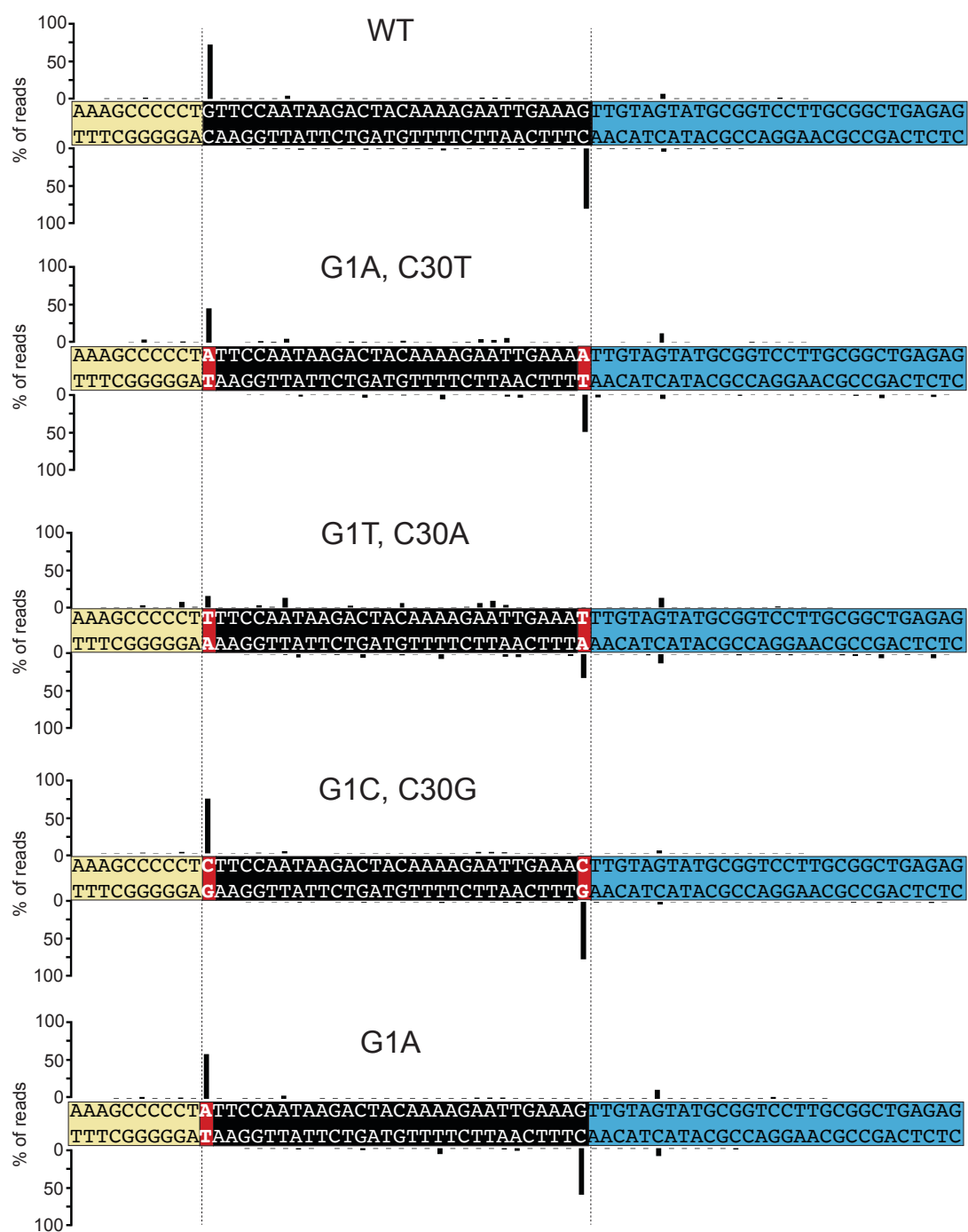
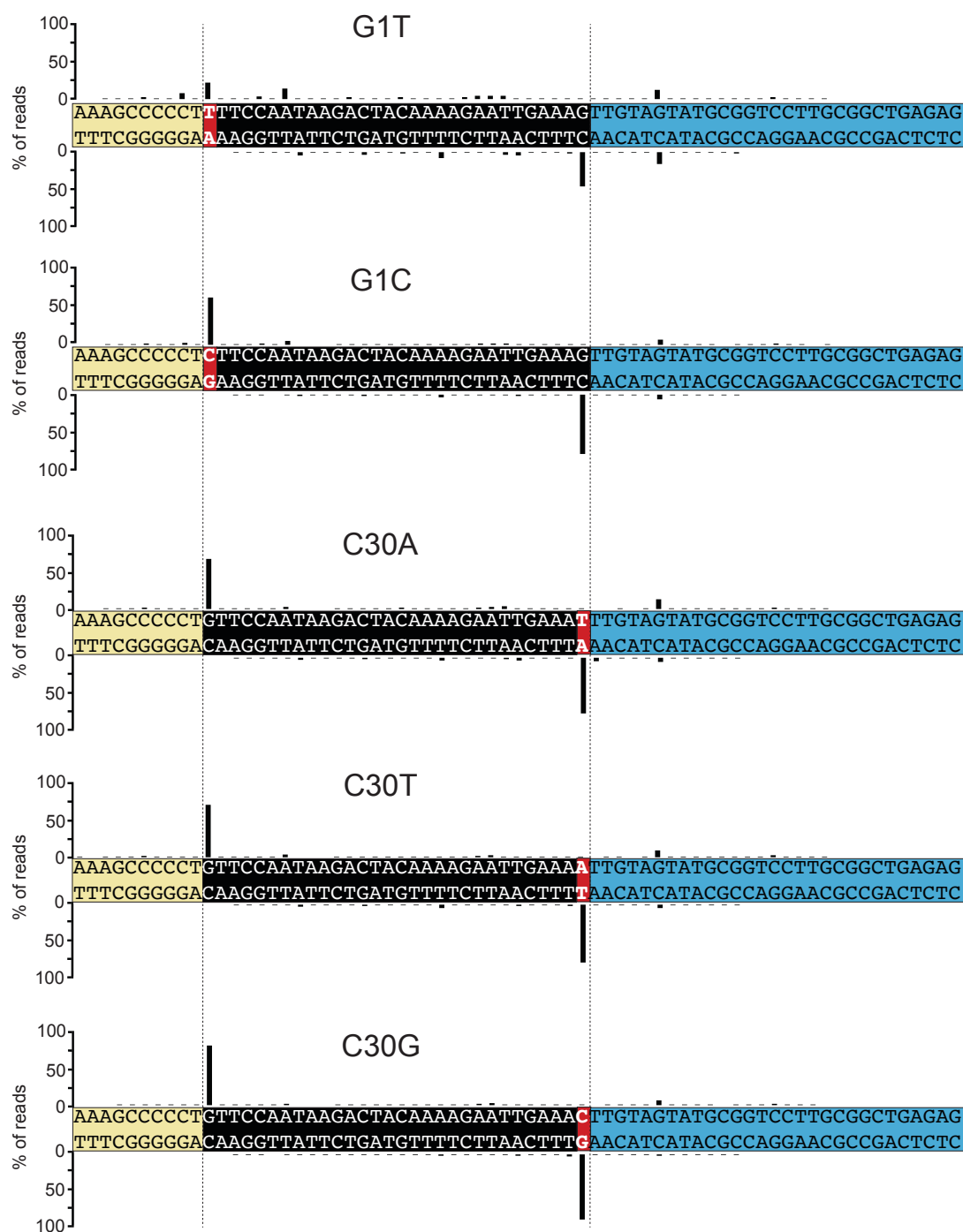
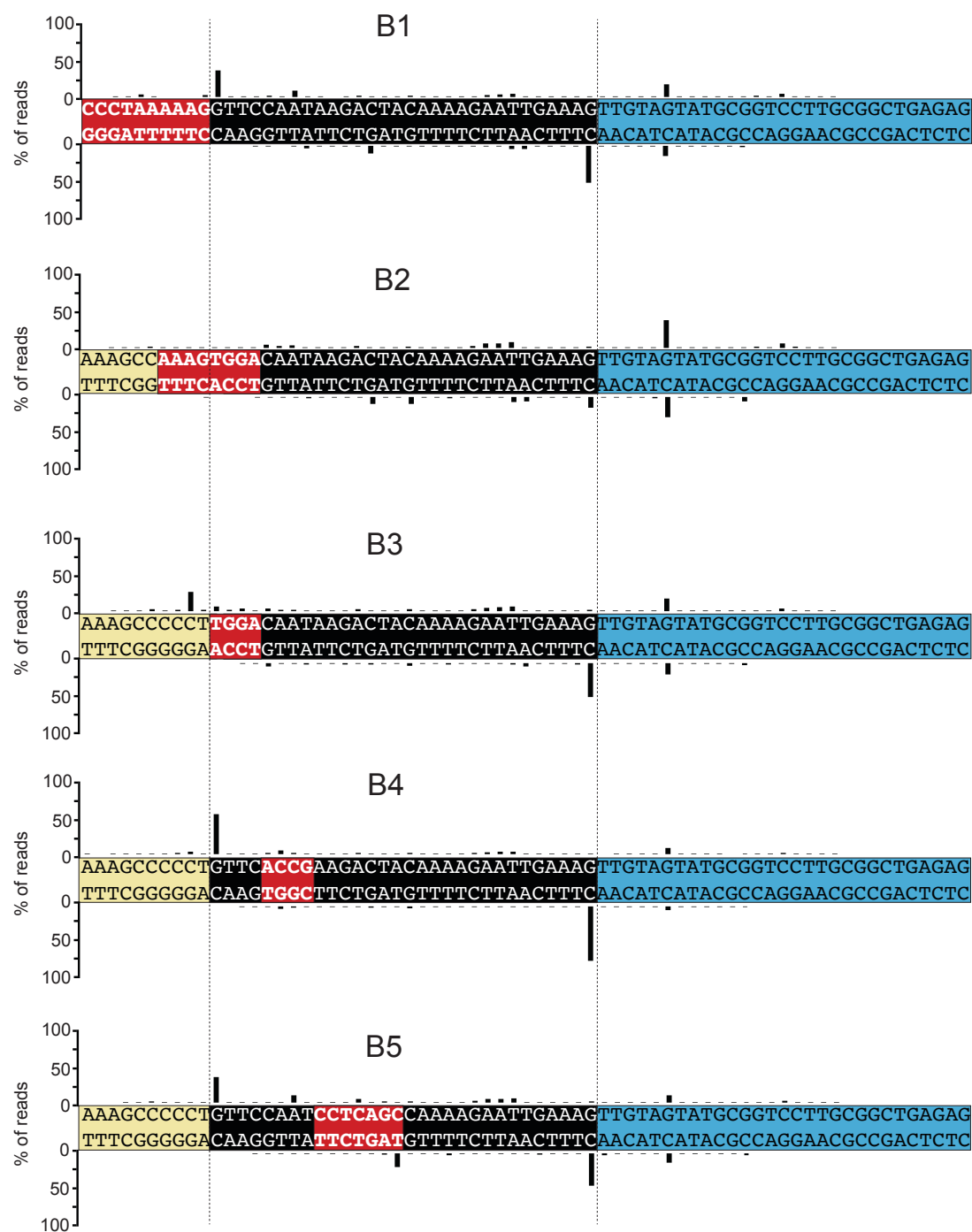
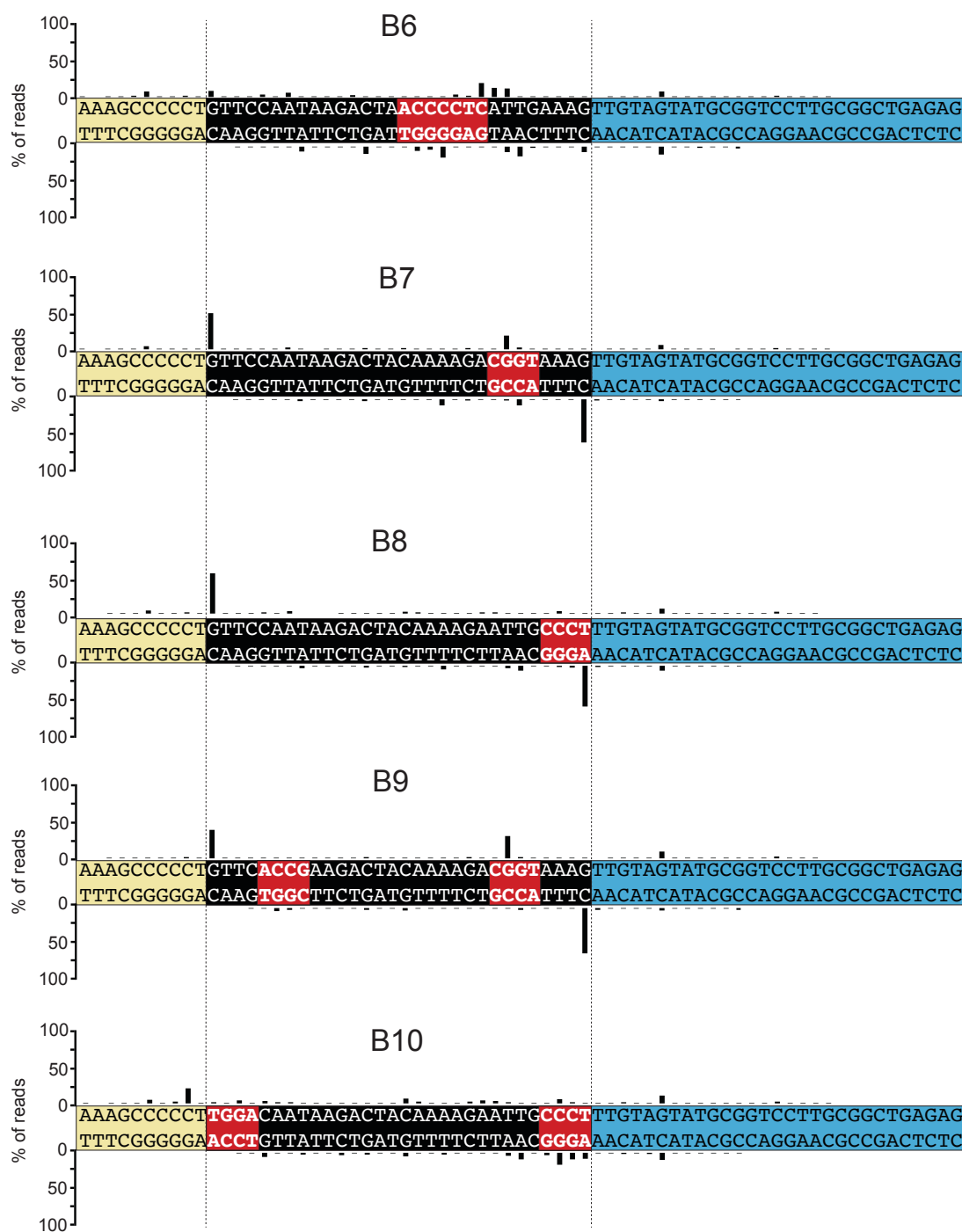


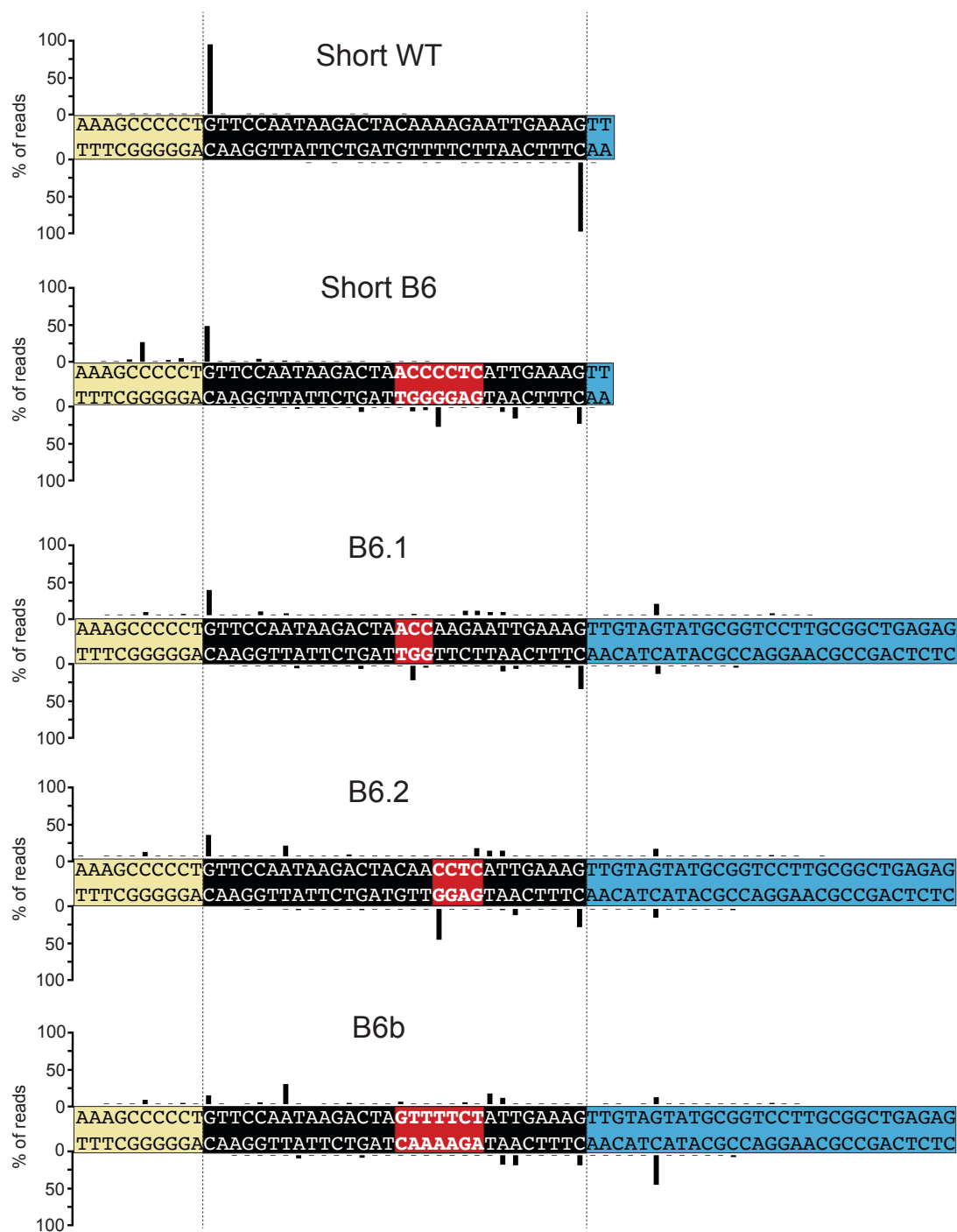
Figure 2.S11. Nucleotide resolution graphs displaying spacer integration sites along minimal CRISPRs used in all experiments. Summary table of % of reads for LR and RS sites can be found in Table 2.S2.











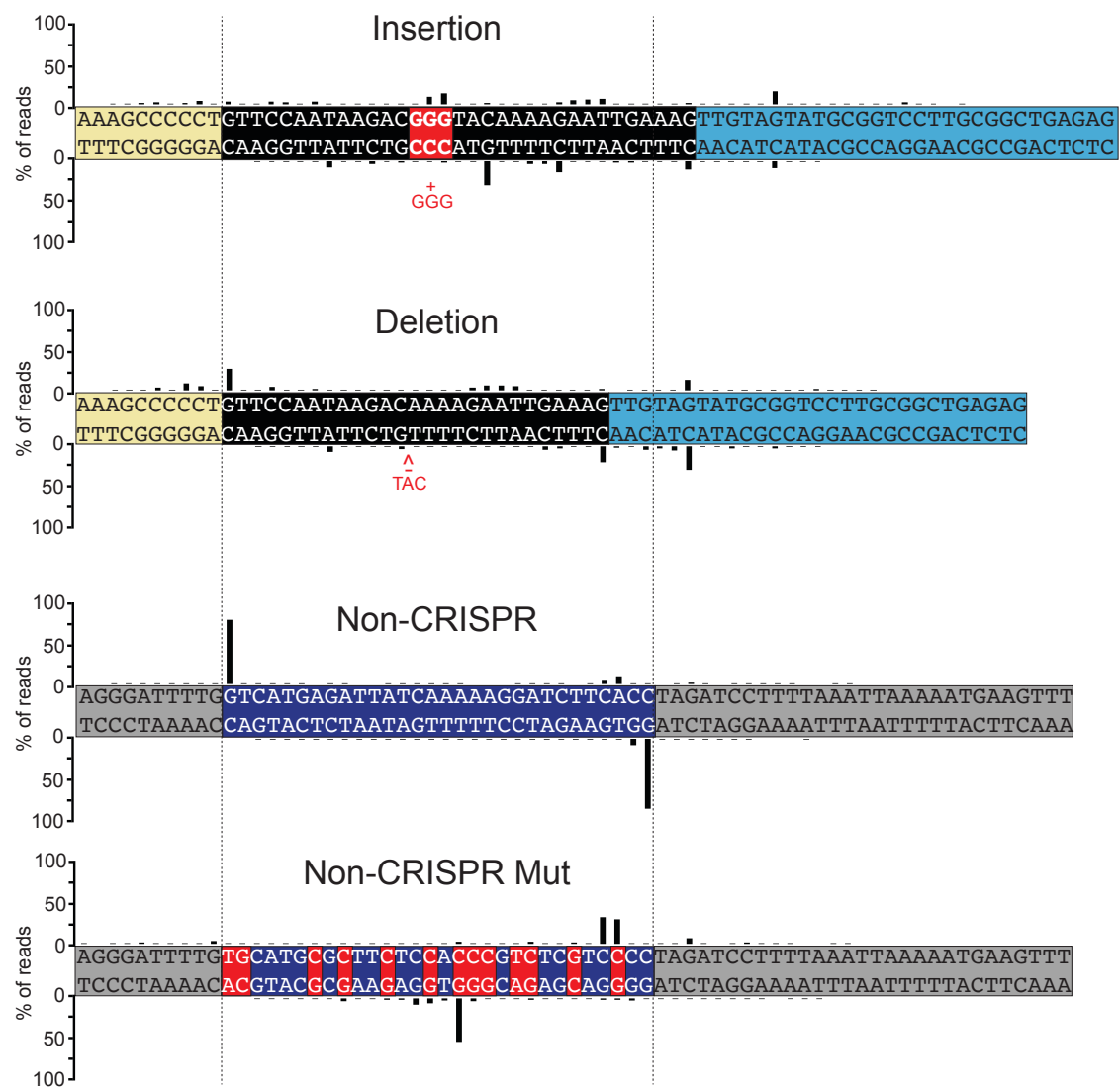


Table 2.S1 Oligonucleotides used in study

Name	Sequence (5'-3')
Pre-spacer top	ATATAAGCAATACGGAACCTTATGAGAACTATT
Pre-spacer bottom	TTCTCATAAGTTCCGTATTGCTTATATTTCAA
LRS linear top	AAAGCCCCCTGTTCCAATAAGACTACAAAAGAATTGAAAGTT GTAGTATGCGGTCCTTGCGGCTGAGAG
LRS linear bottom	CTCTCAGCCGCAAGGACCGCATACTACAACCTTTCAATTCTTT TGTAGTCTTATTGGAACAGGGGGCTTT
LR half-site top	ATATAAGCAATACGGAACCTTATGAGAACTATTGTTCCAATAA GACTACAAAAGAATTGAAAGTTGTAGTATGCGGTCCTTGCG GCTGAG
LR half-site leader	GGATTGGGGCAAAAAAGCCCCCT
LR half-site bottom	CTCAGCCGCAAGGACCGCATACTACAACCTTTCAATTCTTTTG TAGTCTTATTGGAACAGGGGGCTTTTTTGCCCCAATCC
LR half-site spacer	TTCTCATAAGTTCCGTATTGCTTATATTTCAA
RS half-site top	GGATTGGGGCAAAAAAGCCCCCTGTTCCAATAAGACTACAA AAGAATTGAAAGTTGTAGTATGCGGTCCTTGCGGCTGAG
RS half-site bottom	TTCTCATAAGTTCCGTATTGCTTATATTTCAACCTTTCAATTCT TTTGTAGTCTTATTGGAACAGGGGGCTTTTTTGCCCCAATCC
RS half-site S1	CTCAGCCGCAAGGACCGCATACTACAA
RS half-site spacer	ATATAAGCAATACGGAACCTTATGAGAACTATT
LRS spacer hairpin	GCAAAAAAGCCCCCTGTTCCAATAAGACTACAAAAGAATTG AAAGGAGTGTGCGAGGGCTTGGGCCCTGAGGTAACGTGCG ATCGCACGTTACCTCAGGGGCCAAGCCCTCGCACACTCCTT TCAATTCTTTTGTAGTCTTATTGGAACAGGGGGCTTTTTTG
LRS leader hairpin	GCCCTCGCACACTCCTTTCAATTCTTTTGTAGTCTTATTGGA ACAGGGGGGCTTTTTTGCCGATAGGCAAAAAAGCCCCCTGTT CCAATAAGACTACAAAAGAATTGAAAGGAGTGTGCGAGGGC
Non-CRISPR hairpin	GTAAAGGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCA CCTAGATCCTTTTAAATCGATAGATTTAAAGGATCTAGGTG AAGATCCTTTTGTATAATCTCATGACCAAAATCCCTTAAC
G1A, C30T top	AAAGCCCCCTATTCCAATAAGACTACAAAAGAATTGAAATT GTAGTATGCGGTCCTTGCGGCTGAGAG
G1A, C30T bottom	CTCTCAGCCGCAAGGACCGCATACTACAATTTTCAATTCTTT TGTAGTCTTATTGGAATAGGGGGCTTT
G1T, C30A top	AAAGCCCCCTTTTCCAATAAGACTACAAAAGAATTGAAATTT GTAGTATGCGGTCCTTGCGGCTGAGAG
G1T, C30A bottom	CTCTCAGCCGCAAGGACCGCATACTACAATTTTCAATTCTTT TGTAGTCTTATTGGAAGAGGGGGCTTT
G1C, C30G top	AAAGCCCCCTCTTCCAATAAGACTACAAAAGAATTGAACTT GTAGTATGCGGTCCTTGCGGCTGAGAG
G1C, C30G bottom	CTCTCAGCCGCAAGGACCGCATACTACAAGTTTCAATTCTTT TGTAGTCTTATTGGAAGAGGGGGCTTT
G1A top	AAAGCCCCCTATTCCAATAAGACTACAAAAGAATTGAAAGTT GTAGTATGCGGTCCTTGCGGCTGAGAG
G1A bottom	CTCTCAGCCGCAAGGACCGCATACTACAACCTTTCAATTCTTT TGTAGTCTTATTGGAATAGGGGGCTTT

G1T top	AAAGCCCCCTTTTCCAATAAGACTACAAAAGAATTGAAAGTT GTAGTATGCGGTCCTTGCGGCTGAGAG
G1T bottom	CTCTCAGCCGCAAGGACCGCATACTACAACTTTCAATTCTTT TGAGTCTTATTGGAAAAGGGGGCTTT
G1C top	AAAGCCCCCTCTTCCAATAAGACTACAAAAGAATTGAAAGTT GTAGTATGCGGTCCTTGCGGCTGAGAG
G1C bottom	CTCTCAGCCGCAAGGACCGCATACTACAACTTTCAATTCTTT TGAGTCTTATTGGAAGAGGGGGCTTT
C30T top	AAAGCCCCCTGTTCCAATAAGACTACAAAAGAATTGAAAATT GTAGTATGCGGTCCTTGCGGCTGAGAG
C30T bottom	CTCTCAGCCGCAAGGACCGCATACTACAATTTTCAATTCTTT TGAGTCTTATTGGAACAGGGGGCTTT
C30A top	AAAGCCCCCTGTTCCAATAAGACTACAAAAGAATTGAAATTT GTAGTATGCGGTCCTTGCGGCTGAGAG
C30A bottom	CTCTCAGCCGCAAGGACCGCATACTACAAATTTCAATTCTTT TGAGTCTTATTGGAACAGGGGGCTTT
C30G top	AAAGCCCCCTGTTCCAATAAGACTACAAAAGAATTGAACTT GTAGTATGCGGTCCTTGCGGCTGAGAG
C30G bottom	CTCTCAGCCGCAAGGACCGCATACTACAAGTTTCAATTCTTT TGAGTCTTATTGGAACAGGGGGCTTT
Block 1 top	CCCTAAAAAGGTTCCAATAAGACTACAAAAGAATTGAAAGTT GTAGTATGCGGTCCTTGCGGCTGAGAG
Block 1 bottom	CTCTCAGCCGCAAGGACCGCATACTACAACTTTCAATTCTTT TGAGTCTTATTGGAACCTTTTATAGG
Block 2 top	AAAGCCAAAGTGGACAATAAGACTACAAAAGAATTGAAAGTT GTAGTATGCGGTCCTTGCGGCTGAGAG
Block 2 bottom	CTCTCAGCCGCAAGGACCGCATACTACAACTTTCAATTCTTT TGAGTCTTATTGTCCAATTTGGCTTT
Block 3 top	AAAGCCCCCTTGGACAATAAGACTACAAAAGAATTGAAAGTT GTAGTATGCGGTCCTTGCGGCTGAGAG
Block 3 bottom	CTCTCAGCCGCAAGGACCGCATACTACAACTTTCAATTCTTT TGAGTCTTATTGTCCAAGGGGGCTTT
Block 4 top	AAAGCCCCCTGTTACCGAAGACTACAAAAGAATTGAAAGT TGAGTATGCGGTCCTTGCGGCTGAGAG
Block 4 bottom	CTCTCAGCCGCAAGGACCGCATACTACAACTTTCAATTCTTT TGAGTCTTCGGTGAACAGGGGGCTTT
Block 5 top	AAAGCCCCCTGTTCCAATCCTCAGCCAAAAGAATTGAAAGT TGAGTATGCGGTCCTTGCGGCTGAGAG
Block 5 bottom	CTCTCAGCCGCAAGGACCGCATACTACAACTTTCAATTCTTT TGGCTGAGGATTGGAACAGGGGGCTTT
Block 6 top	AAAGCCCCCTGTTCCAATAAGACTAACCCTCATTGAAAGTT GTAGTATGCGGTCCTTGCGGCTGAGAG
Block 6 bottom	CTCTCAGCCGCAAGGACCGCATACTACAACTTTCAATGAGG GGTTAGTCTTATTGGAACAGGGGGCTTT
Block 7 top	AAAGCCCCCTGTTCCAATAAGACTACAAAAGACGGTAAAGT TGAGTATGCGGTCCTTGCGGCTGAGAG
Block 7 bottom	CTCTCAGCCGCAAGGACCGCATACTACAACTTTACCGTCTT TTGAGTCTTATTGGAACAGGGGGCTTT
Block 8 top	AAAGCCCCCTGTTCCAATAAGACTACAAAAGAATTGCCCTTT GTAGTATGCGGTCCTTGCGGCTGAGAG

Block 8 bottom	CTCTCAGCCGCAAGGACCGCATACTACAAAGGGCAATTCTT TTGTAGTCTTATTGGAACAGGGGGGCTTT
Block 9 top	AAAGCCCCCTGTTCAACCGAAGACTACAAAAGACGGTAAAGT TGTAGTATGCGGTCCTTGCGGCTGAGAG
Block 9 bottom	CTCTCAGCCGCAAGGACCGCATACTACAACTTTACCGTCTT TTGTAGTCTTCGGTGAACAGGGGGGCTTT
Block 10 top	AAAGCCCCCTGTTGACAATAAGACTACAAAAGAATTGCCCTTT GTAGTATGCGGTCCTTGCGGCTGAGAG
Block 10 bottom	CTCTCAGCCGCAAGGACCGCATACTACAAAGGGCAATTCTT TTGTAGTCTTATTGTCCAAGGGGGGCTTT
Short WT top	AAAGCCCCCTGTTCCAATAAGACTACAAAAGAATTGAAAGTT
Short WT bottom	AACTTTCAATTCTTTTGTAGTCTTATTGGAACAGGGGGGCTTT
Short Block 6 top	AAAGCCCCCTGTTCCAATAAGACTAACCCTCATTGAAAGTT
Short Block 6 bottom	AACTTTCAATGAGGGGTAGTCTTATTGGAACAGGGGGGCTT T
Block 6.1 top	AAAGCCCCCTGTTCCAATAAGACTAACCAAGAATTGAAAGTT GTAGTATGCGGTCCTTGCGGCTGAGAG
Block 6.1 bottom	CTCTCAGCCGCAAGGACCGCATACTACAACTTTCAATTCTT GGTTAGTCTTATTGGAACAGGGGGGCTTT
Block 6.2 top	AAAGCCCCCTGTTCCAATAAGACTACAACCTCATTGAAAGTT GTAGTATGCGGTCCTTGCGGCTGAGAG
Block 6.2 bottom	CTCTCAGCCGCAAGGACCGCATACTACAACTTTCAATGAGG TTGTAGTCTTATTGGAACAGGGGGGCTTT
Block 6b top	AAAGCCCCCTGTTCCAATAAGACTAGTTTTCTATTGAAAGTT GTAGTATGCGGTCCTTGCGGCTGAGAG
Block 6b bottom	CTCTCAGCCGCAAGGACCGCATACTACAACTTTCAATAGAA AACTAGTCTTATTGGAACAGGGGGGCTTT
Insertion top	AAAGCCCCCTGTTCCAATAAGACGGGTACAAAAGAATTGAA AGTTGTAGTATGCGGTCCTTGCGGCTGAGAG
Insertion bottom	CTCTCAGCCGCAAGGACCGCATACTACAACTTTCAATTCTTT TGTACCCGTCTTATTGGAACAGGGGGGCTTT
Deletion top	AAAGCCCCCTGTTCCAATAAGACAAAAGAATTGAAAGTTGTA GTATGCGGTCCTTGCGGCTGAGAG
Deletion bottom	CTCTCAGCCGCAAGGACCGCATACTACAACTTTCAATTCTTT TGTCTTATTGGAACAGGGGGGCTTT
Non-CRISPR top	AGGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCACCTA GATCCTTTTAAATTAAAAATGAAGTTT
Non-CRISPR bottom	AAACTTCATTTTTAATTTAAAGGATCTAGGTGAAGATCCTTT TTGATAATCTCATGACCAAAATCCCT
Non-CRISPR mutant top	AGGGATTTTGTGCATGAGCTTCTCCACCCGTCTCGTCCCCT AGATCCTTTTAAATTAAAAATGAAGTTT
Non-CRISPR mutant bottom	AAACTTCATTTTTAATTTAAAGGATCTAGGGGACGAGACGG GTGGAGAAGCTCATGCACAAAATCCCT

Table 2.S2. Percent of sequencing reads occurring at leader-repeat (LR) and repeat-spacer (RS) junctions of minimal CRISPR sequences, relative to WT normalized to 100%. Substrate ‘Short Block 6’ is calculated relative to ‘Short WT.’

	% total reads relative to WT	
Substrate	LR	RS
WT	100	100
G1A, C30T	65	70
G1T, C30A	23	46
G1C, C30G	108	112
G1A	84	77
G1T	30	56
G1C	89	95
C30T	87	86
C30A	94	92
C30G	106	103
Block 1	51	62
Block 2	1	18
Block 3	9	56
Block 4	79	95
Block 5	50	55
Block 6	12	9
Block 7	70	76
Block 8	80	72
Block 9	55	80
Block 10	3	11
Short WT	100	100
Short Block 6	51	24
Block 6.1	46	38
Block 6.2	40	30
Block 6b	16	17
Insertion	5	2
Deletion	37	25
Non-CRISPR	110	106
Non-CRISPR mutant	0.01	1

CHAPTER 3

CHARACTERIZING CAS4-1 AND CAS4-2 *IN VITRO* TO DETERMINE THEIR
FUNCTION IN CRISPR-CAS ADAPTATION IN *PYROCOCCUS FURIOSUS*¹

¹Julie Grainy. To be submitted to Nucleic Acids Research.

Abstract

As new DNA spacers are added to CRISPR-Cas loci to be used for prokaryotic defense against mobile genetic elements, the source, size, and orientation of the incoming spacer needs to be regulated. If left unregulated, spacers can be acquired from any DNA source (including the host genome), in a range of sizes, and in both orientations, leading to ineffective targeting of foreign DNA. Here we characterize Cas4-1 and Cas4-2 proteins from *Pyrococcus furiosus* in order to gain insight into their respective roles in CRISPR-Cas adaptation. We found that recombinant Cas4-1 and Cas4-2 proteins interact with the Cas1 and Cas2 proteins responsible for spacer integration into the CRISPR locus. The complex of Cas1, Cas2, Cas4-1, and Cas4-2 was able to facilitate integration of new spacers *in vitro*, but the incoming spacer DNA was not processed to the appropriate size for functional CRISPR-Cas immunity. Moreover, Cas4-2 exhibited *in vitro* nuclease activity against both ssDNA and dsDNA, with apparent 3' to 5' exonuclease activity. Together, these results support previous *in vivo* evidence of Cas4-2 involvement in CRISPR-Cas adaptation, and add insight into the growing model of spacer generation.

Introduction

CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats)-Cas (CRISPR-associated proteins) systems allow prokaryotes to acquire immunity towards viruses (1) and plasmids (2), and are found in about 50% of bacteria and 90% of archaea (3). CRISPR loci found on the host genome contain an array of 25-35 bp direct repeats separated by 30-40 bp DNA “spacers” that are derived from the invader (1). An AT-rich

“leader” sequence lies directly upstream along with nearby CRISPR-associated genes required for immunity (3-5).

CRISPR-Cas immunity is achieved in three distinct stages (6,7): adaptation, CRISPR RNA (crRNA) biogenesis, and invader silencing. New spacers are acquired during adaptation by capturing fragments of the invading nucleic acid and inserting it into the host’s CRISPR array. The CRISPR array is transcribed and matured into crRNAs to be used as complementary guides to invader DNA or RNA (8-10). During invader silencing, the crRNA guides effector Cas proteins to the invader for targeted nucleic acid degradation (11-14).

While the mechanisms of crRNA biogenesis and interference have been heavily studied over the last decade, the mechanisms governing adaption have only recently begun being elucidated. Adaptation is often broken down into two stages, pre-spacer generation and spacer integration into the CRISPR array. Much emphasis has been placed on studying the process of spacer integration, but the underlying mechanisms that coordinate selection and processing of foreign DNA prior to integration is poorly understood. The success of the interference stage of CRISPR-Cas immunity relies on a sequence directly neighboring the target sequence (15-17). This sequence is called the protospacer adjacent motif, because that same sequence needs to be present in the original protospacer from which the spacer was derived. Therefore, it is paramount that newly acquired spacers be processed precisely downstream of PAMs prior to spacer integration at the CRISPR array. It is also key that the new spacer be oriented correctly to ensure that the crRNA is complementary to the strand of the invader that contains the PAM.

While Cas1 and Cas2 proteins are responsible for integrating new spacers into the CRISPR array, it has been shown that Cas4 proteins play a role in preparing spacers for accurate integration (18-24). *Cas4* genes are found in Type I, II, and V CRISPR-Cas systems (25), and have been directly studied in a handful of Type I subtypes. Cas4 knockouts reduce adaptation efficiency in Type I-A and I-B systems (26,27). Most importantly, Cas4 proteins are responsible for the selection of spacers of suitable length, with correct PAMs, and proper orientation *in vivo* (19,21,23,24). PAM processing by Cas4 from a Type I-C system has been confirmed by *in vitro* experimentation (20,22).

While progress has been made in understanding the detailed mechanisms of protospacer generation by Cas4 proteins in the Type I-C CRISPR-Cas system (20,22), the structural and biochemical diversity of Cas4 proteins observed so far (25,28-31) compels further exploration into unstudied systems. Here we aim to describe the roles that the Cas4-1 and Cas4-2 proteins of the hyperthermophilic archaeon *Pyrococcus furiosus* play in the process of adaptation. The *P. furiosus* Cas4-1 nuclease defines the 5'-NGG-3' PAM for the CRISPR-Cas systems of the organism *in vivo*, while Cas4-2 defines an 5'-NW-3' motif at the other flanking sequence of the protospacer (21). This study attempts to characterize the molecular mechanisms of these proteins *in vitro*, in order to advance the model of adaptation for systems containing Cas4 proteins.

Materials and methods

Cloning, expression, and purification

The *cas1* (PF1118), *cas2* (PF1117), *cas4-1* (PF1119), and *cas4-2* (PF1793) genes (Figure 3.1A) were amplified from *Pyrococcus furiosus* COM1 genomic DNA and individually cloned into the following expression vectors: pET24d for N-terminally His-tagged constructs, pET21d for Cas1 or Cas2 C-terminally His-tagged constructs, pET101d for Cas4-1 or Cas4-2 C-terminally His-tagged constructs, and pET21d for untagged constructs). Co-expression constructs were created by cloning the genes including the T7 promoter and T7 terminator regions from the individual expression vectors into the pET24d backbone.

Expression vectors were transformed into *E. coli* BL21 RIPL strain. Cultures were grown at 37°C in 1 L of Terrific Broth to an OD₆₀₀ of 0.4-0.6, and expression of the proteins was induced with 1 mM IPTG during overnight growth at 16°C. Harvested cells were lysed in 40 mM Tris, 500 mM KCl, 10% glycerol, 10 mM imidazole, pH 7.5. Immediately following, the cell lysate was degassed with a gas manifold with 20% CO₂ and 80% N₂ mixed gas for 30 minutes to remove oxygen and remained anaerobic for all remaining steps of purification. Thermal precipitation at 70°C for 30 minutes was performed before the cell lysate was centrifuged at 14,000 rpm at 4°C for 30 minutes and the soluble fraction was collected and filtered (0.8 µm filter pore size Millex filter unit, Millipore). His-tagged proteins were purified by Ni²⁺ affinity column chromatography, using a stepwise increase of imidazole (10, 20, 50, 100, 250, 500 mM) in 40 mM Tris, 500 mM KCl, 10% glycerol, pH 7.5. Peak elution fractions were dialyzed using Vivaspin

500 columns with appropriate molecular weight cut-offs (GE Healthcare Life Sciences or Sartorius) into 40 mM Tris, 200 mM KCl, 10% glycerol, pH 7.5, and stored anaerobically at 4°C prior to use for functional assays.

Site-directed mutagenesis

Putative active site mutation of Cas4-2 (D77A) was made by site-directed mutagenesis. Residue D77 was shown to eliminate Cas4-2 function *in vivo* (21), and aligns with the active site residues reported from other Cas4 proteins *in vitro* (28-30) (Figure 3.1B). Mutations were introduced to the genes on individual or co-expression constructs by PCR with primers containing the desired single nucleotide changes. The template plasmid DNA was degraded with DpnI (NEB), and remaining plasmid was transformed into BL21 RIPL strain for protein expression and purification as described above.

DNA substrate preparation

DNA oligonucleotides were from Eurofins Genomics. The ssDNA substrates were 5' radiolabeled with T4 polynucleotide kinase (NEB) and [γ -³²P]-ATP. The dsDNA substrates were made by annealing labeled ssDNA oligonucleotides with unlabeled ssDNA oligonucleotides in a 1:2 ratio. Annealing was performed by incubating at 95°C for 5 minutes followed by slow cooling until 23°C. The annealed dsDNA substrates were separated by 15% nondenaturing polyacrylamide gel electrophoresis in 1x TBE, detected by autoradiographic exposure, and the band of the expected annealed dsDNA size was excised. The DNA was eluted from the gel slices overnight at 4°C in 500 μ L of reaction buffer (20 mM Tris, 100 mM KCl, 5% glycerol, pH 7.5), and extracted with

phenol/chloroform/isoamyl alcohol (pH 8.0), ethanol precipitated, and resuspended in 20 mM Tris, 100 mM KCl, 5% glycerol, pH 7.5. The dsDNA substrates with a 3' radiolabel were made by annealing oligos in a 1:1 ratio as described above, and 3' radiolabeled by Klenow fragment (Promega) and [α - ^{32} P]-dCTP. For the Klenow reaction, a single nucleotide ssDNA 3' under-hang was used as a template for the dCTP to be incorporated. The DNA was purified by DNA clean and concentrator kit (Zymo) and stored in 20 mM Tris, 100 mM KCl, 5% glycerol, pH 7.5.

Exonuclease activity assay

Unless otherwise stated, 50 nM DNA substrate was mixed with 1 μ M Cas4-1 or Cas4-2 in buffer containing 10 mM MgCl₂, 20 mM Tris, 100 mM KCl, 5% glycerol, 1 mM DTT, pH 7.5 and incubated at 70°C for 30 minutes. To halt the reaction, a final concentration of 1 μ g/ μ L of Proteinase K (Invitrogen), 1% SDS, and 25 mM EDTA were added and incubated for 15 minutes at 50°C. 10 μ L samples were mixed 1:1 with formamide gel loading buffer (47.5% formamide, 9 mM EDTA, and 0.0125% each of SDS, xylene cyanol, and bromophenol blue) and boiled for 5 minutes. Samples were separated on a pre-warmed 15% 8 M urea-PAGE 1x TBE gel. Gels were dried, and radioactivity was detected with phosphorimaging (Storm 840 Scanner GE Healthcare).

Endonuclease activity assay

Reactions were conducted as described above, but with unlabeled pUC19 plasmids DNA as the substrate. The 70°C incubation lasted 24 hours, with samples taken at 1 hour, 5 hours, and 24 hours. Samples were mixed with 1x Purple DNA loading dye (NEB) and

separated by 1% agarose 1x TAE gel electrophoresis. DNA was detected by ethidium bromide.

Spacer integration assay

Spacer integration assays were performed as previously described in Chapter 2. The addition of Cas4-1 and/or Cas4-2 to the assay is described where applicable.

Protospacer trimming and integration assay

Assay was performed as described above (Spacer integration assay), but with unlabeled protospacer or pre-spacer, and a minimal linear CRISPR containing a 3' label on the top strand of the dsDNA substrate.

Results and Discussion

Interactions between CRISPR-Cas adaptation proteins

In order to investigate if stable complexes and/or sub-complexes existed between CRISPR-Cas adaptation proteins, the genes of *cas1*, *cas2*, *cas4-1*, and *cas4-2* from *Pyrococcus furiosus* (Figure 3.1A) were cloned into pET24d expression vectors. Co-expression vectors were constructed to contain either three genes (*cas1*, *cas2*, and *cas4-1*) or four genes (*cas1*, *cas2*, *cas4-1* and *cas4-2*). Proteins were co-expressed in *E. coli* and purified using Ni²⁺ affinity column chromatography. Proteins were co-expressed so that one protein contained a C-terminal 6x-His tag while the other protein(s) were untagged, in order to identify protein-protein interactions through pull-down of untagged proteins.

Each of the untagged proteins were verified to elute from the Ni^{2+} affinity chromatography column at low imidazole concentrations (data not shown), so any untagged proteins detected above 100 mM imidazole are attributed to an interaction with the tagged protein. When expressing the three core genes (*cas1*, *cas2*, and *cas4-1*), a moderate reciprocal interaction was observed between Cas1 and Cas4-1 (Figure 3.2A and 3.2B). This interaction was also observed with the Type I-C Cas1 and Cas4 proteins of *B. halodurans* (20), indicating that this interaction may be functionally important for CRISPR-Cas adaptation at some level.

An even stronger reciprocal pull-down of all four adaptation proteins was also observed when the tagged protein was Cas4-2-His (Figure 3.2C) or Cas1-His (Figure 3.2D). Cas2 weakly interacts, and is not detectible by Coomassie stain in Figure 3.2D, but its presence was confirmed by mass spec analysis (data not shown). When expressed and purified alone, Cas4-2 is highly insoluble (data not shown), but the co-expression with associated adaptation proteins allowed it to remain stable and successfully purify in much larger yields. In light of a recent study that required the presence of dsDNA for complex formation between Cas1, Cas2, and Cas4 of the Type I-C system (22), it is possible that the addition of dsDNA will stabilize the *P. furiosus* Cas1/2-/4-1/4-2 complex even further. It remains to be seen whether or not a Cas1-Cas4 (in the absence of Cas2) complex is responsible for any relevant functions in CRISPR-Cas adaptation or otherwise. The interaction surface in the structure solved for the Cas1-Cas4 complex of *B. halodurans* (20) is different from the interaction in the Cas1-2-4 structure solved by the same investigators (22). The authors indicate that Cas4 would need to completely dissociate from Cas1 in order for Cas2 to be able to be added to the complex. Therefore it

doesn't seem likely that Cas1-Cas4 function upstream to generate pre-spacers before Cas2 is recruited for spacer integration, but the possibility cannot be ruled out until tested directly. Instead, it is more likely that Cas1-2-4 acts as a complete unit in the process of protospacer processing and subsequent spacer integration.

Cas4-2 has 3' to 5' DNase activity with sequence preferences

To determine if Cas4-2 exhibits nuclease activity, a radiolabeled dsDNA substrate was assayed with Cas4-2 in the presence or absence of 2 mM ATP. 10 mM MgCl₂ was included in the reaction, as it is predicted to reside in the active site of the protein. The dsDNA substrate was readily degraded by Cas4-2 in an ATP-independent manner (Figure 3.3A). If there is helicase activity to denature the dsDNA prior to nuclease activity, it does not rely on ATP. A point mutation in the predicted active site of the protein inactivated the nuclease activity, as expected (Figure 3.3A). The nuclease activity was rapid, as 80% of the substrate was degraded by 30 seconds, and 100% was degraded by 60 seconds (Figure 3.3B).

Since Cas4-2 is involved in protospacer processing, a 57 nt protospacer containing a PAM was used as a substrate to assay for trimming of the protospacer in the presence or absence of the other adaptation proteins. There was no change in Cas4-2 activity with the addition of Cas1, Cas2, and or Cas4-1 to the assay (Figure 3.3C). However, it was determined that in the absence of MgCl₂, the nuclease products accumulated as a distinct ladder of intermediate cleavage products (Figure 3.3C). Since the radiolabel is located at the 5' end of the substrate, this indicates 3' to 5' exonuclease or endonuclease activity. The lack of intermediate products with the addition of MgCl₂

suggests that in the presence of metal the 3' to 5' exonuclease activity is so rapid that there are no detected intermediates, or it is capable of 5' to 3' exonuclease activity as well.

The nuclease activity was also detected when using ssDNA substrates, suggesting Cas4-2 can act on dsDNA or ssDNA (Figure 3.3D). The 3' to 5' exonuclease activity of Cas4-2 was observed with Cas4-2 purified alone, or in complex with Cas1, Cas2, and Cas4-1 (Figure 3.3D). The nuclease activity achieved by the complex was solely due to Cas4-2, as Cas4-1 alone showed no activity, and a mutation of Cas4-2 (Figure 3.1B) co-purified with Cas1, Cas2, and Cas4-1 exhibited no activity (Figure 3.3D). While most of the characterized Cas4 proteins exhibit 5' to 3' exonuclease activity (28-30), and one of the Cas4 proteins from *Sulfolobus solfataricus* (Sso1391) exhibits both 5' to 3' and 3' to 5' ssDNA exonuclease activity (30), suggesting these proteins might play homologous roles in their respective systems.

Intriguingly, the sizes of the distinct intermediate nuclease products were consistently different depending on which strand of the protospacer was assayed (Figure 3.3E). The approximate cleavage sites along the 57mers that would result in the nuclease intermediates observed in Figure 3.3E are marked in Figure 3.3F. Since the cleavage sites differed depending on the sequence of the substrate, it is possible that Cas4-2 specifically cuts after certain bases rather than a set distance. Cas4-2 did exhibit 100% cleavage of 37 nucleotide dA and dT ssDNA, as well as the majority of a dC substrate, but disfavored cleavage of dG (Figure 3.3G). The robust DNase activity we observed by Cas4-2 might indicate that Cas4-2 rapidly degrades free DNA ends with blunt ends or overhangs in the cell, with base preferences. Protospacers transformed into the cells of *P. furiosus* were

1000 times less likely to become a spacer in the presence of Cas4-2 (21), indirectly suggesting that Cas4-2 is degrading the protospacer rapidly. Additionally, the NW motif on the 3' end of protospacers *in vivo* is a result of the nuclease activity of Cas4-2 (21). Both of these *in vivo* characteristics are supported by our *in vitro* evidence of rapid DNA degradation by Cas4-2 with robust cleavage after A and T nucleotides. Furthermore, when Cas4-1 (the protein responsible for the 5'-NGG PAM cleavage) is absent *in vivo*, newly acquired spacers processed by Cas4-2 have the NW motif on both flanking regions of the protospacer (21). The slower cleavage of dG substrates by Cas4-2 we observed could implicate a role in making sure Cas4-2 does not trim the PAM side of the protospacer, which needs to be cleaved by Cas4-1 for accurate spacer functionality (21).

Cas4-1 is not active under any of the *in vitro* conditions tested

Abundant yields of highly pure Cas4-1 protein was purified under anaerobic conditions (Figure 3.4A). The Fe-S cluster was presumed to be intact based on the rich golden color of the protein elution. When stored anaerobically, the color remained. However, if stored aerobically, the color would fade over the period of a week, indicating that stability of the protein and maintenance of the Fe-S cluster relies on an anaerobic environment. To determine if Cas4-1 exhibits nuclease activity, a radiolabeled ssDNA substrate was assayed in the presence and absence of 10 mM MgCl₂. No activity was detected from Cas4-1 (Figure 3.4B). A series of divalent metals was tested, to see if the activity relied on a specific ion residing in the active site of Cas4-1, but there was no nuclease activity was detected (Figure 3.4C).

While it is disappointing that Cas4-1 did not exhibit the expected nuclease activity observed by other Cas4 proteins previously characterized (28-30), it is not at all surprising given that Cas4 from the Type I-C system is not active in the absence of Cas1 (22). Due to the apparent physical interaction between Cas4-1 and Cas1, it is possible that Cas4-1 is only activated in the presence of Cas1. However, when Cas1 was titrated into the reaction there was still no detectable nuclease activity (Figure 3.4D). Endonuclease activity was also tested using linear or supercoiled pUC19 plasmid as the substrate. There was no degradation of the linear substrate even after 24 hours (Figure 3.4E). The supercoiled plasmid substrate was relaxed to the open circle confirmation by 5 hours, but this was also observed with the no protein control, suggesting the relaxation was not due to the presence of Cas4-1 (Figure 3.4F). Further experimental conditions will need to be tested to elucidate the mechanism by which Cas4-1 recognizes and cleaves off the 5'-NGG-3' PAM of the protospacer to prepare the DNA fragment for integration into the CRISPR array.

Spacer integration by Cas1 and Cas2 is not affected by the presence of Cas4-1 or Cas4-2 *in vitro*

In order to test the involvement of Cas4-1 and/or Cas4-2 in the process of spacer integration, a spacer integration assay was implemented (Figure 3.5A). A plasmid containing a minimal CRISPR locus was used as the recipient DNA. The CRISPR plasmid (pCR7) contains a full leader, three repeats, and two spacers from *Pyrococcus furiosus* CRISPR locus 7. Radiolabeled pre-spacers that were integrated into the CRISPR plasmid by Cas proteins can be detected by autoradiography. The plasmid will only be

detectable if new spacers have been integrated. Cas1 and Cas2 are sufficient for spacer integration (Figure 3.5B). With the addition of Cas4-1, spacer integration activity appears unaltered (Figure 3.5B). Additionally, if Cas4-1 with a putative-active site mutation is added, there is also no alteration, as expected (Figure 3.5B).

Next, the co-purified complex of Cas1, Cas2, Cas4-1, and Cas4-2 was tested in the spacer integration assay, rather than adding proteins together individually. First, the complex was further purified by subjecting the elution from the Ni^{2+} -affinity chromatography purification to a size-exclusion gel filtration column by our collaborators Scott Bailey and Anita Ramachandran (Johns Hopkins University) (Figure 3.5C). The fractions were sampled and examined by SDS-PAGE to confirm the presence of all four adaptation proteins of the complex (Figure 3.5D). Three fractions in particular were chosen to test in the spacer integration assay: D2, D5, and D12, to represent an early fraction, mid fraction, and late fraction, respectively. One concern was that Cas4-2 present in the complex would cleave the radiolabel off of the pre-spacer before it had a chance to be integrated into the CRISPR repeat, rendering any integration products undetectable by autoradiography. Despite this concern, there did not appear to be cleavage of the radiolabel from the protospacer in the reactions (Figure 3.5E). However, the fractions were also not active for spacer integration, due to the observation that the spacer integration product was only detected when Cas1 and Cas2 were supplemented into the reaction separately (Figure 3.5E). In case the adaptation complex required flanking regions of the spacer to stimulate spacer integration, a protospacer with 10 bp flanking each side of the spacer was tested in the assay, but again spacer integration was only seen when Cas1 and Cas2 were supplemented into the reaction (Figure 3.5E).

Cas4-1 and Cas4-2 do not trim protospacers before *in vitro* integration under conditions tested

Another more sensitive approach to detecting spacer integration is to PCR amplify the integration products. After the spacer integration assay is completed, integration into the substrate can be amplified by using a forward primer specific to the new spacer and reverse primer specific to either the first spacer or leader of the existing CRISPR array (Figure 3.6A). In this particular assay, the CRISPR was reduced to a minimal linear substrate with only 10 bp of the leader, one repeat, and part of spacer 1. Spacer integration was observed with both pre-trimmed spacers and untrimmed protospacers, but there was no trimming of the protospacer by Cas4-1 before being integrated into the CRISPR repeat, based on the size of the product compared to the pre-processed integration product (Figure 3.6B). Additionally, there was no orientation bias for the incoming spacer, as it was seen to be integrated equally well in both orientations (Figure 3.6B and 3.6C).

A final approach to testing protospacer trimming and subsequent integration was developed using a minimal linear CRISPR substrate with a 3' radiolabel on the spacer 1 end of the top strand. This would allow integration products to be distinguished by size depending on the size of the incoming spacer. If the spacer was untrimmed, the product would be 106 nt, but if it was trimmed by the adaptation proteins before being integrated into the repeat, the resulting size of the radiolabeled product would be 96 nt. One explanation as to why the adaptation complex did not previously exhibit spacer integration (Figure 3.5E) is the low concentration of the proteins. So for this assay, remaining fractions from the gel filtration (Figure 3.5C and 3.5D) were pooled and

concentrated into early (A), mid (B), and late (C) fractions. The three fractions were tested in the spacer trimming/integration assay using the radiolabeled minimal CRISPR and either pre-trimmed spacers or protospacers (Figure 3.6D). Cas1 and Cas2 alone were capable of integrating both pre-spacer and protospacer into the CRISPR, but did not trim the protospacer (Figure 3.6E). Fractions B and C of the adaptation complex were capable of integrating pre-spacers, while fraction A was not (Figure 3.6E). When fractions B and C were tested with the protospacer, spacer integration still occurred at the same relative levels, but it did not appear that the protospacer was trimmed in any way before being integrated into the CRISPR repeat (Figure 3.6F). One possible explanation for the lack of trimming could be that the ssDNA PAM on the 3' overhang is too close to the start of the duplex DNA. Cleavage of PAMs by Cas4 in the Type I-C system was recently shown to be optimized when the PAM was at least 4 nt away from the duplex DNA (22).

Conclusions

Spacer acquisition by the CRISPR-Cas system of *Pyrococcus furiosus* relies on Cas1, Cas2, Cas4-1, and Cas4-2 *in vivo* (21). The Cas1 and Cas2 proteins have been demonstrated to be responsible for the integration of new spacers *in vitro* (Chapter 2), but the role of the Cas4 proteins remained elusive for this particular CRISPR-Cas system. Based on our findings, the four adaptation proteins physically interact in a stable complex (Figure 3.2B+C and Figure 3.5C+D). In light of the structure solved of the Cas1-2-4 adaptation complex of a Type I-C system (22), it is likely that the *P. furiosus* adaptation complex would be arranged somewhat similarly. The Type I-C complex has two subunits of the identical Cas4 protein on each side of the complex. Since the *P. furiosus* adaptation

complex includes two unique Cas4 proteins, predictably one of the two Cas4 subunits would be substituted for the other Cas4 protein. A predicted model of the *P. furiosus* adaptation complex and the functional outcomes of this protein architecture is summarized in Figure 3.7.

The robust DNase activity we observed by Cas4-2 could be stalled or regulated by a more precise recognition and cleavage of PAM sequences by Cas4-1. During processing of the protospacer flanking regions, Cas1 and Cas2 could bind the duplex DNA of the spacer to protect Cas4-2 from continuing its processive 3' to 5' exonuclease activity. Our findings demonstrated that the rapid DNase activity by Cas4-2 was not stalled or altered in any way by the addition of Cas1/2/4-1 in complex with Cas4-2 (Figure 3.3C+D), but *in vivo* this promiscuous degradation of DNA could be more tightly regulated to achieve spacers of the correct size for spacer integration. After the pre-spacer has been processed by Cas4-1 on the PAM side and Cas4-2 on the opposite side, the resulting 37 bp spacer is ready for integration into the CRISPR repeat by Cas1 and Cas2. The orientation of spacer integration is likely guided by the asymmetry of the Cas1/2/4-1/4-2 complex. In order to achieve an orientation that will be functional for defense, the end of the new spacer processed by Cas4-1 would need to be integrated at the leader-distal end, while the end processed by Cas4-2 would need to be integrated at the leader-repeat junction. Given the strong interaction observed between Cas1 and Cas4-1, Cas4-2 could dissociate from the complex prior to spacer integration while Cas4-1 remains, which would provide the asymmetry of the integrase complex needed for orienting the spacer. This would leave the Cas1 integrase on that end available and unobstructed to bind to the sequence spanning the leader-repeat junction, which was shown to be

essential for efficient and specific spacer integration by Cas1 and Cas2 *in vitro* (Chapter 2).

References

1. 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.
2. Marraffini, L.A. and Sontheimer, E.J. (2008) CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science*, **322**, 1843-1845.
3. Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J., Barrangou, R., Brouns, S.J., Charpentier, E., Haft, D.H. *et al.* (2015) An updated evolutionary classification of CRISPR-Cas systems. *Nature reviews. Microbiology*, **13**, 722-736.
4. Kunin, V., Sorek, R. and Hugenholtz, P. (2007) Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome biology*, **8**, R61.
5. Makarova, K.S., Wolf, Y.I. and Koonin, E.V. (2013) The basic building blocks and evolution of CRISPR-CAS systems. *Biochemical Society transactions*, **41**, 1392-1400.
6. Makarova, K.S., Aravind, L., Wolf, Y.I. and Koonin, E.V. (2011) Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biology direct*, **6**, 38.

7. van der Oost, J., Jore, M.M., Westra, E.R., Lundgren, M. and Brouns, S.J. (2009) CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends in biochemical sciences*, **34**, 401-407.
8. Brouns, S.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J., Snijders, A.P., 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.
9. 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 & development*, **22**, 3489-3496.
10. 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.
11. 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.
12. 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.
13. Jore, M.M., Lundgren, M., van Duijn, E., Bultema, J.B., Westra, E.R., Waghmare, S.P., Wiedenheft, B., Pul, U., Wurm, R., Wagner, R. *et al.* (2011) Structural basis

- for CRISPR RNA-guided DNA recognition by Cascade. *Nature structural & molecular biology*, **18**, 529-536.
14. 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**, E2579-2586.
 15. Deveau, H., Barrangou, R., Garneau, J.E., Labonte, 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.
 16. 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.
 17. Shah, S.A., Erdmann, S., Mojica, F.J. and Garrett, R.A. (2013) Protospacer recognition motifs: mixed identities and functional diversity. *RNA biology*, **10**, 891-899.
 18. 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.
 19. 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. *Cell Reports*, **22**, 3377-3384.

20. 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.e45.
21. Shiimori, M., Garrett, S.C., Graveley, B.R. and Terns, M.P. (2018) Cas4 Nucleases Define the PAM, Length, and Orientation of DNA Fragments Integrated at CRISPR Loci. *Mol Cell*, **70**, 814-824 e816.
22. Lee, H., Dhingra, Y. and Sashital, D.G. (2019) The Cas4-Cas1-Cas2 complex mediates precise prespacer processing during CRISPR adaptation. *eLife*, **8**, e44248.
23. Zhang, Z., Pan, S., Liu, T., Li, Y. and Peng, N. (2019) Cas4 nucleases can effect specific integration of CRISPR spacers. *Journal of bacteriology*, **201**, e00747-00718.
24. Almendros, C., Nobrega, F.L., McKenzie, R.E. and Brouns, S.J.J. (2019) Cas4-Cas1 fusions drive efficient PAM selection and control CRISPR adaptation. *Nucleic acids research*, **47**, 5223-5230.
25. Hudaiberdiev, S., Shmakov, S., Wolf, Y.I., Terns, M.P., Makarova, K.S. and Koonin, E.V. (2017) Phylogenomics of Cas4 family nucleases. *BMC Evol Biol*, **17**, 232.
26. 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.
27. Liu, T., Liu, Z., Ye, Q., Pan, S., Wang, X., Li, Y., Peng, W., Liang, Y., She, Q. and Peng, N. (2017) Coupling transcriptional activation of CRISPR-Cas system

- and DNA repair genes by Csa3a in *Sulfolobus islandicus*. *Nucleic acids research*, **45**, 8978-8992.
28. Zhang, J., Kasciukovic, T. and White, M.F. (2012) The CRISPR associated protein Cas4 Is a 5' to 3' DNA exonuclease with an iron-sulfur cluster. *PloS one*, **7**, e47232.
 29. 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*. *Journal of the American Chemical Society*, **135**, 17476-17487.
 30. Lemak, S., Nocek, B., Beloglazova, N., Skarina, T., Flick, R., Brown, G., Joachimiak, A., Savchenko, A. and Yakunin, A.F. (2014) The CRISPR-associated Cas4 protein Pcal_0546 from *Pyrobaculum calidifontis* contains a [2Fe-2S] cluster: crystal structure and nuclease activity. *Nucleic acids research*, **42**, 11144-11155.
 31. Kaushik, V., Verma, V.V. and Goel, M. (2018) Functional divergence and comparative in-silico study of Cas4 proteins of DUF83 class. *J Mol Recognit*, **31**, e2694.

Figure 3.1 The Cas4 proteins of *Pyrococcus furiosus*.

(A) The CRISPR associated (Cas) genes of *P. furiosus* are illustrated. The *cas4-1* (PF1119) gene is located next to the adaptation genes, *cas1* and *cas2*. The *cas4-2* (PF1793) gene is not near any other Cas genes. **(B)** Sequence alignment of Cas4 proteins from *P. furiosus* and other organisms. The RecB-like nuclease motifs are highlighted in green, and the conserved cystine residues that coordinate the Fe-S cluster are highlighted in orange. The residue selected for the putative active site mutation of Cas4-2 (D77A) is denoted in red. Figure adapted and modified from Shiimori 2018 (21).

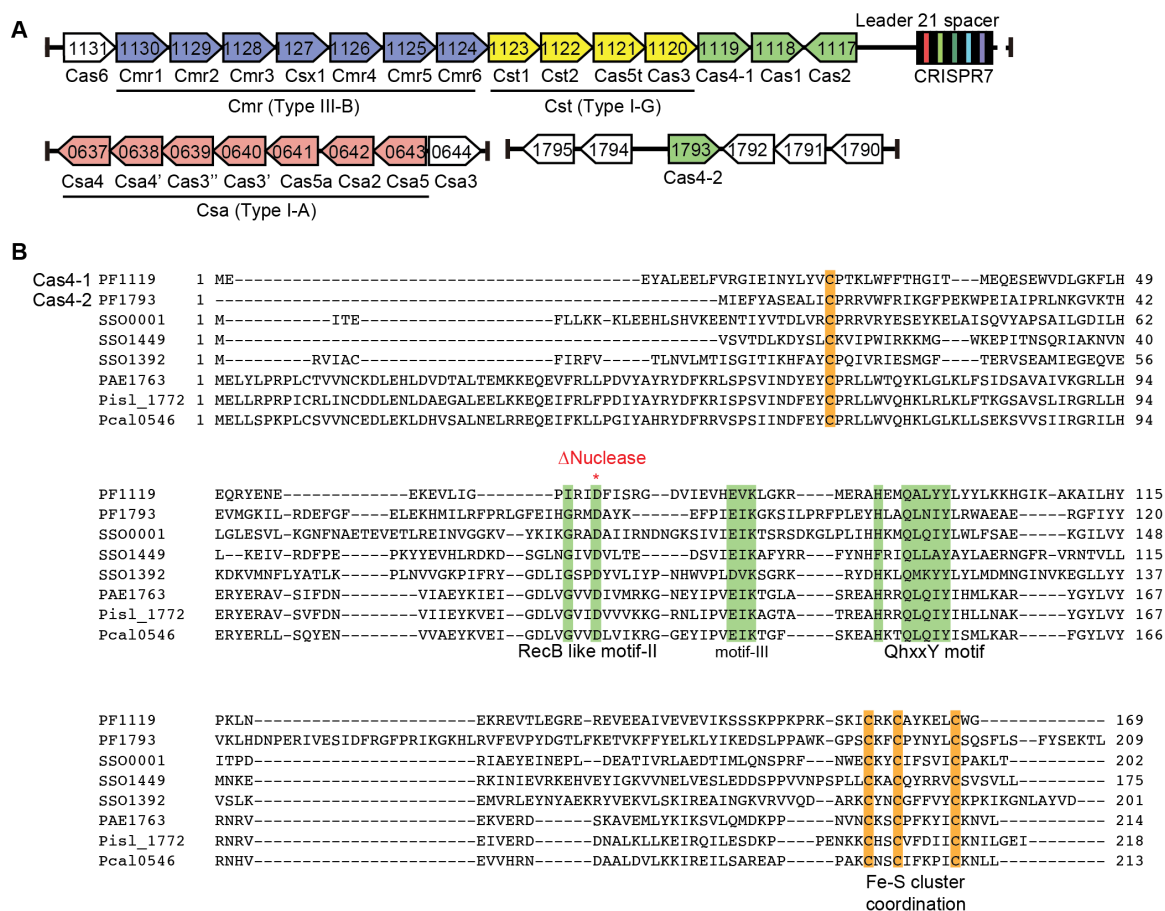


Figure 3.2. Interactions between CRISPR-Cas adaptation proteins of *Pyrococcus furiosus*.

(A) When co-expressed in *E. coli* BL21, Cas1-His pulls down untagged Cas4-1 during Ni²⁺-affinity chromatography. (B) When co-expressed in *E. coli* BL21, Cas4-1-His pulls down untagged Cas1 during Ni²⁺-affinity chromatography. (C) When co-expressed in *E. coli* BL21, Cas4-2-His pulls down untagged Cas1, Cas2, and Cas4-1 during Ni²⁺-affinity chromatography. (D) When co-expressed in *E. coli* BL21, Cas1-His pulls down untagged Cas4-1 and Cas4-2 during Ni²⁺-affinity chromatography.

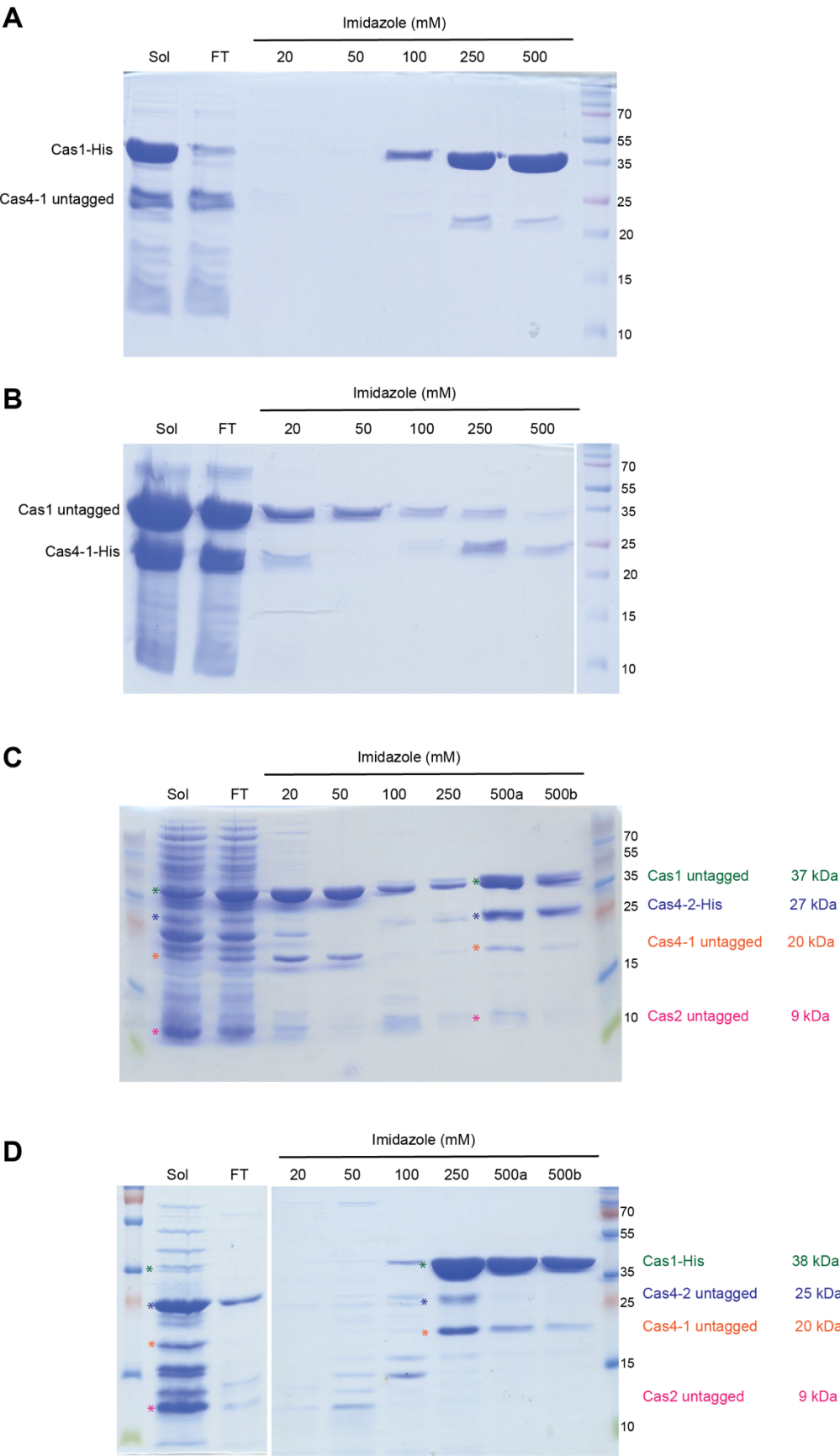


Figure 3.3. Cas4-2 has 3' to 5' DNase activity with sequence preferences

(A) Cas4-2 activity against dsDNA was assayed in the presence or absence of ATP. A 5' radiolabel on the dsDNA substrate allowed for the detection of nuclease products by autoradiography. A putative active site mutation of Cas4-2 was also compared to WT Cas4-2. (B) A time course of Cas4-2 nuclease activity. (C) Cas4-2 activity was assayed in the presents or absence of various combinations of Cas1, Cas2, and Cas4-1. Activity was tested both with and without MgCl₂. (D) Cas4-2 alone was compared to Cas4-2 that was co-purified with Cas1, Cas2, and Cas4-1. (E) Sense and antisense ssDNA substrates were assayed with Cas4-2 with and without MgCl₂. (F) Approximate cleavage sites by Cas4-2 along the sense and antisense ssDNA substrates assayed in Figure 3.3E. (G) Cas4-2 assayed against homopolymer ssDNA substrates with and without MgCl₂.

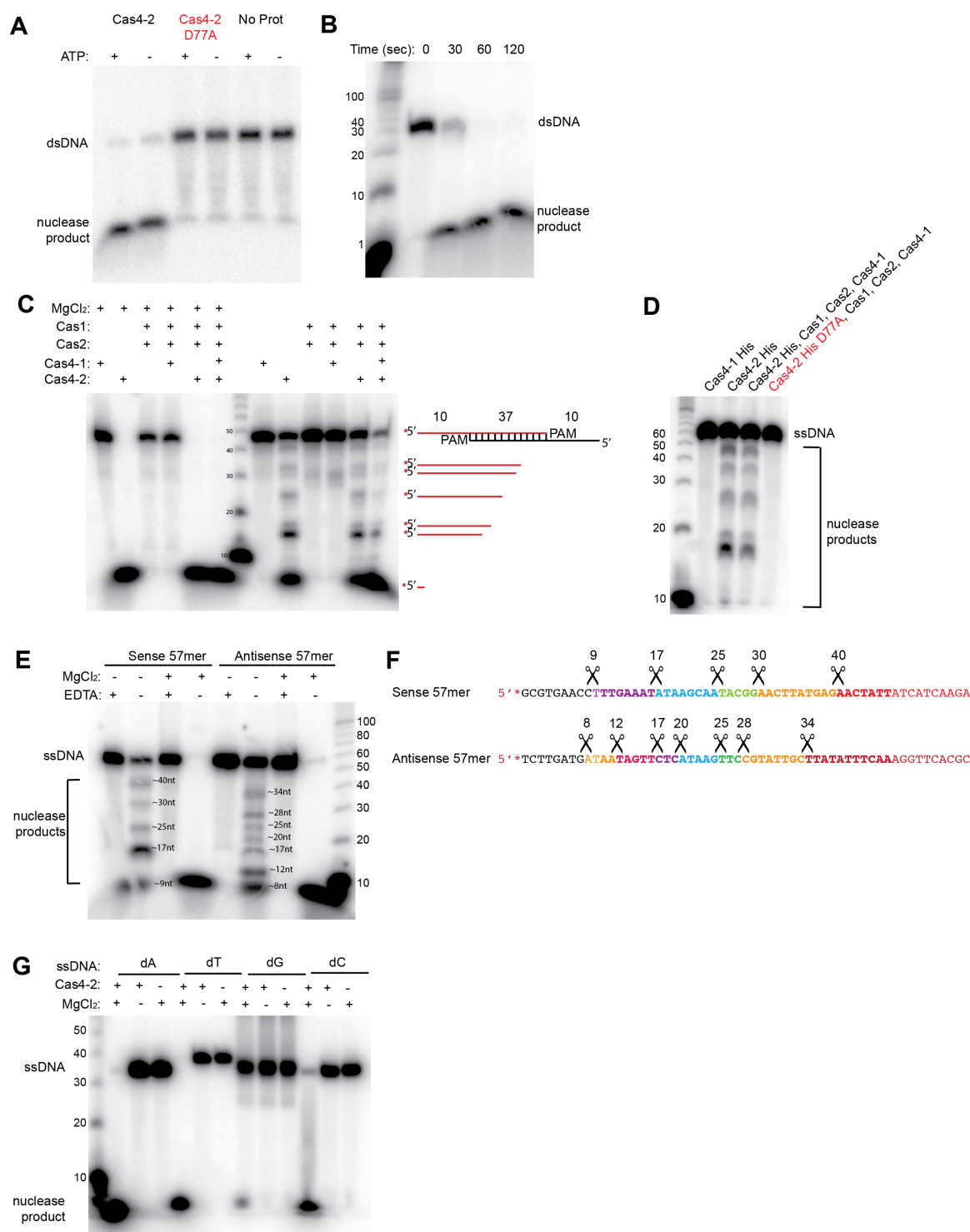


Figure 3.4. Cas4-1 is not active under any of the *in vitro* conditions tested.

(A) Cas4-1-His purified by Ni²⁺-affinity chromatography. (B) Cas4-1 activity against ssDNA was assayed in the presence or absence of MgCl₂. A 5' radiolabel on the ssDNA substrate allowed for the detection of nuclease products by autoradiography. Cas4 from *Thermococcus kodakarensis* was used as a positive control. (C) Cas4-1 activity was tested with a wide range of divalent metals. (D) Cas4-1 (1 µM) activity was tested with increasing concentration of Cas1 (0-2 µM). (E) Endonuclease activity of Cas4-1 was tested with linear pUC19 DNA, with timepoints taken at 1, 5 and 24 hours. (F) Endonuclease activity of Cas4-1 was tested with supercoiled pUC19 DNA, with timepoints taken at 1, 5 and 24 hours.

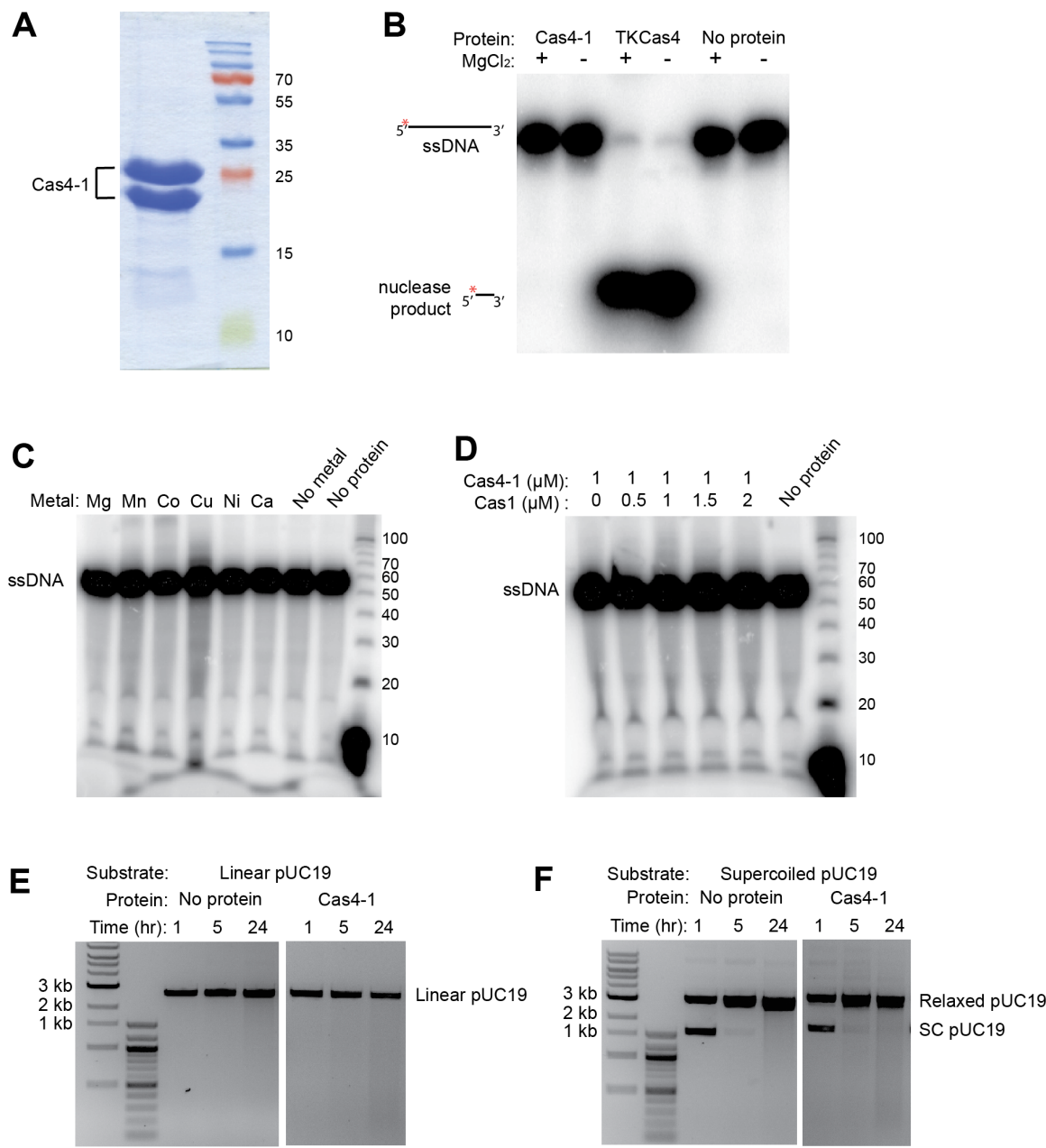


Figure 3.5. Spacer integration by Cas1 and Cas2 is not affected by the presence of Cas4-1 or Cas4-2 *in vitro*.

(A) Schematic of the *in vitro* spacer integration reaction. pCR7 containing a minimal CRISPR array was incubated with radiolabelled pre-spacers and *Pyrococcus furiosus* Cas1 and Cas2 with or without Cas4-1 and/or Cas4-2. Successful integration into the plasmid was detected by autoradiography. (B) Spacer integration by Cas1 and Cas2 is compared to the addition of WT Cas4-1 or Cas4-1 with a putative active site mutation. (C) Peak fractions observed during size-exclusion gel filtration of Cas1-His/Cas2/Cas4-1/Cas4-2 following Ni²⁺-affinity chromatography. (D) Samples of fractions C7 through D12 from Figure 3.5C analysed by SDS-PAGE. (E) Spacer integration assay comparing proteins from gel filtration fractions D2, D5, and D12 from Figure 3.5C. Pre-processed (37 bp) spacer and untrimmed (57 bp) protospacer were both tested. Each fraction was tested alone, as well as with additional Cas1 and Cas2 supplemented into the reaction. The concentration ranges for each fraction tested were as follows: D2 – 60, 155, and 340 nM. D5 – 135, 340, and 740 nM. D12 – 415, 1040, and 2280 nM.

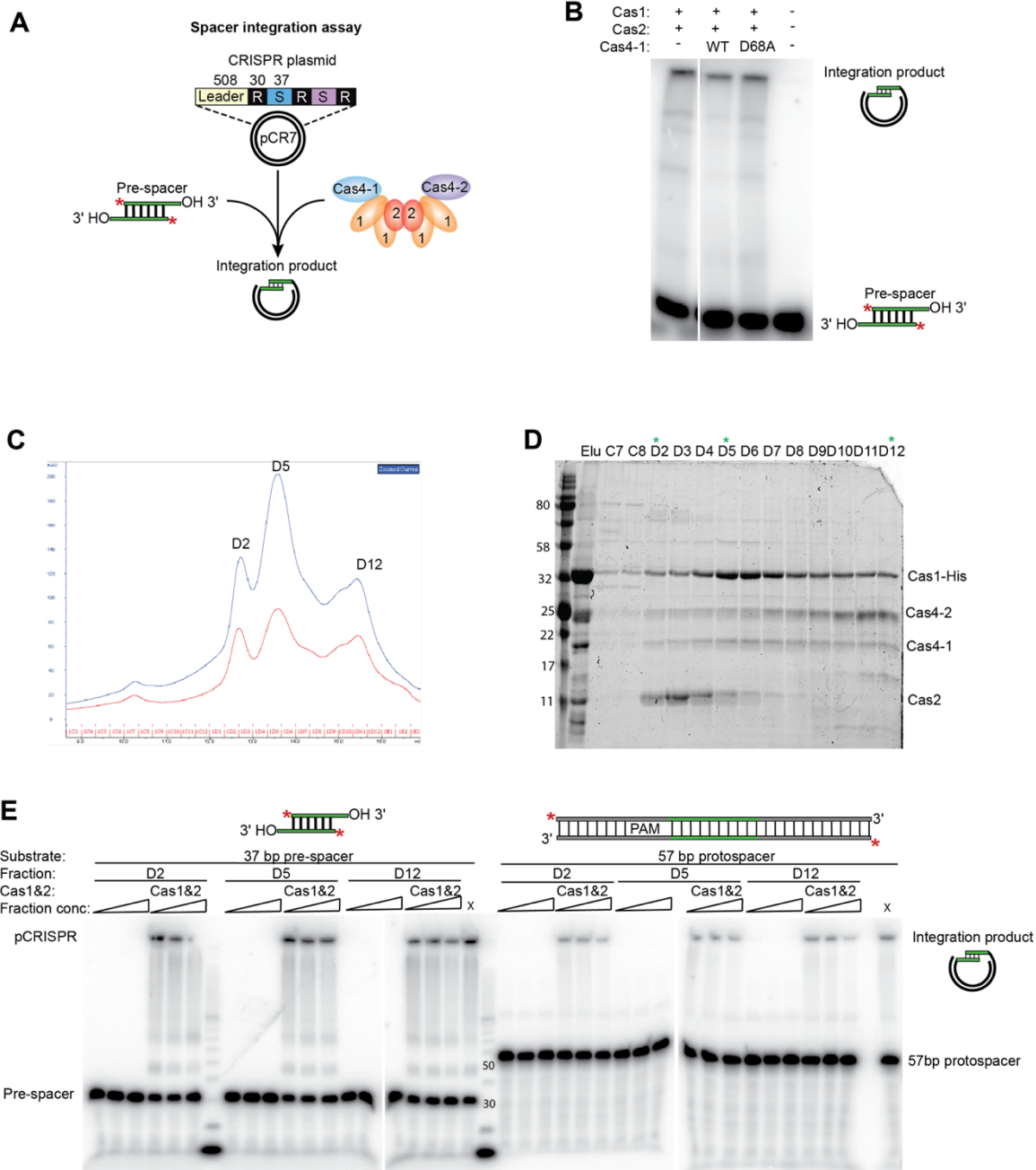


Figure 3.6. Cas4-1 and Cas4-2 do not trim protospacers before *in vitro* integration under conditions tested.

(A) Schematic of the PCR amplification of spacer integration products. Primer 1 sits on the top strand of the correctly oriented new spacer (S0), or the bottom strand of the inverse orientation new spacer. Primer 2 sits on the original spacer (S1) to amplify correctly oriented new spacers. Primer 3 sits in the leader to amplify new spacers that were integrated in the inverse orientation. Control plasmids containing correct or inverse orientation spacers were used as positive control templates for PCR amplification. (B,C) Correct (B) and inverse (C) orientation spacer integration by Cas1 and Cas2 was compared with or without the addition of Cas4-1. The pre-processed spacer (37) was compared to the untrimmed protospacer (57). PCR products were separated by 1% agarose gel electrophoresis and detected by ethidium bromide. (D) Schematic of the protospacer trimming and integration assay. A radiolabel located on the 3' end of the top strand of the CRISPR allowed for the detection of larger products upon spacer integration. The 69 nt CRISPR substrate will become a 96 nt product if a processed 37 nt spacer is integrated, or a 106 nt product if an untrimmed protospacer is integrated. Fractions from early (Fraction A), mid (Fraction B), and late (Fraction C) gel filtration of Figure 3.5C were pooled and concentrated to test in the assay. (E) Gel filtration fractions A, B, and C were tested for pre-spacer (37) integration, compared to Cas1 and Cas2. (F) Gel filtration fractions B and C were tested for protospacer (57) trimming and/or integration, compared to the integration of pre-spacers (37).

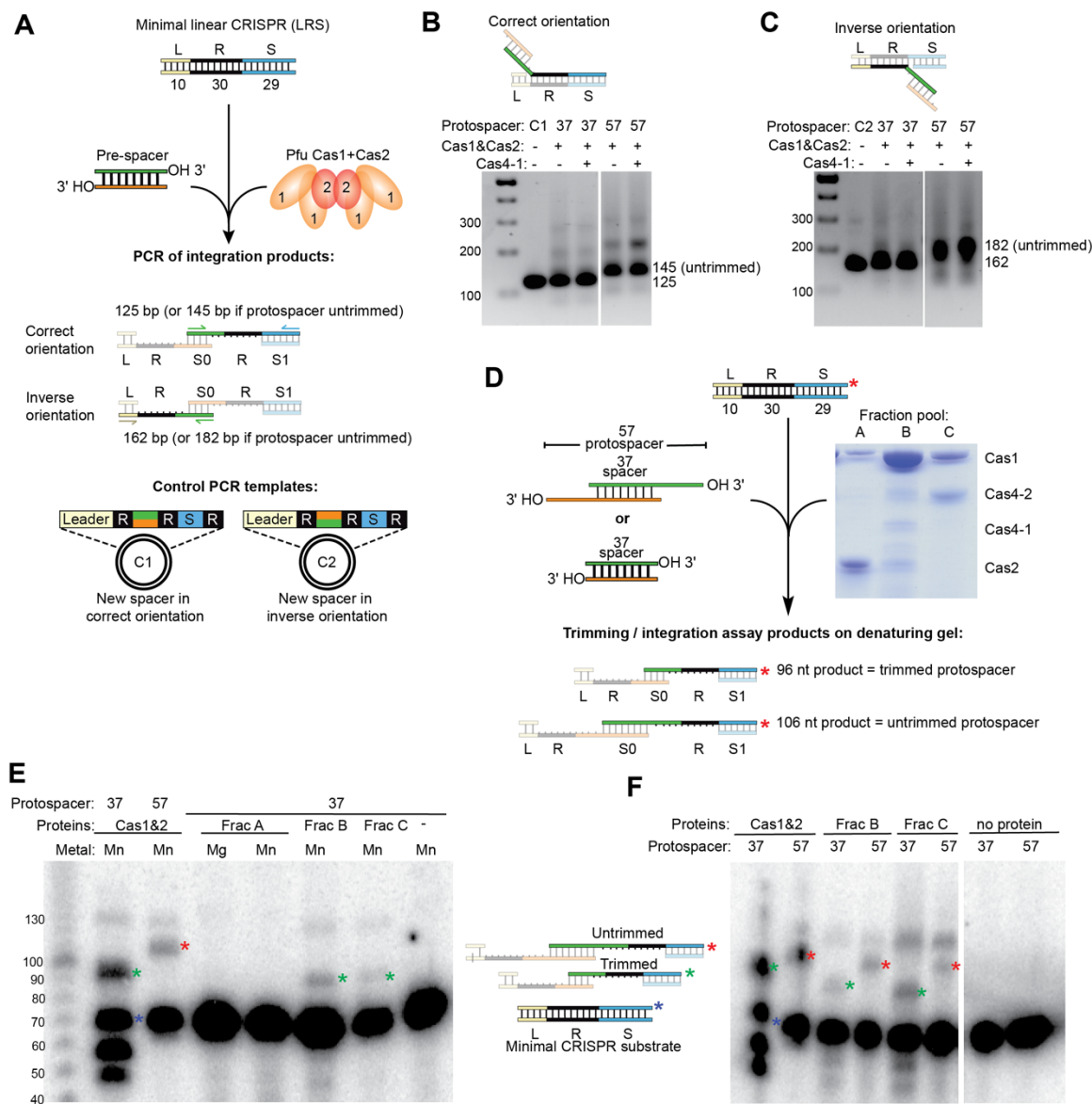
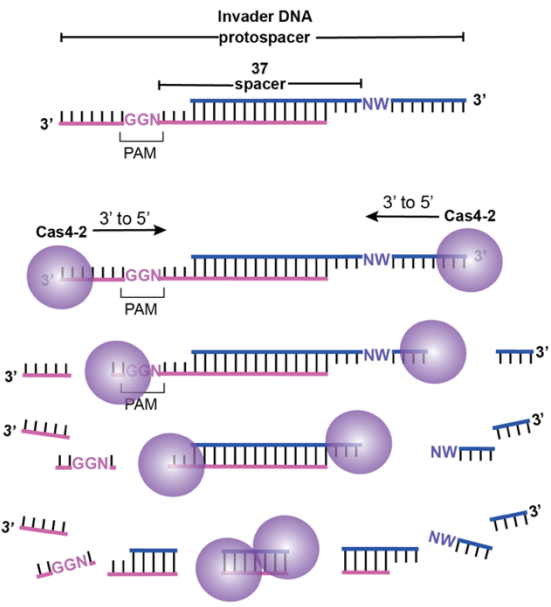


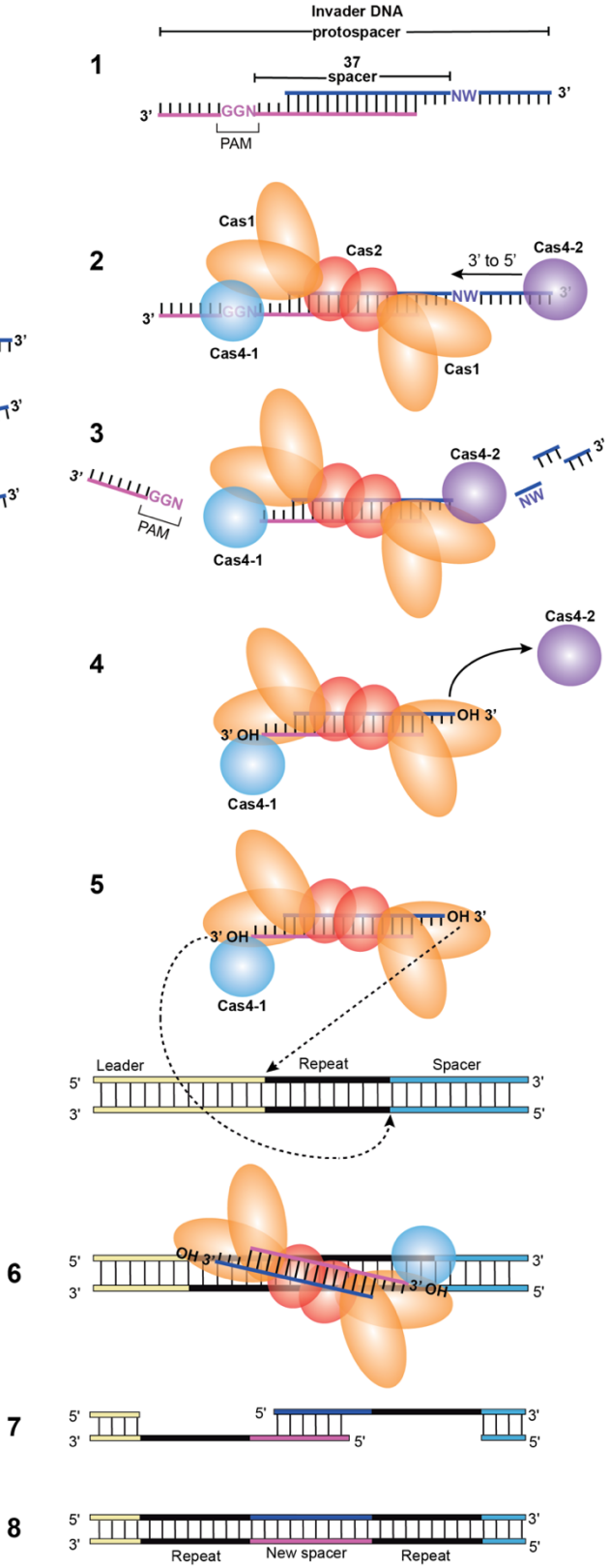
Figure 3.7. Model of *in vivo* Cas4-2 activity in the absence or presence of Cas1, Cas2, and Cas4-1.

(A) Model of unregulated Cas4-2 nuclease activity. Cas4-2 rapidly degrades DNA in a 3' to 5' direction. (B) Model of regulated Cas4-2 nuclease activity in the cell. 1) An example intermediate protospacer that might exist in the cell. 2) Cas4-2 rapidly degrades DNA in a 3' to 5' direction, but is stalled by a complex of Cas1, Cas2, and Cas4-1 bound at a PAM. 3) Once Cas4-2 is recruited to the complex, Cas4-1 cleaves the DNA at the site of the PAM that it is bound to. 4) The remaining pre-spacer fragment is now the appropriate size for spacer integration, dictated by the distance between Cas4-1 and Cas4-2 cleavage sites on the protospacer. At this point Cas4-2 may dissociate from the complex 5) The asymmetric complex leads to an oriented spacer integration at the CRISPR repeat. Cas1 in the absence of Cas4-2 is able to recognize the leader-repeat junction, with the Cas1 that is still tightly bound to Cas4-1 will recognize the leader-distal end of the repeat. 6) Properly oriented spacer integration proceeds. 7) The integrase complex dissociates, leaving single stranded repeats on both sides of the new spacer. 8) The single stranded repeats are filled in by host DNA polymerase and ligase.

A



B



CHAPTER 4

DEVELOPING TOOLS TO STUDY ADAPTATION OF A RNA TARGETING TYPE

III-A CRISPR-CAS SYSTEM *IN VITRO*¹

¹Julie Grainy. To be submitted to Nucleic Acids Research.

Abstract

The diversity of CRISPR-Cas prokaryotic defense mechanisms demands basic research into each unique variant. One particular component of CRISPR-Cas immunity that has remained largely uncharacterized is the adaptation of Type III-A systems capable of targeting RNA. During adaptation, new fragments of foreign nucleic acid are captured and integrated into the CRISPR locus. The system's ability to target RNA begs the question of if/how dsDNA spacers in the CRISPR locus are generated from ssRNA species. Here, we demonstrate the development and validation of the tools necessary to study spacer integration of a Type III-A system *in vitro*. For the first time, we purified recombinant Cas1 and Cas2 adaptation proteins from the Type III-A CRISPR-Cas system of *Streptococcus thermophilus* JIM8232, and developed an assay to test spacer integration. We found that Cas1 was required for spacer integration into a minimal CRISPR locus. Together, the results of this study demonstrate the utility of the developed tools for future investigation of Type III-A spacer acquisition.

Introduction

CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats)-Cas (CRISPR-associated proteins) systems are used by prokaryotes as a defense against invading viruses (1) and plasmids (2), and have been identified in about 90% of archaea and 50% of bacteria (3). Prokaryotes equipped with this defense mechanism contain a genomic CRISPR locus (or multiple loci in some cases) that harbors a series of DNA “spacers,” which are 30-40 bp sequences derived from the invading nucleic acid (1). The spacer DNAs are separated by 25-35 bp direct repeats for which the system is named (1). An AT

rich “leader” sequence is located directly upstream of the repeat-spacer array, and CRISPR-associated (*cas*) genes are typically clustered into one or more operons (3-5).

CRISPR-Cas immunity occurs in three defined stages (6,7): adaptation, CRISPR RNA (crRNA) biogenesis, and invader silencing. New spacers are captured and integrated into the CRISPR array during adaptation. During crRNA biogenesis, guide crRNAs for each spacer are produced from processing the RNA transcript of the CRISPR array (8-10). Invader silencing is achieved when effector Cas proteins are guided to the invader by the crRNA and prompted to cleave the nucleic acid (11-14).

CRISPR-Cas systems have diverse mechanisms to achieve immunity and have been classified according to signature Cas proteins (3,15,16). The highly characterized Type I and Type II systems are only capable of targeting DNA, while Type III systems have been of interest because of their ability to target RNA transcripts in addition to DNA. Beyond targeting RNA transcripts, Type III systems could provide immunity against RNA-based invaders including RNA bacteriophages. While Type III invader silencing mechanisms are steadily being elucidated, the process of adaptation in these systems remains elusive.

For Type I and Type II systems, Cas1 and Cas2 are required for adaptation *in vivo* (17-19). The active site of Cas1 performs the integration of spacers, while Cas2 architecturally bridges the Cas1 proteins (20,21). Spacer integration by Cas1 and Cas2 has also been reconstituted *in vitro* using pre-spacers as donor DNA and plasmid or linear DNA containing a CRISPR locus as acceptor DNA (21-25). While the details of spacer integration has been pioneered by Type I and II system, there is much insight to be gained from investigating Type III RNA targeting systems.

In a subset of Type III CRISPR-Cas systems Cas1 is fused to a reverse transcriptase (RT) (26-31). When the Type III-B system from *Marinomas mediterranea* was studied *in vivo*, it was found that the Cas1-RT fusion is required for acquisition of RNA spacers (30). That same study demonstrated *in vitro* reverse transcription of RNA spacers following ligation of the RNA into the CRISPR array (30). In another subset of Type III systems an individual RT gene is found adjacent to *cas1*, rather than fused, and is expected to lead to the same mechanism of RNA spacer acquisition (28,31).

For Type III systems in organisms that lack this RT, it remains to be seen if/how spacers are acquired from RNA species. The model organism *Streptococcus thermophilus* JIM8232 is a good candidate for studying adaptation by a Type III system lacking a RT. The strain contains a 17 spacer CRISPR array connected with Type III-A Csm invader silencing genes (Figure 4.1A). When expressed in *E. coli*, the crRNA biogenesis protein (Cas6) and the effector proteins (Csm1-Csm5 and Csm6-1) from JIM8232 can achieve invader silencing (32,33). The CRISPR locus also contains the conserved adaptation proteins, Cas1 and Cas2, expected to be involved in spacer integration. There is no RT linked to this CRISPR-Cas system, providing an opportunity to investigate spacer acquisition of an RT-lacking system capable of RNA interference.

This investigation aims to provide and define the necessary tools to study Type III-A CRISPR-Cas spacer integration *in vitro*. The utility of the tools developed here is demonstrated by preliminary evidence showing metal-dependent spacer integration by Cas1.

Materials and methods

Cloning, expression, and purification

The *cas1* and *cas2* genes were amplified from *Streptococcus thermophilus* JIM8232 genomic DNA and individually cloned into the following expression vectors: pET24d for N-terminally His tagged proteins, pET21d for C-terminally His tagged proteins or untagged proteins, and pSAT1 for N-terminal His-SUMO tagged proteins. Co-expression constructs were created by cloning the genes including the T7 promoter and T7 terminator regions from the individual expression vectors into the pET24d backbone.

Expression vectors were transformed into the *E. coli* BL21 star strain. Cultures were grown at 37°C in Luria broth to an OD₆₀₀ of 0.4-0.6, and expression of the proteins was induced with IPTG during overnight growth at 16°C. Unless otherwise stated, harvested cells were lysed in 20 mM Tris, 500 mM NaCl, 10% glycerol, 10 mM imidazole, pH 7.5. The cell lysate was centrifuged at 14,000 rpm at 4°C for 30 minutes, and the soluble fraction was collected and filtered (0.8 µm filter pore size Millex filter unit, Millipore). His-tagged proteins were purified by Ni²⁺ affinity column chromatography, using a stepwise increase of imidazole (10, 20, 100, 250, 500 mM) in 20 mM Tris, 500 mM NaCl, 10% glycerol, 10 mM imidazole, pH 7.5.

His-SUMO tag removal

Peak elution fractions (250 mM imidazole) were dialyzed using Slide-a-lyzer mini dialysis cassettes (Thermo Fisher) into 20 mM Tris, 500 mM NaCl, 10% glycerol, pH 7.5. During dialysis, SUMO-protease was mixed with the protein in a 1:10 molar ratio

(protease:protein). After protease treatment, the proteins were subjected to a second Ni^{2+} affinity chromatography column in order to separate untagged protein from the protease, cleaved tag, and any un-cleaved protein remaining. This was achieved by stepwise increases in the concentration of imidazole (20, 50, 100, 250, 500 mM) in 20 mM Tris, 500 mM NaCl, 10% glycerol, pH 7.5. Fractions with untagged protein were stored at 4°C prior to use for functional assays.

DNA substrate preparation

DNA oligonucleotides were from Eurofins Genomics. Oligonucleotides used to make pre-spacers and CRISPR substrates were separated by 15% denaturing polyacrylamide gel electrophoresis in 1x TBE, detected by ethidium bromide staining, and the band of the expected oligonucleotide size was excised. Oligonucleotides were eluted from the gel slices overnight at 4°C in 500 μL of elution buffer (0.5 M ammonium acetate, 1 mM EDTA (pH 8.0), and 0.1% SDS), extracted with phenol/chloroform/isoamyl alcohol (pH 8.0), ethanol precipitated, and resuspended in 20 mM Tris, 100 mM KCl, 5% glycerol, pH 7.5. Corresponding oligonucleotides were annealed by incubating at 95°C for 5 minutes followed by slow cooling until 23°C. Annealing was confirmed by 10% non-denaturing polyacrylamide gel electrophoresis in 1x TBE. Pre-spacers were 5' radiolabeled with T4 polynucleotide kinase (PNK) and $[\gamma\text{-}^{32}\text{P}]$ ATP.

Spacer integration assay

Unless stated otherwise, a final concentration of 1 μM Cas1, 1 μM Cas2, 1 mM DTT, and 10 mM MnCl_2 was incubated in reaction buffer (20 mM Tris, 100 mM NaCl, 5%

glycerol, pH 7.5) at 4°C for one hour. 20 nM of radiolabeled pre-spacer was added to the reactions and incubated at 4°C for an additional 15 minutes. Finally, plasmid or linear DNA was added to a final concentration of 5 nM for pCR, or 100 nM for minimal CRISPR substrate, and the 20 µL reaction was incubated at 37°C for one hour. Reactions with pCR were quenched with 25 mM EDTA and 1 µL of proteinase K (Life Technologies). 5 µL of the reaction products were mixed with gel loading dye (purple, NEB) and separated on 1% agarose gels in 1x TAE. DNA was detected with ethidium bromide staining and gels were subsequently dried for radioactivity detection by phosphorimaging (Storm 840 Scanner GE Healthcare). 5 µL of the reactions with minimal CRISPR substrates were quenched with an equal volume of Gel Loading Buffer II (Thermofisher) and 25 mM EDTA. Samples were boiled for 5 minutes before separation by 15% denaturing 7M urea-containing polyacrylamide gel electrophoresis in 1x TBE. Gels were dried and radioactivity was detected with phosphorimaging (Storm 840 Scanner GE Healthcare).

Results and Discussion

Building expression vectors for recombinant Cas1 and Cas2 from *S. thermophilus*

The Type III-A CRISPR-Cas system of *S. thermophilus* JIM8232 contains two genes predicted to be involved in adaptation, *cas1* and *cas2* (Figure 4.1A). The *cas1* and *cas2* genes were amplified from *S. thermophilus* genomic DNA and cloned into a variety of expression vectors by seamless cloning. The 15 constructs that were built include individually expressed proteins containing N- and C-terminal 6x His-tags or His-SUMO

tags, as well as co-expressed Cas1 and Cas2 with one protein tagged and the other untagged (Figure 4.1B). The co-expression constructs were built to examine potential pull-down interactions during protein purification, if Cas1 and Cas2 form a stable complex. In some studied systems, Cas1 and Cas2 form a stable complex when mixed together (21,25,34), while others require the presence of DNA to bring them together (35). If a complex does form between the two proteins of this Type III-A system, it is expected to be in a Cas1₄-Cas2₂ configuration with a dimer of Cas2 flanked by two dimers of Cas1, as previously reported (21,25,34).

Protein purification of Cas1 and Cas2

The co-expression constructs were analyzed by small-scale (10 mL culture) expression and solubility tests, and in every case both Cas1 and Cas2 were largely insoluble (Figure 4.2A and 4.2B and data not shown for other five co-expression constructs). To optimize the solubility of the proteins, a series of lysis buffers was tested, including (a) 50 mM NaH₂PO₄ with 300 mM NaCl pH 8, (b) 40 mM Tris with 500 mM KCl pH 7.5, and (c) 20 mM Tris with 500 mM NaCl pH 7.5. Each of these base buffers also contained 10% glycerol and 10 mM imidazole. Additionally, each base buffer was tested with the presence or absence of 5 mM βME or 0.5% Triton X-100. Cas1 and Cas2 were expressed in 100 mL cultures and each divided into nine portions to be treated individually with the nine lysis buffer variations. Three samples were taken and analyzed by SDS-PAGE: total cell lysate after sonication, and insoluble and soluble fractions after high speed centrifugation (Figure 4.2C and 4.2D). It was determined that the optimal buffer for Cas1 was 20 mM Tris, 500 mM NaCl, 10% glycerol, 10 mM imidazole, pH 7.5 (Figure 4.2C).

The addition of β ME or Triton X-100 did not appear to make a difference in this buffer, and was left out of future lysis and purification buffers. Cas2 remained ~80% insoluble regardless of which buffer was being used (Figure 4.2D).

Cas1 and Cas2 were individually expressed with 0.001 mM IPTG in a 1 liter culture and purified by Ni^{2+} affinity column chromatography with the Tris-NaCl buffer. Cas1 N-terminal His-SUMO had robust purification, with peak elution at the 250 mM imidazole wash (Figure 4.2E). Cas2 N-terminal His-SUMO exhibited moderate purification, with peak elution at the first 500 mM imidazole wash (Figure 4.2F). In an attempt to achieve a more robust purification of Cas2, individual Cas2 proteins with N- or C-terminal 6x His tags were also tested under the same conditions, but had low expression and low solubility (data not shown). The His-SUMO tag constructs were designed so that the tag could be cleaved from the protein post-purification using a SUMO-protease. Cleavage of the Cas1 tag was only ~20% efficient, and it proved difficult to separate the remaining tagged protein from the untagged protein (Figure 4.2G). Cleavage of Cas2 appeared to be ~95% efficient, but also proved difficult to separate from the protease and cleaved tag (Figure 4.2H). Moreover, it was difficult to accurately identify the untagged Cas2 protein, as it appeared to run about the same molecular weight as the cleaved tag on an SDS-PAGE. It is also possible that the untagged Cas2 protein is unstable after tag cleavage and that the majority of the band detected at that molecular weight is actually just the cleaved tag. Further optimization of protein purification is recommended in the future in order to be confident in the integrity of Cas2.

Developing a spacer integration assay

One method of detecting spacer integration into the CRISPR array *in vitro* is to radiolabel the incoming pre-spacer so that the unlabeled CRISPR containing substrate becomes radioactive when the pre-spacer has successfully integrated. Part of the CRISPR locus from the *S. thermophilus* Type III-A system was cloned into a plasmid (pCR). The cloned fragment contained the full leader sequence plus two repeats, separated by one natural spacer (Figure 4.3A). The spacer integration assay combines pCR with the radiolabeled pre-spacer, Cas1 and Cas2, and a divalent metal predicted to sit in the active site of the Cas1 integrase (Figure 4.3A). The resulting integration product can be visualized on an agarose gel, running at the speed of the relaxed form of the plasmid as opposed to the supercoiled form.

An alternative approach is to use a minimal linear substrate, which in theory should yield distinctly sized integration products depending on whether the pre-spacer integrates at the leader-repeat junction (LR) or the repeat-spacer junction (RS). A minimal linear CRISPR was designed and produced by annealing two oligonucleotides together. The resulting dsDNA contained 30 bp of the 101 bp leader, one entire repeat, and the first 11 bp of natural spacer one (Figure 4.3B). As a result, when reaction products are run on a denaturing PAGE, spacer integration at the LR junction is represented by a 78 nt product, while spacer integration at the RS junction is a 97 nt product (Figure 4.3C).

When designing the pre-spacers to test in the integration assay, several variations were included in order to identify which would be most favorable. The natural spacers of the *S. thermophilus* Type III-A CRISPR-Cas system range from 35 to 39 bp. Two

different spacer sequences were chosen to test, one being 36 bp and the other being 38 bp. Both of these sequences originated from the pWAR plasmid at protospacer sites that were highly targeted *in vivo*. Variations of each of these sequences were designed as well, to test blunt end or ssDNA overhangs (Figure 4.3D). In case the flanking sequences are required for spacer integration, a protospacer was also designed with 10 bp upstream and downstream of the 38 bp spacer sequence.

Spacer integration by Cas1 requires MnCl₂

Spacer integration into pCR by Cas1 and Cas2 was detected in the presence of MnCl₂ with the 36 nt 3' overhang pre-spacer, while MgCl₂ was not sufficient in the reaction (Figure 4.4A). When the individual proteins were tested, Cas1 alone was capable of integration, while Cas2 was not (Figure 4.4B). The integration activity of Cas1 alone was comparable to that detected in the presence of both proteins, indicating that the integration products observed with Cas1 and Cas2 might be due to Cas1 alone. One explanation for this could be the ambiguity of the Cas2 purification. Another explanation could be that Cas2 is present, but the conditions necessary for integrase complex formation were not met. Although it is possible that spacer integration is occurring *in vivo* via Cas1 alone, this appears unlikely given that Cas2 is essential in previously studied systems (21,36).

When the full range of pre-spacer variations were tested in the spacer integration assay, the 36 nt and 38 nt spacers containing 3' ssDNA overhangs were preferred in the reaction (Figure 4.4C). The 58 bp protospacer did not appear to have an integration product running at the size of the nicked pCR, but it did create a smear down the whole

lane of the agarose gel, indicating that Cas1 and Cas2 might be bound tightly to the reaction substrates, which is not seen to that same degree with the other pre-spacers. The same pre-spacer preference for 3' overhangs was observed when using the minimal linear CRISPR substrates (Figure 4.4D). In contrast to the lack of distinct integration product with the 57 bp protospacer in the plasmid assay, the minimal linear CRISPR assay resulted in a clear integration product with the protospacer. The size of the product relative to that of the pre-processed spacers that were integrated into the linear CRISPR suggest that the protospacer might be getting processed before or during spacer integration. Further detailed analysis of the integration products, including sequencing, will be required to learn more about the details of this reaction.

Conclusion

This report provides the fundamental tools required to study Type III-A spacer integration *in vitro*. The effectiveness of the tools is evidenced by successful spacer integration observed by Cas1. With continued optimization and additional experiments, these tools can lead to a detailed characterization of spacer integration of a Type III-A CRISPR-Cas system.

References

1. 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.

2. Marraffini, L.A. and Sontheimer, E.J. (2008) CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science*, **322**, 1843-1845.
3. Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J., Barrangou, R., Brouns, S.J., Charpentier, E., Haft, D.H. *et al.* (2015) An updated evolutionary classification of CRISPR-Cas systems. *Nature reviews. Microbiology*, **13**, 722-736.
4. Kunin, V., Sorek, R. and Hugenholtz, P. (2007) Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome biology*, **8**, R61.
5. Makarova, K.S., Wolf, Y.I. and Koonin, E.V. (2013) The basic building blocks and evolution of CRISPR-CAS systems. *Biochemical Society transactions*, **41**, 1392-1400.
6. Makarova, K.S., Aravind, L., Wolf, Y.I. and Koonin, E.V. (2011) Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biology direct*, **6**, 38.
7. van der Oost, J., Jore, M.M., Westra, E.R., Lundgren, M. and Brouns, S.J. (2009) CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends in biochemical sciences*, **34**, 401-407.
8. Brouns, S.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J., Snijders, A.P., 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.

9. 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 & development*, **22**, 3489-3496.
10. 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.
11. 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.
12. 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.
13. Jore, M.M., Lundgren, M., van Duijn, E., Bultema, J.B., Westra, E.R., Waghmare, S.P., Wiedenheft, B., Pul, U., Wurm, R., Wagner, R. *et al.* (2011) Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nature structural & molecular biology*, **18**, 529-536.
14. 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**, E2579-2586.
15. Koonin, E.V., Makarova, K.S. and Zhang, F. (2017) Diversity, classification and evolution of CRISPR-Cas systems. *Current opinion in microbiology*, **37**, 67-78.

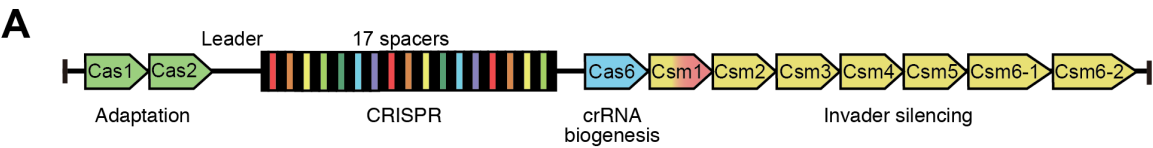
16. Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J., Wolf, Y.I., Yakunin, A.F. *et al.* (2011) Evolution and classification of the CRISPR-Cas systems. *Nature reviews. Microbiology*, **9**, 467-477.
17. Datsenko, K.A., Pougach, K., Tikhonov, 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.
18. Swarts, D.C., Mosterd, C., van Passel, M.W. and Brouns, S.J. (2012) CRISPR interference directs strand specific spacer acquisition. *PloS one*, **7**, e35888.
19. 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.
20. 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.
21. Nunez, 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. *Nature structural & molecular biology*, **21**, 528-534.
22. Nunez, J.K., Lee, A.S., Engelman, A. and Doudna, J.A. (2015) Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature*, **519**, 193-198.

23. Wright, A.V. and Doudna, J.A. (2016) Protecting genome integrity during CRISPR immune adaptation. *Nature structural & molecular biology*, **23**, 876-883.
24. 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.
25. 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.
26. Makarova, K.S., Grishin, N.V., Shabalina, S.A., Wolf, Y.I. and Koonin, E.V. (2006) A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biology direct*, **1**, 7.
27. Toro, N. and Nisa-Martinez, R. (2014) Comprehensive phylogenetic analysis of bacterial reverse transcriptases. *PloS one*, **9**, e114083.
28. Toro, N., Martinez-Abarca, F. and Gonzalez-Delgado, A. (2017) The Reverse Transcriptases Associated with CRISPR-Cas Systems. *Sci Rep*, **7**, 7089.
29. Toro, N., Martinez-Abarca, F., Gonzalez-Delgado, A. and Mestre, M.R. (2018) On the Origin and Evolutionary Relationships of the Reverse Transcriptases Associated With Type III CRISPR-Cas Systems. *Frontiers in microbiology*, **9**, 1317.

30. Silas, S., Mohr, G., Sidote, D.J., Markham, L.M., Sanchez-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.
31. 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.* (2017) On the Origin of Reverse Transcriptase-Using CRISPR-Cas Systems and Their Hyperdiverse, Enigmatic Spacer Repertoires. *mBio*, **8**, e00897-00817.
32. Ichikawa, H.T., Cooper, J.C., Lo, L., Potter, J., Terns, R.M. and Terns, M.P. (2017) Programmable type III-A CRISPR-Cas DNA targeting modules. *PloS one*, **12**, e0176221.
33. Foster, K., Kalter, J., Woodside, W., Terns, R.M. and Terns, M.P. (2019) The ribonuclease activity of Csm6 is required for anti-plasmid immunity by Type III-A CRISPR-Cas systems. *RNA biology*, **16**, 449-460.
34. 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**, 1113-1118.
35. Lee, H., Dhingra, Y. and Sashital, D.G. (2019) The Cas4-Cas1-Cas2 complex mediates precise prespacer processing during CRISPR adaptation. *eLife*, **8**, e44248.
36. Jackson, S.A., McKenzie, R.E., Fagerlund, R.D., Kieper, S.N., Fineran, P.C. and Brouns, S.J. (2017) CRISPR-Cas: Adapting to change. *Science*, **356**, eaal5056.

Figure 4.1. Building expression vectors for recombinant Cas1 and Cas2 from *S. thermophilus*

(A) Organization of the Type III-A CRISPR locus in *S. thermophilus*. **(B)** List of plasmids constructed by cloning *cas1* and *cas2* genes into expression vectors.



B

Tagged protein			
Construct #	Plasmid	Cas1	Cas2
1	pXZ181	N-terminal His	
2	pXZ183	C-terminal His	
3	pXZ185	N-terminal His-SUMO	
4	pXZ187	Untagged	
5	pXZ182		N-terminal His
6	pXZ184		C-terminal His
7	pXZ186		N-terminal His-SUMO
8	pXZ188		Untagged
9	pJG72	N-terminal His	Untagged
10	pJG73	C-terminal His	Untagged
11	pJG74	N-terminal His-SUMO	Untagged
12	pJG75	Untagged	N-terminal His
13	pJG76	Untagged	C-terminal His
14	pJG77	Untagged	N-terminal His-SUMO
15	pJG78	Untagged	Untagged

Figure 4.2. Protein purification of Cas1 and Cas2

(A) Cas2 with an N-terminal His-SUMO tag co-expressed with untagged Cas1 and analyzed for expression and solubility on SDS-PAGE. (B) Cas2 with a C-terminal His tag co-expressed with untagged Cas1 and analyzed for expression and solubility on SDS-PAGE. (C,D) Expression and solubility of Cas1 (C) and Cas2 (D) tested with a range of lysis buffers. (E,F) Protein purification of Cas1 (E) and Cas2 (F) by Ni^{2+} -affinity column chromatography. (G,H) His-SUMO tag cleavage of Cas1 (G) and Cas2 (H). Samples taken before and after protease treatment (“Pre” and “Post”). Protease treated protein was subjected to a second Ni^{2+} -affinity chromatography column, in an attempt to separate tagged proteins (including the protease, un-cleaved Cas protein, and the cleaved tag) from the untagged proteins. Samples were taken after each step-wise increasing imidazole wash and analyzed by SDS-PAGE.

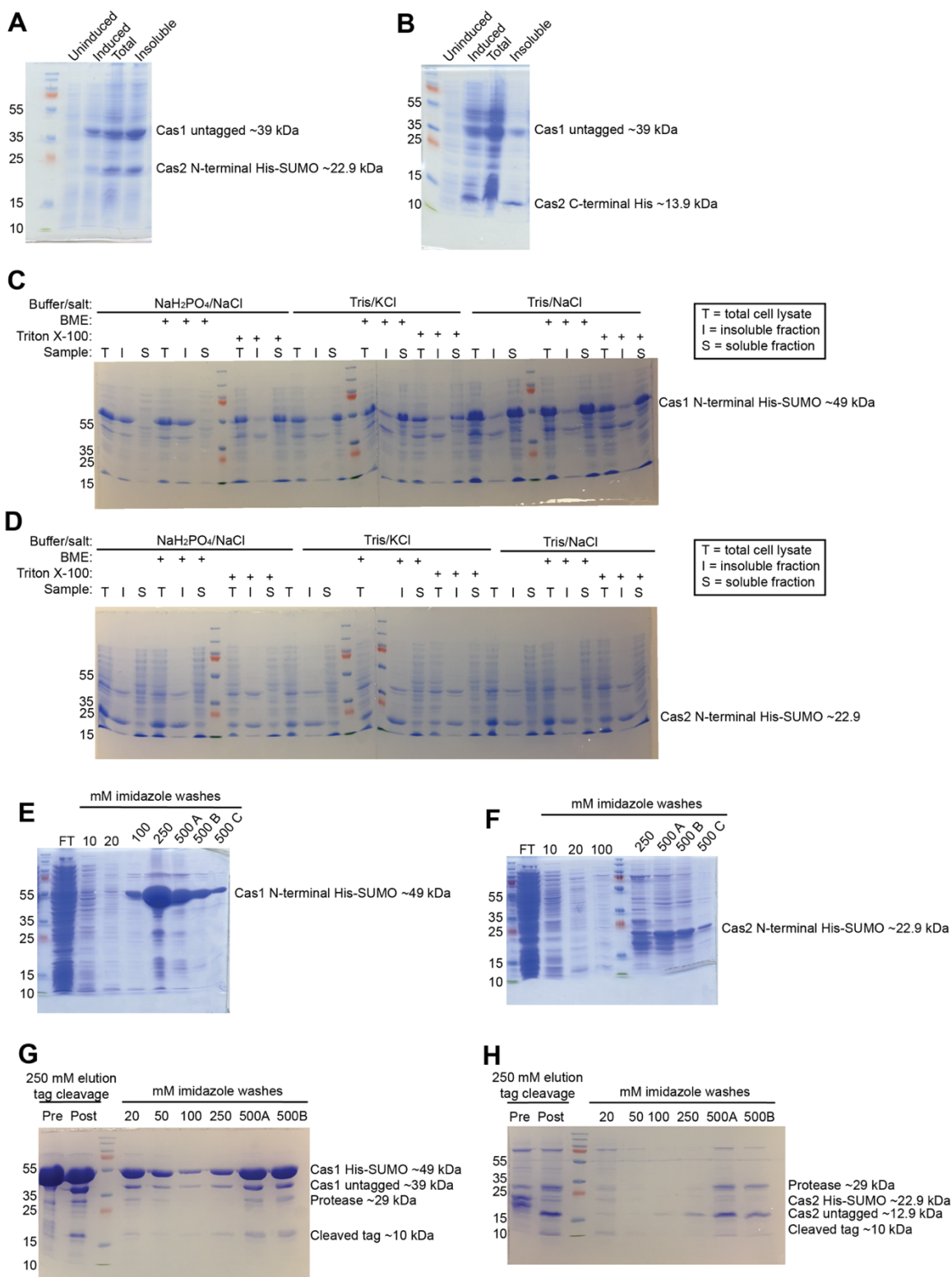


Figure 4.3. Developing a spacer integration assay

(A) Schematic of the spacer integration assay using a plasmid containing a minimal CRISPR locus from the *S. thermophilus* Type III-A CRISPR-Cas system, including a full leader (101 bp), and two 36 bp repeats, separated by one 39 bp natural spacer. The CRISPR plasmid (pCR) is incubated with pre-spacer DNA and Cas1 and Cas2. A 5' radiolabel on the pre-spacer allows for detection of labeled DNA when reaction products are separated on a 1% agarose gel and visualized by autoradiography. pCR will only be detected by autoradiography if the labeled pre-spacer has successfully integrated into the plasmid. (B) Minimal linear CRISPR substrate, with sites of expected spacer integration identified. (C) Schematic of the spacer integration assay using the minimal linear CRISPR substrate. When a 5' radiolabeled pre-spacer that is 36 nt with 5 nt 3' ssDNA overhangs, the resulting products are expected: Leader-repeat (LR) integration products are expected to be 78 nt of a denaturing PAGE, while repeat-spacer (RS) integration products are expected to be 97 nt. (D) Series of pre-spacer and protospacer variations to test in the spacer integration assays.

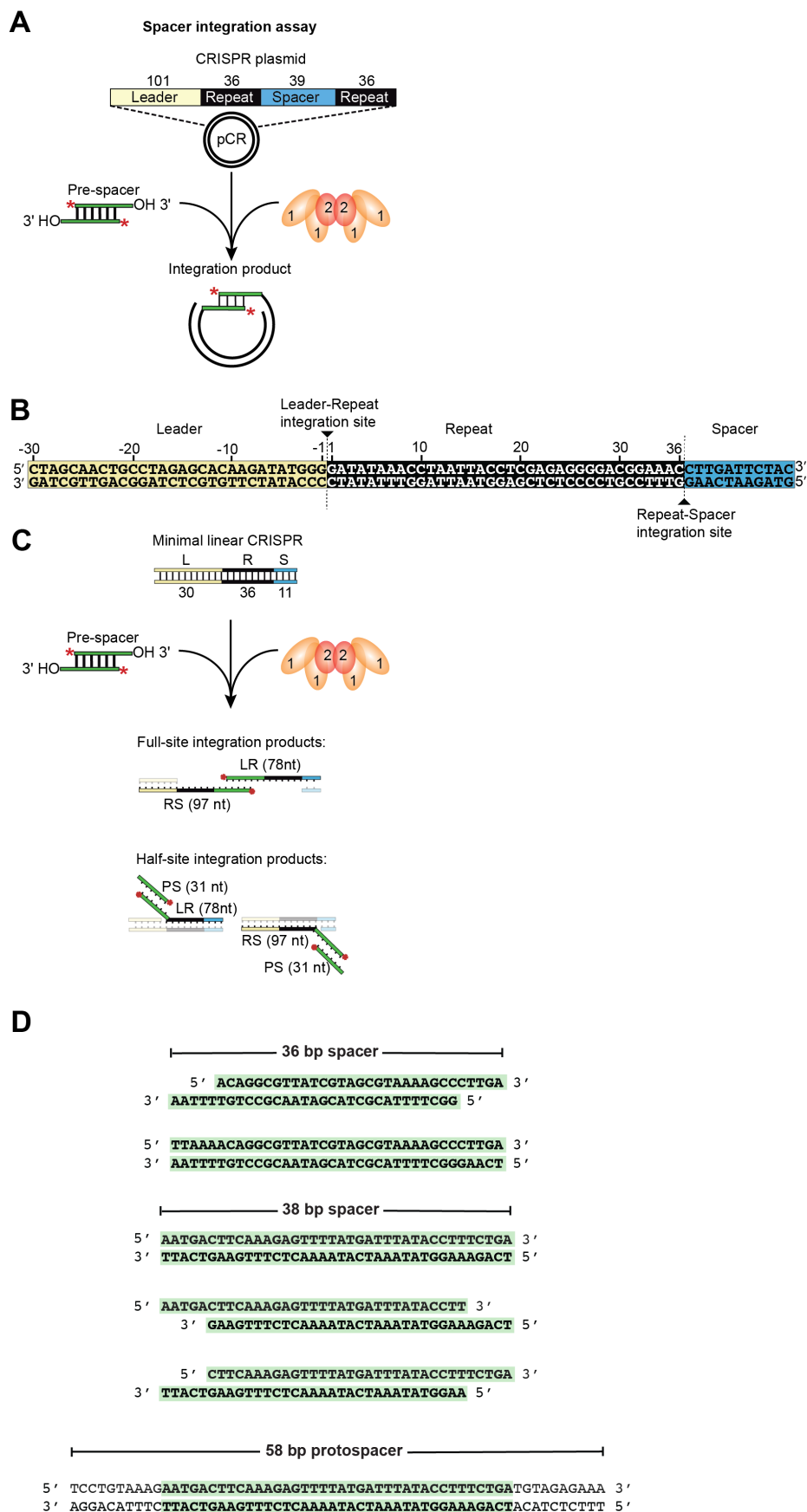
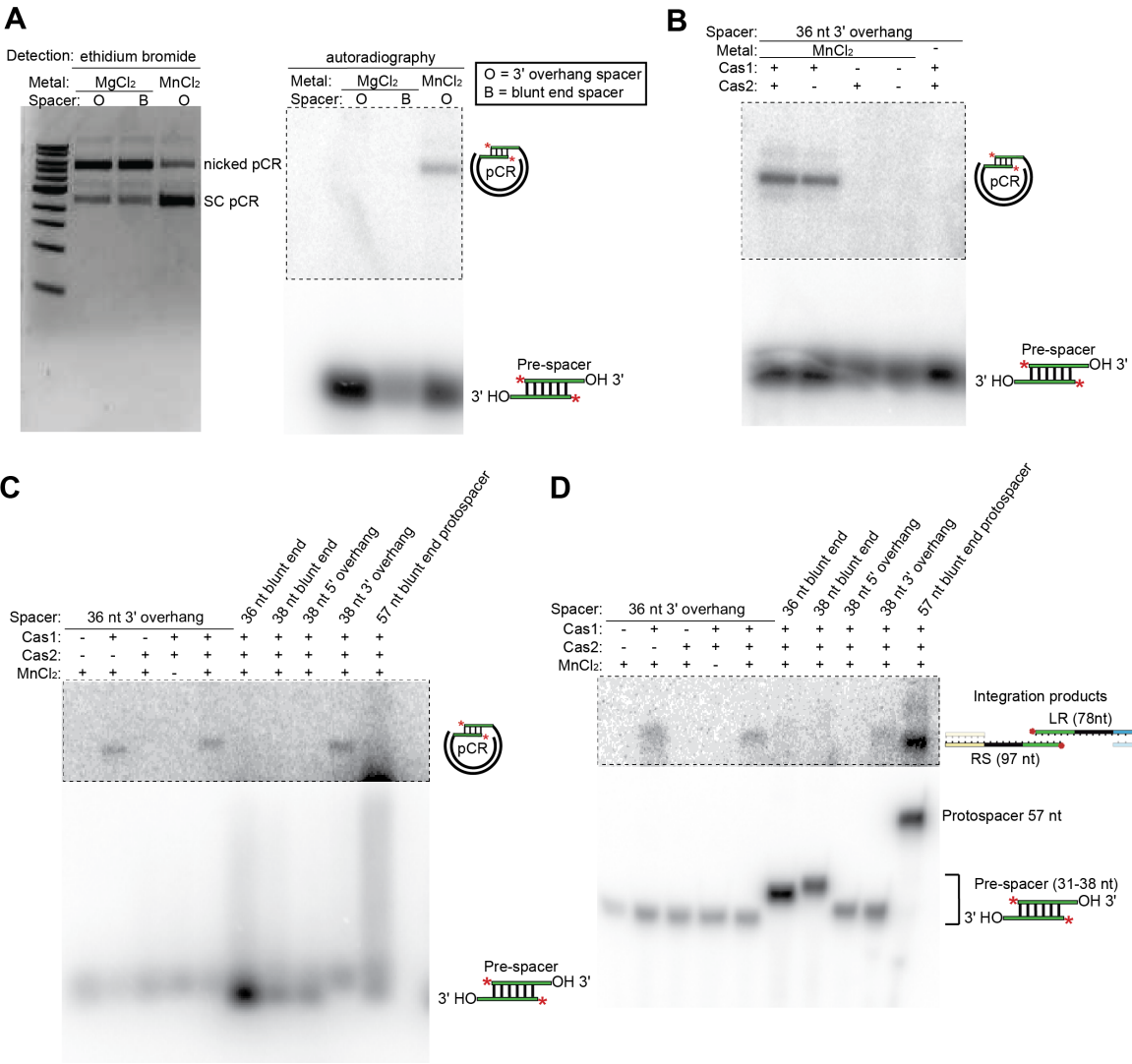


Figure 4.4. Spacer integration by Cas1 requires MnCl₂

(A) Plasmid spacer integration assay products separated by 1% agarose and detected by both ethidium bromide (left) and autoradiography (right). All reactions contained Cas1 and Cas2, but varied in their metal and pre-spacer contents. The metals tested included 10 mM MgCl₂ and 10 mM MnCl₂. The pre-spacers tested included 36 nt with blunt ends, and 36 nt with 5 nt 3' ssDNA overhangs. The region of the gel boxed by dotted lines indicates adjusted levels in order to visualize integration products. (B) Plasmid spacer integration assay comparing Cas1&Cas2, Cas1 alone, and Cas2 alone. Products were separated by 1% agarose and detected by autoradiography. The region of the gel boxed by dotted lines indicates adjusted levels in order to visualize integration products. (C) Plasmid spacer integration assay comparing a range of pre-spacers. Products were separated by 1% agarose and detected by autoradiography. The region of the gel boxed by dotted lines indicates adjusted levels in order to visualize integration products. (D) Linear CRISPR spacer integration assay comparing a range of pre-spacers. Products were separated by 15% denaturing PAGE and detected by autoradiography. The region of the gel boxed by dotted lines indicates adjusted levels in order to visualize integration products.



CHAPTER 5

DISCUSSION

Diversity of CRISPR-Cas adaptation mechanisms

The presented studies have addressed the molecular mechanisms of adaptation for two distinct CRISPR-Cas systems. However, this work represents a small fraction of the diverse adaptation mechanisms previously examined as well as those yet to be explored. The central integrase function of Cas1 has been consistent across all systems studied to date, but beyond that there is much diversity in protein interactions and involvements, including other Cas and non-Cas host proteins. While early adaptation studies examined the Type I-E system, consisting of only Cas1 and Cas2, we have since begun to understand how systems with additional adaptation proteins achieve the same goal of specific spacer integration.

While our understanding of the physical and functional interaction of Cas4 with Cas1 and Cas2 is growing, there is still much diversity between Cas4-containing systems. For example, the system in *P. furiosus* has two Cas4 proteins that coordinate separate essential functions (1). On the other hand, the Type I-C and I-D systems utilize a single Cas4 protein (2-4), and some systems have a Cas4-Cas1 fusion (5). It remains to be seen just how similar the molecular mechanisms will be between the three groups. Having two distinct Cas4 proteins leads to a convenient model for orientation of the spacer due to an

asymmetrical adaptation complex. It will be interesting to see how spacer orientation is achieved in systems with only one Cas4.

There have been striking differences between Type I and Type II spacer acquisition mechanisms, and it is expected that Type III will reveal its own unique properties as well. The unique potential for incorporating ssRNA has already been demonstrated by Type III Cas1 fused to reverse transcriptase (6). It will be exciting to see more and more Type III systems investigated to dissect the molecular mechanisms of this unique capability.

Non-Cas proteins have become increasingly important for CRISPR-Cas adaptation, and the progress to identify these contributing factors and their specific roles is slow but steady. Sometimes there are host proteins that are not required, but enhance, regulate, or specify adaptation in one way or another. Through a series of systematic genetic and biochemical approaches, important host factors have been identified and their molecular contributions characterized. These approaches will need to continue in order to expand the ever-growing model of adaptation.

Pre-spacer generation

When a foreign nucleic acid enters a cell, the steps leading up to spacer integration are still quite enigmatic. Progress has been made in identifying Cas proteins that are involved in PAM recognition and subsequent cleavage, but a host of questions still remain surrounding the process of pre-spacer generation. Many CRISPR-Cas systems have adaptation proteins beyond the Cas1 and Cas2 proteins responsible for spacer integration. A large number of systems have Cas4 proteins, but we have only just started to touch on

the specific roles these proteins have. I found evidence to support the physical interaction between Cas1/Cas2/Cas4-1/Cas4-2 (Chapter 3), which fortifies the *in vivo* reports of Cas4's involvement in adaptation (1). The *in vitro* mechanism of PAM cleavage by Cas4 has started to be evaluated (2,3), and I look forward to seeing the detailed mechanisms of pre-spacer generation unfold as further experimentation continues in a variety of Cas4-containing CRISPR-Cas systems.

Similarly, the Type II-A systems contain Csn2, which is predicted to be involved in adaptation at some level. It has been shown to be required for spacer acquisition *in vivo* (7-9), but the specific contribution is largely unknown. Cas9, the invader-silencing effector nuclease of Type II-A systems, has also been reported to contribute to adaptation by conferring PAM specificity (8-10). The four proteins, Cas1/Cas2/Csn2/Cas9, have been shown to physically interact (8,11,12). Cas9 has not been part of the complex for any of the solved structures to date, but there is evidence that it interacts with Csn2 (11). A cryo-EM structure of Cas1/Cas2/Csn2 revealed a Cas1₈-Cas2₄-Csn2₈ stoichiometry containing ~30 bp duplex DNA bound through a channel in the central region of the complex (12). As the DNA length in the structure is similar to the pre-spacer length used by Type II-A Cas1 and Cas2 (13), a model for spacer length determination was suggested. This observation, combined with RecBCD previously implicated in pre-spacer generation (14), lead the authors to a potential model in which Cas1-Cas2-Csn2 protects the spacer DNA from degradation by RecBCD (or AddAB) (12).

The characterization of Cas proteins beyond Cas1 and Cas2 is providing insight into the pre-spacer generation step of adaptation. Experimental focus on non-Cas proteins

predicted to be involved, such as RecBCD, will provide equally important insight into these mechanisms and offer a bigger picture of adaptation as a whole.

Spacer integration by Cas1 and Cas2

After the invader DNA has been selected and processed into an adequate size, the spacer must be integrated precisely into the CRISPR locus. I found the Cas1 and Cas2 proteins from *Pyrococcus furiosus* to be necessary and sufficient for spacer integration at CRISPR repeats (Chapter 2). This finding aligns with previous *in vitro* integration studies from other CRISPR-Cas systems reporting Cas1 and Cas2 to be necessary for the spacer integration step of adaptation (2,13,15-19). On the other hand, I found that the Cas1 protein from the Type III-A system of *Streptococcus thermophilus* JIM8232 was capable of spacer integration in the absence of Cas2 (Chapter 4). While it would be surprising if Cas1 were to be found acting alone *in vivo*, it is not out of the question given that a recent study reported Type V-C and V-D Cas1 proteins can function as mini integrases without Cas2 (20). It is more likely that the *S. thermophilus* Cas2 tested in the spacer integration assays was either not able to properly interact with Cas1 in the conditions provided, or the integrity of the protein was compromised.

Integration by *P. furiosus* Cas1 and Cas2 was observed at every repeat on pCR7 (Figure 2.1), similar to what is observed in I-E and I-F systems when lacking the leader-binding integration host factor (IHF) (17,21). When IHF is added to the reaction, integration skews towards the first repeat. In the absence of IHF, *in vitro* spacer integration by the I-E and I-F systems requires supercoiled CRISPR plasmid DNA. In contrast, *P. furiosus* Cas1-Cas2 is capable of integrating at CRISPR repeats present in

either supercoiled circular or linear DNA forms. In this way, *P. furiosus* is similar to Type II-A systems, which can also integrate spacers without supercoiled DNA or IHF (13,19). However, II-A systems direct spacer integration to the first repeat *in vitro* without accessory proteins beyond Cas1-Cas2 (13). While *P. furiosus* Cas1 and Cas2 are capable of integration into non-supercoiled DNA, they lack the required specificity for the leader-proximal repeat. Rollie et al. suggested the large leaders of Type I-A archaeal systems could be an important site for an accessory protein to bind in order to optimize integration at the first repeat (22). In *P. furiosus* we are currently working to identify potential non-Cas host proteins involved in this way.

In addition to the promiscuity to integrate at every repeat instead of just the first repeat, I found numerous highly preferred non-CRISPR sites of integration. Non-CRISPR integration has been reported both *in vitro* and *in vivo* (15,17,20-23). Although off-target integration poses potential deleterious effects *in vivo*, evidence suggests that non-canonical spacer integration might serve in the spontaneous generation of CRISPR loci (23). These seemingly irrelevant off-target integration events proved very useful for helping me to identify sequences recognized by Cas1 and Cas2 for efficient and specific spacer integration. Overall, off-target integration events observed in other systems should not be brushed aside, but rather closely analyzed to gain additional insight into the mechanism of spacer integration.

Cis-acting sequences involved in spacer integration

As spacer integration mechanisms continue to be explored, we are quickly learning more about the cis-acting sequences influencing specific spacer integration at the first repeat of

the CRISPR array. The identity of leader and repeat sequences varies widely among CRISPR-Cas systems, but co-evolves with Cas1 (24,25), suggesting that the sequences are important for recognition by Cas1.

Leader sequence

I found that the integration target substrate was able to be minimized to a single repeat with 10 bp of the leader, while maintaining efficiency and specificity with *P. furiosus* Cas1 and Cas2 (Figure 2.3), in the same way that I-C and II-A systems can be reduced to single repeat with a 10-11 bp leader (2,13,19). I observed that the leader of the *S. thermophilus* Type III-A system was also able to be reduced to 30 bp (Figure 4.4). Further testing will need to be done to determine if the leader can be reduced even further or be entirely dispensable for spacer integration for that particular system.

Although most of the leader was dispensable *in vitro* for *P. furiosus* Cas1 and Cas2, the 8 bp sequence spanning the leader-repeat junction was essential for maintaining specificity (Figure 2.5). The importance of the leader-repeat junction has been observed in other systems (17,22,26-30). This feature comes into play often when discussing a model for the two-step transesterification reaction, as the current leading model states that the first site of integration is directed to the leader-repeat junction through sequence-specific docking, while the second site integration event is directed by a ruler mechanism. While we did not find direct evidence against this model, our findings show that for the *P. furiosus* system, the two-transesterification reactions can occur in either direction. An alternative model is discussed below.

Repeat sequence

I found an adenine-rich motif within the repeat that contributes to specificity of spacer integration (Chapter 2). Other *in vivo* studies have identified motifs in the repeat that contribute to accurate second-site integration at the repeat-spacer junction (following first site integration at the leader-repeat junction) (28,31). Shifting the motifs one way or the other by single nucleotide insertions or deletions caused the second site integration location to shift accordingly, leading to a model that these motifs are involved in a ruler mechanism to maintain the original size of the repeat. However, I found that the mid-repeat motif in the *P. furiosus* repeat is contributing to accurate integration at both junctions, rather than just the repeat-spacer site. To me, that suggests that it plays an upstream role of properly docking/orienting the Cas1-Cas2 integrase, as opposed to a secondary binding event following first site integration at the leader-repeat junction. Based on this observation, combined with my results demonstrating rapid full-site spacer integration from either direction, I suggest a possible alternative model. The Cas1-Cas2 integrase carrying a correctly processed pre-spacer is recruited to the CRISPR repeat via the necessary bendability afforded by the adenine-rich motif and other key bases properly arranged within the repeat. The contacts made between Cas1-Cas2 and the repeat sequence align the active sites of Cas1 perfectly with the two sites of spacer integration 30 bp apart (the size of the repeat). The preference for integration at G and C bases helps to ensure accurate spacer integration at the borders of the repeat. In the end, the repeat size is maintained by the architecture of the Cas1-Cas2 integrase bound to the repeat, rather than a ruler mechanism integrating at a set distance from a motif in a sequence-independent manner as previously suggested in other models. Furthermore, *in vivo* Cas1

likely has an affinity for the sequence of the leader directly neighboring the repeat. While we did not see a preference for the repeat neighboring the leader compared to the other repeats *in vitro*, I expect other host proteins and/or Cas4-1 and Cas4-2 to contribute to this feature of spacer integration specificity.

Concluding remarks

The diverse mechanisms that have been discovered for the invader silencing stage of CRISPR-Cas is being echoed through the diversity we are finding in mechanisms of the adaptation stage. The Type I and Type III CRISPR-Cas systems found in *P. furiosus* and *S. thermophilus* have proven to be useful model systems for *in vitro* characterization of the proteins involved in adaptation. The work presented here will provide a foundation for future exploration into the detailed molecular mechanisms involved in pre-spacer generation and spacer integration.

References

1. Shiimori, M., Garrett, S.C., Graveley, B.R. and Terns, M.P. (2018) Cas4 Nucleases Define the PAM, Length, and Orientation of DNA Fragments Integrated at CRISPR Loci. *Mol Cell*, **70**, 814-824 e816.
2. Lee, H., Zhou, Y., Taylor, D.W. and Sashital, D.G. (2018) Cas4-Dependent Prespacer Processing Ensures High-Fidelity Programming of CRISPR Arrays. *Mol Cell*, **70**, 48-59 e45.

3. Lee, H., Dhingra, Y. and Sashital, D.G. (2019) The Cas4-Cas1-Cas2 complex mediates precise prespacer processing during CRISPR adaptation. *eLife*, **8**, e44248.
4. 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. *Cell Rep*, **22**, 3377-3384.
5. Almendros, C., Nobrega, F.L., McKenzie, R.E. and Brouns, S.J.J. (2019) Cas4-Cas1 fusions drive efficient PAM selection and control CRISPR adaptation. *Nucleic Acids Res*, **47**, 5223-5230.
6. Silas, S., Mohr, G., Sidote, D.J., Markham, L.M., Sanchez-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.
7. 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.
8. 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.
9. Wei, Y., Terns, R.M. and Terns, M.P. (2015) Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. *Genes & development*, **29**, 356-361.

10. Heler, R., Wright, A.V., Vucelja, M., Bikard, D., Doudna, J.A. and Marraffini, L.A. (2017) Mutations in Cas9 Enhance the Rate of Acquisition of Viral Spacer Sequences during the CRISPR-Cas Immune Response. *Mol Cell*, **65**, 168-175.
11. Ka, D., Jang, D.M., Han, B.W. and Bae, E. (2018) Molecular organization of the type II-A CRISPR adaptation module and its interaction with Cas9 via Csn2. *Nucleic Acids Res*, **46**, 9805-9815.
12. Wilkinson, M., Drabavicius, G., Silanskas, A., Gasiunas, G., Siksnys, V. and Wigley, D.B. (2019) Structure of the DNA-Bound Spacer Capture Complex of a Type II CRISPR-Cas System. *Mol Cell*, 1-12.
13. Wright, A.V. and Doudna, J.A. (2016) Protecting genome integrity during CRISPR immune adaptation. *Nat Struct Mol Biol*, **23**, 876-883.
14. 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-+.
15. Nunez, J.K., Lee, A.S., Engelman, A. and Doudna, J.A. (2015) Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature*, **519**, 193-198.
16. 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 Res*, **45**, 2714-2723.
17. Nunez, J.K., Bai, L., Harrington, L.B., Hinder, T.L. and Doudna, J.A. (2016) CRISPR Immunological Memory Requires a Host Factor for Specificity. *Mol Cell*, **62**, 824-833.

18. 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**, 1113-1118.
19. 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.
20. Wright, A.V., Wang, J.Y., Burstein, D., Harrington, L.B., Paez-Espino, D., Kyrpides, N.C., Iavarone, A.T., Banfield, J.F. and Doudna, J.A. (2019) A Functional Mini-Integrase in a Two-Protein-type V-C CRISPR System. *Mol Cell*, **73**, 727-737 e723.
21. 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. *Proc Natl Acad Sci U S A*, **114**, E5122-E5128.
22. 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 Res*, **46**, 1007-1020.
23. Nivala, J., Shipman, S.L. and Church, G.M. (2018) Spontaneous CRISPR loci generation in vivo by non-canonical spacer integration. *Nat Microbiol*, **3**, 310-318.
24. Shah, S.A. and Garrett, R.A. (2011) CRISPR/Cas and Cmr modules, mobility and evolution of adaptive immune systems. *Res Microbiol*, **162**, 27-38.

25. Alkhnbashi, O.S., Shah, S.A., Garrett, R.A., Saunders, S.J., Costa, F. and Backofen, R. (2016) Characterizing leader sequences of CRISPR loci. *Bioinformatics*, **32**, i576-i585.
26. 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 Res*, **42**, 7884-7893.
27. Diez-Villasenor, C., Guzman, N.M., Almendros, C., Garcia-Martinez, 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.
28. 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 Res*, **44**, 4266-4277.
29. Wei, Y., Chesne, M.T., Terns, R.M. and Terns, M.P. (2015) Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in *Streptococcus thermophilus*. *Nucleic Acids Res*, **43**, 1749-1758.
30. Yosef, I., Goren, M.G. and Qimron, U. (2012) Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res*, **40**, 5569-5576.
31. 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. *Cell Rep*, **16**, 2811-2818.