

MECHANISMS OF PHAGE RESISTANCE IN STREPTOCOCCUS  
THERMOPHILUS; DNA UPTAKE INTO CRISPR ARRAYS AND THE ROLE OF  
HOST GENES

by

JENNY GI YAE KIM

(Under the Direction of Michael P. Terns)

ABSTRACT

CRISPR-Cas (Clustered regularly interspaced short palindromic repeats-CRISPR-associated) systems are adaptive immune systems found in prokaryotes that defend against viruses and other foreign genetic elements. The first step of CRISPR-Cas defense, termed adaptation, involves two phases: the acquisition of foreign DNA as protospacers, and the incorporation of the DNA as spacers into the CRISPR array. In this dissertation, *in vitro* and *in vivo* approaches were utilized to investigate both phases of adaptation in the Type II-A CRISPR-Cas system of *Streptococcus thermophilus*. First, the mechanism of spacer incorporation into the CRISPR array was investigated by reconstituting the integration reaction *in vitro*. It was determined that Cas1 and Cas2 proteins accurately integrate spacer DNA into a CRISPR locus. Sequences in the CRISPR leader and repeat were identified as important DNA elements that dictate the first site of integration at the leader-repeat junction. Additionally, second-site integration at the repeat-spacer junction was found to be dependent on multiple determinants including a length-defining mechanism that relies on a repeat element proximal to the second site of integration. The

protospacer selection phase of adaptation was also addressed to investigate how foreign DNA is acquired and discriminated towards PAM (protospacer adjacent motif)-adjacent sequences to generate functional spacers. Here, we demonstrate that Csn2 influences the selection of PAM-adjacent sequences for integration by Cas1-Cas2. Additional genetic analyses revealed that loss of a component of the Cas9 ribonucleoprotein, tracrRNA (the trans-activating CRISPR RNA), reduced spacer duplication events observed within the CRISPR array. Furthermore, loss of nuclease activity of DNA repair proteins RexAB was found to negatively impact adaptation frequency, presumably through the reduction in protospacer generation. Lastly, spontaneous mutations in the *S. thermophilus* FtsH protein leads to phage resistance in the absence of a functional CRISPR-Cas system and this effect was bypassed by mutations in a phage tail chaperonin protein. These results provide valuable insight into the mechanism and regulation of CRISPR adaption in the Type II-A CRISPR-Cas system of *S. thermophilus* and contributes to the overall understanding of adaptive immunity against foreign elements and the host-phage evolutionary arms race.

INDEX WORDS: CRISPR; Cas; Cas1; Cas2; Csn2; Cas9; tracrRNA; Type II-A; adaptation; protospacer generation; PAM; RexAB; FtsH; phage; *Streptococcus thermophilus*

MECHANISMS OF PHAGE RESISTANCE IN STREPTOCOCCUS  
THERMOPHILUS; DNA UPTAKE INTO CRISPR ARRAYS AND THE ROLE OF  
HOST GENES

by

JENNY GI YAE KIM

B.S., Georgia Institute of Technology, 2011

M.S., Georgia State University, 2014

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

2020

© 2020

Jenny Gi Yae Kim

All Rights Reserved



MECHANISMS OF PHAGE RESISTANCE IN STREPTOCOCCUS  
THERMOPHILUS; DNA UPTAKE INTO CRISPR ARRAYS AND THE ROLE OF  
HOST GENES

by

JENNY GI YAE KIM

Major Professor:  
Committee:

Michael P. Terns  
David J. Garfinkel  
Christopher West  
Zachary Wood

Electronic Version Approved:

Ron Walcott  
Dean of the Graduate School  
The University of Georgia  
December 2020

## DEDICATION

This dissertation is dedicated to my family. My father and mother, Steve and Mi Hyang Kim, and my brother, Andrew Kim. This work would not have been possible without your love, encouragement and support. Thank you for believing in me and giving me the strength to pursue my dreams.

## ACKNOWLEDGEMENTS

None of this work would have been possible without the guidance and mentorship of Dr. Michael Terns. The continuous support, patience and motivation you have given me has helped me become the scientist that I am today. Thank you for your dedication to helping me succeed and providing me with the opportunities to foster my career. I would also like to express my sincerest gratitude to my committee members Dr. David Garfinkel, Dr. Christopher West, Dr. Zachary Wood and Dr. Claiborne Glover. Your constant guidance and insightful discussions was instrumental in the development of my research and I am grateful for your encouragement over the years.

To past and present members of the Terns lab, thank you. Every single one of you have been an important part of my success and I am forever grateful: Dr. Yunzhou Wei, Dr. Masami Shimorii, Dr. Ryan Catchpole, Dr. Kawanda Foster, Dr. Julie Grainsy, Walter Woodside, Ralph Zhang, Elizabeth Watts, Clare Edwards, Justin Mclean and Landon Clark. To my talented undergraduate students Sonam Brahmabhatt and Riya Gohil, the hard work that you have put into my research projects has truly allowed this work to be possible. I also want to thank Dr. Rebecca Terns and give you my warmest thanks for your mentorship and invaluable advice you have provided me over the years.

I would also like to express my deepest appreciation to Dr. Brenton Graveley and Dr. Sandra Garrett. The work presented here would not have been possible without you. Sandy, I am sincerely grateful for the work that you have provided for these projects.

Everything from experimental design, discussing results to providing helpful comments on this thesis, you have truly played a critical role in my graduate career.

I am also incredibly grateful for my friends both in Athens, Atlanta and all over the country. To name a few, Lauren Adel, Ivette Nuñez, Kawanda Foster, Lydia Aletraris, Carlo Finlay, Leo Finlay, Ann Stoneburner, Robert Wyatt, Woori Koh, Pat Chung, and Sonia Im, thank you. You have given me the strength when I didn't feel like I had enough and continually provided the encouragement to keep going.

Thank you to my family members, Steve Kim, Mi Hyang Kim, Andrew Kim, Joseph Kim, Kristin Kim, Nick Hughey, Songye Hughey, Jonathan Hughey, Tommy Hughey, all of my family in South Korea, my grandmother Tae-Sun Kim and to my late grandmother and grandfathers Kyung-Bae Kim, Ok-Chul Kim and Chung Tae-Sul. I hope I have made you proud.

Last but not least, thank you to Graham Wyatt. You have been a constant source of light and encouragement during countless times of uncertainty. You have believed in me when I didn't believe in myself and have given me the strength to pursue my dreams.

## TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS .....	v
CHAPTER	
1 INTRODUCTION AND LITERATURE REVIEW .....	1
Introduction to CRISPR-Cas Systems .....	1
Classification of CRISPR-Cas Systems.....	2
Mechanism of Adaptation in CRISPR-Cas Systems .....	4
Mechanism of crRNA biogenesis in CRISPR-Cas Systems.....	15
Mechanism of Target interference in CRISPR-Cas Systems .....	16
Type II-A CRISPR-Cas Systems in <i>S. thermophilus</i> .....	18
Dissertation Overview .....	20
References.....	22
Figures.....	44
2 CRISPR DNA ELEMENTS CONTROLLING SITE-SPECIFIC SPACER INTEGRATION AND PROPER REPEAT LENGTH BY A TYPE II CRISPR-CAS SYSTEM.....	58
Abstract .....	59
Introduction.....	59
Materials and Methods.....	65
Results.....	71

Discussion .....	82
References .....	89
Figures.....	99
3 EFFECTS OF CRISPR-ASSOCIATED PROTEINS AND REXAB ON PROTOSPACER GENERATION AND PHAGE RESPONSE IN S. THERMOPHILUS .....	139
Abstract .....	140
Introduction.....	141
Materials and Methods.....	147
Results.....	154
Discussion .....	166
References .....	176
Figures.....	190
4 DISCUSSION .....	208
Pre-spacer generation in Type II CRISPR systems by RexAB .....	208
Role of Cas proteins during spacer acquisition.....	211
Spacer integration .....	217
Phage resistance by non-CRISPR survivors .....	221
Concluding remarks .....	223
References .....	223

# CHAPTER 1

## INTRODUCTION AND LITERATURE REVIEW

### **Introduction to CRISPR-Cas Systems**

The constant arms race between prokaryotes and invading foreign elements such as phages, plasmids and mobile genetic elements has driven the evolution of defense mechanisms to facilitate an effective immune response to an invader (1,2). One specific mechanism of defense is the CRISPR-Cas system. CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) -Cas (CRISPR associated) systems are immune systems that provide protective immunity against invading foreign nucleic acid elements such as bacteriophages (phages) (3,4). These adaptive immune systems are found in the genomes of roughly half of bacteria and almost all archaea species sequenced so far. CRISPR-Cas systems currently fall into 2 classes, 6 types and 33 subtypes (5). Unlike other mechanisms of defense, CRISPR-Cas systems provide an adaptive method of protection against foreign invaders by generating a library of previous infections which enables a highly efficient and effective immune response upon reinfection of the cell – resulting in a heritable memory bank of past invaders (6-9).

The architecture of CRISPR-Cas systems is comprised of the CRISPR-associated genes and a CRISPR locus consisting of an AT-rich, variably-sized leader sequence adjacent to an array of repeats separated by similarly sized, previously incorporated

spacers (Figure 1.1) (10,11). The CRISPR-Cas immune response consists of three stages: adaptation, crRNA (CRISPR RNA) biogenesis and interference (Figure 1.2). In the adaptation stage, foreign invader sequences termed “protospacers” are captured, processed and integrated into the CRISPR array, where they are referred to as “spacers” (3). Since these spacers originate from foreign nucleic acids and are variable sequences, they serve as a memory bank of past invading elements. During crRNA biogenesis, the CRISPR array is transcribed as a single transcript and processed into short, functional mature CRISPR RNAs (crRNAs) that serve as complementary guides to invading nucleic acids (12-14). During interference or invader silencing, an effector protein(s) forms a complex with the functional crRNA species and base-pairs with the invader sequence upon reinfection to degrade the invading nucleic acid (12,15-17).

DNA and RNA-targeting abilities by CRISPR-Cas systems have been repurposed as a revolutionary tool for biological research. For example, re-programming of the Cas9 enzyme has been utilized for medicinal and diagnostics research ranging from genome editing to regulating gene expression and chromatin interactions (18,19). As CRISPR technology becomes increasingly common in advancing scientific research, the need to understand the basic biology of CRISPR-Cas systems has become paramount.

### **Classification of CRISPR-Cas Systems**

The diversity of CRISPR-Cas systems has prompted the development of stringent methodologies to identify unique features that can be used to classify these systems into defined categories. However, with the ongoing discovery of new CRISPR-Cas systems, classification changes and improves over time (7,20,21). Computational strategies have



been a reliable approach for classification and nomenclature of CRISPR-Cas systems. These approaches exploit not only the architecture and composition differences of the CRISPR and *cas* loci but sequence and gene context are also used to distinguish variants of different systems (11). Currently, CRISPR-Cas systems can be grouped into two main classes: Class I and Class II (5,11,20-22). These two classes are distinguished by whether the effector complex is a multi-subunit complex (Class 1) or a single multi-domain protein (Class 2). Classes are further categorized into 6 types. Class 1 consists of Type I, III and IV while Class 2 consists of Type II, V and VI (Figure 1.3) (5,11).

Class 1 CRISPR-Cas systems facilitate interference utilizing a multi-Cas protein complex. In the well-studied Type I CRISPR-Cas system, the multi-subunit complex (termed the Cascade complex) functions together with a signature *cas3* gene, which encodes a helicase domain fused with an endonuclease domain that is responsible for the cleavage of target DNA (23,24). Within the Type I system, sub-types I-C, I-D, I-E and I-F contain a single operon encoding the *cas1*, *cas2* and *cas3* genes together, while others, like subtypes I-A and I-B, contain several clusters of *cas* genes in separate operons (11,25).

Class 2 systems contain a single effector module that consists of a large protein rather than a multi-subunit complex. For the Type II system consists, that large protein is Cas9 while Cas12 and Cas13 are found in Types V and VI, respectively (5). Cas9 is the most notable of the Type II effector enzymes because it is the most well studied effector protein and it is widely developed as a genome editing tool (18,19). This signature protein is a multidomain nuclease that functions in DNA cleavage and was also found to be involved in adaptation (26-28).

Although CRISPR-Cas systems have notable similarities, evolution of these systems has allowed for significant variability among even the closest types and subtypes of CRISPR-Cas systems, and research into the distinct mechanisms of action is ongoing (Figure 1.3). This dissertation focuses on the Type II-A CRISPR-Cas system of the bacterium, *Streptococcus thermophilus*.

### **Mechanism of Adaptation in CRISPR-Cas Systems**

CRISPR adaptation involves the recognition, processing and integration of foreign nucleic acid sequences into the CRISPR loci as a new spacer. Successful integration of invading nucleic acids into the CRISPR loci allows for efficient CRISPR-mediated immunization upon reinfection of the host by the same invader. While CRISPR-Cas immunity through adaptation remains elusive, recent studies have led to an improved understanding of CRISPR adaptation.

CRISPR adaptation requires the Cas1 and Cas2 proteins to integrate spacers into the CRISPR array. Structural studies have shown that Cas1 and Cas2 proteins form a Cas1<sub>4</sub>-Cas2<sub>2</sub> complex consisting of two dimers of Cas1 bridged by a single Cas2 dimer to incorporate new spacers into the CRISPR array in both Type II and Type I systems (29-31). In the widely studied Type I-E system of *Escherichia coli* and Type II-A system of *Streptococcus thermophilus* and *Streptococcus pyogenes*, Cas1-Cas2 mediated spacer integration exhibited a preference for double-stranded DNA (dsDNA) substrates during integration (29-34). During spacer integration, Cas1-Cas2 recognizes the leader and the adjacent repeat sequence with or without additional host factors to mediate polarized integration to the leader-adjacent repeat (31-33,35-42). Upon integration of the new

spacer sequence, DNA polymerase and ligase host factors are required to repair the CRISPR array (43). Therefore, CRISPR adaptation is generally divided into two stages: the capture and processing of foreign DNA (pre-spacer generation), and integration of the DNA into the host CRISPR locus as a new spacer (spacer integration).

### ***Pre-spacer generation***

During the initial step of adaptation, pre-spacer generation involves the acquisition of foreign nucleic acids, often referred to as mobile genetic elements (MGEs) and is further processed for integration into CRISPR arrays. CRISPR adaptation can be classified into two modes: naïve and primed adaptation. Naïve adaptation occurs when a pre-spacer is acquired from an unfamiliar invader to which the host has no prior immunity and integrates the unique spacer into the CRISPR array (44). Primed adaptation occurs when foreign invaders escape CRISPR defense through mutations in the PAM or protospacer resulting in secondary spacer acquisition of additional sequences of the targeted invader (44-46). Pre-spacer generation is required for both modes of adaptation however, the machinery involved in acquiring substrates can vary.

### ***Naïve adaptation***

Adaptation from MGEs that has not yet been cataloged into the existing CRISPR system is termed naïve adaptation (44). During naïve adaptation, a bias for MGEs over host chromosomal DNA is required to prevent self-targeting and cell death. However, despite the risk of autoimmunity, the frequency of spacers acquired from the genome is quite high (28,47). Recent studies have shown that intermediates generated by the host

proteins, RecBCD are one cause of this bias and a main source of pre-spacer substrates in *E. coli* (47,48). The RecBCD enzyme in *E. coli* is a helicase-nuclease complex involved in DNA repair of double-stranded DNA breaks (DSB) (49,50). DSBs that can arise from UV radiation, DNA-damaging agents or stalled replication forks can be lethal to the cell and several modes of repair have evolved as a means of processing the DSB to allow for repair and recombination.

DNA fragments generated by RecBCD activity have since been hypothesized to be captured by Cas proteins. Studies linking RecBCD activity to adaptation have demonstrated that spacer acquisition is replication-dependent and that dsDNA breaks facilitate an adaptation bias for these regions that require processing and repair by RecBCD (47). In this study, chromosomal hotspots for areas of spacer acquisition were additionally bound by Chi (chromosomal hotspot instigator) sequences which are short octameric sequences that regulate activity of RecBCD enzymes to stall the enzymes from further processing (47,50). Additionally, because Chi sequences occur nearly 14 times more frequently in the *E. coli* genome than phage DNA, RecBCD-mediated processing is a method of early defense against MGEs (47).

Whether RecBCD is directly functioning with Cas proteins to generate pre-spacers or if degradation products simply supply pre-spacer substrates is not clear. However, studies involving RecBCD and adaptation have linked both nuclease and helicase activity of RecBCD to be involved in influencing CRISPR adaptation (47,48). In Gram-positive bacteria lacking RecBCD, the homologous AddAB enzymes have been linked to CRISPR adaptation and demonstrated that nuclease activity influences similar adaptation bias seen with RecBCD (51). As details involving pre-spacer source is

becoming more evident, mechanisms of how cell machinery are involved in pre-spacer generation is not well understood. Though details involving the selection of pre-spacers have become clearer, the mechanisms and cell machinery controlling this process are not well understood.

Regardless of spacer source, a short and highly conserved PAM (protospacer adjacent motif) sequence is found flanking each potential protospacer that will be selected for spacer acquisition (52). PAM sequences are typically 2-5 nucleotides in length and have been identified in several Type I and Type II CRISPR-Cas systems (52,53). These motifs are involved not only in spacer uptake but targeting of invader sequences to prevent self-immunization through recognition of the spacer sequence located in the CRISPR array (27,44,54).

In Type I CRISPR-Cas systems, several variants such as Type I-B, I-C, and I-D, Cas4 is involved with PAM-dependent spacer acquisition. In the Type I-D system, Cas4 not only selects PAM-compatible sequences but also processes pre-spacer substrates according to the PAM (55). In the Type I-C system of *Bacillus halodurans*, Cas4 is required for PAM sequence recognition for pre-spacer processing prior to spacer integration (56,57). The CRISPR-Cas systems in *Pyrococcus furiosus*, require two distinct Cas4 nucleases that are essential for spacer acquisition as both are involved in processing of pre-spacers flanked by a PAM and a secondary recognition motif (58). In CRISPR systems missing Cas4 proteins such as the Type I-E system, Cas1-Cas2 integrase complex relies on Cas1 to recognize the PAM sequence (59,60).

Unlike Type I CRISPR-Cas systems, Type II systems utilize the single effector protein, Cas9, to recognize PAM sequences such as those found in the Type II-A systems

of *S. thermophilus* and *S. pyogenes*. The Cas9 enzyme in *S. pyogenes* recognizes a short 5'-NGG-3' PAM element that is identified by a PAM-interacting (PI) domain within Cas9 (52,61,62). Due to its involvement in PAM recognition, Cas9 is involved in target interference and adaptation (26,28).

Unlike Type I and Type II systems, Type III CRISPR-Cas systems lack PAMs and relies on repeat sequences on the crRNA and the flanking sequence of the target RNA for interference (63). Type III systems are unique in that these systems are capable of both DNA and RNA interference activity (Figure 1.3) (63-68). Despite absence of PAM sequences, some Type III systems such as the Type III-B of *Marinomonas mediterranea* have evolved to acquire spacer sequences from RNA. In these CRISPR systems, Cas1 is fused to a reverse transcriptase to allow acquisition of new spacers from RNA to defend against RNA-based invaders (66).

Type IV of Class I systems and Type V and I of Class II systems are less understood and characterized. Type IV systems are different from other CRISPR systems in that that they do not contain Cas1 and Cas2 (5,11,69). Due to the exclusive absence of Cas1 and Cas2 from Type IV systems, it is implicated that Type IV systems exploit the functionality of other CRISPR systems and their Cas1-Cas2 adaptation modules (70). Most Type V systems are considered as minimal CRISPR arrays as some systems such as V-C and V-D lack the gene encoding for Cas2 and have shorter CRISPR arrays with notably less repeat-spacer units. A recent study demonstrated that the Type V-C Cas1 protein alone is functional as an integrase capable of integrating short DNA fragments into the CRISPR array (71). Lastly, Type VI adaptation modules have been shown to be composed of a reverse-transcriptase-Cas1 fusion and Cas2 to acquire spacers from RNA

molecules, a characteristic similarly seen in Type III systems (72). This dissertation, however, focuses on the mechanisms of DNA uptake in the Type II-A CRISPR-Cas system.

### ***Primed Adaptation***

Acquired spacers from MGEs that have never been encountered, allows for CRISPR-Cas systems to generate a targeted response by facilitating base-pairing between the crRNA and the invading DNA. However, mutations with the MGEs in the targeted spacer sequence or PAM can result in a response called ‘primed adaptation’ or ‘priming’ that is initiated by the interference complex. Priming however, is not only limited to mutations in the MGEs and reports have demonstrated that priming can occur with perfect targeting and when the target had an improper PAM as well as mismatches in the seed region (73). This priming response allows hosts to acquire additional defense against MGEs that have been able to evade interference due to secondary mutations. To date, priming has been shown to occur in several Type I CRISPR-Cas systems such as I-B, I-C, I-E and I-F although the molecular details involved in primed adaptation is not well understood (45,74-80). For example, in the widely studied Type I-E system of *E. coli*, primed adaptation involves not only the adaptation proteins Cas1 and Cas2 but the interference (Cascade) complex and Cas3 (45,79,81). In this system, it is thought that weakened binding of the Cascade complex due to mutations in either the PAM sequence or mutations affecting the initial contacts between the crRNA and the target sequence. As a result, this can lower the efficiency of binding of Cascade to the target sequence for interference (46,82). Failure to bind to the target site prevents the recruitment of Cas3 for

DNA degradation. However, during primed adaptation, Cas1-Cas2 recruits Cas3 and translocates along the target DNA to select a new protospacer for spacer acquisition (83-85). Despite evolving several methods to ensure efficient defense against invading MGEs, proper spacer acquisition and integration upon initial contact through naïve adaptation is primarily required to facilitate an immune response.

### ***Role of DNA repair proteins***

In Gram-negative bacteria such as *E. coli*, the RecBCD pathway acts on DSBs to facilitate DNA repair through recombination. RecBCD recognizes blunt dsDNA and processes the duplex ends to generate a 3'-terminated ssDNA strand. The processing activity of the RecBCD complex continues to unwind and degrade dsDNA until an octameric regulatory sequence called a Chi (crossover hotspot instigator) sequence is reached (50). The Chi sequence is located on a single strand of the DNA (5'-GCTGGTGG-3') and upon recognition of the Chi sequence, the biochemical properties of RecBCD is altered and translocation across the dsDNA is stalled to allow the recruitment and loading of the RecA protein to form a filament which further facilitates homologous recombinational repair (Figure 1.4) (50).

An alternative class of the RecBCD helicase-nuclease enzyme is the AddAB (ATP-dependent DNase, and also referred to as RexAB) family found in Gram-positive bacteria. Similar to RecBCD enzymes, both AddAB and RexAB enzymes have helicase-nuclease activity responsible for DNA repair by facilitating homologous recombination (50). Additionally, its function is analogous to RecBCD and have been found to rescue



DNA repair activity in RecBCD deletion strains in *Escherichia coli* (86). This activity was also observed with the AddAB homolog in *Lactococcus lactis*, RexAB. (86,87).

Although the primary structure of RecBCD and AddAB/RexAB are different, conserved regions such as the RecB-like nuclease domain allow functionality of these proteins to be highly similar (Figure 1.5)(50). RecBCD has two helicase domains (RecB and RecD) and one nuclease domain (RecC) while both AddAB and RexAB are made up of a single helicase domain (AddA, RexA) and two nuclease domains (AddA/AddB, RexA/RexB) (50,88,89)). The *E. coli* RecBCD is the most well-studied member of the RecBCD/AddAB family(90). In the RecBCD complex, the RecB contains a superfamily 1 helicase module (SF1) with a 3'-5' directionality(91). The nuclease domain responsible for cleavage of the DNA duplex as well as RecA loading is located in the RecC subunit (92). The RecC subunit recognizes Chi sequences (*E. coli*: GCTGGTGG) and additionally contains a similar SF1 helicase fold, however the helicase motif in RecC has lost the conserved motifs required for functionality deeming it inactive(90). RecD is an additional SF1 helicase with proper functionality that translocates in the 5'-3' direction (91).

The Gram-positive homologs of the three-subunit RecBCD are the two-subunit AddAB/RexAB enzymes. Simply, AddA/RexA corresponds to the RecB subunit of RecBCD and AddB/RexB with RecC. Although AddAB is not as well studied as RecBCD, structures have revealed details on similarities and differences between AddAB and RecBCD (93). AddA contains the same SF1 helicase module as RecB with a 3'-5' directionality and additionally holds a nuclease domain that translocates in the 3'-5'

direction. AddB, similar to RecC, recognizes the Chi sequence (*B. subtilis*: AGCGG) and also contains a nuclease domain that processes from the 5'-3' direction (89,94).

Recent studies have tied both nuclease and helicase activity of the RecBCD enzyme to adaptation (47,48). In *E. coli*, major sources of protospacers were found to be replication-dependent and DSBs originating from stalled replication forks initiate spacer acquisition. Additionally, chromosomal hotspots of acquired spacers were bound by Chi sites suggesting that areas of processing limit spacer acquisition (47). Both nuclease and helicase activity of RecBCD have been tied to naïve adaptation in *E. coli*, however a recent study found that naïve adaptation does not require nuclease activity of RecBCD but rather helicase activity may be important (48). In my work, the nuclease activity of the RecBCD homolog RexAB, in *S. thermophilus* was demonstrated to similarly contribute to spacer acquisition during CRISPR adaptation.

Although RecBCD, AddAB and RexAB possess different primary structures, the functionality of these enzymes is conserved between species especially due to the highly conserved RecB-like nuclease domains (Figure 1.5). The observation of Chi-dependent processing by RecBCD influencing CRISPR spacer uptake in *E. coli* was observed with the homolog AddAB in the Type II system of *S. aureus* (51). Similar to the adaptation bias observed with RecBCD mutations in *E. coli* of the Type I-E system, mutation of the nuclease domain of AddA resulted in an equivalent bias for strong adaptation hotspots in regions bound by the staphylococcal Chi sequence (51). This bias suggests that Type I and Type II systems utilizes intermediates generated by RecBCD or AddAB DNA repair as sources of new spacers and is likely that homologous enzymes such as RexAB in *S. thermophilus* functions in a similar manner (Figure 1.5).

### ***Spacer integration***

Specific integration to the proper junctions of the leader-proximal repeat is necessary to accurately duplicate the repeat between each unique spacer and to generate a consecutive memory bank of past invaders. Proper integration additionally requires spacer sequences to be processed to the correct size and integrated in the CRISPR array in a functional orientation. Orientation of spacers is dependent on PAM-specific spacer integration in which spacers are integrated into the CRISPR array with the respect to the PAM, but the PAM is missing from the final integrated spacer. The resulting new spacer will produce a functional crRNA that are able to target the foreign DNA during interference.

Cas1 and Cas2 are universally found in almost all CRISPR systems and are involved in spacer integration into the CRISPR array (5,11,21). Cas1 is a homodimeric enzyme with a metal-dependent nuclease active site (95). *E. coli* Cas1 proteins have nuclease activity against both double and single-stranded DNA as well as RNA (96). Cas2 proteins have consistently been demonstrated to serve as a structural role by bridging the dimers of Cas1 proteins in the integrase complex of Cas1-Cas2 (30,31,97). Cas2 in *B. halodurans* was demonstrated to have nuclease activity specific to dsDNA substrates (98). However, biochemical studies *in vitro* have shown that the active site of Cas2 of the I-E system is not required for spacer integration (38).

Despite genetic differences amongst CRISPR systems, the method of spacer integration by the Cas1-Cas2 integrase complex is thought to be conserved as Cas1 and Cas2 are found in nearly all CRISPR-Cas systems with very few exceptions (11). Integration of the pre-spacer is facilitated by two nucleophilic attacks by the 3' hydroxyl

ends of the pre-spacer at the borders of leader-adjacent repeat sequence in the CRISPR array (31-34,99-101). The first site of integration occurs at one junction of the repeat to produce a half-site integration intermediate. Second site integration at the other repeat junction results in the progression to a full-site integration product. As the strands of the repeat dissociate, the host DNA polymerase fills in the remaining sequences followed by ligation to seal the generated DNA nicks, resulting in a new repeat-spacer unit (Figure 1.6) (43).

To maintain proper spacer integration and repeat duplication, several DNA elements have been identified to direct site-specific integration. Sequences spanning the leader play a critical role in defining spacer integration polarity to the first repeat in *S. thermophilus* Type II-A systems both *in vivo* and *in vitro* (32,41). Similar leader sequence motifs in other Type II-A systems such as *S. pyogenes* and *E. faecalis* have also been deemed essential during spacer integration (31,34). DNA motifs and elements within the repeat sequence have additionally been observed to regulate site-specific spacer integration. Some studies have demonstrated that the inverted repeats sequences or palindromic repeats act as docking sites to facilitate molecular rulers to define the order of the two-step integration reaction while others exhibit sequences within the repeat essential for integration although the mechanisms are not well understood (31,32,34,36,40,100).

Unlike Type II systems, Type I-E and I-F systems require an integration host factor (IHF) to facilitate polarized integration to the first repeat (38,97,102). IHF binds and bends the CRISPR array at the leader sequence to allow the Cas1-Cas2 integrase to integrate spacers to the first repeat. In CRISPR systems lacking IHF such as those that are

found in *S. solfataricus*, an unidentified ATP-dependent host factor is required to mediate polarized integration (39). The work presented in this dissertation investigates the mechanistic details of spacer integration in the Type II-A CRISPR-Cas system of *Streptococcus thermophilus*.

### **Mechanism of crRNA biogenesis in CRISPR-Cas Systems**

During CRISPR mediated defense, crRNA biogenesis is a critical step in the process of generating mature and functional crRNAs. Although the details of how mature crRNAs are generated, the process of producing functional crRNAs involve two general steps. The first step involves the transcription of the precursor CRISPR RNA molecule (pre-crRNA) under a promoter that is located within the leader sequence. The second step is the maturation of the pre-crRNA transcript into mature crRNAs. This involves the cleavage of repeat sequences within the transcript by either Cas (CRISPR-associated) or host proteins to generate full-length spacer sequences that are flanked by repeat sequences of varying length. Depending on the CRISPR system, these spacer sequences are further processed to remove the flanking repeat sequences to generate functional crRNAs (103).

In Type I CRISPR systems, the pre-crRNA transcript is processed by Cas6 within the Cascade interference complex (or alternatively Cas5 in Type I-C systems) (12,13,104,105). Processing by Cas6 cleaves the repeat sequences in a conserved position upstream of the junction between the repeat and spacer to generate mature crRNAs (12,13). In additional cases, a secondary processing step to remove repeat sequences at the 3' end is required. Similar to Type I systems, Type III systems depend on proteins of

the Cas6 family. Cas6 directly processes the pre-crRNA independent of additional Cas proteins to generate an intermediate product that requires further processing to form mature crRNAs (13,106).

Type II systems have evolved to utilize an additional RNA molecule called the tracrRNA (transactivating CRISPR RNA) that requires Cas9 and a host-encoded RNaseIII enzyme which makes the Type II system unique compared to Type I and Type III systems (14). The tracrRNA forms stable complexes by base pairing with each repeat sequence of the pre-crRNA. This base pairing forms a double stranded RNA substrate that is recognized and further cleaved by endoribonuclease III (RNase III) in the presence of Cas9. Cleavage of the crRNA transcript bound by tracrRNA results in mature crRNA-tracrRNA complexes that bind to Cas9 for target interference (107-110).

### **Mechanism of Target interference in CRISPR-Cas Systems**

Sequence-specific invader silencing by CRISPR-Cas systems is the basis for defense. Following crRNA biogenesis to generate mature crRNAs, sequence complementarity facilitated by the interference protein(s) between the mature crRNA and the target sequence is necessary for target cleavage.

Interference in Class 1 CRISPR-Cas systems involves two general pathways. In the Type I systems, the Cascade (CRISPR-associated complex for antiviral defense) complex is required for target recognition and Cas3 is further recruited for target cleavage (12). The Type I-E system of *E. coli* is the most well characterized system and thus serves as a model to understand target interference in Class 1 systems. In the Type I-E system, the Cascade complex bound to the mature crRNA is composed of a multi-

subunit complex of Cas5<sub>1</sub>-Cas6<sub>1</sub>-Cas7<sub>6</sub>-Cas8<sub>1</sub>-Cas11<sub>2</sub> (12,111). Prior to interference, the PAM sequence is recognized by the Cascade complex and binding of the crRNA to the target strand results in the formation of an R-loop of the target DNA (112,113). Cas3 is recruited to the R-loop resulting in the nicking of the non-target strand(112-115). Similar to Type I systems, Type III systems form large multi-subunit complexes known as Csm and Cmr for Type III-A and III-B, respectively that binds to the mature crRNA to mediate cleavage (116). However, unlike Type I systems, Cas6 is not required for Type III target interference and is additionally interacts with target RNA rather than forming an R-loop structure of the invader DNA to target both RNA and DNA (16,65,68,117). Upon binding of the Csm/Cmr complex to the target transcript, two conserved modes of cleavage occur. The Cas10 subunit facilitates a single-stranded break in the template DNA of the invader while Cas7 (Csm3 (III-A), Cmr4 (III-B)) cleaves the target RNA transcript (16,65,118,119). A second method of interference is triggered by Cas10 generated cyclic oligoadenylates from ATP that further activates Csm6 (III-A) /Csx1 (III-B) RNase activity against non-specific RNA (118,120-124).

Interference by Class II systems is facilitated by a single effector protein. In the well-studied Type II system, Cas9 bound to a crRNA-tracrRNA complex scans the target DNA for the proper PAM sequence and base-pairs the crRNA to the target strand to generate a blunt, DSB (27,125). This activity relies on the formation of an R-loop structure that triggers the cleavage of both the target and non-target strand by two domains (HNH and RuvC) within the Cas9 enzyme (27,126,127). Type V systems utilize the Cas12 enzyme for interference. Unlike Type II systems, Cas12 binds to the crRNA without the presence of a tracrRNA and cleaves both strands of the target strand via a

RuvC-like domain at the PAM-distal end of the target sequence (128-131). Type VI systems, facilitate invader silencing by Cas13. Similar to Type V systems, tracrRNA is not required for target cleavage (129,132,133). However, a key element of Type VI interference is its ability to target complementary ssRNAs (129,132,134,135). This mechanism is facilitated by a PFS (protospacer flanking site) rather than a PAM sequence to induce a conformational change within Cas13 to cleave the target RNA (132,133).

### **Type II-A CRISPR-Cas Systems in *S. thermophilus***

*S. thermophilus*, is a Gram-positive bacterium widely known for its use in the dairy industry as a starter culture making this organism one of the most economically important of all lactic acid bacteria (136). The *S. thermophilus* DGCC 7710 strain is of particular interest for this study because of its role in the discovery of CRISPR-Cas systems (3,137). The DGCC 7710 strain contains four natural CRISPR systems: Type II-A (CRISPR1 and CRISPR3), Type III-A (CRISPR2) and Type I-E (CRISPR4) (Figure 1.7) (138).

CRISPR1 and CRISPR3 systems are Type II-A CRISPR systems. These Type II-A CRISPR systems consist of the universally conserved Cas1 and Cas2 proteins as well as Csn2 and the interference enzyme, Cas9. It has been demonstrated that all four Cas proteins are required for efficient adaptation *in vivo* although the detailed functional roles of these proteins in adaptation has yet to be determined (28). Cas1 and Cas2 are essential for spacer integration in Type II systems as also seen with Type I systems (29-35,42,81,97). Although Csn2 an essential requirement for adaptation, its function is not yet fully understood. Structural studies show that Csn2 is a tetrameric protein that forms a



toroidal structure although its function has only been linked to binding dsDNA ends (139-142). A structural study showing complex formation of Cas1-Cas2-Csn2 *in vitro* has speculated that Csn2 binds to dsDNA ends and assembles into a complex with Cas1 and Cas2 during pre-spacer generation (143). Therefore, a current model suggests that the Cas1-Cas2-Csn2 complex translocates on the DNA until it encounters Cas9 bound to a PAM sequence to initiate cleavage of the pre-spacer to encapsulate a processed substrate of appropriate size (30 bp) for spacer integration (143). The Cas9 nuclease is an essential protein involved in both CRISPR interference and adaptation although nuclease activity is not required for spacer acquisition but rather PAM recognition by Cas9 (26,28). Mutations in the PAM recognition domain of Cas9 resulted in PAM-independent spacer acquisition confirming that Cas9 is directly involved in PAM-dependent spacer acquisition during CRISPR adaptation (26,28).

Despite both CRISPR systems being classified as Type II-A systems, sequence similarities between CRISPR1 and CRISPR3 are low at approximately 33.6% and 41.3% identity in *cas1* and *cas2* (137). Both Type II-A systems are active in spacer uptake against foreign invading elements such as plasmids and phages while CRISPR1 obtains more spacers universally compared to CRISPR3 (3,44,125,137).

The CRISPR2 system is a Csm Type III-A system. These CRISPR systems in *S. thermophilus* strains are thought to have a reduced ability for spacer acquisition due to high proportion of *cas* genes but a low percentage of repeat-spacer region (144). CRISPR4 is a Cse Type I-E system and is found in only a few strains of *S. thermophilus*. The CRISPR4 system of *S. thermophilus* share similar genetic organization with *E. coli*

despite no functional activity *in vivo* although interference activity was characterized *in vitro* suggesting that this system is active during targeting (24,145,146).

While studies focused on adaptation have primarily been targeted to the Type II CRISPR-Cas systems of *S. thermophilus*, details encompassing pre-spacer generation and spacer integration are still not well understood. Current studies in Type II-A systems have provided strong foundational work between homologous systems. However, it is evident that these processes can vastly differ between systems and organism resulting in the need for further research into understanding the details of CRISPR adaptation.

### **Dissertation Overview**

The chapters of this dissertation describe the scientific studies and contributions I have made towards understanding the details underlying CRISPR adaptation in the Type II-A CRISPR-Cas system of *S. thermophilus*.

Chapter 2 investigates the mechanistic details of how spacer sequences are integrated into the CRISPR array. While details vary between different CRISPR-Cas systems, a common requirement for CRISPR immunity requires successful integration of a spacer sequence captured from an invader into the CRISPR array. This results in a memory bank of recent and past invaders that have infected the host cell. These new spacers are integrated at the leader-proximal repeat rather than downstream repeats in a polarized manner to generate a consecutive array of past invader (147-150). This study utilizes a reconstituted *in vitro* reaction using Cas1 and Cas2 proteins to identify DNA elements that control site-specific spacer integration and accurate repeat length in the CRISPR array. In this chapter, we characterized the *in vitro* properties of Cas1 and Cas2

and found that both proteins are required to catalyze full-site spacer integration into the CRISPR array. Unlike Type I systems, the Type II systems exhibit an intrinsic ability to exhibit polarized integration in a directional manner. We also provide the first evidence supporting a molecular ruler-based mechanism in a Type II system that guides the second-site integration reaction to the repeat-spacer junction to maintain repeat length.

Chapter 3 focuses on the upstream processes prior to spacer integration and shifts attention to how protospacers originating from foreign DNA elements are selected and processed for integration into the CRISPR array. In this chapter, we characterized nuclease activity of RexAB in *S. thermophilus* and its effects on spacer acquisition. The RexAB enzyme contains two distinct nucleases in each protein which negatively affects adaptation frequency against both chromosomal DNA and an invading phage when mutated. This provides the first evidence of associating the RexAB DNA repair complex to CRISPR adaptation. We also developed a natural transformation assay to introduce foreign DNA substrates as a way to observe processing and integration of spacers to delineate potential roles of the poorly characterized CRISPR protein, Csn2 as well as RexAB. We provide evidence that Csn2 increases specificity for PAM-adjacent sequences for spacer integration. PAM selection and orientation of spacers during integration *in vivo*. Lastly, by characterizing phage infection survivors in the absence of CRISPR systems, we were able to detect spontaneous mutations within the host genome in a membrane-bound metalloprotease, FtsH that leads to phage resistance and cell survival. To expand on this adaptive model between host and invader, we also identified mutations in phage tail assembly proteins in second-generation phages that apparently enabled the phage to bypass the host ftsH mutations and successfully infect these hosts

that were resistant to infection by wild-type phages. In summary, this dissertation provides insight into the processes involved in adaptation that facilitates CRISPR immunity as well as provides insight into the evolutionary arms race between phages and their bacterial hosts.

## References

1. Bernheim, A. and Sorek, R. (2018) Viruses cooperate to defeat bacteria. *Nature*, **559**, 482-484.
2. Bikard, D. and Marraffini, L.A. (2012) Innate and adaptive immunity in bacteria: mechanisms of programmed genetic variation to fight bacteriophages. *Curr Opin Immunol*, **24**, 15-20.
3. 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.
4. Marraffini, L.A. (2015) CRISPR-Cas immunity in prokaryotes. *Nature*, **526**, 55-61.
5. Makarova, K.S., Wolf, Y.I., Iranzo, J., Shmakov, S.A., Alkhnbashi, O.S., Brouns, S.J.J., Charpentier, E., Cheng, D., Haft, D.H., Horvath, P. *et al.* (2020) Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. *Nat Rev Microbiol*, **18**, 67-83.
6. Koonin, E.V. and Makarova, K.S. (2009) CRISPR-Cas: an adaptive immunity system in prokaryotes. *F1000 Biol Rep*, **1**, 95.

7. 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. *Biol Direct*, **1**, 7.
8. Marraffini, L.A. and Sontheimer, E.J. (2010) CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat Rev Genet*, **11**, 181-190.
9. 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 Biochem Sci*, **34**, 401-407.
10. Kunin, V., Sorek, R. and Hugenholtz, P. (2007) Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol*, **8**, R61.
11. 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.
12. 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.
13. 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.

14. 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.
15. Bailey, S. (2013) The Cmr complex: an RNA-guided endoribonuclease. *Biochem Soc Trans*, **41**, 1464-1467.
16. 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.
17. Westra, E.R., van Erp, P.B., Kunne, T., Wong, S.P., Staals, R.H., Seegers, C.L., Bollen, S., Jore, M.M., Semenova, E., Severinov, K. *et al.* (2012) CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. *Mol Cell*, **46**, 595-605.
18. Adli, M. (2018) The CRISPR tool kit for genome editing and beyond. *Nat Commun*, **9**, 1911.
19. Pickar-Oliver, A. and Gersbach, C.A. (2019) The next generation of CRISPR-Cas technologies and applications. *Nat Rev Mol Cell Biol*, **20**, 490-507.
20. Haft, D.H., Selengut, J., Mongodin, E.F. and Nelson, K.E. (2005) A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput Biol*, **1**, e60.
21. 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. *Nat Rev Microbiol*, **9**, 467-477.
22. Jansen, R., Embden, J.D., Gaastra, W. and Schouls, L.M. (2002) Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol*, **43**, 1565-1575.
  23. Mulepati, S. and Bailey, S. (2011) Structural and biochemical analysis of nuclease domain of clustered regularly interspaced short palindromic repeat (CRISPR)-associated protein 3 (Cas3). *J Biol Chem*, **286**, 31896-31903.
  24. Sinkunas, T., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P. and Siksnys, V. (2011) Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *EMBO J*, **30**, 1335-1342.
  25. Vestergaard, G., Garrett, R.A. and Shah, S.A. (2014) CRISPR adaptive immune systems of Archaea. *RNA Biol*, **11**, 156-167.
  26. 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.
  27. 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.
  28. Wei, Y., Terns, R.M. and Terns, M.P. (2015) Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. *Genes Dev*, **29**, 356-361.

29. Nunez, J.K., Harrington, L.B., Kranzusch, P.J., Engelman, A.N. and Doudna, J.A. (2015) Foreign DNA capture during CRISPR-Cas adaptive immunity. *Nature*, **527**, 535-538.
30. 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.
31. 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.
32. Kim, J.G., Garrett, S., Wei, Y., Graveley, B.R. and Terns, M.P. (2019) CRISPR DNA elements controlling site-specific spacer integration and proper repeat length by a Type II CRISPR-Cas system. *Nucleic Acids Res*, **47**, 8632-8648.
33. 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.
34. Wright, A.V. and Doudna, J.A. (2016) Protecting genome integrity during CRISPR immune adaptation. *Nat Struct Mol Biol*, **23**, 876-883.
35. Diez-Villasenor, C., Guzman, N.M., Almendros, C., Garcia-Martinez, J. and Mojica, F.J. (2013) CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of *Escherichia coli*. *RNA Biol*, **10**, 792-802.



36. 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.
37. 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.
38. 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.
39. 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.
40. 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.
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. 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.

43. Ivancic-Bace, I., Cass, S.D., Wearne, S.J. and Bolt, E.L. (2015) Different genome stability proteins underpin primed and naive adaptation in *E. coli* CRISPR-Cas immunity. *Nucleic Acids Res*, **43**, 10821-10830.
44. 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.
45. Fineran, P.C., Gerritzen, M.J., Suarez-Diez, M., Kunne, T., Boekhorst, J., van Hijum, S.A., Staals, R.H. and Brouns, S.J. (2014) Degenerate target sites mediate rapid primed CRISPR adaptation. *Proc Natl Acad Sci U S A*, **111**, E1629-1638.
46. Semenova, E., Jore, M.M., Datsenko, K.A., Semenova, A., Westra, E.R., Wanner, B., van der Oost, J., Brouns, S.J. and Severinov, K. (2011) Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc Natl Acad Sci U S A*, **108**, 10098-10103.
47. 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.
48. Radovic, M., Killelea, T., Savitskaya, E., Wettstein, L., Bolt, E.L. and Ivancic-Bace, I. (2018) CRISPR-Cas adaptation in *Escherichia coli* requires RecBCD helicase but not nuclease activity, is independent of homologous recombination, and is antagonized by 5' ssDNA exonucleases. *Nucleic Acids Res*, **46**, 10173-10183.

49. 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.
50. Wigley, D.B. (2013) Bacterial DNA repair: recent insights into the mechanism of RecBCD, AddAB and AdnAB. *Nat Rev Microbiol*, **11**, 9-13.
51. 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.
52. Mojica, F.J.M., 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.
53. Sorek, R., Lawrence, C.M. and Wiedenheft, B. (2013) CRISPR-mediated adaptive immune systems in bacteria and archaea. *Annu Rev Biochem*, **82**, 237-266.
54. Shah, S.A., Erdmann, S., Mojica, F.J. and Garrett, R.A. (2013) Protospacer recognition motifs: mixed identities and functional diversity. *RNA Biol*, **10**, 891-899.
55. 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.
56. Lee, H., Dhingra, Y. and Sashital, D.G. (2019) The Cas4-Cas1-Cas2 complex mediates precise prespacer processing during CRISPR adaptation. *Elife*, **8**.

57. 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.
58. 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.
59. 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.
60. Yoganand, K.N., Muralidharan, M., Nimkar, S. and Anand, B. (2019) Fidelity of prespacer capture and processing is governed by the PAM-mediated interactions of Cas1-2 adaptation complex in CRISPR-Cas type I-E system. *J Biol Chem*, **294**, 20039-20053.
61. Anders, C., Niewoehner, O., Duerst, A. and Jinek, M. (2014) Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature*, **513**, 569-573.
62. Nishimasu, H., Ran, F.A., Hsu, P.D., Konermann, S., Shehata, S.I., Dohmae, N., Ishitani, R., Zhang, F. and Nureki, O. (2014) Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell*, **156**, 935-949.
63. Marraffini, L.A. and Sontheimer, E.J. (2010) Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature*, **463**, 568-571.

64. Johnson, K., Learn, B.A., Estrella, M.A. and Bailey, S. (2019) Target sequence requirements of a type III-B CRISPR-Cas immune system. *J Biol Chem*, **294**, 10290-10299.
65. 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.
66. 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.
67. 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.
68. 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.
69. Koonin, E.V., Makarova, K.S. and Zhang, F. (2017) Diversity, classification and evolution of CRISPR-Cas systems. *Curr Opin Microbiol*, **37**, 67-78.
70. Pinilla-Redondo, R., Mayo-Munoz, D., Russel, J., Garrett, R.A., Randau, L., Sorensen, S.J. and Shah, S.A. (2020) Type IV CRISPR-Cas systems are highly

- diverse and involved in competition between plasmids. *Nucleic Acids Res*, **48**, 2000-2012.
71. 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.
  72. Toro, N., Mestre, M.R., Martinez-Abarca, F. and Gonzalez-Delgado, A. (2019) Recruitment of Reverse Transcriptase-Cas1 Fusion Proteins by Type VI-A CRISPR-Cas Systems. *Front Microbiol*, **10**, 2160.
  73. Garrett, S., Shiimori, M., Watts, E.A., Clark, L., Graveley, B.R. and Terns, M.P. (2020) Primed CRISPR DNA uptake in *Pyrococcus furiosus*. *Nucleic Acids Res*, **48**, 6120-6135.
  74. Li, M., Wang, R. and Xiang, H. (2014) *Haloarcula hispanica* CRISPR authenticates PAM of a target sequence to prime discriminative adaptation. *Nucleic Acids Res*, **42**, 7226-7235.
  75. 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 Res*, **42**, 2483-2492.
  76. Rao, C., Chin, D. and Ensminger, A.W. (2017) Priming in a permissive type I-C CRISPR-Cas system reveals distinct dynamics of spacer acquisition and loss. *RNA*, **23**, 1525-1538.
  77. Richter, C., Dy, R.L., McKenzie, R.E., Watson, B.N., Taylor, C., Chang, J.T., McNeil, M.B., Staals, R.H. and Fineran, P.C. (2014) Priming in the Type I-F

- CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. *Nucleic Acids Res*, **42**, 8516-8526.
78. 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. *Nat Commun*, **7**, 12853.
  79. 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.
  80. 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 Res*, **43**, 10831-10847.
  81. 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.
  82. Westra, E.R., Semenova, E., Datsenko, K.A., Jackson, R.N., Wiedenheft, B., Severinov, K. and Brouns, S.J. (2013) Type I-E CRISPR-cas systems discriminate target from non-target DNA through base pairing-independent PAM recognition. *PLoS Genet*, **9**, e1003742.
  83. Blosser, T.R., Loeff, L., Westra, E.R., Vlot, M., Kunne, T., Sobota, M., Dekker, C., Brouns, S.J.J. and Joo, C. (2015) Two distinct DNA binding modes guide dual roles of a CRISPR-Cas protein complex. *Mol Cell*, **58**, 60-70.
  84. 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.
85. Xue, C., Whitis, N.R. and Sashital, D.G. (2016) Conformational Control of Cascade Interference and Priming Activities in CRISPR Immunity. *Mol Cell*, **64**, 826-834.
  86. Kooistra, J., Haijema, B.J. and Venema, G. (1993) The Bacillus subtilis addAB genes are fully functional in Escherichia coli. *Mol Microbiol*, **7**, 915-923.
  87. el Karoui, M., Ehrlich, D. and Gruss, A. (1998) Identification of the lactococcal exonuclease/recombinase and its modulation by the putative Chi sequence. *Proc Natl Acad Sci U S A*, **95**, 626-631.
  88. Quiberoni, A., Biswas, I., El Karoui, M., Rezaiki, L., Tailliez, P. and Gruss, A. (2001) In vivo evidence for two active nuclease motifs in the double-strand break repair enzyme RexAB of Lactococcus lactis. *J Bacteriol*, **183**, 4071-4078.
  89. Yeeles, J.T. and Dillingham, M.S. (2007) A dual-nuclease mechanism for DNA break processing by AddAB-type helicase-nucleases. *J Mol Biol*, **371**, 66-78.
  90. Singleton, M.R., Dillingham, M.S., Gaudier, M., Kowalczykowski, S.C. and Wigley, D.B. (2004) Crystal structure of RecBCD enzyme reveals a machine for processing DNA breaks. *Nature*, **432**, 187-193.
  91. Singleton, M.R., Dillingham, M.S. and Wigley, D.B. (2007) Structure and mechanism of helicases and nucleic acid translocases. *Annu Rev Biochem*, **76**, 23-50.



92. Spies, M. and Kowalczykowski, S.C. (2006) The RecA binding locus of RecBCD is a general domain for recruitment of DNA strand exchange proteins. *Mol Cell*, **21**, 573-580.
93. Saikrishnan, K., Yeeles, J.T., Gilhooly, N.S., Krajewski, W.W., Dillingham, M.S. and Wigley, D.B. (2012) Insights into Chi recognition from the structure of an AddAB-type helicase-nuclease complex. *EMBO J*, **31**, 1568-1578.
94. Chedin, F., Noirot, P., Biaudef, V. and Ehrlich, S.D. (1998) A five-nucleotide sequence protects DNA from exonucleolytic degradation by AddAB, the RecBCD analogue of *Bacillus subtilis*. *Mol Microbiol*, **29**, 1369-1377.
95. Wiedenheft, B., Zhou, K., Jinek, M., Coyle, S.M., Ma, W. and Doudna, J.A. (2009) Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. *Structure*, **17**, 904-912.
96. Babu, M., Beloglazova, N., Flick, R., Graham, C., Skarina, T., Nocek, B., Gagarinova, A., Pogoutse, O., Brown, G., Binkowski, A. *et al.* (2011) A dual function of the CRISPR-Cas system in bacterial antiviral immunity and DNA repair. *Mol Microbiol*, **79**, 484-502.
97. 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.
98. Nam, K.H., Ding, F., Haitjema, C., Huang, Q., DeLisa, M.P. and Ke, A. (2012) Double-stranded endonuclease activity in *Bacillus halodurans* clustered regularly interspaced short palindromic repeats (CRISPR)-associated Cas2 protein. *J Biol Chem*, **287**, 35943-35952.

99. 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.
100. Grainy, J., Garrett, S., Graveley, B.R. and M, P.T. (2019) CRISPR repeat sequences and relative spacing specify DNA integration by *Pyrococcus furiosus* Cas1 and Cas2. *Nucleic Acids Res*, **47**, 7518-7531.
101. 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**.
102. 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.
103. 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.
104. Garside, E.L., Schellenberg, M.J., Gesner, E.M., Bonanno, J.B., Sauder, J.M., Burley, S.K., Almo, S.C., Mehta, G. and MacMillan, A.M. (2012) Cas5d processes pre-crRNA and is a member of a larger family of CRISPR RNA endonucleases. *RNA*, **18**, 2020-2028.
105. Nam, K.H., Haitjema, C., Liu, X., Ding, F., Wang, H., DeLisa, M.P. and Ke, A. (2012) Cas5d protein processes pre-crRNA and assembles into a cascade-like

- interference complex in subtype I-C/Dvulg CRISPR-Cas system. *Structure*, **20**, 1574-1584.
106. 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.
  107. Briner, A.E. and Barrangou, R. (2016) Guide RNAs: A Glimpse at the Sequences that Drive CRISPR-Cas Systems. *Cold Spring Harb Protoc*, **2016**.
  108. Chylinski, K., Makarova, K.S., Charpentier, E. and Koonin, E.V. (2014) Classification and evolution of type II CRISPR-Cas systems. *Nucleic Acids Res*, **42**, 6091-6105.
  109. Chyou, T.Y. and Brown, C.M. (2019) Prediction and diversity of tracrRNAs from type II CRISPR-Cas systems. *RNA Biol*, **16**, 423-434.
  110. Faure, G., Shmakov, S.A., Makarova, K.S., Wolf, Y.I., Crawley, A.B., Barrangou, R. and Koonin, E.V. (2019) Comparative genomics and evolution of trans-activating RNAs in Class 2 CRISPR-Cas systems. *RNA Biol*, **16**, 435-448.
  111. 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. *Nat Struct Mol Biol*, **18**, 529-536.
  112. Hayes, R.P., Xiao, Y., Ding, F., van Erp, P.B., Rajashankar, K., Bailey, S., Wiedenheft, B. and Ke, A. (2016) Structural basis for promiscuous PAM recognition in type I-E Cascade from *E. coli*. *Nature*, **530**, 499-503.

113. Xiao, Y., Luo, M., Hayes, R.P., Kim, J., Ng, S., Ding, F., Liao, M. and Ke, A. (2017) Structure Basis for Directional R-loop Formation and Substrate Handover Mechanisms in Type I CRISPR-Cas System. *Cell*, **170**, 48-60 e11.
114. Jackson, R.N., Golden, S.M., van Erp, P.B., Carter, J., Westra, E.R., Brouns, S.J., van der Oost, J., Terwilliger, T.C., Read, R.J. and Wiedenheft, B. (2014) Structural biology. Crystal structure of the CRISPR RNA-guided surveillance complex from *Escherichia coli*. *Science*, **345**, 1473-1479.
115. Mulepati, S., Heroux, A. and Bailey, S. (2014) Structural biology. Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. *Science*, **345**, 1479-1484.
116. 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.
117. 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.
118. 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 Biol*, **16**, 449-460.
119. You, L., Ma, J., Wang, J., Artamonova, D., Wang, M., Liu, L., Xiang, H., Severinov, K., Zhang, X. and Wang, Y. (2019) Structure Studies of the CRISPR-Csm Complex Reveal Mechanism of Co-transcriptional Interference. *Cell*, **176**, 239-253 e216.

120. 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.
121. Niewoehner, O., Garcia-Doval, C., Rostol, J.T., Berk, C., Schwede, F., Bigler, L., Hall, J., Marraffini, L.A. and Jinek, M. (2017) Type III CRISPR-Cas systems produce cyclic oligoadenylate second messengers. *Nature*, **548**, 543-548.
122. Niewoehner, O. and Jinek, M. (2016) Structural basis for the endoribonuclease activity of the type III-A CRISPR-associated protein Csm6. *RNA*, **22**, 318-329.
123. Rouillon, C., Athukoralage, J.S., Graham, S., Gruschow, S. and White, M.F. (2018) Control of cyclic oligoadenylate synthesis in a type III CRISPR system. *Elife*, **7**.
124. Sheppard, N.F., Glover, C.V., 3rd, Terns, R.M. and Terns, M.P. (2016) The CRISPR-associated Csx1 protein of *Pyrococcus furiosus* is an adenosine-specific endoribonuclease. *RNA*, **22**, 216-224.
125. 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.
126. Jiang, F., Taylor, D.W., Chen, J.S., Kornfeld, J.E., Zhou, K., Thompson, A.J., Nogales, E. and Doudna, J.A. (2016) Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science*, **351**, 867-871.

127. Sternberg, S.H., LaFrance, B., Kaplan, M. and Doudna, J.A. (2015) Conformational control of DNA target cleavage by CRISPR-Cas9. *Nature*, **527**, 110-113.
128. 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.
129. 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.
130. Swarts, D.C., van der Oost, J. and Jinek, M. (2017) Structural Basis for Guide RNA Processing and Seed-Dependent DNA Targeting by CRISPR-Cas12a. *Mol Cell*, **66**, 221-233 e224.
131. Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., Essletzbichler, P., Volz, S.E., Joung, J., van der Oost, J., Regev, A. *et al.* (2015) Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell*, **163**, 759-771.
132. Abudayyeh, O.O., Gootenberg, J.S., Konermann, S., Joung, J., Slaymaker, I.M., Cox, D.B., Shmakov, S., Makarova, K.S., Semenova, E., Minakhin, L. *et al.* (2016) C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science*, **353**, aaf5573.

133. 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.
134. 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.
135. Smargon, A.A., Cox, D.B.T., Pyzocha, N.K., Zheng, K., Slaymaker, I.M., Gootenberg, J.S., Abudayyeh, O.A., Essletzbichler, P., Shmakov, S., Makarova, K.S. *et al.* (2017) Cas13b Is a Type VI-B CRISPR-Associated RNA-Guided RNase Differentially Regulated by Accessory Proteins Csx27 and Csx28. *Mol Cell*, **65**, 618-630 e617.
136. Bolotin, A., Quinquis, B., Renault, P., Sorokin, A., Ehrlich, S.D., Kulakauskas, S., Lapidus, A., Goltsman, E., Mazur, M., Pusch, G.D. *et al.* (2004) Complete sequence and comparative genome analysis of the dairy bacterium *Streptococcus thermophilus*. *Nat Biotechnol*, **22**, 1554-1558.
137. Horvath, P., Romero, D.A., Coute-Monvoisin, A.C., Richards, M., Deveau, H., Moineau, S., Boyaval, P., Fremaux, C. and Barrangou, R. (2008) Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *J Bacteriol*, **190**, 1401-1412.
138. Carte, J., Christopher, R.T., Smith, J.T., Olson, S., Barrangou, R., Moineau, S., Glover, C.V., 3rd, Graveley, B.R., Terns, R.M. and Terns, M.P. (2014) The three major types of CRISPR-Cas systems function independently in CRISPR RNA biogenesis in *Streptococcus thermophilus*. *Mol Microbiol*, **93**, 98-112.

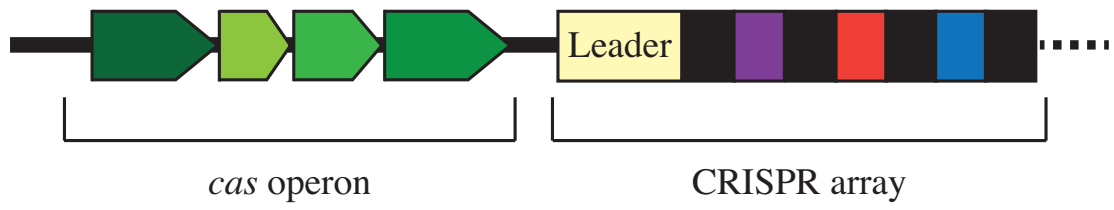
139. Arslan, Z., Wurm, R., Brener, O., Ellinger, P., Nagel-Steger, L., Oesterhelt, F., Schmitt, L., Willbold, D., Wagner, R., Gohlke, H. *et al.* (2013) Double-strand DNA end-binding and sliding of the toroidal CRISPR-associated protein Csn2. *Nucleic Acids Res*, **41**, 6347-6359.
140. Ellinger, P., Arslan, Z., Wurm, R., Tschapek, B., MacKenzie, C., Pfeffer, K., Panjikar, S., Wagner, R., Schmitt, L., Gohlke, H. *et al.* (2012) The crystal structure of the CRISPR-associated protein Csn2 from *Streptococcus agalactiae*. *J Struct Biol*, **178**, 350-362.
141. Lee, K.H., Lee, S.G., Eun Lee, K., Jeon, H., Robinson, H. and Oh, B.H. (2012) Identification, structural, and biochemical characterization of a group of large Csn2 proteins involved in CRISPR-mediated bacterial immunity. *Proteins*, **80**, 2573-2582.
142. Nam, K.H., Kurinov, I. and Ke, A. (2011) Crystal structure of clustered regularly interspaced short palindromic repeats (CRISPR)-associated Csn2 protein revealed Ca<sup>2+</sup>-dependent double-stranded DNA binding activity. *J Biol Chem*, **286**, 30759-30768.
143. 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*, **75**, 90-101 e105.
144. Hao, M., Cui, Y. and Qu, X. (2018) Analysis of CRISPR-Cas System in *Streptococcus thermophilus* and Its Application. *Front Microbiol*, **9**, 257.
145. Horvath, P. and Barrangou, R. (2010) CRISPR/Cas, the immune system of bacteria and archaea. *Science*, **327**, 167-170.



146. Young, J.C., Dill, B.D., Pan, C., Hettich, R.L., Banfield, J.F., Shah, M., Fremaux, C., Horvath, P., Barrangou, R. and Verberkmoes, N.C. (2012) Phage-induced expression of CRISPR-associated proteins is revealed by shotgun proteomics in *Streptococcus thermophilus*. *PLoS One*, **7**, e38077.
147. Amitai, G. and Sorek, R. (2016) CRISPR-Cas adaptation: insights into the mechanism of action. *Nat Rev Microbiol*, **14**, 67-76.
148. 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**.
149. McGinn, J. and Marraffini, L.A. (2019) Molecular mechanisms of CRISPR-Cas spacer acquisition. *Nat Rev Microbiol*, **17**, 7-12.
150. Sternberg, S.H., Richter, H., Charpentier, E. and Qimron, U. (2016) Adaptation in CRISPR-Cas Systems. *Mol Cell*, **61**, 797-808.
151. Terns, R.M. and Terns, M.P. (2014) CRISPR-based technologies: prokaryotic defense weapons repurposed. *Trends Genet*, **30**, 111-118.
152. Zheng, Y., Li, J., Wang, B., Han, J., Hao, Y., Wang, S., Ma, X., Yang, S., Ma, L., Yi, L. *et al.* (2020) Endogenous Type I CRISPR-Cas: From Foreign DNA Defense to Prokaryotic Engineering. *Front Bioeng Biotechnol*, **8**, 62.

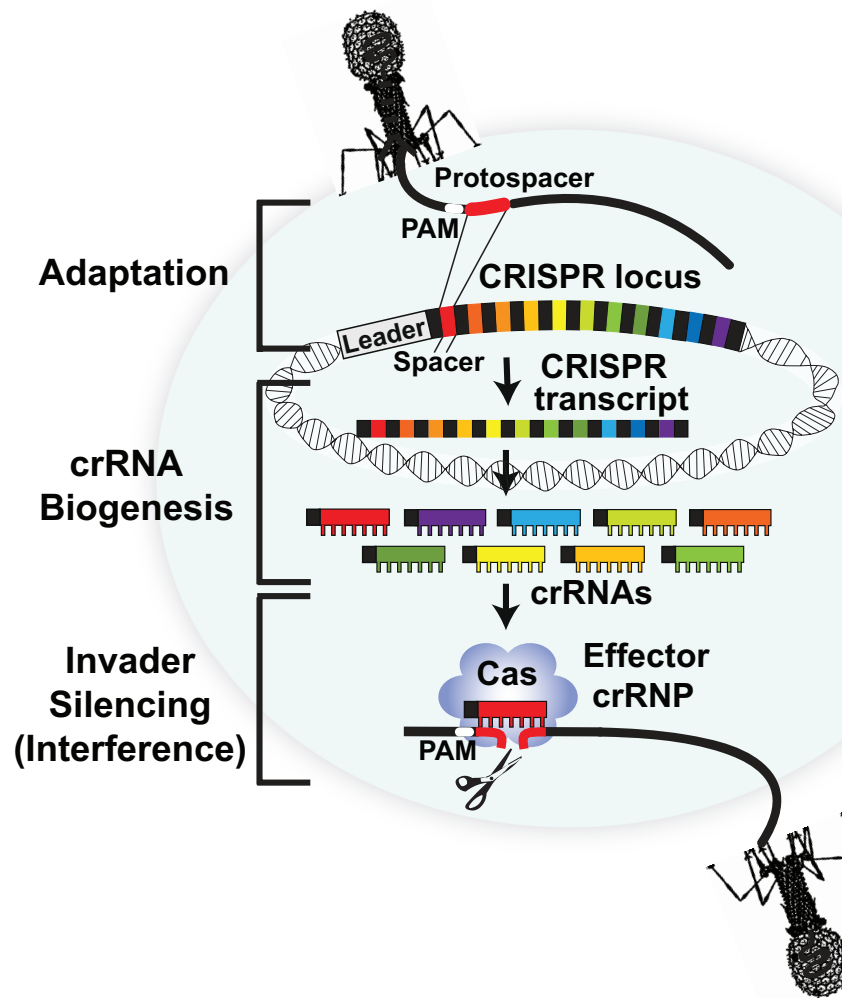
**Figure 1.1. Architecture of CRISPR-Cas loci.**

CRISPR-Cas systems are made up of two components: the *cas* operon (green) and the CRISPR array. The *cas* operon include genes expressing CRISPR-Cas proteins associated with the CRISPR array. The CRISPR array has an AT-rich leader sequence (yellow) that contains the promoter for the transcription of the CRISPR array. Downstream of the leader sequence is the CRISPR array composed of unique spacer sequences (assorted colors) that are interspaced by identical repeat sequences (black).




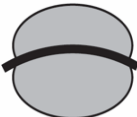
**Figure 1.2. Overview of CRISPR-Cas immunity.**

CRISPR-Cas immunity consists of three stages: Adaptation, crRNA biogenesis and invader silencing (interference). Upon infection of the cell, adaptation involves the selection of a DNA fragment termed “protospacer” from the foreign invader and integrates this fragment into the CRISPR locus at the leader-proximal repeat as a new “spacer” (red). Adjacent to the protospacer is the protospacer adjacent motif (PAM). During crRNA biogenesis, the CRISPR locus is transcribed into a pre-crRNA transcript. This transcript is further processed by Cas or host proteins to generate individual mature crRNAs that bind to CRISPR effector protein(s) to form an effector crRNP. Upon reinfection, if the corresponding PAM is present and the crRNA base-pairs with the foreign DNA or RNA, the foreign nucleic acid is cleaved. Adapted from Terns and Terns, 2014 (151).



**Figure 1.3. Classification of CRISPR-Cas systems.**

CRISPR-Cas systems are classified into 2 classes, 6 types and 33 subtypes. Class 1 systems have effector modules that are composed of multiple Cas proteins that form an effector complex with the crRNA. Class 2 systems have a single, large protein with multiple domains that bind to crRNAs. The figure further characterizes each class into types and subtypes as well as the CRISPR proteins involved in spacer acquisition, crRNA biogenesis, interference and the targeted nucleic acids. Adapted from Zheng, 2020 (152).

Class	Type	Subtype	Spacer Acquisition	crRNA biogenesis	Interference crRNP	Targeted nucleic acids
1 	I	A-G	Cas1, Cas2, Cas4	Cas6/Cas5d	Cascade	DNA
	III	A-F	Cas1, Cas2	Cas6	Csm/Cmr	DNA/RNA
	IV	A-C	Unknown	Csf5	Csf	DNA
2 	II	A-C	Cas1, Cas2, Cas4/Csn2 Cas9	RNaseIII, Cas9	Cas9	DNA
	V	A-I, K	Cas1, Cas2, Cas4	Cas12	Cas12	DNA
	VI	A-D	Cas1, Cas2	Cas13	Cas13	RNA

**Figure 1.4. Double-stranded break (DSB) repair by RecBCD or RexAB.**

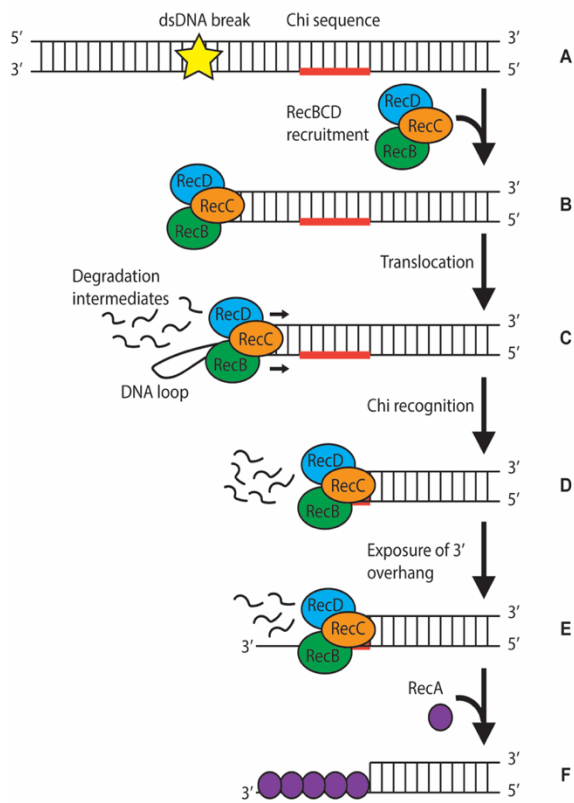
RecBCD (left panel) or RexAB/AddAB (right panel) repairs dsDNA DBSs **(A, B)**

RecBCD/RexAB is recruited to the blunt-ended duplex DNA and translocates along the DNA until a Chi sequence (red) is reached. **(B)** During DNA processing by RecBCD, the RecB (RexA) motor translocates in the 3'-5' direction while the RecD (RexB) motor moves in the 5'-3' direction resulting in the complex moving in the same direction along the DNA. As RecBCD translocates, the nuclease domain of RecB cleaves both DNA strands resulting in degraded DNA as single-stranded DNA (ssDNA) intermediates.

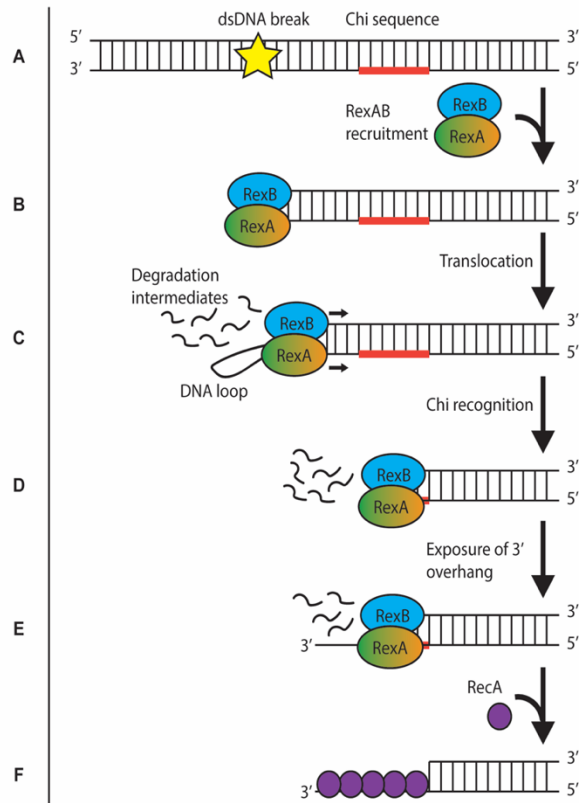
Alternatively, the nuclease domain of both RexA and RexB independently cleaves each strand of the dsDNA. **(C)** The RecD (RexB) nuclease motor translocates at a faster rate than the RecB motor, resulting in a 3' single stranded loop that forms upstream of the enzyme. **(D)** Upon reaching the Chi sequence, a conformational shift occurs within the complex and slows the complex. **(E)** Translocation at the 5' end of the duplex continues to occur to expose a 3'-overhang. **(F)** RecA proteins bind to the 3' overhang strand and generates a RecA filament which initiates homologous recombination. Adapted from Wigley, 2013(50).



### RecBCD mediated DNA repair

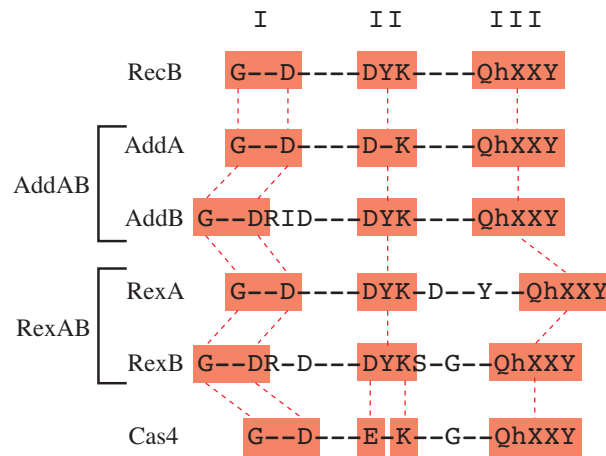
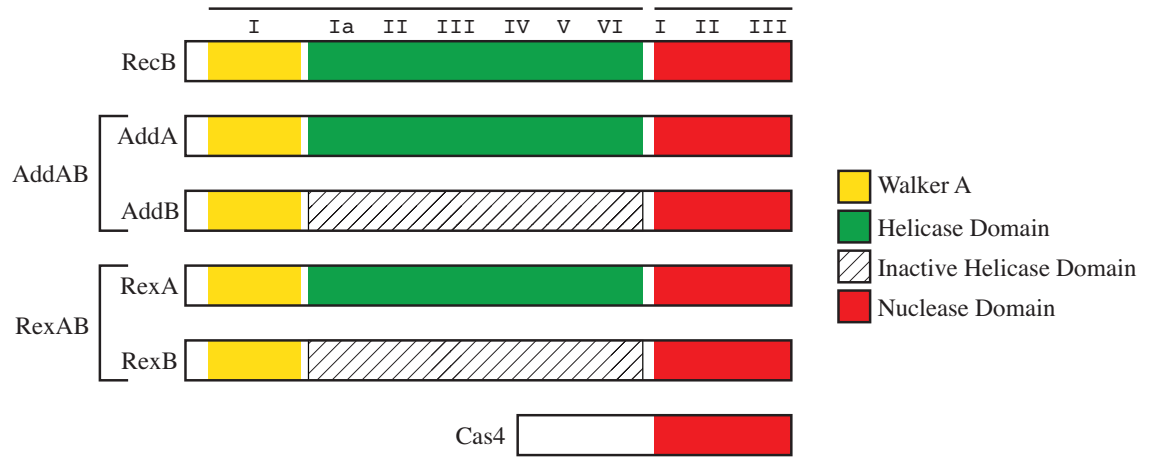


### RexAB/AddAB mediated DNA repair



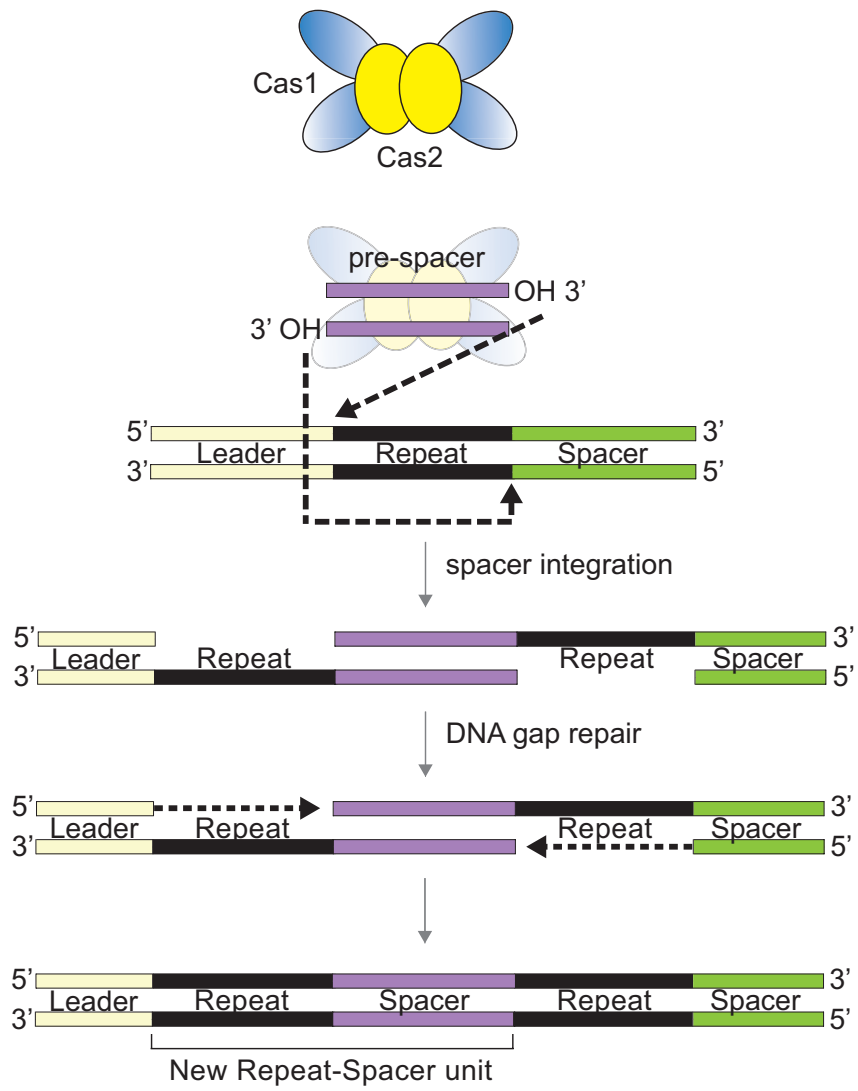
**Figure 1.5. RecB-like nuclease domains in AddAB, RexAB and Cas4.**

RecBCD and its homologs AddAB (*B. subtilis*) and RexAB (*S. thermophilus*), contain several overlapping motifs in both the helicase domain (containing the Walker A motif) and nuclease domain. The helicase domain consists of a conserved ATP-binding Walker A motif (yellow, I) at the C-terminal end of RecB and its homologs. The centralized helicase domain (green) is defined by several conserved residues (Ia, II-VI). These residues are required for functional helicase activity (green) and the absence of these motifs result in an inactive helicase (white). The most conserved domain between RecBCD and its homologs is the nuclease domain. The nuclease domain consists of 3 main motifs I (G—D), II (DYK), and III (QhXXY) that are conserved between each protein with RecB-like nuclease activity including the Cas4 CRISPR-Cas protein.



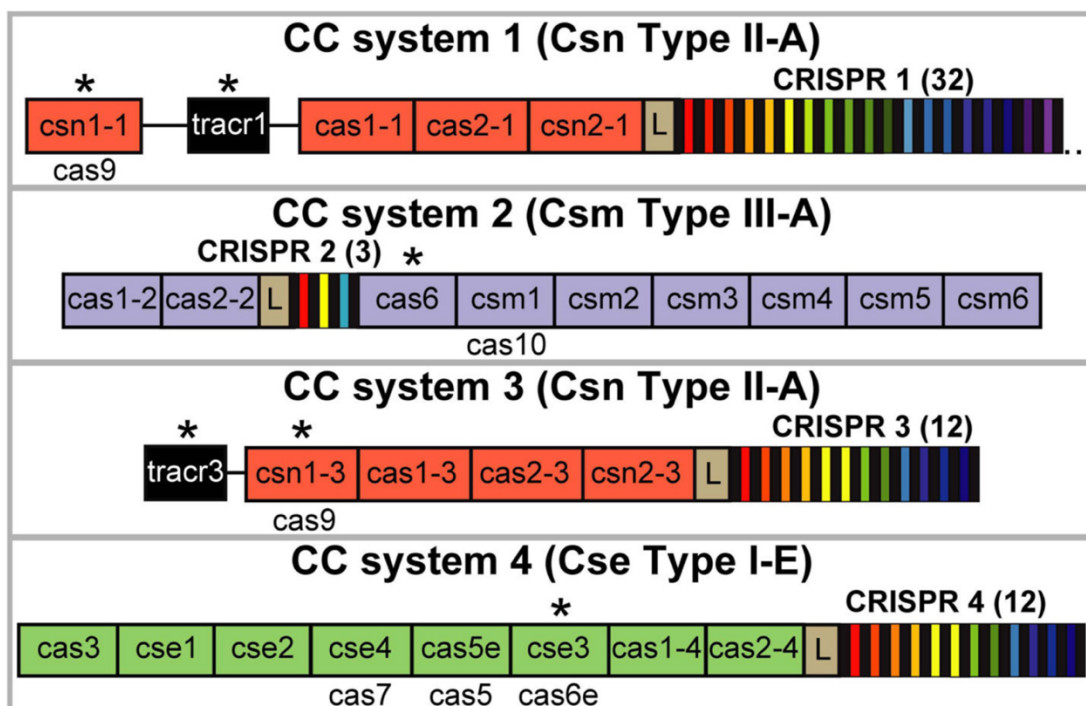
**Figure 1.6. Integration of spacers into the CRISPR array.**

The Cas1-Cas2 integrase complex binds to a dsDNA pre-spacer substrate for spacer integration into the CRISPR array at the leader-proximal repeat. During spacer integration, the 3' hydroxyl groups of the pre-spacer attacks the junctions of the repeat on the top and bottom strands. The resulting single-stranded repeats are filled in by host proteins (DNA polymerase and ligase) to allow for DNA gap repair resulting in a new repeat-spacer unit.



**Figure 1.7. CRISPR-Cas systems in *S. thermophilus* DGCC7710.**

There are 4 CRISPR-Cas systems found in *S. thermophilus* DGCC7710 strain: CRISPR1 and CRISPR3 that are both Type II-A systems, CRISPR2 a Type III-A system and CRISPR4 a Type I-E system. The CRISPR arrays are annotated with spacers (assorted colors) interspersed with repeat sequences (black). Above each array indicates the total number of spacers for each locus. All 4 CRISPR arrays contain *cas* genes associated with that CRISPR systems and genes involved in crRNA biogenesis are indicated with an asterisk. Adapted from Carte, 2014 (138).



CHAPTER 2

CRISPR DNA ELEMENTS CONTROLLING SITE-SPECIFIC SPACER

INTEGRATION AND PROPER REPEAT LENGTH BY A TYPE II CRISPR-CAS

SYSTEM<sup>1</sup>

---

<sup>1</sup>Jenny G. Kim, Sandra Garrett, Yunzhou Wei, Brenton G. Graveley, Michael P. Terns.  
Accepted by Nucleic Acids Research. Reprinted here with permission of publisher.



## Abstract

CRISPR-Cas systems provide heritable immunity against viruses by capturing short invader DNA sequences, termed spacers, and incorporating them into the CRISPR loci of the prokaryotic host genome. Here, we investigate DNA elements that control accurate spacer uptake in the type II-A CRISPR locus of *Streptococcus thermophilus*. We determined that purified Cas1 and Cas2 proteins catalyze spacer integration with high specificity for CRISPR repeat junctions. We show that 10 bp of the CRISPR leader sequence is critical for stimulating polarized integration preferentially at the repeat proximal to the leader. Spacer integration proceeds through a two-step transesterification reaction where the 3' hydroxyl groups of the spacer target both repeat borders on opposite strands. The leader-proximal end of the repeat is preferentially targeted for the first site of integration through recognition of sequences spanning the leader-repeat junction. Subsequently, second-site integration at the leader-distal end of the repeat is specified by multiple determinants including a length-defining mechanism relying on a repeat element proximal to the second site of integration. Our results highlight the intrinsic ability of type II Cas1/Cas2 proteins to coordinate directional and site-specific spacer integration into the CRISPR locus to ensure precise duplication of the repeat required for CRISPR immunity.

## Introduction

CRISPR-Cas (Clustered regularly interspaced short palindromic repeats-CRISPR-associated) systems are diverse prokaryotic defense systems that provide immunity against viruses and plasmids (1,2). These adaptive immune systems are found in roughly

half of bacteria and almost all archaea and fall into six distinct CRISPR-Cas types (I-VI) and over thirty subtypes that each utilize distinct components and mechanisms to achieve defense outcomes (3,4). CRISPR-Cas systems provide a heritable and sequence-specific method of protection against foreign invading elements by generating a memory of previous infections to elicit an effective immune response upon reinfection of the cell (5-8). Short invader-derived sequences are captured within the host CRISPR loci and used as templates to create short CRISPR RNAs that guide Cas proteins to recognize and cleave foreign genetic elements (9-15)

We are only now beginning to understand the detailed molecular mechanisms governing the capture of invader-derived sequences. This initial step in the CRISPR-Cas immune pathway is responsible for providing new, heritable immunity, and is referred to as ‘adaptation’. While details vary for the different types of CRISPR-Cas systems investigated thus far, adaptation generally involves the capture of foreign DNA and incorporation of that DNA into the host CRISPR locus, where the DNA fragments is then referred to as a ‘spacer’. The spacer sequence acts as a memory of the corresponding sequence within the foreign genome (called the protospacer). The foreign DNA must undergo processing steps prior to integration, during which time it is referred to as a pre-spacer (2,16). The CRISPR locus consists of a variably-sized leader sequence flanking an array of repeats separated by the similarly sized, previously incorporated spacers. The leader sequence harbors promoter elements used for crRNA expression as well as elements that guide the integration of new spacers at the leader-proximal repeat. (8,17-20). After addition of a new spacer, it has been shown that DNA repair machinery fills in DNA gaps and ensures faithful duplication of the CRISPR repeat such that a full repeat-

spacer unit is added to the CRISPR locus (6,21) and Figure 2.S1). It is essential that both the sequence and length of the new repeat be preserved since the CRISPR repeats function both at the RNA level in crRNA biogenesis/function (18,22,23) and at the DNA level as they are the recipient site for addition of new spacers at the CRISPR locus. New spacers are added to CRISPR arrays in a polarized manner with the vast majority of spacers being incorporated at the leader-proximal repeat rather than downstream repeats (5,6,21,24). CRISPR-captured spacers located adjacent to the leader are often more highly expressed and more efficient in mediating interference against the invading nucleic acid than spacers present near the trailer end of the CRISPR array (25,26).

While several proteins have been shown to participate in the adaptation process in the various types of CRISPR systems (27-41), Cas1 and Cas2 are core components required for spacer integration *in vivo* and *in vitro* in nearly all CRISPR systems examined to date (3). Cas1 functions as the integrase that catalyzes spacer integration into the CRISPR repeat while Cas2 appears to primarily serve a structural role in the formation of a stable Cas1-Cas2 integrase complex (42-49). In both type I and type II systems, structural studies revealed that the integrase complex consists of a Cas2 dimer sandwiched between two Cas1 dimers; the complex binds pre-spacer DNA substrates and catalyzes integration of the two pre-spacer ends into the borders of a CRISPR repeat (44,46-50). Although Cas1 and Cas2 are the most highly conserved Cas proteins, *cas1* and *cas2* gene sequences vary, and *cas1* gene variability is one important basis for classifying CRISPR systems into types and subtypes (3). Sequence differences between Cas1 and Cas2 proteins likely underlie observed functional variability observed for integration in distinct CRISPR-Cas systems.

Recent *in vitro* studies with Cas1 and Cas2 from various type I and type II systems have provided key insight into the mechanisms of spacer integration into CRISPR repeats. The Cas1 and Cas2 complex can catalyze integration of the two ends of the pre-spacer independently, with a single integrated end referred to as a half-site event (17,42,48,49). *In vitro*, half-site integrations of spacer DNA can either proceed to full-site integrations (which results in complete insertion of spacer DNA into a CRISPR repeat) or they can be reversed by Cas1-mediated disintegration (7,42,51). Productive, full-site pre-spacer integration requires two concerted transesterification reactions in which the 3'-OH groups of the pre-spacer DNA carry out nucleophilic attacks at the 5' ends of the repeat borders (17,42,49,50) and Figure 2.S1). Whether the two nucleophilic attacks required for full-site integration proceed with a set directionality or not for a given system is the subject of ongoing investigation. Recent 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 (5,24).

Both leader and repeat sequences are relatively conserved among related CRISPR-Cas systems and *in vivo* and *in vitro* mutational analyses have provided evidence for a role of leader and repeat elements in specifying accurate integration of spacer DNAs into CRISPR arrays (7,20,26,42,49,52). Depending upon the system under investigation, polarized addition of spacers at the leader-proximal repeat can be mediated either by a protein factor, illustrated by the requirement for integration host factor (IHF) in type I-E and I-F systems (28,38,48,53,54) or by intrinsic properties of the Cas1-Cas2 integrase complex (7,49). Specific elements within the repeats have been shown to govern pre-spacer integration, but motifs vary for the types of repeats investigated

(7,8,17,20,42,49,52,55-57). Of note, there is *in vitro* evidence from type I-E (56) and I-B (52) systems for key regions within the repeat that serve as ‘molecular rulers’ to guide integration to a defined distance away from these elements. It is unknown what mechanisms other CRISPR systems (e.g. type II) employ to ensure accurate integration and to maintain repeat length within a CRISPR array.

The seminal discovery that CRISPR-Cas systems function as adaptive immune systems was made following phage infections of the bacterium, *Streptococcus thermophilus* (1). *S. thermophilus* strain DGCC7710 remains one of the few organisms shown to incorporate spacer DNAs from invading viral (phage) or plasmid DNAs under laboratory conditions and without a need to overexpress adaptation proteins (1,41,58). We and others have focused attention on determining the detailed mechanism of adaptation in this organism which harbors four distinct CRISPR-Cas systems (two type II-A systems, a type III-A system and a I-E system). Both type II-A systems (CRISPR1 and CRISPR3) are active in the adaptation process (59-61). Our *in vivo* genetic analyses of the CRISPR1 system revealed that robust spacer acquisition requires Csn2 and Cas9 in addition to Cas1 and Cas2. However, it was not clear whether Csn2 and Cas9 were influencing upstream steps such as protospacer selection and processing, downstream integration, or both. (29,41). Recent *in vitro* studies with the related type II-A system in *S. pyogenes* found that Cas1 and Cas2 are sufficient for spacer integration and that Csn2 and Cas9 likely play a role in an upstream process of protospacer generation rather than being directly involved in spacer integration into CRISPR loci (7).

In *S. thermophilus*, our *in vivo* mutagenesis experiments revealed that the repeat-proximal 10 bp of the leader were necessary and sufficient to guide integration of spacers

to the leader-proximal repeat (20). Moreover, mutations at the leader-repeat junction disrupted adaptation while mutations at the repeat-spacer junction were tolerated (20). A detailed mutational analyses of the *S. thermophilus* CRISPR repeat has not yet been performed to understand the role that repeat sequences play in specifying high fidelity integration of spacer DNA precisely at the repeat borders. Figure 2.S1 displays a provisional model of *S. thermophilus* adaptation based on *in vivo* experiments conducted with *S. thermophilus* (20) and *in vitro* experiments with type I and II systems (7,44,50). To gain a more in-depth understanding of the mechanisms governing *S. thermophilus* type II-A CRISPR spacer acquisition, we reconstituted and characterized the pre-spacer integration reaction *in vitro*. Our results show that Cas1 and Cas2, likely functioning as a Cas1-Cas2 integrase complex, have an intrinsic ability to recognize sequences to catalyze pre-spacer integration with high specificity for the identical repeat junction utilized *in vivo* (20). The spacer integration reaction is a two-step process and proceeds in a directional manner whereby integration of spacer DNA at the leader-repeat junction precedes integration at the repeat-spacer junction. Our findings indicate that each integration relies on the recognition of distinct elements within the leader or repeat of the CRISPR array. Our results underscore the intrinsic capacity of *S. thermophilus* Cas1 and Cas2 to coordinate specific and directional spacer integration during the adaptation stage of CRISPR-Cas immunity.

## Materials and Methods

### *Plasmid construction*

The leader sequence and two repeat-spacer units of the CRISPR array was PCR-amplified from the *S. thermophilus* genome and cloned into the pWAR228 backbone plasmid by overlap PCR to generate pCRISPR. Leader sequence mutations were generated via inverse PCR and ligation of linearized plasmid using pCRISPR as the template. All plasmid constructs were verified by DNA sequencing and are listed in Supplemental Table 2.S1.

### *Protein purification*

The *cas1*, *csn2* and *cas9* genes were amplified by PCR from the *S. thermophilus* genome and cloned into pET expression vectors to generate 6x-histidine-tagged proteins at the C-terminus (pET21d; Cas1 and Cas9) or N-terminus (pET24d; Csn2). The *cas2* gene was subcloned into pBAT4 expression vector to generate 6x-histidine-tagged SUMO Cas2 proteins at the N-terminus (pSAT1 and pSENP kindly provided by Dr. Scott Bailey, Johns Hopkins University). Expression vectors were transformed into *E. coli* BL21-Star cells (DE3, Stratagene). Cells were grown at 37°C in 1 L cultures of Luria broth to an OD<sub>600</sub> of 0.6, and protein expression was induced overnight at room temperature by the addition of isopropylthio-β-D galactoside (IPTG) to a final concentration of 1 mM. The cells were pelleted, resuspended in lysis buffer (20 mM Tris, 500 mM NaCl, 10% glycerol, 20 mM imidazole, and 5 mM 2-Mercaptoethanol (BME), pH 7.5) and disrupted by sonication (Misonix Sonicator 3000). The lysate was cleared by centrifugation at 3,500 rpm for 20 min at 4°C and His-tagged proteins were purified by

Ni<sup>2+</sup> affinity column chromatography (1.5 mL of HisPur Ni-NTA Resin (Thermo Scientific)) using a stepwise increase of imidazole (20, 50, 100 and 500 mM). The protein samples were dialyzed at 4°C in dialysis buffer (20 mM Tris, 150 mM KCl, 10% glycerol, 5 mM 2-Mercaptoethanol (BME), pH 7.5) prior to performing activity assays. Purified proteins were analyzed by SDS-PAGE followed by Coomassie blue staining (Supplemental Figure 2.S2).

### ***Generation of DNA substrates***

DNA oligonucleotides were from Eurofins MWG Operon with the exception of hairpin DNA substrates used in Figure 2.4, which were from Integrated DNA Technologies and the sequences are given in Supplemental Table 2.S2. Oligonucleotides were annealed by an incubation temperature gradient for 1 min at 95°C decreasing by 1°C each minute, down to 23°C. Annealed double-stranded substrates were run on a non-denaturing 15% polyacrylamide gel containing 1X TBE (89 mM Tris base, 89 mM Boric acid, 2 mM EDTA, pH 8.0), followed by ethidium bromide post-staining to verify proper annealing prior to radiolabeling. The annealed DNA substrates used as pre-spacers were 5' end-labeled with T4 polynucleotide kinase (New England Biolabs (NEB)) in a 20 µL reaction containing 20 pmol oligonucleotide, 150 µCi of [ $\gamma$ -<sup>32</sup>P] ATP (6000 Ci/mmol; Perkin Elmer), 1X T4 PNK buffer, and 10 U of T4 kinase (NEB).

### ***Integration assay with radiolabeled pre-spacer***

For plasmid integration assays, individually purified recombinant Cas1 and Cas2 proteins at 2.5 µM each were added to a reaction containing 5 nM plasmid DNA, 20 nM



5' [ $\gamma$ - $^{32}\text{P}$ ] ATP-radiolabeled DNA pre-spacer substrate, and integration buffer (20 mM Tris (pH 7.5), 100 mM KCl, 10 mM  $\text{MnCl}_2$ , 5 mM 2-Mercaptoethanol). This reaction was incubated at 37°C for 1 hour and then quenched by the addition of 1  $\mu\text{g}$  Proteinase K (ThermoFisher Scientific), 0.5% SDS, 1 mM EDTA and incubated at 50°C for 30 min. The products were analyzed on a 0.8% agarose gel pre-stained with ethidium bromide. After gel electrophoresis, the gels were dried on blot absorbent filter paper (Bio-Rad) overnight at room temperature using a vacuum gel dryer (Bio-Rad, Model 583 Gel Dryer). Radioactivity was detected with a phosphorimager (Storm 840 Scanner GE Healthcare).

For linear DNA target integration assays, individually purified recombinant Cas1 and Cas2 proteins both at 250 nM were added to a reaction containing 100 nM DNA CRISPR target, and integration buffer (described above). This reaction was incubated at 25°C for 5 min. and then 20 nM 5' [ $\gamma$ - $^{32}\text{P}$ ] ATP-radiolabeled DNA pre-spacer substrates were added and 10  $\mu\text{L}$  samples were removed at 15 sec, 1 min and 15 min or incubated at 25°C for 1 hour. Reactions were quenched by the addition of equal volume (10  $\mu\text{L}$ ) of 95% formamide and 50  $\mu\text{M}$  EDTA and incubated at 98°C for 5 minutes and separated on a 12% (8.0 M urea) denaturing polyacrylamide gel. Radiolabeled Decade Markers (Life Technologies) were used to determine the size of observed products. After gel electrophoresis, the gels were dried for 1 hr at 90°C (Bio-Rad, Model 583 Gel Dryer) and radioactivity was detected by phosphorimaging as described above.

### ***Repeat mutation adaptation assay in vivo***

For *in vivo* integration assays, pCas1/Cas2/Csn2/Cas9 with a minimal CRISPR array (pCRISPR) was used as template as previously described (41) and inverse PCR was used to introduce both insertions and deletions of the repeat sequence. Plasmid constructions were verified by sequencing and transformed into *S. thermophilus* DGCC7710 strain via electroporation (59). *S. thermophilus* harboring the plasmids were grown in LM17 liquid medium supplemented with 2 µg/mL chloramphenicol for 16 hours. Cells from each strain were harvested, pelleted and genomic DNA was extracted using the Zymo Research Quick-DNA Fungal/Bacterial Miniprep Kit (Zymo Research, Irvine CA) and used as PCR template. Primers matching the leader and plasmid sequence were used for PCR amplification of the CRISPR array on the plasmid. PCR products were run on 2.5% TAE-agarose gels, pre-stained with ethidium bromide to assess CRISPR array expansion. Bands representing an expanded CRISPR array were gel excised using the Zymo Gel Extraction DNA Recovery Kit (Zymo Research, Irvine CA), purified and sequenced by high-throughput sequencing. Plasmid constructs are listed in Supplemental Table 2.S1.

### ***Pre-spacer integration high-throughput sequencing***

#### ***Library preparation.***

To sequence integration events, the spacer integration assay was performed as described above using unlabeled pre-spacer. After 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). Illumina adapter sequence with an N<sub>10</sub> random primer was annealed to the plasmid DNA and extended (thermocycler conditions: 98°C for 30 sec, 25°C for 30 sec, 35°C for 30 sec, 45°C for 30 sec, and 72°C for 5 min). Excess adapter was then removed using AMPure beads, and PCR was performed 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. The resulting amplicons captured both full-site and half-site integration events with no apparent discrimination. Illumina barcodes and adapter sequences were added with a final PCR and the resulting library was separated on a 1% agarose gel. DNA in a 400 to 700 bp size range was selected and isolated using the Zymo Gel DNA Recovery Kit (Zymo Research, Irvine CA). Sequencing was performed on an Illumina MiSeq with 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 µL 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 µL 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 and aligned to the appropriate plasmid reference using Bowtie (62). To visualize the distribution of integration events, alignment output files were converted into coverage files using bedtools (63) and displayed on a custom UCSC genome browser track hub (<https://www.genome.ucsc.edu>). 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 with bedtools and used to make sequence logos (64). 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. Integration events are displayed as the percent of total reads for each position along the CRISPR target.

### *Characterizing in vivo spacer integration into pCRISPR with repeat mutations.*

Size selected and purified array amplicon libraries were sequenced on an Illumina MiSeq with a 250 by 50 cycle run (250bp Read 1 data used in this study). Samples were de-multiplexed by barcode and then analyzed with custom python scripts to determine how new spacers were integrated. Briefly, the leader-repeat junction and the beginning of the second repeat were located in each read. The beginning of the second repeat was defined as the 3' end of a set of hypothetical spacers, which ranged in size from 27 to 33 bp. This size range captures 99.9% of new type II-A spacers observed in

spacer uptake assays with wildtype *S. thermophilus*. Each of the seven hypothetical spacers was aligned to a reference sequence including the genome and plasmid sequences using bowtie (62). Alignment outputs were then examined to determine the longest hypothetical spacer that aligned with no mismatches. This hypothetical spacer was considered the “true” new spacer and its length was used to locate the position of the repeat-spacer junction, thereby allowing us to identify the integration site for each read. The number of reads supporting integration at each position along the pCRISPR array was counted and summarized and events are displayed as the percent of total reads for each position along the pCRISPR array.

## Results

### ***S. thermophilus* Cas1 and Cas2 accurately integrate spacer DNA at the leader-proximal repeat *in vitro***

To investigate mechanisms directing spacer DNA uptake into CRISPR loci by the *S. thermophilus* type II-A CRISPR-Cas system, we established an *in vitro* system capable of accurately integrating pre-spacer (PS) donor DNA substrates into CRISPR DNA recipient molecules (Figure 2.1). Purified recombinant *S. thermophilus* Cas1 and Cas2 (Figure 2.S2) were incubated with 5'-radiolabeled double-stranded DNA pre-spacers with 5 nt 3'-overhangs and a plasmid (pCRISPR) containing a minimal CRISPR array consisting of the full, 157 bp leader and two repeat-spacer units (Figure 2.1A). The leader used in pCRISPR was either wildtype or contained blocks of transition mutations upstream of the first repeat (-32 to -21 bp (L1), -20 to -11 bp (L2) and -10 to -1 (L3); Figure 2.1C). The pre-spacer design was based on a 30 bp substrate originating from the

frequently acquired S4 sequence of the 2972 lytic phage (58), with overhangs to mimic a processed pre-spacer prior to integration. Consistent with type II-A (7) and type I-E (45) *in vitro* spacer integration assays, blunt-ended 30 bp substrates resulted in a less efficient spacer integration reaction compared to 3'-overhang substrates (data not shown). A plasmid lacking a CRISPR array (pControl) was used to observe any off-target spacer integration events. Spacer integration, as evidenced by incorporation of radiolabeled pre-spacer DNA into the recipient plasmid substrates, was observed for pCRISPR, all mutant leader variants of pCRISPR as well as pControl (devoid of a CRISPR array) (Figure 2.1B, lower panel, lanes 4-8). The formation of integration products was also deduced from changes in plasmid conformation: strand nicking during either half-site or full-site spacer integration (Figure 2.1A) converts the supercoiled (SC) plasmid into relaxed (R) forms (Figure 2.1B, upper panel, lanes 1-8). As expected, the majority of the radiolabeled integration products co-migrated with the relaxed form of the plasmid but a minor signal is observed at the position of the supercoiled form and likely reflects integration prior to relaxation of the supercoiled plasmid DNA (Figure 2.1B, lower panel, lanes 4-8). Both Cas1 and Cas2 were necessary for efficient integration, although very low levels of integration were reproducibly observed with Cas1 alone (Figure 2.1B, lower panel, compare lane 2 with lane 5) as has been observed in other *in vitro* integration studies (38,42,50).

The precise sites of all spacer integrations for each of the tested plasmids were determined by high-throughput DNA sequencing (Figure 2.1C and Figure 2.S3). Specifically, we used primers targeting the pre-spacer to make strand-specific amplicon libraries. Integration into pCRISPR occurred with high specificity for the first (leader-

proximal) repeat, occurring at the same top strand and bottom strand repeat junctions as is observed *in vivo* (Figure 2.1C (WT)) (20). Upstream leader mutations did not disrupt this specificity (L1 and L2), but integration at the leader proximal repeat was dramatically impaired when the repeat-proximal 10 bp of the leader was mutated (L3) (Figure 2.1C; Figure 2.S3), revealing that this region of the leader is critical for guiding integration to the appropriate leader-adjacent repeat. These results show that Cas1 and Cas2 are sufficient to faithfully recapitulate spacer integration at the leader-adjacent repeat of a CRISPR array as is observed *in vivo* and that the adjacent 10 bp of the leader region is critical for guiding integration to the appropriate repeat.

In addition to specific integration at the leader-proximal CRISPR repeat, we also observed low levels of integration at non-CRISPR sites that were broadly distributed throughout the plasmid backbone in both pCRISPR (containing a CRISPR array) and pControl (lacking a CRISPR array) (Figure 2.2A and Figure 2.S3). Analyses of these off-target sites, which likely represent half-site integrations, revealed a strong preference for guanine which is in agreement with the nucleotide identity of the natural *S. thermophilus* CRISPR repeat borders (Figure 2.2B). The base preference of integration was guanine (pCRISPR: 56.5%; pControl: 54.1%) followed by adenine (pCRISPR: 22.1%; pControl: 21.1%) and then cytosine (pCRISPR: 13.7%; pControl: 17.1%) and thymine (pCRISPR: 7.7%; pControl: 7.7%) (Figure 2.2B). In addition, there is an apparent preferred upstream and downstream sequence context for off-target integrations (Figure 2.2C) that resembles a leader-repeat junction sequence, further supporting an intrinsic sequence recognition by Cas1-Cas2.

### ***S. thermophilus* Cas1-Cas2 integrates pre-spacers into linear CRISPR targets**

The plasmid integration assay described above (Figure 2.1) demonstrated that *S. thermophilus* Cas1 and Cas2 show high specificity for integrating spacers at the leader-proximal repeat, but it did not allow us to distinguish half-site vs. full-site spacer integration or reveal the potential order of the two nucleophilic attacks. To address these questions, we employed a minimal linear CRISPR target consisting of 10 bp of the leader sequence, a single 36 bp repeat, and a single 20 bp spacer (Figure 2.3A). We observed specific integration of radiolabeled pre-spacers at the repeat borders of this linear CRISPR target as evidenced by bands of the expected sizes for spacer integration at the LR and RS junctions (Figure 2.3B and C). The sites of integration at the LR and RS borders were also analyzed by high-throughput sequencing, again using a strand-specific amplicon approach. Sequencing reads revealed that integration occurred precisely at the first and last nucleotides of the repeat (Figure 2.3D). We previously found that four proteins (Cas1, Cas2, Csn2, and Cas9) are required for new spacer addition to CRISPR arrays *in vivo* (41). We tested the importance of each protein for carrying out *in vitro* spacer integration (Figure 2.3C and purified proteins shown Figure 2.S2) and we found that Cas1 and Cas2 proteins are sufficient for specific integration (Figure 2.3C, lane 6). The integration levels observed with Cas1 and Cas2 appeared to be unaffected by addition of Csn2 (lane 5) and slightly enhanced by Cas9 (lane 4), through an unknown mechanism. High-throughput sequencing of the integration reaction products showed that Cas1 and Cas2 alone are capable of highly specific integration at the repeat borders with at least 92% of all pre-spacers mapping to the LR and RS repeat junctions (Figure 2.3D). These results show that sequence specificity of the *S. thermophilus* Cas1-Cas2 integrase



complex is sufficient for accurate integration into a minimal linear CRISPR target *in vitro*.

### **Full-site integration of pre-spacers by Cas1-Cas2 is directional**

We next examined if *S. thermophilus* Cas1-Cas2 was capable of catalyzing full-site and accurate spacer integrations and if there was a preference for first site integration at the leader-repeat or repeat-spacer junction (Figure 2.4). These experiments were conducted using CRISPR targets with a DNA hairpin structure at either the spacer end or leader end to enable full-site products to be distinguished from half-site products on the basis of size (Figures 2.4A and B). In addition, we compared integration patterns for pre-spacers having natural 3' hydroxyl end groups (capable of executing two nucleophilic attacks for full-site integration) with those containing a single 3' dideoxy (dd) group on one strand or the other (can undergo just one site of integration) or having dideoxy groups at both ends (to block all 3' hydroxyl-catalyzed-mediate integrations). Pre-spacers with 3'-OH termini underwent full-site integration at both LR and RS borders (Figure 2.4A). In contrast, for pre-spacers with a single modified dideoxy terminus, the majority of the integration products were leader-repeat half-site intermediates for both spacer and leader hairpin targets (Figure 2.4A, lanes 3, 4, 8 and 9) indicating that the first nucleophilic attack occurs at the leader-repeat junction rather than the repeat-spacer junction. As expected, integrations were blocked when the dideoxy was present on both strands of the pre-spacer (the low background of observable integration is likely due to lack of dideoxy groups on a small fraction of the pre-spacer substrates or a less efficient nucleophile in the reaction (e.g. H<sub>2</sub>O)) (Figure 2.4A, lanes 2 and 7).

A time course analysis of the integration reaction with pre-spacers with 3'-OH termini provided additional evidence that the leader-repeat junction is preferentially recognized vs. the repeat-spacer junction (Figure 2.4B). Within 6 seconds of initiating the reaction, leader-repeat half-site intermediates were the most abundant product and accumulated prior to the appearance of repeat-spacer half-site intermediates and full-site (LR + RS) integration products (Figure 2.4B, top panel). Quantification of the half-site integration intermediates and full-site products with time showed that leader-repeat half-site intermediates are the most abundant products throughout the reaction and that progression to full-site integration correlated with a steady decrease in leader-repeat half-site integrations (Figure 2.4B, bottom panel). Together, these results show that full-site spacer integration reactions facilitated by *S. thermophilus* Cas1-Cas2 proceed in a sequential manner with the first reaction occurring at the leader-repeat junction followed by a subsequent second integration at the repeat-spacer junction.

### **Important elements of the CRISPR repeat**

Having determined the important role of the first 10 bp of the leader sequence in directing spacer integrations at the leader-proximal repeat (Figure 2.1), we next investigated sequence determinants within the repeat important for guiding integration by the Cas1-Cas2 complex (Figure 2.5). We introduced a series of block substitution mutants to a minimal linear CRISPR target (Figure 2.5A) and evaluated the effects of each mutation relative to the wildtype repeat, on spacer integration efficiency (level of integration products observed by gel separation and autoradiography; Figure 2.5B) and specificity (location of integration determined and quantified through strand-specific

sequencing; Figure 2.5C and see Figure 2.S4 for detailed mapping of integration sites). Mutation of sequences spanning the leader-repeat junction (mutant B1) abolished spacer integration at both LR and RS junctions with only a relatively moderate reduction in overall integration efficiency. Likewise, mutation of leader-proximal region of the repeat (mutant B2) also resulted in a similar loss of specificity at both junctions of the repeat and significantly impaired the efficiency of integration. Mutation of a mid-repeat sequence block towards the leader (mutant B3) did not significantly impact integration specificity or integration efficiency. However, mid-repeat sequence mutations towards the spacer end of the repeat (mutant B4) as well as for mutations in one (mutants IR 1 and IR 2) or both (mutant IR 3) of the palindromic repeats did not significantly impact integration specificity, despite leading to a significant reduction in the efficiency of spacer integration at the second site of the repeat-spacer junction. Mutation of a sequence block adjacent to the repeat-spacer border (mutant B5) resulted in a loss of specificity at the second site of the repeat-spacer junction but not the first (LR) site of integration while efficiency at the second site was significantly reduced. We note that mutations that affect specificity of the first site of integration at the LR junction (e.g. mutants B1, B2 and to a lesser extent B3) also resulted in a loss of specificity at the second site of integration at the RS border. The inverse was not true as illustrated by the block 5 mutant (B5) which is capable of integration at the LR but not RS junctions. Furthermore, we observed a relatively prominent aberrant integration eight bases downstream of the RS border at a guanine in the spacer region occurring for both B1 and B2 (and to a reduced degree with B3) mutants that is not observed with the WT repeat (Figure 2.5C and Figure 2.S4). Together, the results support a role for several repeat sequences in determining the

efficiency and/or specificity of integrations by Cas1-Cas2 and also provide support for a two-step integration reaction whereby accurate first step integration is a prerequisite for achieving accurate full-site integration.

Next, to understand if the identity of the two nucleotides which serve as the sites of nucleophilic attack during spacer integration (G1 on the top strand and G36 on the bottom strand), is important for directing integration by Cas1-Cas2, we assayed each possible combination of nucleotides at these two positions (Figure 2.6 and see Figure 2.S4 for detailed mapping of integration site). Mutation of the guanine in position 1 to a cytosine (G1C) or adenine (G1A) did not impact integration specificity at either LR or RS junction, while a moderate defect in specificity at both junctions was observed for the thymine substitution (G1T) at position 1 (Figure 2.6C) and this led also to aberrant integration within the spacer region at a guanine (Figure 2.6C and Figure 2.S4). A reduction in integration efficiency was observed for both the G1C and G1T mutations but not the G1A mutation (Figure 2.6B). At the last position of the repeat, mutation from a guanine to all other nucleotides (G36C, G36A and G36T) did not significantly affect integration specificity at either junction of the repeat or integration efficiency at the first site (LR border). However, all three changes to nucleotide 36 impaired efficiency of integration at the second site (RS border) (Figure 2.6B, 2.6C). Thus, the identity of the base at the sites of transesterification attack on the CRISPR repeat is an important component for specifying efficient and/or specific integration at a CRISPR repeat by the Cas1-Cas2 complex.

## **Second-site integration is defined by a molecular ruler-based mechanism**

Our findings support a model where full-site spacer integration proceeds in a directional manner such that integration at the leader-repeat junction (site 1) occurs prior to integration at the repeat-spacer junction (site 2) with the first integration being governed by sequence-specific interactions of Cas1-Cas2 and sequences spanning the leader-repeat junction. We next investigated how the second site of integration at the far end of the repeat (G36) is orchestrated (Figure 2.7). Similar to what has been observed for *in vivo* type I-E and I-B adaptation studies (52,56), we tested whether the site of the second integration would be directed by a region within the repeat that acts as a molecular ruler to determine the distance of the second nucleophilic attack in a sequence-independent fashion. In the type I studies, the repeat regions determined to act as molecular rulers were identified by testing the effects of strategically located nucleotide insertions or deletions within repeats on defining the site of the second step of integration. Altering the length of the repeat upstream of the ruler element shifted the second integration site upstream or downstream a fixed distance dictated by the length of the insertion or deletion. In contrast, insertions or deletions downstream of the ruler element resulted in second-site integrations occurring at a fixed short distance (typically 8-10 bp depending on the system) downstream of the motif to a common location (52,56). Accurate integration precisely at the two repeat borders is required to maintain repeat length which in turn is critical for generating a functional CRISPR array capable of producing active crRNAs as well as accepting spacers from new viral invaders.

To test the hypothesis that type II repeats harbor an element that serves as a molecular ruler defining the second site of integration (in this case G36), single cytosine

residues were inserted at regular intervals across the 36 bp repeat and the sites of integration for each mutant were quantified to determine any effects on the choice of second site integration (Figure 2.7A and see Figure 2.S4 for detailed integration site mapping). None of the C insertions impacted accurate integration at the leader-repeat border (site 1 at G1). However, significant differences were observed in the location of the second site of integration (site 2) depending upon if the C were inserted before or after position 28 (Figure 2.7A). Specifically, we found that when C insertions were introduced at locations upstream of position 28 (mutants C5, C9, C14, C19, C24, C28), then second-site integration occurred one base further down (i.e. position 37) than WT repeat (position 36) and there were spurious sites not observed with WT repeats (see Figure 2.S4 for locations of all sites of integration). Moreover, deletion of a C upstream of position 28 (mutant Del14 at position 14) resulted in a shift in the second site of integration one base upstream (i.e. position 35) than WT repeat (position 36). In contrast, when the C insertions were performed at or downstream of position 33 (mutants C33, C36,), integrations occurred at the same site as the WT repeat (position 36). Additionally, we found that the second site of integration remained at the 36<sup>th</sup> position of the repeat even when single, double or triple insertion (mutants Ins36, Ins36-37, Ins36-38) or deletions (mutants Del35, Del34-35, Del33-35) were introduced with low levels of aberrant integration (Figure 2.7A and B). Together, the results indicate that the upstream region of the repeat can tolerate changes in length, but downstream of the 28-32 region, insertions or deletions result in off-site integration at any nucleotide 8 base-pairs away from position 28 of the repeat.

### ***In vivo* evidence that the second-site integration step is governed by a molecular ruler-based mechanism**

Finally, we tested if the ruler-based mechanism governing the site of the second step of *in vitro* integrations also operates *in vivo*. Similar to our *in vitro* mutational analysis (Figure 2.7), we introduced nucleotide insertions or deletions, both upstream and downstream of the ruler element located between position 28 and 32, into repeats on pCRISPR (Figure 2.8). Plasmids were then transformed into *S. thermophilus* cells and new spacer acquisition was determined using a PCR based approach (20) combined with high-throughput sequencing of the expanded CRISPR arrays. Expanded arrays were observed for all mutants, however the overall efficiency of spacer integration was often noticeably reduced (Figure 2.S5). None of the insertion or deletion mutations affected the accuracy of integration at the first leader-repeat (LR) border and the downstream spacer-repeat (SR) border was also preserved (Figure 2.8A). In contrast, differences in the second site of integration at the repeat-spacer (RS) border were observed for some of the repeat mutants (Figure 2.8A). Similar to what we observed *in vitro* (Figure 2.7), single nucleotide insertion (Ins C24) or deletion (Del A23) upstream of the ruler element resulted in a corresponding shift in the site of integration by one nucleotide downstream (position 37) or upstream (position 35) compared to WT (position 36), respectively. We note that integration at position 36 for these two mutants was also observed at a relatively high level compared to what was observed *in vitro*, indicating that compensatory mechanisms appear to operate *in vivo* to find the natural RS junction despite the introduced point mutations within the repeats. For the insertion mutant (Ins C24), we also observed significant aberrant second-site reactions (marked “other” in Figure 2.8B) that

mostly correlate to an attack within the spacer at position 45 which is the same guanine observed with our *in vitro* mutational results (eight bases downstream of the RS border in the spacer region, see Figure 2.S6B and see Figure 2.S4 for detailed integration site mapping). This improper second-site reaction results in partial duplication of the spacer during full-site integration, and is not observed in WT (Figure 2.S6A). As predicted for the molecular ruler model, insertions (Ins C33, Ins CG 33-34) or deletions (Del C33, Del CA 33-34) downstream of the ruler element did not affect site of integration at the second-site and maintained a preference similar to WT for position 36. Sequencing results showed that positioning was maintained by either the loss of 1 nucleotide from the 3' end of the mutant repeat or the addition of 1 nucleotide corresponding to the first base of the previous spacer. Together, these results show that *in vivo*, second-site integration is influenced by a ruler-based distance mechanism.

## Discussion

Successful acquisition and integration of new spacers into the CRISPR locus is a fundamental step for heritable CRISPR-Cas immunity against viruses and other potentially harmful or lethal mobile genetic elements. With each new spacer acquired, there is an accompanying duplication of the repeat due to DNA repair of the gapped DNA intermediate containing the integrated spacer flanked on either side by single-stranded repeat sequences (Figure 2.S1). When spacer integration occurs accurately at 5' nucleotides that comprise the leader-repeat (LR) and repeat-spacer (RS) borders, DNA repair processes (polymerase fill-in and ligation reactions) yield a new repeat that is a perfect copy of the original repeat (8,30). Subsequently, the newly generated repeat at the



leader end of the CRISPR array is competent to function as the recipient structure for subsequent addition of the next spacer. The periodicity of the repeat-spacer units of the entire CRISPR repeat is maintained even after multiple novel spacer additions. In addition to its role in permitting accurate array expansion, the repeat could also influence the biogenesis of functional crRNAs. Transcribed type II repeat sequences must match and bind tracrRNA, be processed by RNase III, and ultimately portions of the repeat RNA (referred to as 5' or 3' 'tags' or 'handles') are key elements of mature crRNAs and are critical for crRNA-Cas protein assembly and function in crRNA-guided invader nucleic acid destruction (11,12,65,66). Thus, imprecise full-site integration of spacers has the potential to lead to inactive CRISPR arrays and/or non-functional crRNAs.

Our work provides the first *in vitro* characterization for spacer integration for the type II-A CRISPR-Cas system of *Streptococcus thermophilus*. We established an *in vitro* system capable of accurately integrating full-site spacer DNA at the proper junctions of *S. thermophilus* CRISPR repeats and importantly, our characterization of the reaction revealed key mechanistic information for how Cas1 and Cas2 accurately integrate new spacers in a polarized manner at the leader-adjacent repeat. Our approach of analyzing *in vitro* integration products through gel electrophoresis to address efficiency, combined with sequencing to address integration site specificity, provided a more comprehensive approach to studying integration than previous studies that relied on either gel analysis or sequence analysis alone (7,49,52,56).

## Model for type II spacer integration at CRISPR arrays

Our results are consistent with a model (Figure 2.9) whereby *S. thermophilus* Cas1 and Cas2, likely functioning as a Cas1-Cas2 integrase complex that binds the spacer substrates (49), catalyze spacer integration specifically at the leader-proximal repeat through a two-step transesterification reaction. The 3' hydroxyl groups of the DNA spacer each carry out nucleophilic attacks at the borders of the first repeat sequence, on opposite strands (Figure 2.9A, B and C). Several lines of evidence indicate that there is an apparent obligate order to the two nucleophilic attacks whereby the first attack occurs on the top strand at the guanine of the leader-repeat junction (LR) and the second attack is made at the guanine of the repeat-spacer junction (RS) on the bottom strand (Figure 2.9B and C). For example, integration occurred selectively at the LR rather than at the RS junction when pre-spacers had only a single unmodified dideoxy terminus available for nucleophilic attack. (Figure 2.4). Furthermore, LR integrations temporally precede both RS and full-site integration when the reactions were performed with pre-spacers capable of catalyzing both transesterification reactions (Figure 2.4). Additionally, repeat mutations that prevented LR integration also resulted in loss of accurate RS integrations. Moreover, we observed mutations that preserved LR integrations but blocked or altered the site of RS integration but never *vice versa* in both our *in vitro* (Figures 2.5, 2.6 and 2.7) and *in vivo* analyses (Figure 2.8). Finally, off-target (non-CRISPR) plasmid DNA integrations mapped to a short stretch of sequences that match the LR junction and flanking upstream and downstream nucleotides rather than the RS junction and surrounding sequences, indicating that *S. thermophilus* Cas1-Cas2 integrase exhibits intrinsic sequence recognition of sequences spanning the LR border (Figure 2.2). A

similar preference for integration at the LR vs. RS site was observed *in vitro* for *Streptococcus pyogenes* (7) and *Enterococcus faecalis* (49) type II Cas1-Cas2 integrases. Together, the findings indicate that spacer integration into type II CRISPR repeats normally proceeds with directionality such that the LR junction is initially selected for half-site integration and additional determinants (discussed below) govern the next attack at the RS site that results in a full-site spacer integration at the repeat (Figure 2.9C).

### **First-site integration: Leader-repeat junction**

Our results indicate that recognition of a DNA element at the leader-repeat junction, composed of at most 10 bp of the leader and 5 bp of the repeat, is critical for guiding *S. thermophilus* Cas1-Cas2 to make the first step of the two-step, full-site integration reaction at the leader-proximal repeat (Figure 2.9A). Mutational analyses both *in vivo* (20) and *in vitro* (Figure 2.1C) demonstrated that 10 bp of the leader proximal to the repeat are necessary and sufficient for directing integration. Moreover, specific block mutations within the repeat immediately downstream of the LR junction disrupted overall efficiency and specific integration at the LR (and RS) junctions while block mutations elsewhere did not prevent accurate LR integrations (Figure 2.5). As described above, off-target integration events revealed that the *S. thermophilus* Cas1-Cas2 integrase targets sequences that mimic the leader-repeat junction (guanine) which includes ~ 5 bp of the upstream leader and 5 bp of the downstream repeat (Figure 2.2C). The recognition of this leader-repeat element by a type II Cas1-Cas2 integrase has been captured by recent X-ray crystallographic structures and revealed base-specific DNA contacts of Cas1 at positions -1 through -4 of the leader as well as +1 and +2 of the repeat (49). The preference for a

guanine for the site of integration at both off-target and LR and RS junctions (Figure 2.2 and 2.6) appears to be a common determinant for Cas1 proteins of diverse CRISPR systems (7,38,42,49-52,67). In agreement with our findings, other type II-A studies showed that the first 5 bp of the leader sequence specifies sites of integration *in vivo* and that mutations of the leader-proximal repeat sequences affects integration efficiency *in vitro* (7,26,49). Collectively, the results provide strong evidence that type II Cas1-Cas2 proteins have evolved to integrate at the leader-proximal repeat rather than downstream repeats of the CRISPR array via direct recognition of sequences spanning the leader-repeat junction. This contrasts the mechanisms revealed for other (type I) systems that that rely on additional factors such as IHF (integration host factor) that bind at the leader and direct Cas1-Cas2 to integrate at the leader-proximal (first) repeat (28,38,48,53,54).

### **Second-site integration: Repeat-spacer junction**

Once the first step of integration is complete, the remaining 3'-OH terminus of the covalently linked spacer normally performs the second nucleophilic attack precisely at the guanine of the RS border on the opposite strand (Figure 2.9C and Figure 2.S1). Our results suggest that the second site of nucleophilic attack is influenced by multiple factors: 1) it is likely sterically restrained to a relatively narrow range of nucleotides a set distance from the first integration position, 2) it is further specified by a preference for guanine over other bases (Figure 2.6), 3) it depends upon several determinants within the repeat that likely make contacts with the Cas1-Cas2 integrase (49) to lead to directionality and specificity of the second nucleophilic attack, and 4) it is influenced by an element located just upstream of the RS junction (between positions 28-32) that

defines integration a fixed distance of 8 bp downstream of the 5' border of the element both *in vitro* and *in vivo* (Figures 2.7, 2.8 and 2.9C).

Structural studies of a type II Cas1-Cas2 integrase bound to spacer and target DNA during full-site integration (49) suggest that bending of the repeat is necessary for accurate second-site integration (Figure 2.9C). Moreover, the structural information showed that contacts between the Cas1-Cas2 integrase and the majority of the repeat are mediated by sugar-phosphate backbone interactions rather than base-specific contacts. Second-site recognition appears to be reliant on an accurate first integration step and further guided by multiple determinants distributed throughout the repeat that likely influence repeat bending and positioning of the Cas1 active site at the appropriate guanine residue at the RS border. We noted that the structure showed contact between a non-catalytic Cas1 and repeat residues that correspond to positions 28-29 of the repeat in our experiments. In light of this structural information, it is conceivable that the ruler element that we identified through mutational analysis may represent the breakpoint between the region of the repeat that interacts with the Cas1-Cas2 integrase and the region of the repeat that projects out towards the catalytic Cas1 for second-site integration. The spacing between the non-catalytic Cas1 contact point (with positions 28-29) and the active site of the catalytic Cas1 may correspond to the 8 bp ruler element that we observe for our repeat sequence.

### **Type II pre-spacer integration**

In summary, we have characterized the *in vitro* properties of *Streptococcus thermophilus* Cas1 and Cas2 and found that these two proteins collaborate to catalyze

accurate and full-site spacer integration into CRISPR arrays. Our results revealed that type II systems appear to be unique from well-studied type I systems in that the type II Cas1-Cas2 integrases exhibit an intrinsic specificity for LR junctions that drives integration into the leader-proximal repeat instead of downstream repeats and an intrinsic directionality such that the first transesterification reaction is at the LR junction and step two follows at the RS junction. We provide the first evidence supporting a molecular ruler-based mechanism in a type II system that helps guide the second step a fixed distance downstream and functions to maintain the repeat length (Figures 2.7, 2.8 and 2.9). Such second-site, molecular ruler elements were previously demonstrated to function within type I systems (52,56). Understanding the molecular basis of the ruler-based mechanism that guides the second integration step is an important future goal that will likely require structural and molecular analyses. Future studies are also needed to understand key steps that function upstream of CRISPR spacer integration. For example, there is a gap of knowledge in understanding how viral or plasmid protospacers are recognized and properly processed prior to binding by the Cas1-Cas2 integrase. Furthermore, there is a need for determining the specific roles that Cas9 (29,41) and Csn2 (47,68,69) and perhaps additional host factors play in protospacer to pre-spacer generation and precise PAM removal required for directing spacer integration in a functional orientation in type II CRISPR arrays.

## 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. Koonin, E.V., Makarova, K.S. and Zhang, F. (2017) Diversity, classification and evolution of CRISPR-Cas systems. *Curr Opin Microbiol*, **37**, 67-78.
4. 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.
5. McGinn, J. and Marraffini, L.A. (2019) Molecular mechanisms of CRISPR-Cas spacer acquisition. *Nat Rev Microbiol*, **17**, 7-12.
6. Sternberg, S.H., Richter, H., Charpentier, E. and Qimron, U. (2016) Adaptation in CRISPR-Cas Systems. *Mol Cell*, **61**, 797-808.
7. Wright, A.V. and Doudna, J.A. (2016) Protecting genome integrity during CRISPR immune adaptation. *Nat Struct Mol Biol*, **23**, 876-883.
8. 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.

9. 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.
10. 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.
11. 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.
12. 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.
13. 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.
14. Marraffini, L.A. (2015) CRISPR-Cas immunity in prokaryotes. *Nature*, **526**, 55-61.
15. 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.



16. 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.
17. 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.
18. Hale, C., Kleppe, K., Terns, R.M. and Terns, M.P. (2008) Prokaryotic silencing (psi)RNAs in *Pyrococcus furiosus*. *RNA*, **14**, 2572-2579.
19. Jansen, R., Embden, J.D., Gaastra, W. and Schouls, L.M. (2002) Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol*, **43**, 1565-1575.
20. 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.
21. Amitai, G. and Sorek, R. (2016) CRISPR-Cas adaptation: insights into the mechanism of action. *Nat Rev Microbiol*, **14**, 67-76.
22. Carte, J., Christopher, R.T., Smith, J.T., Olson, S., Barrangou, R., Moineau, S., Glover, C.V., 3rd, Graveley, B.R., Terns, R.M. and Terns, M.P. (2014) The three major types of CRISPR-Cas systems function independently in CRISPR RNA biogenesis in *Streptococcus thermophilus*. *Mol Microbiol*, **93**, 98-112.
23. 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.
24. 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**.
  25. 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.
  26. 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.
  27. Dillard, K.E., Brown, M.W., Johnson, N.V., Xiao, Y., Dolan, A., Hernandez, E., Dahlhauser, S.D., Kim, Y., Myler, L.R., Anslyn, E.V. *et al.* (2018) Assembly and Translocation of a CRISPR-Cas Primed Acquisition Complex. *Cell*, **175**, 934-946 e915.
  28. 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.
  29. 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.

30. Ivancic-Bace, I., Cass, S.D., Wearne, S.J. and Bolt, E.L. (2015) Different genome stability proteins underpin primed and naive adaptation in *E. coli* CRISPR-Cas immunity. *Nucleic Acids Res*, **43**, 10821-10830.
31. 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.
32. Krivoy, A., Rutkauskas, M., Kuznedelov, K., Musharova, O., Rouillon, C., Severinov, K. and Seidel, R. (2018) Primed CRISPR adaptation in *Escherichia coli* cells does not depend on conformational changes in the Cascade effector complex detected in Vitro. *Nucleic Acids Res*, **46**, 4087-4098.
33. Kunne, T., Kieper, S.N., Bannenberg, J.W., Vogel, A.I., Mielliet, 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.
34. 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.
35. 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.
36. Radovic, M., Killelea, T., Savitskaya, E., Wettstein, L., Bolt, E.L. and Ivancic-Bace, I. (2018) CRISPR-Cas adaptation in *Escherichia coli* requires RecBCD helicase but not nuclease activity, is independent of homologous recombination,

- and is antagonized by 5' ssDNA exonucleases. *Nucleic Acids Res*, **46**, 10173-10183.
37. 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.
  38. 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.
  39. Rollins, M.F., Chowdhury, S., Carter, J., Golden, S.M., Wilkinson, R.A., Bondy-Denomy, J., Lander, G.C. and Wiedenheft, B. (2017) Cas1 and the Csy complex are opposing regulators of Cas2/3 nuclease activity. *Proc Natl Acad Sci U S A*, **114**, E5113-E5121.
  40. 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.
  41. Wei, Y., Terns, R.M. and Terns, M.P. (2015) Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. *Genes Dev*, **29**, 356-361.
  42. Grainy, J., Garrett, S., Graveley, B.R. and M, P.T. (2019) CRISPR repeat sequences and relative spacing specify DNA integration by *Pyrococcus furiosus* Cas1 and Cas2. *Nucleic Acids Res*.
  43. Lee, H., Dhingra, Y. and Sashital, D.G. (2019) The Cas4-Cas1-Cas2 complex mediates precise prespacer processing during CRISPR adaptation. *Elife*, **8**.

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. *Nat Struct Mol Biol*, **21**, 528-534.
45. 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.
46. 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.
47. 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*.
48. 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.
49. 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.
50. Nunez, J.K., Harrington, L.B., Kranzusch, P.J., Engelman, A.N. and Doudna, J.A. (2015) Foreign DNA capture during CRISPR-Cas adaptive immunity. *Nature*, **527**, 535-538.

51. 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**.
52. 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.
53. 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.
54. 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.
55. Diez-Villasenor, C., Guzman, N.M., Almendros, C., Garcia-Martinez, J. and Mojica, F.J. (2013) CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of *Escherichia coli*. *RNA Biol*, **10**, 792-802.
56. 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.
57. 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.

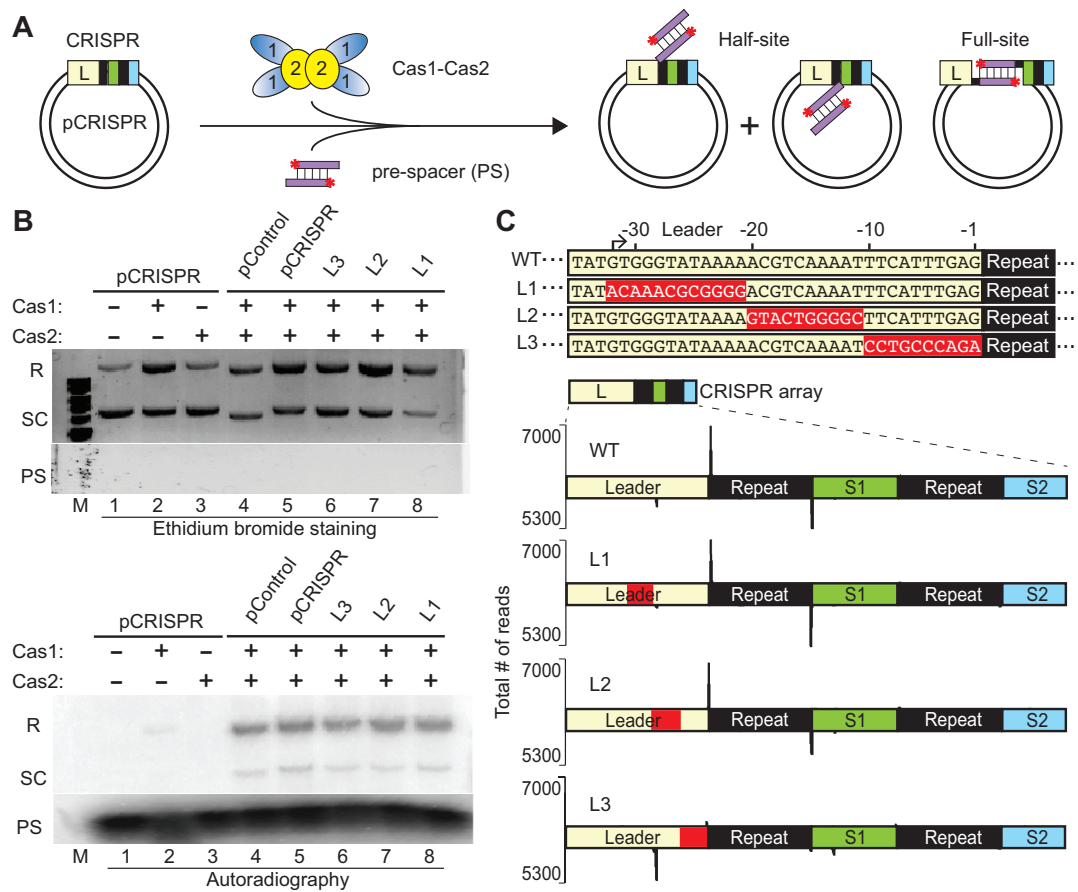
58. Deveau, H., Barrangou, R., Garneau, J.E., Labonte, J., Fremaux, C., Boyaval, P., Romero, D.A., Horvath, P. and Moineau, S. (2008) Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J Bacteriol*, **190**, 1390-1400.
59. Garneau, J.E., Dupuis, M.E., Villion, M., Romero, D.A., 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.
60. Horvath, P., Romero, D.A., Coute-Monvoisin, A.C., Richards, M., Deveau, H., Moineau, S., Boyaval, P., Fremaux, C. and Barrangou, R. (2008) Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *J Bacteriol*, **190**, 1401-1412.
61. Magadan, A.H., Dupuis, M.E., Villion, M. and Moineau, S. (2012) Cleavage of phage DNA by the *Streptococcus thermophilus* CRISPR3-Cas system. *PLoS One*, **7**, e40913.
62. 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.
63. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841-842.
64. Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) WebLogo: a sequence logo generator. *Genome Res*, **14**, 1188-1190.

65. 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.
66. 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.
67. 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.
68. Arslan, Z., Wurm, R., Brener, O., Ellinger, P., Nagel-Steger, L., Oesterhelt, F., Schmitt, L., Willbold, D., Wagner, R., Gohlke, H. *et al.* (2013) Double-strand DNA end-binding and sliding of the toroidal CRISPR-associated protein Csn2. *Nucleic Acids Res*, **41**, 6347-6359.
69. Ka, D., Lee, H., Jung, Y.D., Kim, K., Seok, C., Suh, N. and Bae, E. (2016) Crystal Structure of *Streptococcus pyogenes* Cas1 and Its Interaction with Csn2 in the Type II CRISPR-Cas System. *Structure*, **24**, 70-79.



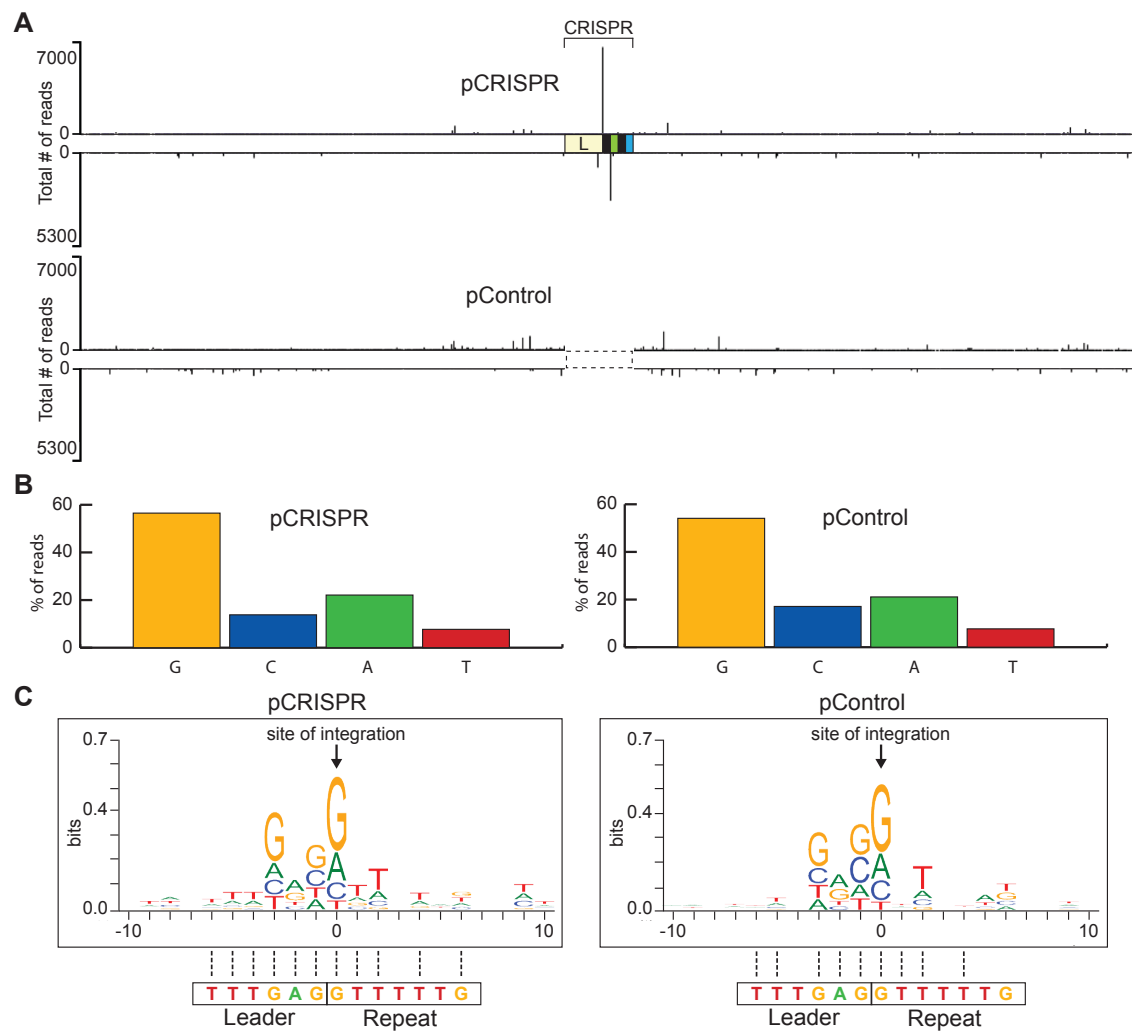
**Figure 2.1. *S. thermophilus* Cas1 and Cas2 accurately integrates pre-spacers *in vitro* and 10 bp of the leader sequence is essential for polarized integration.**

**(A)** Schematic of pre-spacer (PS) integration by Cas1-Cas2 into a plasmid target containing a minimal CRISPR array (pCRISPR). Integration of pre-spacers can occur as half-site intermediates at either junction of the repeat or as full-site products. **(B)** Integration assays with Cas1-Cas2 and radiolabeled pre-spacers visualized with ethidium bromide staining and autoradiography. Integration products corresponding to relaxed plasmids (R), unintegrated supercoiled plasmid (SC) and free pre-spacers (PS) are indicated. **(C)** Variants of the leader sequence mutations (L1, L2, L3) engineered on pCRISPR. Sites of spacer integration were identified by high-throughput sequencing and mapped to the plasmids on the plus (upper) and minus (lower) strands.



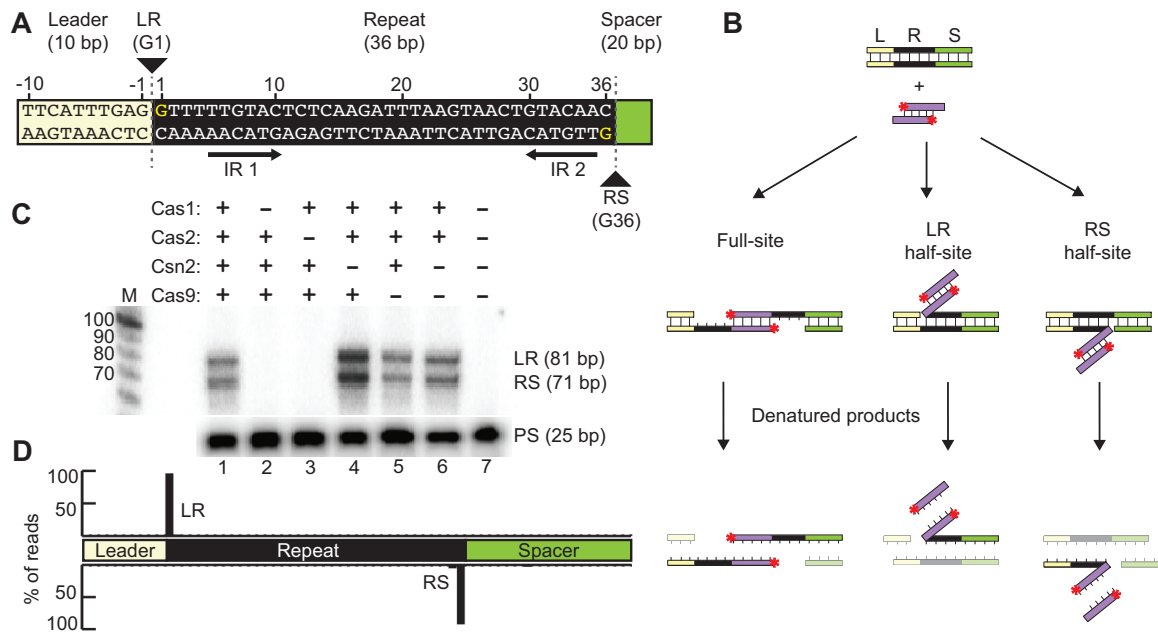
**Figure 2.2. Non-CRISPR integration sites resemble a leader-repeat junction.**

**(A)** Sites of spacer integration for pCRISPR and pControl were identified by high-throughput sequencing. **(B)** Percent of spacer integration sites occurring at a guanine, cytosine, adenine, and thymine on pCRISPR and pControl. **(C)** WebLogo of all non-CRISPR integration sequences. Sequence homology to the actual leader-repeat junction sequence is indicated with dotted lines.



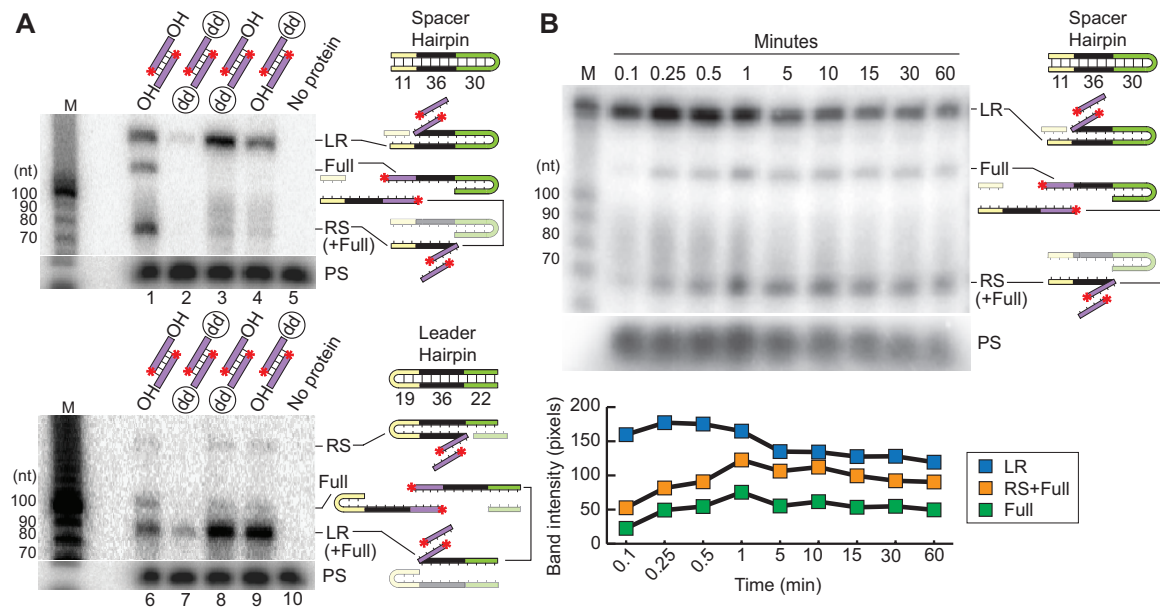
**Figure 2.3. Specific pre-spacer integration by Cas1-Cas2 into linear dsDNA targets.**

**(A)** Representation of the linear CRISPR target. Target consists of a leader sequence (yellow), repeat (black) and spacer (green). Leader-repeat (LR) and repeat-spacer junctions (RS) and points of integration (G1 and G36) are indicated. Palindromic inverted repeat sequences (IR 1 and IR 2) are marked. **(B)** Schematic of spacer integration *in vitro* with a minimal linear CRISPR target. Integration products include full-site integration, or half-site integrations at either the leader-repeat junction (LR) or repeat-spacer junction (RS). **(C)** *In vitro* spacer-integration assay with Cas1, Cas2, Csn2 and/or Cas9. Expected integration products corresponding to both junctions (LR or RS) are indicated. **(D)** High-throughput sequencing analysis of integration products represented as percent of total reads mapped throughout the linear CRISPR target.



**Figure 2.4. Cas1-Cas2 spacer integration reaction is directional.**

**(A)** Detection of full-site and half-site integration products with spacer hairpin CRISPR targets (top panel) and leader hairpin CRISPR targets (bottom panel) with unmodified (OH/OH) and modified (OH/dd, dd/OH, dd/dd) pre-spacers show site of the first transesterification reaction. **(B)** Time-course of spacer integration assay using unmodified pre-spacer and spacer hairpin CRISPR target. Quantification of **B** (bottom panel).





**Figure 2.5. Repeat sequence mutations affect efficiency and specificity during spacer integration.**

**(A)** Annotation of leader and repeat sequence mutations on the linear CRISPR target.

Leader-repeat junction (LR), Repeat-spacer junction (RS) and inverted repeats (IR1 and IR2) are indicated. **(B)** Integration reaction with mutated CRISPR targets taken at time

points: 15 sec, 1 min and 15 min. **(C)** High-throughput sequencing analysis of strand-specific integration products; peaks represent the percent of total reads mapped

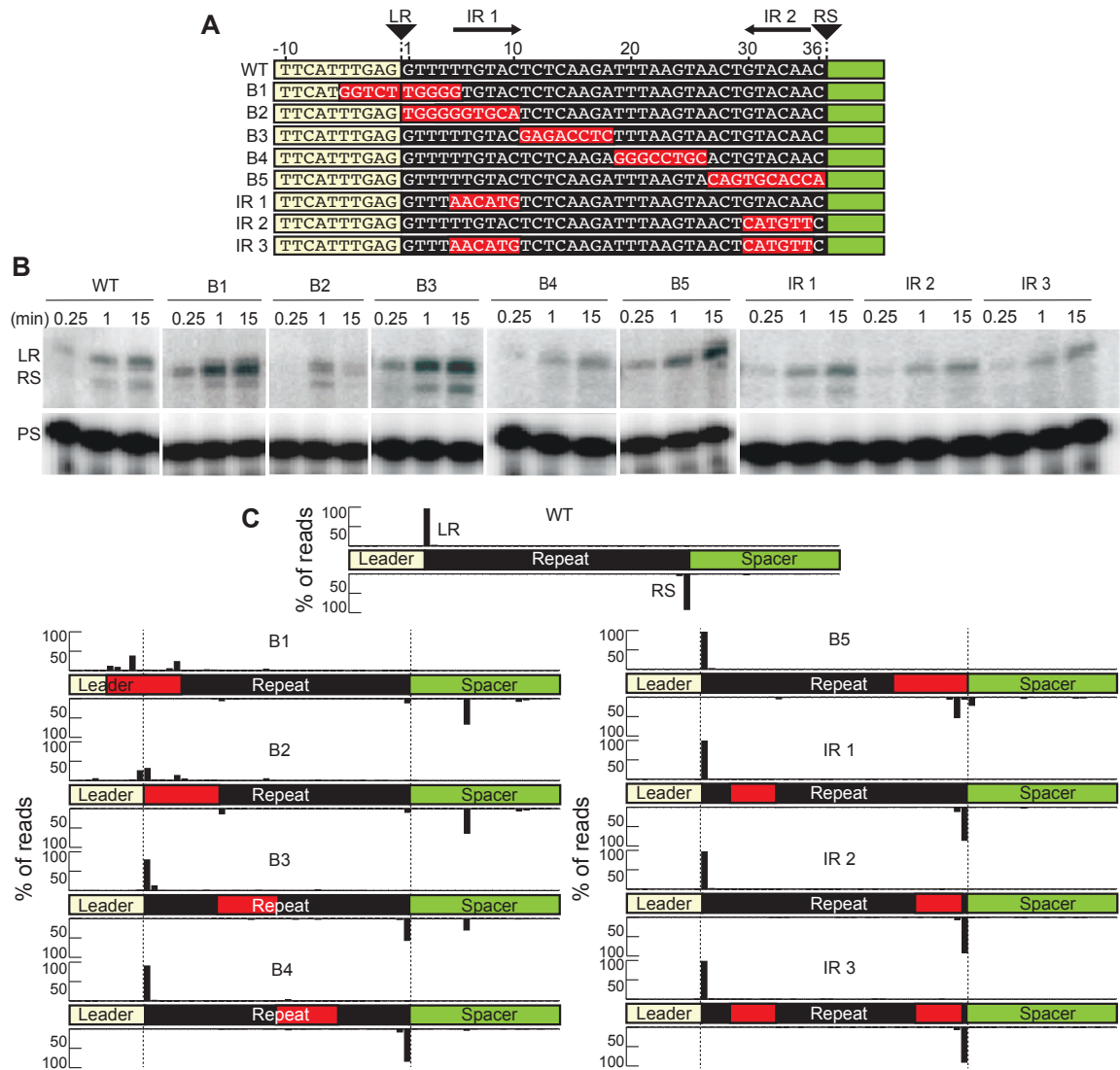
throughout the linear CRISPR target. Top strand and bottom strand libraries were

prepared separately and read counts are normalized across strands. Red boxes indicate

mutated sequences in the CRISPR target. Nucleotide level resolution of high-throughput

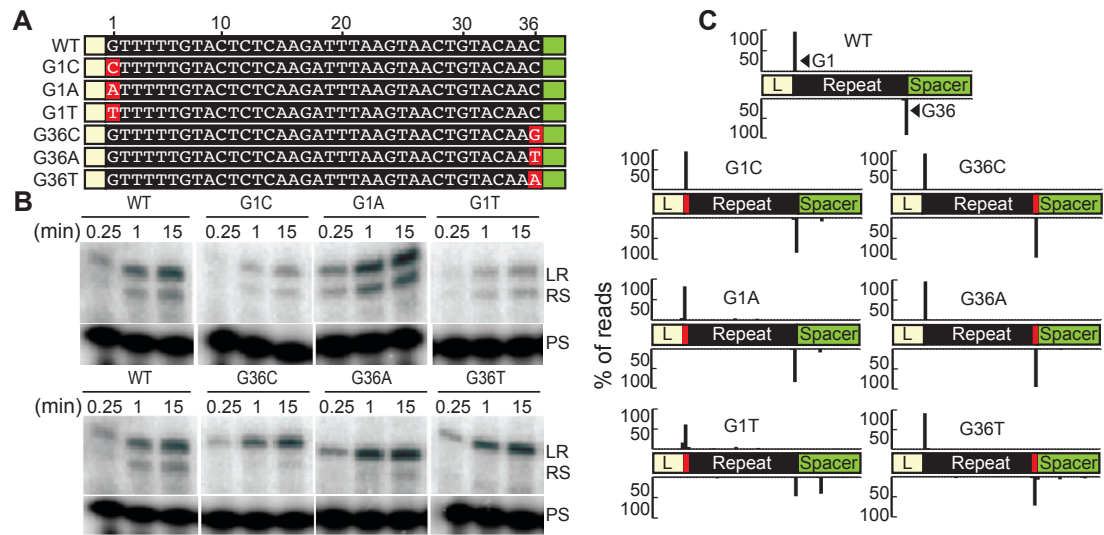
sequencing data is provided in Supplemental Figure 2.S4. Range of total number of reads

(4,711-12,359).



**Figure 2.6. Identity of the first and last nucleotide of the repeat affect efficiency and specificity during spacer integration.**

**(A)** Annotation of repeat sequence mutation of guanine at position 1 and position 36. **(B)** Integration reaction with mutated CRISPR targets taken at time points: 15 sec, 1 min and 15 min. **(C)** High-throughput sequencing analysis of integration products represented as percent of total reads mapped throughout the linear CRISPR target. Guanine at position 1 and 36 of the repeat are displayed as G1 and G36. Nucleotide level resolution of high-throughput sequencing data is provided in Supplemental Figure 2.S4. Range of total number of reads (6,114-16,093).



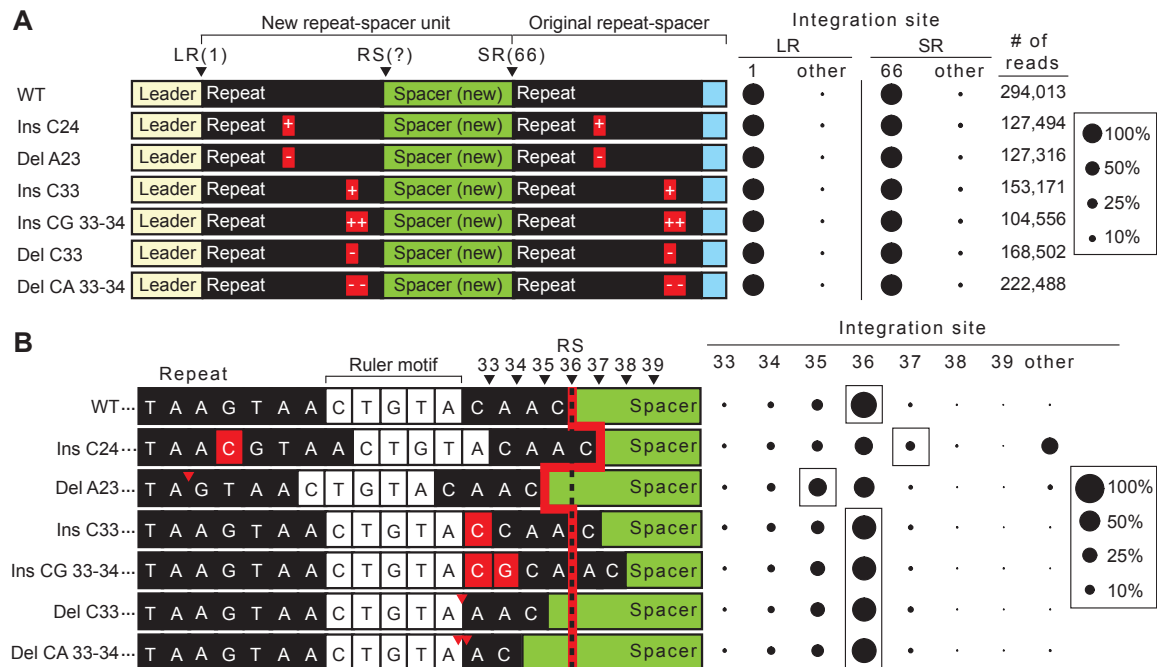
**Figure 2.7. Second-site integration at the repeat-spacer junction is defined by a molecular ruler.**

**(A)** (Left panel) Single, double and triple nucleotide insertion and deletion mutations made to the minimal CRISPR target. Mutations in red boxes are indicated. Predicted 5'-CTGTA-3' ruler element defining second-site transesterification attack 8 bp from the 5'-C is marked. Dotted line marks position 36 of the repeat. Grey arrow indicates the preferred site of integration for each CRISPR target. (Right panel) Sites of integration represented as percent of total mapped reads at the leader-repeat junction (LR) and positions spanning the repeat-spacer junction (RS) qualitatively represented (right panel). Nucleotide level resolution of high-throughput sequencing data is provided in Supplemental Figure 2.S4. **(B)** Integration sites for single, double and triple insertion and deletion mutations mapped to the minimal CRISPR target at nucleotide resolution. Position 36 is indicated with a dotted line. Range of total number of reads (14,884-293,723).



**Figure 2.8. Ruler-based mechanism influences second-site integration *in vivo*.**

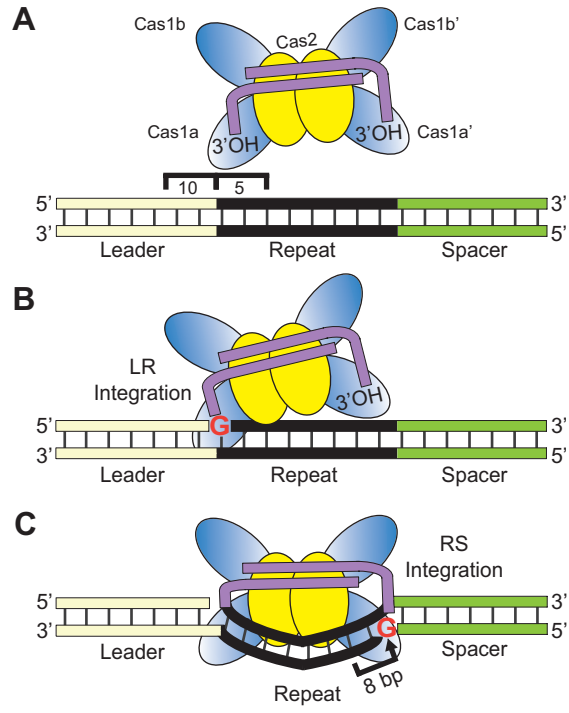
**(A)** (left panel) Annotated sequence of expanded CRISPR array *in vivo* with single and double insertion and deletion mutations in the repeat sequence. (Right panel) Sites of integration represented as percent total of mapped reads at the leader-repeat junction (LR) and spacer-repeat junction (SR). **(B)** (left panel) Sequences spanning the repeat-spacer junction and site of integration. Red line highlights expected site of integration relative to insertions and deletions upstream and downstream of the ruler element (5'-CTGTA-3') as indicated by red arrows or boxes. (Right panel) Integration site at the positions spanning the repeat-spacer junction.





**Figure 2.9. *S. thermophilus* type II-A spacer integration model.**

**(A)** The integrase complex (Cas1-Cas2 bound by a pre-spacer) recognizes 15 bp of the sequences spanning the leader-repeat junction to direct specific integration to the first repeat. **(B)** Identity of a guanine at position 1 of the repeat facilitates localization of integration to direct the first transesterification attack at the leader-repeat junction, resulting in a half-site intermediate. **(C)** DNA bending of the repeat sequence initiates the second-site transesterification attack at the repeat-spacer junction measuring 8 nucleotides upstream of a molecular ruler localized near the repeat-spacer junction (Figure 2.7 and 2.8). The integration complex additionally relies on a guanine at position 36 to progress to full-site integration of a new spacer into the CRISPR array.



**Table 2.S1. Plasmids used in this study**

Plasmid name	Plasmid description
pControl	pWAR derived from pWAR228 lacking CRISPR sequence
pCRISPR	pWAR derived from pWAR228 with minimal CRISPR array LRSRS
L1	pCRISPR leader mutation; nucleotides 21 - 32
L2	pCRISPR leader mutation; nucleotides 11 - 20
L3	pCRISPR leader mutation; nucleotides 1 - 10
pCas1/Cas2/Csn2/Cas9+CRISPR	pCas1/Cas2/Csn2/Cas9+minimal CRISPR array
Ins C24	pCas1/Cas2/Csn2/Cas9+CRISPR; insert C, +24
Del A23	pCas1/Cas2/Csn2/Cas9+CRISPR; delete A, +23
Ins C33	pCas1/Cas2/Csn2/Cas9+CRISPR; insert C, +33
Ins CG 33-34	pCas1/Cas2/Csn2/Cas9+CRISPR; insert CG, +33-34
Del C33	pCas1/Cas2/Csn2/Cas9+CRISPR; delete C, +33
Del CA 33-34	pCas1/Cas2/Csn2/Cas9+CRISPR; delete CA, +33-34

**Table 2.S2. Oligonucleotides used in this study**

Oligos	Sequence (5' – 3')
S4-OS5-F	TCGTTACTGGTGAACCAGTTTCAAT
S4-OS5-R	AACTGGTTCACCAGTAACGACTGAG
WT-F	TTCATTTGAGGTTTTGTACTCTCAAGATTTAAGTAACTGTACAACCTGTTTGACAGCAAATCAAGA
WT-R	TCTTGATTTGCTGTCAAACAGTTGTACAGTTACTTAAATCTTGAGAGTACAAAAACCTCAAATGAA
B1-F	TTCAT <b>GGTCTTGGGG</b> TGTACTCTCAAGATTTAAGTAACTGTACAACCTGTTTGACAGCAAATCAAGA
B1-R	TCTTGATTTGCTGTCAAACAGTTGTACAGTTACTTAAATCTTGAGAGTACA <b>CCCAAGACC</b> ATGAA
B2-F	TTCATTTGAG <b>TGGGGGTGCA</b> TCTCAAGATTTAAGTAACTGTACAACCTGTTTGACAGCAAATCAAGA
B2-R	TCTTGATTTGCTGTCAAACAGTTGTACAGTTACTTAAATCTTGAGAT <b>TGCACCCCACT</b> CAAATGAA
B3-F	TTCATTTGAGGTTTTGTAC <b>GAGACCTC</b> TTTAAAGTAACTGTACAACCTGTTTGACAGCAAATCAAGA
B3-R	TCTTGATTTGCTGTCAAACAGTTGTACAGTTACTTAA <b>AGGTCTC</b> GTACAAAAACCTCAAATGAA
B4-F	TTCATTTGAGGTTTTGTACTCTCAAG <b>AGGCCGTGCA</b> CTGTACAACCTGTTTGACAGCAAATCAAGA
B4-R	TCTTGATTTGCTGTCAAACAGTTGTACAG <b>GCAGGCCCT</b> CTTGAGAGTACAAAAACCTCAAATGAA
B5-F	TTCATTTGAGGTTTTGTACTCTCAAGATTTAAGT <b>ACAGTGCACCA</b> TGTTTGACAGCAAATCAAGA
B5-R	TCTTGATTTGCTGTCAAACA <b>TGGTGCAC</b> TGTACTTAAATCTTGAGAGTACAAAAACCTCAAATGAA
IR 1-F	TTCATTTGAGGTTT <b>AACATG</b> TCTCAAGATTTAAGTAACTGTACAACCTGTTTGACAGCAAATCAAGA
IR 1-R	TCTTGATTTGCTGTCAAACAGTTGTACAGTTACTTAAATCTTGAG <b>ACATGTT</b> AAACCTCAAATGAA
IR 2-F	TTCATTTGAGGTTTTGTACTCTCAAGATTTAAGTAACT <b>CATGTT</b> CTGTTTGACAGCAAATCAAGA
IR 2-R	TCTTGATTTGCTGTCAAACAG <b>AACATG</b> AGTTACTTAAATCTTGAGAGTACAAAAACCTCAAATGAA
IR 3-F	TTCATTTGAGGTTT <b>AACATG</b> TCTCAAGATTTAAGTAACT <b>CATGTT</b> CTGTTTGACAGCAAATCAAGA
IR 3-R	TCTTGATTTGCTGTCAAACAG <b>AACATG</b> AGTTACTTAAATCTTGAG <b>ACATGTT</b> AAACCTCAAATGAA
G1C-F	TTCATTTGAG <b>C</b> TTTTGTACTCTCAAGATTTAAGTAACTGTACAACCTGTTTGACAGCAAATCAAGA
G1C-R	TCTTGATTTGCTGTCAAACAGTTGTACAGTTACTTAAATCTTGAGAGTACAAAA <b>GCT</b> CAAATGAA
G1A-F	TTCATTTGAG <b>A</b> TTTTGTACTCTCAAGATTTAAGTAACTGTACAACCTGTTTGACAGCAAATCAAGA
G1A-R	TCTTGATTTGCTGTCAAACAGTTGTACAGTTACTTAAATCTTGAGAGTACAAAA <b>TCT</b> CAAATGAA
G1T-F	TTCATTTGAG <b>T</b> TTTTGTACTCTCAAGATTTAAGTAACTGTACAACCTGTTTGACAGCAAATCAAGA
G1T-R	TCTTGATTTGCTGTCAAACAGTTGTACAGTTACTTAAATCTTGAGAGTACAAAA <b>ACT</b> CAAATGAA
G36C-F	TTCATTTGAGGTTTTGTACTCTCAAGATTTAAGTAACTGTACAA <b>GT</b> GTTTGACAGCAAATCAAGA
G36C-R	TCTTGATTTGCTGTCAAACA <b>C</b> TGTGTACAGTTACTTAAATCTTGAGAGTACAAAAACCTCAAATGAA
G36A-F	TTCATTTGAGGTTTTGTACTCTCAAGATTTAAGTAACTGTACAA <b>T</b> GTTTGACAGCAAATCAAGA
G36A-R	TCTTGATTTGCTGTCAAACA <b>A</b> TGTGTACAGTTACTTAAATCTTGAGAGTACAAAAACCTCAAATGAA
G36T-F	TTCATTTGAGGTTTTGTACTCTCAAGATTTAAGTAACTGTACAA <b>AT</b> GTTTGACAGCAAATCAAGA
G36T-R	TCTTGATTTGCTGTCAAACA <b>T</b> TGTGTACAGTTACTTAAATCTTGAGAGTACAAAAACCTCAAATGAA
C5-F	TTCATTTGAGGTTT <b>C</b> TTGTACTCTCAAGATTTAAGTAACTGTACAACCTGTTTGACAGCAAATCAAGA
C5-R	TCTTGATTTGCTGTCAAACAGTTGTACAGTTACTTAAATCTTGAGAGTACAA <b>GA</b> AAACCTCAAATGAA
C9-F	TTCATTTGAGGTTTTGT <b>C</b> ACTCTCAAGATTTAAGTAACTGTACAACCTGTTTGACAGCAAATCAAGA
C9-R	TCTTGATTTGCTGTCAAACAGTTGTACAGTTACTTAAATCTTGAGAGT <b>G</b> ACAAAAACCTCAAATGAA
C14-F	TTCATTTGAGGTTTTGTACTCT <b>C</b> CAAGATTTAAGTAACTGTACAACCTGTTTGACAGCAAATCAAGA
C14-R	TCTTGATTTGCTGTCAAACAGTTGTACAGTTACTTAAATCTTG <b>G</b> AGAGTACAAAAACCTCAAATGAA
C19-F	TTCATTTGAGGTTTTGTACTCTCAAG <b>C</b> TTTAAAGTAACTGTACAACCTGTTTGACAGCAAATCAAGA
C19-R	TCTTGATTTGCTGTCAAACAGTTGTACAGTTACTTAA <b>AGT</b> CTTGAGAGTACAAAAACCTCAAATGAA

C24-F	TTCATTTGAGGTTTTTGTACTCTCAAGATTTAA <u>C</u> GTAAGTGTACAAGTGTGACAGCAAATCAAGA
C24-R	TCTTGATTTGCTGTCAAACAGTTGTACAGTTAC <u>G</u> TAAATCTTGAGAGTACAAAAACCTCAAATGAA
C28-F	TTCATTTGAGGTTTTTGTACTCTCAAGATTTAAGTAA <u>C</u> CTGTACAAGTGTGACAGCAAATCAAGA
C28-R	TCTTGATTTGCTGTCAAACAGTTGTACAG <u>G</u> TACTTAAATCTTGAGAGTACAAAAACCTCAAATGAA
C33-F	TTCATTTGAGGTTTTTGTACTCTCAAGATTTAAGTAACTGTAC <u>C</u> CAAGTGTGACAGCAAATCAAGA
C33-R	TCTTGATTTGCTGTCAAACAGTTG <u>G</u> TACAGTTACTTAAATCTTGAGAGTACAAAAACCTCAAATGAA
C36-F	TTCATTTGAGGTTTTTGTACTCTCAAGATTTAAGTAACTGTACAA <u>C</u> CTGTGACAGCAAATCAAGA
C36-R	TCTTGATTTGCTGTCAAACAG <u>G</u> TGTACAGTTACTTAAATCTTGAGAGTACAAAAACCTCAAATGAA
Ins36-F	TTCATTTGAGGTTTTTGTACTCTCAAGATTTAAGTAACTGTACAA <u>A</u> CTGTGACAGCAAATCAAGA
Ins36-R	TCTTGATTTGCTGTCAAACAG <u>T</u> TGTACAGTTACTTAAATCTTGAGAGTACAAAAACCTCAAATGAA
Ins36-37-F	TTCATTTGAGGTTTTTGTACTCTCAAGATTTAAGTAACTGTACAA <u>AT</u> CTGTGACAGCAAATCAAGA
Ins36-37-R	TCTTGATTTGCTGTCAAACAG <u>AT</u> TGTACAGTTACTTAAATCTTGAGAGTACAAAAACCTCAAATGAA
Ins36-38-F	TTCATTTGAGGTTTTTGTACTCTCAAGATTTAAGTAACTGTACAA <u>ATA</u> CTGTGACAGCAAATCAAGA
Ins36-38-R	TCTTGATTTGCTGTCAAACAG <u>TAT</u> TGTACAGTTACTTAAATCTTGAGAGTACAAAAACCTCAAATGAA
Del35-F	TTCATTTGAGGTTTTTGTACTCTCAAGATTTAAGTAACTGTACACTGTTTGACAGCAAATCAAGA
Del35-R	TCTTGATTTGCTGTCAAACAGTGTACAGTTACTTAAATCTTGAGAGTACAAAAACCTCAAATGAA
Del34-35-F	TTCATTTGAGGTTTTTGTACTCTCAAGATTTAAGTAACTGTACCTGTTTGACAGCAAATCAAGA
Del34-35-R	TCTTGATTTGCTGTCAAACAGGTACAGTTACTTAAATCTTGAGAGTACAAAAACCTCAAATGAA
Del33-35-F	TTCATTTGAGGTTTTTGTACTCTCAAGATTTAAGTAACTGTACTGTTTGACAGCAAATCAAGA
Del33-35-R	TCTTGATTTGCTGTCAAACAGTACAGTTACTTAAATCTTGAGAGTACAAAAACCTCAAATGAA
Spacer-	TTTCATTTGAGGTTTTTGTACTCTCAAGATTTAAGTAACTGTACAAGTGTGACAGCAAATCAAGATT
Hairpin	CGAATCGATAGATTCTGAATCTTGATTTGCTGTCAAACAGTTGTACAGTTACTTAAATCTTGAGAGTACA AAAACCTCAAATGAAA
Leader-	AATCTTGATTTGCTGTCAAACAGTTGTACAGTTACTTAAATCTTGAGAGTACAAAAACCTCAAATGAAA
Hairpin	TTTTGCGATAGCAAATTTTCATTTGAGGTTTTTGTACTCTCAAGATTTAAGTAACTGTACAAGTGTG ACAGCAAATCAAGATT

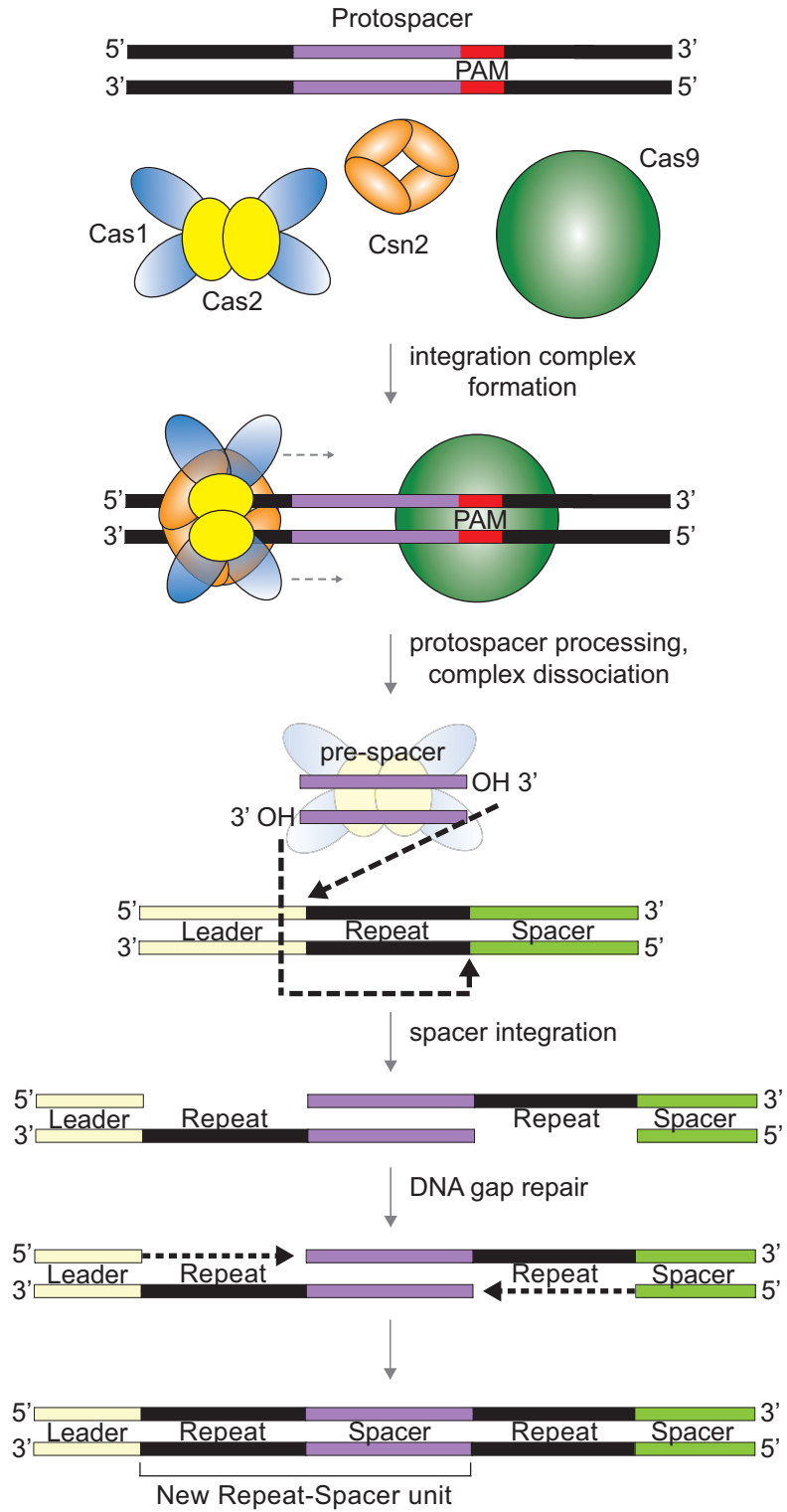
---

\*underlined annotates mutations

**Figure 2.S1. Adaptation model for *S. thermophilus*.**

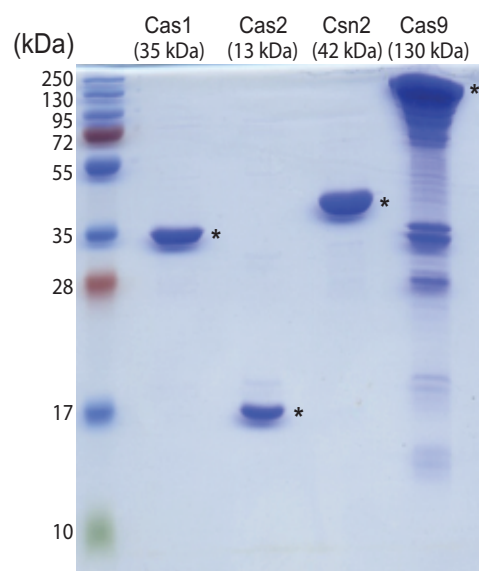
CRISPR adaptation involves the capture of foreign invader sequences (protospacers) and generation of pre-spacers in a PAM-dependent manner by the adaptation proteins Cas1, Cas2, Csn2, Cas9 and/or other host factors. Site-directed incorporation of the sequences (pre-spacers) into the host CRISPR locus involves a two-step concerted transesterification reaction by Cas1 and Cas2 in which the 3'-OH groups of the pre-spacer DNA carries out nucleophilic attacks at the 5' ends of the repeat borders. DNA repair events fill in DNA gaps and ensure faithful duplication of a CRISPR repeat.





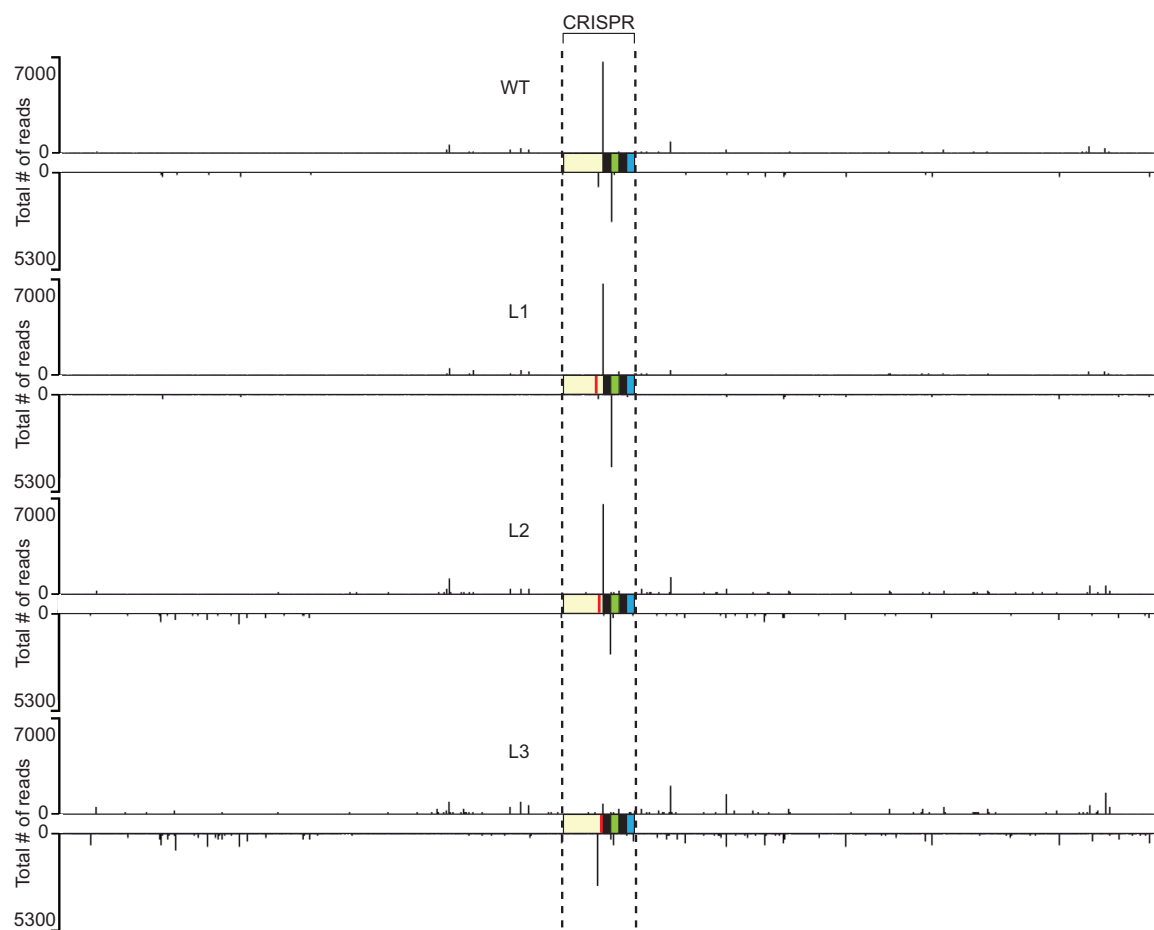
**Figure 2.S2. Purified *S. thermophilus* Cas proteins.**

Purification of individual Cas proteins by Ni<sup>2+</sup> affinity column chromatography. Products were separated by an SDS-PAGE gel followed by Coomassie blue staining. Bands corresponding to each protein are indicated by a black asterisk while molecular weights of the proteins are listed above.



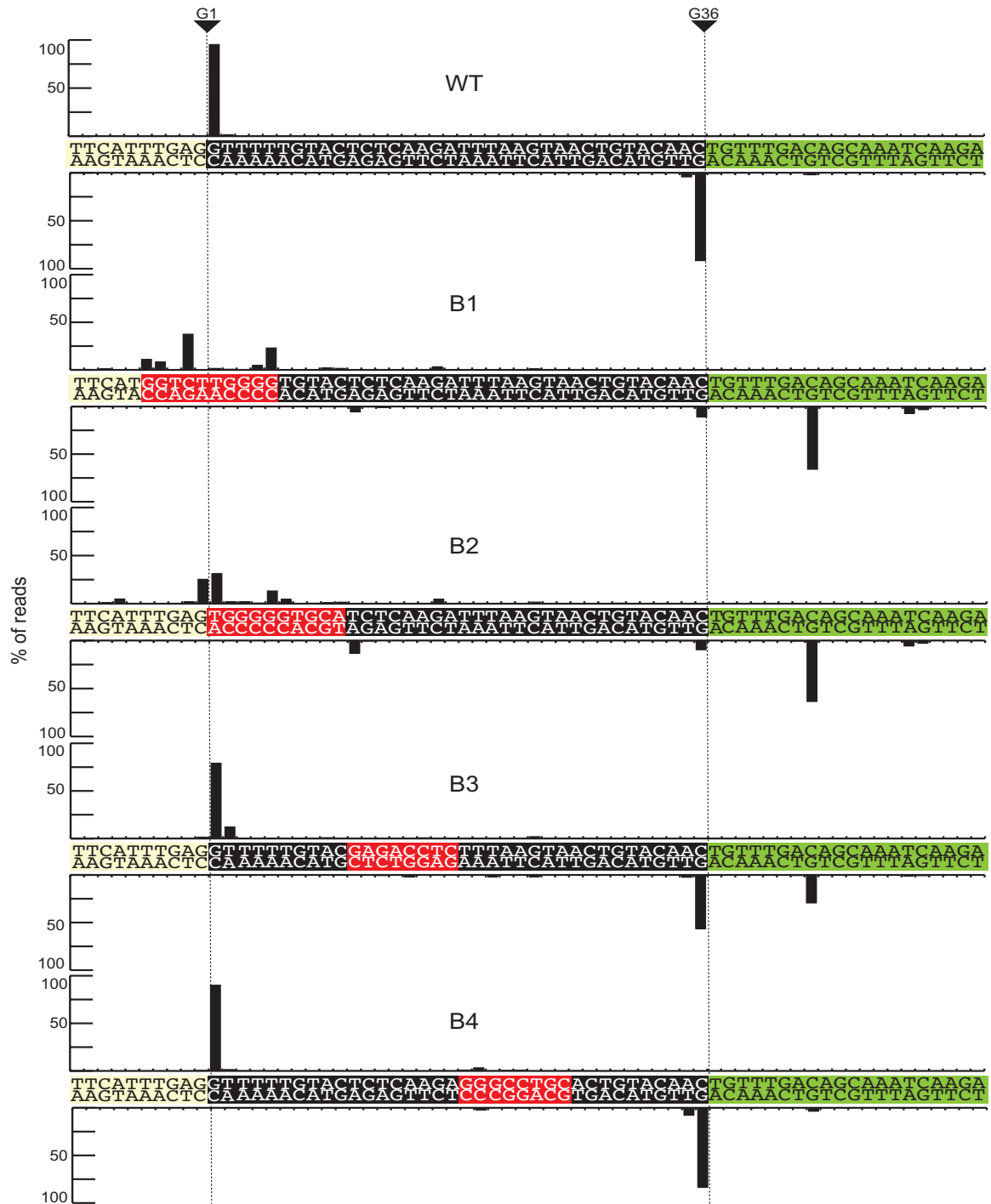
**Figure 2.S3. Full plasmid mapping of integration sites on mutated pCRISPR.**

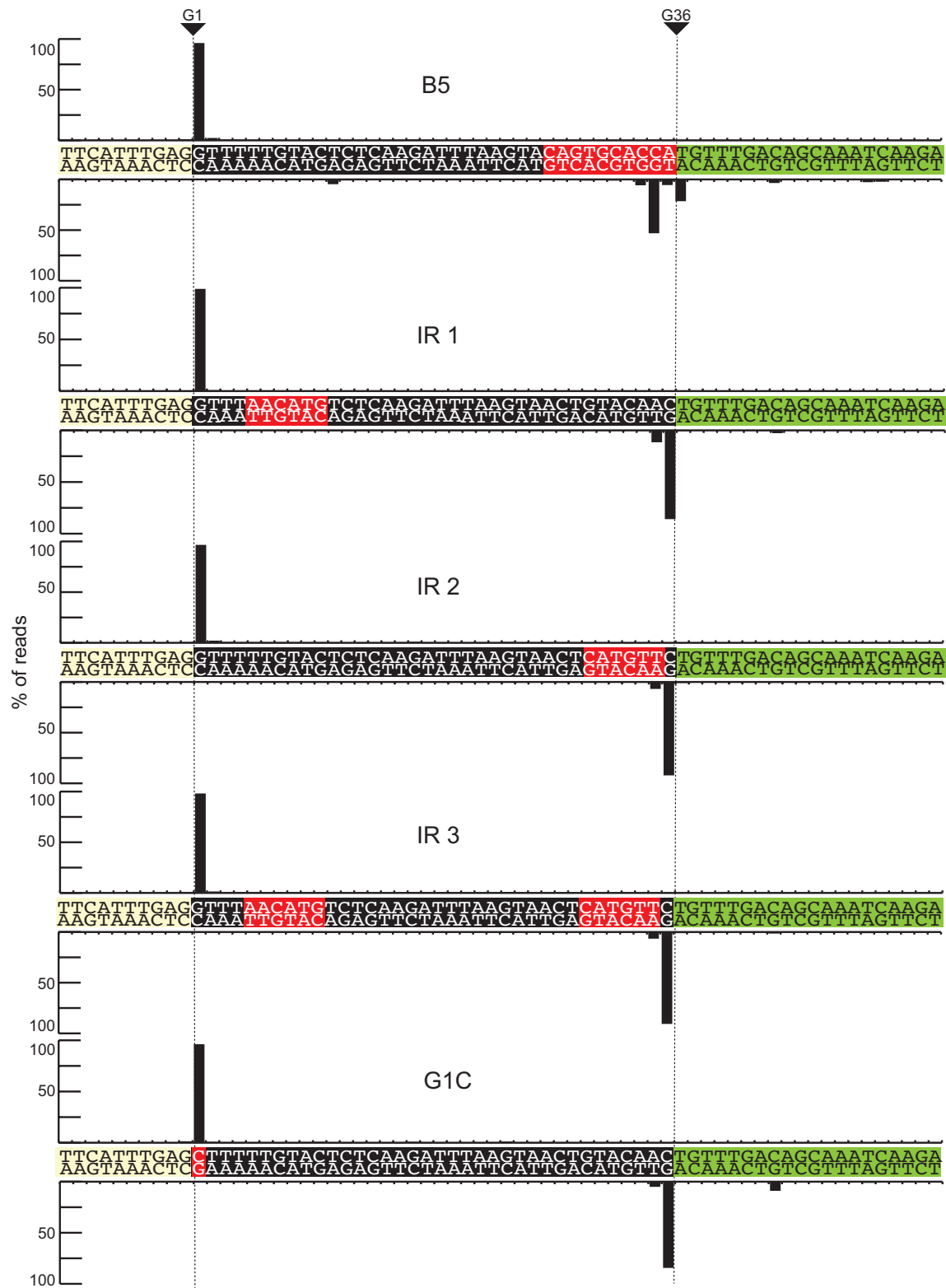
High-throughput sequence mapping of integrated pre-spacers in mutated pCRISPR plasmids with mutated leader sequences (L1, L2, L3).



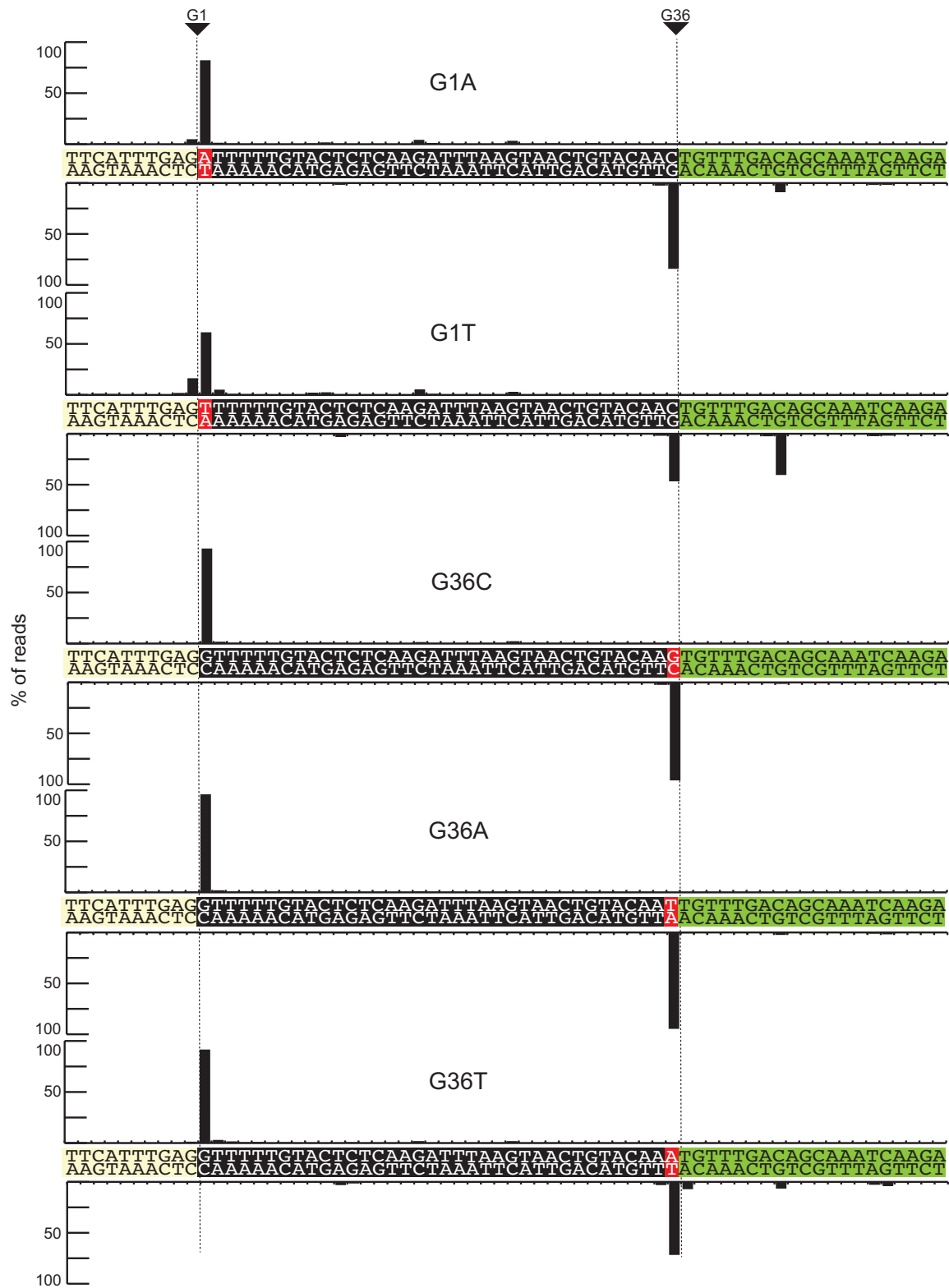
**Figure 2.S4. Nucleotide resolution and detailed mapping of integration sites.**

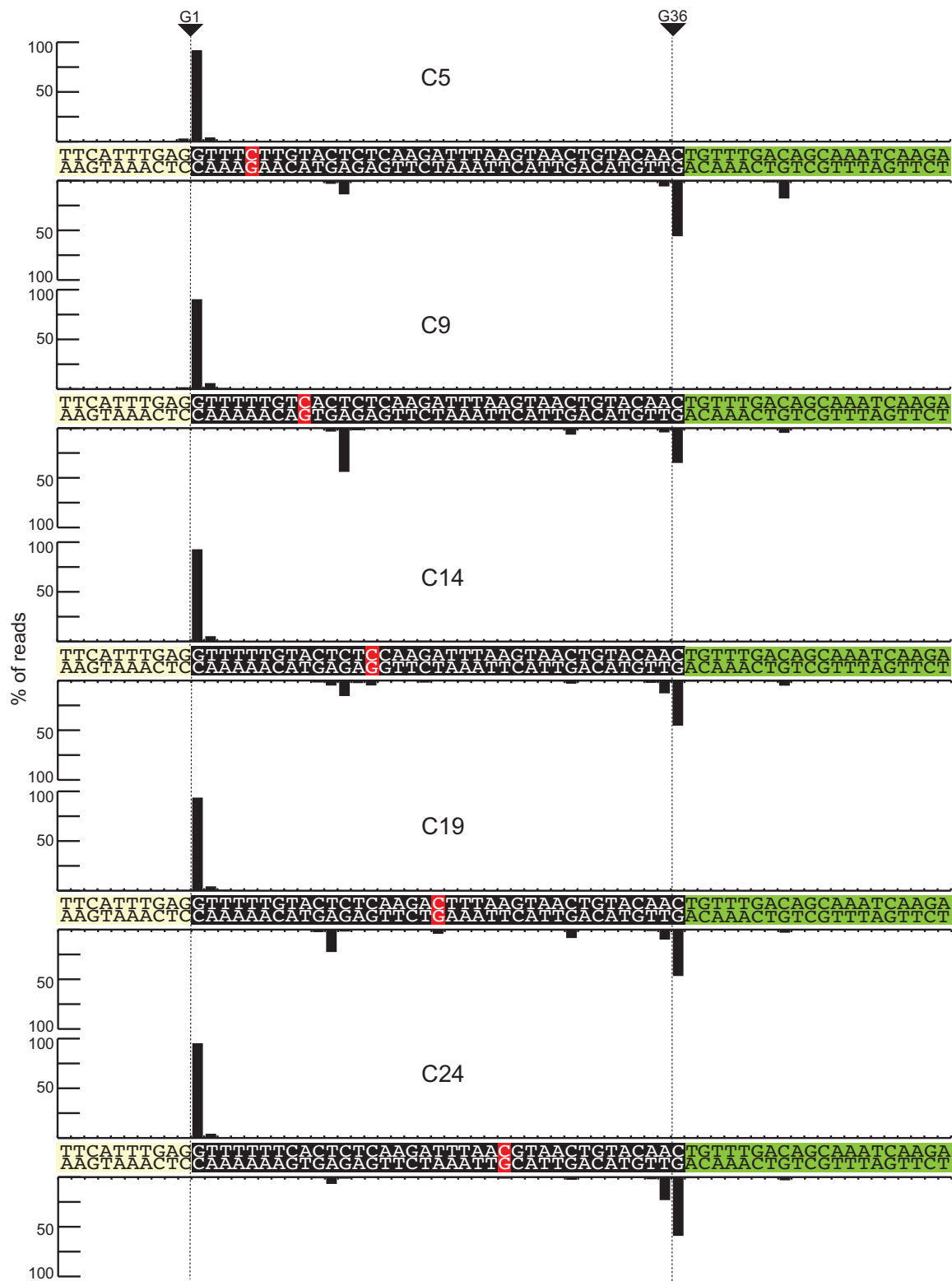
Histogram plots of high-throughput sequencing analysis of integrated pre-spacers in mutated linear CRISPR targets.

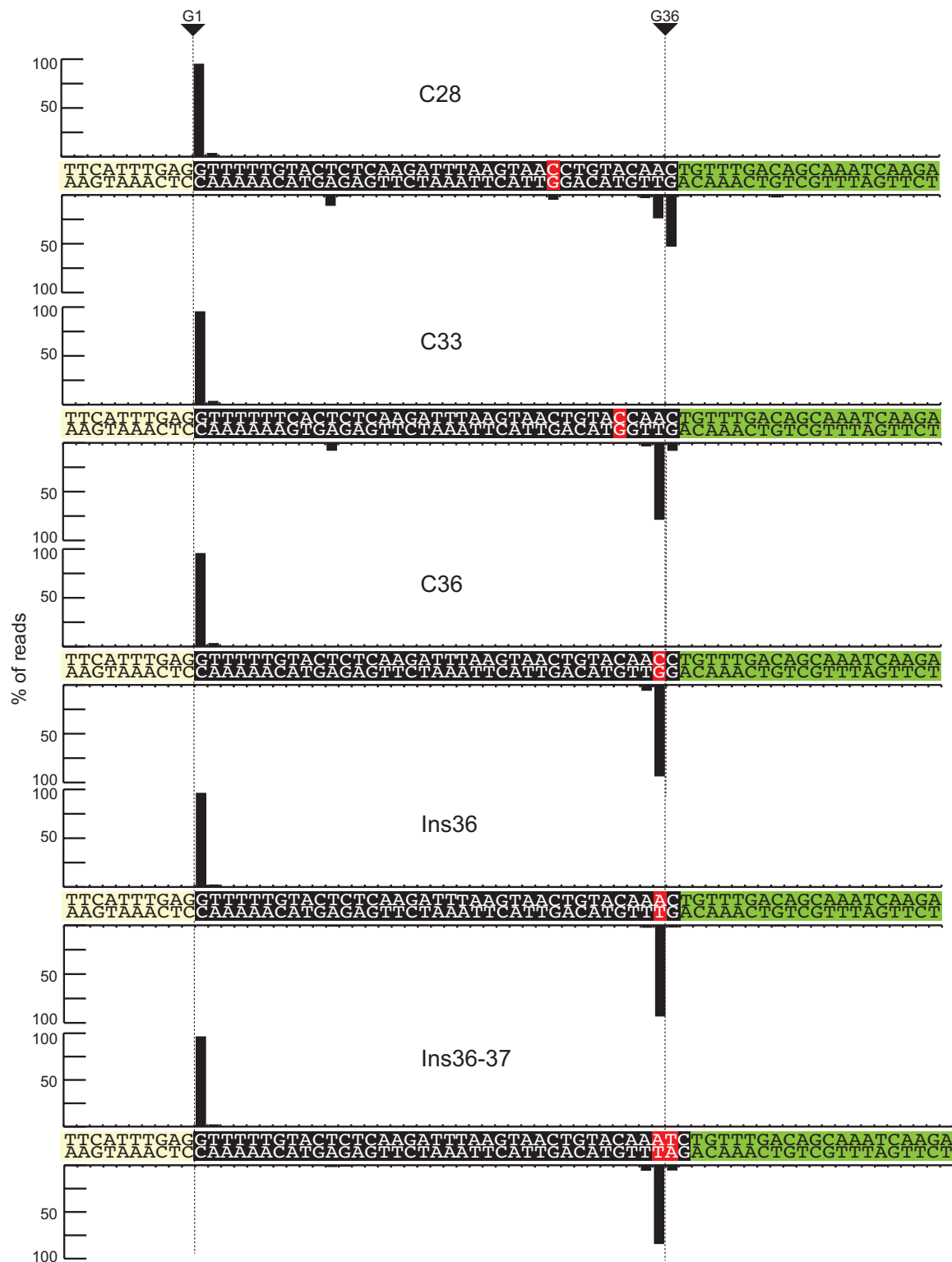


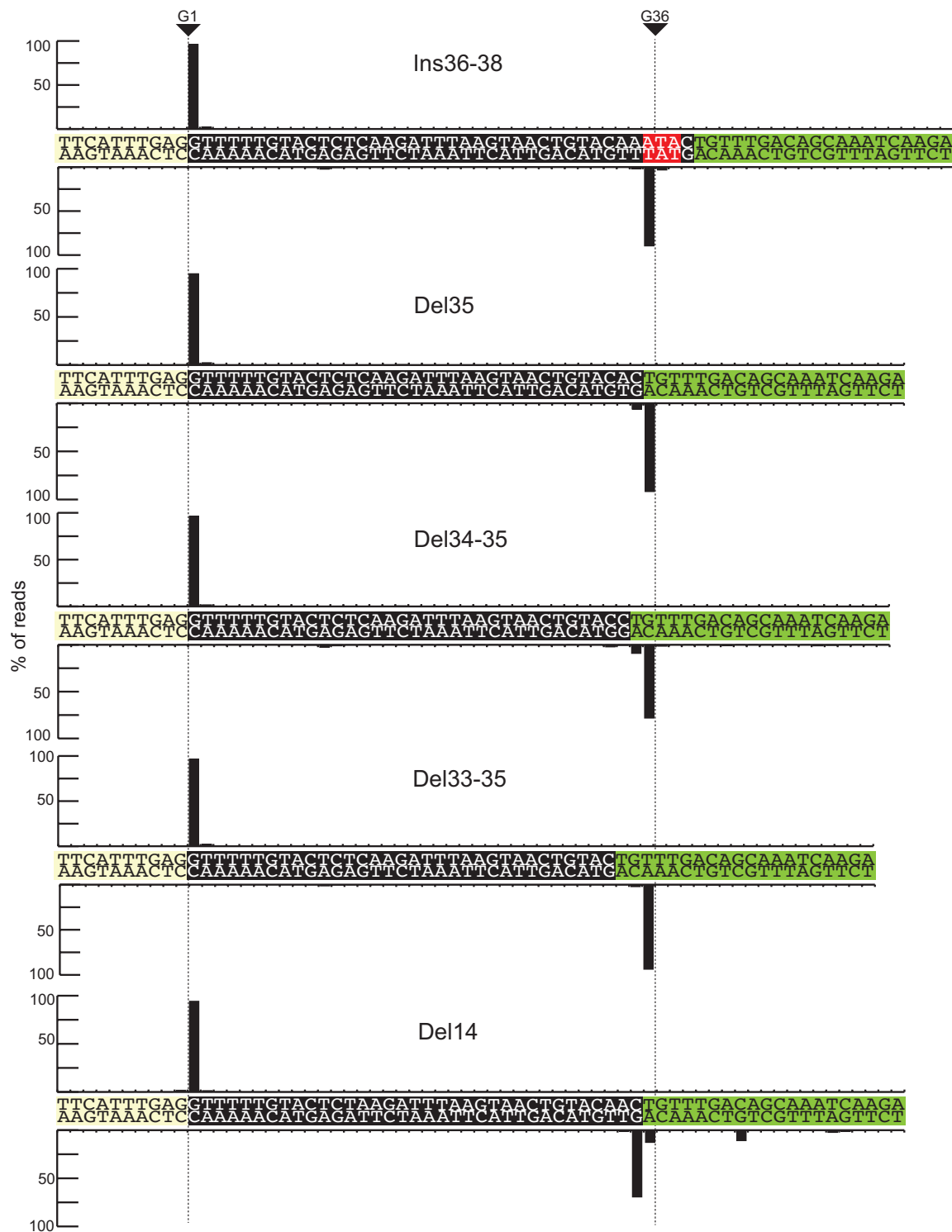








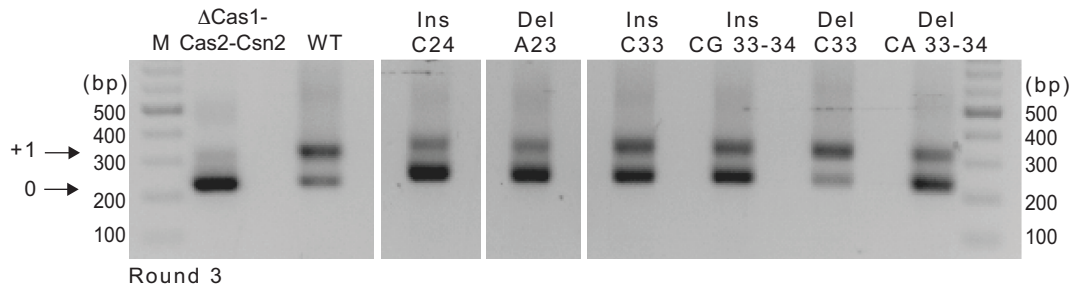
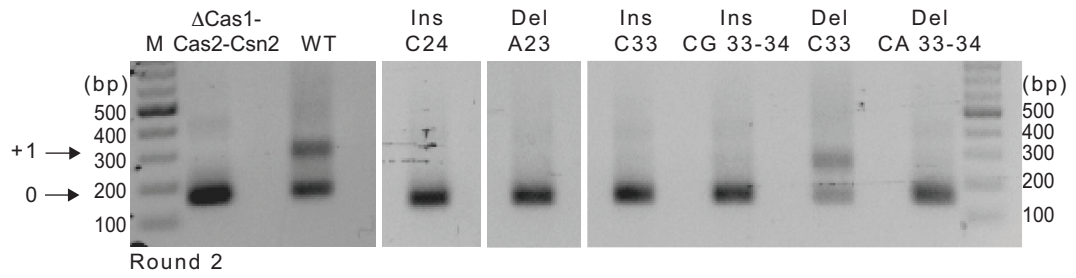
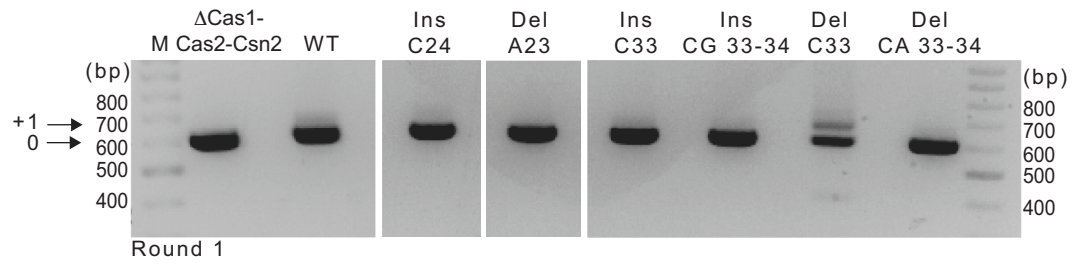




**Figure 2.S5. Repeat sequence insertions and deletions reduce adaptation *in vivo*.**

(Top panel) PCR amplification of the CRISPR array with either wildtype or indicated repeat sequence mutations. Regions in the agarose gel containing expanded CRISPR arrays (+1) were gel excised away from unexpanded CRISPR arrays (0). Extracted DNA was used as template for re-amplification of the array (middle and bottom panels).

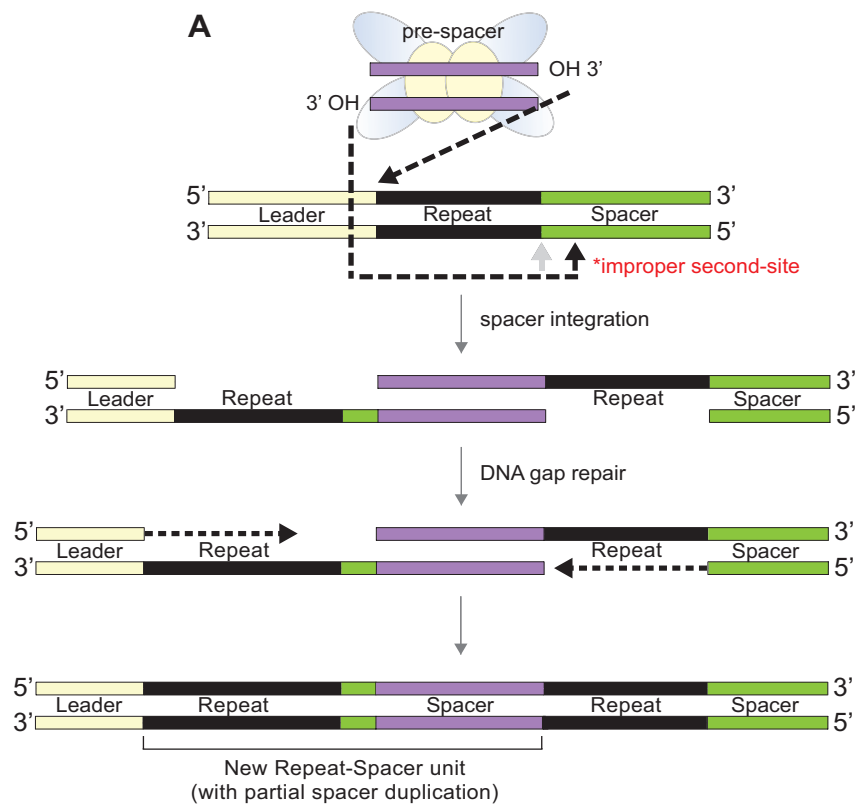
Adaptation null strain ( $\Delta$ Cas1-Cas2-Csn2) containing the pCRISPR plasmid served as a negative control.



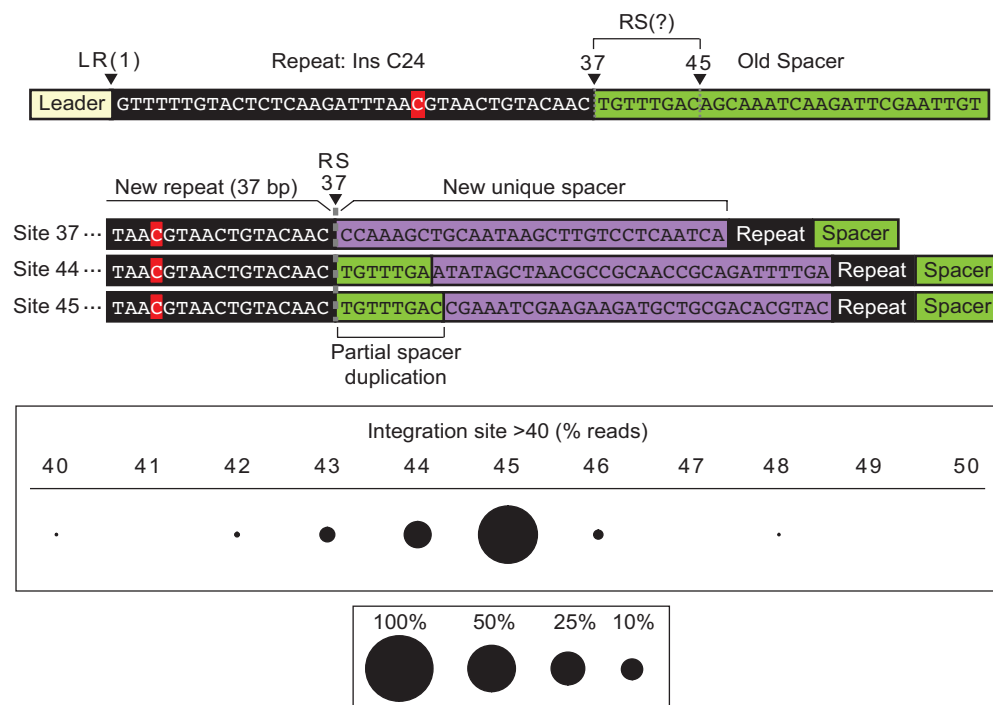
**Figure 2.S6. Incorrect second-site integration of the repeat results in partial spacer duplication *in vivo*.**

**(A)** Schematic of improper second-site recognition during pre-spacer integration.

Misrecognition of the repeat-spacer junction and integration at sites downstream in the adjacent spacer results in partial spacer sequence duplication upstream of the new spacer sequence after DNA gap repair. **(B)** Analysis of “other” sites of integration in Ins C24 repeat mutation strain from Figure 2.8. Off-target second-site integration events occur predominately at site 44 and 45 within the spacer sequence resulting in partial spacer duplication upstream of the unique spacer of the expanded array.



**B**





CHAPTER 3

EFFECTS OF CRISPR-ASSOCIATED PROTEINS AND REXAB ON  
PROTOSPACER GENERATION AND PHAGE RESPONSE IN *S. THERMOPHILUS*<sup>1</sup>

---

<sup>1</sup>Jenny G. Kim, Sandra Garrett, Yunzhou Wei, Brenton G. Graveley, Michael P. Terns.  
To be submitted to Nucleic Acids Research.

## Abstract

CRISPR-Cas immunity is acquired through the integration of foreign genetic elements, termed protospacers, into the host CRISPR loci as new spacers. Selection of protospacers requires the presence of a protospacer adjacent motif (PAM) next to the sequence. Mechanisms detailing how CRISPR-associated proteins and host DNA-repair proteins RexAB are involved in PAM-dependent processing of protospacer into pre-spacers remain a mystery. Here, we investigate the role of the Type II-A CRISPR-Cas proteins (Cas1, Cas2, Csn2, Cas9) as well as the trans-activating CRISPR RNA (tracrRNA) during protospacer selection. Additionally, we investigate the function of DNA repair proteins RexAB in influencing the acquisition of new spacer sequences from foreign invaders. Expression of Cas1-Cas2 in the absence Csn2 reduced the number of spacer sequences with the proper PAM suggesting a role of Csn2 in selecting PAM-containing sequences. We also determined that loss of nuclease activity of RexAB reduces adaptation *in vivo*. Moreover, we demonstrate that truncation of the tracrRNA to prevent its pairing with crRNA and possibly repeat DNA reduced spacer duplication events of pre-existing spacers in the CRISPR array. Lastly, we challenged *S. thermophilus* cells lacking a functional CRISPR-Cas systems to determine if cells can resist lytic phage infections through a mechanism other than CRISPR-Cas. We determined that rare non-CRISPR *S. thermophilus* phage resistant survivors had various mutations that each are predicted to inactivate a membrane protease FtsH, implicating this protein in phage life cycle. This effect was bypassed by mutations in a phage tail chaperonin protein that permitted phage sensitivity in the phage resistant survivors. Taken together, our findings not only highlight the intrinsic ability of Cas and non-Cas

proteins to coordinate protospacer selection with proper discrimination to ensure effective immunity, but also provide insight into the understanding of the host-phage evolutionary arms race.

## **Introduction**

CRISPR-Cas (Clustered regularly interspaced short palindromic repeats and CRISPR-associated genes) systems are defense mechanisms that provide heritable immunity against invading foreign elements such as viruses and plasmids (1,2). There are six distinct CRISPR-Cas types (I-VI) and more than thirty subtypes, and these adaptive immune systems are found in the genomes of roughly half of bacteria and almost all archaea sequenced to date (3,4). CRISPR systems function in providing protection against foreign invading elements by integrating fragments of DNA and incorporating them into the CRISPR array of the host's own genome. This generates a memory bank of prior infections to elicit an effective immune response upon reinfection of the cell. CRISPR-Cas immunity involves three main stages: adaptation, CRISPR RNA (crRNA) biogenesis, and interference. During the adaptation stage, pieces of foreign DNA termed "protospacers" are recognized, captured and further processed into "pre-spacers" for incorporation into the CRISPR array as "spacers" (2,5). The crRNA biogenesis stage involves the transcription of the CRISPR array and processing of the transcript into mature crRNAs, which form complexes with the Cas effector proteins (6). Interference or invader silencing occurs upon binding of the effector complex to the target complementary DNA or RNA during reinfection and cleaving the foreign nucleic acid (7-9).

We are only now beginning to unfold the mechanistic details involved in foreign DNA capture and integration for CRISPR-mediated defense. The initial steps of adaptation in CRISPR-Cas immunity can be broken down into two main stages. The first stage, protospacer generation, involves the capture of foreign nucleic acids as protospacers and further processing of these substrates into pre-spacers. The second stage, spacer integration, is the integration of fully-processed pre-spacer sequences into the CRISPR array as spacers. The molecular details of how spacer sequences are integrated into the CRISPR array have now been described for several different CRISPR-Cas systems. In contrast, fewer clear models for the process of protospacer generation and acquisition by CRISPR-Cas exist, although some mechanisms have been uncovered. For example, in the widely studied Type I-E CRISPR system in *E. coli*, adaptation is stimulated by host DNA repair proteins, RecBCD, presumably because these proteins generate substrates for CRISPR uptake (10,11). RecBCD is a nuclease-helicase complex and is biologically critical in end-processing of blunt-ended double stranded breaks to stimulate homologous recombination for DNA repair (12-21). A critical element during this process is the regulatory chi sequence (chromosomal hotspot instigator) a short octameric sequence that controls and modulates translocation of the enzyme across the DNA (22-29). Upon Chi recognition by a Chi-recognition domain in RecC, exonuclease activity of RecBCD is altered and nuclease polarity is switched; resulting in attenuation of strand degradation (16). Since chi sequences occur nearly 14 times more frequently in the *E. coli* genome than phage DNA, RecBCD tends to produce protospacers that are phage-derived and thus useful for defense (10,16).

To dissect the role of RecBCD on CRISPR adaptation, a previous study demonstrated that deletion of RecBCD proteins in *E. coli* resulted in a shift in adaptation biases for acquired spacers and acquisition hotspots were confined by chi sites (10). Similar homologs of RecBCD in Gram-negative bacteria such as AddAB, have also been linked to adaptation frequency and spacer acquisition patterns around chi sites. This effect however, was simplified to the nuclease activity of the primary protein AddA (30). While nuclease activity of repair proteins have been linked to CRISPR adaptation, weak evidence suggests that helicase activity of RecBCD, as well as several other host proteins, can contribute to adaptation, so utilization of non-Cas proteins and their natural function for host survival mediated by CRISPR is not uncommon in organisms (11,31). In DNA repair exo/hel (exonuclease/helicase) enzymes, sequence differences amongst bacterial species suggest functional variability however, conservation of active-site domains likely suggest overlapping roles. Given that RecBCD nuclease activity contributes to CRISPR spacer acquisition, will its homologs behave in similar ways? The RecBCD analog, RexAB, is commonly found in several Gram-positive species such as *Staphylococcal*, *lactococcal* and *streptococcal* species including *Streptococcus thermophilus* (32-35). *In vivo* studies suggest that RexAB is involved in DNA repair and has been identified to behave similarly to the extensively studied, two-subunit enzyme AddAB present in other Gram-positive bacteria (32-35). To date, the relationship between RexAB and CRISPR adaptation has not been explored in any organism.

During the capture of foreign nucleic acids as protospacers, proper processing into pre-spacers is critical for the integration of functional spacers into the CRISPR array. A key characteristic of spacer acquisition is the selection of pre-spacers with a flanking

PAM (protospacer-adjacent motif) sequence (36). By targeting protospacers with a PAM, the system can prevent self-recognition or autoimmunity against spacers in the CRISPR locus (whose repeats are devoid of PAMs) by the crRNA-effector complex during interference. In Type I-E systems, the Cas1-Cas2 complex composed of two Cas1 dimers and one Cas2 dimer, mediates PAM recognition during protospacer capture via Cas1 (37,38). The pre-spacer is further processed at the respective ends to generate a substrate with 3'-OH groups for spacer integration by Cas1-Cas2 (37,39). Other Type I systems such as the Type I-A, I-C, I-D and I-U require Cas4 to recognize PAM-containing pre-spacers (40-42). However, Cas4 is not present in all CRISPR systems, suggesting that this role in the identification of PAM sequences is handled by other CRISPR or potentially non-CRISPR proteins.

In Type II-A systems, spacer acquisition *in vivo* requires all four CRISPR-Cas proteins, which are Cas1, Cas2, Csn2 and Cas9 (Cas9:tracrRNA:crRNA) plus a non-coding RNA known as the trans-activating CRISPR RNA (tracrRNA), that base-pairs with each crRNA to facilitate crRNA maturation and is an integral component of the Cas9-crRNA complex that cleaves target DNA (43,44). Currently, the mechanistic details of how Cas1, Cas2, Csn2, Cas9 proteins and the tracrRNA function together to select PAM-containing pre-spacers and integrate them into the CRISPR array remains to be elucidated. However, previous studies have provided evidence of functional roles for each protein. Cas1 and Cas2 are universally conserved in almost all CRISPR systems and are directly involved in integration of a new spacer into the CRISPR array (37,45-50). Although deletion of Csn2 has been shown to significantly reduce adaptation frequency (43,44), the mechanism for this is yet unclear. Csn2 is a tetramer that forms a toroidal

structure that has been shown to bind to double-stranded DNA ends (45,51-54).

Structural analysis of the Type II-A Cas proteins suggests that Csn2 may direct Cas1-Cas2 to pre-spacer substrates bound by Cas9 in a PAM-dependent manner (54). Unlike in Type I systems, PAM recognition is carried out by the interference protein Cas9 through the same PAM-interacting domain that is required for PAM-specific DNA targeting (43). However, the double-stranded DNase activity of Cas9 required for interference is not required for adaptation. Specifically, mutations in the Cas9 RuvC and HNH nuclease active centers prevent Cas9 from cutting DNA and providing protection against mobile genetic elements, but these mutations do not impair the function of Cas9 in adaptation (43,44). Drawing these observations together into a cohesive model of spacer uptake will require additional information.

The finding that adaptation in the Type II-A systems requires all four Cas proteins was determined by observing changes in adaptation efficiency with gene deletions and mutations. Due to the low frequency of adaptation, it is often difficult to detect spacer uptake in a population of cells, so both studies utilized the overexpression of the Cas proteins to increase adaptation and permit detection of spacer uptake after a single-round PCR amplification of the CRISPR array and visualization by gel electrophoresis. These early studies therefore did not address how Cas deletion affects adaptation under endogenous expression levels. In this study, we used a sensitive high-throughput sequencing approach (CAPTURE) to analyze the effects of gene deletions on adaptation in a background of endogenous expression levels.

Previous work has shown that in *S. thermophilus*, Type II adaptation is the dominant mechanism leading to cell survival when challenged by lytic phage infection

(1,55,56). Other methods of defense such as restriction modification and abortive infection are also observed in other bacteria and these different mechanisms of phage-resistance may interfere with several stages of the phage life cycle such as phage DNA replication and release. For example, evidence shows that *L. lactis* can sense bacteriophages and mount a response targeting the host cell envelope stress response pathway (57). This response was additionally demonstrated to contribute to the regulation of transcriptional factors of the lytic phage  $\lambda$  in *E. coli*. In this study, we searched for additional mechanisms of phage resistance (besides Type II CRISPR-based resistance) by creating *S. thermophilus* strains wherein the Type II CRISPR systems was genetically deleted and then infecting them with lytic phage. Using this host-phage system, characterization of non-CRISPR mediated phage response offered the possibility of identifying new factors and mechanisms that reduce phage infections.

In this study, we build on what is currently known about the functional role of both non-CRISPR associated DNA repair proteins and CRISPR-Cas proteins during adaptation and phage resistance. Expression of Cas1-Cas2 in the absence of Csn2 resulted in increased spacer integration of sequences without the proper PAM suggesting that Csn2 selects PAM-adjacent substrates upstream of integration by Cas1-Cas2. We also observe spacer duplication of pre-existing spacers within the CRISPR array. These duplication events were observed to occur at increased frequency when adaptation efficiency was low. Moreover, deletion of the trans-activating RNA, *tracrRNA*, reduced spacer duplication frequency and resulted in hotspot patterns of spacer acquisition at the ribosomal RNA gene loci providing evidence of increased self-targeting spacer uptake by Cas9 in its apo-form. Addressing the upstream processes of adaptation, our results show



that RexAB nuclease activity contributes to spacer acquisition likely through the generation of protospacers. Furthermore, we gained insight into the evolutionary arms race of host versus phage in the absence of CRISPR systems. Together, our work contributes to the current efforts to understand the molecular mechanisms governing CRISPR adaptation as well as further elucidate the diverse mechanisms involved in bacterial survival against invading foreign elements.

## **Materials and methods**

### ***S. thermophilus* strains and growth conditions**

*S. thermophilus* DGCC7710 was kindly provided by Dr. Sylvain Moineau. The *cas2*, *csn2*, *cas9* and *tracr* deletion strains were constructed using the pINTRS plasmid (58). The *csn2/cas9* deletion strain and *rexA* and *rexB* mutation strains were constructed using ComS-dependent natural transformation (59). Briefly, 800 bp upstream and downstream of the site of mutation was PCR-amplified from the *S. thermophilus* genome and ligated via splicing overlap extension PCR to generate a linear dsDNA fragment. The final PCR product was gel purified and transformed with 10 mM ComS peptide into natural competent cells as previously described (59,60). All strain mutations were verified by DNA sequencing and sequences of oligonucleotides are listed in Supplementary Table 3.S2. Overexpression plasmids expressing *S. thermophilus* CRISPR1 proteins were previously described (47). Plasmid constructions were verified by sequencing and transformed into *S. thermophilus* DGCC7710 strains via electroporation or ComS-dependent natural transformation (61). Plasmids used in this study are listed in Supplementary table 3.S1. *S. thermophilus* strains were grown in

LM17 (HiMedia) at 37°C. *S. thermophilus* harboring plasmids were grown in LM17 liquid medium supplemented with 2 µg/mL chloramphenicol for 16 hours.

### ***Population-based adaptation assay***

*S. thermophilus* cells harboring the indicated plasmids were grown in LM17 supplemented with 2 µg/mL chloramphenicol for 16 hours. Cells from each strain were harvested, pelleted and genomic DNA was extracted using the Zymo Research Quick-DNA Fungal/Bacterial Miniprep Kit (Zymo Research, Irvine CA) and used as PCR template. The CRISPR array was amplified using primers matching the leader sequence and first spacer of CRISPR1. Oligonucleotides used in this study are listed in Supplementary table 3.S2. PCR products were run on 2.5% TAE-agarose gels, pre-stained with ethidium bromide and examined under UV light to visualize CRISPR array expansion.

### ***Spacer Acquisition high-throughput sequencing***

CRISPR arrays were amplified by using a pair of primers in which the forward primer annealed within the leader region of the CRISPR array and the reverse primer annealed within the existing spacer closest to the leader (spacer 1). If a new spacer was integrated into the CRISPR array, the resulting PCR product was longer because of the additional repeat and spacer sequence, and the CRISPR array was considered to be expanded. These larger, expanded PCR products were separated from unexpanded products by 2.5% agarose gel electrophoresis using TAE buffer followed by DNA recovery (Zymogen DNA Gel Recovery Kit, Zymo Research). A second PCR reaction

utilizing CAPTURE repeat primers was used to amplify expanded arrays only. Single amplified products were gel extracted under similar conditions and used as a template for the final round of PCR. The repeat PCR primers included an overhang corresponding to part of the adapter necessary for Illumina sequencing. After size selection of the PCR product containing adapter sequences, the final PCR reaction was performed to add additional sequences corresponding to Illumina adapters and barcodes. Each experimental condition and replicate received a unique barcode (index) for multiplexing. The sequences of oligonucleotides used are available in Supplementary Table 3.S2. For each strain, at least four biological replicates were prepared, and from each of these replicates, an amplicon library was prepared from the CRISPR1 array. Final gel-purified amplicon libraries were pooled and generated for Illumina sequencing (59,62,63).

### ***Phage infection assay and survivor genotyping***

*S. thermophilus* DGCC7710 strains and derivatives were incubated at 42°C in LM17 until they reached an OD<sub>600</sub> of 0.3. Lytic 2972 phages were amplified and propagated in LM17 media supplemented with 10 mM CaCl<sub>2</sub> as described previously (64). Purified 2972 phage was added to 200 µL cells at a MOI of 2 in 3 mL of top agar (0.75% LM17 agar) supplemented with 10 mM CaCl<sub>2</sub> and poured over 1.5% LM17+10 mM CaCl<sub>2</sub> plates and incubated at 42°C. Surviving colonies were isolated on 1.5% LM17 plates incubated at 42°C. Individual liquid cultures of each colony was made by inoculating 1 mL of LM17 with cells and grown overnight at 42°C. CRISPR spacer uptake into each of the four CRISPR arrays was assayed (using 1 µL of the overnight liquid culture) by PCR using primers matching to the leader sequence and first spacer of

each CRISPR array. Oligonucleotides used for the PCR tests are listed in Supplementary Table 3.S2.

### ***Isolation of BIMs***

Bacteriophage insensitive mutants (BIMs) were generated by challenging sensitive *S. thermophilus* strains and derivatives with virulent 2972 phages. Similar to the phage infection assay defined above, *S. thermophilus* strains were grown in LM17 at 42°C until an OD<sub>600</sub> of 0.3 was reached. 2972 phage was added to 200 µL cells at a MOI of 2 and mixed with 3 mL of top agar (0.75% LM17 agar) supplemented with 10 mM CaCl<sub>2</sub> and poured over 1.5% LM17+10 mM CaCl<sub>2</sub> plates and incubated at 42°C. To ensure independence, single survivor colonies were genotyped using primers matching to the leader sequence and first spacer of each CRISPR array to confirm phage spacer acquisition. Phage sensitivity of the isolated BIMs were tested by a spot test using a range of phage titers (10<sup>9</sup> -10<sup>2</sup> pfu/mL).

### ***Phage Spotting Assays***

*S. thermophilus* strains were grown in LM17 at 42°C until an OD<sub>600</sub> of 0.3 was reached. 1.5% LM17 + 10 mM CaCl<sub>2</sub> agar plates were pre-warmed in a 42°C incubator. 400 µL of cell culture was mixed in 4 mL top agar (0.75% LM17 agar) supplemented with 10 mM CaCl<sub>2</sub> and poured over the agar plate to create a lawn of cells. Five-fold serial dilutions of phage were spotted onto the plates (8 µL of each dilution) to the surface of the agar plate and incubated overnight at 42°C. Agar plates with phage plaques

were imaged using a Gel Doc system (Bio-Rad Gel Doc XR+ Gel Documentation System).

### ***Isolation and purification of mutant phages***

Phage mutants were generated from purification of single-plaques after re-infection of 2972 phage onto BIMs originating from the CRISPR1 null strain. Individual phage plaques from the CRISPR1 null strain were individually isolated using a blunt-ended 1 mL pipette tip and the gel piece was resuspended in 500  $\mu$ L LM17 + 10 mM  $\text{CaCl}_2$ . 200  $\mu$ L of the resuspended phage was used to re-infect 5 mL of *S. thermophilus* cells grown to a starting OD600 of 0.3. The liquid culture containing cells and phages were incubated in a 42°C shaker until complete cell lysis. Upon cell lysis indicated by full clearing of the cell culture, the culture was centrifuged at 3,000 rpm for 15 minutes. The cleared lysate was filtered using a 0.22  $\mu$ m syringe filter to remove cell lysis debris and the titer of purified phage lysate was quantified by infecting *S. thermophilus* cells with a range of diluted phage lysate and counting the total number of phage plaques per mL of phage.

### ***Non-CRISPR survivor (NCS) and phage genomic DNA extraction and sequencing***

#### ***S. thermophilus non-CRISPR survivor genome extraction***

*S. thermophilus* strains were grown in 100 mL of LM17 and incubated overnight at 37°C. Following overnight growth, cells were pelleted at 3,500 xg for 20 minutes. Cell pellets were flash frozen using liquid nitrogen and used for genomic DNA extraction. Bacterial genomic DNA was extracted from frozen cell pellets using the DNeasy

PowerBiofilm extraction kit (Qiagen, Germantown, MD, USA) following manufacturer's instructions.

#### *Phage genome extraction*

Genomic DNA of phage 2972 was isolated using a phenol-chloroform based extraction method. Purified phage lysate (2 ml) was treated with 7.5 µg/mL of both DNase I (Thermo Scientific, EN0523) and RNase A (Thermo Scientific, EN0531) for 30 min at 37°C to remove host nucleic acids. Following DNase/RNase treatment, phages were concentrated by ultracentrifugation at 40,000 xg for 2 hours at 4°C (Beckman XL90, SW41Ti rotor, Beckman 344059 Ultra-Clear 13.2 mL open top ultracentrifugation tubes). The supernatant was discarded, and remaining phage pellet was resuspended in 500 µL phage buffer (50 mM Tris-HCl pH 7.5, 100 mM NaCl, 8 mM MgSO<sub>4</sub>) supplemented with 20 mM EDTA pH 8.0, 50 µg/mL proteinase K (New England Biolabs), and 0.5% SDS, and incubated for 1 hour at 56°C. Following incubation, equal volume of phenol:chloroform:isoamyl alcohol (Fisher Scientific, 25:24:1) was added and gently inverted prior to centrifugation at 3,000 xg for 5 min at 25°C and supernatant was carefully collected for subsequent wash steps. This step was repeated twice. Equal volume of chloroform was then added to the collected supernatant and gently inverted prior to centrifugation (3,000 xg for 5 min at 25°C) (Eppendorf 5424 Microcentrifuge). The supernatant was collected and 10% volume of 3M NaOAc (pH 7.5) was added and resuspended. 100% cold ethanol was added and incubated at 25°C for 30 minutes. The samples were centrifuged at 20,000 xg for 20 min at 4°C and the supernatant was removed. 500 µL 70% ethanol was used to resuspend the pellet and centrifuged (20,000

xg for 20 min at 4°C). The supernatant was removed, and pellets were dried in an open microcentrifuge tube for 30 minutes at 25°C. DNA pellets were resuspended in 50 µL water and quantitated.

#### *Library preparation and sequencing*

Sequencing libraries were prepared using the Illumina DNA prep kit (Illumina, San Diego, CA, USA) and sequenced on an Illumina MiSeq instrument with reagents and protocol set to generate 150 bp unpaired reads (first set of bacterial samples) or 2 x 300 bp paired-end reads (phage samples and all additional sets of bacterial samples), both with 8 bp dual indexing.

After sequencing, reads were de-multiplexed and adapter-trimmed using the generate FASTQ module within the MiSeq Reporter analysis package provided with the instrument. Reads were aligned to the appropriate reference sequence (phage 2972 or *Streptococcus thermophilus* DGCC7710 genome) using bowtie2 (65) in local mode with default settings. Aligned reads were sorted and indexed with samtools, default settings (66). Short sequence variants (SNPS, short insertions and deletions) were identified using bcftools mpileup and bcftools call (67), with the ploidy option set to one and default settings. To search for larger mutations, alignment files were used to generate custom genome coverage tracks (66,68), which were then visualized on the UCSC Genome browser (69). Tracks were examined individually for any coverage gaps or anomalies, which are indicative of deletions, duplications, or inversions.

## Results

### ***S. thermophilus* Csn2 controls PAM-dependent spacer acquisition *in vivo***

To investigate the *in vivo* roles of each Type II-A Cas protein and the tracrRNA in directing PAM-dependent protospacer recognition and processing, we generated strains in which one or more of the Type II-A Cas or tracrRNA genes ( $\Delta$ Cas9,  $\Delta$ Csn2,  $\Delta$ Cas9-Csn2,  $\Delta$ Cas2,  $\Delta$ Cas1-Cas2-Csn2,  $\Delta$ tracrRNA) was deleted from the chromosome (Figure 3.1A). RexAB nuclease active site mutants were additionally generated to determine if the nuclease activities of these DNA repair proteins influenced adaptation as has been observed for related RecB-like nucleases in other systems (42). We then characterized newly integrated spacer sequences into the CRISPR1 array of *S. thermophilus* and the adjacent downstream PAM in both wild-type (WT) and mutant strains by using high-throughput sequencing (Figure 3.1B). To maximize the capture of expanded CRISPR arrays, we employed the “CAPTURE” method (Figure 3.1C) (70).

As expected, spacers taken up into the CRISPR1 locus of the wildtype *S. thermophilus* strain are derived from protospacer DNA exhibiting a conserved 5'-NNAGAAW-3' PAM sequence located immediately downstream (Figure 3.1D) (44,61,71). Loss of Csn2 and Cas1-Cas2-Csn2 severely reduced efficiency of adaptation, as indicated by the total number of unique spacers detected in the CRISPR array at position 1 (Figure 3.1D, left panel). We note that deletion of adaptation proteins (Cas1, Cas2, Csn2 and Cas9) and the resulting loss of adaptation was expected based on our earlier work (44) and further confirms the important role of each Cas protein for efficient adaptation. Despite the much lower number of unique new spacers, we examined the sequences upstream and downstream of the protospacers for evidence of PAMs.



Consistent with previous findings, deletion of Cas9 or Cas1, Cas2 and Csn2 abolished adaptation (data not shown) confirming the importance of these four Cas proteins in adaptation. Deletion of Csn2 alone however, did result in spacer levels above background levels when compared with the Cas1, Cas2, Csn2 null strain that should not be active in spacer uptake and integration (43,72). Loss of Csn2 resulted in a reduction of new protospacers with canonical PAMs compared to wild-type.

To identify a role for tracrRNA in adaptation, a tracrRNA deletion strain was generated whereby the region of tracrRNA that recognizes the repeat region of the crRNA was removed, which is expected to render the tracrRNA inactive due to inability to base-pair with crRNA repeats. This tracrRNA truncation mutant did not decrease the efficiency of adaptation, but far fewer protospacers had a PAM as compared to the wildtype strain (Figure 3.1D, right panel).

To determine whether RexAB nuclease activity contributes to PAM specification during spacer acquisition, adjacent downstream sequences were analyzed for proper PAM sequences in RexAB mutation strains. The mutations created in the RexA and RexB subunits were designed to disrupt the highly conserved motifs (DYK and GIID) within the nuclease domain of both enzymes. Neither mutation in either subunit appeared to affect efficiency of adaptation or PAM recognition (Figure 3.1D, right panel). These results indicate that the predicted nuclease activities of RexAB are not involved in the recognition and selection of PAM during protospacer generation and this process is likely a functional role of Cas9 and Csn2.

Given our observation that loss of Csn2 did not result in a significant phenotype due to major effects on the reduction of adaptation frequency, we elected to identify

phenotypes when the proteins were overexpressed to gain insight on the functional role of Csn2. To test this, we included an overexpression plasmid in the WT strain of *S. thermophilus* to express excess Cas1, Cas2 and Csn2 proteins and analyzed new spacer sequences for frequency of protospacer selection from PAM-containing DNA regions (Figure 3.1E). Under endogenous levels of all CRISPR-Cas proteins, a strong consensus for the proper 5'-NNAGAAW-3' PAM was observed for spacers originating from both the genome and plasmid. However, overexpression of Cas1-Cas2 alone resulted in a severe reduction in the number PAM-adjacent sequences that were acquired for integration regardless of spacer source. Although the consensus was low, the PAM sequence pattern was still maintained. We believe that the endogenous levels of Cas1, Cas2, Csn2 and Cas9 are functioning properly *in vivo* to maintain PAM recognition. Additionally, the total number of spacers were significantly higher in the strain overexpressing Cas1-Cas2 than in the wild-type strain with endogenous levels of all Cas proteins. To test a potential regulatory role of Csn2 during spacer integration by Cas1-Cas2, we compared the properties of spacer integration when just Cas1 and Cas2 were overexpressed with the scenario when Cas1, Cas2 and Csn2 were overexpressed. We found that overexpression of Cas1-Cas2 only not only resulted in a higher number of total acquired spacers, recognition for the 5'-NNAGAAW-3' PAM was severely diminished. When Csn2 was additionally expressed with Cas1 and Cas2, the number of total acquired spacers decreased compared to the overexpression of Cas1-Cas2 but higher than the no plasmid wild-type strain. High specificity for the correct PAM was also observed in the Cas1-Cas2-Csn2 overexpression strain at similar levels to the no plasmid wild-type strain compared to the Cas1-Cas2 overexpression strain. Together, these results not only

confirm the importance of Cas proteins for efficiency adaptation but indicates that Csn2 plays a role in selecting PAM-containing sequences that does not occur with Cas1-Cas2 alone.

### **Spacer duplication is influenced by tracrRNA and reduced spacer acquisition**

In this work, the *S. thermophilus* strains with truncated tracrRNA did not result in a loss of adaptation efficiency (Figure 3.1D). However, a slight reduction in PAM selection was observed. These results suggested that the apo-form of Cas9 is functioning differently than the holo-Cas9 bound by tracrRNA and the crRNA. To address the question of whether tracrRNA can influence spacer acquisition by Cas9 in *S. thermophilus*, we investigated the phenotype of the tracrRNA deletion strain with respect to spacer origin and the distribution of self-targeting protospacers across the *S. thermophilus* genome. We also looked for protospacer clusters or hotspots across the *S. thermophilus* genome in the  $\Delta$ tracrRNA strain. The distribution of acquired protospacers was the same for all mutant strains compared to WT, however the genome protospacer hotspots in the  $\Delta$ tracr strain were specifically located at the rRNA (ribosomal RNA) gene cluster (Figure 3.S1). The rRNA cluster includes the 5S, 16S, 23S rRNA and Asn-, Ala-tRNA genes. Previous studies have implicated that high rates of transcription of rRNA are necessary for optimal growth (73). Additionally, highly transcribed regions have been observed as targeted regions for spacer uptake in CRISPR-Cas systems (74). Our results suggest that tracrRNA not only acts as a co-factor for Cas9 but appears to influence the spacer acquisition patterns. A second cluster of new spacers was additionally detected and found to represent spacer duplications of pre-existing spacers in the CRISPR array.

During CRISPR-Cas immunity, adaptation requires the capture and integration of new spacer sequences at the leader-proximal repeat of the CRISPR array (Figure 3.2A). Under circumstances that are not well understood such as integration errors or recombination events within the CRISPR array, we demonstrate that duplications of pre-existing spacers can occur to generate multiple copies of an already existing spacer. We analyzed the origin of the unique spacers for each strain and mapped them back to the location on the *S. thermophilus* genome. Any new spacer that aligned to a pre-existing spacer was considered a duplication event. In WT, the majority of new spacers were acquired from unique regions of the genome although 22% of expanded arrays were the result of spacer duplications (Figure 3.2B). Interestingly, in strains missing one or more adaptation proteins (Cas1, Cas2, Csn2 and or Cas9), the frequency of spacer duplication in the expanded arrays was nearly 100% for each strain (>98%) indicating that spacer duplications at CRISPR arrays primarily occur independently of Cas protein function. We found that RexAB nuclease-deficient strains, exhibited high relative levels of duplicated spacers (RexA GIIA: 55.2%, RexA AYK: 74.7%, RexB GIIA: 34.2%, RexB AYK: 82.4%). The frequency of spacer duplication in the RexAB mutant strains was more than WT but less than the adaptation null strains. This suggests that the role of RexAB nuclease activity may have an intermediate effect on the efficiency of adaptation compared to the Cas protein deletion strains. Surprisingly, in the tracrRNA deletion strain, spacer duplication frequency was lower than WT (5.1% compared to 22%). These results suggest that there are likely functional differences of Cas9 in the absence or presence of its co-factor during spacer acquisition or selection of functional protospacer sequences. In contrast, the wild-type tracrRNA may independently bind to the DNA

repeat region of the CRISPR repeat to catalyze sequence duplication and thus partial deletion of tracrRNA could hinder the duplication reaction.

To further investigate the pattern of spacer duplication events, particularly in the  $\Delta$ tracrRNA strain, we examined the aligned spacers within the CRISPR1 array for WT, the CRISPR1 null ( $\Delta$ Cas1-Cas2-Csn2) and  $\Delta$ tracrRNA strains (Figure 3.2C). The heights of the peaks indicate the relative number of reads supporting each spacer duplication. Based on these heights, we observed a strong preference for duplication of the first two spacers regardless of the mutation strain (CRISPR1 null or  $\Delta$ tracrRNA) and the first three spacers in WT (Figure 3.2C, top panel). Every spacer in the CRISPR1 array was found to have been duplicated at low levels in the cell population except for a single mid-array spacer. Certain spacers were duplicated with higher frequency near the middle of the array, such as spacers 11, 13, 14 and 21 (for all strains). Additionally, spacer 19 exhibited no duplication in all strains and replicates for unknown reasons. As noted before, the number of duplication events was significantly higher in strains where adaptation is severely reduced (CRISPR1 null: up to 16,000 total reads) compared to those with functional adaptation proteins (WT and  $\Delta$ tracr: up to 3,000 total reads) (also described in Figure 3.2B).

Analysis of unique spacer duplication sequences also revealed that improper duplication of spacers in the  $\Delta$ tracr strain occurred more frequently than in WT and CRISPR1 null (Figure 3.2C, bottom panel). Spacers were more likely duplicated in the reverse orientation (pink bars) and sequences of the spacer or repeat were deleted arbitrarily which would result in non-functional crRNAs. The results also suggest that spacer duplication involves functional adaptation giving spacer duplication a potential

biological role for maintaining immunity of relevant spacers. In contrast, these events may additionally occur due to spontaneous nicking, recombination of the CRISPR array, or replication errors derived from mechanisms similar to trinucleotide repeat expansions that results in the duplication of sequences.

### **RexAB nuclease activity promotes spacer acquisition**

Alignment of several exonuclease/helicase enzymes reveal poor homology over most of the protein, the RecB nuclease motifs in RecBCD-like proteins are highly conserved. In RecBCD, these motifs correspond to the nuclease activities of RecB enzyme (16,75,76). We show that this motif is present in not only the homologous two-subunit exo/hel enzymes AddAB, but in the similar and not widely characterized RexAB species (Figure 3.3A). Characterization of the nuclease motifs have indicated that nuclease activity of the AddAB two-subunit nucleases are primarily found within the DYK RecB motif-III motif (75-77). This domain in AddA was found to influence spacer acquisition patterns in the Type II system in *S. aureus* (30). These corresponding motifs are additionally found in the CRISPR Cas4 proteins (78). Previous studies in *Pyrococcus furiosus* show that mutation of a single motif in the GIID RecB motif-II, abolishes Cas4 nuclease activity (42). This mutation clearly showed that nuclease activity of Cas4 proteins are essential in defining PAM, length and orientation of DNA fragments prior to spacer integration into the CRISPR array (42). Both DYK and GIID motifs are conserved in RexAB (Figure 3.3A). Therefore, the conservation of the RecB-like nuclease motifs between RexAB and Cas4 lead us to ask whether RexAB proteins are similarly involved in CRISPR-Cas adaptation. To test this, we generated two independent mutations in each

subunit of RexAB. The aspartic acid in the DYK motif of RexA and RexB were mutated to an alanine (RexA D1164A; RexB D892A) as well as the aspartic acid in the GIID motif of RexA and RexB (RexA D1151A; RexB D878A). Each of the RexAB mutant strains were generated in a Cas9 deletion strain and transformed with a plasmid expressing Cas1-Cas2-Csn2-Cas9 (pCas). In the presence of the over-expression plasmid, adaptation can be observed following PCR amplification of the region between the leader and first spacer (Figure 3.3B, lane 1 and 2, spacer uptake indicated by the additional of a single repeat-spacer unit (+1)). The RexAB mutant strains demonstrated either no spacer acquisition (lanes 3, 4, and 6), or reduced spacer acquisition (see faint band, RexB GIIA mutation strain, lane 5). In the presence of a catalytically defective Cas9 on the over-expression plasmid (dCas9), targeting ability by Cas9 is eliminated and spacer accumulation is significantly increased with nearly all cells having acquired at least 2 new spacers (Figure 3.3C, lane 2 compared to lane 3). In the dCas9 background, the RexAB mutant strains exhibited either severely reduced spacer acquisition with the majority of the cells acquiring 1 new spacer (RexA GIIA, (data not shown); RexB AYK, lane 6), or moderately reduced spacer acquisition (RexA AYK, lane 4; RexB GIIA, lane 5). All RexAB mutation strains resulted in reduced adaptation as evidenced by increased signal for cells with unexpanded arrays (0) compared to the WT strain. These results indicate that RexAB mutation strains in both motifs of the nuclease domain reduce adaptation levels in the context of plasmid-based overexpression of Cas1-Cas2-Csn2-Cas9.

To investigate whether the RexAB mutation strains affect spacer acquisition from phages, WT and each RexAB mutant strain was infected with lytic phage 2972 and

surviving colonies were assayed for an expanded array representing spacer acquisition against the phage. WT *S. thermophilus* infected with the lytic phage resulted in 11/15 surviving colonies having picked up a new spacer in the CRISPR1 array (Figure 3.4). Additionally, a single survivor adapted in the CRISPR3 array with no expansion in the CRISPR2 and CRISPR4 arrays. These results with the WT strain are consistent with numerous previous studies showing that Type II-A CRISPR-Cas system affiliated with CRISPR1 array is the dominant system but that the second Type II-A system affiliated with CRISPR3 array is also weakly active, while no activity is observed with the Type III-A system (CRISPR2 array) or Type I-E system (CRISPR4 array) (1,55,56). Mutations in RexA (GIIA and AYK) significantly reduced the ability of cells to adapt against the phage with only 1/15 survivors having undergone spacer acquisition in CRISPR1 for both strains. Reduced spacer acquisition in the CRISPR1 array was additionally observed with RexB mutant strains (5/15 in the RexB GIIA strain and 3/15 in the RexB AYK strain) although the effect was not as significant as the RexA mutant strains. Taken together, the results indicate that RexAB nuclease activity contributes to spacer acquisition. Additionally, the spacer characteristics for the RexAB mutant strains such as PAM sequence and spacer size (data not shown), were the same as wild-type (Figure 3.1). These results further suggest that the role RexAB during adaptation is likely upstream of protospacer processing and new spacers are acquired as degradation intermediates of RexAB processing.



### **Mutations in FtsH hinder phage infectivity in non-CRISPR survivors**

In the phage infection assay described above, we note that some cells that survived phage infection did not acquire a spacer in any CRISPR array (Figure 3.4), and thus are non-CRISPR survivors (NCS). In addition, a significant number of phage infection survivors are observed in a CRISPR1 null strain ( $\Delta$ CRISPR1 array) that did not adapt a new spacer against the infecting phage (Figure 3.4). We examined the context of these non-CRISPR survivors (NCS) in the phage infection assay (Figure 3.5A) by characterizing the total number of expanded Type II CRISPR arrays (CRISPR1 and CRISPR3). Of the genotyped survivors, over 75% of the survivors in the wild-type *S. thermophilus* cells expanded in the CRISPR1 array with very little expansion observed in the CRISPR3 array (Figure 3.5B). As expected, the CRISPR1 null strain (deletion of all CRISPR1 repeat-spacer units) did not exhibit expanded (or unexpanded) CRISPR1 arrays in its survivors and also had very few expanded CRISPR3 arrays. We confirmed that despite similar number of phage infection survivors (not shown) between the wildtype and the CRISPR null strain, nearly 100% of the survivors for the CRISPR null strain were non-CRISPR survivors (Figure 3.5C).

The numbers suggested that non-CRISPR associated survival mechanisms may be an important part of phage resistance and we sought to determine their genetic basis. Starting with the standard phage infection assay to characterize *S. thermophilus* survivors, we generated NCS isolates and second-generation mutant phages (Figure 3.5A). Strains devoid of an active CRISPR system were generated and tested against WT lytic 2972 phages, resulting in survivors that were not a result of spacer acquisition in a CRISPR array. These new NCS strains were initially insensitive to the WT 2972 phage

(Figure 3.5D, Figure 3.S2). We re-infected these NCS strains and, surprisingly, observed plaque formation on agar plates, which is indicative of viable phages (Figure 3.S2). The phages from these plaques were isolated, deemed “mutant 2972” phages, and used to challenge both WT and NCS strains. WT *S. thermophilus* was still sensitive to the mutant 2972 phage. The NCS strains were infected with the mutant 2972 phages and plaque formation was observed, demonstrating that the mutant 2972 overcame phage resistance in the NCS isolates (Figure 3.S2, Figure 3.5F).

In all, we isolated and prepared 11 NCS strains for genetic analysis, along with two isolates of mutant 2972 phages that could overcome NCS resistance. Eight NCS strains contained mutations in the gene coding for an ATP-dependent metallopeptidase FtsH/Yme1/Tma family protein (FtsH). A different mutation was found in each of the eight NCS strains and the mutations were located throughout the coding region of the gene. Mutations consisted of pre-mature stop codons (Stop-96, Stop-240), single amino acid substitutions (Q97R, G224D, G289V, T306I, R359L, M539I) and a frameshift mutation (FS-512) (Figure 3.5E). Additionally, NCS strains exhibited moderate growth defects compared to WT *S. thermophilus* (data not shown).

To explore the effects of FtsH on phage resistance as observed with the NCS strains, we tested phage susceptibility of WT and NCS strains to both WT and two individually isolated second-generation mutant 2972 phages. Additionally, WT FtsH was complemented on an expression plasmid (Figure 3.5F). WT *S. thermophilus* was sensitive to both WT and mutant phages, but was more sensitive to the WT phage (Figure 3.4E, top left panel). The non-CRISPR survivor originating from a CRISPR1 null strain was resistant to the WT phage but was sensitive to both isolates of the mutant phages (top

right panel). In the presence of an empty plasmid (pEmpty) with the non-CRISPR survivor strain, phage sensitivity and resistance was comparable to the parental strain without the plasmid (bottom left panel). However, expression of WT FtsH on the plasmid (pFtsH) in the same non-CRISPR survivor strain rescued susceptibility of this strain to the WT phage (bottom right panel). These results suggest that mutation in the FtsH gene is contributing to phage resistance in the absence of CRISPR-mediated defense, and they demonstrate, for the first time, evidence of a host's response to phage infection in *S. thermophilus* through mutations predicted to alter the expression or function of a membrane protease.

Next, we characterized the two isolated second-generation mutant phages to identify genetic changes. The mutant phages were purified, and genomic DNA was extracted for whole-genome sequencing to identify common mutations. Interestingly, two mutations were identified in a hypothetical protein within the coding region of tail proteins: a one nucleotide SNP was found in mutant phage 1, G10473C, and a two nucleotide SNP, GG10485TT was found in mutant phage 2 (Figure 3.5G). No other SNPs were observed. Further alignment studies suggested that the hypothetical protein is likely a tail assembly chaperone. Both SNPs would result in amino acid substitutions (G100R and G104F) in the C-terminal end of the tail assembly chaperone. Given the evidence that tail proteins provide specificity of host cell recognition, we reason that this adaptation observed by the invader could arise from evolution between host and invader. Our findings provide strong evidence that under some environments, secondary mechanisms such as self-mutation of genomes is an adapting survival mechanism that is independent of CRISPR-Cas systems.

## Discussion

Spacer acquisition is an essential step of CRISPR adaptation that contributes to effective protection against foreign nucleic acids. Successful acquisition by CRISPR-Cas systems requires sequence recognition and processing of foreign DNA elements for integration into the CRISPR array. This sequence recognition is not only dependent on the discrimination of foreign nucleic acids but is reliant on selection of protospacers in a PAM-dependent process. However, the processes underlying how protospacers are recognized, generated and selected are poorly understood. Here, we report the first evidence that the nuclease activity of DNA repair proteins, RexAB contributes to adaptation in *S. thermophilus*. Our results indicate that RexAB is not involved in PAM-dependent processing of pre-spacers but is likely involved in CRISPR adaptation through the generation of substrates for spacer uptake. Spacer acquisition is also a PAM-dependent process to ensure discrimination between host and invader during target silencing. Our results are consistent with previous findings that Cas9 is important for adaptation and is likely the primary protein recognizing the PAM sequence of newly acquired spacers (43,44). However, our work further reveals a role for Csn2 in the selection PAM-flanking spacer sequences prior to spacer integration. Given these results, we propose a speculative model that Cas9 scans protospacer sequences (likely generated as degradation intermediates by RexAB and by other processes) in a PAM-dependent manner until the correct 5'-NNAGAAW-3' sequence is recognized. Upon PAM recognition, Cas9 binds tightly to the protospacer. Cas1-Cas2-Csn2 bind as a complex, engaging free DNA ends until the complex encounters Cas9 bound to the PAM. By an unknown mechanism, the protospacer is processed by nucleases to remove the PAM. The

resulting pre-spacer is integrated as a new spacer into the CRISPR array by Cas1-Cas2. Additionally, in the context of spacer acquisition and integration, our data demonstrates that duplication of pre-existing spacer sequences occurs at high frequency in the absence of Cas proteins required for new spacer integration. The frequency of duplication of pre-existing spacers was additionally reduced in the absence of the Cas9 co-factor, tracrRNA, and suggest that either the apo-form of Cas9 functions very differently from the holo-Cas9 during adaptation or that tracrRNA may function independently of Cas9 association. This work also implicates a metalloprotease membrane protein, FtsH, in the life cycle of phage 2972 as loss of fully functional FtsH confers resistance in the host. Mutation of a tail assembly chaperon gene in the phage can overcome the FtsH mutation in the host demonstrating adaptive mechanisms between host and invader. Collectively, our work reveals several key factors involved in the acquisition of functional spacer sequences as well as critical functions for CRISPR-independent factors in mediating phage defense.

### **Substrate generation by RexAB nucleases for spacer uptake**

The main functional role of RexAB in Gram-positive bacteria is the repair of dsDNA breaks. However, other studies have linked similar exonuclease/helicase enzymes to anti-viral defense (10, 16). The process of utilizing the cellular functions of RecBCD-like enzymes as a means of antiviral defense systems likely involves degradation of linear phage DNA following infection. is through the function of the enzyme to degrade linear DNA- a key characteristic of phage DNA upon infection. In this process, the RecBCD-like RexAB binds to double-stranded DNA ends and translocates along the linear

fragment until a chi sequence is reached (33). This species-specific cis-regulatory element dramatically alters the properties of RexAB nuclease and helicase activity to pause further degradation (33,35). This process of regulation has evolved to act as a method of distinguishing between self (chromosomal DNA) and non-self (invading phage DNA) as chi sites are more abundant in genomes of bacteria such as *E.coli* which are found at an average of every 5 kb compared to the genomes of phages (27).

In the Type I-E system in *E.coli* and Type II-A system of *S. aureus*, RecBCD and its homolog AddAB influence adaptation (10,11,30). In particular, the nuclease activity of the main subunit AddA contributes to spacer acquisition. Additionally, characterization of protospacer hotspots revealed boundaries defined by chi-sequences suggesting that spacer substrates are derivatives of AddAB end-processing (30). The role of AddAB during spacer acquisition upon infection initially requires recognition of the phage DNA upon entry. Previous data showed that new spacers are acquired immediately following injection of viral free DNA ends into the cell and end-processing by AddAB likely contributes to the generation of substrates for acquisition by CRISPR proteins (30). Our results provide the first evidence that the RecBCD homolog, RexAB contributes to adaptation in *S. thermophilus* through similar pathways (Figure 3.3, 3.4). Mutation of two motifs within the nuclease domain of RexA and RexB reduced adaptation frequency of self-derived genome spacers in the context of over-expressed Cas proteins (Figure 3.3) as well as in the context of an invader under endogenous levels of Cas proteins (Figure 3.4). However, a decrease in the number of unique spacers was not observed in the sequencing assay with endogenous levels of Cas proteins (Figure 3.1). We speculate that nuclease activity of RexAB supplies protospacer substrates but is not involved in processing of

protospacers. Therefore, it is likely that availability of Cas1, Cas2, Csn2 and Cas9 limits adaptation and RexAB mutations are curtailing spacer uptake when Cas proteins are no longer limiting this supply through overexpression. Additionally, despite recent evidence of RecB-like nuclease domains in Cas4 functioning in PAM-dependent protospacer processing, mutation of this motif did not impact PAM specificity of newly acquired spacers (Figure 3.1D). Without an impact on PAM recognition and simply the reduction of acquired spacers during adaptation, the role of RexAB is very likely contributing to CRISPR adaptation through the generation of pre-spacer intermediates.

### **Role of Csn2 in maintaining PAM-dependent pre-spacer capture**

During spacer generation, proper PAM recognition is required for cleavage of foreign nucleic acids. Additionally, specificity for PAM-flanking protospacers during spacer acquisition further differentiates between invader and integrated spacers within the CRISPR array. Therefore, targeting of protospacers with the proper flanking PAM sequences is a critical process during spacer capture. Sequencing analysis of newly acquired spacer sequences in the Type II-A CRISPR-Cas system demonstrated that spacer uptake depends on recognition of the 5'-NNAGAAW-3' PAM (Figure 3.1D). Our results support previous findings that Cas9 is necessary for efficient spacer uptake (data not shown) and is likely contributing to PAM-dependent targeting similar to *S. pyogenes* (43,72). Additionally, preference for pre-spacers containing the proper PAM sequence was influenced by the presence of Csn2 (Figure 3.1E). In the absence of Csn2, PAM-specific spacer acquisition was maintained despite the low number of acquired spacers suggesting that PAM recognition is not dependent on Csn2 (Figure 3.1D). However,

preference for PAM-containing spacers for integration was demonstrated by the over-expression of Csn2 in the presence of excess Cas1-Cas2 (Figure 3.1E). We hypothesize that the dilution of PAM in newly acquired spacers in the presence of excess Cas1-Cas2 may result from promiscuous spacer integration by the excess Cas1-Cas2 integrase complex. Though adaptation was described to require all four Cas proteins (Cas1, Cas2, Csn2 and Cas9) *in vivo*, recent studies have only visualized adaptation efficiency using gel electrophoresis (43,44). Additionally, integration activity of Cas1-Cas2 has been observed *in vitro* suggesting that although all four Cas proteins are required for functional adaptation *in vivo*, spacer integration of dsDNA substrates may still occur with only Cas1 and Cas2 (47,50,79). Together, these findings contribute to the on-going hypothesis that Csn2 is contributing to adaptation by regulating Cas1-Cas2 spacer integration until the complex encounters a Cas9 protein bound to a PAM-containing pre-spacer (54).

### **Spacer duplication as a potential CRISPR-mediated method of defense**

During spacer integration, Cas proteins integrate a new spacer into the CRISPR array at the leader-proximal repeat. *In vitro* studies have provided mechanistic details as to how new spacers are integrated into the array as a new repeat-spacer unit. However, in the absence of CRISPR-Cas proteins to perform this function, our data has demonstrated duplication of pre-existing spacer sequences occurring at high frequencies. Our approach of studying spacer duplication in a cultured population of cells in mutation strains devoid of the Type II-A Cas proteins provides us with a snapshot of events that occur when adaptation is non-functional. These results demonstrate that duplication of pre-existing spacers are occurring at relatively low levels in wildtype, with 22% of captured expanded



arrays showing spacer duplication. Interestingly, in the absence of CRISPR-Cas proteins, nearly 100% of captured expanded arrays were as a result of spacer duplication (Figure 3.2B). Analysis of duplicated spacers demonstrated that almost all spacers in the CRISPR array were found to have been duplicated at least one time at the leader-proximal repeat. However, a strong preference for duplication of the first three spacers was clearly evident (Figure 3.2C). Though these specific events have not been described previously and the biological role of spacer duplication has not been investigated, spacer duplication may contribute to supporting long-term CRISPR-mediated survival. Our data leads us to hypothesize that pre-existing spacer sequences are continuously duplicated and positioned at the first position of the array to maintain a constant level of immunity when new spacers are not being actively acquired. We cannot dismiss however, that spacer duplication is simply a result of incomplete spacer integration through array nicking and host repair or a phenomenon arising from recombination within the CRISPR array. Another potential cause of spacer duplication could be a result of slippage during DNA replication, also known as triplet repeat expansion or trinucleotide repeat expansion. Due to the repetitive nature of the repeat sequences in the CRISPR array, the repeat sequences may form unusual DNA loop structures during DNA synthesis that causes misalignment between the synthesized and template DNA resulting in sequence expansions (80,81). In addition to occurring during DNA synthesis, trinucleotide repeat expansion events can occur during DNA repair. The pathways involved in the repair or synthesis of DNA and its influence on spacer duplication is not well documented. However, these findings raise an interesting possibility that spacer duplication may be driven by the frequency of spacer uptake as the frequency increases in the absence of adaptation.

## **tracrRNA as a functional co-factor of Cas9 in discriminating between self and non-spacers**

In *S. thermophilus*, all four Type II-A CRISPR-Cas proteins (Cas1, Cas2, Csn2 and Cas9) as well as tracrRNA (trans-activating crRNA) are required for adaptation, however nuclease activity of Cas9 is not (44). Despite Cas9 being widely characterized during CRISPR-mediated immunity, the specific role that Cas9 plays in protospacer to pre-spacer generation is not well understood. During target interference, Cas9 bound to tracrRNA:crRNA form an effector complex that base-pairs to the target sequence to induce sequence-specific cleavage (61). However, in its apo-form, structural analysis revealed that in the absence of its crRNA + tracrRNA co-factors, the PAM-recognition domain is in an inactive configuration (82). Interestingly, our results demonstrated that truncation of tracrRNA did not abolish PAM-recognition but resulted in a severe dilution of PAM-adjacent spacers. The 5'NNAGAAW-3' motif was weakly conserved, and the majority of new spacer sequences did not contain the canonical PAM (Figure 3.1D). This may suggest that Cas9 in the apo-form has less efficient PAM-recognition activity. In the similar species *S. pyogenes*, full deletion of tracrRNA prevented spacer acquisition, suggesting that the apo-form of Cas9 is not sufficient to promote spacer acquisition without the presence of the crRNA and tracrRNA (43,83).

Moreover, the observed hotspot region for the preferred site of spacer acquisition in the context of self-acquired spacers in the tracrRNA deletion strain occurred within the rRNA gene cluster (containing 5S, 16S, 23S, tRNA-Asn and tRNA-Ala) (Figure 3.S1). rRNAs are highly transcribed genes in both bacteria and archaea, and high transcription levels have been implicated in the generation of protospacers (73,74,84-86). Our results

demonstrate that the apo-Cas9 enzyme targets these highly transcribed regions suggesting that in the absence of its co-factors, fidelity of self-avoidance during spacer acquisition may be affected. It has been shown in a previous study that in the presence of a nuclease-defective Type II-A *S. thermophilus* Cas9, the majority of all newly acquired spacers were from the host genome (44). Additionally, previous studies demonstrating non-specific DNA binding of the apo-Cas9 may reveal underlying details about the regulatory function of Cas9, its co-factors such as tracrRNA, and how a functional Cas9 contributes to ensuring proper activity of the enzyme during spacer recognition (87). The apparent requirement for tracrRNA for specific Cas9 function may have additionally evolved to play a role in spacer duplication. Although the biological role or mechanisms of spacer duplication is not known, we speculate that these events are either rare occurrences that are a result of recombination due to the numerous tandem direct repeats, or an evolved mechanism to enhance immunity by increasing copy of number of specific crRNAs produced from a given array. Collectively, the findings implicate a role of tracrRNA as a Cas9 co-factor important for Cas9 fidelity during multiple stages of adaptation to ensure effective CRISPR-Cas immunity in downstream processes.

### **CRISPR-independent phage survival by FtsH inactivation**

Our findings indicate that FtsH plays a major role in phage sensitivity in *S. thermophilus*. These findings are consistent with similar work performed in *Lactococcus lactis* in which the *ftsH* gene impacts the life cycle of temperate phage TP712 through unknown mechanisms (57). In *E. coli*, FtsH plays a role in lambda phage infection by proteolysis of phage protein CII, a transcription factor which modulates phage lysis

versus lysogeny (88-90). Accordingly, the *S. thermophilus* FtsH protease may normally act on key regulatory phage proteins required for the 2972 phage life cycle. More work is required to understand how the *S. thermophilus* FtsH membrane protease hinders cell infection by bacteriophage 2972.

Our data revealed that in a CRISPR1 null strain, frequency of cell survival after phage infection was not lower than in wild-type, despite the absence of an active CRISPR system (Figure 3.4). Whole genome sequencing revealed that *S. thermophilus* non-CRISPR survivors were resisting phage infection through distinct mutations in the membrane-bound metalloprotease *ftsH* gene predicted to lead to loss or functional inactivation of FtsH protein (Figure 3.5E, 3.5F)). A study in *L. lactis* observed that FtsH did not hinder phage adsorption, DNA delivery or activation of the lytic cycle (57). Therefore, based on the function of FtsH in maintaining membrane integrity in Gram-negative bacteria, we speculate that mutation of FtsH is a means to prevent cell lysis during phage infections in *S. thermophilus* (91), though the exact mechanism is not clear. It is possible that mutation of FtsH in *S. thermophilus* results in physiological changes to the cell wall structure resulting in failed recognition of cell surface receptors by phages, blocks DNA injection, or impairs phage assembly of cell lysis. A previous study in *S. thermophilus* demonstrated that a host gene coding for a metalloprotease, methionine aminopeptidase (metAP), is necessary for phageDT1 infection (92). Similar to effects observed in FtsH mutant strains, phage adsorption, DNA replication and protein expression were not affected by the mutation, suggesting a wide-range of host adaptations in evading phage infections.

To understand the molecular mechanism underlying phage resistance in the NCS strains, second-generation mutation phages were isolated and sequenced for common mutations (Figure 3.5A, 3.5F, 3.S2). Whole-genome sequencing of the mutant phages revealed mutations in the same hypothetical protein gene, which we predicted to be a tail assembly chaperone (TAC) (Figure 3.5G). Genes encoding TAC are found in the majority of long-tailed phage genomes and are flanked by genes encoding the tail tube and tape measure proteins that are related by a translational frameshift (93). The requirement of frameshifting events for tail formation has been observed in several *E. coli* phages and suggests that these events are internally controlled for the assembly of tails (94,95). In the context of the mutation of TACs to infect NCS strains, the evolutionary adaptation by the phages is suggested as a mechanism by which TAC could alter tail assembly production. This simple adjustment may allow surface interaction of the phage in the NCS (mutant FtsH) host. An additional proposed function of the mutated TAC is to avoid specific proteolytic degradation by host surface proteases, allowing phage adsorption. Overall, our data demonstrates an example of adaptive mechanisms between host and invader, further contributing to the understanding of phage resistance in hosts as well as phage response in an evolutionary context.

In summary, we have characterized the role of CRISPR-Cas and RexAB proteins of *S. thermophilus* during spacer uptake for adaptation. Our results contribute to the current model that RecBCD-like enzymes such as RexAB contribute to spacer generation through a CRISPR-independent process. The function of these enzymes is likely upstream of CRISPR-mediated spacer uptake and it is suggested here that dsDNA processing provides substrates for PAM-dependent spacer recognition by Cas proteins.

Furthermore, we determine the specific role of Cas9 in PAM-dependent spacer uptake. Additionally, a regulatory role of Csn2 in ensuring Cas1-Cas2 spacer uptake is PAM-specific was demonstrated. We provide the first evidence illustrating spacer duplication of pre-existing CRISPR array spacers and a possible function of tracrRNA and apo-Cas9 in host survival through spacer duplication and discriminating against self-targeted spacers. Lastly, we provide evidence of host-regulated mutagenesis as a method of phage defense and evolutionary adaptations of the invader to overcome these mutations. Understanding the molecular basis of RexAB during spacer substrate generation is an important future goal that will likely require more in-depth molecular analysis and *in vitro* characterization. For example, there is a gap of knowledge in understanding if RexAB is capable of processing substrates after they are captured and bound by CRISPR-Cas proteins. Furthermore, there is a need for determining the specific role of Csn2, tracrRNA, as well as the apo-Cas9 versus holo-Cas9 during protospacer to pre-spacer generation and PAM-dependent processing. Finally, additional studies are necessary to address how mutations in FtsH of the host and TAC of the phage can affect the phage life cycle through adsorption, DNA delivery, and activation of the lytic cycle.

## 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. Koonin, E.V., Makarova, K.S. and Zhang, F. (2017) Diversity, classification and evolution of CRISPR-Cas systems. *Curr Opin Microbiol*, **37**, 67-78.
4. 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.
5. 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.
6. 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.
7. 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.
8. Marraffini, L.A. (2015) CRISPR-Cas immunity in prokaryotes. *Nature*, **526**, 55-61.
9. 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.

10. 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.
11. Radovcic, M., Killelea, T., Savitskaya, E., Wettstein, L., Bolt, E.L. and Ivancic-Bace, I. (2018) CRISPR-Cas adaptation in *Escherichia coli* requires RecBCD helicase but not nuclease activity, is independent of homologous recombination, and is antagonized by 5' ssDNA exonucleases. *Nucleic Acids Res*, **46**, 10173-10183.
12. Bianco, P.R., Brewer, L.R., Corzett, M., Balhorn, R., Yeh, Y., Kowalczykowski, S.C. and Baskin, R.J. (2001) Processive translocation and DNA unwinding by individual RecBCD enzyme molecules. *Nature*, **409**, 374-378.
13. Boehmer, P.E. and Emmerson, P.T. (1992) The RecB subunit of the *Escherichia coli* RecBCD enzyme couples ATP hydrolysis to DNA unwinding. *J Biol Chem*, **267**, 4981-4987.
14. Braedt, G. and Smith, G.R. (1989) Strand specificity of DNA unwinding by RecBCD enzyme. *Proc Natl Acad Sci U S A*, **86**, 871-875.
15. Chung, C. and Li, H.W. (2013) Direct observation of RecBCD helicase as single-stranded DNA translocases. *J Am Chem Soc*, **135**, 8920-8925.
16. 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.
17. Dillingham, M.S., Spies, M. and Kowalczykowski, S.C. (2003) RecBCD enzyme is a bipolar DNA helicase. *Nature*, **423**, 893-897.



18. Jockovich, M.E. and Myers, R.S. (2001) Nuclease activity is essential for RecBCD recombination in *Escherichia coli*. *Mol Microbiol*, **41**, 949-962.
19. Taylor, A.F. and Smith, G.R. (2003) RecBCD enzyme is a DNA helicase with fast and slow motors of opposite polarity. *Nature*, **423**, 889-893.
20. Wiktor, J., van der Does, M., Buller, L., Sherratt, D.J. and Dekker, C. (2018) Direct observation of end resection by RecBCD during double-stranded DNA break repair in vivo. *Nucleic Acids Res*, **46**, 1821-1833.
21. Wilkinson, M., Chaban, Y. and Wigley, D.B. (2016) Mechanism for nuclease regulation in RecBCD. *Elife*, **5**.
22. Anderson, D.G. and Kowalczykowski, S.C. (1997) The recombination hot spot chi is a regulatory element that switches the polarity of DNA degradation by the RecBCD enzyme. *Genes Dev*, **11**, 571-581.
23. Bianco, P.R. and Kowalczykowski, S.C. (1997) The recombination hotspot Chi is recognized by the translocating RecBCD enzyme as the single strand of DNA containing the sequence 5'-GCTGGTGG-3'. *Proc Natl Acad Sci U S A*, **94**, 6706-6711.
24. Cho, C.C., Chung, C. and Li, H.W. (2018) How Chi Sequence Modifies RecBCD Single-Stranded DNA Translocase Activity. *Chemphyschem*, **19**, 243-247.
25. Dabert, P., Ehrlich, S.D. and Gruss, A. (1992) Chi sequence protects against RecBCD degradation of DNA in vivo. *Proc Natl Acad Sci U S A*, **89**, 12073-12077.

26. Dohoney, K.M. and Gelles, J. (2001) Chi-sequence recognition and DNA translocation by single RecBCD helicase/nuclease molecules. *Nature*, **409**, 370-374.
27. El Karoui, M., BiauDET, V., Schbath, S. and Gruss, A. (1999) Characteristics of Chi distribution on different bacterial genomes. *Res Microbiol*, **150**, 579-587.
28. Taylor, A.F., Amundsen, S.K., Guttman, M., Lee, K.K., Luo, J., Ranish, J. and Smith, G.R. (2014) Control of RecBCD enzyme activity by DNA binding- and Chi hotspot-dependent conformational changes. *J Mol Biol*, **426**, 3479-3499.
29. Taylor, A.F. and Smith, G.R. (1999) Regulation of homologous recombination: Chi inactivates RecBCD enzyme by disassembly of the three subunits. *Genes Dev*, **13**, 890-900.
30. 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.
31. Ivancic-Bace, I., Cass, S.D., Wearne, S.J. and Bolt, E.L. (2015) Different genome stability proteins underpin primed and naive adaptation in E. coli CRISPR-Cas immunity. *Nucleic Acids Res*, **43**, 10821-10830.
32. Clarke, R.S., Bruderer, M.S., Ha, K.P. and Edwards, A.M. (2019) RexAB is essential for the mutagenic repair of Staphylococcus aureus DNA damage caused by co-trimoxazole. *Antimicrob Agents Chemother*.
33. el Karoui, M., Ehrlich, D. and Gruss, A. (1998) Identification of the lactococcal exonuclease/recombinase and its modulation by the putative Chi sequence. *Proc Natl Acad Sci U S A*, **95**, 626-631.

34. Halpern, D., Gruss, A., Claverys, J.P. and Karoui, M.E. (2004) rexAB mutants in *Streptococcus pneumoniae*. *Microbiology*, **150**, 2409-2414.
35. Quiberoni, A., Biswas, I., El Karoui, M., Rezaiki, L., Tailliez, P. and Gruss, A. (2001) In vivo evidence for two active nuclease motifs in the double-strand break repair enzyme RexAB of *Lactococcus lactis*. *J Bacteriol*, **183**, 4071-4078.
36. Mojica, F.J.M., 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.
37. 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.
38. 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.
39. 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.
40. 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.
41. 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.

42. 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.
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. Wei, Y., Terns, R.M. and Terns, M.P. (2015) Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. *Genes Dev*, **29**, 356-361.
45. Arslan, Z., Wurm, R., Brener, O., Ellinger, P., Nagel-Steger, L., Oesterhelt, F., Schmitt, L., Willbold, D., Wagner, R., Gohlke, H. *et al.* (2013) Double-strand DNA end-binding and sliding of the toroidal CRISPR-associated protein Csn2. *Nucleic Acids Res*, **41**, 6347-6359.
46. Grainy, J., Garrett, S., Graveley, B.R. and M, P.T. (2019) CRISPR repeat sequences and relative spacing specify DNA integration by *Pyrococcus furiosus* Cas1 and Cas2. *Nucleic Acids Res*, **47**, 7518-7531.
47. Kim, J.G., Garrett, S., Wei, Y., Graveley, B.R. and Terns, M.P. (2019) CRISPR DNA elements controlling site-specific spacer integration and proper repeat length by a Type II CRISPR-Cas system. *Nucleic Acids Res*, **47**, 8632-8648.
48. Nunez, J.K., Harrington, L.B., Kranzusch, P.J., Engelman, A.N. and Doudna, J.A. (2015) Foreign DNA capture during CRISPR-Cas adaptive immunity. *Nature*, **527**, 535-538.

49. 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.
50. Wright, A.V. and Doudna, J.A. (2016) Protecting genome integrity during CRISPR immune adaptation. *Nat Struct Mol Biol*, **23**, 876-883.
51. Ellinger, P., Arslan, Z., Wurm, R., Tschapek, B., MacKenzie, C., Pfeffer, K., Panjikar, S., Wagner, R., Schmitt, L., Gohlke, H. *et al.* (2012) The crystal structure of the CRISPR-associated protein Csn2 from *Streptococcus agalactiae*. *J Struct Biol*, **178**, 350-362.
52. Lee, K.H., Lee, S.G., Eun Lee, K., Jeon, H., Robinson, H. and Oh, B.H. (2012) Identification, structural, and biochemical characterization of a group of large Csn2 proteins involved in CRISPR-mediated bacterial immunity. *Proteins*, **80**, 2573-2582.
53. Nam, K.H., Kurinov, I. and Ke, A. (2011) Crystal structure of clustered regularly interspaced short palindromic repeats (CRISPR)-associated Csn2 protein revealed Ca<sup>2+</sup>-dependent double-stranded DNA binding activity. *J Biol Chem*, **286**, 30759-30768.
54. 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*, **75**, 90-101 e105.
55. Carte, J., Christopher, R.T., Smith, J.T., Olson, S., Barrangou, R., Moineau, S., Glover, C.V., 3rd, Graveley, B.R., Terns, R.M. and Terns, M.P. (2014) The three

- major types of CRISPR-Cas systems function independently in CRISPR RNA biogenesis in *Streptococcus thermophilus*. *Mol Microbiol*, **93**, 98-112.
56. Horvath, P., Romero, D.A., Coute-Monvoisin, A.C., Richards, M., Deveau, H., Moineau, S., Boyaval, P., Fremaux, C. and Barrangou, R. (2008) Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *J Bacteriol*, **190**, 1401-1412.
57. Roces, C., Wegmann, U., Campelo, A.B., Garcia, P., Rodriguez, A. and Martinez, B. (2013) Lack of the host membrane protease FtsH hinders release of the *Lactococcus lactis* bacteriophage TP712. *J Gen Virol*, **94**, 2814-2818.
58. Renye, J.A., Jr. and Somkuti, G.A. (2009) Insertion of a heterologous gene construct into a non-functional ORF of the *Streptococcus thermophilus* chromosome. *Biotechnol Lett*, **31**, 759-764.
59. Fontaine, L., Boutry, C., de Frahan, M.H., Delplace, B., Fremaux, C., Horvath, P., Boyaval, P. and Hols, P. (2010) A novel pheromone quorum-sensing system controls the development of natural competence in *Streptococcus thermophilus* and *Streptococcus salivarius*. *J Bacteriol*, **192**, 1444-1454.
60. Letort, C. and Juillard, V. (2001) Development of a minimal chemically-defined medium for the exponential growth of *Streptococcus thermophilus*. *J Appl Microbiol*, **91**, 1023-1029.
61. 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.

62. Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) WebLogo: a sequence logo generator. *Genome Res*, **14**, 1188-1190.
63. 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.
64. 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.
65. Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods*, **9**, 357-359.
66. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Genome Project Data Processing, S. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-2079.
67. Li, H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987-2993.
68. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841-842.
69. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. (2002) The human genome browser at UCSC. *Genome Res*, **12**, 996-1006.

70. McKenzie, R.E., Almendros, C., Vink, J.N.A. and Brouns, S.J.J. (2019) Using CAPTURE to detect spacer acquisition in native CRISPR arrays. *Nat Protoc*, **14**, 976-990.
71. 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.
72. Anders, C., Niewoehner, O., Duerst, A. and Jinek, M. (2014) Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature*, **513**, 569-573.
73. Klumpp, S. and Hwa, T. (2009) Traffic patrol in the transcription of ribosomal RNA. *RNA Biol*, **6**, 392-394.
74. 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. *Nat Commun*, **7**, 12853.
75. Yeeles, J.T., Gwynn, E.J., Webb, M.R. and Dillingham, M.S. (2011) The AddAB helicase-nuclease catalyses rapid and processive DNA unwinding using a single Superfamily 1A motor domain. *Nucleic Acids Res*, **39**, 2271-2285.
76. Yeeles, J.T. and Dillingham, M.S. (2007) A dual-nuclease mechanism for DNA break processing by AddAB-type helicase-nucleases. *J Mol Biol*, **371**, 66-78.
77. Amundsen, S.K., Fero, J., Salama, N.R. and Smith, G.R. (2009) Dual nuclease and helicase activities of *Helicobacter pylori* AddAB are required for DNA repair, recombination, and mouse infectivity. *J Biol Chem*, **284**, 16759-16766.



78. 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 Res*, **42**, 11144-11155.
79. 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.
80. Salinas-Rios, V., Belotserkovskii, B.P. and Hanawalt, P.C. (2011) DNA slip-outs cause RNA polymerase II arrest in vitro: potential implications for genetic instability. *Nucleic Acids Res*, **39**, 7444-7454.
81. Usdin, K., House, N.C. and Freudenreich, C.H. (2015) Repeat instability during DNA repair: Insights from model systems. *Crit Rev Biochem Mol Biol*, **50**, 142-167.
82. Jiang, F. and Doudna, J.A. (2017) CRISPR-Cas9 Structures and Mechanisms. *Annu Rev Biophys*, **46**, 505-529.
83. 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.
84. DiRuggiero, J., Achenbach, L.A., Brown, S.H., Kelly, R.M. and Robb, F.T. (1993) Regulation of ribosomal RNA transcription by growth rate of the hyperthermophilic Archaeon, *Pyrococcus furiosus*. *FEMS Microbiol Lett*, **111**, 159-164.

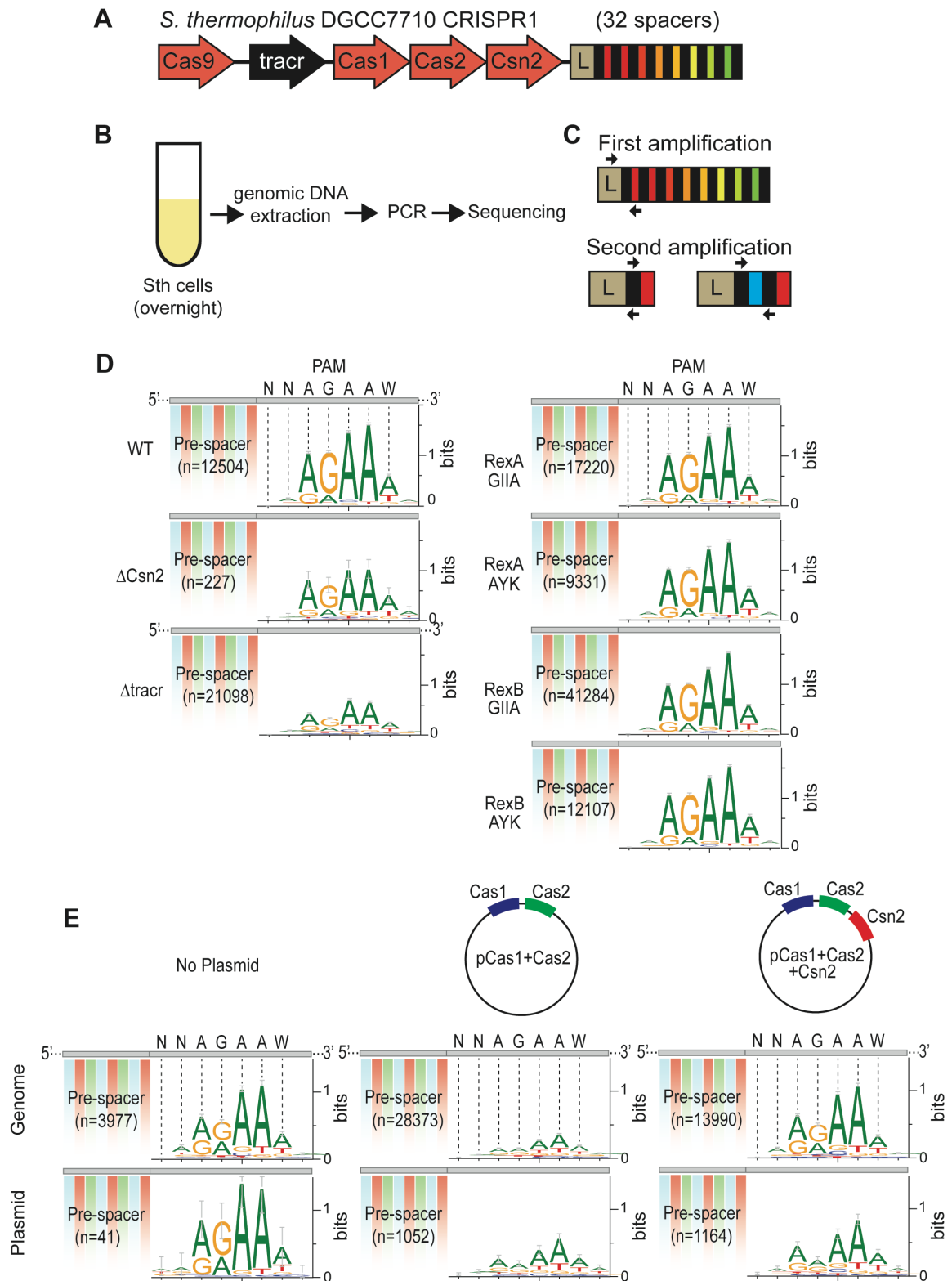
85. Shiimori, M., Garrett, S.C., Chambers, D.P., Glover, C.V.C., 3rd, Graveley, B.R. and Terns, M.P. (2017) Role of free DNA ends and protospacer adjacent motifs for CRISPR DNA uptake in *Pyrococcus furiosus*. *Nucleic Acids Res*, **45**, 11281-11294.
86. Yoon, S.H., Reiss, D.J., Bare, J.C., Tenenbaum, D., Pan, M., Slagel, J., Moritz, R.L., Lim, S., Hackett, M., Menon, A.L. *et al.* (2011) Parallel evolution of transcriptome architecture during genome reorganization. *Genome Res*, **21**, 1892-1904.
87. Sternberg, S.H., Redding, S., Jinek, M., Greene, E.C. and Doudna, J.A. (2014) DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature*, **507**, 62-67.
88. Bandyopadhyay, K., Parua, P.K., Datta, A.B. and Parrack, P. (2010) *Escherichia coli* HflK and HflC can individually inhibit the HflB (FtsH)-mediated proteolysis of lambdaCII in vitro. *Arch Biochem Biophys*, **501**, 239-243.
89. Kihara, A., Akiyama, Y. and Ito, K. (1997) Host regulation of lysogenic decision in bacteriophage lambda: transmembrane modulation of FtsH (HflB), the cII degrading protease, by HflKC (HflA). *Proc Natl Acad Sci U S A*, **94**, 5544-5549.
90. Langklotz, S., Baumann, U. and Narberhaus, F. (2012) Structure and function of the bacterial AAA protease FtsH. *Biochim Biophys Acta*, **1823**, 40-48.
91. Katz, C. and Ron, E.Z. (2008) Dual role of FtsH in regulating lipopolysaccharide biosynthesis in *Escherichia coli*. *J Bacteriol*, **190**, 7117-7122.
92. Labrie, S.J., Mosterd, C., Loignon, S., Dupuis, M.E., Desjardins, P., Rousseau, G.M., Tremblay, D.M., Romero, D.A., Horvath, P., Fremaux, C. *et al.* (2019) A

mutation in the methionine aminopeptidase gene provides phage resistance in *Streptococcus thermophilus*. *Sci Rep*, **9**, 13816.

93. Xu, J., Hendrix, R.W. and Duda, R.L. (2004) Conserved translational frameshift in dsDNA bacteriophage tail assembly genes. *Mol Cell*, **16**, 11-21.
94. Christie, G.E., Temple, L.M., Bartlett, B.A. and Goodwin, T.S. (2002) Programmed translational frameshift in the bacteriophage P2 FETUD tail gene operon. *J Bacteriol*, **184**, 6522-6531.
95. Levin, M.E., Hendrix, R.W. and Casjens, S.R. (1993) A programmed translational frameshift is required for the synthesis of a bacteriophage lambda tail assembly protein. *J Mol Biol*, **234**, 124-139.

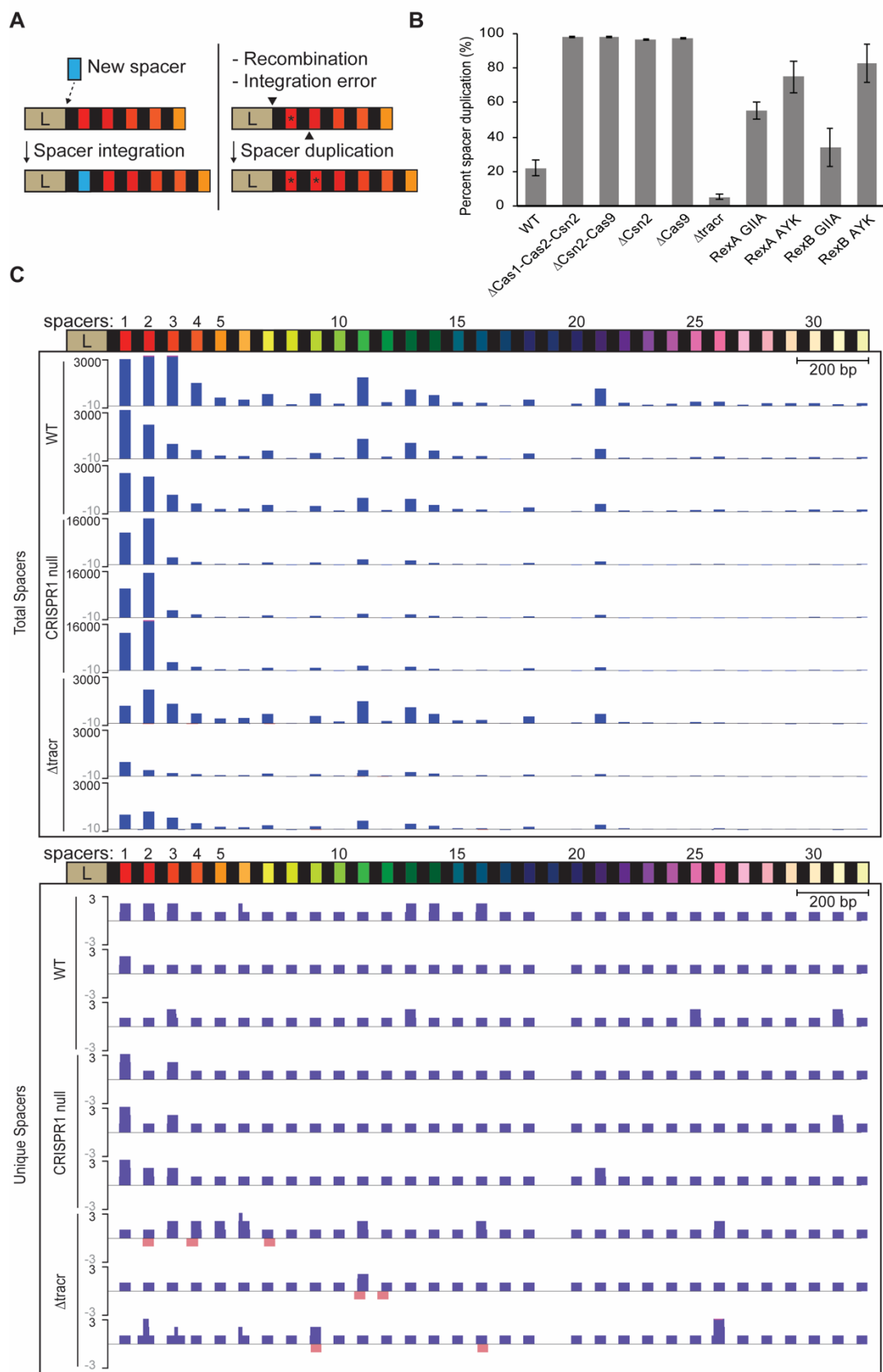
**Figure 3.1. Analysis of PAM-dependent spacer acquisition in *S. thermophilus* during adaptation.**

**(A)** Architecture of the CRISPR1 loci of *S. thermophilus* DGCC7710. Adaptation genes (orange) and *tracr* (black) are indicated. The CRISPR leader (L) is followed by alternating repeat sequences (black) and spacer (colored) units. **(B)** Graphical representation of the adaptation assay. **(C)** Illustration of the CRISPR locus and amplification of the CRISPR array for the adaptation assay. The leader to the first spacer region of the CRISPR array was amplified using primers indicated (black arrow) in the first amplification reaction. The second amplification reaction was performed using the first reaction DNA as template with repeat-specific primers to amplify expanded arrays. **(D)** Newly-acquired spacers were identified on the genome to find the consensus 5'-NNAGAAW-3' PAM sequence 8 bp downstream of the protospacer in WT and deletion strains. **(E)** PAM-analysis of Cas1-Cas2 and Cas1-Cas2-Csn2 over-expression strains.



**Figure 3.2. Spacer duplication frequency is influenced by reduced adaptation levels and tracrRNA.**

**(A)** Illustration of new spacer integration during adaptation. (left panel) New spacer sequences (blue) are integrated at the leader-proximal position of the CRISPR array. (right panel) Recombination, spacer integration errors and other unknown factors may contribute to spacer duplication events with previously existing spacers (\*) being duplicated within the CRISPR array. **(B)** Proportion of expanded arrays resulting from spacer duplication in WT and mutant strains. Pooled data from four replicates. **(C)** Distribution of total and unique duplicated spacers in the CRISPR1 array of three replicates of WT, CRISPR1 null ( $\Delta$ Cas1-Cas2-Csn2) and  $\Delta$ tracrRNA. Bars show total number of new spacers originating from each pre-existing spacer (colored blocks). Spacers on the plus and minus strand are indicated in blue and pink. Spacer duplication mapping from CRISPR1 in the WT, adaptation null ( $\Delta$ Cas1-Cas2-Csn2) are compared to the  $\Delta$ tracr strain is shown here; other mutation strains and WT show similar distributions.



**Figure 3.3. RexAB nuclease activity is important for adaptation.**

**(A)** Alignment of conserved RecB nuclease motifs present in each subunit of two-subunit exonuclease/helicase enzymes. Highly conserved motifs are enclosed in rectangles, and the black arrow over the sequence DYK and GIID corresponds to the amino acid mutation from aspartic acid to alanine generated in both RexA and RexB. Sth, *Streptococcus thermophilus*; Lla, *Lactococcus lactis*; Spnu, *Streptococcus pneumoniae*; Bsub, *Bacillus subtilis*; Ecoli, *Escherichia coli*. **(B)** Analysis of adaptation in a Cas9 deleted WT RexAB (left) and RexAB mutation strains (center) with over-expression conditions of Cas1, Cas2, Csn2 and Cas9 from a plasmid (pCas) for CRISPR1. Adaptation null ( $\Delta$ Cas1-Cas2-Csn2) strain did not contain a plasmid to represent a baseline for no adaptation. PCR products corresponding to an expanded array of a single repeat-spacer unit is indicated with a “+1”, or “+2” for two repeat-spacer units. **(C)** Analysis of adaptation in a Cas9 deleted WT RexAB (left) and RexAB mutation strains) with overexpression conditions of Cas1, Cas2, Csn2 and dCas9 from a plasmid (pCas) for CRISPR1.

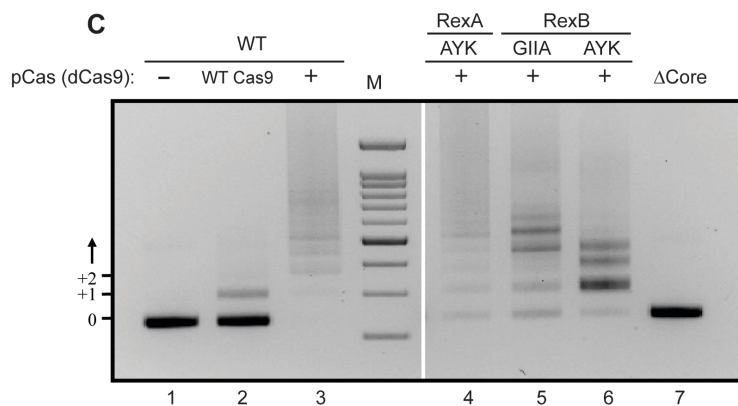
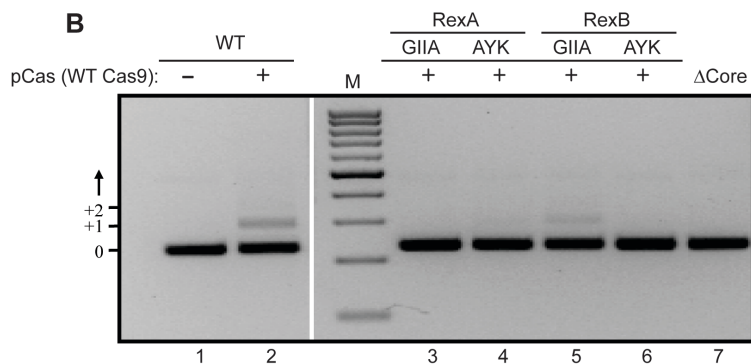


**A RecB nuclease domain**

RexA_Sth:	PASKEDFVVRGIIDGYLLLED----	RIVLFDYKTNRFTHP-----	SELKERYKGM	MSLYAKALSQA	1193
RexB_Sth:	LLANSIKITGIIDRVDRDRL-TDG--	ALGVVDYKTKGNVFDIQKF-----	YNGLSPLQ	LVTYLEALRQT	923
RexA_Lla:	EFAKEQYIVRGICDGFVKLAD----	KIILFDYKTDRTFTNV-----	SAISEIKERYK	DMNLYSEALQKA	1177
RexB_Lla:	NFSVDDIYLRGRIDRLDQL-STD--	YLGAIIDYKSSAHSFKLQEA-----	YDGLSLQF	MTYLDVIKQA	941
RexA_Spnu:	QKSQEDFVVRGILDGYLLYEN----	KIVLFDYKTDTRYDEP-----	SQLVDRYRG	QLALYEEALSRA	1193
RexB_Spnu:	LDNGRSVFVRGKVDRIDRLKANG--	AIGVVVDYKSSLTQFQFPFH-----	FNLNSQLPT	TYLAALKRE	933
AddA_Bsub:	HEADPELLVQGIIDCLYETED----	GLYLLDYKSDRIEGKFQHG--	FEGAAPILK	KRYETQIOLYTKAVEQI	1211
AddB_Bsub:	LKNGCTMELVGRIDRDVKAESSKGLLLRIVDYKSSDKGLDLAEV-----	YYGLALQMLTYLDLSITH	992		
RecB_Ecoli:	EFMQVRGMLKGFIDLVRHEG----	RYYLLDYKSNWLGEDSSAYTQQAMAAAMQAHRYDLYQYLYTLALHRY	1121		

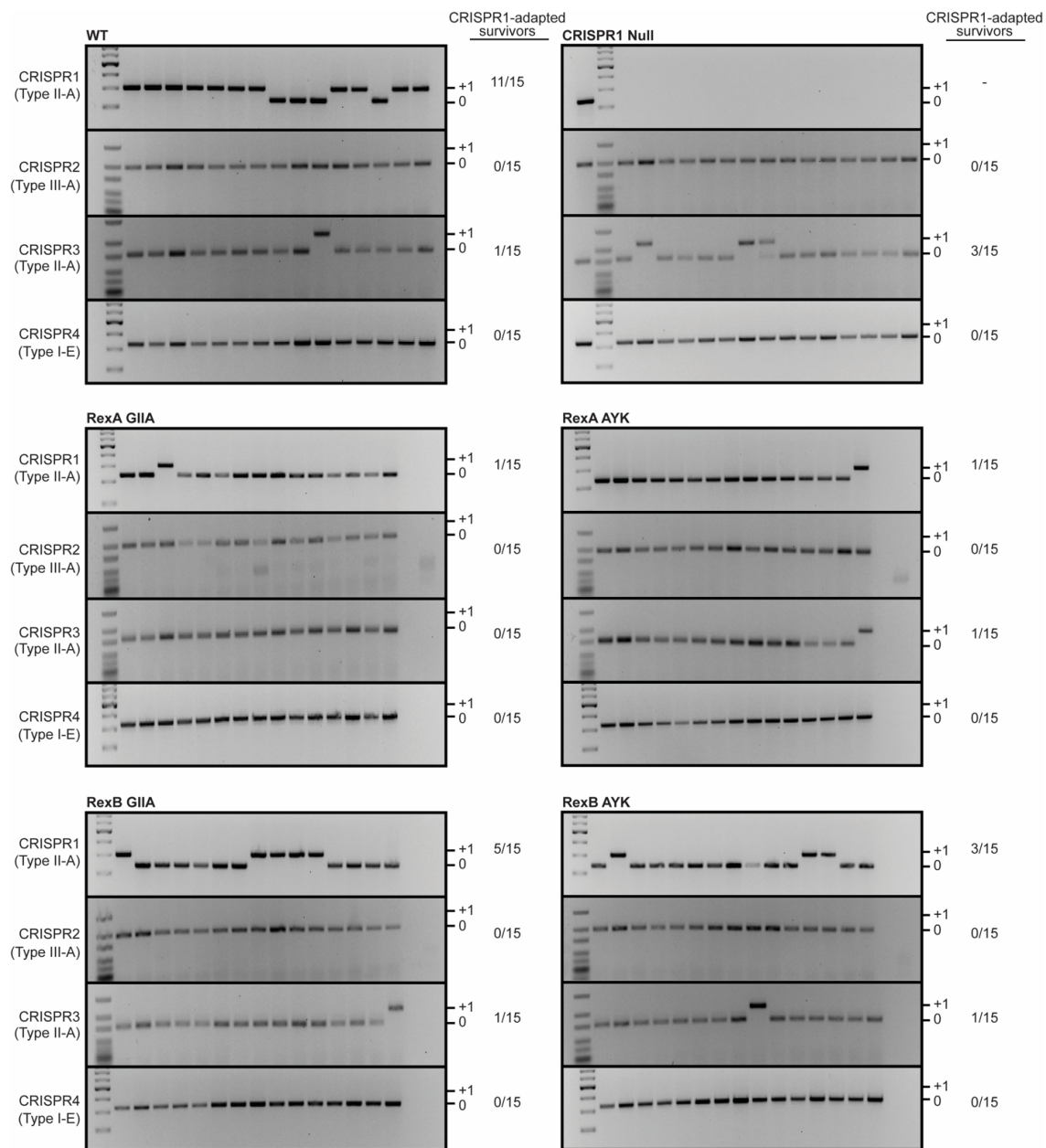
: \* \*
.\*\*\*:
\* \* \*

RecB motif-II
motif-III
QhxxY motif



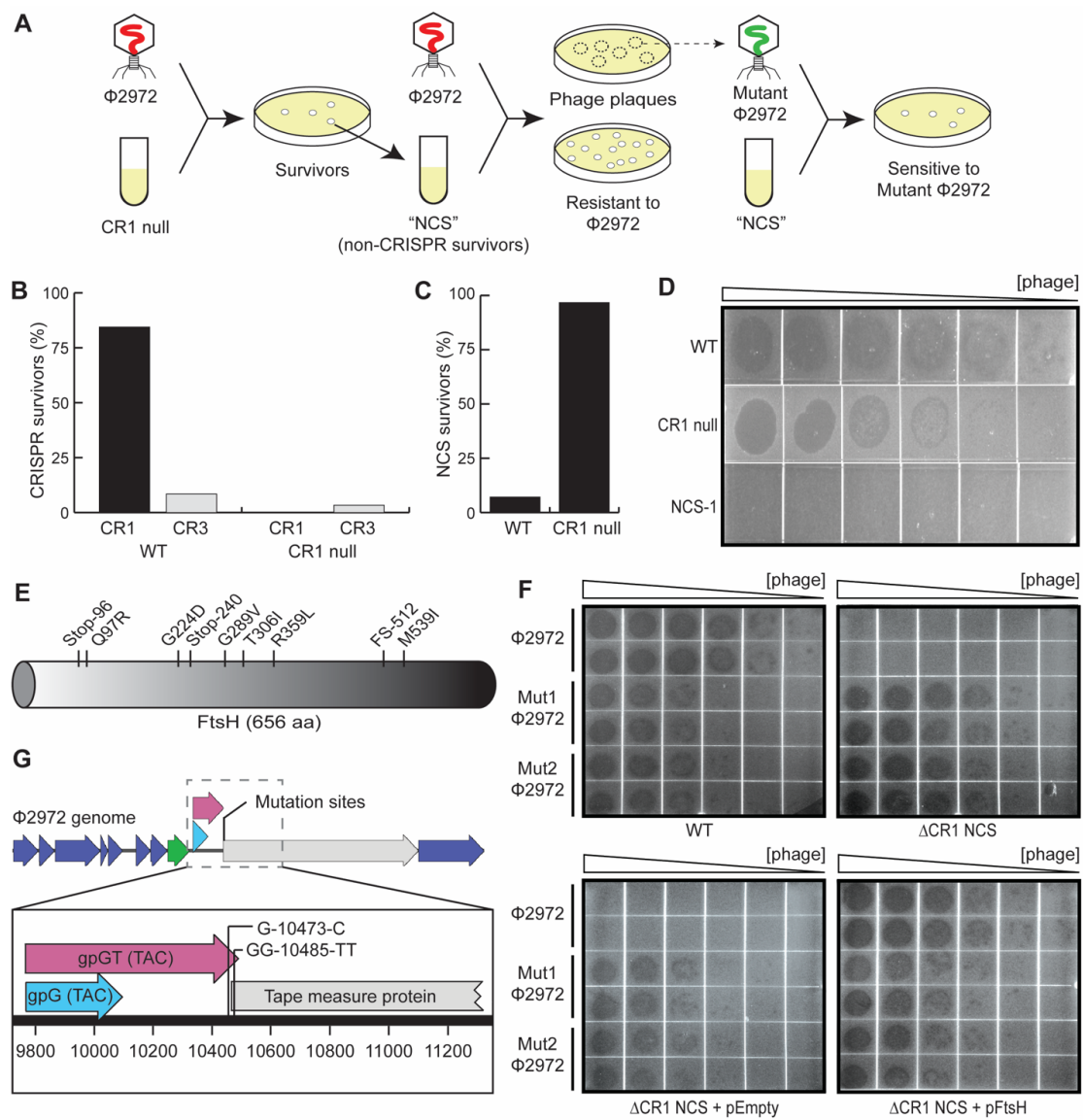
**Figure 3.4. Loss of RexAB nuclease activity reduces adaptation against phages.**

**(A)** Genotyping of phage infection survivors in WT, CRISPR1 null, and RexAB mutation strains. The leader-proximal region of CRISPR1, CRISPR2, CRISPR3 and CRISPR4 was PCR-amplified with primers indicated in Figure 3.1. Array expansion of survivors by a single repeat-spacer unit is indicated with a “+1” and unexpanded arrays in the survivors are indicated with a “0”.  $\Delta$ CRISPR1 array strain includes a control left of the marker, representing an unexpanded band to control for PCR. Total number of CRISPR1-adapted survivors corresponding on PCR amplification is presented on the right.



**Figure 3.5. Lack of FtsH contributes to phage-resistance in non-CRISPR survivors.**

**(A)** Illustration for non-CRISPR survivor mutant and phage mutant generation. Infection of a strain devoid of a CRISPR array (CR1 null) with a WT lytic phage ( $\phi 2972$ ) generates non-CRISPR survivors (NCS) that survive infection independent of CRISPR-mediated defense. Re-infection of NCS with the WT phage results in either phage-resistant survivors or phage plaques representing mutated phage (mutant  $\phi 2972$ ). **(B)** Percentage of phage-resistant survivors due to array expansion in CRISPR1 versus CRISPR3 in WT and a CRISPR1 array deletion strain. **(C)** Percentage of phage-resistant survivors that did not have expanded CRISPR arrays. **(D)** 5-fold dilutions of WT lytic phage ( $\phi 2972$ ) were spotted to lawns of WT Sth, Adaptation null ( $\Delta\text{Cas1-Cas2-Csn2}$ ) and a  $\Delta\text{CRISPR1}$  array non-CRISPR survivor (NCS-1). **(E)** Mutations in *ftsH* gene of NCS strains. Mutation positions are of 9 independent NCS strains. Frameshift mutations (FS), premature stop codons (STOP) and single amino acid mutations at each position are shown. **(F)** WT ( $\phi 2972$ ) and two mutant phages (Mut  $\phi 2972$ ) were spotted in 5-fold serial dilutions to WT Sth,  $\Delta\text{CRISPR1}$  NCS,  $\Delta\text{CRISPR1}$  NCS with an empty plasmid, and  $\Delta\text{CRISPR1}$  NCS with an FtsH complementation plasmid. **(G)** Sites of mutation from mutant phages (Mut  $\phi 2972$ ) following infection of NCS strains. Schematic representation of sites of mutation shows mutations within the tail assembly chaperonin (blue and pink arrow) and tape measure protein (grey arrow) of the  $\phi 2972$  phage genome.



**Table 3.S1. Plasmids used in this study**

Plasmid name	Plasmid description
pEmpty (pWAR)	pWAR derived from pWAR228
O/E Cas1-Cas2	pWAR + ppgm-Cas1-Cas2
O/E Cas1-Cas2-Csn2	pWAR + ppgm-Cas1-Cas2-Csn2
pCas(WT Cas9)	pWAR + ppgm-Cas1-Cas2-Csn2+PromCas9-Cas9
pCas(dCas9)	pWAR + ppgm-Cas1-Cas2-Csn2+PromCas9-dCas9
pFtsH	pWAR + ppgm-wildtype FtsH

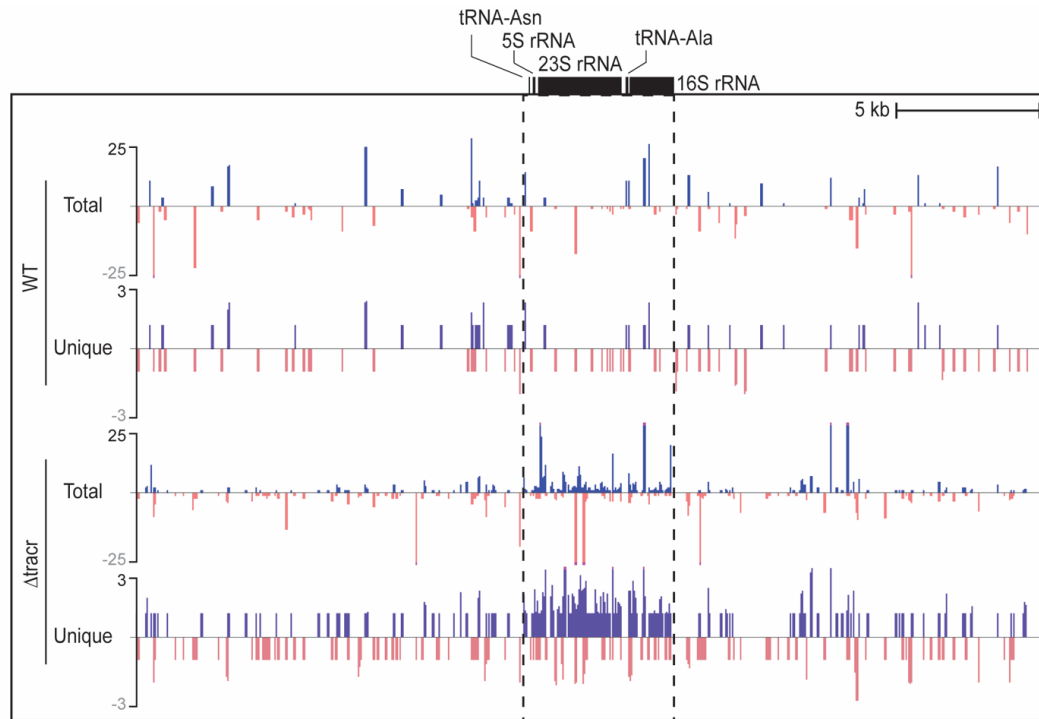
**Table 3.S2. Oligonucleotides used in this study**



Oligos	Sequences (5' – 3')
CR1 Leader-F	GCC-CTCGAG –CCTAAATCAGCTGTTTCATTTTAG
CR2 Leader-F	CATTCTCTTCTTCTAAGCCTTTATAGACC
CR3 Leader-F	TAAAATTGGAATTATTTTGAAGCTGAAGTC
CR4 Leader-F	GAAAGATGCTAGACTAATCTATC
CR1 Spacer1-R	CAAT-CTCGAG-TTCGAATCTTGATTTGCTGT
CR2 Spacer1-R	TTTCTAGGAATGGGTAATTATAGCGAGCTAGAAAGC
CR3 Spacer1-R	CCTCTTCCTCTTTAGCGTTTAG
CR4 Spacer1-R	CTATTCGCCGATAATACAGG
CC1-Capture-L	GTGGGTATAAAAACGTCAAAATTTTCATTTGAG
CC1-Capture-S1	ACAATTCGAATCTTGATTTGCTGTCAAACA
CC1-Capture-R1	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
	CTCTCAAGATTTAAGTAACTGTACAAC
CC1-Capture-R2	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
	CAGTTACTTAAATCTTGAGAGTACAAAAAC
RexA-800-F	GCCTCATGACCATTCAAGTCCAAG
RexA-800-R	CTTTACGGCTATGATTTTCTCTTGTAAGTCTAGG
RexB-up-F	CAATGAGGCGACGGATGAGC
RexB-800-R	GTCCAATCATATCAAGGATACGCTCAACC
RexA GIID-F	GTCGTCCGTGGTATCATCGA
RexA GIIA-F	GTCGTCCGTGGTATCATCGC
RexA DYK-F	GAAGACCGTATTGTCCTCTTTGA
RexA AYK-F	GAAGACCGTATTGTCCTCTTTGC
RexB GIID-F	GCATCAAGATTACTGGGATTATTGA
RexB GIIA-F	GCATCAAGATTACTGGGATTATTGC
RexB DYK-F	GATGGTGCTCTGGGTGTTGTTGA
RexB AYK-F	GATGGTGCTCTGGGTGTTGTTGC
RexB Seq-F	GCTCTGACGACCTTCTATAACAAC
FtsH-Up-f	GCAAAGCAAGCTCAGTAAAAATTGC
FtsH-Dn-r	GATTATGAATGCACTGGAAGTTCC
FtsH487-508	GCCATGAACTTTGGCCGTAATC
FtsH1081-1104	GCCATCTTGAAAGTACATGCTAAG
FtsH1387-1409	GTTCACAAAGTTACTATTGTTCC

**Figure 3.S1. Protospacers are enriched for ribosomal RNA (rRNA) in  $\Delta$ tracr strains.**

(A) Protospacers are significantly enriched around the 5S, 23S and 16S rRNA encoding regions. Plus and minus strands are indicated in blue and pink respectively. Protospacers enrichment for total and unique spacers was analyzed by aligning new acquired spacers into the CRISPR1 array in both WT and  $\Delta$ tracr. Spacer mapping from CRISPR1 in the WT compared to the  $\Delta$ tracr strain is shown here; other mutation strains and WT show similar distributions.

**A**

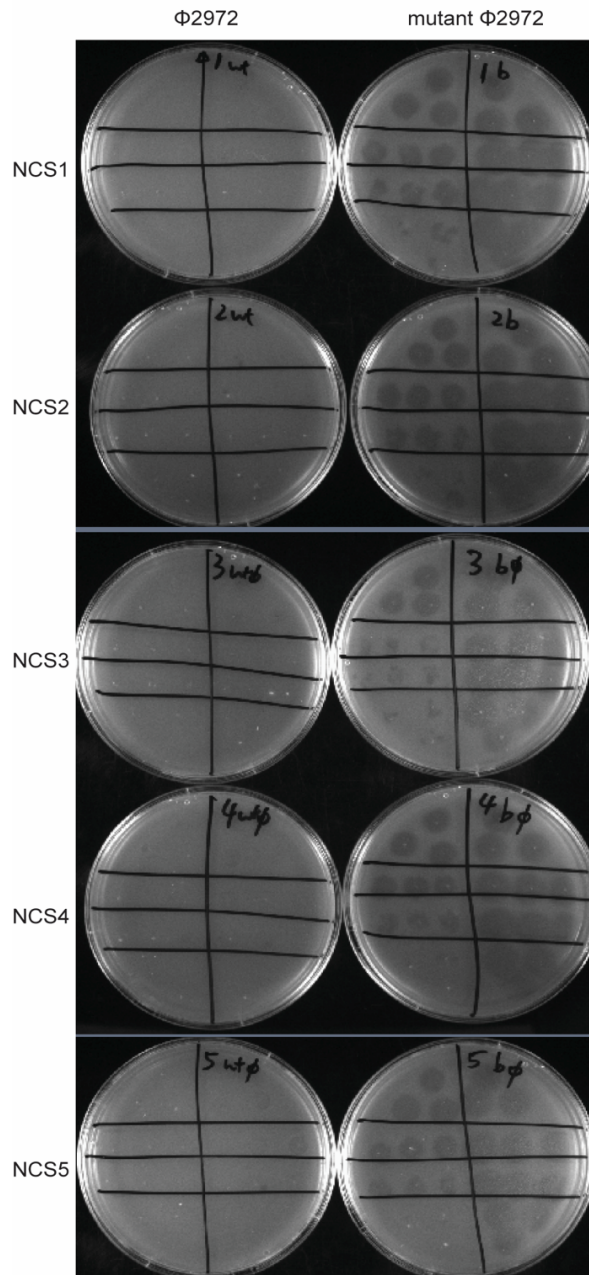
**Figure 3.S2. NCS survivor susceptibility to wild-type and mutant 2972 phages.**

**(A)** Top panel, Phage spot test using titrating concentration of phages from  $10^9$  to  $10^2$  pfu/mL. Bottom panel, wild-type 2972 phage and mutant 2972 (derived from NCS infections) were spotted in triplicates on 5 different NCS strains originating from CRISPR1 null strains at phage pfu/mL dilutions from  $10^9$  to  $10^2$  pfu/mL.

**A**

Phage spotting  
(pfu/mL)

$10^5$	$10^9$
$10^4$	$10^8$
$10^3$	$10^7$
$10^2$	$10^6$



## CHAPTER 4

### DISCUSSION

CRISPR-Cas systems are elaborate bacterial and archaeal adaptive immune systems that encompasses complex machinery to elicit an immune response against foreign genetic elements. The Type II-A CRISPR-Cas system in *Streptococcus thermophilus* is notable in that adaptive immunity against bacteriophages was first discovered in this system (1). Efforts have primarily focused on understanding the specific details governing the steps involved in CRISPR adaptation. The work presented here is a contribution to these efforts by providing information on how pre-spacers are generated and integrated into the Type II-A CRISPR-Cas system of *S. thermophilus*.

#### **Pre-spacer generation in Type II CRISPR systems by RexAB**

During CRISPR adaptation, capture of foreign DNA and incorporation of the DNA into the CRISPR array is an essential first step in CRISPR-Cas mediated immunity. However, the machinery and steps involved in the upstream processes of protospacer generation such as recognition of foreign DNA as a source of spacer sequences and PAM-dependent spacer acquisition are not well understood.

Recent studies have demonstrated the involvement of DNA repair machinery RecBCD, during CRISPR adaptation in the Type I system of Gram-positive bacteria

*Escherichia coli* (2,3). Additionally, a single study characterizing AddAB, the functional homolog of RecBCD in Gram-positive organisms, in the Type II system of *S. pyogenes* revealed similar adaptation hotspots bound by the staphylococcal Chi sequence. This bias was further identified to be a result of nuclease activity in the AddA subunit (4). Nuclease activity of the AddB was not tested in that study suggesting that AddA nuclease activity may be sufficient to reduce spacer acquisition. Here, we provide novel evidence that the RecBCD homolog, RexAB, influences adaptation in *S. thermophilus*. We demonstrate that nuclease activity of both RexAB subunits are contributing to spacer acquisition during adaptation in the context of not only self-targeting spacers from the host chromosome, but from plasmids and invading bacteriophage in the Type II system of *S. thermophilus*.

Bioinformatics analysis revealed that a set of highly conserved motifs found in RecBCD and AddAB are also found in the RexAB enzymes of *S. thermophilus* (Figure 3.3A). An *in vitro* study in *Lactococcus lactis* showed that RexAB nuclease activity is attributed to the conserved DYK motif within the nuclease domain (5). Mutational analysis of the DYK motif in both subunits of RexAB was tested here to determine the effects of both RexA and RexB nuclease activity on adaptation (Figure 3.3A).

Additionally, further sequence alignments of RecBCD homologs containing RecB-like nuclease domains revealed two additional motifs, G--D and QhXXY, all within the metal-binding pocket of these enzymes (Figure 3.3A). Furthermore, the CRISPR-Cas protein Cas4 contains the same conserved RecB-like nuclease motifs. Cas4 plays a significant role during pre-spacer generation by controlling pre-spacer processing and PAM selection *in vivo* (6). The mutation that disrupted this functionality *in vivo* was in

the less-studied G--D motif in the RecBCD nuclease domain. Structural analysis indicated that all three motifs were highly conserved and overlapped between Cas4 and the homologous AddAB enzymes, suggesting a potential role in pre-spacer processing in RexAB that is not associated with the conserved DYK motif that have been previously characterized in similar two-subunit homologs of RexAB. Therefore, we tested the DYK and GIID motifs of RexAB for effects on adaptation specifically spacer generation and PAM-recognition (Figure 3.3B, 3.3C).

Mutations of the aspartic acid in either the DYK motif (AYK) or GIID motif (GIIA) of both subunits of RexAB resulted in a reduction in adaptation frequency in the context of both host chromosome and plasmids (Figure 3.3B). This system expressed a high levels of all Type II-A CRISPR-Cas proteins (Cas1, Cas2, Csn2, Cas9) to provide increased levels of spacers acquisition that is detectable with amplification of the CRISPR array. These effects are consistent with the reduced adaptation frequency observed with RecBCD and AddAB studies in the Type I-E and II-A systems (2-4). We also mutated the GIID domain (to GIIA) to address questions of potential Cas4-like processing activity in RexAB. A similar pattern of reduced adaptation frequency was observed in a strain with RexB mutated in the GIID motif (Figure 3.3B, 3.3C). PAM-recognition of new spacer sequences, however, was not affected in both the DYK to AYK and GIID to GIIA mutation strains of RexA and RexB (Figure 3.1D). Our findings suggest that the nuclease activity of RexAB is likely contributes to adaptation not through direct PAM sequence processing, but rather through its natural role in the host as an DNA end-processing complex. This DNA end-processing activity is predicted to generate spacer substrate intermediates that are available for spacer uptake by CRISPR proteins



and this natural process contributes to this process significantly given the measurable reduction in CRISPR adaptation. To determine whether the reduction in adaptation is observed in the context of a natural invader, phage infections were carried out in the RexAB nuclease-deficient strains and spacer uptake of phage DNA was monitored (Figure 3.3C, 3.S3). In a study of lytic phage infection of *S. thermophilus* in the same RexAB mutations strains, array expansion following infection was reduced in all strains (Figure 3.3, 3.S3). This reduction coincides with the patterns observed with spacer uptake from chromosomal and plasmid DNA, indicating that RexAB degradation intermediates are likely spacer substrates for CRISPR adaptation rather than direct processing of substrates by RexAB.

### **Role of Cas proteins during spacer acquisition**

The work presented in this dissertation not only focuses on the potential role of RexAB during pre-spacer generation but attempts to identify the role of CRISPR-Cas proteins during this stage. In the Type I-E system of *E. coli*, Cas1 and Cas2 are the only two Cas proteins sufficient for spacer integration into the CRISPR array (7,8). The Type II CRISPR system of *S. thermophilus* consists of four proteins: Cas1, Cas2, Csn2, Cas9 and two co-factors of Cas9 being tracrRNA and a crRNA (9,10). All four proteins are required for adaptation to occur while CRISPR interference involves the Cas9:tracrRNA:crRNA complex (9-11). However, the specific role for each protein during spacer generation and PAM-dependent processing is not well understood.

### ***Cas1-Cas2***

In *S. thermophilus*, Cas1 and Cas2 have been characterized to function as an integrase complex *in vitro* (12-14). *In vivo* however, whether the functional role of Cas1-Cas2 that was observed *in vitro* is directly translatable is unknown. In the Type I CRISPR system of *E. coli*, Cas1 was demonstrated to be a metal-dependent nuclease that is required for spacer acquisition while Cas2 enzymatic activity was not (15). In similar Type II systems like *S. pyogenes*, whether Cas1 nuclease activity is required for adaptation is still not understood, however crystal structures revealed that metal-binding required for nuclease activity is not necessary for spacer acquisition (9,16). Although we did not characterize metal-binding or nuclease activity in this study, we demonstrate that Cas1-Cas2 is not involved in PAM recognition and both proteins are required for efficient adaptation *in vivo* (Figure 3.1D). In addition, over-expression of Cas1-Cas2 resulted in a significant increase in the number of newly acquired spacers as well as a loss in PAM-specificity (Figure 3.1E). PAM-specific DNA uptake into CRISPR arrays was restored to that of the wild-type strain upon addition of Csn2 indicating that Csn2 plays a key role in ensuring that PAM-containing protospacers are selected (Figure 3.1E).

### ***Csn2***

It has been demonstrated that in *S. thermophilus*, Csn2 is an essential player in CRISPR adaptation (10). However, aside from functioning in binding dsDNA ends, its role during adaptation is not known (17-19). Recent crystal structures have provided potential roles by demonstrating that Csn2 forms a complex with both Cas1 and Cas2

proteins in both *S. pyogenes* and in *S. thermophilus* (16,19). Additionally, Cas9 was found to weakly interact with Csn2 suggesting a functional role of Csn2 in regulating PAM-specific spacer acquisition during the integration step of adaptation (16,19). This study shows that excess Cas1-Cas2 proteins result in promiscuous spacer integration of spacers originating from any source as long as the substrate length is of the proper size (Figure 3.2). The promiscuity of Cas1-Cas2 spacer acquisition is evident in the random sampling of chromosomal DNA without a conserved PAM sequence (Figure 3.2). Conservation of the downstream 5'-NNAGAAW-3' PAM was also significantly reduced in the presence of excess Cas1-Cas2. This promiscuity in PAM selection and spacer source is significantly reduced in the presence of excess Csn2. It is likely that the non-specific nature of Cas1-Cas2 is controlled by Csn2 through direct interaction with Cas1-Cas2 leading to modulation of integration activity. We speculate that this Cas1-Cas2-Csn2 complex represents an early state of spacer capture. This provides a potential link between the two main stages of adaptation: spacer capture and spacer integration. The genetic and biochemical data presented here adds to the understanding of a higher order assembly of Cas1, Cas2, Csn2 and Cas9 (and potential specific subcomplexes of two or three components) during pre-spacer generation. However, it is also clear that further studies are required to understand the details of how the pre-spacers are processed for integration and what role these proteins play in spacer integration *in vivo*. Given our current understanding, we speculate that cellular nucleases cleave the DNA after protospacer capture by Cas proteins.

## ***Cas9***

In Type II CRISPR-Cas systems, Cas9 is required for both adaptation as well as interference (9-11). *In vivo*, Cas9 forms a complex with tracrRNA and crRNA to target foreign DNA during interference and binding of Cas9 to its cognate co-factors are required for its activity during adaptation or interference (11,20). Previous work in *S. pyogenes* demonstrated that the major role of Cas9 during adaptation, however, is the recognition of PAM sequences of acquired spacers (9). The work presented here further contributes to the understanding that the role of Cas9 is important amongst Type II CRISPR systems during adaptation particularly through defining PAM-dependent spacer acquisition. In the absence of Cas9, adaptation was severely reduced (data not shown). This suggests that the role of Cas9 during both adaptation and interference is highly conserved and likely contributes to high specificity for spacer targeting.

The experiments performed here additionally investigated the role of the tracrRNA during adaptation and PAM-dependent spacer acquisition. Similar to conformational changes induced by Cas9 bound by crRNA, tracrRNA is an essential co-factor required for the activation of Cas9 (11,20). Our results demonstrated that in the absence of tracrRNA, adaptation appears to occur with a similar frequency as WT although specificity for the proper PAM sequence was diminished. This suggests that apo-Cas9 can function in adaptation but the lack of bound tracrRNA reduces its ability of the PAM-interacting domain of Cas9 to recognize PAM-sequences (Figure 3.1D). Interestingly, protospacer hotspot regions of acquired spacers exhibited a strong preference for the rRNA gene cluster on the host genome (Figure 3.S2). Recent work demonstrated that although Cas9 nuclease activity is not required for spacer acquisition

(10). We speculate that the apo-Cas9 functions differently from the holo-Cas9 in the context of spacer sequence targeting. Highly transcribed regions have been linked to sites of preferred regions of acquired spacers suggesting that CRISPR adaptation may be a costly system if self versus non-self discrimination is not controlled for.

Lastly, mutational analysis of Cas9 and its co-factors revealed spacer duplication of pre-existing spacers in the CRISPR (Figure 3.2A). Our data demonstrated that in the absence of functional adaptation, the frequency of pre-existing spacer sequences being duplicated to the first position of the CRISPR array was nearly 100% (Figure 3.2B). The patterns of duplicated spacers revealed that not only were almost all 32 spacers in the CRISPR1 array duplicated at least one time, spacers at the leader-proximal end of the CRISPR array were duplicated more frequently than mid-array spacers. This suggested that spacer duplication may be an adaptive response when host spacer acquisition is not functioning. We cannot, however, dismiss that spacer duplication is a result of improper spacer integration following incomplete nicking and integration of a new spacer into the CRISPR array, or the result of recombination of the many closely spaced, direct repeat units in the CRISPR array. It is surprising however, that the deletion of *tracrRNA* significantly reduced the frequency of spacer duplication (Figure 3.2B). Additionally, the apo-Cas9 in the absence of *tracrRNA* affected the fidelity of duplication with a higher number of reverse oriented spacer sequences and sequence deletions (Figure 3.2C, 3.S1). It remains unclear whether the apo-Cas9 functions differently than the holo-Cas9 during spacer acquisition. Alternatively, the *tracrRNA* may function independent of its association with Cas9. Of note, a large portion of the *tracrRNA* molecule is complementary to one strand of the CRISPR DNA repeats. Perhaps novel *tracrRNA*-

repeat DNA base-paired interactions normally enhances recombination of CRISPR spacers as well as influences directionality of integration of incoming spacers.

Our results indicate that pre-spacer generation in the Type II CRISPR system depends on several factors. Based on our findings, we propose a model that involves host proteins RexAB, the functional analog of RecBCD, in *S. thermophilus*. Each subunit of the RexAB enzymes, similar to AddAB, encodes a nuclease domain capable of end-processing dsDNA (5,21-23). The *S. thermophilus* Type II CRISPR-Cas system does not contain a Cas4, so it is likely that RexAB simply functions in generating substrates for spacer acquisition. However, further testing is required to understand the mechanistic details of RexAB is contributing to spacer generation by analyzing protospacer boundaries and whether end-processing by RexAB could be utilized for protospacer trimming. Our general model then invokes that all four Type II-A CRISPR-Cas proteins: Cas1, Cas2, Csn2 and Cas9 are involved in the selection, processing and integration of spacer sequences. Pre-spacer selection is initiated by Cas9 binding at a PAM. Following pre-spacer selection, a complex of Cas1-Cas2-Csn2 binds to the DNA duplex and translocates along the DNA until the Cas9 protein is reached to initiate further processing. However, the mechanisms of how this pre-spacer is processed when all four Cas proteins come into contact with each other is still not well understood. It is possible that the CRISPR proteins are capable of pre-spacer cleavage as Cas1 has nuclease active sites and has been shown to cut DNA in vitro(24-27). Although given the already characterized function of these proteins, it is likely that host nucleases (such as RexAB) are recruited for processing of a pre-bound spacer.

## Spacer integration

Following foreign DNA capture, the spacer sequence must undergo spacer integration into the CRISPR locus. The work presented here contributes to ongoing research that identifies the functional role Cas1-Cas2 during spacer integration as well as cis-acting elements that are important for site-specific integration in *S. thermophilus*.

### *Cas1-Cas*

A goal of this dissertation was to understand how the Type II-A Cas1-Cas2 proteins facilitate spacer integration into the CRISPR locus. First, we identified that spacer integration by *S. thermophilus* CRISPR1 Cas1-Cas2 proteins were both required for sufficient integration into the CRISPR array (Figure 2.1, 2.3). This finding aligns with previous *in vitro* studies characterizing similar Type II CRISPR-Cas systems as well as other well-studied Type I systems(7,12-14,28-32). Low levels of integration were observed with Cas1 in the absence of Cas2, suggesting that Cas1 may either be acting alone *in vivo* or these integration reaction intermediates never progress to full-site integration products. However, adaptation in the Type II-A CRISPR-System in *S. thermophilus* requires all four CRISPR-Cas proteins *in vivo* (10). Therefore, it is likely that Cas1-Cas2 is minimally required for the most simplified integration reaction of a pre-processed, properly sized substrate and Csn2 and Cas9 are required in upstream steps involved in pre-spacer processing.

The Type II-A Cas1-Cas2 proteins in *S. thermophilus* were also shown to have an innate ability to recognize and prefer sequences spanning the leader-repeat junction of the CRISPR array without the requirement of accessory host proteins (Figure 2.1, 2.2). This

ability by the integrase complex to recognize specific sequences is different from Type I-E and I-F systems, which required an integration host factor (IHF) protein that bends the leader sequence within the CRISPR array to direct polarized integration to the first repeat (30,33). Type I-A CRISPR-Cas systems in archaea have additionally been demonstrated to require an unknown ATP-dependent host factor to recruit Cas1-Cas2 to the first repeat (34). It is evident that Type II Cas1-Cas2 proteins have evolved to self-recognize sequences to dictate site-specific integration without the presence of a host factor that are required in other systems.

### **Cis-acting sequences involved in spacer integration**

As CRISPR-Cas systems rapidly evolve, the diversity of these systems is ever expanding. Despite CRISPR-Cas diversity, studies have shown that there is co-evolution with the CRISPR leader sequence, the repeat, adaptation modules and the cognate PAM (35-37). These evolutionary studies further support an innate recognition ability by CRISPR-Cas proteins involved in spacer integration and cis-acting sequences in the leader and repeat sequences to dictate site-specific integration.

### ***Leader sequence***

The Type II-A leader sequence in *S. thermophilus* was minimized to an essential 10 base-pairs (bp) at the repeat-proximal end that allowed for adaptation against phages to occur *in vivo* (38). However, recognition of this sequence by Cas1-Cas2 *in vitro* was addressed in the work presented (Chapter 2).



*In vitro*, the 10 bp at the repeat-proximal end of the leader was found to be critical in recruiting Cas1-Cas2 to the first repeat. This essential region is consistent with what was observed *in vivo* and further supports the idea that recognition of the sequences spanning the leader-repeat sequence is an evolved mechanism by Cas1-Cas2 adaptation proteins to ensure proper spacer integration to the leader-adjacent repeat. Characterizing non-CRISPR integration sites on a plasmid and observing leader-repeat-like sequences flanking the site of integration further demonstrated this sequence specificity (Chapter 2, Figure 2.2).

The importance of the leader-repeat junction is further emphasized in Chapter 2 with additional *in vitro* work involving a minimal CRISPR target. Utilizing modified hair-pin target sequences and 3'dideoxy-termini modified pre-spacer substrates demonstrated that the spacer integration reaction is highly directional, with the first site of integration almost always occurring at the leader-repeat junction (Figure 2.4). Recognition of the leader and repeat sequence by the adaptation modules is an innate ability to control for site-specific integration to the specific CRISPR array. Our findings show that this recognition is highly specific and acts as an initial check point to ensure proper first-site integration.

### ***Repeat sequence***

Having found the important role of the first 10 bp of the leader sequence in directing polarized integration, sequences within the repeat were investigated in this study to identify DNA elements affecting site-specific integration at the second site (repeat-spacer junction). Proper integration at the repeat-spacer junction is critical in

ensuring that proper repeat duplication occurs after the integration of a new spacer. Improper repeat duplication has downstream implications such as generating non-functional crRNAs. The results reported in this dissertation solidifies the importance of repeat sequences in directing site-specific integration to the second site.

Previous studies have demonstrated that internal sequences within the repeat are important elements although how these sequences dictate integration is not understood. In several studies, both adaptation and integration efficiency was affected by the presence of a mutated repeat sequence *in vitro* and *in vivo* (12-14,38-42). Despite these studies, specific implication for these repeat sequences were not understood until the Type I-B and I-E systems were revealed to contain mid-repeat motifs that acted as docking sites for a molecular ruler dictating site-specific integration (39,42). The molecular ruler was thought to measure a known distance from the docking sites to determine the sites of integration at the borders of the repeat. However, a limitation to all previous *in vitro* studies is that integration efficiency does not directly translate to integration specificity. For example, a reduction in integration efficiency at one junction of the repeat can still maintain specificity. This limitation was addressed in this work as both integration efficiency and specificity were observed for each repeat mutation tested.

We identified several DNA elements within the repeat sequence that are important for integration. First, the identity of the base at the sites of transesterification attack on the CRISPR repeat is an important component for both efficient and specific integration (Figure 2.6). Second, although the role is not yet known, mid-repeat mutations influence integration efficiency of the integration reaction, though specificity is not disrupted. This suggests that the structural integrity of the repeat may influence

integration (Figure 2.5). Lastly, for the first time in Type II systems, evidence for a molecular ruler similar to the ruler observed in the Type I-E and I-B systems was demonstrated (Figure 2.7, 2.8). Evidence for the molecular ruler was additionally observed *in vivo* although with the same insertions and deletions, spacer integration at the second-site was not entirely disrupted as observed with the *in vitro* reactions (Figure 2.8). This suggests that although the molecular ruler influences spacer integration at the second site, second-site integration is not dependent on this mechanism *in vivo*. Further studies are required to understand the molecular basis of the ruler-based mechanism that guides second-site integration, likely necessitating structural and molecular analyses

### **Phage resistance by non-CRISPR survivors**

Phage-host relationships have evolved extensively over time leading to an evolutionary arms-race. Bacteria have therefore evolved to develop several strategies to evade phage infections such as CRISPR-Cas, restriction modification, and abortive infection (43). Although not a common method of phage resistance, host regulation of membrane proteins has been demonstrated to provide immunity in bacterial species (44-46). Our findings indicate that a membrane-bound metalloprotease in *S. thermophilus* is a host regulated protein that plays a major role in phage resistance during infection. In the absence of an active CRISPR1 systems, we observed similar number of phage infection survivors compared to the WT strain (Figure 3.4A). Genotyping of these survivors revealed that strains harboring functional CRISPR systems were surviving (12 out of 15) due to a newly acquired spacer from the phage into the CRISPR array. However, in the absence of a functional CRISPR system, genotyped survivors (non-CRISPR survivors,

NCS) revealed that almost no survivors (1 out of 15) adapted a spacer sequence (Figure 3.4A, 3.S3). This suggested that these survivors were resisting phage infection through another mechanism. Whole-genome sequencing revealed a common mutation in the *ftsH* gene of all non-CRISPR survivors (Figure 3.4D, 3.4D). In *E. coli*, FtsH, a membrane protein that regulates lipopolysaccharide biosynthesis, plays a role in viral infection by proteolysis of phage proteins CII, which modulates phage lysis and lysogeny (44,47,48). However, the role of similar membrane proteases in *S. thermophilus* hindering cell infection by bacteriophage 2972 has not been investigated. A previous study in *S. thermophilus* have additionally demonstrated that a host gene coding for a metalloprotease, methionine aminopeptidase (metAP) is necessary for phageDT1 infection (45). It is possible that mutation of FtsH in *S. thermophilus* results in physiological changes to the cell wall structure resulting in impaired cell lysis or failed recognition of cell surface receptors by phages.

Second-generation phages capable of infecting NCS strains revealed an adaptive mutation to infect originally resistant host strains (Figure 3.4B). Whole-genome sequencing of the mutant phages revealed common mutations in a hypothetical protein predicted to be a tail assembly chaperone (TAC) (Figure 3.4F). Although this adaptive method has not yet been observed in previous studies, we speculate that this simple adjustment may allow surface interaction of the phage to the newly adapted host. Overall, our data further illustrates the constant adaptive nature between host and invader, contributing to the understanding of phage resistance and response in an evolutionary prospective.

## Concluding remarks

CRISPR-Cas systems have diversified over time to defend against rapidly evolving invaders. The large diversity of these defense systems presents a challenge to fully determine the mechanisms involved in CRISPR-Cas immunity. Although significant work has been done to elucidate the downstream stages of CRISPR-Cas systems, the mechanisms involved in adaptation are not well understood. The work presented here provides information to further the understanding of the Type II-A CRISPR-Cas system in *S. thermophilus*. First, we demonstrate that Cas1 and Cas2 have an intrinsic ability to exhibit polarized integration in a directional manner. We also provide the first evidence supporting a molecular-based mechanism that maintains repeat length during spacer integration. Additionally, we characterize both CRISPR and non-CRISPR proteins involved in adaptation. Lastly, we provide novel details behind host-regulated mechanisms of phage resistance in the absence of CRISPR-Cas systems. In summary, this dissertation contributes novel insight into the processes involved in phage resistance by CRISPR-Cas systems in *S. thermophilus*.

## 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. 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.

3. Radovic, M., Killelea, T., Savitskaya, E., Wettstein, L., Bolt, E.L. and Ivancic-Bace, I. (2018) CRISPR-Cas adaptation in *Escherichia coli* requires RecBCD helicase but not nuclease activity, is independent of homologous recombination, and is antagonized by 5' ssDNA exonucleases. *Nucleic Acids Res*, **46**, 10173-10183.
4. 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.
5. Quiberoni, A., Biswas, I., El Karoui, M., Rezaiki, L., Tailliez, P. and Gruss, A. (2001) In vivo evidence for two active nuclease motifs in the double-strand break repair enzyme RexAB of *Lactococcus lactis*. *J Bacteriol*, **183**, 4071-4078.
6. 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.
7. 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.
8. 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.
9. 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.

10. Wei, Y., Terns, R.M. and Terns, M.P. (2015) Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. *Genes Dev*, **29**, 356-361.
11. 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.
12. Kim, J.G., Garrett, S., Wei, Y., Graveley, B.R. and Terns, M.P. (2019) CRISPR DNA elements controlling site-specific spacer integration and proper repeat length by a Type II CRISPR-Cas system. *Nucleic Acids Res*, **47**, 8632-8648.
13. Wright, A.V. and Doudna, J.A. (2016) Protecting genome integrity during CRISPR immune adaptation. *Nat Struct Mol Biol*, **23**, 876-883.
14. 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.
15. 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.
16. Ka, D., Lee, H., Jung, Y.D., Kim, K., Seok, C., Suh, N. and Bae, E. (2016) Crystal Structure of *Streptococcus pyogenes* Cas1 and Its Interaction with Csn2 in the Type II CRISPR-Cas System. *Structure*, **24**, 70-79.
17. Arslan, Z., Wurm, R., Brener, O., Ellinger, P., Nagel-Steger, L., Oesterhelt, F., Schmitt, L., Willbold, D., Wagner, R., Gohlke, H. *et al.* (2013) Double-strand DNA end-binding and sliding of the toroidal CRISPR-associated protein Csn2. *Nucleic Acids Res*, **41**, 6347-6359.

18. Nam, K.H., Kurinov, I. and Ke, A. (2011) Crystal structure of clustered regularly interspaced short palindromic repeats (CRISPR)-associated Csn2 protein revealed Ca<sup>2+</sup>-dependent double-stranded DNA binding activity. *J Biol Chem*, **286**, 30759-30768.
19. 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*, **75**, 90-101 e105.
20. Nishimasu, H., Ran, F.A., Hsu, P.D., Konermann, S., Shehata, S.I., Dohmae, N., Ishitani, R., Zhang, F. and Nureki, O. (2014) Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell*, **156**, 935-949.
21. Halpern, D., Gruss, A., Claverys, J.P. and Karoui, M.E. (2004) rexAB mutants in *Streptococcus pneumoniae*. *Microbiology*, **150**, 2409-2414.
22. Wigley, D.B. (2013) Bacterial DNA repair: recent insights into the mechanism of RecBCD, AddAB and AdnAB. *Nat Rev Microbiol*, **11**, 9-13.
23. Yeeles, J.T. and Dillingham, M.S. (2007) A dual-nuclease mechanism for DNA break processing by AddAB-type helicase-nucleases. *J Mol Biol*, **371**, 66-78.
24. Babu, M., Beloglazova, N., Flick, R., Graham, C., Skarina, T., Nocek, B., Gagarinova, A., Pogoutse, O., Brown, G., Binkowski, A. *et al.* (2011) A dual function of the CRISPR-Cas system in bacterial antiviral immunity and DNA repair. *Mol Microbiol*, **79**, 484-502.
25. Beloglazova, N., Brown, G., Zimmerman, M.D., Proudfoot, M., Makarova, K.S., Kudritska, M., Kochinyan, S., Wang, S., Chruszcz, M., Minor, W. *et al.* (2008) A



- novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. *J Biol Chem*, **283**, 20361-20371.
26. Kim, T.Y., Shin, M., Huynh Thi Yen, L. and Kim, J.S. (2013) Crystal structure of Cas1 from *Archaeoglobus fulgidus* and characterization of its nucleolytic activity. *Biochem Biophys Res Commun*, **441**, 720-725.
  27. Wiedenheft, B., Zhou, K., Jinek, M., Coyle, S.M., Ma, W. and Doudna, J.A. (2009) Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. *Structure*, **17**, 904-912.
  28. Lee, H., Zhou, Y., Taylor, D.W. and Sashital, D.G. (2018) Cas4-Dependent Pre-spacer Processing Ensures High-Fidelity Programming of CRISPR Arrays. *Mol Cell*, **70**, 48-59 e45.
  29. 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.
  30. 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.
  31. 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.
  32. 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.
33. 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.
34. 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.
35. 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.
36. Diez-Villasenor, C., Guzman, N.M., Almendros, C., Garcia-Martinez, J. and Mojica, F.J. (2013) CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of *Escherichia coli*. *RNA Biol*, **10**, 792-802.
37. 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.
38. 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.

39. 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.
40. Grainy, J., Garrett, S., Graveley, B.R. and M, P.T. (2019) CRISPR repeat sequences and relative spacing specify DNA integration by *Pyrococcus furiosus* Cas1 and Cas2. *Nucleic Acids Res*, **47**, 7518-7531.
41. 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.
42. 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.
43. Stern, A. and Sorek, R. (2011) The phage-host arms race: shaping the evolution of microbes. *Bioessays*, **33**, 43-51.
44. Kihara, A., Akiyama, Y. and Ito, K. (1997) Host regulation of lysogenic decision in bacteriophage lambda: transmembrane modulation of FtsH (HflB), the cII degrading protease, by HflKC (HflA). *Proc Natl Acad Sci U S A*, **94**, 5544-5549.
45. Labrie, S.J., Mosterd, C., Loignon, S., Dupuis, M.E., Desjardins, P., Rousseau, G.M., Tremblay, D.M., Romero, D.A., Horvath, P., Fremaux, C. *et al.* (2019) A mutation in the methionine aminopeptidase gene provides phage resistance in *Streptococcus thermophilus*. *Sci Rep*, **9**, 13816.
46. Shotland, Y., Koby, S., Teff, D., Mansur, N., Oren, D.A., Tatematsu, K., Tomoyasu, T., Kessel, M., Bukau, B., Ogura, T. *et al.* (1997) Proteolysis of the

phage lambda CII regulatory protein by FtsH (HflB) of Escherichia coli. *Mol Microbiol*, **24**, 1303-1310.

47. Bandyopadhyay, K., Parua, P.K., Datta, A.B. and Parrack, P. (2010) Escherichia coli HflK and HflC can individually inhibit the HflB (FtsH)-mediated proteolysis of lambdaCII in vitro. *Arch Biochem Biophys*, **501**, 239-243.
48. Langklotz, S., Baumann, U. and Narberhaus, F. (2012) Structure and function of the bacterial AAA protease FtsH. *Biochim Biophys Acta*, **1823**, 40-48.