UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title

Structures and Mechanisms of CRISPR-Cas Adaptive Immunity

Permalink

https://escholarship.org/uc/item/1h547661

Author

Wang, Joy Y.

Publication Date

2022

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Structures and Mechanisms of CRISPR-Cas Adaptive Immunity

By

Joy Y. Wang

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Chemistry

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Jennifer A. Doudna
Professor Christopher J. Chang
Professor David F. Savage

Spring 2022

Abstract

Structures and Mechanisms of CRISPR-Cas Adaptive Immunity

by

Joy Y. Wang

Doctor of Philosophy in Chemistry

University of California, Berkeley

Professor Jennifer A. Doudna, Chair

In the ongoing arms race between prokaryotes and phage, the emergence of CRISPR-Cas (clustered regularly interspaced short palindromic repeats - CRISPR-associated) adaptive immunity provided bacteria and archaea a unique basis of defense against genetic invaders. CRISPR systems record immunological memory of previous infection in the host CRISPR genomic locus, consisting of alternating short repeat sequences and foreign-derived spacers. During CRISPR adaptation, fragments of foreign genetic material (prespacers) are captured and processed into short sequences (protospacers) that are integrated into the CRISPR array. After transcription of the CRISPR array and maturation of CRISPR RNAs (crRNAs), CRISPR interference proteins will associate with individual crRNAs to carry out target identification and destruction. This work describes the origins and evolution of the CRISPR adaptation module and the diversity of enzymes and mechanisms used in spacer acquisition.

The core component of the CRISPR adaptation module is the CRISPR integrase, which in most systems, consists of the Cas1 and Cas2 proteins and performs the integration of new spacers. We identify a functional CRISPR mini-integrase comprising of only the type V-C Cas1 that may represent the ancestral CRISPR integrase prior to Cas2 adoption. We determine that the mini-integrase forms a tetramer for integration and has selectivity for short 17-19 bp substrates. Based on the evolutionary model, the mini-integrase may mark the emergence of a precise ruler mechanism dictating spacer length, a hallmark of CRISPR integrases.

In certain CRISPR systems, a reverse transcriptase (RT) is another key component of the adaptation module and enables spacer acquisition from RNA sources. Using cryo-electron microscopy, we uncover the structure of a naturally occurring fusion

complex that combines an RT, a Cas1—Cas2 integrase, and Cas6 maturase and through structure-guided mutagenesis, we reveal crosstalk between the three active sites. Based on a unique RT conformation and other domain interactions, we suggest structural rearrangements that may coordinate sequential enzymatic activities enabling spacer acquisition from RNA and DNA substrates.

A third class of enzymes involved in CRISPR adaptation are the enzymes that aid in the processing of prespacers into viable integration substrates. In many systems, this involves the removal of the protospacer adjacent motif (PAM), which serves to distinguish between self and non-self. Using a natural DEDDh exonuclease fusion to a Cas1—Cas2 integrase, we elucidate the stepwise coordination of prespacer processing and integration and reveal a gate-keeping ruler mechanism providing exquisite precision in PAM removal. These results underscore the diverse structures and mechanisms employed in the generation of CRISPR immunological memory.

Acknowledgements

The journey to this thesis has been a transformative experience and it would not have been possible without the guidance and contributions from scientific mentors and collaborators and the support from friends and family. To everyone who has helped me along the way, thank you and I am eternally grateful.

I am thankful for all the people who have served as mentors during this time. I would like to especially thank Jennifer Doudna for giving me the invaluable opportunity to learn and grow as a scientist in your lab and the freedom to explore my scientific interests and take on new challenges. The environment that you've established in your lab has produced an invigorating and collaborative space for scientific research. Here, I need to thank Addison Wright, my rotation mentor, who took me under his wing when I was a fresh graduate student who was new to biochemistry research. To Addison, thank you for the crash-course on biochem basics and more importantly, for introducing me to CRISPR adaptation and showing me how fun the puzzle-solving nature of mechanistic biochemistry can be. In my first few years in lab, I've also learned a great deal from Gavin Knott, Audrey Lapinaite, and Junjie Liu who are not only examples of stellar scientists but also wonderful teachers. To Marco Lobba and Josh Cofsky, the more senior grad students, you've always been so welcoming, willing to help, and incredibly generous with your time. For all of your guidance, I am deeply grateful.

During my time here, I've also gotten a chance to work with many wonderful collaborators. I want to especially thank Christopher Hoel: our serendipitous run-in at the cryoEM facility led to a collaboration that was one of the highlights of my PhD. It was truly a pleasure working with you and Stephen Brohawn and I am immensely appreciative of our collaboration. I would also like to thank Anthony Iavarone, Basem Al-Shayeb, Petr Skopintsev, and Patrick Pausch for working with me on various projects and providing important insights. To my rotation students, Gary Li and Owen Tuck, and undergraduate students, Charles Alza, Mohak Jain, and Julia Zhou, thank you for all your hard work and contributions and I look forward to hearing about your future successes.

I also need to thank others in the Doudna lab for making it such a supportive lab environment. To Kaihong Zhou and Jenny Ye, you've been instrumental to making my research go as smoothly as possible - thank you both for keeping the lab so well managed and organized. Kaihong, thank you for all the great conversations and advice and for taking care of us in lab. To Brady Cress, Ben Rubin, Tina Liu, Kasia Soczek, Carolyn Huang, Liz O'Brien, Amy Eggers, Brittney Thornton, Enbo Ma, Dylan Smock, Emily Meier, David Colognori, Connor Tsuchida, Jenny Hamilton, Kai Chen, Honglue Shi, Haridh Shivram, and Kyle Watters, thank you for all of the wonderful scientific discussions, technical advice, and insightful perspectives that you've offered. To my bay mates in Bayoncé, Connor Tsuchida, Jenny Hamilton, Kasia Soczek, Audrey Lapinaite,

Cindy Sandoval Espinoza, and Riley McGarrigle, you've all made our bay such a friendly and supportive environment, and I've thoroughly enjoyed working next to you. Thanks for all the great conversations and fun Bayoncé outings.

To dear friends both in and outside of lab, I want to give special thanks. Carolyn, thank you for swinging by my bay all the time to say hi - your visits always brighten my day. Also, thank you for constantly feeding me - your culinary skills are out of this world and I am always grateful whenever you share your cooking / baked goods with me. To Connor, David, and Josh, thanks for all the hangouts at Victory Point, movie nights, and other fun shenanigans that have made these past years so enjoyable. To Connor, who has been my bay mate, late night lab companion, and tennis hitting partner since the very beginning, and Liz, who is a fellow stuffed animal lover and still a great friend even after leaving lab, you've both been constant sources of support and I will forever treasure your friendship. You've helped me celebrate the good times and been there for me during the bad, and my grad school experience has been that much better because of you. To Vasil Vasilev, my dear friend and chem lab partner from undergrad, I'm so glad we ended up going to Berkeley together and you were a friendly face when I didn't know anyone else here. I'd also like to thank my friends since childhood, Shweta Sen, Tori Seto, Ana Veskovic, and Whitney Zhou. Thank you so much for keeping in touch all of these years despite living in different parts of the country - I've always felt your support from wherever you are.

Finally, thank you to my parents, Yadong Wang and Xiaoyi He, and my partner, Lewis Hom, for all of your endless love and support - I am so, so lucky to have you in my life. To my parents, your unconditional love has always been one of the greatest sources of strength for me. To my mom, thank you for being my constant companion - despite the physical distance between us, you're always just a phone call away. To my dad, thank you for always rooting for me - you've been instrumental in guiding me down this path from the very beginning, and I would not be here today without your help. To Lewis, you have been the greatest gift. Thank you for bearing with me all these years doing everything you can to make my grad school life as comfortable as possible, providing an endless source of warmth and encouragement, helping me with countless errands, and making me gourmet meal preps. The daily routines of life are immeasurably more enjoyable with you by my side. I can't wait to start this next chapter of our lives together!

TABLE OF CONTENTS

Abstract ……1
Acknowledgements ……i
Table of Contents ……iii
List of Figures ……vii
List of Tables……viii

Chapter 1. Evolution and diversity of CRISPR adaptation ……1
1.1 Abstract ……2
1.2 Introduction ……2
1.3 Origins and mechanisms of CRISPR spacer capture ……4
1.3.1 Origins and evolution of the CRISPR integrase ……4
1.3.2 Recruitment of reverse transcriptases for RNA spacer acquisition ….7
1.3.3 Mechanisms of CRISPR protospacer integration ……8
1.3.4 CRISPR integration substrate biogenesis, selection, and Processing ……9
1.4 Conclusion and outlook ……13
1.5 Acknowledgements ……14
1.6 Contributions ……14

Chapter 2. A functional mini-integrase in a two-protein type V-C CRISPR System ……15
2.1 Abstract ……16
2.2 Introduction ……16
2.3 Results ……17
2.3.1 Type V-C and V-D systems have CRISPR arrays with short repeats and spacers and lack cas2 genes ……17
2.3.2 Type V-C Cas1 catalyzes DNA integration without Cas2 ……18
2.3.3 Promiscuous DNA integration by V-C Cas1 ……22
2.3.4 Type V-C Cas1 catalyzes full-site integration ……25
2.3.5 The Cas1 mini-integrase forms a tetramer for integration ……28
2.4 Discussion ……34
2.5 Materials and methods ……35
2.5.1 Cloning / protein purification ……35
2.5.2 Integration assays ……36
2.5.3 High-throughput sequencing ……40
2.5.4 Full-site selection ……40
2.5.5 protein complex formation / light scattering / native nanoelectrospray ionization mass spectrometry ……41

2.6 Accession codes ……42
2.7 Acknowledgements ……42
2.8 Author contributions ……42

Chapter 3. Structural coordination between active sites of a CRISPR reverse transcriptase-integrase complex ……43
3.1 Abstract ……44
3.2 Introduction ……44
3.3 Results……45
3.3.1 Architecture of the Cas6-RT-Cas1—Cas2 complex……45
3.3.2 Substrate preferences of Cas6-RT-Cas1—Cas2 for cleavage-ligation……52
3.3.3 Cas6-RT-Cas1—Cas2 exhibits length selectivity for ssDNA and dsDNA substrates ……57
3.3.4 Cas6-RT-Cas1—Cas2 catalyzes full-site integration of dsDNA substrates ……58
3.3.5 Unique RT conformation results in close contact with Cas1 active site……60
3.3.6 Structural elements contribute to crosstalk between Cas6, RT, and Cas1 active sites ……63
3.4 Discussion ……67
3.5 Materials and methods ……70
3.5.1 Cas6-RT-Cas1 protein identification……70
3.5.2 Residue conservation analysis……70
3.5.3 Protein purification……70
3.5.4 DNA and RNA substrate preparation ……71
3.5.5 Integration assays ……75
3.5.6 RT and Cas6 activity assays ……76
3.5.7 Electrophoretic mobility shift assays ……77
3.5.8 Grid preparation ……77
3.5.9 Cryo-EM data acquisition ……78
3.5.10 Cryo-EM data processing ……79
3.5.11 Modeling, refinement, and analysis ……80
3.6 Accession codes ……81
3.7 Acknowledgements ……81
3.8 Author contributions ……82

Chapter 4. Molecular mechanisms of prespacer processing for CRISPR adaptation ……83
4.1 Abstract ……84

4.2 Introduction ……..……84
4.3 Results……..……86
4.3.1 Cas1—Cas2-DEDDh processes prespacers to correct size for integration ……86
4.3.2 Delayed PAM removal biases orientation of integration……89
4.3.3 Cas1—Cas2-DEDDh uses gate-keeping ruler mechanism for PAM processing……93
4.3.4 PAM motif in prespacer results in lower integration efficiency *in vitro* ……..…….95
4.3.5 Substrate preferences for selective processing by Cas1—Cas2-DEDDh ……97
4.4 Discussion ……..101
4.5 Materials and methods ……..105
4.5.1 Protein purification ……..……105
4.5.2 DNA substrate preparation ……..……..106
4.5.3 Processing assays ……..……108
4.5.4 Ligation assays with pCRISPR integration target plasmid ……..…109
4.5.5 Full-site integration assays ……..……109
4.5.6 CRISPR locus bioinformatic analysis……..110
4.6 Acknowledgements ……..……110
4.7 Author contributions ……..……110

Bibliography ……..……..…111

LIST OF FIGURES

Figure 1.1 CRISPR systems provide bacteria and archaea adaptive immunity……..3
Figure 1.2 Modular organization of CRISPR-Cas systems……...4
Figure 1.3 Evolution and diversity of the CRISPR integrase architecture ……7
Figure 1.4 Protospacer integration mechanism ……..9
Figure 1.5 Mechanisms of CRISPR prespacer biogenesis and orientation ……12

Figure 2.1 Type V-C and V-D CRISPR systems have short repeats and spacers and lack cas2 genes ……18
Figure 2.2 Type V-C Cas1 integrates DNA fragments of expected length in vitro …..19
Figure 2.3 Determination of type V-C Cas1 protospacer preferences ..……21
Figure 2.4 Integration of protospacers occurs at many off-target sites…….23
Figure 2.5 Integration sites along pUC19 ……24
Figure 2.6 Integration of Protospacer into CRISPR Repeat and lacZ Hotspot ..……25
Figure 2.7 Detection of full-site integration by type V-C Cas1 ……27
Figure 2.8 Protospacer sequence influences orientation during full-site integration ..28
Figure 2.9 Type V-C Cas1 forms a multimeric complex for integration……29
Figure 2.10 Dynamic light scattering size measurements of apo Cas1 compared to Cas1 with added pseudo-full-site substrate ...……30
Figure 2.11 Detection of type V-C Cas1 tetramer by native mass spectrometry ……..31
Figure 2.12 Native mass spectra of apo V-C Cas1 and V-C Cas1 complexed with pseudo-full-site substrate in native state ……32
Figure 2.13 Model for type V-C mini-integrase ……..33

Figure 3.1 A natural Cas6-RT-Cas1 fusion protein within a type III CRISPR-Cas system ……..…46
Figure 3.2 Complex formation ……47
Figure 3.3 Cryo-EM processing pipeline……..……...48
Figure 3.4 Cryo-EM validation ……49
Figure 3.5 Architecture of Cas1, RT, and Cas6 domains……...51
Figure 3.6 Cas6-RT-Cas1—Cas2 catalyzes cleavage-ligation reactions with dsDNA, ssDNA, and ssRNA substrates in vitro ……53
Figure 3.7 Cas6-RT-Cas1—Cas2 prefers DNA and DNA/RNA hybrid substrates over ssRNA substrates for cleavage-ligation ……54
Figure 3.8 Cas6-RT-Cas1—Cas2 catalyzes ligation of DNA and RNA protospacers into the CRISPR array ……..56
Figure 3.9 Cas6-RT-Cas1—Cas2 shows length selectivity for dsDNA and ssDNA substrates for integration ……58
Figure 3.10 Cas6-RT-Cas1—Cas2 conducts site-specific full-site integration …….59

Figure 3.11 Cas6-RT-Cas1 catalyzes cDNA synthesis with both RNA and DNA templates ……...61
Figure 3.12 Cas6-RT-Cas1 catalyzes cDNA synthesis with minimal primer-template homology ……..……62
Figure 3.13 Crosstalk between RT and Cas1 and Cas6 active sites ……64
Figure 3.14 Ligation and crRNA processing by Cas6-RT-Cas1—Cas2……66
Figure 3.15 Model for RT activation……68

Figure 4.1 Cas1—Cas2-DEDDh trims duplex prespacers to integration-competent size ……..….87
Figure 4.2 Kinetics of exonuclease trimming and ligation by Cas1—Cas2-DEDDh *in vitro* ……..……....89
Figure 4.3 TT PAM motif is initially protected from trimming ……90
Figure 4.4 PAM motif is processed for full-site integration……92
Figure 4.5 DEDDh is responsible for PAM processing…….…..95
Figure 4.6 Presence of PAM motif lowers integration efficiency from prespacer *in vitro*……..……...…97
Figure 4.7 Cas1—Cas2-DEDDh selectively performs PAM-dependent trimming on certain ssDNA substrates but not dsDNA ……..…...99
Figure 4.8 Cas1—Cas2-DEDDh exhibits strict requirement for pre-duplexed 23-bp DNA duplex prespacer for functional ruler-guided trimming …..……...100

LIST OF TABLES

Table 2.1 IMG genome sequences of type V-C Cas1 from mouse cecum and V-D Cas1 from beetle gut ……...36
Table 2.2 DNA oligonucleotides used in this study……38

Table 3.1 Sequences of type III *Thiomicrospira* Cas6-RT-Cas1 and Cas2 proteins ……..71
Table 3.2 DNA and RNA oligonucleotides used in this study ……..73
Table 3.3 Cryo-EM data collection, processing, and refinement……79

Table 4.1 Sequences of the type I-E *Megasphaera NM10_related* Cas1 and Cas2-DEDDh proteins used in this study……..……..106
Table 4.2 DNA oligonucleotides used in this study ..……107

CHAPTER 1

Evolution and diversity of CRISPR adaptation

A portion of the work presented in this chapter is part of the following manuscript currently accepted at *Nat. Rev. Microbiol.*: Wang, J.Y., Pausch, P. & Doudna, J.A. Structural biology of CRISPR immunity and genome editing enzymes. *Nat. Rev. Microbiol.* (2022).

1.1 Abstract

In this review, we examine recent advances in understanding the evolution and diversity of molecular structures and mechanisms of enzymes involved in CRISPR adaptation. These structural insights reflect close evolutionary links between CRISPR systems and mobile genetic elements, including the origins and evolution of CRISPR systems from DNA transposons, retrotransposons, and toxin-antitoxin modules. We discuss how the evolution and diversity of CRISPR adaptation proteins explains their functional complexity in achieving robust immunity against foreign invaders while avoiding autoimmunity.

1.2 Introduction

Clustered regularly interspaced short palindromic repeats (CRISPRs) and CRISPR-associated (Cas) proteins provide microbes with RNA-guided adaptive immunity[1]. These systems typically consist of the CRISPR array, an adaptation module, and a crRNA biogenesis and DNA/RNA-interference module (Fig. 1.1, 1.2) (reviewed in refs[2,3]). To provide adaptive and heritable immunity, the CRISPR array stores the genetic information of parasitic elements as 'spacer' sequences (typically ~25-50 base-pair (bp) in size, though can range from ~17-72 bp[4,5]) inserted between short palindromic repeats (reviewed in ref[6]). The Cas1—Cas2 adaptation machinery captures a segment of viral or plasmid DNA, the protospacer, in a bacterial cell and integrates it into the CRISPR array (Fig. 1.1). In DNA-targeting CRISPR systems, protospacer selection depends on the presence of a 3-5 bp long protospacer adjacent motif (PAM) that is not integrated into the CRISPR array and serves to distinguish self from non-self target sequences (reviewed in ref.[7]).

Once the CRISPR array is transcribed, dedicated pre-crRNA processing Cas-ribonucleases, or trans-activating (tra)crRNA and 'host factor' RNAses, associate with and cleave the pre-crRNA repeat sequence to release individual crRNAs (Fig. 1.1) (reviewed in ref.[8]). Depending on the CRISPR class, several Cas proteins (Class 1), or a single Cas protein (Class 2), recruit the mature crRNA guide for DNA or RNA interference (Fig. 1.1). To prevent self-targeting of the complementary spacer within the CRISPR array, DNA targeting systems recognize a PAM motif before interrogation of double-stranded DNA. Hybridization of the crRNA spacer sequence to the unwound target nucleic acid results in a conformational shift in the interference complex, activating the Cas-nuclease for DNA or RNA cutting and invader destruction. Currently, the two CRISPR classes are subdivided into six types and >30 subtypes (reviewed in ref.[9]), differentiated by their interference module composition and nucleic acid target specificity (Fig. 1.2).

Advances in both biochemical and imaging methods have accelerated structural studies of CRISPR-Cas proteins and complexes. We focus here on the process by which CRISPR systems provide adaptive immunity in microbes by capturing and integrating foreign DNA into CRISPR sequence arrays. We start with the origins of and evolution of the CRISPR integrase. We then turn to the diversity and roles of other CRISPR adaptation proteins, focusing on the recruitment of reverse transcriptases for RNA spacer acquisition. Finally, we explore the mechanisms that govern spacer

acquisition and the biogenesis and processing of integration substrates in order to achieve robust immunity against foreign invaders while avoiding autoimmunity.

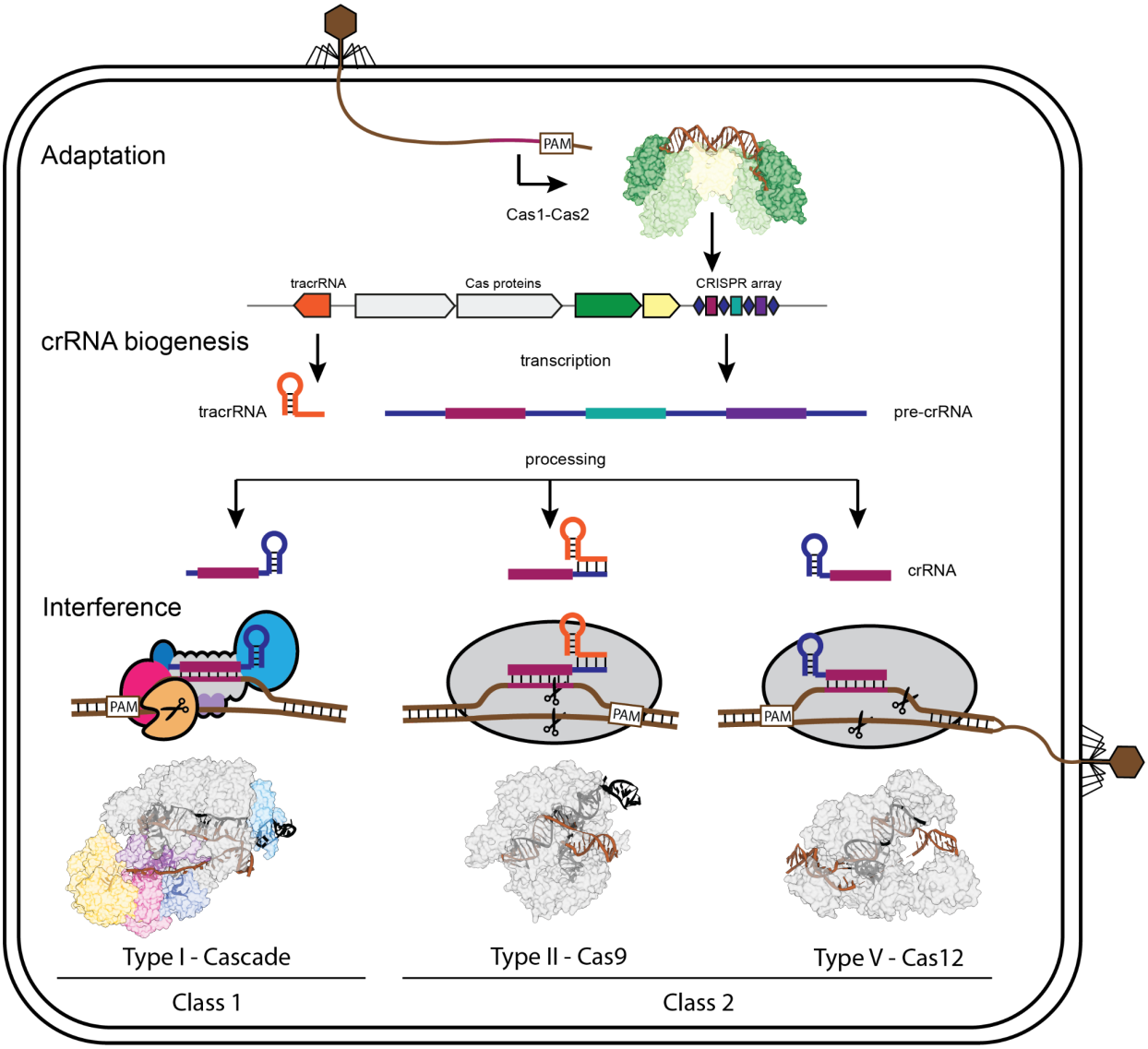


Figure 1.1 | CRISPR systems provide bacteria and archaea adaptive immunity. Schematic of the three stages of CRISPR immunity: adaptation, crRNA biogenesis, and interference. In adaptation, Cas1—Cas2 (PDB-ID 5DS4)[10] inserts protospacers, derived from foreign genetic elements, into the CRISPR array as new spacers (represented as different colored rectangles) that are separated by CRISPR repeats (represented as blue diamonds). During crRNA biogenesis, the CRISPR array is transcribed into pre-crRNA, which is processed into mature crRNAs that each have a single spacer. The crRNA (or in some cases, the dual crRNA-tracrRNA) assembles with the effector protein/complex to form a surveillance complex that recognizes and degrades foreign genetic elements complementary to the crRNA spacer during interference. Class 1 systems have multi-subunit effector complexes, while Class 2 systems have single-subunit effector proteins. Target cleavage by the Class I type I Cascade-Cas3 effector complex (left), the Class 2 type II Cas9 effector (center), and the Class 2 type V Cas12a effector (right) is depicted schematically, and representative structures of the effector complexes are shown beneath: Cascade-Cas3 bound to crRNA and target DNA (PDB-ID 6C66)[11], Cas9 bound to guide RNA

(gRNA; composed of crRNA and tracrRNA) and target DNA (PDB-ID 4UN3)[12], Cas12a bound to crRNA and target DNA (PDB-ID 5NFV)[13]. Nucleic acids in the structures are color-coded: DNA, brown; RNA, black.

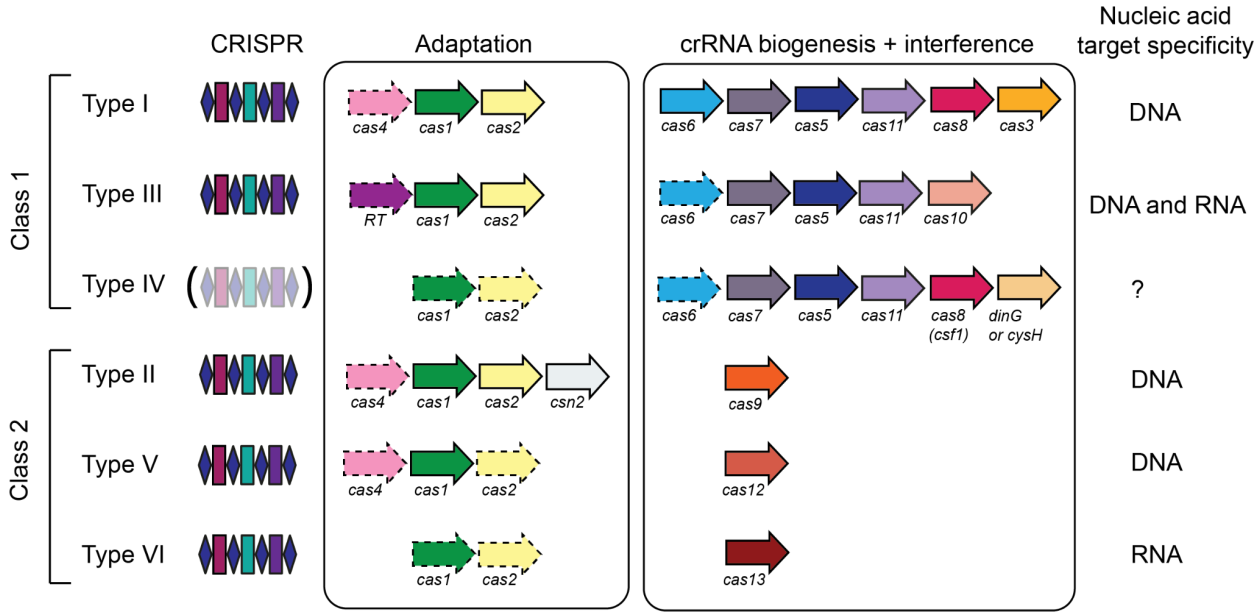


Figure 1.2 | Modular organization of CRISPR-Cas systems. CRISPR systems are classified into two classes and six types. The functional modules involved in adaptation and crRNA biogenesis and interference for each type are illustrated, and the nucleic acid target specificity is indicated. Genes only present in some subtypes are marked by dashed outlines.

1.3 Origins and mechanisms of CRISPR spacer capture

1.3.1 Origins and evolution of the CRISPR integrase

The emergence of the CRISPR adaptation module was a key event in the origins of prokaryotic adaptive immunity, enabling bacteria and archaea to record previous infections in the form of short foreign sequences inserted as spacers in host cell CRISPR arrays (Fig. 1.1). Phylogenetic studies suggest that the CRISPR adaptation module evolved from ancestral casposons, a novel class of DNA transposons that encode a Cas1 homolog, the caposase, as the transposase[14]. The remarkable parallels in structure and function between the casposase and CRISPR integrase provide insight about the elements necessary for the emergence of CRISPR adaptation (Fig. 1.3a). Cas1 and *M. mazei* casposase share a canonical two-domain architecture with an eight-stranded N-terminal domain and a C-terminal domain containing the conserved active site residues E-H-E[15]. The casposase's binding modes to the integration target and the ssDNA casposon ends resemble that of Cas1 bound to the CRISPR repeat and 3'-overhangs of the protospacer, respectively[10,16–18]. Interestingly, the casposase contains an additional C-terminal helix-turn-helix (HTH) domain, although it is unclear what role it plays in integration. The structural resemblance between the casposase and Cas1 translates to many functional similarities. The *M. mazei* casposase's unique ability among transposases to integrate a variety of non-sequence-specific substrates provides

the groundwork for the emergence of an adaptive immune system that needs to acquire diverse sequences to maintain a genetic record of prior infections[15]. Upon integration, the casposase generates 14-15 bp target site duplications (TSDs)[19,20], similar to how the CRISPR integrase generates 25-50 bp CRISPR repeat duplications[21–24]. Strikingly, the casposase's mode of target recognition also parallels that of the CRISPR integrase. Unlike other DNA transposons that insert DNA into random locations[19], the casposase inserts substrates into a preferred target site[15,20], which is a key characteristic of the CRISPR integrase[25,26]. For the *M. mazei* casposase, a conserved target motif within the TSDs and a sequence motif 12-17 bp upstream of the TSD are critical for directing integration[15]. This is analogous to CRISPR spacer acquisition, which depends on both a specific repeat sequence and sequence motifs within the upstream leader[21,25,26]. These similarities give insight into the origins of many key elements of CRISPR adaptation.

The casposase structure also provides clues to how the Cas1 integrase evolved to accommodate Cas2, a key subunit in most CRISPR integrases, as well as other binding partners involved in CRISPR adaptation (Fig. 1.3a)[16,17,27]. The casposase forms a tetramer for integration, composed of two dimers holding the target site between them[15]. While the casposase is structurally impeded from simultaneous tetramerization and binding to Cas2, similarities between the relevant interface of the casposase dimer and the Cas2-interacting interface of Cas1 suggest that surface mutations could allow the casposase dimer to bind to a Cas2 protomer. Introduction of a bridging Cas2 would maintain the orientation of the catalytic monomers toward each other and provide an extended binding surface for short DNA integration substrates. In addition to Cas2, the casposase also has interfaces that could accommodate binding to the CRISPR adaptation proteins Csn2 and Cas4, which assist Cas1—Cas2 in spacer acquisition in certain systems and are discussed in greater detail in later sections of this review[28]. The relevant binding interface with the auxiliary protein Csn2 is the eight-stranded N-terminal domain, which is remarkably similar in the Cas1 and Casposase structures[15,29]. At the dimer level, the casposase also appears able to accommodate interactions with a Cas4 nuclease and even has a conserved asparagine residue that is important for Cas4 binding[15,30]. However, interestingly, in both the casposase and CRISPR Cas4-Cas1—Cas2 integrase structures, the relevant interface for Cas4 binding overlaps with the interface involved in integration target binding. Further work is required to elucidate when Cas1 acquired these binding partners in its evolution from the casposase into a functional CRISPR integrase.

Phylogenetic analyses imply the existence of an ancestral CRISPR integrase composed of only Cas1 prior to the adoption of Cas2[4,31]. This is the focus of Chapter 2 of this thesis, which proposes that this Cas1 integrase may mark the emergence of a precise ruler mechanism that defines the length of integration substrates[32] (Fig. 1.3a,b). In a Cas1 phylogenetic tree from a recent review, the V-C and V-D Cas1 genes along with solo Cas1 genes form a branch rooted near the ancestral casposon branch[33]. These systems lack the Cas2 protein entirely and have unusually short spacers (17-20 bp) and repeats (25 bp) compared to other CRISPR systems. Chapter 2 of this thesis reveals that, like the casposase, the V-C Cas1 also forms a tetramer for integration[32]. However, while the casposase has a wide tolerance for different substrate lengths[15], the V-C integrase has a precise ruler mechanism that favors short integration substrates. More structural characterization is required to determine the evolutionary path from the

casposase to a functional CRISPR integrase and the emergence of the ruler mechanism. Interestingly, compared to the casposase, the V-C Cas1 integrase shows greater promiscuity for the integration site. Across different CRISPR systems, CRISPR integrases display varying levels of intrinsic specificity for the integration site, and some rely on host factors such as the integration host factor (IHF) to guide them to the correct site (reviewed in ref.[28]). For systems that rely on IHF, the phase of the IHF binding motifs appears to be critical to integration efficiency, as uncovered by a recent study examining the diversity and evolution of different CRISPR leader sequences[34]. It is unclear what host factors the V-C Cas1 integrase may require for specific integration. Further investigation could shine light on the evolution of the mechanisms by which CRISPR systems identify the preferred integration site and coevolution of the adaptation module with their supporting host factors.

A key event in the evolution of the CRISPR integrase is the acquisition of Cas2, transforming the tetrameric structure of the casposase and V-C Cas1 into the canonical heterohexameric Cas1—Cas2 complex observed in most CRISPR systems. Based on comparative sequence studies, the CRISPR adaptation module most likely adopted Cas2 from toxin-antitoxin (TA) modules (Fig. 1.3a)[2,35]. Structural comparisons with the functional TA module *H. influenzae* VapXD provide further insight into the evolutionary connection between TA systems and CRISPR adaptation[36]. The VapD protein is structurally related to Cas2: both form homodimers with a core ferredoxin fold and share conserved structural elements, including a similar nuclease active site, although Cas2 nuclease activity does not appear to be required for CRISPR acquisition[27,37,38]. Further work is required to understand VapD's mode of action in contributing to pathogenicity in the TA module and whether and how it is relevant to Cas2's role in CRISPR systems. The main structural differences between VapD and Cas2 occur in regions involved in interactions with the VapX antitoxin and Cas1[36]. VapD has two additional helices involved in VapX antitoxin binding, which, based on alignments with the Cas1—Cas2 structure bound to protospacer DNA, would sterically occlude protospacer binding[10,16–18,36]. Cas2, on the other hand, has an additional β-hairpin at the C terminus responsible for Cas1 interaction. These architectural differences hint at the loss of VapX as a binding partner and the adoption of Cas1 as a new binding partner in the transformation of VapD from a toxin in a TA system to a critical component of the CRISPR adaptation module. There is a potential evolutionary advantage for the adoption of Cas2 as a structural unit in the integrase structure, which could provide increased stability and would lengthen the internal ruler to enable integration of longer sequences, offering greater targeting specificity of the crRNA for downstream CRISPR immunity processes[32]. However, the loss of the VapX antitoxin as a binding partner could have toxic effects, which the bacteria would need to account for in order to recruit VapD into the CRISPR integrase.

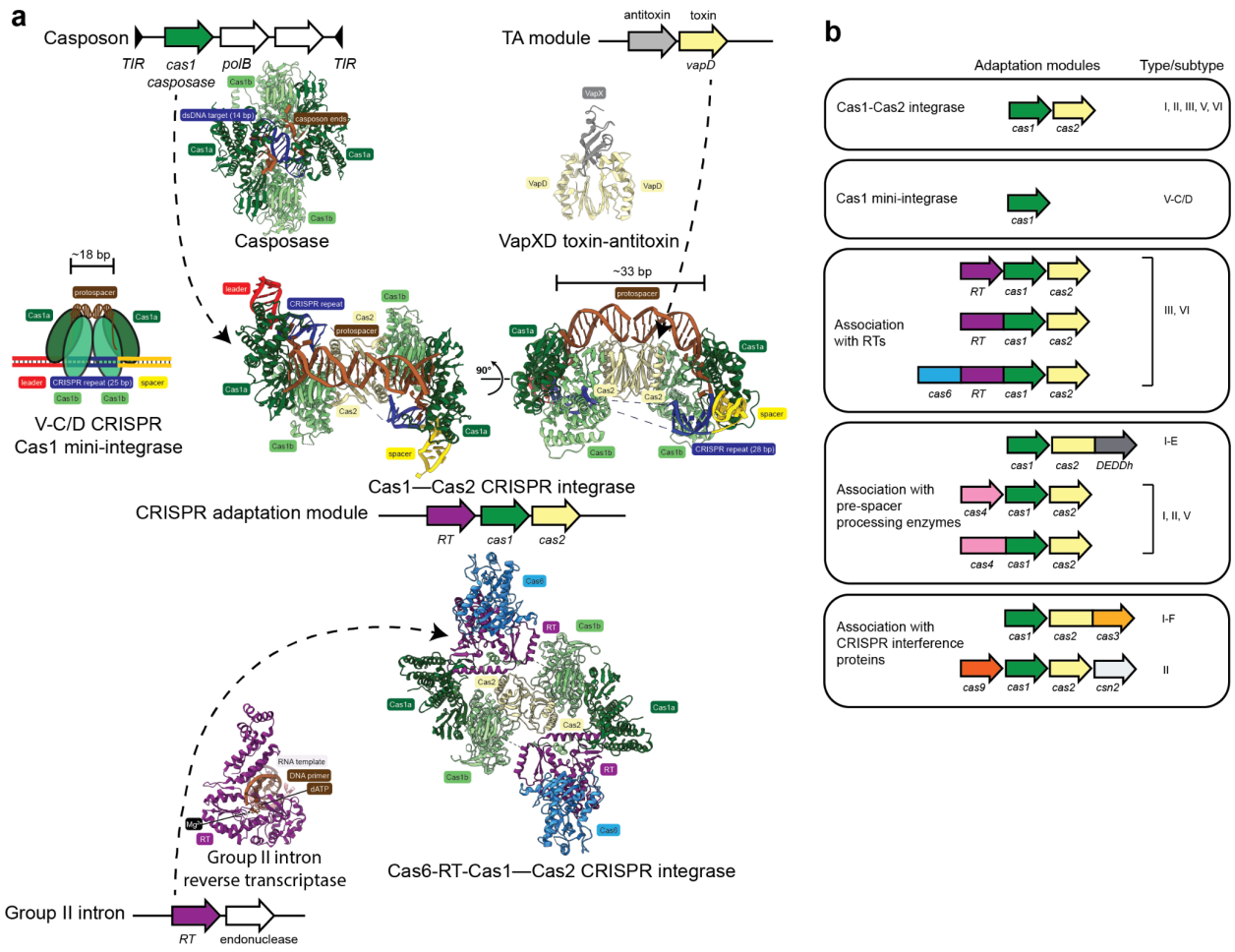


Figure 1.3 | Evolution and diversity of the CRISPR integrase architecture. (a) Contributions of casposons, TA modules, and group II introns to origins and evolution of the CRISPR adaptation module and depiction of homologous structures. Top left: casposase bound to integration product mimic (PDB-ID 6OPM)[15]. DNA substrate is color-coded: dsDNA target, dark blue; casposon ends, brown. Top right: VapXD toxin-antitoxin complex (PDB-ID 6ZN8)[36]. Center: V-C/D CRISPR Cas1 mini-integrase cartoon (left) and Cas1—Cas2 CRISPR integrase bound to full-site integration product mimic (right, PDB-ID 5VVK)[18]; DNA substrates are color-coded: leader, red; CRISPR repeat, dark blue; spacer, yellow; protospacer, brown. Bottom left: structure of group II intron reverse transcriptase bound to RNA template / DNA primer and dATP (PDB-ID 6AR3)[39]. Nucleic acids are color-coded: RNA, lavender; DNA brown. Bottom right: structure of Cas6-RT-Cas1—Cas2 CRISPR integrase (PDB-ID 7KFU)[40]. Two of the four Cas6-RT domains are missing. Protein structures are colored by domain: catalytic Cas1a monomer, dark green; non-catalytic Cas1b monomer, light green; Cas2/VapD, light yellow; VapX, gray; Cas6, light blue; RT, magenta. TIR represents terminal inverted repeat sequence. Protospacer lengths are indicated for the V-C/D Cas1 integrase and Cas1—Cas2 integrase. **(b)** Diversity of domain organizations of CRISPR adaptation module.

1.3.2 Recruitment of reverse transcriptases for RNA spacer acquisition

In the evolution of CRISPR spacer capture, some acquisition modules apparently recruited reverse transcriptases (RTs) from group II introns, enabling these systems to acquire spacers from RNA in addition to DNA (Fig. 1.3a,b)[41–43]. In several of these systems, the RT is fused with either Cas1 or with the Cas6 maturase, enabling

coordinated integration and crRNA generation (Fig. 1.3b)[40,44]. The *M. mediterranea* Cas6-RT-Cas1—Cas2 integrase has been shown to carry out precise ligation of short ssRNA substrates into the CRISPR array *in vitro* and to conduct target primed reverse transcription using the free 3' DNA end generated after RNA ligation[41]. Interestingly, for these fusions, the Cas6 domain is required for RNA spacer acquisition, and analysis of RT-containing systems suggests coevolution of the Cas6 and RT domains[44]. Chapter 3 of this thesis reveals the architecture of the *Thiomicrospira* Cas6-RT-Cas1—Cas2 complex and shows structural interactions between the Cas1—Cas2 integrase, the RT, and the Cas6 maturase, enabling functional crosstalk between the three active sites[40](Fig. 1.3a). A connecting alpha helix links the RT and Cas1 active sites and serves as a potential regulator of the two activities. Although much of the RT domain aligns closely to the group II intron RT structure[39], its palm region containing the catalytic center appears to be in an inactive conformation[40]. The distinct RT conformation and the occluded Cas1 active site hint at larger conformational changes that may coordinate the different enzymatic activities of the complex. Further structural characterization is required to understand how the RT and Cas1 achieve CRISPR sequence integration from RNA substrates as well as Cas6's role in this process. How the RNA substrate is reverse transcribed, whether ssRNA ligation represents an actual intermediate step *in vivo*, and how this intermediate is resolved into a full-site integration product are questions that remain to be addressed. It is also not known whether RNA spacer acquisition confers immunity against RNA phages and other foreign RNA elements.

1.3.3 Mechanisms of CRISPR protospacer integration

Biochemical and structural studies in the past decade have elucidated the protospacer integration mechanism[28,45]. Integration consists of two transesterification reactions by the 3'-OH ends of the canonical integration substrate, a ~23 bp duplex DNA with short single-stranded 3' overhangs, on both sides of the leader proximal repeat[23]. Cas1 provides the catalytic activity, while Cas2 plays an important structural role by bridging the two Cas1 dimers, orienting the two catalytic Cas1 active sites toward each other, and stabilizing the protospacer DNA during integration[10,16](Fig. 1.4a,b). Cas1—Cas2 has exquisite control over the length of the duplex region: the *E. coli* integrase caps the ends of a 23-bp duplex with tyrosine residues[10,16](Fig.1.4c). This capping is similarly observed in other Cas1—Cas2s, such as the *E. faecalis* and *Synechocystis* Cas1—Cas2, which use histidine and aspartate residues, respectively, to cap a 22-bp duplex[17,46]. The single-stranded 3' overhangs are threaded into the Cas1 active site channels[10,16,17](Fig. 1.4b). The first nucleophilic attack results in a branched intermediate called the half-site integration intermediate, and the second nucleophilic attack by the other protospacer strand results in the fully integrated spacer, known as the full-site integration product[24](Fig. 1.4d). The integrated spacer is flanked on both ends by the repeat sequence, ensuring precise repeat duplication. The resulting gaps are then filled and ligated by host enzymes.

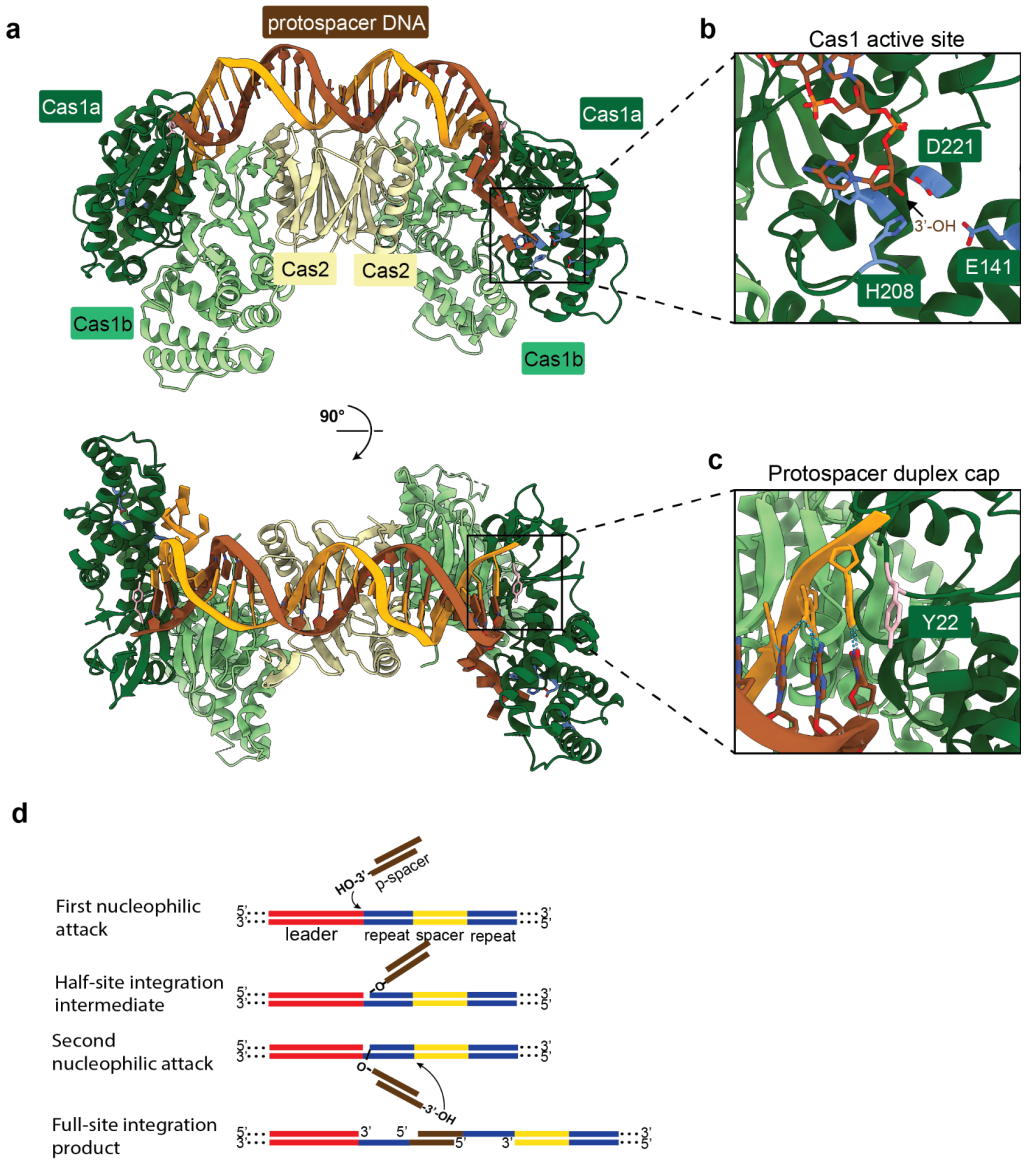


Figure 1.4 | Protospacer integration mechanism. (a) Structure of *E. coli* Cas1—Cas2 integrase (PDB-ID 5DS4) bound to protospacer DNA and 90° rotation. Catalytic Cas1a, dark green; noncatalytic Cas1b, light green; Cas2, yellow; protospacer strands, brown and orange. Cas1 active site residues are colored in blue. Cas1 tyrosine capping the protospacer duplex is colored pink. **(b)** Close-up of Cas1 active site. The nucleophilic 3'-OH is indicated with an arrow. **(c)** Close-up of the protospacer duplex cap showing the stacking interaction between the Cas1 tyrosine residue and the last 5' nucleotide. **(d)** Protospacer integration mechanism, showing the two nucleophilic attacks by the protospacer strands into the CRISPR array. Leader, red; repeat, blue; spacer, yellow; protospacer, brown. The nucleophilic attacks by the 3' OH ends of the protospacer are indicated by curved arrows.

1.3.4 CRISPR integration substrate biogenesis, selection, and processing

While the CRISPR protospacer integration mechanism has been well-characterized, the processes upstream of integration, including substrate biogenesis and processing, are less well understood. This is the topic of the last chapter

of this thesis. In the current model for spacer acquisition, RecBCD or AddAB generates DNA degradation products from foreign genetic material that are the precursors for CRISPR integration[47,48]. The molecular details of CRISPR substrate biogenesis remain unknown, though recent studies have elucidated how substrate precursors are selected, processed to the correct length, and integrated in the correct orientation with respect to the PAM position to yield functional crRNA generation for downstream CRISPR immunity processes[16,30,49–53].

Cas1—Cas2 binds to suitable precursor DNA substrates, which are then processed into mature integration-competent molecules. The *E. coli* Cas1—Cas2, which has a PAM binding pocket formed by the groove of the C-terminal domain of the catalytic Cas1a and the C-terminal tail of the non-catalytic Cas1b, preferentially selects partially duplexed DNA substrates with long 3' single-stranded overhangs that contain a PAM sequence in the correct location flanking the eventually integrated spacer[16,49](Fig. 1.5a,b). Cas1—Cas2 likely predetermines the length of the eventual spacer by protecting the correctly sized region of the substrate from being processed. In certain systems, Cas2 is naturally fused to a DEDDh exonuclease, a DnaQ-like exonuclease, which processes long 3' overhangs of precursor prespacers to the correct length for integration when complexed with Cas1[54]. Chapter 4 of this thesis describes how a Cas1—Cas2-DEDDh complex trims prespacers in a ruler-guided and PAM-dependent manner and carries out delayed PAM processing via a gate-keeping ruler mechanism. We propose that similar type I-E systems without such a natural fusion utilizes the same mechanism with host exonucleases. For the type I-E *E. coli* system, DnaQ exonucleases including DNA PolIII and ExoT have been shown to catalyze DNA substrate processing when provided in concert with Cas1—Cas2[49,50]. Further structural characterization is required to understand the molecular details of the interaction between these processing enzymes and the Cas1—Cas2 integrase.

In certain type I, II, and V CRISPR systems, the Cas4 nuclease works together with Cas1—Cas2 for spacer acquisition and assists in PAM selection and substrate maturation[30,51,52,55]. In these systems, Cas4, rather than Cas1, contains the PAM binding pocket that sequesters the PAM-containing 3'-overhang[30](Fig. 1.5a,b). For the *G. sulfurreducens* system, only the presence of a short DNA molecule containing a PAM triggers assembly of a functional integrase complex consisting of the natural Cas4-Cas1 fusion and its associated Cas2, suggesting that Cas4 acts as a gatekeeper in selecting suitable substrate precursors. Cas4 further functions as an endonuclease that catalyzes precise cleavage of PAM sequences before the substrate becomes fully integrated[52,55]. This step is critical for preventing self-targeting of newly acquired CRISPR spacer sequences to the host cell CRISPR array when the immune system is active.

For both the *E. coli* Cas1—Cas2 and various Cas4-Cas1—Cas2 integrases, PAM binding establishes the directionality of integration. Sequestration of the PAM prevents its trimming by processing nucleases and temporarily blocks that side of the substrate from integration (Fig. 1.5a). Meanwhile the non-PAM 3' side of the DNA is trimmed efficiently and freed for integration into the leader-proximal end of the repeat, the preferred first integration site for both systems[30,49]. In the *E. coli* system, half-site integration enables further trimming of the PAM[49], although there have previously been ambiguous results if this is carried out by processing enzymes or by Cas1[16,49]. Chapter 4 of this thesis provides more clarification, suggesting that for these type I-E systems,

host exonucleases process the PAM and coordinate with Cas1—Cas2's gate-keeping ruler mechanism to ensure PAM removal. In Cas4-containing systems, half-site integration activates Cas4 to process the PAM via endonucleolytic cleavage, as shown in the recently characterized *G. sulfurreducens* system[30]. The structure of the *G. sulfurreducens* Cas4-Cas1—Cas2 suggests that a sub-optimally placed catalytic lysine residue in the conserved DEK motif in the Cas4 active site initially inhibits PAM cleavage and a conformational change resulting from half-site integration can reposition the lysine to activate endonucleolytic PAM cleavage (Fig. 1.5c). In both *E. coli* and *G. sulfurreducens*, PAM-processing frees that DNA strand for full integration of the prespacer. These studies show that conformational dynamics of PAM sequestration and delayed trimming is a recurring theme in the tightly regulated stepwise processing and directional integration of prespacers. Whether systems that lack PAMs coordinate spacer orientation remains to be determined.

In type II-A CRISPR systems including those found in *S. pyogenes* and *S. thermophilus*, which do not encode Cas4, Cas9 is responsible for PAM recognition for spacer acquisition[56,57]. In these systems, all the genes encoded in the CRISPR locus (tracrRNA, *cas9*, *cas1*, *cas2*, and *csn2*) are required for spacer acquisition[56]. While Cas9's PAM-binding motif is not necessary to acquire new spacers, it is required to select functional PAM-adjacent spacers that would provide immunity[56]. Cas9's HNH domain appears to have exonuclease activity independent of a guide RNA sequence that can trim DNA precursors to the correct size for integration[58]. Other host nucleases can also carry out this function since Cas9's nuclease activity is not required for spacer acquisition *in vivo*[56]. Cas9 has been shown to form a complex with Cas1—Cas2 and the auxiliary protein Csn2[29,56,58]; however, further structural characterization is necessary to establish how these domains coordinate DNA substrate selection and integration.

In addition to naive adaptation, resulting from infection of previously unencountered foreign material, some CRISPR systems have evolved another pathway for spacer acquisition known as primed adaptation to efficiently adapt to previously encountered invading DNA that has acquired escape mutations. The current understanding of primed adaptation has been recently reviewed[45]. In type I systems, this involves the formation of a primed acquisition complex (PAC) by Cas1—Cas2, Cascade and Cas3, which translocates along DNA and hands off Cas3 cleavage products to Cas1—Cas2 for integration[22,53,59,60]. While fluorescence and chromatin immunoprecipitation experiments have provided insight into the molecular mechanisms of type I priming[53,60,61], the nature of Cascade's conformational changes during priming, PAC assembly, and substrate hand-off from Cas3 to the Cas1—Cas2 integrase are not yet known. In type II systems, priming is directly correlated with target cleavage by Cas9[62,63]. Future structural and single-molecule data will also be required to understand substrate hand-off from Cas9 to the integrase in type II priming and whether it involves the Cas9—Csn2—Cas1—Cas2 supercomplex[29].

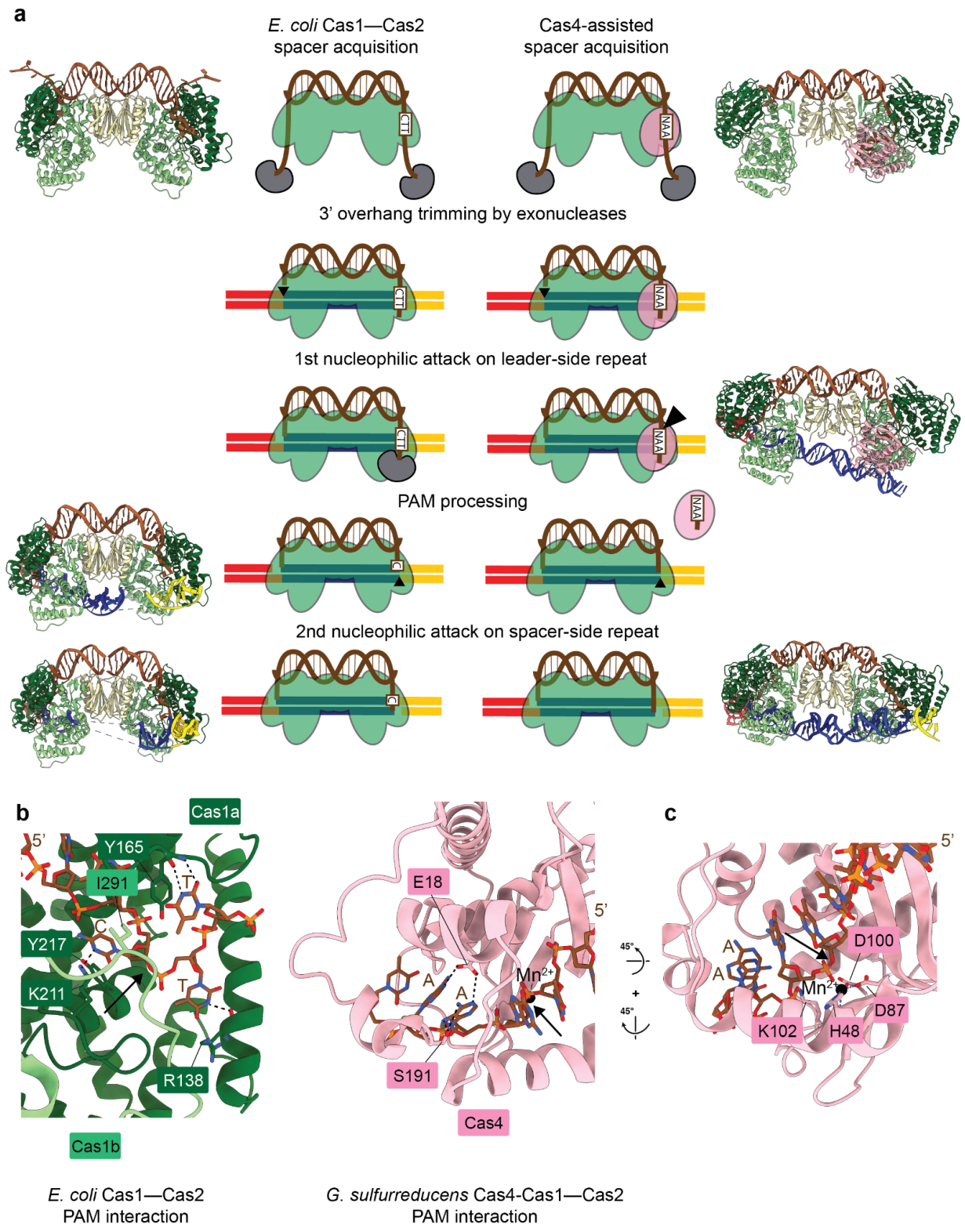


Figure 1.5 | Mechanisms of CRISPR prespacer biogenesis and orientation. (a) Schematic of prespacer processing and directional integration by *E. coli* Cas1—Cas2 and Cas4-assisted CRISPR adaptation systems. In the schematic cartoons, the DNA prespacer (brown) is bound to Cas1—Cas2

(green) and the 3' ends are trimmed by exonucleases (gray). The PAM (shown in brown boxes) for each system is initially protected from trimming and the non-PAM 3' end is integrated into the leader-side of the CRISPR array (leader, red; repeat, dark blue; spacer, yellow). PAM-processing is carried out by exonucleases or Cas1 in the *E. coli* system and by Cas4 (pink) in Cas4-assisted acquisition and the processing sites are indicated by arrowheads. After PAM-processing, the 3' end is freed for spacer-side integration. The associated structures of *E. coli* Cas1—Cas2 (left) and *G. sulfurreducens* Cas4-Cas1—Cas2 (right) at different stages of spacer acquisition are shown next to the cartoons. *E. coli* Cas1—Cas2 structures from top to bottom: Cas1—Cas2 bound to dual-PAM prespacer DNA (PDB-ID 5DQZ)[16], Cas1—Cas2 bound to half-site integration product mimic (PDB-ID 5VVJ)[18], Cas1—Cas2 bound to full-site integration product mimic (PDB-ID 5VVK)[18]. As a point of clarification, the first *E. coli* structure uses a dual-PAM prespacer DNA as a substrate. It is expected that the natural substrate in cells only has a single PAM, shown in the cartoon model. If the cell does encounter a dual-PAM prespacer DNA, it is unclear how PAM protection on both sides affects PAM processing and integration. *G. sulfurreducens* Cas4-Cas1—Cas2 structures from top to bottom: Cas4-Cas1—Cas2 bound to single-PAM prespacer DNA (PDB-ID 7MI5)[30], Cas4-Cas1—Cas2 bound to half-site integration product mimic with Cas4 still engaged in PAM recognition (PDB-ID 7MIB)[30], Cas4-Cas1—Cas2 bound to full-site integration product mimic (PDB-ID 7MI9)[30]. In the top two *G. sulfurreducens* Cas4-Cas1—Cas2 structures, the three non-PAM-interacting Cas4s are missing. In the Cas4-Cas1—Cas2 bound to full-site integration product mimic structure, all four Cas4s are missing. **(b)** Close-up of Cas1 PAM interaction (left) in *E. coli* Cas1—Cas2 (PDB-ID 5DQZ)[16] and Cas4 PAM interaction (right) in *G. sulfurreducens* Cas4-Cas1—Cas2 (PDB-ID 7MI4)[30]. PAM-interacting residues and sequence-specific hydrogen-bonding interactions are depicted. Arrow depicts PAM-processing cleavage site. **(c)** Close-up of Cas4 active site, with arrow depicting PAM-processing cleavage site. In the integrase structures, Cas4 is shown in pink, and the other domains follow color-coding from Fig. 1.3. The DNA substrates are color-coded: leader, red; repeat, dark blue; spacer, yellow; prespacer, brown.

1.4 Conclusion and outlook

Fifteen years after the demonstration of CRISPR immunological memory in bacteria[1], structural and analytical studies have uncovered a plethora of intricate mechanisms that govern the selective capture and processing of foreign genetic material for spacer acquisition and discernment of self vs. nonself. In the arms race between prokaryotes and phage, CRISPR systems have evolved sophisticated machinery to carry out their functions with exquisite control. From a biological perspective, the architectures of CRISPR adaptation machinery reflect different drivers of genome evolution and reveal evolutionary relationships between CRISPR systems and mobile genetic elements (MGEs). CRISPR systems have recruited MGEs to aid in their defense against invaders on multiple occasions in the evolutionary history of CRISPR-Cas, as observed in the origins of the Cas1—Cas2 adaptation module from casposons and TA modules and the recruitment of reverse transcriptases from retrotransposons.

Knowledge of the diversity of the machinery and mechanisms employed by CRISPR adaptation modules can also help advance the development of CRISPR adaptation-based molecular recording tools, which have been utilized for *in vivo* data storage in bacteria[64–67]. A greater understanding of the elements required for integration substrate selection and processing may be the key to increasing the efficiency of these tools and expanding their use in other microbes and mammalian cells to study horizontal gene transfer and lineage tracing.

1.5 Acknowledgements

We thank J. C. Cofsky, K. M. Soczek and G. J. Knott for sharing the coordinates of Cas9 bound to linear and bent dsDNA prior to publication. J.Y.W. is supported by the US National Science Foundation Graduate Fellowship and previously by the Berkeley Graduate Fellowship. P.P. receives funding from the European Regional Development Fund under grant agreement No 01.2.2-CPVA-V-716-01-0001 with the Central Project Management Agency (CPVA), Lithuania. This material is based upon work supported by the National Science Foundation under award number 1817593 and by the Somatic Cell Genome Editing Program of the Common Fund of the National Institutes of Health under award number U01AI142817-02. J.A.D. is an HHMI investigator. We thank G. J. Knott and A. Lapinaite for helpful discussions.

1.6 Contributions

All authors contributed to writing, reviewing, and editing the manuscript.

CHAPTER 2

A functional mini-integrase in a two-protein type V-C CRISPR system

A portion of the work presented in this chapter has been published previously as part of the following paper: 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. A functional mini-integrase in a two-protein type V-C CRISPR system. *Mol. Cell* **73**, 727–737.e3 (2019).

* These authors contributed equally.

2.1 Abstract

CRISPR-Cas immunity requires integration of short, foreign DNA fragments into the host genome at the CRISPR locus, a site consisting of alternating repeat sequences and foreign derived spacers. In most CRISPR systems, the proteins Cas1 and Cas2 form the integration complex and are both essential for DNA acquisition. Most type V-C and V-D systems lack the cas2 gene and have unusually short CRISPR repeats and spacers. Here, we show that a mini integrase comprising the type V-C Cas1 protein alone catalyzes DNA integration with a preference for short (17–19 base pair) DNA fragments. The mini-integrase has weak specificity for the CRISPR array. We present evidence that the Cas1 proteins form a tetramer for integration. Our findings support a model of a minimal integrase with an internal ruler mechanism that favors shorter repeats and spacers. This minimal integrase may represent the function of the ancestral Cas1, prior to Cas2 adoption.

2.2 Introduction

CRISPR-Cas (clustered regularly interspaced short palindromic repeats, CRISPR associated) systems function as a form of adaptive immunity for bacteria and archaea[1]. In these systems, a dedicated integrase incorporates segments of foreign DNA, termed protospacers, into the host CRISPR locus in a process known as acquisition[21,68,69]. The CRISPR locus, which consists of direct repeats separated by viral-derived spacers of similar length, is transcribed and processed into mature CRISPR RNAs (crRNAs) that are used by an effector complex to identify and destroy any phage or plasmid bearing sequence complementarity to the virally derived segment of the crRNA[70–75].

While the Cas proteins and mechanisms involved in CRISPR immunity vary widely, the acquisition module is largely conserved[33]. In all studies conducted so far, Cas1 and Cas2 have been found to be necessary for acquisition[69]. These two proteins form a heterohexameric complex that catalyzes the integration of new protospacers at one end of the CRISPR array[1,24,27]. Cas1 is the catalytic subunit, while Cas2 plays an important structural role by bridging the two Cas1 dimers and stabilizing both the bound protospacer DNA and the CRISPR repeat during integration[16–18,21–23,27].

This architecture provides an internal ruler mechanism that dictates the 25–50 base pair length of the protospacer for integration[10,16]. Integration occurs when the 3' ends of the protospacer make a nucleophilic attack at each end of the CRISPR repeat sequence, resulting in a half-site intermediate after the first attack that proceeds to a full-site product[23,24]. Specific recognition of the CRISPR repeat is important to avoid damage to the genome resulting from off-target integration. Integration is specific to the CRISPR locus with the preferred site being the first repeat adjacent to the leader, an upstream sequence that also contains the promoter for the CRISPR array[21,24,26,27,76].

Cas1—Cas2 integrases have intrinsic target specificity, although host factors have been shown to improve integration specificity in some hosts[77–81].

Since this Cas1—Cas2 module is conserved across most systems, the recent discovery of a potential non-canonical acquisition module provides an opportunity to investigate the activity of a more minimal integrase[4,31]. Two type V CRISPR-Cas subtypes, V-C and V-D, have been shown to lack cas2 genes entirely and have notably shorter CRISPR spacers and repeats. These observations are especially exciting given recent studies suggesting that Cas1 originated from a superfamily of DNA transposons called casposons[14,19,33,82]. Casposons encode a Cas1 homolog, called a casposase, which has transposase activity and some sequence preference for its integration sites; notably, like the type V systems, casposons do not include Cas2[14,20]. Determining whether and how the type V systems carry out acquisition without Cas2 and where these systems fit in the evolutionary history of CRISPR-Cas can shed light on the origins and diversity of CRISPR-based anti-viral immunity.

We hypothesized that the Cas1 protein of CRISPR type V-C and type V-D systems functions on its own as a DNA integrase whose architecture might confer a preference for the characteristic short spacer sequences in the CRISPR arrays of these systems. Here, we reconstitute the type V-C integration reaction and show that type V-C Cas1 is sufficient to catalyze full-site integration of 18-base pair DNA substrates. Furthermore, our experiments reveal that V-C Cas1 has an intrinsic mechanism to orient protospacers in a biased manner during integration. We observed considerably more off-target integration with this mini-integrase relative to other CRISPR integrases, suggesting the need for additional host factors to provide target site specificity. We present evidence for Cas1 multimerization triggered by association with half-site and full-site integration product mimics. Our findings support a model for a minimal CRISPR integrase in which Cas1 provides both DNA substrate length selectivity and catalytic activity necessary for foreign DNA acquisition and which may represent the ancestral CRISPR adaptation module.

2.3 Results

2.3.1 Type V-C and V-D systems have CRISPR arrays with short repeats and spacers and lack *cas2* genes

To date, type V-C and type V-D CRISPR-Cas systems have been found exclusively in uncultured microbes. Type V-C systems have been identified in marine and gut metagenomes, encoded on contigs which were too short to determine their host, and type V-D were identified in the genomes of small, putatively symbiotic bacteria from groundwater samples[4,31]. In both systems, adaptation modules lack cas2 and are characterized by short spacers (17–20 bp) (Fig. 2.1a,b). To better characterize type V-C

and V-D systems, we searched for previously unidentified systems within assembled metagenomes available in the IMG/M database, which incorporates a wide variety of environments[83]. We report here the identification of a new V-C system from a mouse cecum sample and a V-D system from beetle gut (Table 2.1). In these systems, cas2 genes were also absent and the spacers were considerably shorter compared to spacers from other CRISPR-Cas systems (Fig. 2.1c).

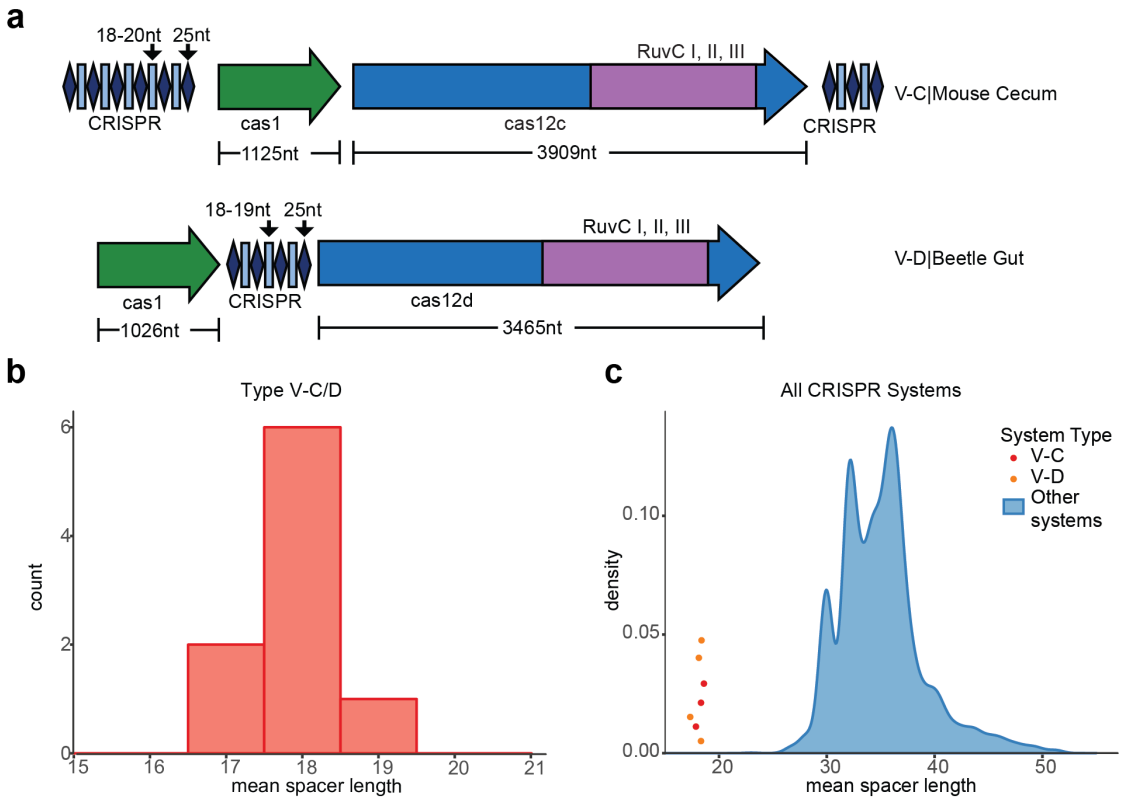


Figure 2.1 | Type V-C and V-D CRISPR systems have short repeats and spacers and lack cas2 genes. (a) CRISPR-Cas loci of the effector proteins of the type V-C system from mouse cecum and V-D system from beetle gut with lengths of genes and repeats and spacers. See Table 2.1 for protein sequences. **(b)** Distribution of mean spacer lengths of type V-C/D systems. **(c)** Comparison of mean spacer length for V-C/D vs. other CRISPR-Cas systems. The blue density plot depicts the distribution of mean spacer length in CRISPR-Cas systems, excluding V-C/D systems. The mean spacer lengths of published V-C and V-D systems are noted in red and orange dots, respectively, with arbitrary y-values to set them apart.

2.3.2 Type V-C Cas1 catalyzes DNA integration without Cas2

We set out to determine whether type V-C CRISPR-Cas1 can catalyze DNA substrate integration into a target DNA molecule. We expressed and purified the Cas1 protein derived from both the new type V-C and type V-D CRISPR systems and conducted integration assays with the well-characterized Escherichia coli type I-E

Cas1—Cas2 integrase as a positive control. These acquisition assays were performed with a supercoiled target plasmid containing the cognate CRISPR locus for each protein and a 33-bp or 18-bp DNA substrate for the *E. coli* Cas1—Cas2 complex or type V Cas1, respectively (Table 2.2). We evaluated integration by analyzing changes in plasmid topology, using gel electrophoresis to separate the open-circle integration products from the non-integrated supercoiled plasmid (Fig. 2.2a)[24].

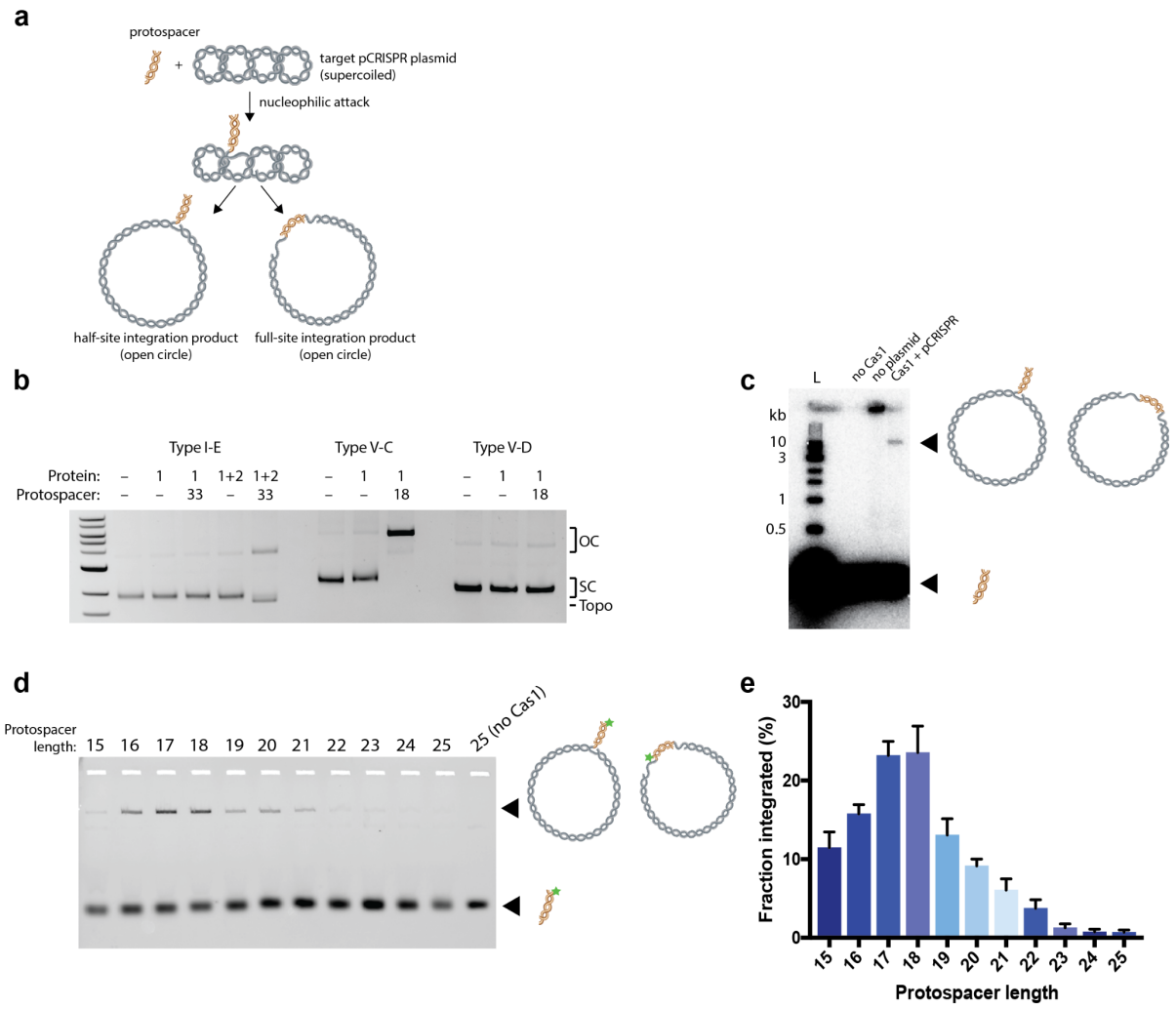


Figure 2.2 | Type V-C Cas1 integrates DNA fragments of expected length *in vitro*. (a) Schematic of *in vitro* integration of protospacer into a target supercoiled pCRISPR plasmid. Nucleophilic attack by the protospacer yields two different open-circle products: the half-site and full-site integration products. **(b)** Integration assay comparing type I-E, V-C, and V-D systems. A 33-nt protospacer is used for the type I-E system and an 18-nt protospacer for the type V-C and V-D systems. The open-circle integration products (OC), supercoiled target plasmid (SC), and topoisomers (Topo) are indicated. **(c)** Integration assay with type V-D Cas1, radiolabeled protospacer, and plasmid target. Integration product and free protospacer are indicated and schematized. **(d)** Integration assay with variable length fluorescent protospacers from 15-bp to 25-bp long. Star indicates 6-carboxyfluorescein label. **(e)** Quantification of **d** demonstrating the effect of protospacer length on type V-C Cas1 integration. The fraction integrated is calculated as the fraction of the fluorescent protospacer that has been integrated into the target plasmid. Experiments were

carried out in triplicate; the bars represent mean values, with error bars depicting standard deviations. See Table 2.2 for nucleotide sequences.

For the *E. coli* proteins, both Cas2 and protospacer DNA were required for integration, consistent with previous experiments (Fig. 2.2b)[24]. For the type V-C Cas1, we observed efficient production of open-circle plasmid only in the presence of the 18-bp DNA substrate. These results show that the type V-C Cas1 catalyzes plasmid nicking and/or ligation of short DNA fragments to plasmid DNA in the absence of Cas2 (Fig. 2.2b). Notably, type V-C Cas1 does not generate any observable plasmid topoisomers, which are visible as a band that migrates ahead of the supercoiled target plasmid for the *E. coli* Cas1—Cas2 integration reactions[24]. These products result from abortive integration (integration followed by disintegration), and their absence suggests that the type V-C Cas1 is less prone to reversible integration under the tested conditions[24]. Together, these results show that the type V-C Cas1 protein alone possesses catalytic activity in the presence of short DNA fragments, consistent with its potential function as a mini-integrase.

We did not observe generation of open-circle plasmid products by the type V-D Cas1 in these experiments, which could indicate a complete lack of activity or simply a lower reaction efficiency. To increase the sensitivity of the assay, we repeated the experiment using radiolabeled 18-bp DNA substrates[24]. When the type V-D Cas1 was incubated with the labeled DNA and a target plasmid, the radiolabel was incorporated into a product the size of an open-circle plasmid, suggesting the V-D Cas1 is indeed capable of protospacer integration, albeit at reduced levels under the studied reaction conditions (Fig. 2.2c).

We next wondered whether the type V-C Cas1, like other previously studied CRISPR integrases, also has an internal ruler mechanism that dictates the spacer length and whether its preferred DNA substrates *in vitro* would match the *in vivo* distribution of spacers for these systems. We explored the DNA substrate length requirements for the type V-C Cas1 plasmid addition activity by performing integration assays with fluorescent protospacers with lengths ranging from 15–25 bp (Table S2). The results showed that V-C Cas1 is most active with substrates that are 18 bp long, consistent with the predominant spacer length for native V-C CRISPR arrays (Fig. 2.1b, 2.2d,e, 2.3a). Although the extent of integration decreased as the DNA fragment length diverged from 18 bp, a significant amount of integrated protospacer was detected for substrates 16–21 bp. We observed a generally similar distribution when identifying integration through open-circle formation (Fig. 2.3b,c). In both assays, integration was also observed with substrates outside the range observed for spacers *in vivo*. While V-C may be able to integrate a range of protospacer lengths, an alternative possibility is that V-C Cas1 processes longer oligos to preferred length and then inserts them. To examine whether integration of long protospacers is processing-dependent, we carried out integration reactions with a 23-bp protospacer with phosphorylated 3’ ends and

compared the results to regular 23bp protospacer integration. The 3' phosphate is expected to prevent integration unless one or more terminal nucleotides are removed to expose a 3' OH[80]. The results show very low levels of conversion to open-circle products (less than 5% after 60 minutes) with the phosphorylated substrate compared to much greater conversion rates (over 15% after 60 minutes) with the regular 23-bp substrate (Fig. 2.3d,e). This suggests that most integration of long protospacers is processing-independent, though this Cas1 is capable of protospacer processing. A more likely explanation for the integration of long protospacers is promiscuity of the half-site reaction, as previous work in a type II-A has shown that protospacer length preference is more pronounced for the full-site reaction[26].

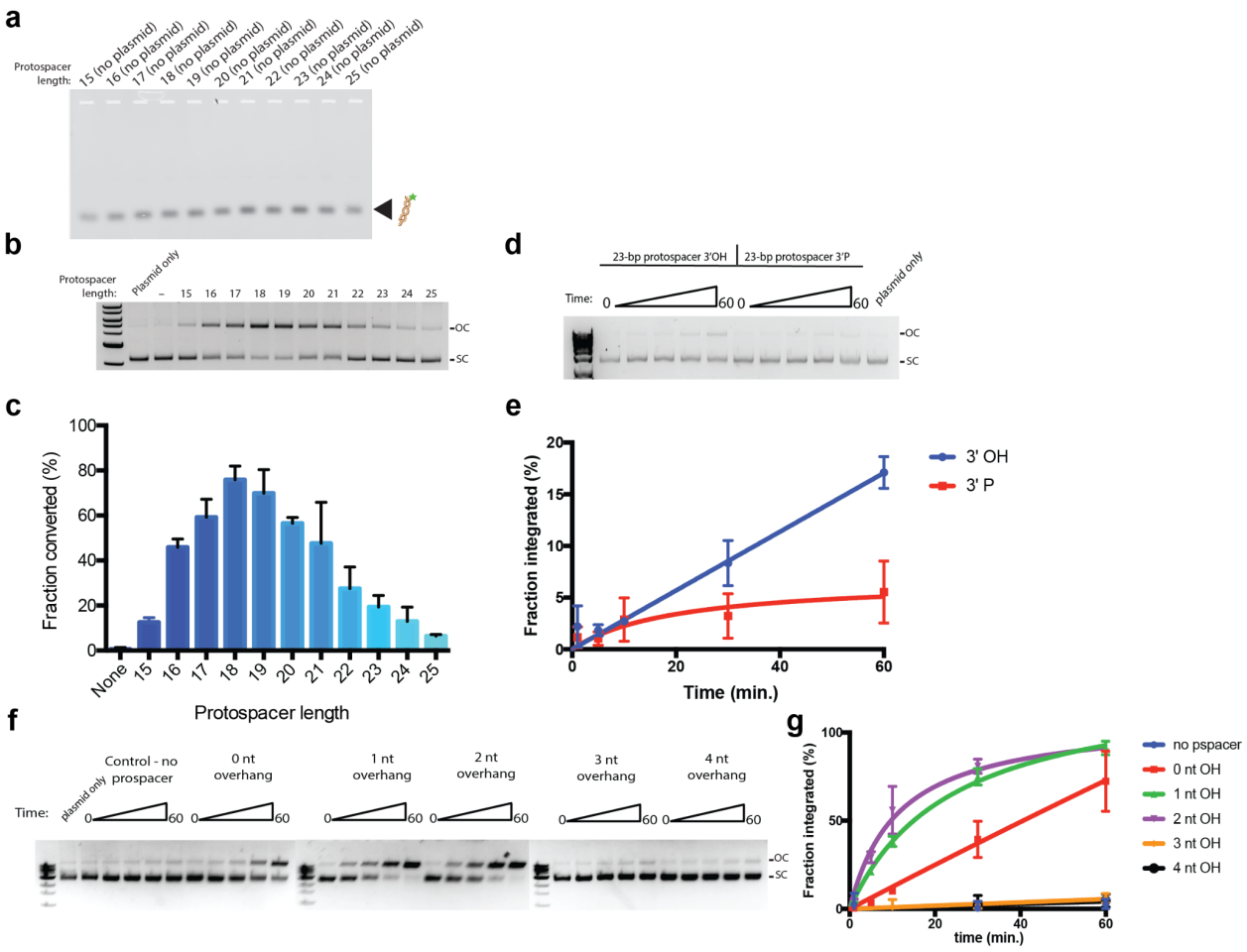


Figure 2.3 | Determination of type V-C Cas1 protospacer preferences. (a) Control integration assay with variable length fluorescent protospacers with no plasmid. **(b)** Integration assay with variable length protospacers from 15-bp to 25-bp long. **(c)** Quantification of **b** demonstrating the effect of protospacer length on the conversion of supercoiled plasmid to open-circle products by type V-C Cas1. The fraction converted is calculated as the fraction of open-circle products relative to all plasmid. Experiments were carried out in triplicate; the bars represent mean values, with error bars depicting standard Deviations. **(d)** Time-course integration reactions with a 23-bp protospacer and a 23-bp protospacer with

3’ phosphate modifications on both strands. **(e)** Quantification of **d**. Plotted points represent the mean of three independent experiments. Error bars represent standard deviation. **(f)** Time-course integration reactions with protospacers with 0- to 4-nt overhangs. **(g)** Quantification of **f**. Plotted points represent the mean of three independent experiments. Error bars represent standard deviation. See Table 2.2 for nucleotide sequences.

In addition to varying DNA fragment length, we also tested substrates with different single-stranded 3’ overhangs from 0–4 nt long, since other CRISPR integrases prefer substrates with single-stranded 3’ overhangs (Table 2.2)[10,16,26]. We found that substrates bearing 1–2 nt 3’ overhangs on either end were integrated into the target plasmid at faster rates than those with blunt ends, with 2-nt-overhang substrates yielding a 5-fold increase in the amount of open-circle plasmid relative to blunt substrates after a 10 minute reaction (Fig. 2.3f,g). No integration was observed for any protospacers with overhangs longer than 2 nt (Fig. 2.3f,g). From these results, we determined that the preferred substrate of type V-C Cas1 is an 18-bp protospacer with 2-nt overhangs. Together, our results suggest that V-C Cas1 has an internal ruler mechanism similar to Cas1—Cas2 complexes, but it favors shorter length substrates.

2.3.3 Promiscuous DNA integration by V-C Cas1

The Cas1—Cas2 integrase from other CRISPR systems catalyzes DNA insertion selectively at the leader-proximal repeat in the CRISPR array[24,68,80,81]. In the type V-C CRISPR system under investigation here, two CRISPR arrays flank the two cas genes in the host genome (Fig. 2.1a). Based on sequence comparisons, we were unable to determine whether one array is preferred and which side of each array contains the leader sequence that directs new acquisition events. To examine the preferred integration sites for type V-C Cas1, we carried out high-throughput sequencing of the plasmid integration products[24]. To prepare the sequencing library, integration reactions were conducted with the type V-C Cas1, an 18-bp protospacer, and either pUC19 (a control lacking a CRISPR array) or a plasmid bearing shortened versions of both natural V-C CRISPR arrays (Table 2.2). After fragmentation of the integration products, an extension with a barcoded protospacer-specific primer was carried out to enrich for fragments containing integration sites (Fig. 2.4a). Following removal of PCR duplicates, we mapped ~2.8 million reads to pUC19 and the CRISPR array-bearing plasmid (Fig. 2.4b,b; Fig. 2.5a,b).

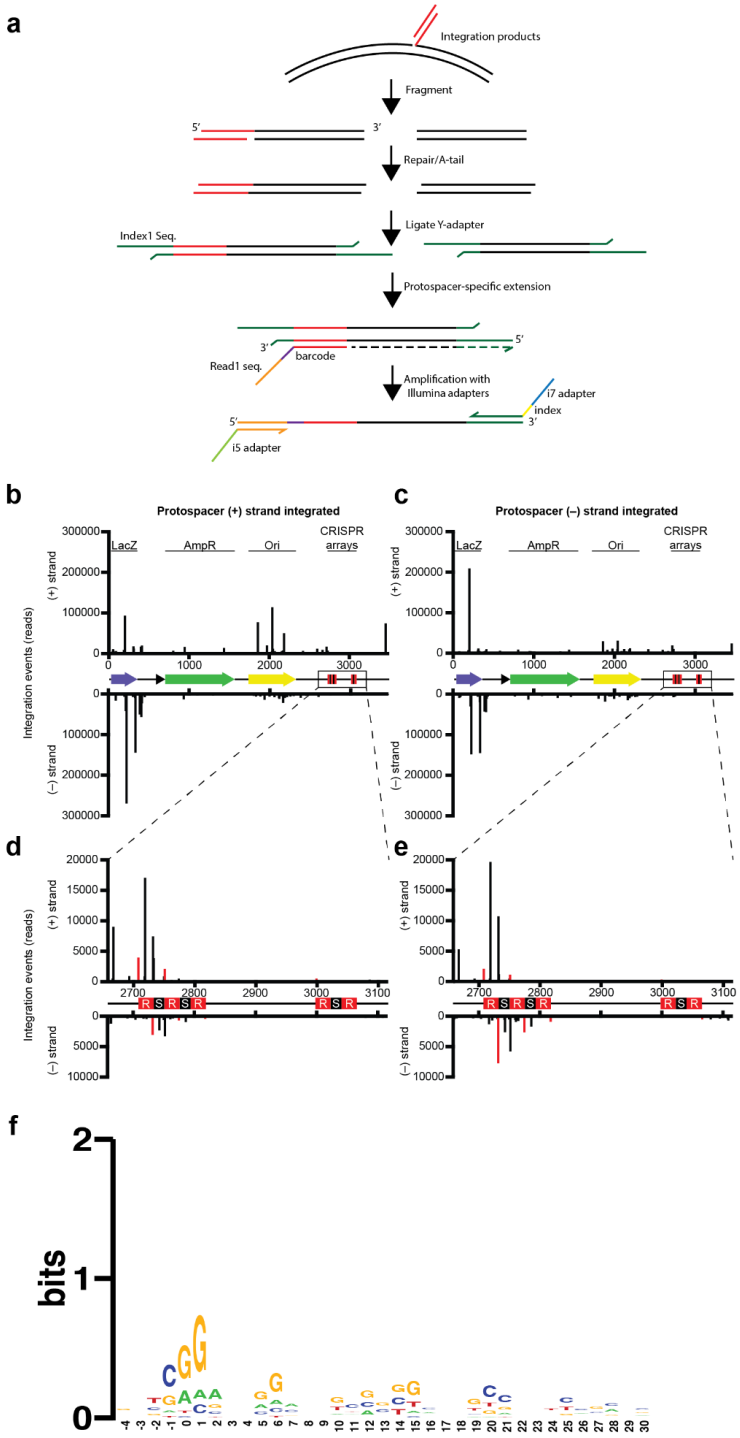


Figure 2.4 | Integration of protospacers occurs at many off-target sites. (a) Schematic of library preparation protocol for high-throughput sequencing. Integration products are fragmented, end-repaired, A-tailed, and ligated with Y-adapters. A protospacer-specific extension is carried out before amplification with Illumina primers. **(b,c)** Integration sites along pCRISPR. Results are separated based on the orientation of the protospacer that is integrated: the plots show integration by the (+) strand of the protospacer **(b)** and the (-) strand of the protospacer **(c)**. **(d,e)** Magnified view of integration in the

CRISPR arrays by the (+) strand of the protospacer **(d)** and the (−) strand of the protospacer **(e)**. See Table 2.2 for nucleotide sequences. **(f)** WebLogo for integration sites in pCRISPR. The first nucleotide following integration is numbered 1, with the upstream nucleotides labeled -4 through 0.

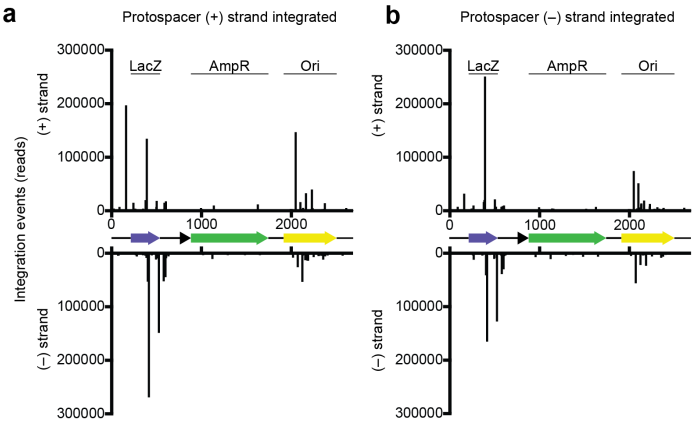


Figure 2.5 | Integration sites along pUC19. (a,b) Integration sites along pUC19 for protospacer (+) strand **(a)** and (–) strand **(b)**.

Surprisingly, most integration events did not occur at the CRISPR locus. Instead, we observed frequent off-target integration in the lacZ gene and the plasmid origin for both pUC19 and pCRISPR (Fig. 2.4b,c; Fig. 2.5a,b). The repeat borders did support levels of accurate integration above background, but several off-target sites within the repeats were even more frequently used (Fig. 2.4d,e). The repeat-specific integration was largely restricted to the upstream (relative to the cas1 gene) array and, in particular, the first repeat of the array, suggesting that this might be the preferred array and that the upstream sequence may be the leader (Fig. 2.1a). The overall lack of specificity suggests that a host factor or possibly sequence elements outside of the available assembled sequence are required for specificity. A sequence logo generated from all pCRISPR integration sites revealed a strong preference for G in both the −1 and +1 position relative to the integration site, while the sequence at the −1 – +1 site of the first repeat is “TC,” further suggesting that V-C Cas1 relies on factors other than its intrinsic sequence preference to recognize the repeat *in vivo* (Fig. 2.4f). We tested the ability of V-C Cas1 to insert fragments into short DNA targets containing either a single repeat or the sequence surrounding the dominant off-target peaks in lacZ and in both cases observed many products, consistent with a lack of strict selection for a target site (Fig. 2.6).

We observed trends in the sequencing data indicating that V-C Cas1 integrates DNA substrates with biased orientation. For repeat-specific integration events, the (+) strand of the protospacer (CTCTCCGAGGCCAGCGTG) was integrated primarily in the (+) strand of the plasmid, likely corresponding to leader-side integration, while integration of the (–) protospacer strand was biased toward spacer-side integration (Fig. 2.4d,e). We also noted paired integration peaks at the dominant off-target site in lacZ,

with a spacing and strand orientation consistent with full-site integration. Data at this site also suggested biased orientation of the inserted sequence, with most protospacer (+) strand integration events occurring in the (–) strand of the plasmid (Fig. 2.4b,c).

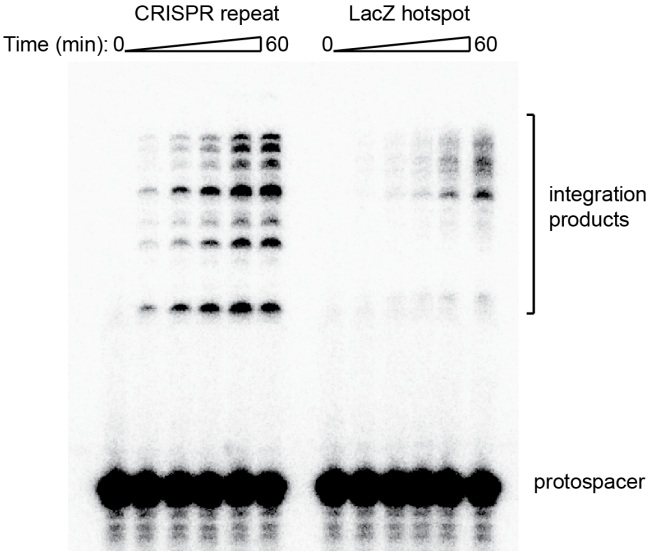


Figure 2.6 | Integration of Protospacer into CRISPR Repeat and lacZ Hotspot. Integration reactions are conducted with radiolabeled protospacer and target plasmids containing a CRISPR repeat or the lacZ hotspot identified from high-throughput sequencing. See Table 2.2 for nucleotide sequences.

2.3.4 Type V-C Cas1 catalyzes full-site integration

Given that low levels of half-site integration have been observed for Cas1 alone in at least one canonical Cas1—Cas2 system, we investigated whether type V-C Cas1 could complete the full integration reaction alone[81]. Since bulk integration assays and sequencing data cannot differentiate between half-site and full-site integration events and previous full-site assays require precise integration to be interpretable, we devised a different method to test whether V-C Cas1 can perform full-site integration[26]. We used a selection based on chloramphenicol resistance to recover plasmids with full-site integration events in the target region[84]. A reporter plasmid was designed with the 5' CRISPR repeat and putative leader sequence immediately upstream of an out-of-frame chloramphenicol resistance gene (CmR + 2) such that the open reading frame is restored by the 43 bp insertion resulting from full-site integration (18-bp spacer + duplication of a 25-bp repeat) (Fig. 2.7a). *In vitro* integration reactions were carried out with this reporter plasmid and an 18-bp protospacer, and the products were electroporated into *E. coli* (Table 2.2). Transformants were plated on media containing chloramphenicol, and surviving colonies were sequenced to confirm full-site integration and determine the location of integration sites.

The results from the selection show that type V-C Cas1 carries out full-site integration *in vitro*. The sequencing alignments show insertion of the expected 18-bp

spacer and duplication of the 25 bp directly adjacent to the spacer, which is consistent with full-site integration in other CRISPR systems. The locations of the integration events are scattered near the leader-repeat junction, with peaks at the expected integration site as well as 15, 27, and 36 nucleotides upstream (Fig. 2.7b,c). The full-site integration results appear to show a strong bias for protospacer orientation, with all integration events at the leader-repeat junction having the protospacer oriented with the TT end toward the leader (Fig. 2.7c). To further explore the potential sequence dependence of the protospacer orientation, we varied the 3' terminal nucleotides of the protospacer and also tested another protospacer with a completely randomized sequence (Table 2.2). Of the new protospacers designed, only the protospacer with the TA 3' end has the possibility to generate a stop codon upon integration and was therefore excluded from our analysis. For all of the other protospacers, both orientations result in an open reading frame. For all of the tested 3' ends, integration at the repeat occurred almost exclusively with the same relative orientation as observed for the initial protospacer (Fig. 2.8). Different 3' end nucleotides led to different apparent orientation preferences at off-target sites, but the orientation bias at the repeat appears to depend on the internal sequence rather than terminal nucleotides. The randomized sequence, however, resulted in no integration events at the repeat. The orientation of integrated spacers is critical *in vivo* for ensuring that the resulting crRNA targets a protospacer with a PAM on the correct side[85–87]. This data suggest that V-C Cas1 has an intrinsic mechanism to correctly orient protospacers during integration, similar to what has been observed for the *E. coli* integrase, though it is unclear how this mechanism functions in the native context to ensure fidelity[16]. Surprisingly, the 3' terminal nucleotides do appear to influence specific target recognition of the leader-repeat junction. The protospacers with the AA + TT 3' ends and TC + AG 3' ends were integrated at the leader-repeat junction in more than 40% of integration events on average, while some other protospacers were properly integrated in less than 5% of integration events (Fig. 2.8). The relevance of this variable specificity for the *in vivo* reaction is unclear, but it further supports the model that the protospacer sequence affects the integration reaction.

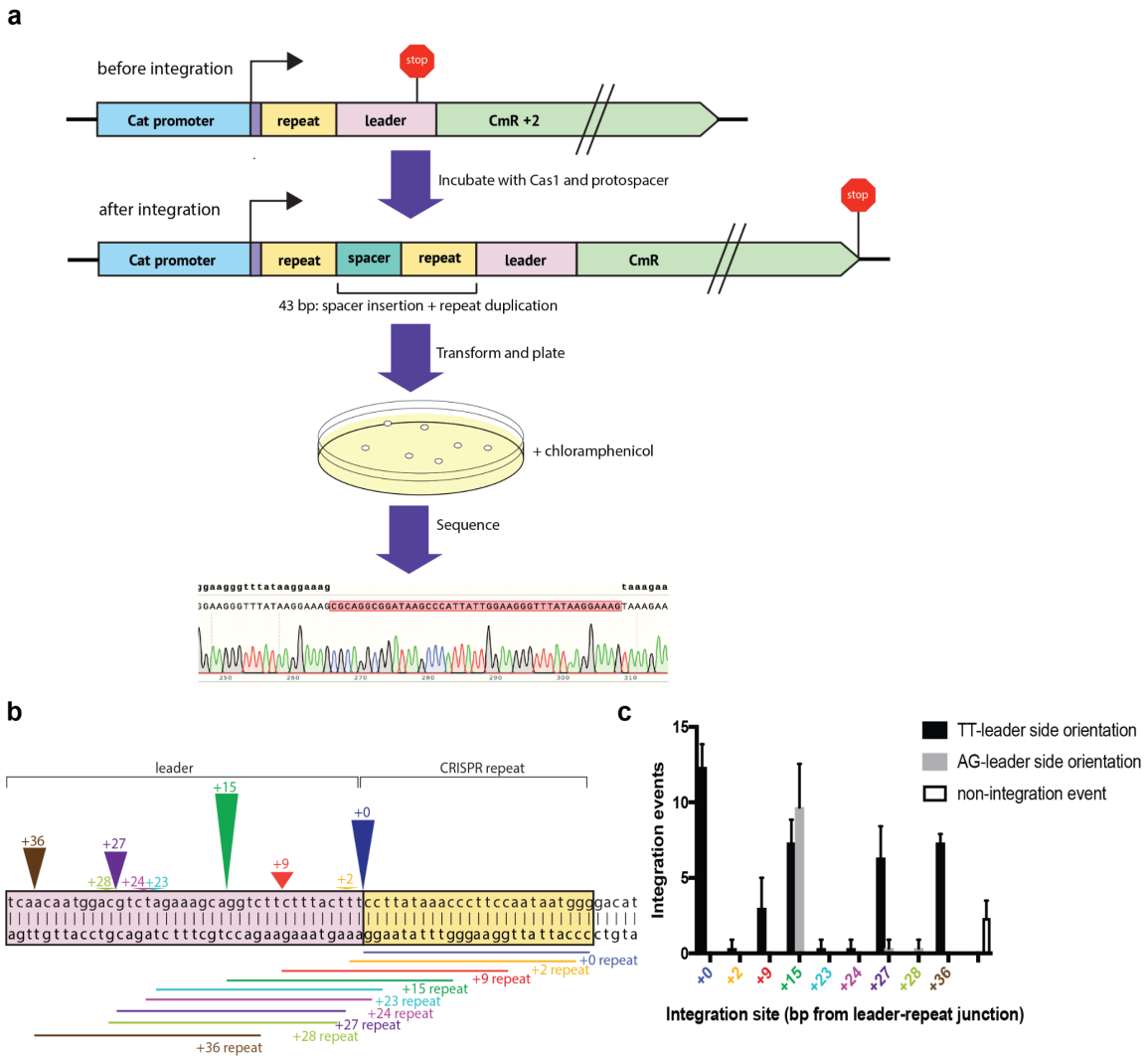


Figure 2.7 | Detection of full-site integration by type V-C Cas1. (a) Schematic representation of the chloramphenicol resistance turn-on screen to detect full-site integration near the leader-repeat junction. The construct contains a CRISPR repeat and leader upstream of an out-of-frame chloramphenicol resistance gene (CmR + 2). Translation of the transcript generated by Pcat begins upstream of the repeat (arrow) and ends in the leader (stop sign). Full-site integration of an 18-nt protospacer restores the open reading frame for the CmR coding sequence. Transforming and plating on chloramphenicol plates allows for positive selection of clones that have the inserted spacer. Sanger sequencing is used to confirm full-site integration. **(b)** Visual representation of identified full-site integration events near the leader-repeat junction. The colored arrowheads designate the position of the integration sites, with the number written above the arrowhead representing the number of base pairs from the leader-repeat junction. The corresponding colored lines designate the sequence that is duplicated upon spacer insertion. The arrowhead height is scaled to the total integration events at that site. **(c)** Number of integration events with specified spacer orientation at each integration site. The number of non-integration events (as a result of deletions) is also indicated. The mean and standard deviation of three independent replicates are shown. See Table 2.2 for nucleotide sequences.

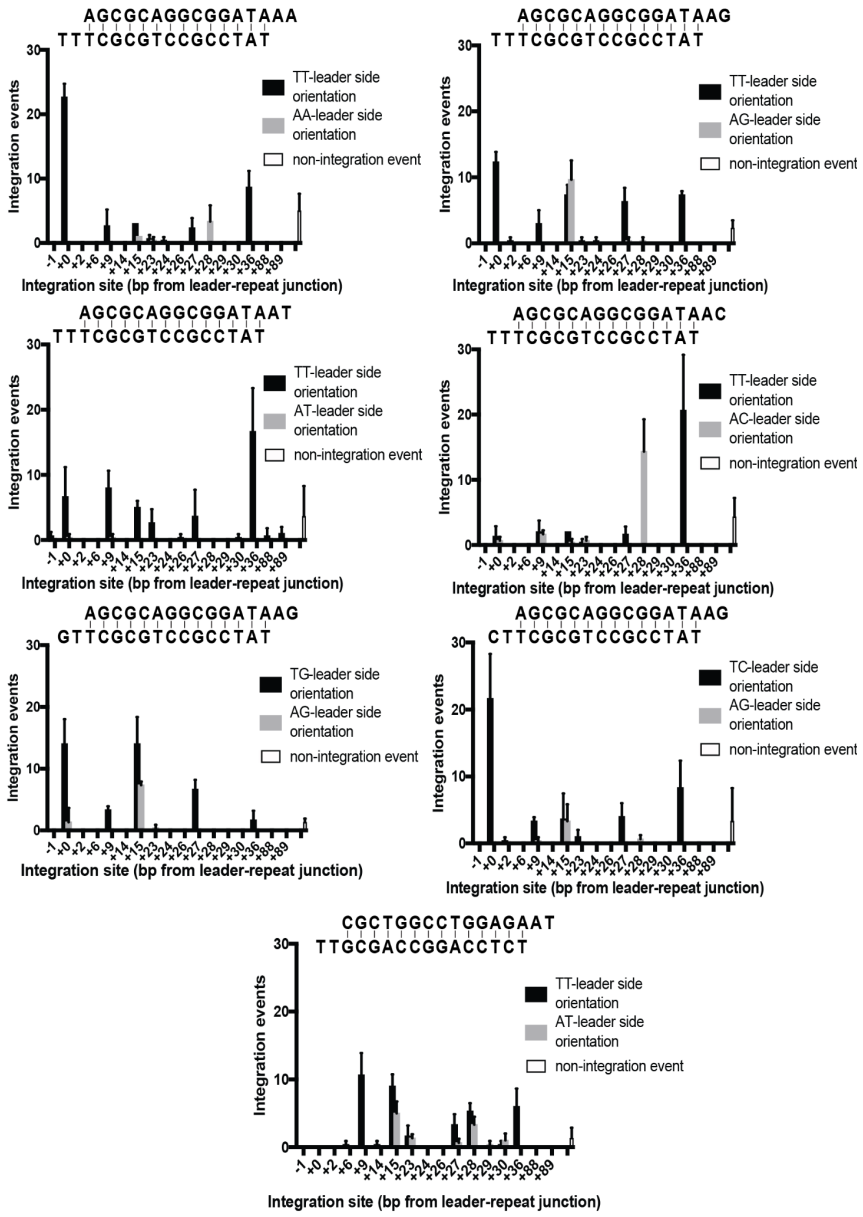


Figure 2.8 | Protospacer sequence influences orientation during full-site integration. Graphs show the number of integration events with specified spacer orientation at each integration site. The protospacer sequence is listed above each graph. The number of nonintegration events (as a result of deletions) is also indicated. Mean and standard deviation of three independent replicates are plotted. See Table 2.2 for nucleotide sequences.

2.3.5 The Cas1 mini-integrase forms a tetramer for integration

Given that Cas1—Cas2 integrases function as heterohexameric complexes containing four copies of Cas1 and two copies of Cas2, we wondered whether the V-C Cas1 integrase also functions as a multimeric complex[16,27]. While V-C Cas1 presumably associates with the protospacer in the conditions used for the integration reaction, the

salt concentrations necessary to maintain solubility in the presence of protospacer prevent stable association for analytical measurement of the complex. When run on a size-exclusion column, the protospacer and protein are observed to elute as separate species (Fig. 2.9a). In an effort to capture the active form of the type V-C Cas1, we designed a full-site product mimic (pseudo-full-site substrate) and a half-site product mimic that might result in a more stable complex[18]. The pseudo-full-site substrate includes a section of the leader sequence, the repeat, and a protospacer with a break in the middle to give the integrase access to the repeat. The half-site substrate has only one end of the protospacer covalently connected at the leader-repeat junction (Table 2.2).

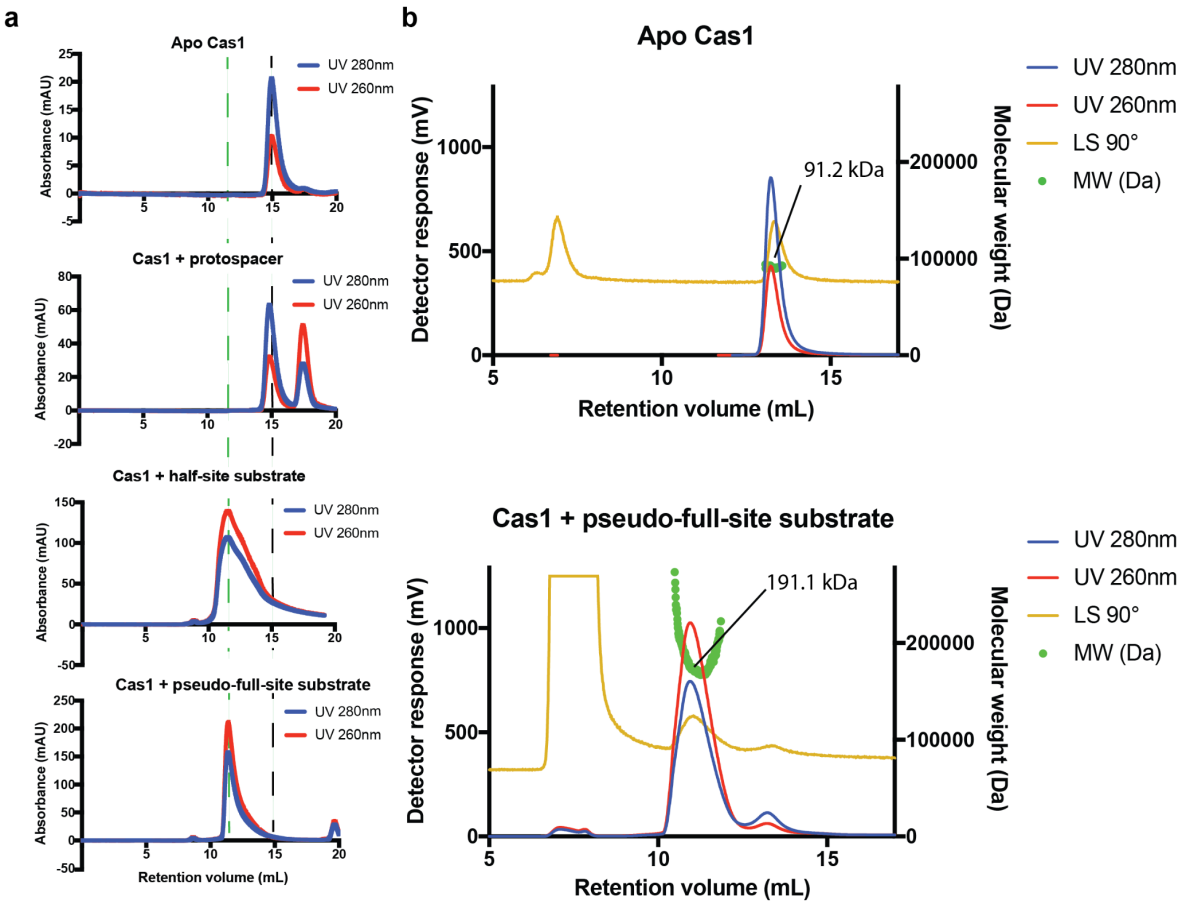
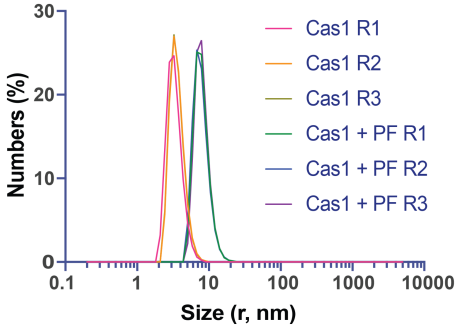


Figure 2.9 | Type V-C Cas1 forms a multimeric complex for integration. (a) Chromatograms of size-exclusion runs of type V-C Cas1, Cas1 with protospacer, Cas1 bound to half-site substrate, and Cas1 bound to pseudo-full-site substrate. Dashed black line indicates elution volume of free Cas1 dimer. Dashed green line indicates elution volume of Cas1 complex bound to full-site substrate. **(b)** Molecular weight characterization of the apo Cas1 and Cas1 bound to pseudo-full-site substrate by size exclusion chromatography coupled with dual angle light scattering. The experimental Mp for each peak is indicated. See Table 2.2 for nucleotide sequences.

Size-exclusion chromatography suggests that V-C Cas1 elutes as a dimer when in isolation but elutes as a larger multimer when bound to either the pseudo-full-site or the half-site substrate (Fig. 2.9a,b). The shift in size after association with DNA is further supported by dynamic light scattering results. The average radius calculated across three runs was 3.8 ± 0.8 nm for the apo V-C Cas1 sample and 8.2 ± 2.0 nm for the V-C Cas1-pseudo-full-site substrate complex, with a polydispersity index between 0.39 and 0.45 (Fig. 2.10). These results show that in the presence of pseudo-full-site substrate, V-C Cas1 forms a complex that is a little more than twice the size of its apo state. Dual angle light scattering analyses of the apo Cas1 and the Cas1 bound to pseudo-full-site substrate reveal peak molecular weights (91.2 and 191.1 kDa) that correspond closely to free dimer and tetramer bound to DNA (86 and 207 kDa), respectively (Fig. 2.9b). The intense light scattering peak at 7 mL is due to small amounts of aggregate in the void. Together, these results are consistent with the conclusion that the apo V-C Cas1 primarily exists as a dimer but, in the presence of the pseudo-full-site DNA substrate, and to a lesser extent the half-site substrate, forms a tetramer.



Sample	Mean size (r, nm)	PdI
Cas1 R1	3.8 ± 0.8	0.39
Cas1 R2	3.8 ± 0.8	0.40
Cas1 R3	3.9 ± 0.7	0.45
Cas1 + PF R1	7.7 ± 1.9	0.34
Cas1 + PF R2	8.8 ± 1.9	0.42
Cas1 + PF R3	8.0 ± 2.1	0.43

Figure 2.10 | Dynamic light scattering size measurements of apo Cas1 compared to Cas1 with added pseudo-full-site substrate. Measurements were collected from three runs. The mean size with standard deviations and PdI for each sample run are provided below the distribution plots. See Table 2.2 for nucleotide sequences.

To confirm whether the multimer was indeed a tetramer bound to a single DNA substrate, we used native nanoelectrospray ionization mass spectrometry (nanoESI-MS) measurements. Apo Cas1 and the Cas1-DNA complex were initially cross-linked before measuring by nanoESI-MS. While the predominant detected species was cross-linked dimer for both samples, we observed tetramer in both samples as well, with much more cross-linked tetramer in the Cas1-DNA complex sample (Fig. 2.11a,b). This suggests V-C Cas1 may form a tetramer in the absence of

DNA, but the presence of DNA favors additional tetramer formation (Fig. 2.11c). A possible explanation for the significant amount of cross-linked dimer and free DNA detected in the Cas1-DNA sample is that the two Cas1 dimers in the tetramer may not support efficient cross-linking at the interface and are only weakly associated, resulting in disassembly during buffer exchange to the nanoESI-MS buffer or during the nanoESI-MS measurements.

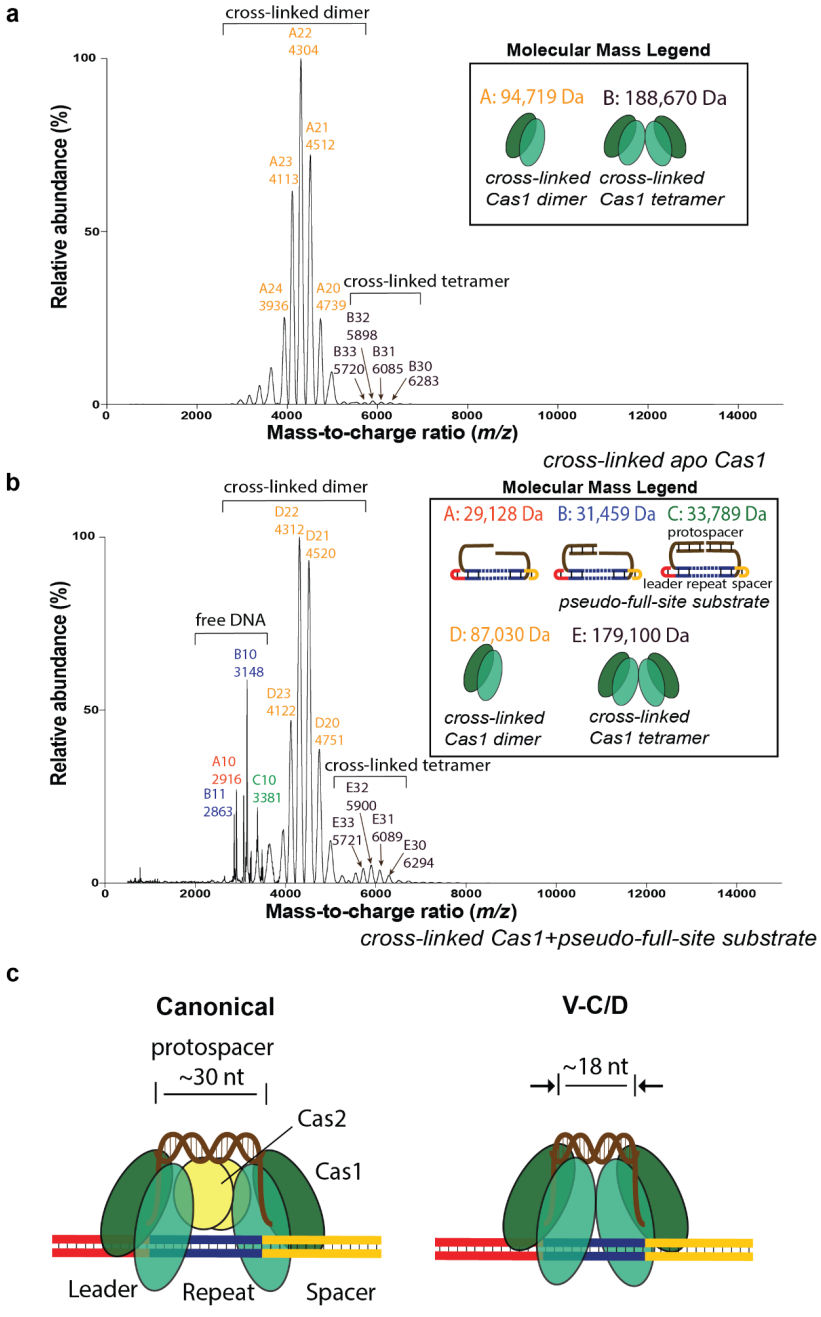


Figure 2.11 | Detection of type V-C Cas1 tetramer by native mass spectrometry. (a) NanoESI mass spectra of cross-linked samples of apo Cas1. Measured molecular masses and corresponding cartoons of

each species are listed on the top right corner. Ions of different species are labeled with different colors. **(b)** NanoESI mass spectra of cross-linked samples of Cas1 after complexing with pseudo-full-site substrate. **(c)** Comparison between type V-C mini-integrase and canonical Cas1—Cas2 integrase. See Table 2.2 for nucleotide sequences.

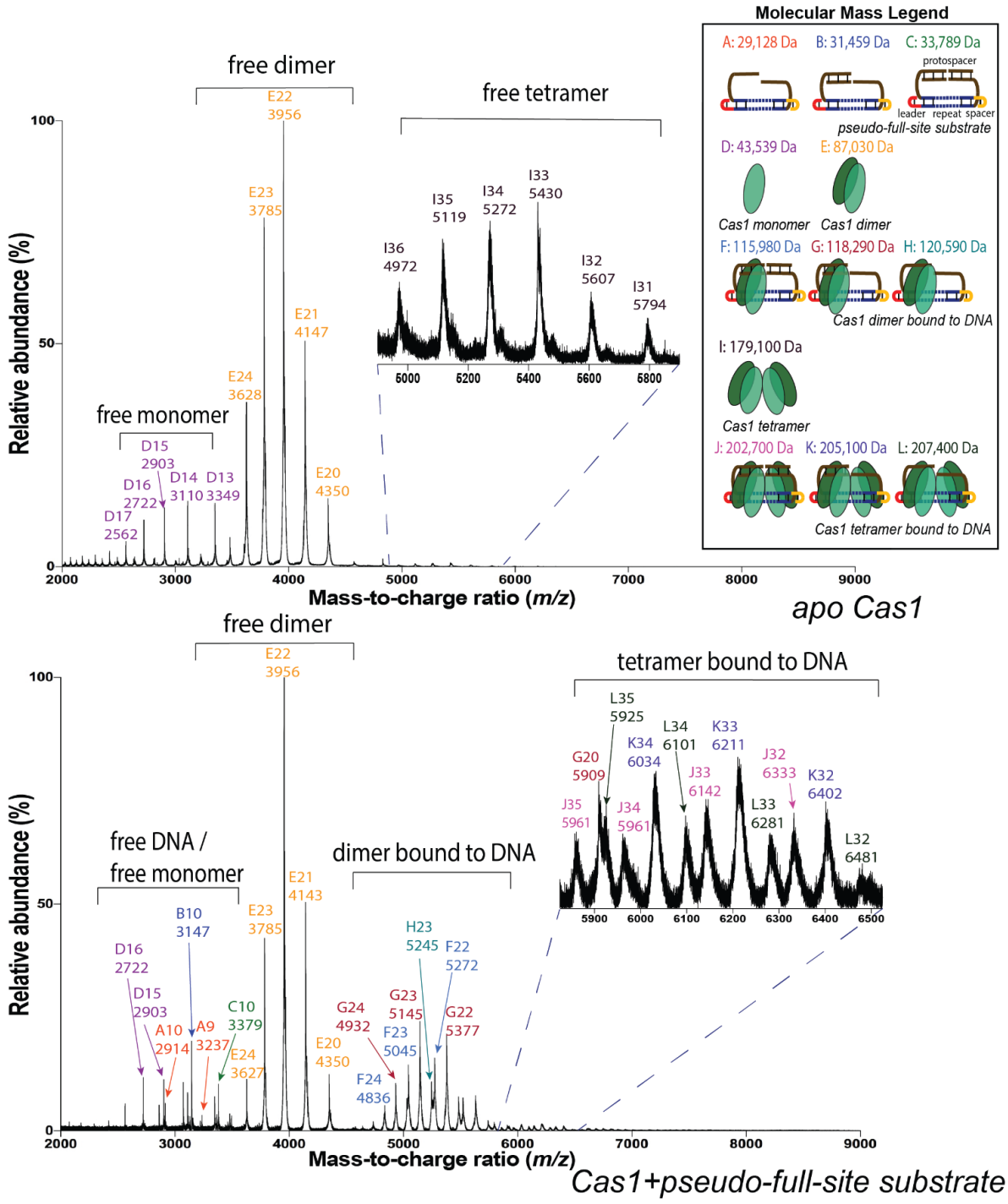


Figure 2.12 | Native mass spectra of apo V-C Cas1 and V-C Cas1 complexed with pseudo-full-site

substrate in native state. The molecular masses of the detected species are listed on the top right. The ions of different species are labeled with different colors. See Table 2.2 for nucleotide sequences.

In an effort to capture the tetramer bound to DNA in its native state without cross-linking, we also performed native MS on complexes purified by size-exclusion chromatography. For the Cas1-DNA complex, we detected masses that correspond exactly to tetramer bound to a single copy of pseudo-full-site substrate (Fig. 2.12). Observed in greater abundance were free dimer, dimer bound to a single copy of pseudo-full-site substrate, free monomer, and free DNA, which again suggests that the Cas1 tetramer-DNA complex disassembles either during the buffer exchange to the nanoESI-MS buffer or during the nanoESI-MS measurements. A small amount of tetramer was also observed for the apo Cas1 sample, but the amount detected was lower than that of tetramer bound to pseudo-full-site substrate in the DNA-bound sample. Notably, no free tetramer was detected from the DNA-containing sample. Together with the size-exclusion and light scattering results, the data suggest that there exists an equilibrium between the dimer and tetramer in solution that is biased toward the dimer in the apo state but, in the presence of the pseudo-full-site DNA substrate, shifts to a tetramer.

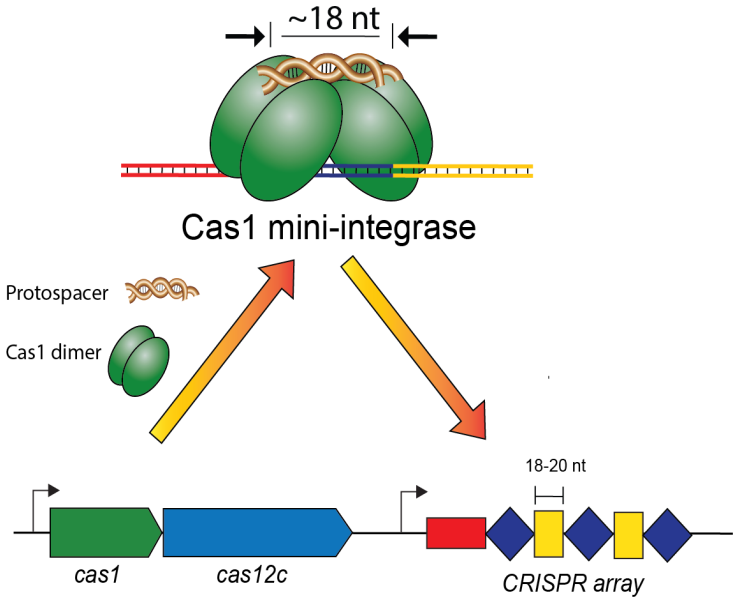


Figure 2.13 | Model for type V-C mini-integrase. The mini-integrase from V-C systems, comprised of a single protein Cas1, inserts unusually short fragments of viral DNA into the bacterial genome as part of the CRISPR-Cas adaptive immunity pathway. This minimal integrase may represent the function of the ancestral Cas1, prior to Cas2 adoption.

Together, our results support a model in which the type V-C Cas1 mini-integrase tetramerizes for integration, providing the architecture for a flexible internal ruler mechanism that favors shorter DNA integration substrates (Fig. 2.13). For the

Cas1—Cas2 integrases, the Cas2 protein dimer at the center of the integrase complex acts as a bridge between the two catalytic Cas1 dimers (Fig. 2.11c). A condensed tetramer structure without a Cas2 bridge explains why the type V-C integrase prefers ~18-bp protospacers rather than the longer protospacers that are preferred by the other Cas1—Cas2 integrases[10,16].

2.4 Discussion

Type V-C and V-D CRISPR systems are unique among microbial adaptive immune systems in having unusually short CRISPR array spacers and repeats and lacking a cas2 gene previously thought to be essential for CRISPR sequence acquisition[4,31]. Although both Cas1 and Cas2 are essential for new DNA integration in most CRISPR systems, we demonstrate that type V-C and V-D Cas1 alone can carry out integration of DNA substrates, including full-site integration in the case of V-C Cas1. Furthermore, type V-C Cas1 appears to have an intrinsic mechanism to orient protospacers during integration, though further work is required to understand how this activity cooperates with protospacer processing to result in PAM-specific orientation *in vivo*. Our experiments revealed that type V-C Cas1 has many off-target integration sites, even for full-site integration, which typically has greater substrate specificity than half-site integration in other CRISPR systems[68]. However, we have reason to suspect that varying the 3’ end nucleotides of the protospacer can yield more specific target recognition. Our results strongly suggest that additional host factors or mechanisms are involved either in the protospacer selection or during integration to achieve specific target recognition *in vivo* and avoid deleterious off-target integration.

Our findings are consistent with a model in which the V-C Cas1 minimal integrase forms a tetramer for the integration of new protospacers. Without a Cas2 dimer acting as a bridge, the tetramer model for the type V-C Cas1 would result in a shortened internal ruler, thus explaining the preference for short protospacers of ~18 bp. While V-C Cas1 exists predominantly as a dimer in solution, binding to the protospacer might stabilize transient tetramerization events to produce an active complex, resulting in the tetramer observed in the intermediate- and product-bound complexes.

Although it is possible that the type V-C and V-D systems are the unique outcome of a loss event of the cas2 gene, a recent study examining the Cas1 phylogeny indicated that type V-C and V-D Cas1 proteins may represent an important stage in the evolutionary history of CRISPR-Cas adaptation[33]. V-C and V-D Cas1 genes were found to form a branch along with solo Cas1 genes, found in the absence of apparent CRISPR systems or casposons, that was rooted near the casposon branch. This phylogeny is consistent with a model wherein the type V-C and V-D Cas1 proteins represent an extant version of an ancestral Cas1 that evolved from the casposase, which has integrase activity and similar target specificities as CRISPR-Cas spacer acquisition and also forms a tetramer of Cas1s for integration[14,15,20]. If this is correct,

then the first CRISPR systems likely appeared much like extant V-C/D systems, with an effector protein or complex alongside an isolated Cas1 integrase, as opposed to the previous model wherein Cas2 was a part of the ancestral casposon prior to the evolution of a functional CRISPR system[88]. We propose that this minimal integrase marks the emergence of the ruler mechanism that is characteristic of CRISPR integrases and not observed in the tetrameric casposase[15,17,46], although further structural characterization is required to elucidate the structural basis for this emergence. This minimal integrase could have later adopted Cas2, potentially from a toxin-antitoxin system, for greater structural stability, resulting in the canonical Cas1—Cas2 integrase complex used by type I and II systems[88]. The addition of Cas2 to the complex also allows for the acquisition of longer spacers, which could in turn allow for greater targeting specificity of crRNA-effector complexes, giving systems bearing a Cas1—Cas2 integrase an advantage. A greater knowledge of the structure and mechanisms of these different integrases gives us more insight into the evolutionary history of CRISPR adaptation modules and their unique functions.

2.5 Materials and methods

2.5.1 Cloning / protein purification

CRISPR loci from the Cas12c system from mouse cecum (ORF Ga0073908_100052011 from IMG genome 3300005460) and the Cas12d system from Passlidae beetle gut (ORF IMNBL3_1000107225 from IMG genome 3300000114) were ordered as G-blocks. Protein-coding genes were codon-optimized for *E. coli* expression and arrays were reduced. To generate target plasmids, arrays and flanking intergenic sequences were amplified by PCR and inserted into a pUC19 backbone by Gibson Assembly. Cas1 genes were PCR amplified and inserted into a pET-3a plasmid with an N-terminal 10xHis-MBP-TEV tag. Proteins were expressed in BL21(DE3) Star cells. Cells were grown to an OD600 of ~0.6 and induced overnight at 16°C with 0.5 mM isopropyl-β-D-thiogalactopyranoside (IPTG). Cells were harvested and resuspended in lysis buffer (20 mM HEPES, pH 7.5, 1 M NaCl, 10 mM imidazole, 0.1% Triton X-100, 2 mM Tris (2-carboxyethyl)phosphine (TCEP), Complete EDTA-free protease inhibitor (Roche) and 10% glycerol. Cells were lysed by sonication, and the lysate was cleared by centrifugation. The supernatant was incubated on Ni-NTA resin (Qiagen), and the resin was washed with buffer containing 20 mM HEPES, pH 7.5, 1 M NaCl, 10 mM imidazole, 1 mM TCEP, and 5% glycerol. The protein was eluted with wash buffer supplemented to 300 mM imidazole. The protein was digested with TEV protease overnight. The salt concentration was reduced by dilution to 400 mM NaCl, the cleaved MBP tag was removed with an MBPTrap column (GE Healthcare), and the protein was bound to a HiTrap heparin HP column (GE Healthcare). The protein was eluted with a

gradient from 400 mM to 1 M NaCl. The proteins were further purified on a Superdex 200 (16/60) column with Storage Buffer (20 mM HEPES, pH 7.5, 500 mM KCl, 1 mM TCEP, and 5% glycerol).

Source	Identifier	Sequence
Passalidae beetle gut sample Sequenced at the JGI in 2012. Public data. PI: David Sherman	**Taxon ID:** 3300000114 **Scaffold ID:** IMNBL3_c10001072 **Gene ID:** IMNBL3_1000107225 **GOLD Analysis Project ID:** Ga0026078	MAHRIPLWLPYLSEVNVSASGVIHYVYKGGEDSVRLKDISS IMIYGETDASFSFKTFEKISRAGIPVIYTRRNIGSPMIIHGGL RYDAVDTVTAQILARENKKKQCHIARQLLKAKMNSMKHLA PVFELPDNANINELRNIEAKHAKKYWREYYTYIGHPEWTR RGEETGKNPVSDTLNLTSKFICGIILRWTLYHHLSPSHGYF HVMTDYQSLVYDLIEPYRALFDIELIRLFKDNAHLEENWPG MAINLMKNKLNKKCYCPLTRQVVTYHELTHGIVLSLKSYLIG RQRKFHIPLPGSPNGGRPSKVEFKLYARNAGKTDFWKVA KEVADLDFACLSVG*
Mouse cecum sample Sequenced at the JGI in 2016. Public data. PI: Michael Wagner	**Taxon ID:** 3300005460 **Scaffold ID:** Ga0073908_1000520 **Gene ID:** Ga0073908_100052011 **GOLD Analysis Project ID:** Ga0073908	MDQGNQTIENQTINCDQHPDFLWTWKSNKRGSRVSVWL PYFSQAKKIPRSKKWSVAYNGGSIEFDLKETDLIMFYGATG ELPLEFLDDASKNGVMILIHRRNVLQPYVFYPSVIGDEEDIL TKQIQFRTNERKRLYIAKTLIKKRLENMGSTIPISAPLLRQLS AAKSIDEVRAIEANTTARYWNKWYENLNIETTRRKDHPINS ALDAGSKFIYGVILRWLVFHRFSPNHGFMHQPTSYPSLVY DLMEPFRYMIENVCSAAWKRGERENSKIVALSLSFLKEEL DKPCYVPATRQYVRKKNLLHGAVLALRSYLIGDMRKLVFP SEGVPNGGRPIKASYKLPGSMYDVG RKPPEIKQKDEICFDEVSQEESEE*

Table 2.1 | IMG genome sequences of type V-C Cas1 from mouse cecum and V-D Cas1 from beetle gut.

2.5.2 Integration assays

All integration assays were performed in buffer containing 20 mM HEPES, pH 7.5, 25 mM KCl, 10 mM $MgCl_2$, 1 mM DTT, 0.01% Nonidet P-40, and 10% DMSO and carried out at room temperature unless otherwise noted. Pre-complexing of the *E. coli* Cas1—Cas2 was carried out as previously described[24]. For experiments with unlabeled protospacer, 50 nM protospacer was added to 250 nM Cas1 (or Cas1—Cas2) and incubated at room temperature for 15 minutes, followed by addition of target plasmid (pVC_CRISPR1 or pVD_CRISPR) to 20 ng/μL (~10 nM). For the initial experiment alongside *E. coli* Cas1—Cas2, the reaction was carried out at 37°C for one hour. The protospacer length variation experiment with unlabeled substrates was carried out at room temperature for 30 minutes. For all assays using unlabeled substrates, reactions

were quenched by the addition of 0.4% SDS and 25 mM EDTA and analyzed on a 1% ethidium bromide-stained agarose gel. Quantifications were made using Image Lab (BioRad). Integration of radiolabeled protospacer into plasmid by V-D Cas1 was carried out with 400 nM Cas1, 1 nM protospacer that was previously 5' labeled with [γ−32P]ATP (PerkinElmer) and T4 polynucleotide kinase (New England Biolabs), and 20 ng/μL plasmid at 37°C for one hour. Products were separated on a 1.5% agarose gel. The gel was dried and visualized through phosphorimaging. Integration of radiolabeled protospacer into oligonucleotide targets by V-C Cas1 was carried out at room temperature with 100 nM Cas1, 10 nM 5'radiolabeled protospacer, and 100 nM double-stranded target DNA. Reactions were quenched by the addition of an equal volume of 95% formamide and 50 mM EDTA, and products were separated on 12% urea-PAGE. The gel was dried and imaged with phosphorimaging. For the fluorescent protospacer experiments, one strand of each protospacer was ordered with a 5' 6-carboxyfluorescein (FAM) attachment and hybridized with its complementary unlabeled strand. Integration of fluorescent protospacers was carried out at room temperature with 250 nM Cas1, 10 nM protospacer, and 20 ng/μL plasmid for 1 hour. Reactions were quenched by the addition of 0.4% SDS and 25 mM EDTA and treated with proteinase K for 30 minutes at room temperature before analysis on a 1.5% agarose gel. The gel was imaged with a Typhoon FLA gel imaging scanner.

Fluorescent bands were quantified using ImageQuantTL. The fraction integrated was calculated as the ratio of the integration product fluorescence intensity to the total intensity of the free fluorescent protospacer and the integration product. When examining integration by open-circle formation, gel bands were visualized by ChemiDoc MP (BioRad) and quantified using Image Lab (BioRad). The fraction integrated was calculated as the ratio of the open-circle integration product band intensity to the total intensity of both the open-circle band and the supercoiled plasmid band. Statistical analyses were performed using Prism GraphPad. Data were presented as the mean ± SD (error bars) of three independent experiments.

Name/Description	Sequence	Figure
15mer protospacer	GCGCAGGCGGATAAG	2.3
Fluorescent 15mer protospacer	(5’ 6-FAM) GCGCAGGCGGATAAG	2.2
RC[1]	CTTATCCGCCTGCGC	2.2, 2.3
16mer protospacer	AGCGCAGGCGGATAAG	2.3
Fluorescent 16mer protospacer	(5’ 6-FAM) AGCGCAGGCGGATAAG	2.2
RC	CTTATCCGCCTGCGCT	2.2, 2.3
17mer protospacer	AAGCGCAGGCGGATAAG	2.3
Fluorescent 17mer protospacer	(5’ 6-FAM) AAGCGCAGGCGGATAAG	2.2
RC	CTTATCCGCCTGCGCTT	2.2, 2.3
18mer protospacer	AAAGCGCAGGCGGATAAG	2.3
Fluorescent 18mer protospacer	(5’ 6-FAM) AAAGCGCAGGCGGATAAG	2.2
RC	CTTATCCGCCTGCGCTTT	2.2, 2.3
18mer protospacer for sequencing (+) strand	CACGCTGGCCTCGGAGAG	2.4, 2.5, 2.6
RC ((–) strand)	CTCTCCGAGGCCAGCGTG	2.4, 2.5, 2.6
19mer protospacer	GAAAGCGCAGGCGGATAAG	2.3
Fluorescent 19mer protospacer	(5’ 6-FAM) GAAAGCGCAGGCGGATAAG	2.2
RC	CTTATCCGCCTGCGCTTTC	2.2, 2.3
20mer protospacer	AGAAAGCGCAGGCGGATAAG	2.3
Fluorescent 20mer protospacer	(5’ 6-FAM) AGAAAGCGCAGGCGGATAAG	2.2
RC	CTTATCCGCCTGCGCTTTCT	2.2, 2.3
21mer protospacer	AGAAAGTCGCAGGCGGATAAG	2.3
Fluorescent 21mer protospacer	(5’ 6-FAM) AGAAAGTCGCAGGCGGATAAG	2.2
RC	CTTATCCGCCTGCGACTTTCT	2.2, 2.3
22mer protospacer	AGAAAGTCGCAGGCTGGATAAG	2.3
Fluorescent 22mer protospacer	(5’ 6-FAM) AGAAAGTCGCAGGCTGGATAAG	2.2
RC	CTTATCCAGCCTGCGACTTTCT	2.2, 2.3
23mer protospacer	AGAAAGTCCGCAGGCTGGATAAG	2.3
Fluorescent 23mer protospacer	(5’ 6-FAM) AGAAAGTCCGCAGGCTGGATAAG	2.2
RC	CTTATCCAGCCTGCGGACTTTCT	2.2, 2.3
23mer protospacer with 3’ P modification	AGAAAGTCCGCAGGCTGGATAAG (3’P)	2.3
RC with 3’ P modification	CTTATCCAGCCTGCGGACTTTCT (3’P)	2.3
24mer protospacer	AGAAAGTCCGCAGGCATGGATAAG	2.3
Fluorescent 24mer protospacer	(5’ 6-FAM) AGAAAGTCCGCAGGCATGGATAAG	2.2

RC	CTTATCCATGCCTGCGGACTTTCT	2.2, 2.3
25mer protospacer	AGAAAGTCCGGCAGGCATGGATAAG	2.3
Fluorescent 25mer protospacer	(5' 6-FAM) AGAAAGTCCGGCAGGCATGGATAAG	2.2
RC	CTTATCCATGCCTGCCGGACTTTCT	2.2, 2.3
33mer protospacer	GCGAGAATTACTACTCGTTCTGGTGTTTCTCGC	2.2
RC	GCGAGAAACACCAGAACGAGTAGTAATTCTCGC	2.2
1-nt overhang protospacer	AAGCGCAGGCGGATAAG	2.3
RC	TTATCCGCCTGCGCTTT	2.3
2-nt overhang protospacer (G)	AGCGCAGGCGGATAAG	2.7, 2.9, 2.3, 2.8
RC (T)	TATCCGCCTGCGCTTT	2.7, 2.9, 2.3, 2.8
RC (A)	TATCCGCCTGCGCTTA	2.8
RC (C)	TATCCGCCTGCGCTTC	2.8
RC (G)	TATCCGCCTGCGCTTG	2.8
3-nt overhang protospacer	GCGCAGGCGGATAAG	2.3
RC	ATCCGCCTGCGCTTT	2.3
4-nt overhang protospacer	CGCAGGCGGATAAG	2.3
RC	TCCGCCTGCGCTTT	2.3
2-nt overhang protospacer (A)	AGCGCAGGCGGATAAA	2.8
2-nt overhang protospacer (C)	AGCGCAGGCGGATAAC	2.8
2-nt overhang protospacer (T)	AGCGCAGGCGGATAAT	2.8
Pseudo-full-site half-protospacer - repeat - spacer - hairpin - spacer (pseudo-full-site substrate)	GGCGGCGGACCCATTATTGGAAGGGTTTATAAGGCTCAC GAAGTGAG	2.9, 2.10, 2.11, 2.12
Pseudo-full-site half-protospacer - repeat - leader - hairpin - leader (pseudo-full-site substrate)	GGCGGCGGTCCTTATAAACCCTTCCAATAATGGGAGACT GAAAGTCT	2.9, 2.10, 2.11, 2.12
Pseudo-full-site half-protospacer	CCGCCGCC	2.9, 2.10, 2.11, 2.12
Half-site protospacer-repeat-spacer	ACGCTGGCCTCGGAGAGCCCATTATTGGAAGGGTTTATA AGGATCAA	2.9
Half-site protospacer	TCTCCGAGGCCAGCGTG	2.9
Half-site spacer-repeat-leader-hairpin-leader	TTGATCCTTATAAACCCTTCCAATAATGGGAGACTGAAAGT CT	2.9
CRISPR repeat target	AAGCAGGTCTTCTTTACTTTCCTTATAAACCCTTCCAATAA TGGGAGACTAGCAAATATCTATTTGAAAGTCAAT	2.6
RC	ATTGACTTTCAAATAGATATTTGCTAGTCTCCCATTATTGGA AGGGTTTATAAGGAAAGTAAAGAAGACCTGCTT	2.6
lacZ hotspot target	CAGGCTGCGCAACTGTTGGGAAGGGCGATCGGTGCGGG CCTCTTCGCTATTACGCCAGCTGGCGAAAGGGGGATG	2.6
RC	CATCCCCCTTTCGCCAGCTGGCGTAATAGCGAAGAGGCC CGCACCGATCGCCCTTCCCAACAGTTGCGCAGCCTG	2.6

Sequencing Y-adapter strand 1	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T	2.4, 2.5
Sequencing Y-adapter strand 2	GATCGGAAGAGCTGT	2.4, 2.5
(+)-strand protospacer-specific barcoding primer	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTAC ACGACGCTCTTCCGATCTNNWNNWNNCACGCTGGCCTC GGAGA*G	2.4, 2.5
(–)-strand protospacer-specific barcoding primer	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTAC ACGACGCTCTTCCGATCTNNWNNWNNCTCTCCGAGGCC AGCGT*G	2.4, 2.5

Table 2.2 | DNA oligonucleotides used in this study. Sequences are written 5' to 3'. RC represents reverse complementary strand of the previous oligonucleotide. Asterisk indicates phosphorothioate bond.

2.5.3 High-throughput sequencing

Integration reactions into pUC19 or pVC-CRISPR were carried out with 250 nM Cas1, 50 nM protospacer, and 20 ng/μL plasmid at room temperature for one hour. Reactions were quenched by the addition of phenol-chloroform-isoamyl alcohol, chloroform extracted, and ethanol precipitated. The products were purified with a DNA Clean & Concentrator column (Zymo) to remove free protospacer. Products were fragmented with dsDNA fragmentase (New England Biolabs) and end-repaired. A single round of extension with Taq polymerase was performed on the products with one of the two protospacer-specific primers. Free primer was removed with Exonuclease I (New England Biolabs, RRID:AB_2293772). Annealed adapters were ligated with T4 DNA ligase (New England Biolabs). Libraries were amplified with Q5 polymerase and NEBNext universal and index primers for Illumina (New England Biolabs), size-selected with AMPure XP beads (Beckman Coulter), and sequenced on an Illumina MiSeq with 150-nt single reads.

3'-adapter sequences were removed from reads with Cutadapt[89]. Reads with identical barcodes and read sequences were removed. The barcode and protospacer sequence were removed with Cutadapt, and the resulting reads were mapped to the appropriate plasmid with Bowtie[90]. Consensus integration sequences were generated with WebLogo from the 5 nucleotides upstream and 30 nucleotides downstream of an integration site, weighted by how many reads were measured at the integration site[91].

2.5.4 Full-site selection

The reporter plasmid construct pVC_CRISPR_Cat was designed by cloning a cat promoter and chloramphenicol resistance gene into a pUC19 plasmid followed by subsequent insertion of a sequence carrying 59 bp of the 5' leader from pVC-CRISPR, the adjacent CRISPR repeat, 2 bp of the spacer adjacent to the repeat, and a start codon by Gibson Assembly. Twelve 50 uL integration reactions were carried out as

described above with the reporter construct for 1 hr. The integration products were purified using the Qiagen MinElute PCR Purification Kit and electroporated into DH10B cells. Transformants were plated on LB agar containing chloramphenicol and 50 of the surviving colonies were sequenced. Numbers of integration events were presented as the mean ± SD (error bars) of three independent experiments with 50 sequenced colonies for each experiment.

2.5.5 Protein complex formation / light scattering / native nanoelectrospray ionization mass spectrometry

DNA substrates were made as previously described[18]. Cas1 and DNA substrates were complexed by mixing 50 μM Cas1 and 12.5 μM DNA half-site or pseudo-fullsite substrates in Storage Buffer and dialyzing in Complex Buffer (10 mM HEPES, pH 7.5, 5 mM EDTA, 250 mM KCl, and 1 mM TCEP) for 2 hours using the Slide-A-Lyzer MINI Dialysis Devices and concentrated to 100 μM Cas1.

To characterize the molecular weight of the peaks off size-exclusion by light scattering, the samples were run on a Superdex 200 (16/60) column attached to an Agilent 1260 Infinity Multi-Detector system. Light scattering was performed using a 658nm laser. The UV signals at 280 nm and 260 nm and the light scattering signal at 90° were collected. Data analysis and Mp calculations were performed using the Bio-SEC software (Agilent).

Dynamic light scattering measurements were collected on a Malvern Zetasizer Nano instrument. Cas1 and DNA substrates were complexed as described above in 35 μM final Cas1 concentration and 8.75 μM final DNA concentration before filtering with a 0.22 μm Corning SpinX Centrifuge Tube Filter. Three runs with 70 measurements per run were collected for each sample. The following parameters were set up: viscosity, 0.891 cP; temperature, 25°C; refractive index, 1.375[92]. Sizes were presented as the mean ± SD of three runs with 70 measurements per run collected for each sample.

To cross-link the protein samples for native nanoelectrospray ionization mass spectrometry, Cas1 and DNA substrates were complexed as described above by dialyzing in Complex Buffer using 10 μM Cas1 and 2.5 μM DNA half-site or pseudo-full-site substrates. The samples were then cross-linked with 1 mM BS(PEG)5 (PEGylated bis(sulfosuccinimidyl)suberate) final concentration for 30 minutes before quenching with 30 mM final concentration Quenching Buffer (1M Tris·HCl, pH 8.0) for 15 minutes. The cross-linked samples were then subjected to five rounds of buffer exchange in to Mass Spectrometry Buffer A (436 mM ammonium acetate, 13.5 mM ammonium bicarbonate, pH 7.5, and 5% glycerol) using 10,000 MWCO (molecular weight cut-off) Amicon Ultra-0.5 Centrifugal Filter Units and concentrated to 20 μM final Cas1 concentration. For non-cross-linked samples, Cas1 and pseudo-full-site substrate samples were complexed and purified as described above on a Superdex 200 (16/60)

column. The major peak was concentrated to an A280 of 3.0 using Nanodrop and subjected to five rounds of buffer exchange in to Mass Spectrometry Buffer B (242 mM ammonium acetate, 7.5 mM ammonium bicarbonate, pH 7.5, and 5% glycerol). Native mass spectrometry measurements were performed using a Synapt G2-Si mass spectrometer equipped with a nanoelectrospray ionization source, as described previously[93].

2.6 Accession codes

Sequencing reads have been deposited in the Sequence Read Archive under project number PRJNA506050.

2.7 Acknowledgements

A.V.W. is supported by a US National Science Foundation Graduate Fellowship. J.Y.W is supported by the NIH T32 066698 training grant and the Berkeley Graduate Fellowship Program. This work was funded by US National Science Foundation grant number 1244557. This work used a mass spectrometer that was purchased using NIH funding (grant 1S10OD020062–01) and the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 OD018174 Instrumentation Grant. This work was partly conducted by the US Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, under contract number DE-AC02–05CH11231. J.A.D. is an investigator of the Howard Hughes Medical Institute. We thank the Francis lab for use of their Agilent 1260 Infinity Multi-detector system and Malvern Zetasizer Nano instrument and M.J. Lobba for help in data collection with these instruments. We thank G.J. Knott for input on the manuscript. Correspondence and requests for materials should be addressed to J.A.D. (doudna@berkeley.edu).

2.8 Author contributions

A.V.W., J.Y.W., D.B., and J.A.D. designed research. A.V.W. and J.Y.W. performed the biochemical experiments and analyzed data. L.B.H. performed the protein purification. D.B., D.P.-E., and N.C.K. performed the bioinformatics experiments and analyses. J.F.B. supervised the bioinformatics research and provided funding for it. A.T.I performed mass spectrometry analysis. A.V.W., J.Y.W., D.B., and J.A.D. wrote the paper.

CHAPTER 3

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

A portion of the work presented in this chapter has been published previously as part of the following paper: Wang, J.Y., Hoel, C.M., Al-Shayeb, B., Banfield, J.F., Brohawn, S.G., & Doudna, J.A. Structural coordination between active sites of a CRISPR reverse transcriptase-integrase complex. *Nat. Commun.* **12**, 2571 (2021).

3.1 Abstract

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

3.2 Introduction

CRISPR-Cas (clustered regularly interspaced short palindromic repeats-CRISPR associated) systems provide adaptive immunity against viruses for bacteria and archaea[1,94]. This process begins when an integrase comprising the conserved Cas1 and Cas2 proteins incorporates segments of foreign DNA—protospacers—into the host CRISPR locus, which consists of direct repeats separated by virally derived spacers[6,16–18,21–24,26,27,77]. Transcription and transcript processing generate CRISPR RNAs (crRNAs), which assemble with Cas proteins to form effector complexes that identify and destroy foreign nucleic acids bearing sequence complementarity to the ~20-nucleotide (nt) spacer segment of the crRNA[70,72–75,85].

Although most CRISPR spacer sequences are derived from phage DNA, a minority of CRISPR-Cas systems include a reverse transcriptase (RT), enabling spacer acquisition from RNA sources[9,41,43,66,95–100]. In some cases, the RT occurs as a fusion protein together with Cas1 on the RT's C-terminus that can facilitate DNA-based storage of transcriptional information and has been used as a tool to record cellular transcriptional activity[66]. RT-Cas1s occur primarily in type III CRISPR-Cas systems that target both RNA and DNA, raising the interesting possibility that these variants provide adaptive immunity against RNA phage or other RNA elements[75,101–104]. Notably, many of these RT-Cas1 fusion proteins include an N-terminal Cas6 domain[43,44,99]. The endoribonuclease Cas6 functions as a maturase that recognizes the CRISPR repeat and processes CRISPR transcripts into mature crRNAs[105–107]. A recent study showed that the Cas6 domain of a Cas6-RT-Cas1 fusion is required for RNA spacer acquisition

and co-evolves with the RT domain[106]. However, the molecular basis for functional coordination between Cas1, RT, and Cas6 remains unknown.

In this work, we determine the 3.7 Å resolution structure of a naturally occurring *Thiomicrospira* Cas6-RT-Cas1—Cas2 complex using single particle cryo-electron microscopy (cryo-EM) and investigate its activities *in vitro* to elucidate how RT activity might coordinate with the CRISPR integrase and with the Cas6 maturation nuclease in the context of a Cas6-RT-Cas1 fusion protein. We show that in spite of a difference in arrangement of the Cas1—Cas2 disposition compared to the well-studied *E. coli* integrase, the Cas6-RT-Cas1—Cas2 complex catalyzes site-specific full-site CRISPR spacer integration of double-stranded DNA (dsDNA) substrates. The unique RT conformation observed in the context of the CRISPR integrase fusion protein, in addition to direct interactions between the RT, integrase, and maturase active sites, suggests that structural rearrangements coordinate different activities and substrates during CRISPR sequence acquisition. It is unclear whether RNA serves as the primary source of spacers for these systems, and surprisingly, we find that the RT-integrase favors DNA as an integration substrate and conducts complementary DNA (cDNA) synthesis using both RNA and DNA templates that have minimal homology between primer and template. Together, these findings underscore the versatility of CRISPR RT-integrase proteins to provide for expanded rather than specialized CRISPR spacer acquisition.

3.3 Results

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

To investigate the structure of a Cas6-RT-Cas1—Cas2 complex and its activities, we purified a Cas6-RT-Cas1 and its associated Cas2 protein, which are part of a *Thiomicrospira* type III CRISPR locus identified from a groundwater, CO2-driven Geyser dataset (Fig. 3.1a). To form the complex, the Cas6-RT-Cas1 and Cas2 proteins were assembled with a DNA substrate designed to mimic a genomic integration intermediate10 (Fig. 3.1a,b, Fig. 3.2a,b). The structure of the resulting Cas6-RT-Cas1—Cas2 complex was determined by single particle cryo-EM at an overall resolution of 3.7 Å (Fig. 3.1b–d, Fig. 3.3a,b, Fig. 3.4a–h). Approximately half of the complex displayed relatively lower local resolution, and refinement using a mask that excluded this region generated a reconstruction at 3.4 Å resolution with improved interpretability of high-resolution features. (In the remainder of the text, we refer to these two reconstructions as full complex and masked partial complex, respectively.) A model corresponding to two Cas2 chains and portions of four Cas6-RT-Cas1 chains was de novo built and refined into the higher resolution masked partial complex reconstruction. Portions of three Cas6-RT-Cas1 chains that were excluded by the mask were then rigid

body docked and refined in the full complex reconstruction. The final model consists of two entire Cas2 chains, two entire Cas6-RT-Cas1 chains (excluding five poorly resolved loops), and the Cas1 domain from two additional Cas6-RT-Cas1 chains.

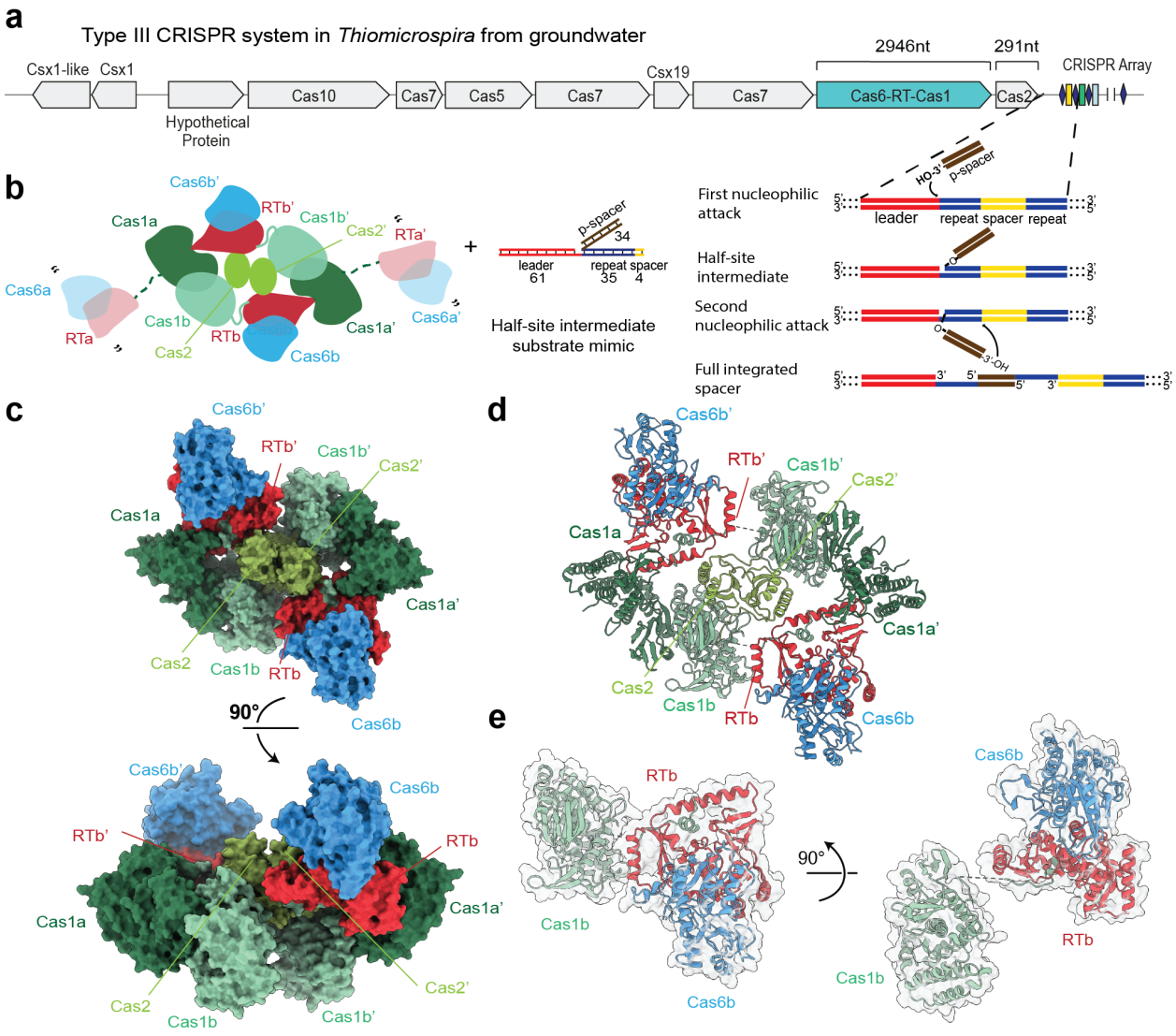


Figure 3.1 | A natural Cas6-RT-Cas1 fusion protein within a type III CRISPR-Cas system. (a) Illustration of the CRISPR genomic locus encoding the proteins used in this study and schematic of the CRISPR spacer integration reaction. Leader, red; repeat, blue; spacer, yellow; dsDNA protospacer (p-spacer), brown. **(b)** Cartoon showing Cas6-RT-Cas1 and Cas2 proteins and the DNA substrate, mimicking the half-site intermediate depicted in **a**, used in complex assembly. Proteins colored by domain: Cas6, blue; RT, red; Cas1a, dark green; Cas1b, turquoise; Cas2, lime green. Visible RT-Cas1 linkers are shown as solid turquoise lines. Missing Cas6-RT domains (Cas6a/RTa and Cas6a'/RTa') that are present in the complex but not visible are depicted as semi-transparent, connected to the rest of the structure by dashed dark green lines. Quotation marks indicate presumed mobility. Lengths of the DNA substrate strands are indicated. **(c)** Overall structure of the complex depicted in surface representation and 90° rotation, with color-coding as shown in **b**. **(d)** Overall structure of the complex, depicted in ribbon

representation. (**e)** Architecture of a single Cas6-RT-Cas1 protomer and 90° rotation.

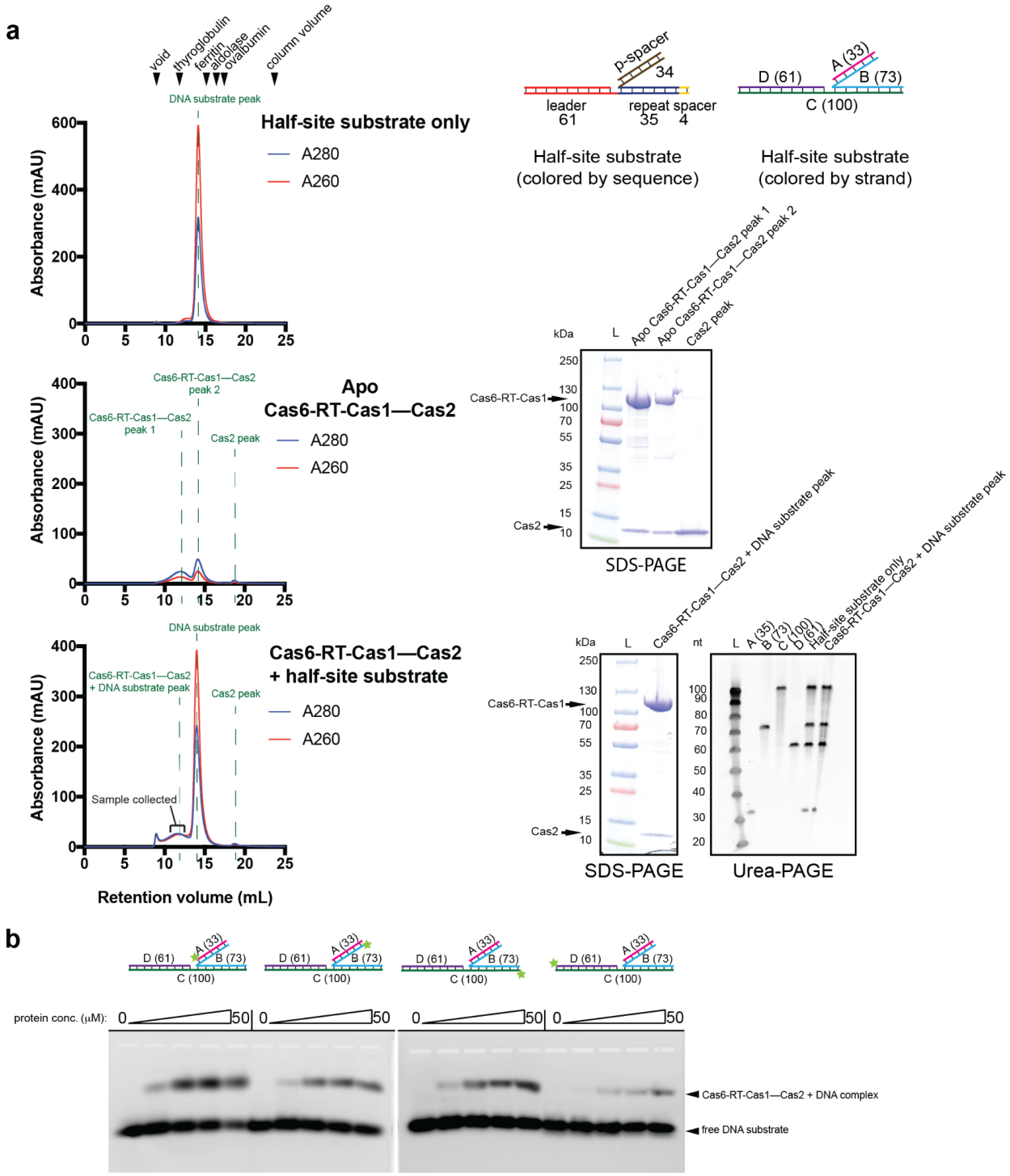


Figure 3.2 | Complex formation. (a) Gel filtration runs of DNA half-site substrate only, apo Cas6-RT-Cas1—Cas2, and Cas6-RT-Cas1—Cas2 + DNA half-site substrate. Cartoons of half-site substrate are shown colored by sequence (leader, red; repeat, blue; spacer, yellow; dsDNA protospacer, brown) and colored by strand with lengths of DNA strands shown (A, pink; B, light blue; C, green; D,

purple). The absorbance at 280 nm (blue) and 260 nm (red) are shown and the void volume, column volume, and elution volumes of standard proteins are indicated by black arrowheads. The peaks representing the Cas6-RT-Cas1—Cas2 complexes, the sample used for cryo-EM data collection, and free DNA substrate are indicated by green dashed lines. SDS-PAGE gels and urea-PAGE gel of the elution samples are provided. **(b)** Electrophoretic mobility shift assays using fluorescent half-site substrates colored by strand with color-coding from **a**. Star indicates 6-carboxyfluorescein label. Results are representative of 3 independent experiments.

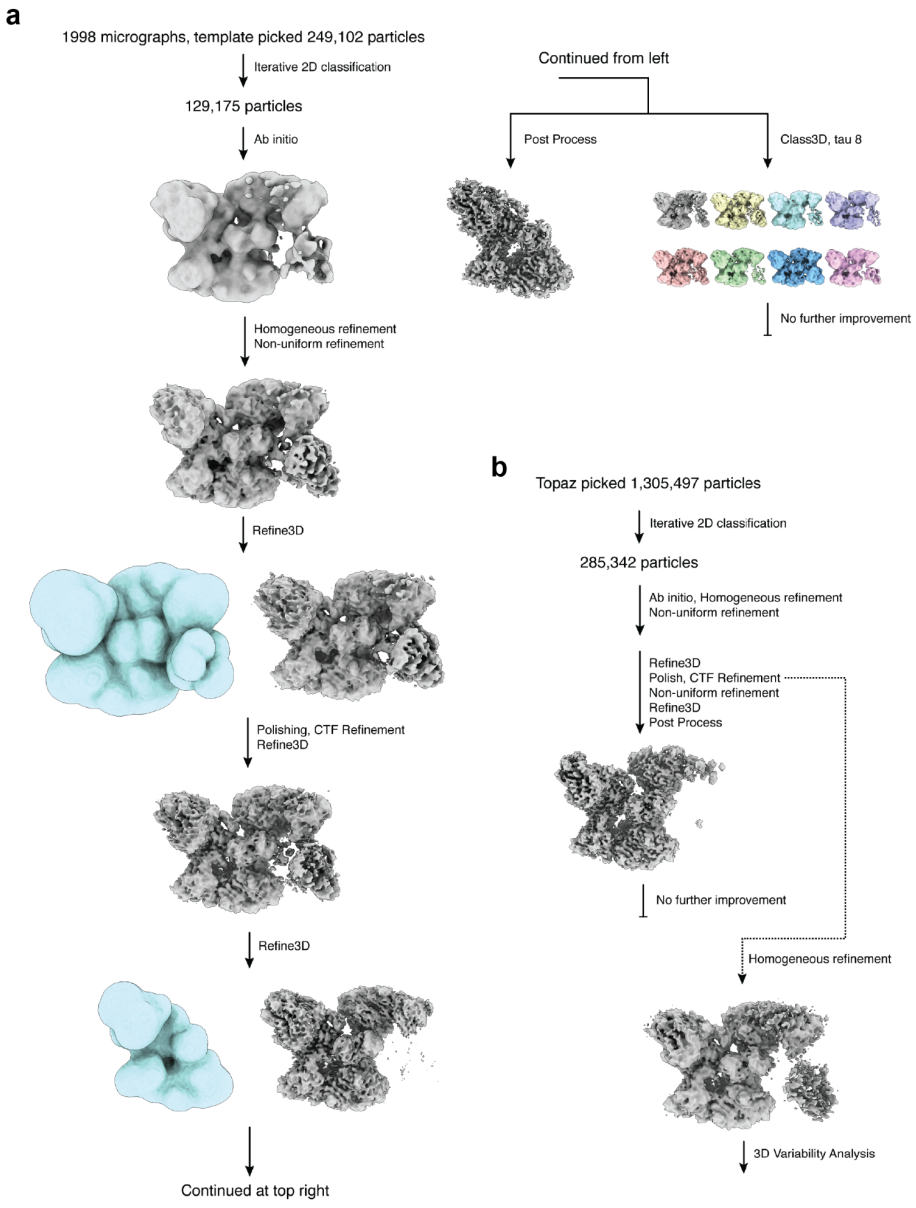


Figure 3.3 | Cryo-EM processing pipeline. (a,b) Cryo-EM data processing pipeline in RELION and cryoSPARC from RELION picked particles and Topaz picked particles respectively. Cryo-EM maps are colored gray, and masks used in RELION colored light blue. See Methods for details.

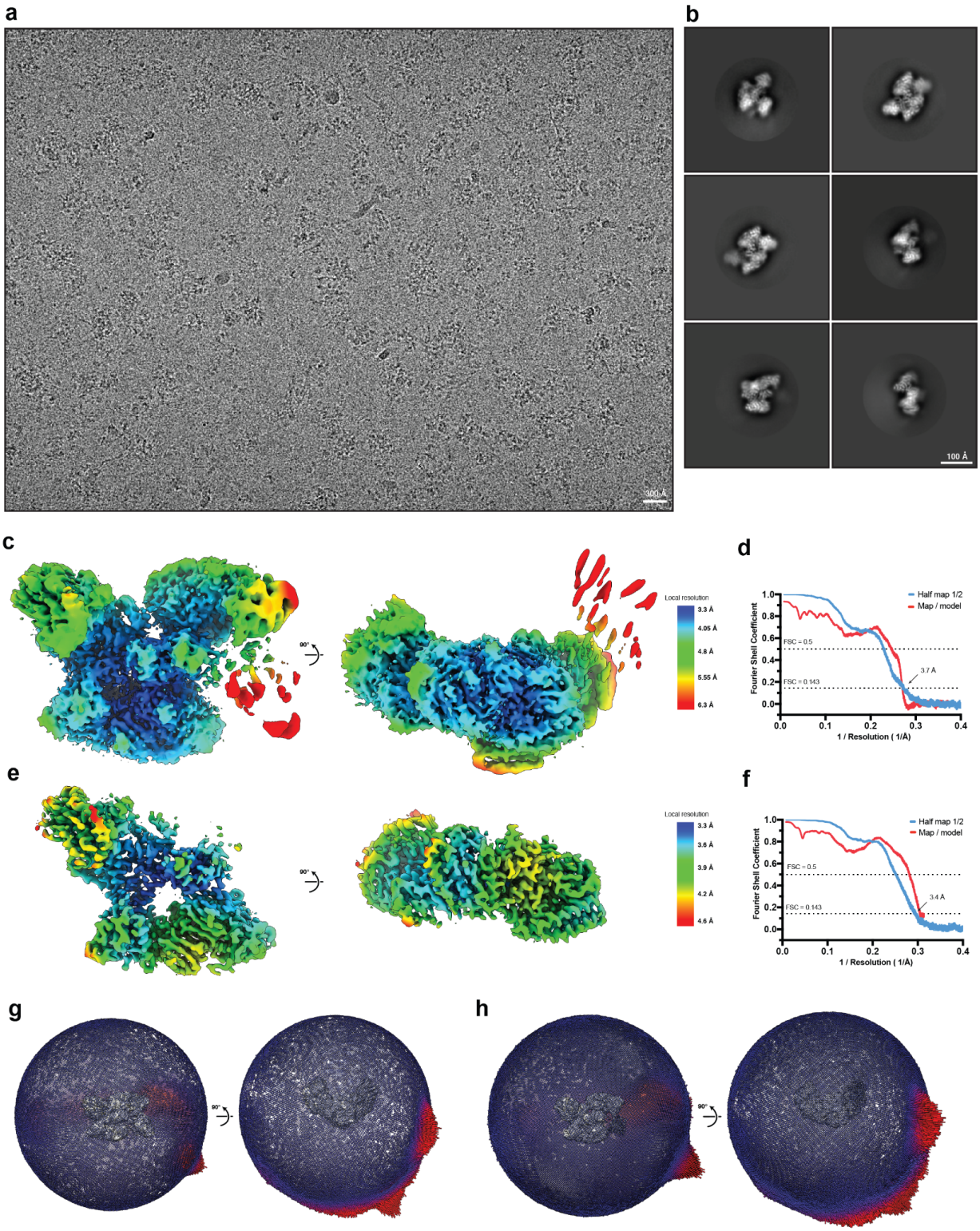


Figure 3.4 | Cryo-EM validation. (a) Representative micrograph with 10 Å lowpass filter, particles which contributed to the final reconstruction are circled. **(b)** Selected 2D class averages. Results for **a** and **b** are representative of 1996 micrographs (Table 3.3). All micrographs are deposited in EMPIAR-10642. **(c,e)** Local resolution estimated in RELION3.1 with map surfaces colored in rainbow as indicated, for full and

partial complexes respectively. **(d,f)** Fourier Shell Correlation (FSC) for the two unfiltered half-maps (light blue) and final map versus model (red), for full and partial complexes respectively. **(g,h)** Angular distribution of particles with corresponding cryo-EM map, for full and partial complexes respectively.

As observed for previous CRISPR integrase structures[10,16,18], the Cas6-RT-Cas1—Cas2 complex is heterohexameric, consisting of a central Cas2 dimer and two distal Cas6-RT-Cas1 dimers (Fig. 3.1b–d). The heterohexameric architecture of the core Cas1—Cas2 is a hallmark of CRISPR integrases and is critical to their internal ruler mechanism dictating the length of the spacers that are integrated and their ability to catalyze full-site integration[10,16,18]. Extending from the Cas1—Cas2 core are the Cas6-RT lobes, which are positioned orthogonal to the Cas1 domains relative to the central Cas2 dimer. The Cas6-RT lobe is connected to its associated Cas1 domain via a flexible linker, with the RT domain proximal to the central Cas2 dimer and the Cas6 domain closely abutting the RT domain (Fig. 3.1c–e). We observed density for the Cas6-RT lobe in two of the four Cas6-RT-Cas1 protomers (Cas1b and b'), presumably due to flexibility of the Cas6-RT domains in the other two Cas6-RT-Cas1 protomers (Cas1a and a'). No density was observed for the DNA substrate despite co-migration during size exclusion chromatography and indication of binding from electrophoretic mobility shift assays, suggesting that DNA was bound in only some of the complexes collected or that the DNA was lost during grid preparation or is too flexible to see (3.2a,b).

The Cas1 and Cas2 domains largely resemble those of the *E. coli* Cas1–Cas2 complex; however, there are some notable differences. When compared to the *E. coli* Cas1—Cas2 complex structure, the C-terminal α-helical domains of a given Cas1 are very similar (2.3 Å RMSD), while the linker and N-terminal β-sheet domains are offset by ~5–7 Å (3.5a,b). Each Cas2 protomer has a ferredoxin-like fold, as observed in previous Cas2 structures[108,109]. However, in contrast to the *E. coli* Cas1—Cas2 complex[27], the β-sheet in Cas2 is composed of three β-strands, and the C-terminal domain following the third β-strand crosses the central axis to pack against the face of the opposing subunit, resulting in an altered dimer interface. Additionally, the C-terminal segment (comprising amino acids 88–96) interacts with Cas1b rather than with Cas1b' as in the *E. coli* Cas1—Cas2 structure.

Previous studies have shown that CRISPR-associated RTs are most closely related to the RTs encoded by group II introns[41,43,98,99]. We observe that the RT domain shares the characteristic palm and fingers regions of the canonical right-handed fold in other RT structures but is missing the thumb/X subdomain that is present in retroviral and group II intron RTs[39,110,111] (3.5c). Compared to the group II intron RT structure[39], the finger domains are very similar with an overall RMSD of 1.9 Å, while the palm domains are strikingly different. Interestingly, in the present structure, the palm domain is reoriented such that the FADD motif, containing active site residues conserved across diverse RT subtypes, points away from the substrate binding cleft, suggesting that this

conformation may not be competent for reverse transcription.

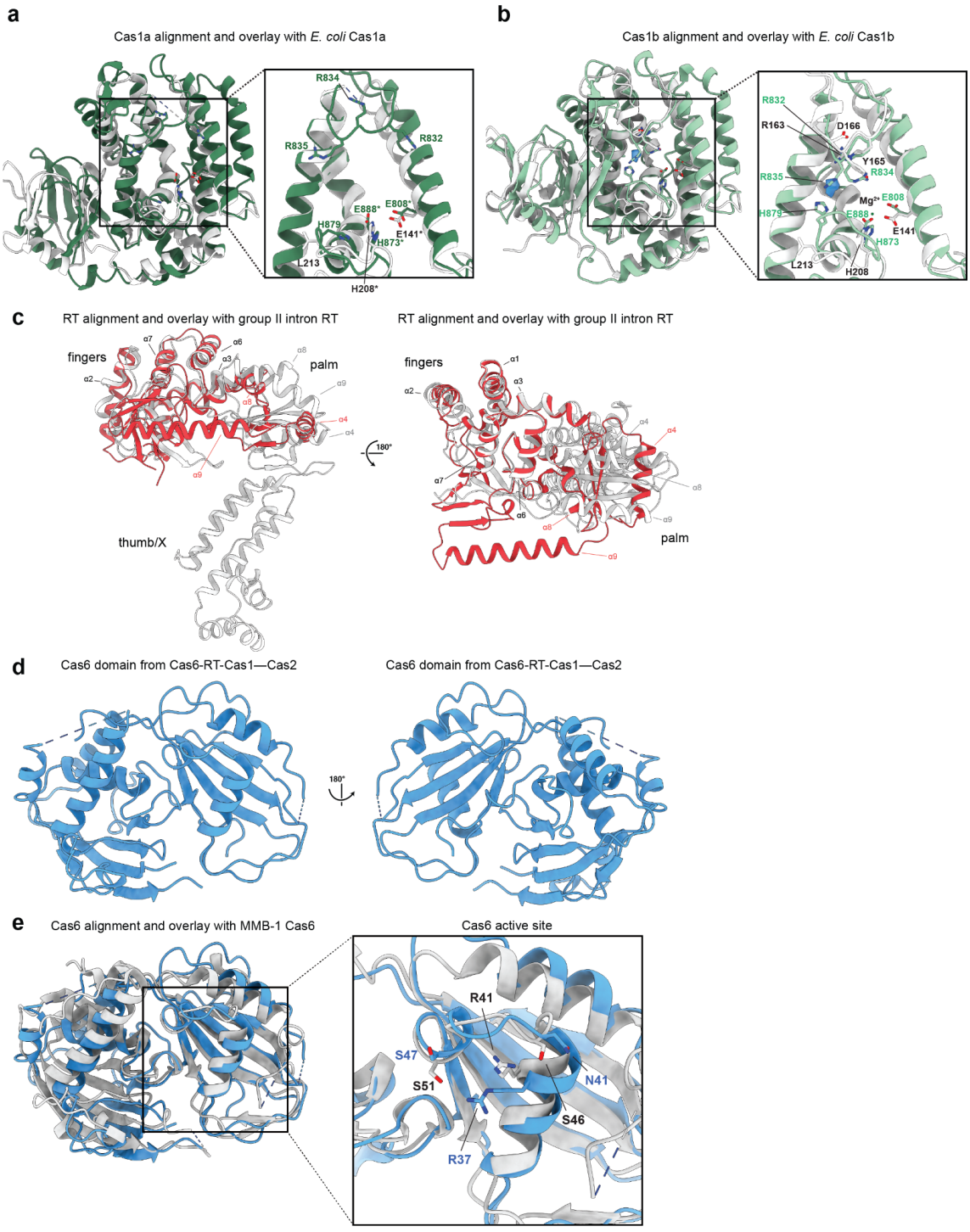


Figure 3.5 | Architecture of Cas1, RT, and Cas6 domains. (a) Alignment and overlay of Cas1a domain (dark green) from Cas6-RT-Cas1—Cas2 structure with *E. coli* Cas1a (white, PDB-ID 5DS5) and closeup of active site and surrounding residues, shown in stick configuration. Cas6-RT-Cas1—Cas2 residues are labeled in dark green, *E. coli* Cas1 residues labeled in black. Active site residues are labeled with an

asterisk *. **(b)** Alignment and overlay of Cas1b domain (turquoise) with *E. coli* Cas1b (white, PDB-ID 5DS5) and closeup of active site and surrounding residues. **(c)** Alignment and overlay of RT domain (red) and group II intron RT (white, PDB-ID 6AR1), with helices from the finger and palm domains labeled. Closely-aligned helices labeled in black, offset helices labeled in red (Cas6-RT-Cas1—Cas2 RT domain) and gray (group II intron RT). **(d)** Architecture of Cas6 domain from Cas6-RT-Cas1—Cas2 structure and 90 rotation. **(e)** Alignment and overlay of Cas6 domain from Cas6-RT-Cas1—Cas2 structure (blue) with Cas6 domain from MMB-1 Cas6-RT-Cas1 fusion (PDB-ID 6DD5, white) and closeup of active site residues, shown in stick configuration. Cas6-RT-Cas1—Cas2 residues are labeled in blue, MMB-1 Cas6 residues labeled in black.

Finally, the structure of the Cas6 domain closely resembles those of previous Cas6 structures[12,44,105,112–114] (3.5d). When compared directly to the structure of the Cas6 domain from the Marinomonas mediterranea (MMB-1) Cas6-RT-Cas1 protein[44], the two structures align with RMSD of 2.4 Å, with the greatest variance in the β-strand of the C-terminal RNA recognition motif (RRM) fold, perhaps due to the subsequent connection to the RT domain, absent in the MMB-1 Cas6 structure (3.5e).

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

Previous biochemical and genomic experiments demonstrated both RT and integration activities for the MMB-1 RT-Cas1 fusion protein[41,44]. To test the integrase function of the *Thiomicrospira* Cas6-RT-Cas1—Cas2 complex, we first investigated its substrate preferences for integration into a target DNA molecule and the structural elements that may be involved in the cleavage-ligation reactions. We conducted integration assays by incubating the purified Cas6-RT-Cas1 and Cas2 proteins with a 5'-fluorophore-labeled 35-base pair (bp) dsDNA or a 35-nt single-stranded DNA (ssDNA) or a single-stranded RNA (ssRNA) substrate and a supercoiled plasmid containing the CRISPR locus that functions as a target for spacer integration. Deoxynucleoside triphosphates (dNTPs) were supplied to reactions with RNA substrates. We observed protospacer ligation by examining the incorporation of the fluorophore-labeled substrate into the CRISPR locus in the target plasmid (Fig. 3.6a). The results show that the Cas6-RT-Cas1—Cas2 complex can catalyze ligation of all three substrates into the target plasmid (Fig. 3.6b). Interestingly, the Cas6-RT-Cas1 protein alone is able to catalyze protospacer ligation with DNA substrates. While Cas2 alone shows no activity, its presence increases the cleavage-ligation efficiency of the complex. Time-course experiments that quantified the extent of plasmid-nicking in each reaction showed >99% and 77% conversion from supercoiled plasmid target to open circular product with dsDNA and ssDNA, respectively, 25% with ssRNA, and 13% with no protospacer over 2 h (Fig. 3.6c, Fig. 3.7a). We conducted additional time-course experiments with two DNA/RNA hybrid substrates: one hybrid consisting of a 35-nt RNA oligonucleotide and a shorter 25-nt DNA strand and the second hybrid consisting of the same 35-nt RNA oligonucleotide and its fully complementary 35-nt DNA strand. This

resulted in an 89% and 91% conversion from supercoiled plasmid target to open circular product with the first and second hybrid, respectively, after 2 h in the absence of dNTPs and a 95% and 97% conversion with the first and second hybrid, respectively, after 2 h in the presence of dNTPs (Fig. 3.7b). These results show that this *Thiomicrospira* Cas6-RT-Cas1—Cas2 integrase is significantly more efficient with DNA and DNA/RNA hybrid substrates than ssRNA substrates for cleavage-ligation and that dNTPs do not seem to be required for cleavage-ligation activity.

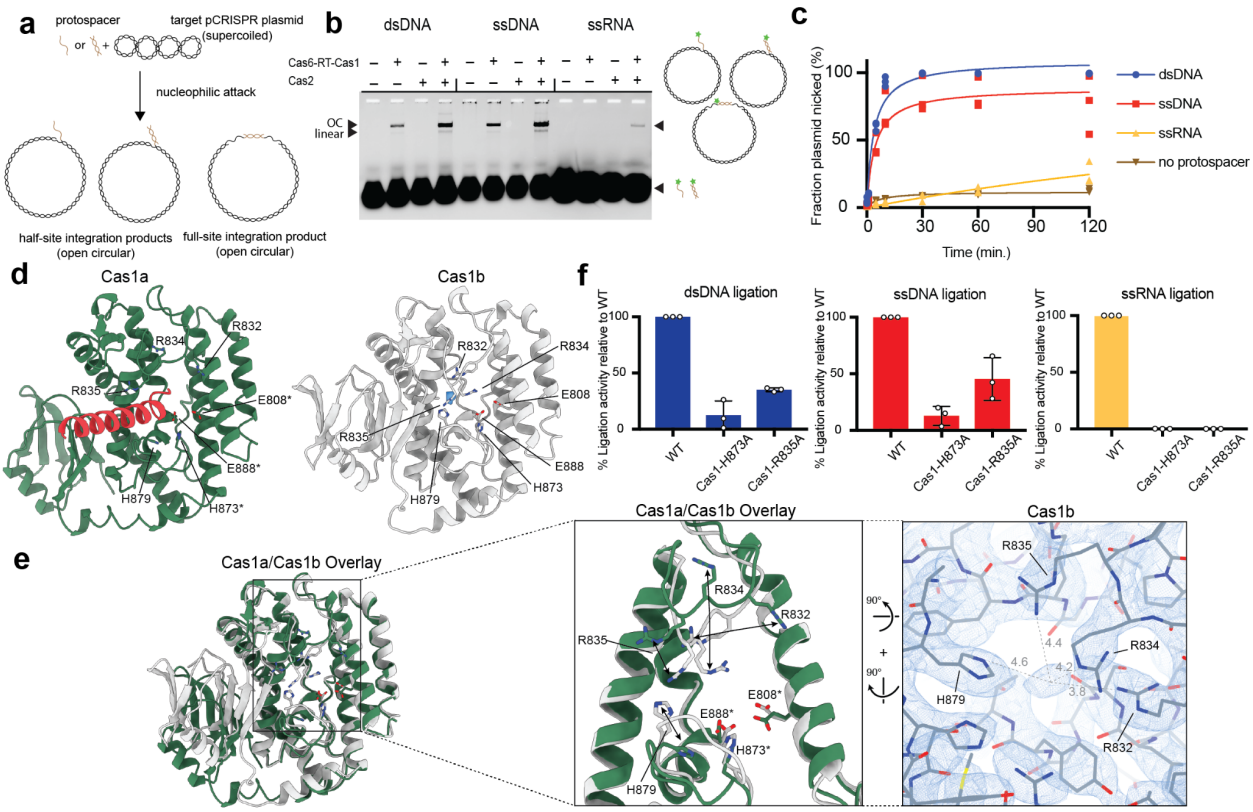


Figure 3.6 | Cas6-RT-Cas1—Cas2 catalyzes cleavage-ligation reactions with dsDNA, ssDNA, and ssRNA substrates *in vitro*. (a) Schematic of *in vitro* integration reaction of DNA or RNA protospacer (brown) into a supercoiled pCRISPR plasmid (black) that functions as the target. **(b)** Integration assay with fluorescent 35-nt dsDNA (0.5 μM), ssDNA (0.5 μM), and ssRNA (4 μM) protospacers. Open circular (OC) and linear ligation products are indicated. Star indicates 6-carboxyfluorescein label. Results are representative of 3 independent experiments. **(c)** Time-course integration assay comparing extent of plasmid-nicking with dsDNA, ssDNA, and ssRNA protospacers (35 nt; 4 μM). Fraction plasmid nicked is calculated as the fraction of open circular products relative to all plasmid (n = 3 biologically independent experiments). Experimental fits are shown as solid lines. Statistical significance of fraction plasmid nicked after 2 h was assessed using unpaired, two-tailed T tests (α = 0.05) (P = 0.000092, dsDNA vs. ssRNA; P = 0.0173, ssDNA vs. ssRNA). Representative gels are shown in Fig. 3.7a. **(d)** Comparison of Cas1a (dark green) and Cas1b (white), with RT-helix (red) and unknown density (blue). Active site and surrounding residues are shown in stick configuration. Active site residues are labeled with an asterisk *. **(e)** Cas1a and Cas1b overlay and closeup of active site. Double-headed arrows indicate large

conformational differences. Single-headed arrows indicate 90° rotations. Cryo-EM density shown in blue mesh, distances in angstroms. **(f)** Ligation of 35-nt fluorescent dsDNA (0.5 μM), ssDNA (0.5 μM), and ssRNA (4 μM) protospacers into target pCRISPR by mutant Cas6-RT-Cas1—Cas2 proteins. Percent ligation activity is calculated as the fraction of fluorescent products from the mutant Cas6-RT-Cas1—Cas2 relative to that from the wild-type (error bars, mean ± sd, n = 3 biologically independent experiments).

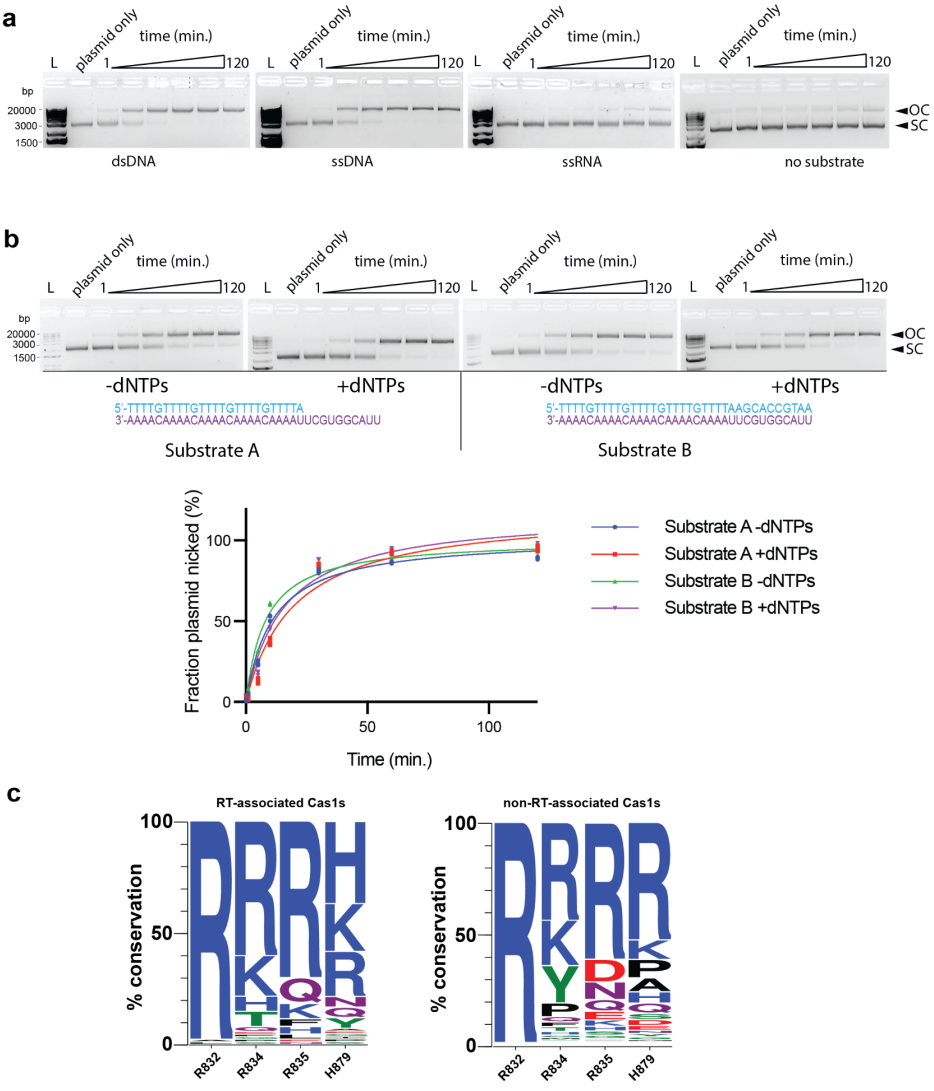


Figure 3.7 | Cas6-RT-Cas1—Cas2 prefers DNA and DNA/RNA hybrid substrates over ssRNA substrates for cleavage-ligation. (a) Time-course integration reactions with 35-nt dsDNA, ssDNA, ssRNA protospacers (4 μM) and a control with no protospacer. Supercoiled pCRISPR plasmid that functions as the target for integration (SC) and open-circle integration products (OC) are indicated. Results are representative of 3 independent experiments. **(b)**Time-course integration reactions comparing extent of plasmid-nicking with DNA/RNA hybrid substrates (4 μM) with and without dNTPs and quantification. Substrate sequences are shown (DNA, blue; RNA, purple). Fraction plasmid nicked is calculated as the fraction of open circular products relative to all plasmid (n = 3 biologically independent experiments). Experimental fits are shown as solid lines. Statistical significance of fraction plasmid nicked after two hours was assessed by comparing results to the results with ssRNA from Fig. 3.6c using

unpaired, two-tailed T tests (α = 0.05) (P = 0.0002, Substrate A - dNTPs; P = 0.0001, Substrate A + dNTPs; P = 0.0002, Substrate B - dNTPs; P = 0.0001, Substrate B + dNTPs). **(c)** Sequence logos indicating the most conserved residues of the R832, R834, R835, and H879 motif in RT-associated Cas1s (left) and non-RT-associated Cas1s (right) following sequence alignment across a diverse selection of Cas1s.

We wondered if there are elements near the Cas1 active site that enable RT-associated CRISPR integrases to catalyze cleavage-ligation chemistry with RNA substrates, a process not observed for non-RT-associated Cas1s. We found a cluster of residues (R832, R834, R835, and H879) adjacent to the active site that undergoes a large conformational change between Cas1a and Cas1b. In the non-catalytic Cas1b domains, these residues appear to coordinate an unknown density, potentially corresponding to a metal ion or water molecule (Fig. 3.6d,e). In the catalytic Cas1a domains, this density is not observed and the residues of interest are pushed away from each other, seemingly to accommodate the insertion of an α-helix from the RT domain. Sequence alignments across a diverse sample of RT-associated Cas1 variants show that there is ~100%, ~60%, ~70%, and ~35% conservation for R832, R834, R835, and H879, respectively, across RT-associated Cas1s (Fig. 3.7c). The latter three residues show only a slightly lower sequence conservation across the selection of non-RT-associated Cas1s, though they are not conserved in the classical *E. coli* Cas1 (Fig. 3.5a,b). Given our structural observations, we wondered if this cluster of residues may play an important role in the function of Cas6-RT-Cas1.

To examine the potential role of R835, one of the residues that coordinates the unknown density and undergoes a large conformational change, in protospacer ligation, we conducted integration assays with the Cas1-R835A mutant and the Cas1 active site mutant Cas1-H873A. We quantified the relative amounts of protospacer ligation to parallel wild-type (WT) controls. As expected, the Cas1-H873A mutant abolishes almost all protospacer ligation (Fig. 3.6f). Interestingly, the Cas1-R835A mutant also affected ligation activity, with a 65% and 55% reduction for dsDNA and ssDNA ligation and >99% reduction for ssRNA. Together, these findings identify R835 as a potential structural element that contributes toward Cas6-RT-Cas1—Cas2's ability to accommodate RNA substrates for protospacer ligation.

To determine whether the *Thiomicrospira* Cas6-RT-Cas1—Cas2 complex can catalyze cleavage-ligation in a site-specific manner (at the ends of the repeat sequence in the CRISPR array) with all three substrates[24,26], we conducted integration reactions using a short linear dsDNA molecule comprising 49 bp of the leader sequence, the 35-bp repeat, and 15 bp of the adjacent spacer as the target for integration of the substrate (Fig. 3.8a). The results show that the predominant cleavage-ligation product corresponds to nucleophilic attack of the protospacer at the spacer end of the repeat (the expected site for integration) for all three protospacers (Fig. 3.8b–d). This suggests that the integrase complex has some intrinsic target recognition, although the numerous

products resulting from off-target nucleophilic attacks indicate that other factors may be required to improve specificity. Curiously, while previous work on the MMB-1 system suggests that the cleavage-ligation products are potential substrates for target-primed reverse transcription[41,115,116], we do not observe any bands that represent extension of the 3' end of the DNA after cleavage-ligation when dNTPs are supplied (Fig. 3.8c). Furthermore, unlike what is observed for the MMB-1 system, dNTPs are not required for RNA ligation. The similar pattern of ligation products for all three substrates suggests that the mode for target recognition is likely the same with both RNA and DNA protospacers, although other factors likely contribute to its difference in efficiency.

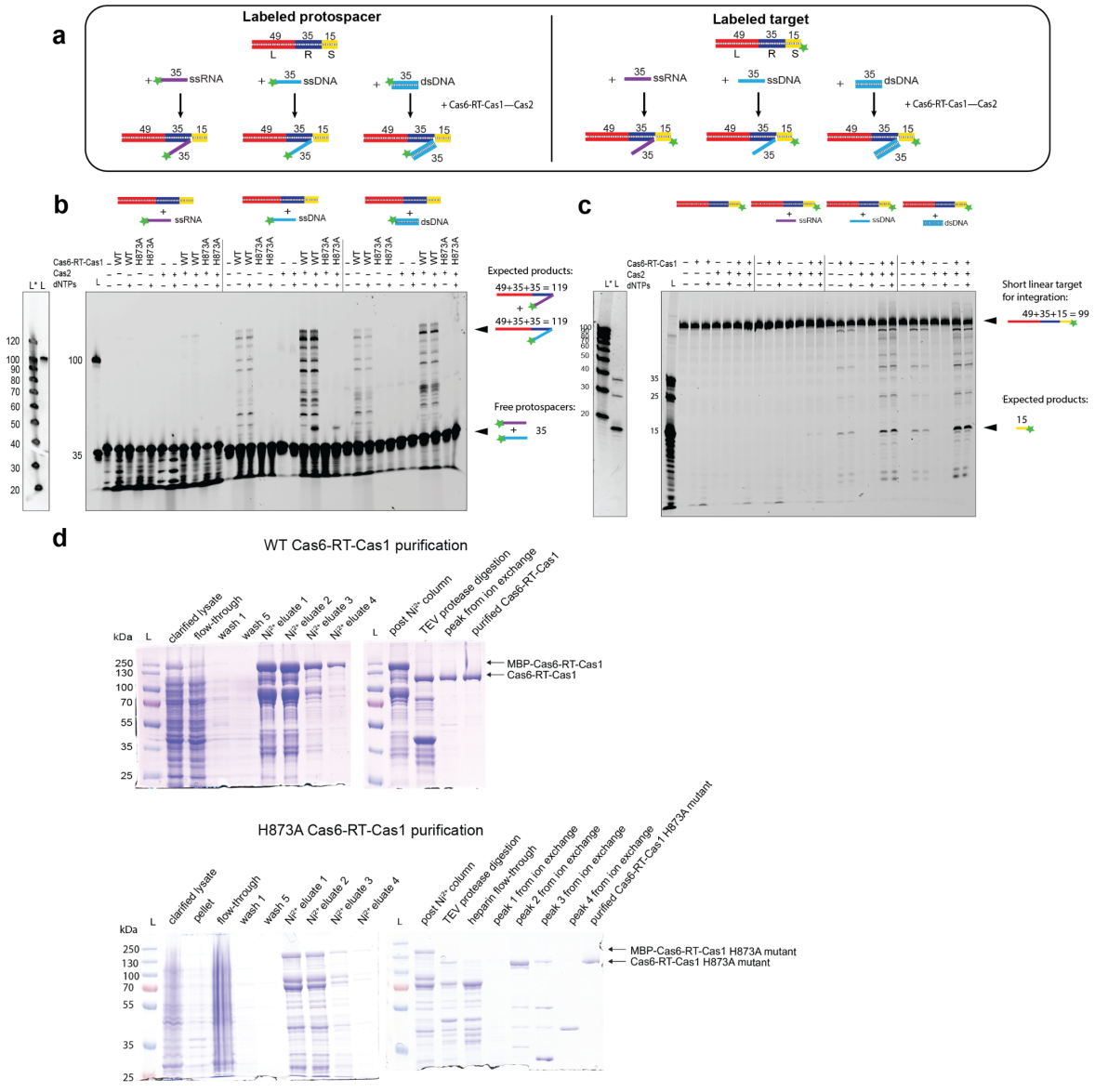


Figure 3.8 | Cas6-RT-Cas1—Cas2 catalyzes ligation of DNA and RNA protospacers into the CRISPR array. (a) Schematic of *in vitro* integration reaction of DNA or RNA protospacer into a short

linear dsDNA containing the CRISPR repeat as the target for integration. The lengths of the leader (L, red), repeat (R, dark blue), spacer (S, yellow), and protospacers (DNA, light blue; RNA, purple) are indicated. Star indicates 6-carboxyfluorescein label. **(b)** Cleavage-ligation assays conducted with fluorescent dsDNA, ssDNA, and ssRNA protospacers (1 μM) and unlabeled short linear target containing the CRISPR repeat. Expected products and their lengths are indicated. On the left, a non-fluorescent ladder L* is shown next to the fluorescent ladder L, visualized after SYBR Gold poststaining. **(c)** Cleavage-ligation assays conducted with fluorescent target (bottom strand labeled). Expected products and their lengths are indicated. Results for **b** and **c** are representative of 3 independent experiments. **(d)** SDS-PAGE gels of WT Cas6-RT-Cas1 and H873A Cas6-RT-Cas1 mutant purifications.

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

At the heart of the Cas6-RT-Cas1—Cas2 structure lies the heterohexameric Cas1—Cas2 core, which has some interesting differences from the *E. coli* integrase structure, raising questions about its functions as a CRISPR integrase. Two positively charged regions of the *E. coli* Cas1—Cas2 structure are critical for protospacer substrate binding: the Arginine Channel, which contacts the DNA substrate where the duplex terminates and the single-stranded overhang enters the active site, and the Arginine Clamp, which stabilizes the middle of the duplex[10]. Although similar charged regions occur in the Cas1 and Cas2 dimers of the Cas6-RT-Cas1—Cas2 structure, the Cas2s and Cas1a'/Cas1b' dimer are rotated further away from the Cas1a/Cas1b such that these charged regions are no longer in the same linear plane as observed in the *E. coli* Cas1—Cas2 structures[10,16,18] (Fig. 3.9a,b). The differences in the arrangement of the Cas1—Cas2 components relative to their observed positioning in the *E. coli* Cas1—Cas2 integrase led us to test whether the Cas6-RT-Cas1—Cas2 complex retains the intrinsic ruler mechanism characteristic of CRISPR integrases. Results of integration assays performed with 15- to 115-nt ssDNA and dsDNA substrates showed that the complex has some length selectivity for ssDNA and dsDNA. It is most active with 15- to 55-nt substrates and its activity decreases with substrates longer than 55 nt (Fig. 3.9c,d). This distribution is larger than the range of spacers observed in the CRISPR array (30–47 nt), though it is not surprising since half-site reactions tend to accept a larger variety of substrate lengths than full-site reactions4. What is more surprising is that the length distribution is similar for ssDNA and dsDNA protospacers, suggesting that the complex is able to narrow the selection of protospacer lengths in a manner that does not depend on interaction with a DNA duplex or access to two 3' nucleophilic ends.

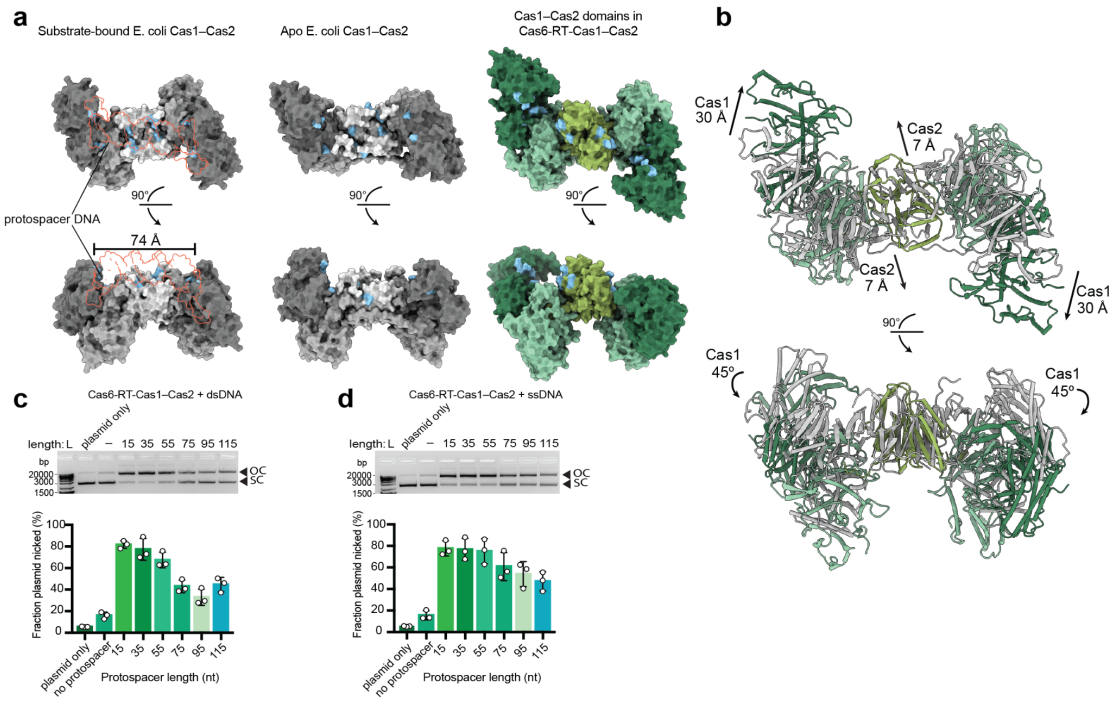


Figure 3.9 | Cas6-RT-Cas1—Cas2 shows length selectivity for dsDNA and ssDNA substrates for integration. (a) Comparison of Cas1—Cas2 disposition within the Cas6-RT-Cas1—Cas2 complex (color-coding from Fig. 3.1) to substrate-bound *E. coli* Cas1—Cas2 integrase (PDB-ID 5DS5) and apo *E. coli* Cas1—Cas2 integrase (PDB-ID 4P6I) (Cas1a, dark gray; Cas1b, medium gray; Cas2, light gray), shown in surface representation. Two views are shown related by a 90° rotation. Arginine clamp and arginine channel residues are colored in blue. Protospacer substrate is depicted in red outline. **(b)** Overlay of Cas1—Cas2 domains of Cas6-RT-Cas1—Cas2 complex (color-coding from Fig. 3.1) with apo *E. coli* Cas1—Cas2 integrase (PDB-ID 4P6I) (gray), aligned via Cas2 dimer, shown in ribbon representation. Two views are shown related by a 90° rotation. Arrows indicate conformational difference between Cas1—Cas2 domain of Cas6-RT-Cas1—Cas2 complex and that of *E. coli* Cas1—Cas2. **(c)** Cas6-RT-Cas1—Cas2 integration assays with variable length dsDNA protospacers (15–115 bp). Supercoiled pCRISPR plasmid that functions as target for integration (SC) and open circular products (OC) are indicated. Fraction plasmid nicked is calculated as the fraction of open circular products relative to all plasmid (error bars, mean ± sd, n = 3 biologically independent experiments). **(d)** Cas6-RT-Cas1—Cas2 integration with variable length ssDNA protospacers (15–115 nt), quantifying fraction plasmid nicked (error bars, mean ± sd, n = 3 biologically independent experiments).

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

We next set out to determine whether the Cas6-RT-Cas1—Cas2 complex can catalyze full-site CRISPR spacer integration. We devised an assay using chloramphenicol resistance as a selection marker for plasmids with protospacer insertion events near the target region[32,84]. The reporter construct was designed with 163 bp of the leader sequence and the CRISPR repeat sequence immediately upstream of a chloramphenicol resistance gene with a missing ribosomal binding site (RBS) sequence and start codon (Fig. 3.10a). Insertion of a spacer supplying the missing RBS

sequence and start codon to the target region allows translation of the chloramphenicol resistance gene transcript. We performed *in vitro* integration assays with this reporter plasmid and a protospacer with an RBS sequence and start codon, transformed the integration products into *E. coli*, plated the transformants on agar containing chloramphenicol, and sequenced the surviving colonies.

The results show that the Cas6-RT-Cas1—Cas2 complex catalyzes full-site integration with dsDNA substrates. Full-site integration events, characterized by spacer insertion followed by duplication of the adjacent 35 bp, represent the majority of the insertion events (14 out of 25) (Fig. 3.10b,c). The other insertion events appear to be the result of incomplete or abortive integration, characterized by partial/whole spacer insertions followed by a deletion of variable length of the adjacent sequence. While the incomplete/abortive integration events occurred at many off-target locations, full-site integration events were highly specific, with 12 out of the 14 full-site integration sites located at the leader-repeat junction. This suggests that while the complex might attempt integration at off-target sites, it maintains specificity for full-site integration. We did not observe full-site integration with ssDNA or ssRNA protospacers. It is possible that the Cas6-RT-Cas1—Cas2 complex alone can only conduct half-site reactions with single-stranded protospacers *in vitro* and other host factors may be necessary to support full-site integration with those substrates.

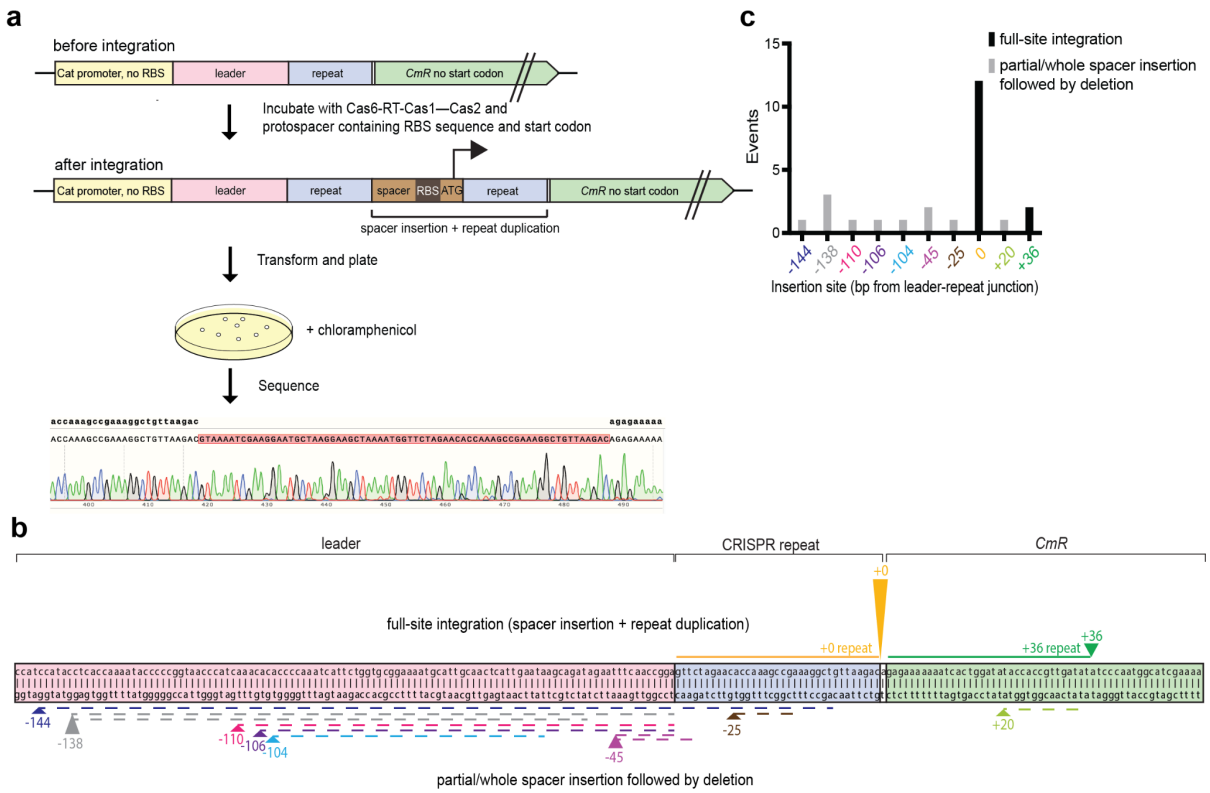


Figure 3.10 | Cas6-RT-Cas1—Cas2 conducts site-specific full-site integration. (a) Schematic of chloramphenicol selection screen for full-site integration events near the leader-repeat junction. The

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

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

Although the RT domain has been implicated in RNA-derived CRISPR spacer integration[41], its mechanism and possible coordination with Cas1 is not known. Similar to other RT structures[39,110,111], the RT active site of the Cas6-RT-Cas1—Cas2 complex resides in the palm region and consists of three conserved aspartate residues located on a three-stranded antiparallel β-sheet (Fig. 3.11a–c). Structural alignment with the group II intron RT reveals that despite close alignment in the fingers region, the palm region of the RT domain is dramatically offset, resulting in a ~90° rotation of the β-sheet containing the active site[39] (Fig. 3.11b, c). Possibly as a result, the two active site aspartates between β6 and β7 are further removed from the third active site aspartate between β5 and α5. This rotation is surprising given that there is comparatively little variation among the group II intron RT and other retroviral RT active sites[39,110,111], raising the possibility that the current conformation may represent an inactive conformation.

In solution, however, Cas6-RT-Cas1 alone catalyzes reverse transcription, an activity that is enhanced upon addition of Cas2 (Fig. 3.11d). Using substrates consisting of a 25-nt labeled DNA primer annealed to a 35-nt DNA or RNA template, we observe extension of the primer from 25 to 35 nt, corresponding to the full template length, when dNTPs are supplied (Fig. 3.11e). There are some unexpected dNTP-independent products around 50–55 nt that do not form when the Cas1 active site mutant is used, suggesting that they are cleavage-ligation products (Fig. 3.11e,f, Fig. 3.12a). Interestingly, in the lane that has only the DNA primer as a substrate, a smeary dNTP-dependent band forms around 45–48 nt, which matches the length of a cDNA product that forms when a second copy of the 25-nt DNA substrate is used as a template for primer extension with only 1–2 bp homology. A follow-up experiment using a 5'-labeled 35-nt DNA or RNA oligonucleotide with the original template sequence also

resulted in template-dependent extension of the substrate with only 1 bp homology (Fig. 3.12b). These results suggest that the Cas6-RT-Cas1—Cas2 complex is able to catalyze primer extension with DNA and RNA templates with minimal homology requirements.

To examine RT fidelity, we conducted dNTP drop-out experiments and supplied specific dideoxynucleoside triphosphates (ddNTPs) to terminate primer extension. The results show that Cas6-RT-Cas1—Cas2 complex is faithful to the RNA template, terminating cDNA synthesis at the correct position along the sequence where a necessary dNTP is missing or after a ddNTP is incorporated (Fig. 3.11f). Comparisons between experiments using the DNA template and those using the RNA template indicate that the complex is more efficient in carrying out primer extension with the DNA template than the RNA template; however, the complex is slightly more error-prone with the DNA template with bands suggesting some misincorporation of additional dNTPs. The results demonstrate that while the visible RT conformation is likely inactive, the complex can carry out primer extension with both RNA and DNA templates, with slightly higher fidelity for RNA than DNA.

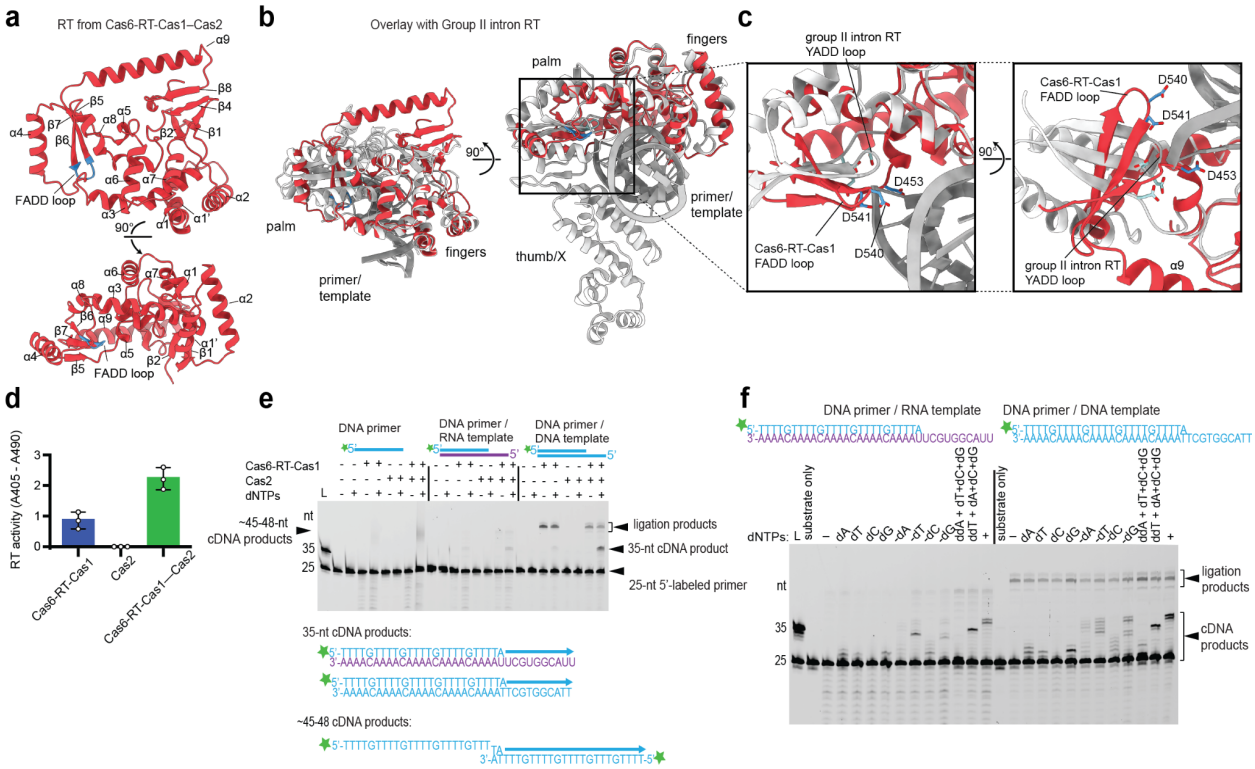


Figure 3.11 | Cas6-RT-Cas1 catalyzes cDNA synthesis with both RNA and DNA templates. (a) Architecture of RT domain from Cas6-RT-Cas1—Cas2 (red) and 90° rotation, with FADD motif (blue) indicated. **(b)** Alignment and overlay of RT domain from Cas6-RT-Cas1—Cas2 (red) with group II intron RT (white) bound to primer/template substrate (PDB-ID 6AR1) and 90° rotation, FADD motif from Cas6-RT-Cas1—Cas2 colored blue and YADD motif from group II intron RT colored light blue. Palm,

fingers, and thumb/X regions and primer/template are indicated. **(c)** Closeup of RT active site residues of Cas6-RT-Cas1—Cas2 (blue) and group II intron RT (light blue), shown in stick configuration. Two views are shown related by a 90° rotation. **(d)** RT activity assays comparing Cas6-RT-Cas1 alone and Cas6-RT-Cas1—Cas2. RT activity is measured using an ELISA-based colorimetric reverse transcriptase activity assay (error bars, mean ± sd, n = 3 biologically independent experiments) (Catalog No. 11468120910, Roche Diagnostics, Indianapolis, IN). **(e)** Template-driven cDNA synthesis reactions off a fluorescent DNA primer annealed to DNA and RNA templates. Substrates are schematized (DNA, blue; RNA, purple). Expected cDNA synthesis reactions are indicated. Star indicates 6-carboxyfluorescein label. Results are representative of three independent experiments. **(f)** Template-driven cDNA synthesis reactions in the absence of different dNTPs and in the presence of added ddNTPs. Results are representative of three independent experiments.

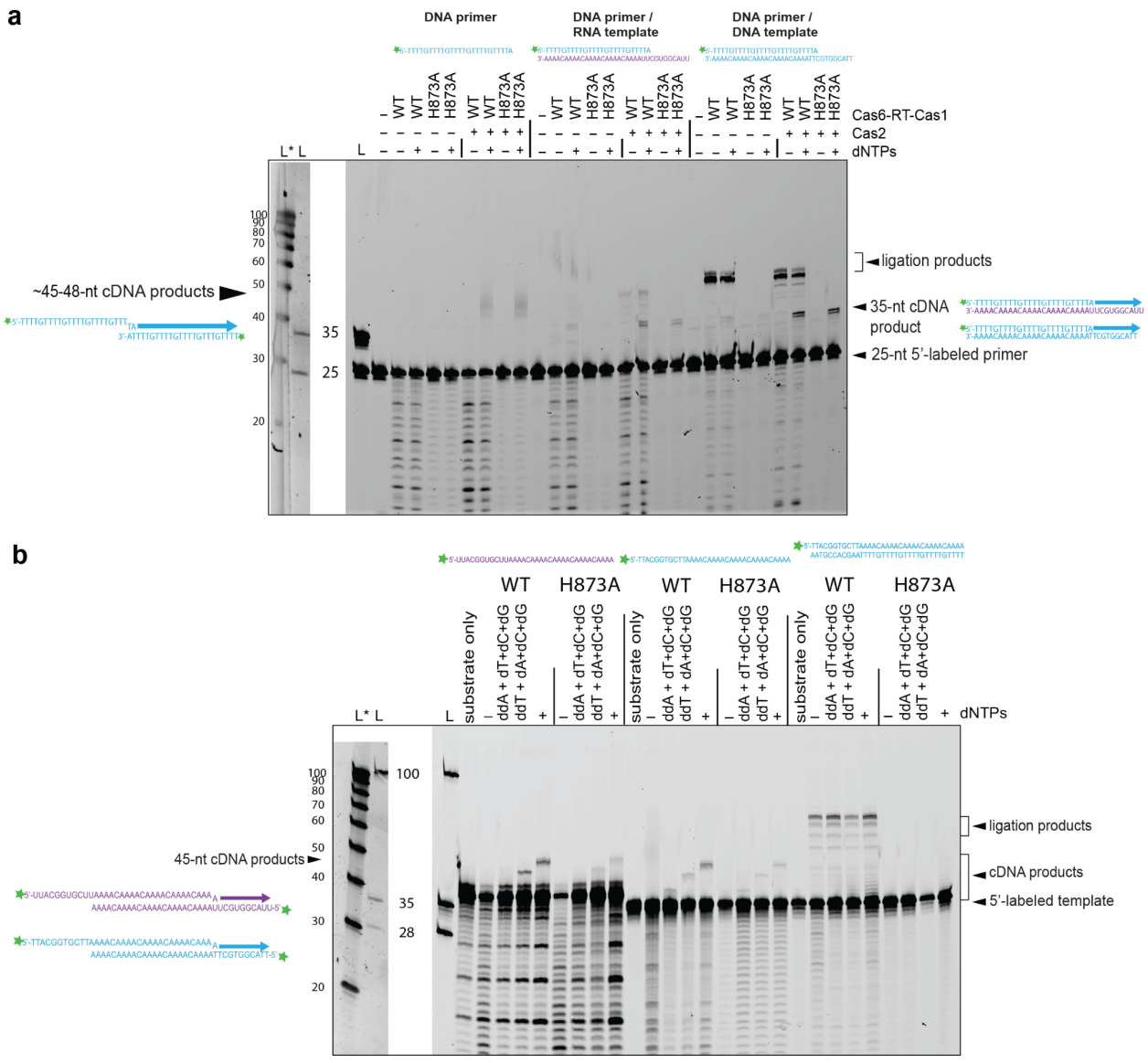


Figure 3.12 | Cas6-RT-Cas1 catalyzes cDNA synthesis with minimal primer-template homology. (a) Template-driven cDNA synthesis reactions off a fluorescent DNA primer annealed to DNA and RNA templates with WT and H873A Cas6-RT-Cas1 complexed with Cas2. Substrate sequences are shown

(DNA, blue; RNA, purple) and arrows indicate expected cDNA synthesis reactions. Star indicates 6-carboxyfluorescein label. On the left, a non-fluorescent ladder L* is shown next to the fluorescent ladder L, visualized after SYBR Gold poststaining. **(b)** Template-driven cDNA synthesis reactions in the absence of different dNTPs and in the presence of added ddNTPs with DNA or RNA template with no pre-annealed primer. Arrows indicate expected cDNA synthesis reactions. Results for **a** and **b** are representative of 3 independent experiments.

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

The structure reveals intriguing elements suggesting potential interplay between the Cas1, RT, and Cas6 domains. The most notable is an RT α-helix that serves as a direct connection between the RT and Cas1 active sites. While the N-terminal end of the RT-helix α9 is attached to the β-sheet containing the RT active site, the C-terminal end points into the catalytic Cas1a active site center (Fig. 3.13a,b). The helix insertion may be directly related to the conformational difference between Cas1a and Cas1b shown earlier. Structural alignments with other substrate-bound Cas1—Cas2 structures suggest that the helix insertion would result in considerable steric obstruction for the protospacer and target, perhaps even blocking access to the Cas1 active site entirely (Fig. 3.14a,b). The RT-helix would likely need to pull out in order for Cas1 to catalyze integration. We hypothesize that this RT-helix motion may be associated with the rotation of the connected β-sheet, which may bring the RT active site closer to the canonical conformation seen in other RT structures[39,110,111]. A closer look at the region around the RT-helix reveals additional clues on how this movement could potentially be regulated. The flexible RT-Cas1 linker wraps around the RT-helix, raising the possibility of its involvement in aiding the helix's movement. The RT-helix also interacts with Cas2, suggesting that rearrangement of the Cas2s could also impact the position of the RT-helix and the attached β-sheet containing the RT active site.

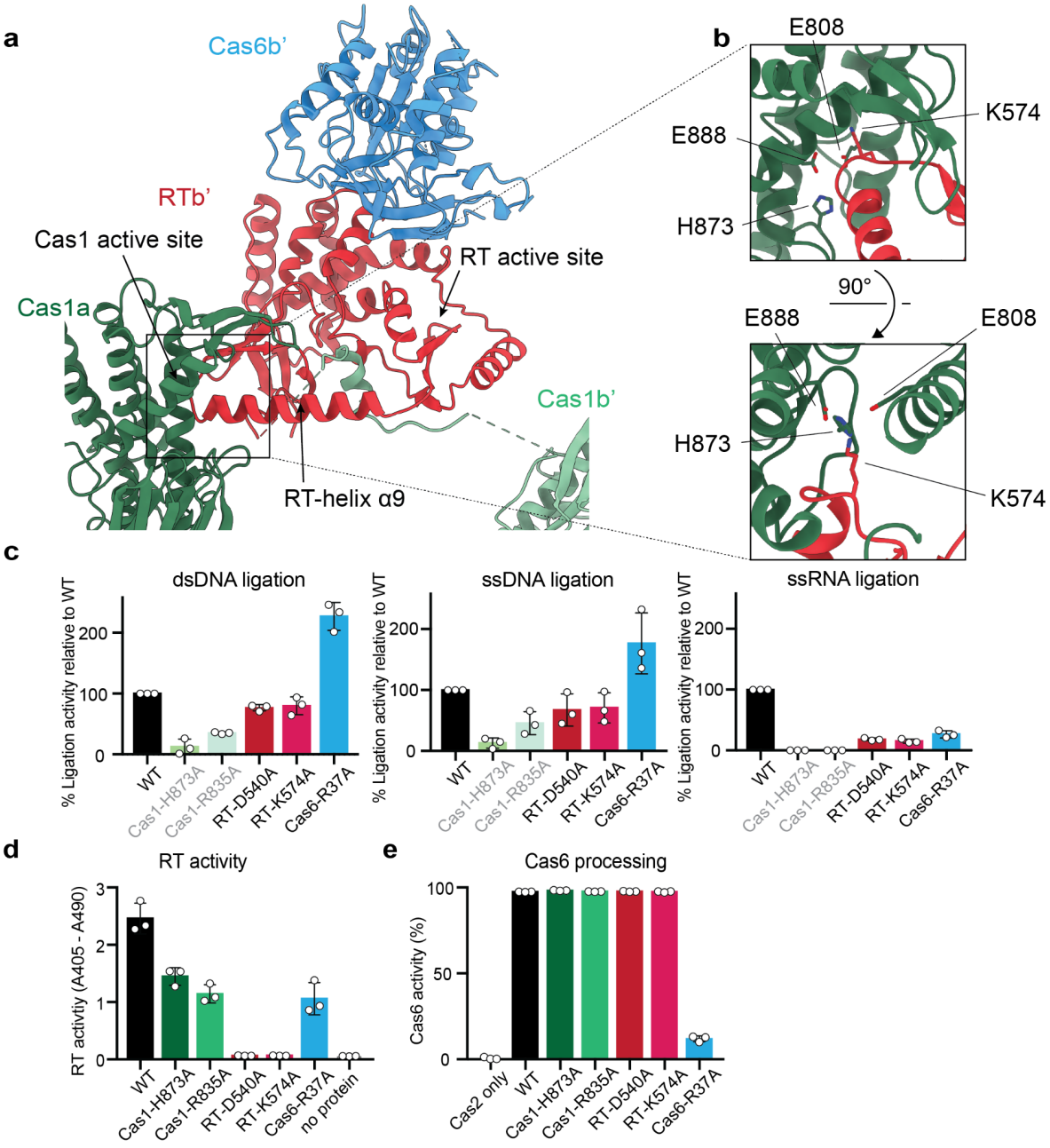


Figure 3.13 | Crosstalk between RT and Cas1 and Cas6 active sites. (a) Interaction between RT-helix α9 and Cas1 active site (color-coding from Fig. 3.1). **(b)** Closeup of RT-helix in Cas1a and 90° rotation. Cas1 active site residues and RT-helix K574 are shown in stick configuration. **(c)** Ligation of fluorescent 35-nt dsDNA (0.5 μM), ssDNA (0.5 μM), and ssRNA (4 μM) protospacers into target pCRISPR by mutant Cas6-RT-Cas1—Cas2s. The percent ligation activity is calculated as the fraction of fluorescent products from the mutant complex relative to that from the WT complex (error bars, mean ± sd, n = 3 biologically independent experiments). Results are represented by colored bars: WT, black; Cas1 mutants, greens; RT mutants, reds; Cas6 mutant, blue. Cas1 mutant data are the same data shown in Fig. 3.6f and are depicted in pale greens with gray labels. Representative gels are shown in Fig. 3.14c. **(d)** RT activity assays comparing the WT and mutant Cas6-RT-Cas1—Cas2s (error bars, mean ± sd, n = 3 biologically independent experiments; Catalog No. 11468120910, Roche Diagnostics, Indianapolis, IN). Statistical significance was assessed by comparing samples to WT control using unpaired, two-tailed T tests

(α = 0.05) (P = 0.0048, Cas1-H873A; P = 0.0019, Cas1-R835A; P = 0.0001, RT-D540A; P = 0.0001, RT-K574A; P = 0.0034, Cas6-R37A). Same color-coding is used from **c** except Cas1 mutant data are shown in darker greens with black labels. **(e)** Cas6 activity assays comparing the WT and mutant Cas6-RT-Cas1—Cas2 proteins. The percent cleavage is calculated as the fraction of the fluorescent CRISPR repeat RNA that has been cleaved, with color-coding from **d** (error bars, mean ± sd, n = 3 biologically independent experiments). A representative gel is shown in Fig. 3.14e.

To investigate the functional implications of the RT-helix in the crosstalk between the RT and Cas1 domains, we generated a panel of Cas6-RT-Cas1 mutants and tested them in protospacer ligation and RT activity assays. In addition to the Cas1 residues H873 and R835, we also mutated the RT active site residue D540 and the RT-helix lysine (K574), which points into the Cas1a active site. Here, ligation activity is shown as the percent ligation into the target plasmid by the mutant Cas6-RT-Cas1—Cas2 complex relative to a parallel WT control (Fig. 3.13c, Fig. 3.14c). Interestingly, while the RT-D540A and RT-K574A mutants result in a 20–33% reduction in DNA ligation, they reduce RNA ligation by more than 80%. It appears that impairing either RT activity or the RT-helix-Cas1a interaction substantially disrupts RNA ligation but not DNA ligation.

The results from the RT activity assays show further evidence of crosstalk between the RT and Cas1 domains. Reverse transcriptase activity levels for the different mutant Cas6-RT-Cas1—Cas2 complexes showed that mutating either D540 or K574 abolishes RT activity. The Cas1 domain mutations also show significant effects: the Cas1-H873A and Cas1-R835A mutants result in a 41% and 53% reduction in dNTP incorporation, respectively, after 2 h (Fig. 3.13d). These results show that despite being far away from the RT active site, the RT-helix lysine as well as the Cas1 residues potentially involved in the RT-helix—Cas1 interaction are significant for RT activity, supporting the hypothesis that the RT-helix plays an essential role in regulating RT activity. Together, these results suggest that there is bidirectional crosstalk between the Cas1 and RT domains.

We next explored the potential crosstalk between the Cas1/RT domains and the Cas6 domain, which lies on the opposite side of the RT. The results show that the Cas6 active site mutant Cas6-R37A reduces RT activity by 57% compared to the WT (Fig. 3.13d). Interestingly, the Cas6-R37A mutant increases DNA ligation to more than 176% that for the WT while reducing RNA ligation by 74% compared to the WT (Fig. 3.13c). Though the increase in DNA ligation is puzzling, the Cas6 active site mutant may reduce RNA ligation by way of its effect on RT activity, which is more critical for RNA ligation than for DNA ligation. These results show that Cas6 is not just operating on its own and influences the other domains in the complex.

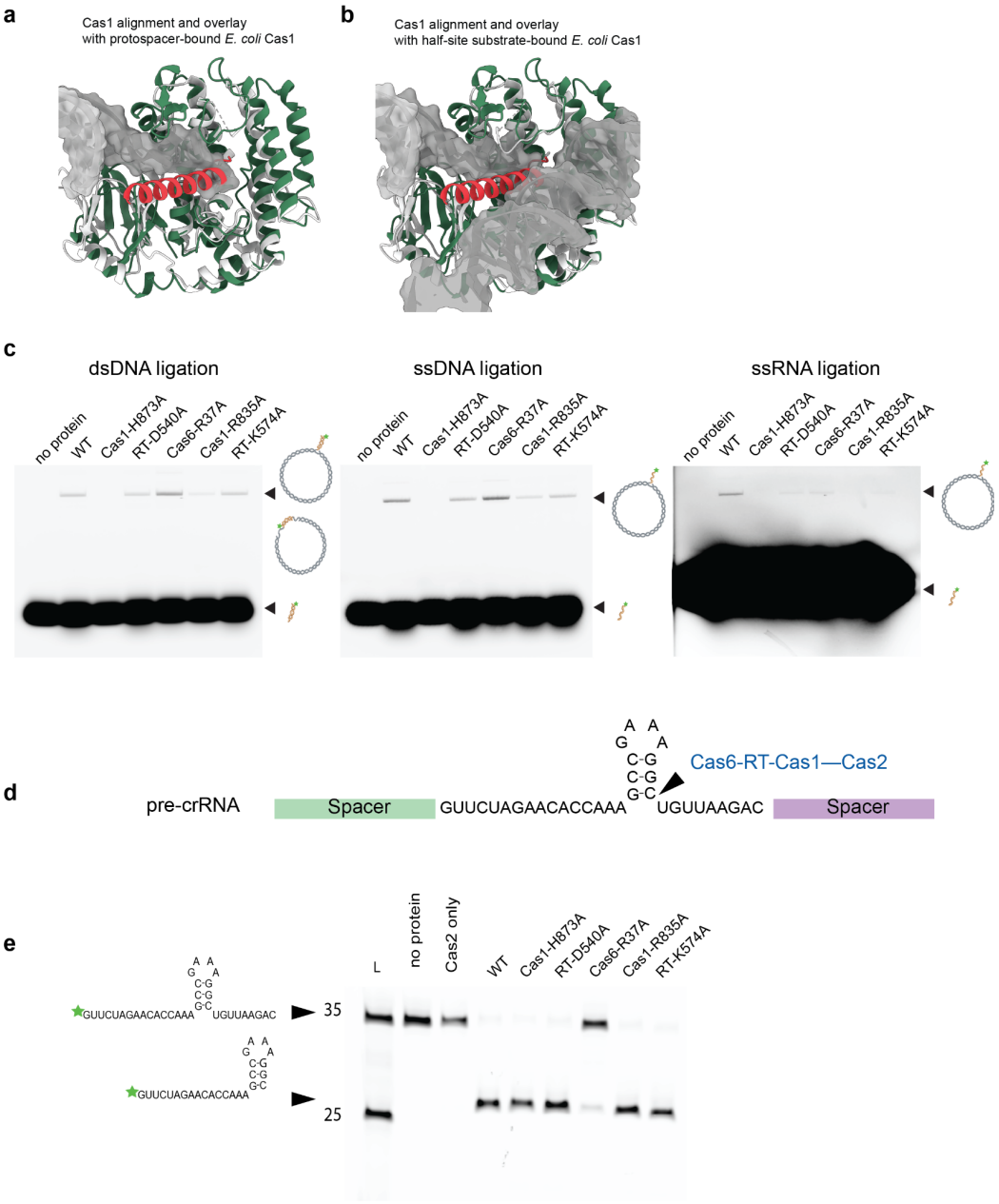


Figure 3.14 | Ligation and crRNA processing by Cas6-RT-Cas1—Cas2. (a) Alignment and overlay of Cas1a domain (dark green) with Cas1a from protospacer-bound *E. coli* Cas1—Cas2 structure (white, PDB-ID 5DS5 [https://doi.org/10.2210/pdb5ds5/pdb]). Protospacer substrate shown in gray surface representation. RT-helix is colored red. **(b)** Alignment and overlay of Cas1a domain (dark green) with Cas1a from half-site substrate-bound *E. coli* Cas1—Cas2 structure (white, PDB-ID 5VVJ [https://doi.org/10.2210/pdb5VVJ/pdb]). Half-site substrate shown in gray surface representation. **(c)** Ligation of fluorescent 35-nt dsDNA (0.5 μM), ssDNA (0.5 μM), and ssRNA (4 μM) protospacer into target pCRISPR by WT and mutant Cas6-RT-Cas1—Cas2s. Integration products and free protospacer are indicated and schematized. Star indicates 6-carboxyfluorescein label. **(d)** Sequence and predicted secondary structure of repeat with predicted cleavage site indicated. **(e)** crRNA processing activity assay comparing WT and mutant Cas6-RT-Cas1—Cas2s. Sequence and predicted secondary structure of CRISPR repeat RNA and the predicted product are indicated. Star indicates 6-carboxyfluorescein label.

Results for **c** and **e** are representative of 3 independent experiments.

We finally tested the mutants alongside the WT Cas6-RT-Cas1 in Cas6 processing activity assays. The Cas6-RT-Cas1—Cas2 complex was incubated with a 5'-labeled 35-nt RNA substrate corresponding to the CRISPR repeat sequence. The results show that the WT Cas6-RT-Cas1—Cas2 processes the 35-nt RNA substrate down to a 26-nt product with 97% efficiency after 2 h (Fig. 3.13e, Fig. 3.14d,e). Mutating the Cas6 active site residue R37 reduces processing efficiency to only 12%. The other mutants show no difference in processing relative to the WT, suggesting that the crosstalk between the Cas6 domain and the other two domains is unidirectional, with the Cas6 active site mutation affecting ligation and RT activity but the Cas1 and RT domain mutations having no effect on Cas6 RNA processing activity.

3.4 Discussion

RT-Cas1 and associated Cas2 proteins represent a unique CRISPR adaptation module shown to be involved in the acquisition of CRISPR spacers directly from RNA in a fascinating example of information flow from RNA to DNA[9,41,43,66,96–100,115]. In this study, we identify a type III *Thiomicrospira* Cas6-RT-Cas1—Cas2 complex containing three active sites for CRISPR spacer integration, reverse transcription, and RNA processing, and we observe all three activities in our reconstituted system. The structure of this Cas6-RT-Cas1—Cas2 complex revealed interesting differences between its Cas1—Cas2 integrase core and that observed for the *E. coli* Cas1—Cas2 integrase. An RT-helix links the RT and Cas1 active sites but also appears to block the position of substrate binding. Functional interactions between the Cas6, RT, and Cas1 domains show bidirectional crosstalk between the RT and Cas1 domains and unidirectional crosstalk from Cas6 to the other two domains.

Although the *Thiomicrospira* Cas6-RT-Cas1—Cas2 complex catalyzes site-specific cleavage-ligation chemistry with dsDNA, ssDNA, and ssRNA substrates, we only observe full-site integration with dsDNA substrates. Its inability to fully integrate single-stranded substrates even in the presence of dNTPs implies that additional host factors may be required to carry out this process *in vivo*. It is still unclear when the RNA substrate is reverse transcribed into cDNA during spacer acquisition. We found that the RT faithfully synthesizes cDNA from primer/template hybrids with short ~35-nt RNA or DNA templates and with minimal homology requirements; however, the preferred template and primer for these systems *in vivo* remain unknown. The difference in efficiency between the DNA polymerase activity and the RT activity is intriguing, because it suggests the possibility that the RT domain is primarily functioning as a DNA polymerase, although this result may also simply be a reflection of the tradeoff between efficiency and accuracy. Though previous work on the MMB-1 system suggested that the cleavage-ligation products could be substrates for target-primed reverse

transcription similar to group II intron retrohoming[41,115,116], we did not see evidence of this in biochemical assays. If such a process were to occur *in vivo*, it is interesting to speculate whether the twisted arrangement of the Cas1—Cas2 components relative to their observed positioning in the *E. coli* Cas1—Cas2 integrase might be advantageous to stall integration at some intermediate step to allow access for the RT.

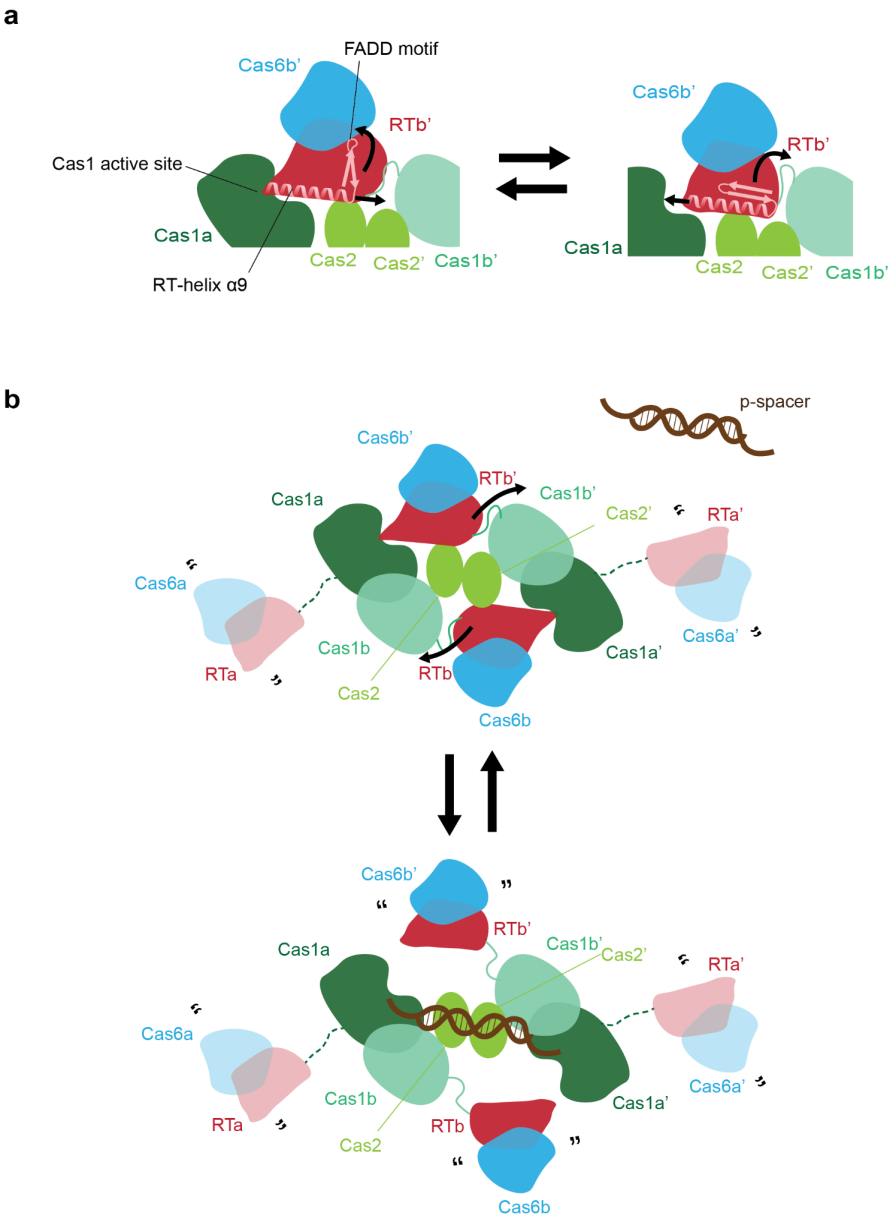


Figure 3.15 | Model for RT activation. (a) Cartoon showing hypothesized motion of RT-helix α9 in and out of the Cas1a active site and the hypothesized motion of the β-strands attached to the FADD motif containing the active site residues. Proteins colored by domain: Cas6, blue; RT, red, Cas1a, dark green; Cas1b, turquoise; Cas2, lime green. Visible RT-Cas1 linkers are shown as solid turquoise lines. RT-helix α9 and connecting β-strands containing FADD motif are schematized and shown in pink. Hypothesized

motions are indicated by curved arrows. **(b)** Cartoon showing hypothesized conformational change of the Cas6-RT-Cas1—Cas2 complex to accommodate a protospacer substrate, with color-coding from **a**. Missing Cas6-RT domains (Cas6a/RTa and Cas6a'/RTa') that are present in the complex but not visible are depicted as semi-transparent, connected to the rest of the structure by dashed dark green lines. Quotation marks indicate presumed mobility. Protospacer substrate is schematized in brown. Hypothesized motions are indicated by curved arrows.

The structural and functional interactions detected between the Cas1, RT, and Cas6 active sites suggest a model in which the linking RT-helix serves as a regulator of both RT and Cas1 activities through its movement in and out of the Cas1 active site (Fig. 3.15a,b). Given the RT-helix connection to the β-sheet containing the RT active site, we speculate that the motion of the RT-helix drives the RT to adopt a conformation more closely resembling the active conformation of other RT structures, or vice versa, that the conformational change of the β-sheet drives the motion of the RT-helix (Fig. 3.15a). We suggest that the Cas6b-RTb and Cas6b'-RTb' domains may be able to adopt two states: one where the Cas6-RT lobe is freely moving, only flexibly attached by a linker to the Cas1 domain in the same chain and another where it docks onto the opposite Cas1a domain, blocking Cas1 from carrying out integration (Fig. 3.15b). Having on-off switches for the RT and Cas1 active sites may be critical to the process of RNA spacer acquisition where intermediate substrates may pass from one active site to the other. In addition, the unidirectional crosstalk between the Cas6 to the other two domains also implies a possible multi-step process in which coordination between the different enzymatic functions creates an ordered process for spacer integration and CRISPR RNA processing.

Our results imply that Cas6-RT-Cas1—Cas2 functions as a coordinated system to enable spacer acquisition from both DNA and RNA sources. Questions remain about the relative efficiency of these reactions in light of our observations that the complex strongly prefers DNA over RNA substrates for cleavage-ligation, suggesting that RNA may not be the primary source of spacers for this system or that additional host factors are required to improve the efficiency of RNA integration. Previous literature examining correlation between the frequency of spacer acquisition and overall gene expression level does indicate that there is variation among these systems *in vivo*[41,66,100]. It is also possible that the efficiency in integrating RNA substrates varies depending on the substrate as a part of a selection process to distinguish between self and non-self. Future research on RNA spacer acquisition by these RT-Cas1 and Cas6-RT-Cas1 fusions can more broadly shed light on how these systems may potentially generate immunity against RNA invaders, which is still a topic that is not very well understood. Furthermore, deeper insight into these fusions can help make strides toward advancing tools that utilize these proteins to record transcriptional information in cells. By presenting a high-resolution structure of the Cas6-RT-Cas1—Cas2 complex, our work provides key details toward achieving a greater understanding of RT-associated integrases and their functions.

3.5 Materials and methods

3.5.1 Cas6-RT-Cas1 protein identification

Hidden Markov Models were built from RT-Cas1 fusion proteins or Cas1 sequences affiliated with RTs obtained from a previous study[99] and used to search for more RT-Cas1 representatives. Metagenomic contigs harboring fusion proteins consisting of RT, Cas1, and Cas6 domains were selected from a groundwater, CO2-driven Geyser dataset. The Cas6-RT-Cas1 and Cas2 proteins and their cognate CRISPR array from a *Thiomicrospira* genome assembly became the subject of this study following manual genome curation[117]. Sequencing reads from the original dataset were mapped to the de novo assembled sequences using Bowtie2 v.2.3.4.15. We retained unplaced mate pairs of mapped reads and mappings were manually checked to identify local misassemblies. Sequence ends prematurely terminated due to the presence of repetitive elements were extended using unplaced or incorrectly placed paired reads. In some cases, extended ends were used to recruit new scaffolds that were then added to the assembly. The extension accuracy and accuracy of local assembly changes were verified by subsequent read mapping.

3.5.2 Residue conservation analysis

A diverse set of 968 Cas1 proteins and 92 Cas1s fused to or associated with RTs (and in some cases, Cas6) was used for analysis of the conservation of the cluster of residues consisting of R832, R834, R835, and H879. An alignment was generated from the aforementioned set with the Cas6-RT-Cas1 protein of interest using the MAFFT v7.407 LINSI setting with 1000 iterations[118]. The positions of interest were extracted and a sequence logo was generated to display the conservation of residues across the relevant Cas1 proteins.

3.5.3 Protein purification

The Cas6-RT-Cas1 and Cas2 genes were codon optimized for *E. coli* expression and ordered as G-blocks. Cas6-RT-Cas1 and Cas2 were PCR amplified and cloned separately into a pET-based expression vector with an N-terminal 10xHis-MBP-TEV tag. Plasmids were transformed into chemically competent Rosetta cells. Cells were grown to an OD600 of ~0.6 and induced overnight at 16 °C with 0.5 mM isopropyl-β-D-thiogalactopyranoside (IPTG). Cells were harvested and resuspended in lysis buffer (20 mM HEPES, pH 7.5, 500 mM NaCl, 10 mM imidazole, 0.1% Triton X-100, 1 mM Tris (2-carboxyethyl)phosphine (TCEP), Complete EDTA-free protease inhibitor (Roche), 0.5 mM phenylmethylsulfonyl fluoride (PMSF), and 10% glycerol).

After cell lysis by sonication, lysate was clarified by centrifugation and the supernatant was incubated on Ni-NTA resin (Qiagen). The resin was washed with wash buffer (20 mM HEPES, pH 7.5, 500 mM NaCl, 10 mM imidazole, 1 mM TCEP, and 5% glycerol), before the protein was eluted with wash buffer supplemented with 300 mM imidazole and then digested with TEV protease overnight. The salt concentration was diluted to 335 mM NaCl using ion-exchange buffer A (20 mM HEPES, pH 7.5, 1 mM TCEP, and 5% glycerol), the cleaved MBP tag was removed with an MBPTrap column (GE Healthcare), and the protein was bound to a HiTrap heparin HP column (GE Healthcare), before elution with a gradient from 335 mM to 1 M KCl. The protein was then concentrated and purified on a Superdex 200 (16/60) column with storage buffer (20 mM HEPES, pH 7.5, 500 mM KCl, 1 mM TCEP, and 5% glycerol). The same purification protocol was used for Cas6-RT-Cas1 (wild-type and mutants) and Cas2. Sequences of proteins are shown in Table 3.1.

Protein	Sequence
Cas6-RT-Cas1	MILPSFPDLTGLVVNLKFTARAEFSLNHEMAVDAFLRHSLNLGESYSHHLSII TPENGRLFYREGDTYRFVVIAMGNQQQTNSIWHTLINHLRKLPDSAPITDK QAPLRNNIKLESLNDLFDGIPVSSKESLDAYTLQRAMEQGLAWHKAANLTE QPLDIQWYWQSTVRILHADHKQHKGEQRYCRDAVQLTPLLLLKRIYETLNN VATYFGLKTNKNTTENHQAWLKEQAQYIEIQHPDLYWIDTPYFGKDAEKNTL GMAGNFTLSLKPGIEPGLLAMLILTQMVGVGQRRTSGLGKYWLKHSLKHAH LILGLKPNRVTRSQTLDCIIQPHIISQAIAEIEKKTNIDTLNERTLSQVQSAIGQL RKHQYQAPKLQGFTIPKKDGTERLLAVSPLYDRILQKAAAIVLTPGLDAIMSQ ASYGYRKGLSRQQVRYEIQNAYRQGYHWVYESDIEDFFDAVYRPQLINRLK SLLGNDPLWEQIESWLGQDIHIKDTIIERTPNLGLPQGSPLSPLLANFILDDFD SDLETHGFKIIRFADDFIILCKSQHEAQQAAHAVEQSLKEVKLSINVEKTHIIQL NQGFRFLGYLFREDHAIEIAGEKSDGRTTFAAEQTPTNLPPWLANLGTKSP QPLADDDLPKKSYGQIETQGTHLVLAGDAQIITTDNQNLIVKKDDKITHKISLE QLHAVTLIGLHTMTLPAKHRLLEHKIPVHIADRTGRYLGAVTSFQPAQNNYK NWFIQLQMCDREPFAHAIAQQIVISRIHNQRQTLLKRKAHRKQLQQTLSNLK KLQYKVTAATKRSSLNGLEGSATREYFQQFNLFLPEWAHFSKRTRRPPKDP FNVLLSLGYTILYSHTDAILQSAGFITWKGIYHQQSAAHAALASDIMESYRHL VERYAIYIINHGQIKQDDFRQEKDHLGQDTIRLSAEARRRYVGGLINRFQKFS KDKTLHQHLYQQAQQLKNAMHNQQSSQFQVWKELK*
Cas2	MKHYLICFDVQHDKTRAKLSRLLEKYGPRVQGSVFEVSFKTPDRKRQLEYK IHQIIKQSNTEENNIRFYNLNKDTIKHSHDINGNPIAQLPAAIVL*

Table 3.1 | Sequences of type III *Thiomicrospira* Cas6-RT-Cas1 and Cas2 proteins.

3.5.4 DNA and RNA substrate preparation

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

generated by cloning a cat promoter and chloramphenicol resistance gene into a pUC19 plasmid, removing the RBS sequence and start codon, and then inserting a sequence carrying 163 bp of the leader from pCRISPR and the 35 bp repeat before the chloramphenicol resistance gene. DNA and RNA oligos used in this study were ordered from Integrated DNA Technologies (IDT) and purified on 6% (for >35-nt oligos) or 14% urea-PAGE (for ≤35-nt oligos). Protospacers, dsDNA targets, primer/template hybrids, and the half-site substrate were formed by heating at 95 °C for 5 min and slow cooling to room temperature in either HEPES hybridization buffer (20 mM HEPES, pH 7.5, 25 mM KCl, and 10 mM $MgCl_2$) or Tris hybridization buffer (20 mM Tris, pH 7.5, 25 mM KCl, and 10 mM $MgCl_2$), as consistent with the integration buffer in which the substrate will be used. For the half-site substrate, hybridization was carried out with a 1.25-fold excess of the shortest strand before purification on 8% native PAGE. Sequences of cloning primers and DNA and RNA substrates are shown in Table 3.2.

Name/Description	Sequence	Figure
Half-site substrate protospacer-repeat-spacer	GTAAAATCGAAGGAATGCTAAGGAAGCTAAAATGGTTCT AGAACACCAAAGCCGAAAGGCTGTTAAGACAGCT	3.1, 3.2
Half-site substrate protospacer	ATTTTAGCTTCCTTAGCATTCCTTCGATTTTAC	3.1, 3.2
Half-site substrate spacer-repeat-leader	AGCTGTCTTAACAGCCTTTCGGCTTTGGTGTTCTAGAA CTCCGGTTGAAATTCTATCTGCTTATTCAATGAGTTGCA ATGCATTTTCCGCACCAGAATGA	3.1, 3.2
Half-site substrate leader	TCATTCTGGTGCGGAAAATGCATTGCAACTCATTGAATA AGCAGATAGAATTTCAACCGGA	3.1, 3.2
Fluorescent half-site substrate protospacer-repeat-spacer	/56-FAM/GTAAAATCGAAGGAATGCTAAGGAAGCTAAAAT GGTTCTAGAACACCAAAGCCGAAAGGCTGTTAAGACAG CT	3.2
Fluorescent half-site substrate protospacer	/56-FAM/ATTTTAGCTTCCTTAGCATTCCTTCGATTTTAC	3.2
Fluorescent half-site substrate spacer-repeat-leader	/56-FAM/AGCTGTCTTAACAGCCTTTCGGCTTTGGTGTT CTAGAACTCCGGTTGAAATTCTATCTGCTTATTCAATGA GTTGCAATGCATTTTCCGCACCAGAATGA	3.2
Fluorescent half-site substrate leader	/56-FAM/TCATTCTGGTGCGGAAAATGCATTGCAACTCAT TGAATAAGCAGATAGAATTTCAACCGGA	3.2
Fluorescent 35-mer ssDNA protospacer and (+) strand for dsDNA protospacer and DNA template used for cDNA synthesis reactions	/56-FAM/TTACGGTGCTTAAAACAAAACAAAACAAAACAA AA	3.6, 3.8, 3.12 3.13, S8
RC for fluorescent and non-fluorescent dsDNA protospacer	TTTTGTTTTGTTTTGTTTTGTTTTAAGCACCGTAA	3.6, 3.7, 3.8, 3.12, 3.13, 3.14
Fluorescent 35-mer RNA protospacer and RNA template used for cDNA synthesis reactions	/56-FAM/UUACGGUGCUUAAAACAAAACAAAACAAAACA AAA	3.6, 3.7, 3.8, 3.13, 3.14
35-mer ssDNA protospacer and (+) strand for dsDNA protospacer and DNA template used for cDNA synthesis reactions	TTACGGTGCTTAAAACAAAACAAAACAAAACAAAA	3.6, 3.7, 3.8, 3.11, 3.12
35-mer RNA protospacer and RNA template used for cDNA synthesis reactions	UUACGGUGCUUAAAACAAAACAAAACAAAACAAAA	3.6, 3.7, 3.8, 3.11, 3.12
25-mer DNA strand for RNA/DNA hybrid	TTTTGTTTTGTTTTGTTTTGTTTTA	3.12
15-mer ssDNA protospacer and (+) strand for dsDNA	AGCTGTTGATAAGCA	3.9
RC for dsDNA protospacer	TGCTTATCAACAGCT	3.9
35-mer ssDNA protospacer and (+) strand for dsDNA	AGCTGTTAAGTACCCTTGGATCAGGCGGATAAGCA	3.9
RC for dsDNA protospacer	TGCTTATCCGCCTGATCCAAGGGTACTTAACAGCT	3.9
55-mer ssDNA protospacer and (+) strand for dsDNA	AGCTGTTAAGTACCAATGTTCCAGACCATCCCTTCTTGG ATCAGGCGGATAAGCA	3.9
RC for dsDNA protospacer	TGCTTATCCGCCTGATCCAAGAAGGGATGGTCTGGAAC ATTGGTACTTAACAGCT	3.9
75-mer ssDNA protospacer and (+) strand for dsDNA	AGCTGTTAAGTACCAATGTTCCAGACCATCCCTTCAAGT CTTTGTACTGCGTCCATTGGATCAGGCGGATAAGCA	3.9
RC for dsDNA protospacer	TGCTTATCCGCCTGATCCAATGGACGCAGTACAAAGAC TTGAAGGGATGGTCTGGAACATTGGTACTTAACAGCT	3.9
95-mer ssDNA protospacer and	AGCTGTTAAGTACCAATGTTCCAGACCATCCCAGTTAAC	3.9

(+) strand for dsDNA	GCATCCATAGCAATTCAAGTCTTTGTACTGCGTCCATTG GATCAGGCGGATAAGCA	
RC for dsDNA protospacer	TGCTTATCCGCCTGATCCAATGGACGCAGTACAAAGAC TTGAATTGCTATGGATGCGTTAACTGGGATGGTCTGGAA CATTGGTACTTAACAGCT	3.9
115-mer ssDNA protospacer and (+) strand for dsDNA	AGCTGTTAAGTACCAATGTCACGTACCATTCCAGACCAT CCCAGTTAACGCTTACGATTCAATCCATAGCAATTCAAG TCTTTGTACTGCGTCCATTGGATCAGGCGGATAAGCA	3.9
RC for dsDNA protospacer	TGCTTATCCGCCTGATCCAATGGACGCAGTACAAAGAC TTGAATTGCTATGGATTGAATCGTAAGCGTTAACTGGGA TGGTCTGGAATGGTACGTGACATTGGTACTTAACAGCT	3.9
Full-site assay protospacer	GTAAAATCGAAGGAATGCTAAGGAAGCTAAAATG	3.10
RC	CATTTTAGCTTCCTTAGCATTCCTTCGATTTTAC	3.10
Fluorescent 25-mer DNA primer used for cDNA synthesis reactions	/56-FAM/TTTTGTTTTGTTTTGTTTTGTTTTA	3.11, 3.12
Fluorescent RNA repeat	/56-FAM/GUUCUAGAACACCAAAGCCGAAAGGCUGUUA AGAC	3.13, 3.14
Linear DNA target (+) strand	TTGGTCTTAACAGCTGTCTTAACAGCCTTTCGGCTTTGG TGTTCTAGAACTCCGGTTGAAATTCTATCTGCTTATTCAA TGAGTTGCAATGCATTTTCC/3ddC/	3.8
RC for fluorescent and non-fluorescent linear DNA target	GGAAAATGCATTGCAACTCATTGAATAAGCAGATAGAAT TTCAACCGGAGTTCTAGAACACCAAAGCCGAAAGGCTG TTAAGACAGCTGTTAAGACCAA/3ddC/	3.8
Fluorescent linear DNA target (+) strand	/56-FAM/TTGGTCTTAACAGCTGTCTTAACAGCCTTTCGG CTTTGGTGTTCTAGAACTCCGGTTGAAATTCTATCTGCT TATTCAATGAGTTGCAATGCATTTTCC/3ddC/	3.8
pCRISPR_full-site cloning primer 1	AAAGCCGAAAGGCTGTTAAGACAGAGAAAAAAATCACT GGATATACCACCGTTG	
pCRISPR_full-site cloning primer 2	CACCATATTTTAATGTTGATTTAGCTCCTGAAAATCTCGA TAACTCAAAAAATACG	
pCRISPR_full-site cloning primer 3	TTTTTGAGTTATCGAGATTTTCAGGAGCTAAATCAACATT AAAATATGGTGCGGGAACT	
pCRISPR_full-site cloning primer 4	AACGGTGGTATATCCAGTGATTTTTTTCTCTGTCTTAACA GCCTTTCGGCTTT	
pCRISPR cloning primer 1	TAACGCGGCTGTCTTGTATAACAATTTCACACAGGAAAC AGCTATGACC	
pCRISPR cloning primer 2	GATAACTTTGTGGCGTTTTGGTGCCGCTGTGCGGTTTG CGTATTGGGC	
pCRISPR cloning primer 3	AAGCGGAAGAGCGCCCAATACGCAAACCGCACAGCGG CACCAAAACG	
pCRISPR cloning primer 4	AGCCTTTCGGCTTTGGTGTTCTAGAACAGTACAAAGAC TTGAAGGGATG	
pCRISPR cloning primer 5	AGTCTTTGTACTGTTCTAGAACACCAAAGCCGAAAGGC TGTTAAGACTC	
pCRISPR cloning primer 6	TCATAGCTGTTTCCTGTGTGAAATTGTTATACAAGACAG CCGCGTTAGC	
pCas6-RT-Cas1_expression cloning primer 1	TTCAAGTATGGAAGGAGTTGAAGTGAATTGGAAGTGGA TAACGGATCCG	
pCas6-RT-Cas1_expression cloning primer 2	GGGAAAGACGGAAGAATCATAGCATTGGATTGGAAGTA CAGGTTTTCCT	
pCas6-RT-Cas1_expression cloning primer 3	GAAAACCTGTACTTCCAATCCAATGCTATGATTCTTCCG TCTTTCCCCG	
pCas6-RT-Cas1_expression	GATCCGTTATCCACTTCCAATTCACTTCAACTCCTTCCA	

cloning primer 4	TACTTGAAAC	
pCas2_expression cloning primer 1	AATTACCTGCTGCAATCGTTCTATAAATTGGAAGTGGATAACGGATCCG	
pCas2_expression cloning primer 2	CAAATCAGGTAGTGTTTCATAGCATTGGATTGGAAGTACAGGTTTTCCT	
pCas2_expression cloning primer 3	TGTACTTCCAATCCAATGCTATGAAACACTACCTGATTTGCTTTGATGT	
pCas2_expression cloning primer 4	ATCCGTTATCCACTTCCAATTTATAGAACGATTGCAGCAGGTAATTGAG	
pCas6-RT-Cas1H873A_expression cloning primer 1	GCACAGCAGTCGGCCGC	
pCas6-RT-Cas1H873A_expression cloning primer 2	ATAGATGCCTTTCCATGTGATAAACCC	
pCas6-RT-Cas1D540A_expression cloning primer 1	GCAGATTTCATTATTTTATGTAAGTCACAACATGAAGCCC	
pCas6-RT-Cas1D540A_expression cloning primer 2	AGCAAAACGGATGATCTTGAACCC	
pCas6-RT-Cas1R37A_expression cloning primer 1	GCACACTCTTTAAATTTGGGCGAGAGCTA	
pCas6-RT-Cas1R37A_expression cloning primer 2	TAAGAATGCATCCACAGCCATTTCG	
pCas6-RT-Cas1K574A_expression cloning primer 1	GCAACGCACATCATTCAGCTGAAC	
pCas6-RT-Cas1K574A_expression cloning primer 2	GCAACGCACATCATTCAGCTGAAC	
pCas6-RT-Cas1R835A_expression cloning primer 1	GCACCTCCTAAAGATCCTTTCAATGTGCTG	
pCas6-RT-Cas1R835A_expression cloning primer 2	ACGTGTGCGCTTCGAAAAATG	

Table 3.2 | DNA and RNA oligonucleotides used in this study. RC indicates the complementary strand of the previous oligonucleotide. Sequences are written 5' to 3'.

3.5.5 Integration assays

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

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

For full-site integration assays, pCRISPR_full-site is used as the target plasmid. Sixteen 50 uL integration reactions were carried out as described above with the reporter construct for 2 h. Instead of quenching the reaction, the integration products were purified using the Qiagen MinElute PCR Purification Kit and electroporated into DH10B cells. Transformants were plated on LB agar containing chloramphenicol and 95 of the surviving colonies were sequenced using Sanger Sequencing and analyzed using SnapGene Version 5.0.8.

3.5.6 RT and Cas6 activity assays

RT activity was measured using an ELISA-based colorimetric reverse transcriptase activity assay (Catalog No. 11468120910, Roche Diagnostics, Indianapolis, IN). The HEPES integration buffer was used for the reaction. Cas6-RT-Cas1 and Cas2 were first pre-complexed at 2 μM at room temperature for 30 min. The reactions were conducted in triplicates using the supplied template/primer

hybrid poly (A) × oligo (dT)15 at 37 °C for 2 h. After following the assay kit's ELISA procedure, the RT activity was measured in absorption units (A405 nm–A490 nm) using a fluorescence plate reader (BioTek). Statistical analyses were performed using Prism 7 version 7.0c (GraphPad). Data were presented as the mean ± SD (error bars) of three independent experiments. Significance was assessed by comparing samples to the respective WT controls using unpaired, two-tailed T tests (α = 0.05). For template-driven cDNA synthesis reactions off a fluorescent DNA primer annealed to a DNA or RNA template, Cas6-RT-Cas1 and Cas2 were first pre-complexed at 2 μM at room temperature for 30 min before incubating with 250 nM primer/template substrate and 1 mM dNTPs, where indicated, at 37 °C for 2 h. For Cas6 activity assays, Cas6-RT-Cas1 and Cas2 were first pre-complexed at 2 μM at room temperature for 30 min. before incubating with the 5'-labeled 35-nt RNA substrate corresponding to the repeat sequence at 37 °C for 2 h. The template-driven cDNA synthesis reactions and Cas6 processing reactions were quenched by adding 2 vol. quench buffer and heating at 95 °C for 4 min., before analysis on a 14% urea-PAGE gel. Fluorescent bands were visualized by Typhoon FLA gel imaging scanner and intensities were quantified using ImageQuantTL 8.2. The percent Cas6 activity is quantified as the ratio of the product band intensity to the total intensity of both the product band and the unprocessed RNA substrate band.

3.5.7 Electrophoretic mobility shift assays

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

3.5.8 Grid preparation

Cas6-RT-Cas1—Cas2 and the half-site DNA substrate were complexed by mixing 50 μM Cas6-RT-Cas1, 50 μM Cas2, and 12.5 μM DNA half-site substrates in storage buffer and dialyzing in complex buffer (10 mM HEPES, pH 7.5, 5 mM EDTA, 500 mM KCl, 0.5 mM TCEP, and 0.5% glycerol) for 2 h using the Slide-A-Lyzer MINI Dialysis Devices at room temperature. The complex was concentrated to 100 μM Cas6-RT-Cas1—Cas2 and purified over a Superose 6 Increase 10/300 GL column. For

freezing grids, a 3 µl drop of protein was applied to freshly glow discharged Holey Carbon, 300 mesh R 1.2/1.3 gold grids (Quantifoil, Großlöbichau, Germany). A FEI Vitrobot Mark IV (ThermoFisher Scientific) was used with 4°C, 100% humidity, one blot force, and a 3 second blot time. Grids were then clipped and used for data collection.

3.5.9 Cryo-EM data acquisition

Grids were clipped and transferred to a FEI Titan Krios electron microscope operated at 300 kV. In total, 50 frame movies were recorded on a Gatan K3 Summit direct electron detector in super-resolution counting mode with pixel size of 0.5953 Å. The electron dose was 10.50 e^{-} $Å^2$ s^{-1} and total dose was 49.98 e^{-} $Å^2$. Initial data collection was performed with a GIF Quantum energy filter inserted, however after 516 micrographs had been collected it became unstable and was removed for the final 2729 micrographs. Nine movies were collected around a central hole position with image shift and defocus was varied from −1 to −3 µm through SerialEM[119]. See Table 3.3 for data collection statistics.

	Cas6-RT-Cas1—Cas2 (masked partial complex) (EMD-22855) (PDB 7KFT)	Cas6-RT-Cas1—Cas2 (full complex) (EMD-22856) (PDB 7KFU)
Data collection and processing		
Magnification	64,000x	64,000x
Voltage (kV)	300	300
Electron exposure (e–/Å^2)	49.98	49.98
Defocus range (μm)	1.0 - 3.0	1.0 - 3.0
Pixel size (Å)	1.187	1.187
Symmetry imposed	C1	C1
Initial particle images (no.)	249,102	249,102
Final particle images (no.)	129,175	129,175
Map resolution (Å)	3.4	3.7
FSC threshold	0.143	0.143
Map resolution range (Å)	3.35-5.68	3.39-9.07
Refinement		
Initial model used (PDB code)	7KFT	7KFU
Model resolution (Å)	3.3 / 3.5	3.7 / 4.0
FSC threshold	FSC = 0.143 / FSC = 0.5	FSC = 0.143 / FSC = 0.5
Map sharpening *B* factor (Å^2)	-111.2	-66.03
Model composition		
Non-hydrogen atoms	11680	21848
Protein residues	1427	2674
Ligands	0	0
B factors (Å^2)		
Protein	57.06	183.8
R.m.s. deviations		
Bond lengths (Å)	0.005	0.006
Bond angles (°)	0.824	0.821
Validation		
MolProbity score	1.63	1.68
Clashscore	5.54	6.27
Ramachandran plot		
Favored (%)	95.24	95.15
Allowed (%)	4.76	4.85
Disallowed (%)	0	0

Table 3.3 | Cryo-EM data collection, processing, and refinement.

3.5.10 Cryo-EM data processing

Motion-correction and dose-weighting were performed on all 3330 movies 2x binned to 1.187 Å per pixel using RELION 3.0's implementation of MotionCor2[120,121]. CTFFIND-4.1 was used to estimate the contrast transfer function (CTF) parameters[122].

Micrographs were then manually examined to remove subjectively bad micrographs, such as empty or contaminated holes, resulting in 2035 micrographs[123]. In addition, micrographs with a CTF maximum resolution lower than 4 Å were discarded, yielding 1998 micrographs. Template-free auto-picking of particles was performed with RELION3.1's Laplacian-of-Gaussian (LoG) filter to generate an initial set of particles which were iteratively classified to find a subset of particles that were subsequently used to template-based auto-pick 249,102 particles.

Template picked particles were iteratively 2D-classified in RELION3.0 resulting in a set of 129,175 particles. These particles were imported to cryoSPARC v2. An initial map was generated with an ab-initio job and refined with subsequent homogeneous and non-uniform refinement jobs[124,125]. 'csparc2star.py' from UCSF pyem was used to convert the resulting data to a RELION star file[126]. These particles were then refined in RELION3.0 resulting in a map with ~4.1 Å overall resolution. Subsequent refinement following Bayesian particle polishing, and CTF refinement resulted in a map with ~3.66 Å overall resolution[127]. A mask surrounding the stable side of the complex was generated and supplied to a refinement resulting in a map with ~3.42 Å overall resolution. Further classification in 3D was attempted but did not result in an improved map, nor did the application of C2 symmetry during refinement.

This set of 129,175 particle coordinates were used for training in the Topaz particle-picking pipeline[128]. Training, picking, and extraction in Topaz resulted in 1,305,497 particles. These particles were then processed using a similar pipeline to the above Relion template-picked particles, resulting in 285,342 particles after 2D classification and yielding a ~3.8 Å map before, and a ~3.3 Å map following Bayesian particle polishing. The final maps were not distinguishably improved compared to the template-picked particle set. As with the prior particle set, further classification in 3D could not identify a subset of particles yielding an improved map. A consensus homogeneous refinement of these particles in cryoSPARCv2 was used as input for 3D variability analysis in cryoSPARCv2 (filter resolution 5 Å)[129].

3.5.11 Modeling, refinement, and analysis

De novo modeling of the best-resolved portion of the complex was performed in Coot using the output map from a masked 3D refinement in Relion sharpened with Phenix.auto_sharpen[130,131]. The de novo modeled structure, corresponding to Cas2 chain A (amino acids 1–90), Cas2 chain B (amino acids 1–96), Cas6-RT-Cas1 chain C (amino acids 1–634), Cas6-RT-Cas1 chain D (amino acids 650–981), and Cas6-RT-Cas1 chain E (amino acids 641–981), was refined using Phenix.real_space_refine. Molprobity was used to guide iterative rounds of manual adjustment in Coot and refinement in Phenix[132]. Subsequently, portions of the refined model were copied and rigid-body docked into density corresponding to less-well

resolved portions of the complex (corresponding to Cas2 chain A (amino acids 91–96), Cas6-RT-Cas1 chain C (amino acids 641–981), Cas6-RT-Cas1 chain E (amino acids 1–634), and Cas6-RT-Cas1 chain F (amino acids 650–981)) in a postprocessed Relion 3D refinement output map that was masked to include the entire visible complex. A final round of rigid-body refinement was performed in Phenix.real_space_refine. This model consists of two entire Cas2 protomers (amino acids 1–96), two entire Cas6-RT-Cas1 protomers (amino acids 1–981), and two partial Cas6-RT-Cas1 protomers with only the Cas1 domain modeled (amino acids 650–981). Six loops were unmodeled in Cas6-RT-Cas1 protomers: three in the Cas6 domain (amino acids 95–110, 248–257, and 212–221), two in the RT domain (amino acids 383–388 and 593–616), and a linker between the RT and Cas1 domains (amino acids 635–640). Figures were prepared using Chimera, ChimeraX, Prism 7 version 7.0c (GraphPad), GNU Image Manipulation Program, and Adobe Illustrator 2020 software[133,134].

3.6 Accession codes

The atomic model of the partial Cas6-RT-Cas1—Cas2 complex masked for the stable density is in the Protein Data Bank (PDB) under 7KFT and the corresponding map is deposited in the Electron Microscopy Data Bank (EMDB) under EMD-22855. The atomic model of the full Cas6-RT-Cas1—Cas2 complex is in the Protein Data Bank (PDB) under 7KFU and the corresponding map is deposited in the Electron Microscopy Data Bank (EMDB) under EMD-22856. The original micrograph movies and final particle stack are deposited in the Electron Microscopy Public Image Archive (EMPIAR) under EMPIAR-10642.

3.7 Acknowledgements

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

3.8 Author contributions

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

CHAPTER 4

Molecular mechanisms of prespacer processing for CRISPR adaptation

A portion of the work presented in this chapter is part of the following manuscript currently in preparation: Wang, J.Y., Skopintsev, P., Tuck, O., Li, G. & Doudna, J.A. Molecular mechanisms of prespacer processing for CRISPR adaptation.

4.1 Abstract

CRISPR-Cas immunity protects bacteria from foreign invaders by recording memory of previous infections as spacers in the host CRISPR array. The coordinated selection and processing of DNA fragments flanked by a protospacer adjacent motif (PAM) for spacer acquisition is fundamental for the generation of CRISPR immunological memory. Here, we investigate prespacer processing and integration by a natural DEDDh exonuclease fusion to the Cas1—Cas2 CRISPR integrase and provide a model for how host exonucleases aid in spacer acquisition. We determine that the exonuclease and Cas1—Cas2 work together to process prespacers in a ruler-guided and PAM-dependent manner to generate viable integration substrates and carry out integration in a stepwise manner. We elucidate the mechanism of PAM processing and subsequent full-integration, which is exquisitely controlled by a gate-keeping ruler mechanism preventing PAM insertion. This represents an alternative model for PAM processing employed by CRISPR systems lacking the Cas4 endonuclease. Finally, we probe into the substrate preferences for selective PAM-dependent capture and processing by Cas1—Cas2-DEDDh and provide insight into the nuances of substrate selection for CRISPR spacer acquisition.

4.2 Introduction

CRISPR (clustered regularly interspaced short palindromic repeats) adaptive immune systems protect bacteria from viral or phage infection[135]. These systems generate immunological memory of previous infection by integrating small DNA segments—protospacers—from invading genetic material into the CRISPR array, consisting of short repeats and foreign-derived spacers (~25-50bp)[1,6,21,69]. The CRISPR array is transcribed and processed into mature CRISPR RNAs (crRNAs), which are recruited by a single effector or effector complex to trigger a robust immune response upon reinfection[70–75].

In almost all CRISPR systems, the Cas1—Cas2 CRISPR integrase carries out protospacer integration[17,24,26]. The Cas1—Cas2's heterohexameric structure provides an internal ruler mechanism that dictates the length of integrated spacers[10,16,17]. Although the mechanisms for protospacer integration have been characterized extensively[6,45], the processes upstream of integration, including the capture and processing of foreign DNA fragments, called prespacers, are not as well understood. In many CRISPR systems, protospacer selection depends on the presence of a flanking 3-5 bp protospacer adjacent motif (PAM) that is removed prior to integration and enables the downstream interference complex to distinguish between self and non-self during target identification[7,85,87]. PAM processing before integration into the CRISPR array is thus critical to avoid an autoimmune response. Prior to its removal, the PAM helps direct the

orientation of spacer insertion in order to produce functional crRNAs for the downstream immune response[50,64,136]. How sequences from invading genetic material are selected and processed for integration is fundamental to the generation of CRISPR immunological memory.

Despite being the most conserved component of CRISPR-Cas systems, the CRISPR adaptation module has recruited different enzymes and evolved diverse mechanisms for PAM selection and prespacer processing across different CRISPR systems[9,49–51,54,55]. For a number of CRISPR systems, the Cas4 endonuclease performs the role of PAM selection and PAM processing[30,51,52,55]. However, Cas4 is absent in many CRISPR systems. How systems lacking Cas4 select prespacers and process the PAM is still unclear. It has been shown that in certain CRISPR systems, such as the type I-E *E. coli* system, the Cas1 contains a PAM-binding pocket and is believed to play a role in prespacer selection[16,49]. However, there have been ambiguous results whether Cas1 in these systems is capable of removing the PAM in a similar manner to Cas4 or if it relies on host exonucleases to perform this function[16,49]. To produce viable integration substrates, CRISPR systems need to first process prespacers to the integration-competent size. Recent *in vitro* studies have identified certain exonucleases capable of aiding Cas1—Cas2 to process prespacer substrates to the correct size for integration[49,50]. There are also a handful of systems containing a natural Cas2-DEDDh exonuclease fusion[54,137], further suggesting a functional link between exonucleases and the CRISPR integrase. These systems provide a good model for studying the coordination between host exonucleases and the CRISPR integrase during prespacer selection and processing in CRISPR systems that lack Cas4.

In this study, we reconstitute ruler-guided prespacer processing by a naturally occurring *Megasphaera NM10_related* Cas1—Cas2-DEDDh complex. We find that the complex initially protects the PAM from trimming but removes the PAM for full-site integration. We identify that the DEDDh active site is responsible for both the initial trimming of the excess 3’ overhangs and removing the PAM in the half-site intermediate, providing some clarification on previously ambiguous results about whether the Cas1 or host exonucleases are responsible for this function in similar CRISPR systems[16]. Our experiments show that PAM processing is necessary to facilitate full-site integration. Despite the lack of sequence specificity of the exonuclease trimming, the DEDDh and integrase coordinate their activities to fully remove the PAM and produce a specific integration product, suggesting a gate-keeping mechanism that simultaneously prevents PAM insertion and precisely specifies spacer size. Our results show that delayed PAM processing is a common theme across many CRISPR systems even though different enzymes and mechanisms are utilized to perform this function. Finally, we examine prespacer selection by the Cas1—Cas2-DEDDh complex *in vitro* and surprisingly, find that in the canonically preferred prespacer form (a 23-bp duplex with long single-stranded 3’ overhangs), the complex exhibited lower integration efficiency from

PAM-containing prespacers compared to PAM-deficient ones, contrary to the previous model. To gain insight into PAM selection, we investigate Cas1—Cas2-DEDDh's substrate preferences for PAM-dependent capture and processing and find that Cas1—Cas2-DEDDh can selectively capture and process certain PAM-containing ssDNA substrates. We propose that PAM selection occurs further upstream during prespacer biogenesis and involves additional nuances. These findings underscore that CRISPR adaptation has evolved many elegant mechanisms during prespacer biogenesis, selection, and processing to maximize efficiency in producing robust immunity against foreign invaders while avoiding autoimmunity.

4.3 Results

4.3.1 Cas1—Cas2-DEDDh processes prespacers to correct size for integration

To investigate the coordination between the DEDDh exonuclease and Cas1—Cas2 integrase during prespacer processing, we expressed and purified the Cas1 and Cas2-DEDDh proteins from a type I-E *Megasphaera NM10_related* CRISPR system and tested their processing activities *in vitro* (Fig. 4.1a-b). Based on the size of the spacers in the CRISPR array and the substrate preferences of the closely related type I-E *E. coli* Cas1—Cas2, we hypothesized that the preferred integration substrate is likely a 23-bp DNA duplex with 5-nt single-stranded 3' overhangs[10,16]. For the processing assay, we tested various 5' fluorophore-labeled prespacer substrates (containing a 23-bp duplex region and extended single-stranded 3' overhangs of varying lengths) and examined trimming of the 3' overhangs by the Cas1 and Cas2-DEDDh proteins (Fig. 4.1c-d). The results show that the Cas1—Cas2-DEDDh complex processes the prespacers to the correct size for integration. It trims the excess nucleotides in the 3' overhangs of both strands and generates a 23-bp duplex with 5-6 nt single-stranded 3' overhangs, the likely preferred integration substrate. Interestingly, Cas1 and Cas2-DEDDh each exhibit some exonuclease activity alone, yielding distinct products that do not appear to be functionally relevant (Fig. 4.1d). Only the full Cas1—Cas2-DEDDh complex exhibits the ruler mechanism producing integration-competent substrate sizes. To confirm that the DEDDh active site is responsible for the exonuclease processing activity in the complex (as opposed to the Cas1 active site), we expressed and purified a catalytically dead DEDDh mutant, Cas2-DEDDh(D132A), complexed it to Cas1, and tested the complex's prespacer processing activity. We do not observe any ruler-guided processing activity with this mutant complex (Fig. 4.1e). Together, these results indicate that the full Cas1—Cas2-DEDDh complex is necessary for prespacer processing to the correct size and that the DEDDh active site provides the exonuclease activity for prespacer processing.

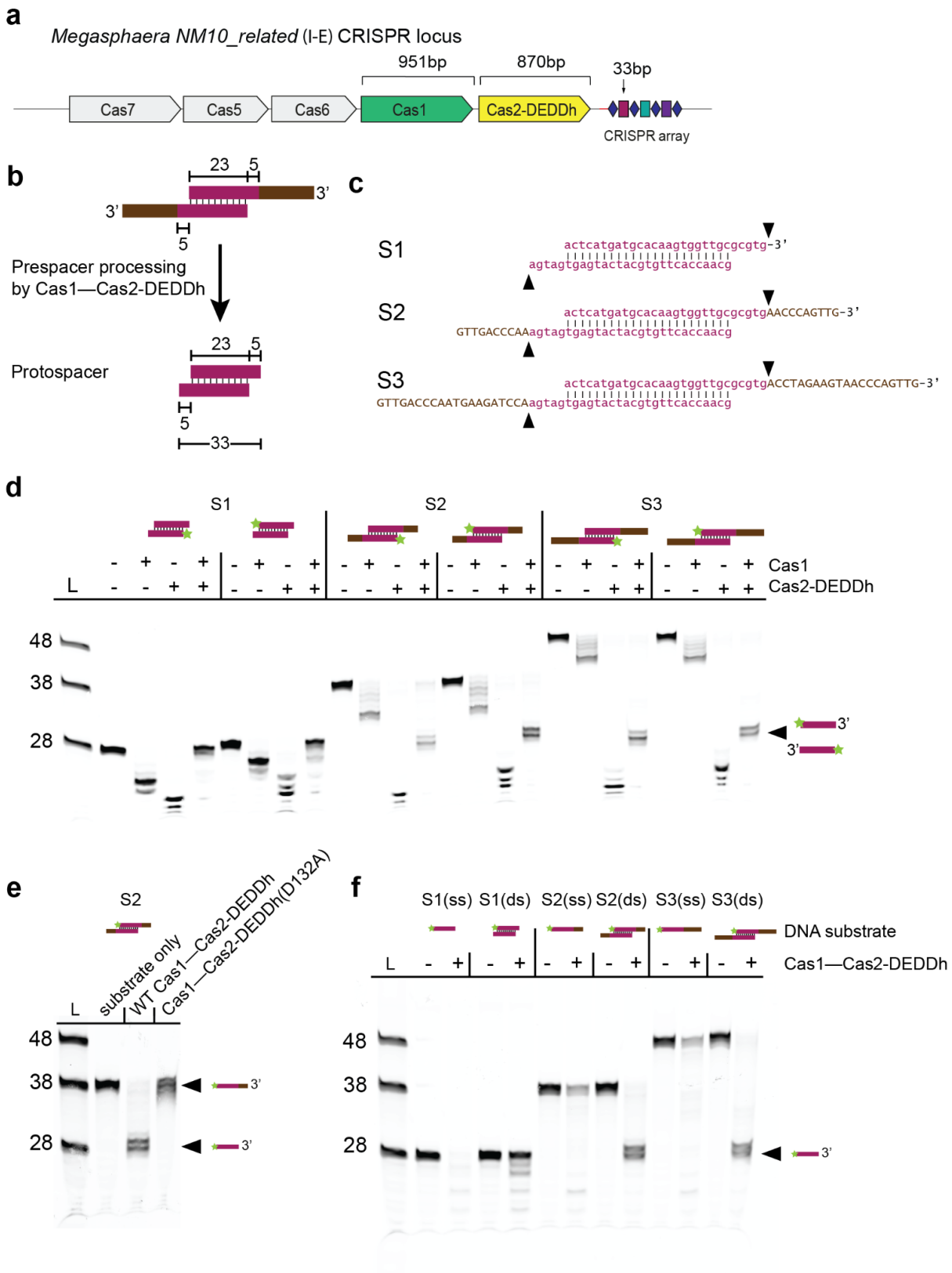


Figure 4.1 | Cas1—Cas2-DEDDh trims duplex prespacers to integration-competent size. (a) Illustration of the CRISPR genomic locus encoding the proteins used in this study. **(b)** Schematic of

processing reaction by Cas1—Cas2-DEDDh. Protospacer sequence is shown in magenta with lengths of duplex and single-stranded regions indicated. Excess nucleotides in single-stranded 3' overhangs of prespacers are shown in brown. **(c)** Substrates used in processing assay, following the same coloring scheme as in **b**. Arrowhead indicates how far substrate strand gets trimmed during processing reaction. **(d)** Processing assay with substrates represented in **c**. Star indicates 6-carboxyfluorescein label. Expected processing products of integration-competent size are indicated on the right. **(e)** Processing assay with wild-type (WT) and mutant Cas1—Cas2-DEDDh (DEDDh-D132A) with substrate S2, represented in **c**. Schematized substrate and product strands are indicated on the right. **(f)** Processing assay with single-stranded controls (ss represents top strand of substrate represented in **c**, ds represents duplex substrate represented in **c**.

To examine whether Cas1—Cas2-DEDDh can use ruler-guided trimming to process ssDNA substrates, we tested processing of the top strands of each prespacer substrate as single-stranded controls (Fig. 4.1f). The results show no indication of ruler-guided trimming of the ssDNA substrates. The Cas1—Cas2-DEDDh degrades the 28-nt ssDNA completely but is less efficient at degrading the longer ssDNA substrates. The results suggest that the duplex region of the prespacer is important for Cas1—Cas2-DEDDh's ruler-guided processing, potentially due to better binding of the duplex DNA to the protein complex compared to ssDNA.

We next examined the kinetics of ruler-guided exonuclease trimming and substrate ligation into an integration target plasmid by Cas1—Cas2-DEDDh. Time-course assays show similar processing efficiencies for the two prespacer substrates with varying overhang lengths: about 30% of the substrates are processed to the correct size for integration after 10 minutes and 60% of the substrates are processed to the correct size after 1 hour (Fig. 4.2a,b). For the ligation assay, we used fluorescent prespacers and examined the incorporation of the fluorophore into an integration target plasmid pCRISPR containing a shortened version of the natural CRISPR array (Fig. 4.2c). As expected, Cas1—Cas2-DEDDh shows relatively higher ligation efficiency with the canonically preferred integration substrate (the 23-bp duplex with 5-nt single-stranded 3' overhangs) compared to the prespacers with extended overhangs (Fig. 4.2d). For the canonical integration substrate, we observe a faint band representing the ligation product appearing after 2 min. For the substrates with the extended overhangs, we do not observe the ligation product until after 10 min. The results are consistent with the model that Cas1—Cas2-DEDDh trims prespacers to the integration-competent size before ligation into the CRISPR array.

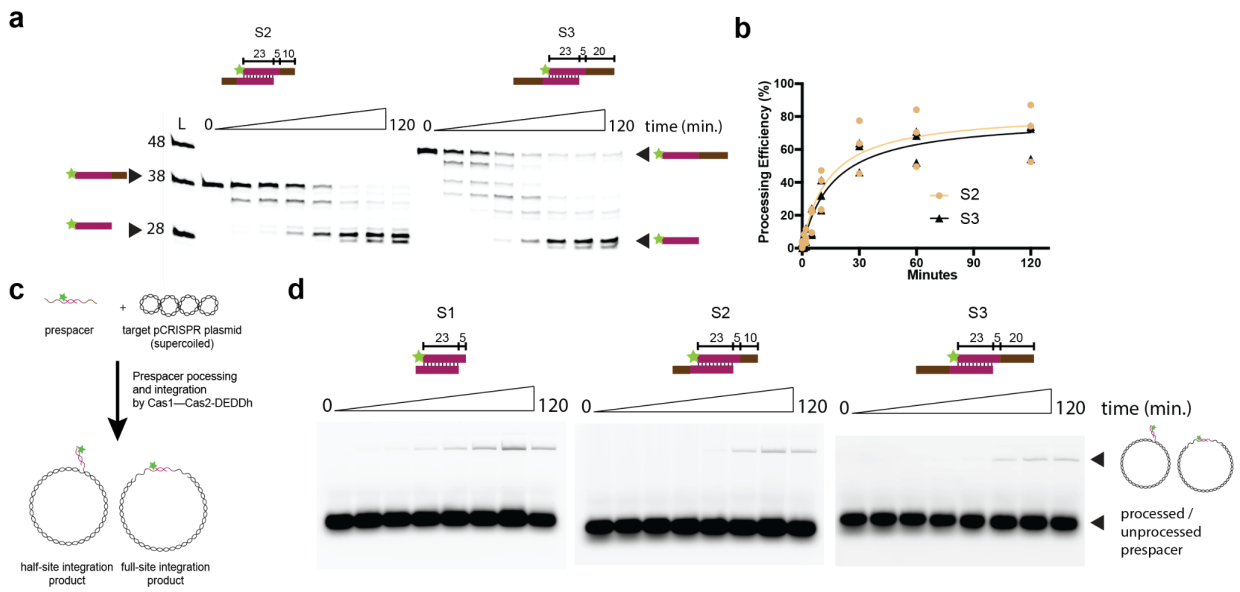


Figure 4.2 | Kinetics of exonuclease trimming and ligation by Cas1—Cas2-DEDDh *in vitro*. (a) Time-course reactions of ruler-guided trimming by Cas1—Cas2-DEDDh using substrates S2 and S3 from Fig. 4.1c over two hours. **(b)** Quantification of time-course reactions of ruler-guided trimming shown in **a**. Processing efficiency is calculated as the percentage of fully processed product formation at 28-29 nt (n = 3 biologically independent experiments). **(c)** Schematic of *in vitro* ligation assay with integration target pCRISPR plasmid and prespacer substrate. Star indicates 6-carboxyfluorescein label. **(d)** Time-course reactions of ligation assay with substrates from Fig. 4.1c over 2 hours. Ligation products are indicated.

4.3.2 Initial PAM protection delays PAM processing for integration

We next investigated how the presence of a PAM motif would affect prespacer processing in this system. To test this, we varied the sequence of the 3' overhang in the prespacers, where we expect the PAM to be and tested the substrates in a processing assay (Fig. 4.3a). The results show that the Cas1—Cas2-DEDDh recognizes a TT PAM motif in the PAM site (adjacent to the protospacer sequence, 5 nt away from the duplex region) in the single-stranded 3' overhang and protects the PAM from trimming (Fig. 4.3b). In the absence of a TT PAM, the prespacer strands are trimmed fully to the integration competent size (28 nt), guided by the ruler mechanism observed earlier. However, when there is a TT PAM in the correct position (nucleotide positions 29 and 30 relative to the 5' end), the results show only partial trimming of the PAM-containing strand. Trimming ends precisely 3 nt away from the PAM, resulting in a 33-nt product. We hypothesize that the partial trimming can be attributed to a PAM binding pocket in Cas1, which protects the PAM and the 3 adjacent nucleotides from trimming by the exonuclease. Sequence alignment shows the Cas1 in this system closely resembles the *E. coli* Cas1 and importantly, contains the conserved PAM binding residues that have sequence specific interactions with the TT sequence in the *E. coli* PAM (CTT)[16]. PAM sequestration by Cas1 can explain the observed PAM-dependent trimming of the

prespacer substrates by Cas1—Cas2-DEDDh.

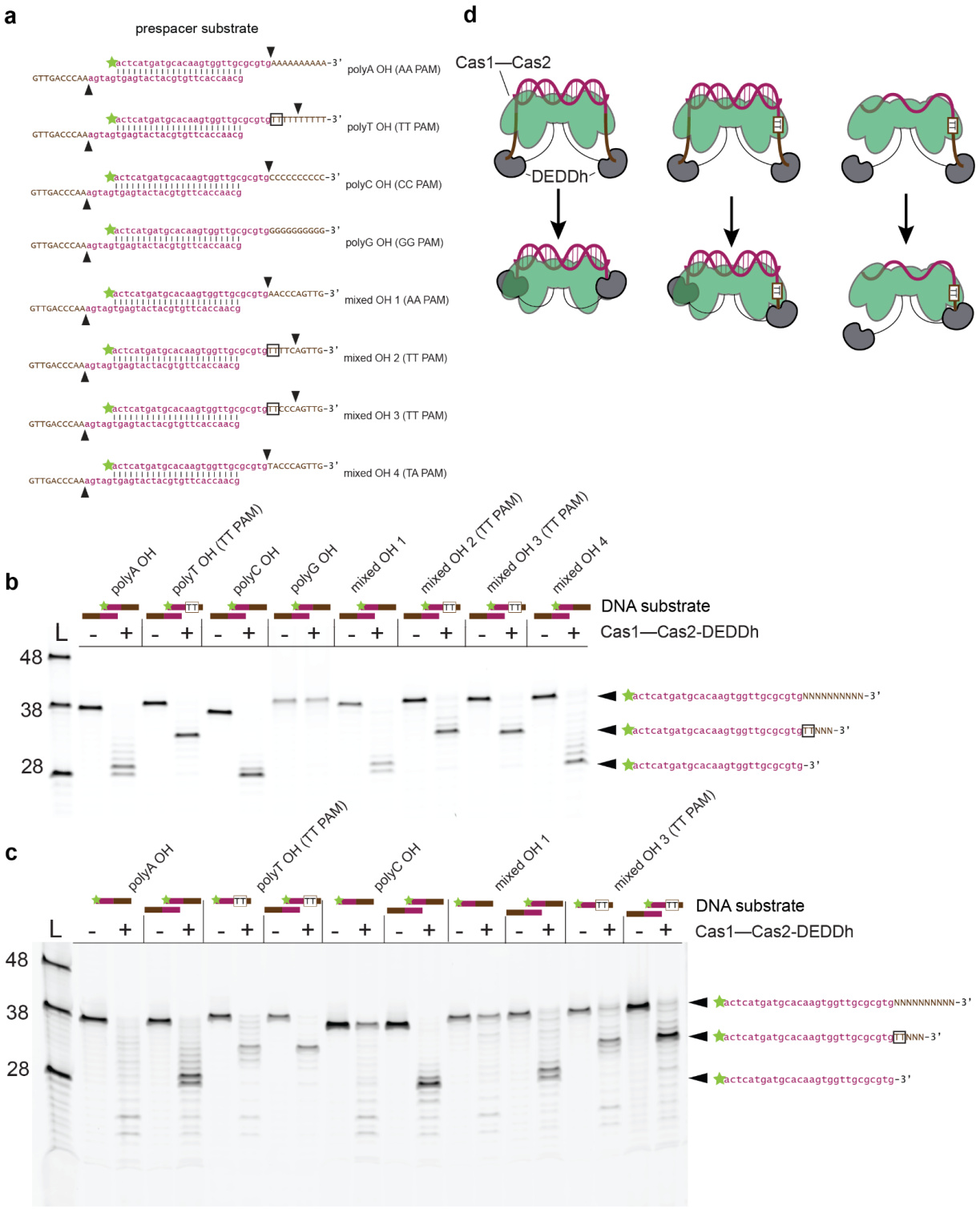


Figure 4.3 | TT PAM motif is initially protected from trimming. (a) Substrates used in processing assay with varied 3' overhang sequences. TT PAM is boxed. Arrowhead indicates how far substrate

strand gets trimmed during processing reaction. Star indicates 6-carboxyfluorescein label. **(b)** Processing assay with substrates shown in **a**. Substrates are schematized and TT PAM is indicated. Processing products are depicted on the right, TT PAM is boxed. **(c)** Processing assay with select substrates in **a** and their single-stranded DNA controls, TT PAM is boxed. **(d)** Model of ruler-guided and PAM-dependent trimming by Cas1—Cas2-DEDDh.

Given the importance of the duplex region for ruler-guided trimming in the previous assay, we wondered whether it was important to have a duplex region adjacent to the PAM in order for Cas1—Cas2-DEDDh to perform PAM-dependent trimming. To investigate this, we took a subset of the prespacer substrates, some containing the TT PAM and others without, and tested them alongside their single-stranded DNA (ssDNA) controls: the top strand of each substrate. Surprisingly, we found that Cas1—Cas2-DEDDh also seems to perform TT-dependent trimming of ssDNA, recognizing the same TT motif in the PAM position (designated as nucleotide positions 29 and 30 relative to the 5' end) as in the associated duplex (Fig. 4.3c). In the absence of a TT motif in the PAM position, Cas1—Cas2-DEDDh performs ruler-guided trimming of the duplex prespacers but no ruler-guided or sequence-dependent trimming of the ssDNA controls. However, when the TT motif is present in the designated PAM position, we observe the same TT-dependent trimming of both the duplex prespacer and the ssDNA control: trimming stops precisely 3 nts away from the TT motif (Fig. 4.3c,d). Curiously, even if there are multiple TT sequences in the substrates, the complex only appears to recognize a specific TT motif in the designated PAM position. It is unclear how the complex is directed to the specific TT motif, especially in the ssDNA substrates where there is no proximal duplex to help define the PAM position. We hypothesize there are other sequence motifs or structural features that contribute to specific PAM recognition and PAM-dependent trimming in ssDNA. Together, these results indicate that Cas1—Cas2-DEDDh processes prespacers in a ruler-guided and PAM-dependent manner (Fig. 4.3d).

Although the prespacer processing assays show PAM protection from trimming, the PAM needs to be removed prior to insertion into the CRISPR array in order to avoid self-targeting by the effector complex in the downstream immune response[7]. To investigate whether the Cas1—Cas2-DEDDh complex carries out PAM processing for integration, we carried out an *in vitro* integration assay with the target integration plasmid pCRISPR and a strategically designed unprocessed prespacer (a 23-bp duplex with 15-nt single-stranded overhangs) containing the TT PAM motif (Fig. 4.4a). The integration assay uses a chloramphenicol-based selection[58]. The prespacer is designed with BsaI cut sites in the duplex region such that a chloramphenicol selection cassette can be inserted after integration and be used to select for integration events. We performed the *in vitro* integration assay, subjected the integration products to gap-filling and ligation, and used Golden Gate cloning to insert the chloramphenicol selection cassette. We then transformed the products into *E. coli*, plated the transformants on

agar containing chloramphenicol, and sequenced the clones using Sanger sequencing. The results show that Cas1—Cas2-DEDDh complex carries out both prespacer processing and integration into the CRISPR array (Fig. 4.4b-c). The complex is specific for the CRISPR array and all integration events occur at the repeat borders. However, integration occurs at all three repeats present in the array and there is no specificity for the leader-proximal repeat, as seen in many CRISPR systems *in vivo*[1,68]. The results are indicative of the potential involvement of a host factor like the *E. coli* integration host factor (IHF) to aid in integration target recognition *in vivo*[18,77]. The integrated spacer sequences confirm prespacer processing before integration: all integrated spacers are between 32-35 bp long. The excess 3' overhangs are trimmed to within 1-2 nt of the expected length and importantly, the PAM motif is absent in all integrated sequences (Fig. 4.4c). These findings indicate that while the Cas1—Cas2-DEDDh complex initially protects the PAM motif from trimming, it removes the PAM prior to full integration.

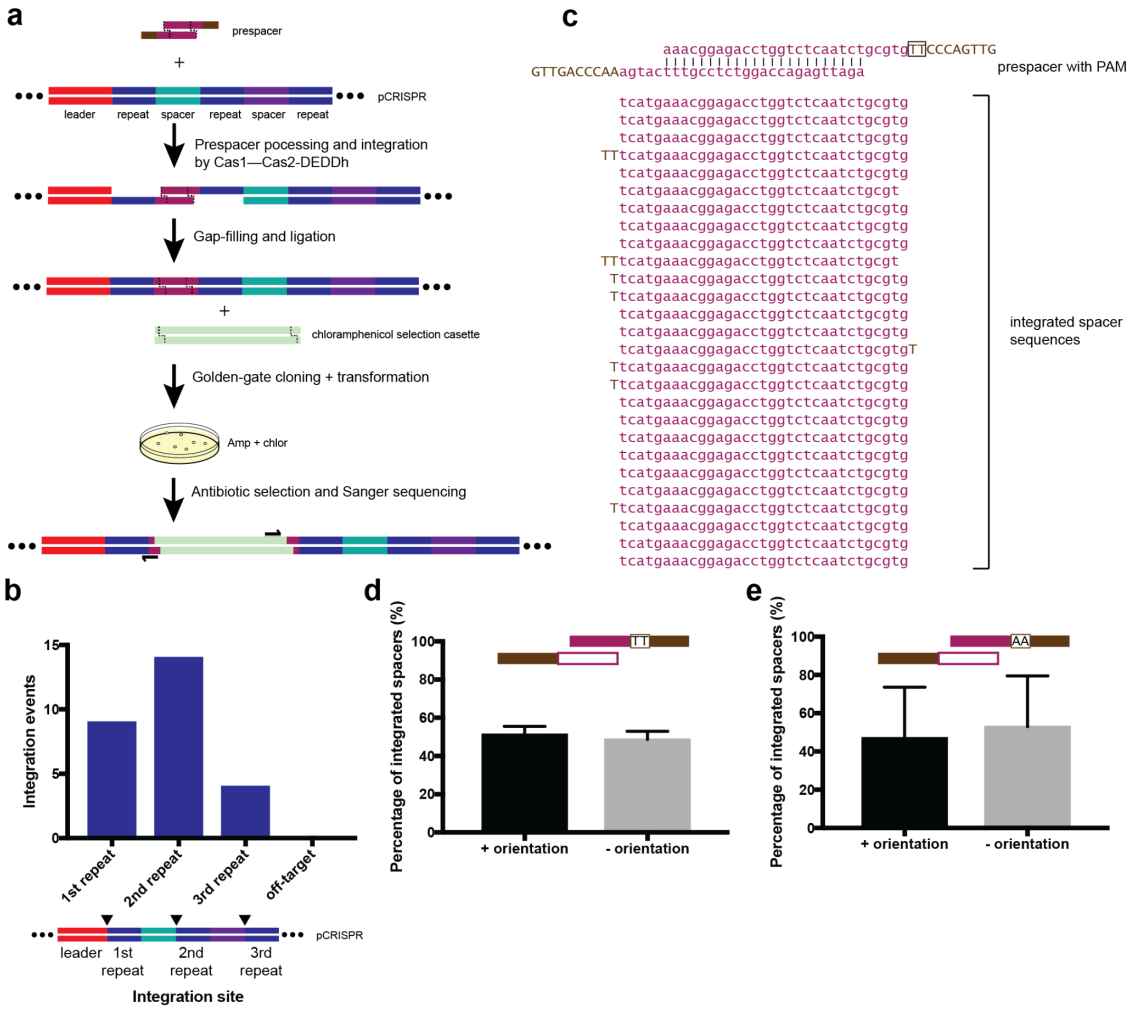


Figure 4.4 | PAM motif is processed for full-site integration. (a) Schematic of *in vitro* full-site integration assay with integration target pCRISPR plasmid. Integration events are selected using a

chloramphenicol selection screen. Prespacers are designed with BsaI cut sites in the protospacer, indicated by dashed lines. Protospacer region is shown in magenta, excess nucleotides in single-stranded 3' overhangs of prespacers are shown in brown. Following prespacer processing and integration, integration products are subjected to gap-filling and ligation. A chloramphenicol selection cassette is introduced by Golden Gate cloning. Transformants are plated on agar containing chloramphenicol and clones are sequenced using Sanger sequencing. **(b)** Number of integration events at each integration site from sequenced clones. Integration sites are depicted by arrows at repeat borders. **(c)** Fully integrated spacer sequences from a prespacer containing the PAM motif. Original unprocessed prespacer substrate is depicted above the integrated spacer sequences. Expected protospacer sequence is shown in magenta and excess nucleotides are shown in brown with the PAM motif boxed. **(d)** Orientation of spacer insertion from prespacer containing PAM motif, + orientation orients original TT PAM containing end toward the leader, - orientation orients original TT PAM containing end away from the leader. **(e)** Orientation of spacer insertion from control prespacer without PAM motif (the TT PAM in **d** is replaced with AA). The mean and standard deviation of three independent replicates are shown.

We next wondered whether the PAM motif in the prespacer would affect spacer orientation, which is critical to produce functional spacers for the downstream immune response. There is only one functional spacer orientation that would allow the resulting crRNA to target the complementary sequence in the foreign invader with the PAM on the correct side[7]. In order to direct the orientation of spacer insertion, the integration complex must both discern between the PAM and non-PAM sides of the prespacer and discern between the leader-side and spacer-side of the repeat. In previous systems, delayed PAM processing biases the orientation of integration by promoting the fully processed non-PAM-containing strand of the prespacer to attack the leader-side of the repeat for half-site integration[30,49]. To examine if this is the case for our system, we assessed the orientation of integration from a PAM-containing prespacer and separately from a control PAM-deficient prespacer substrate. We observed no orientation bias among the resulting integrated spacers from either substrate (Fig. 4.4d,e). We suspect that the complex requires the aid of a host factor similar to IHF to correctly orient the spacers *in vivo*. In *E. coli*, the addition of IHF serves to direct the first nucleophilic attack to the leader-side of the leader-proximal repeat through specific contacts with the leader sequence[18,77]. Based on the earlier processing results, it appears that the Cas1—Cas2-DEDDh alone is able to discern between the PAM and non-PAM sides of the prespacer (Fig.4.3b). Despite the results from our *in vitro* integration assays, we suspect that *in vivo*, where the system likely has higher specificity for the correct integration target site, the complex may utilize delayed PAM processing as the basis for determining the orientation of integration, as observed in other CRISPR systems[30,49,50].

4.3.3 Cas1—Cas2-DEDDh uses gate-keeping ruler mechanism for PAM processing

We next set out to investigate the mechanistic details of PAM removal and which active site was responsible for this activity. Based on earlier studies in other systems, we hypothesize that PAM processing is delayed until after half-site integration of the

fully processed PAM-deficient strand[30,49](Fig. 4.5a). To examine PAM removal and subsequent full-site integration, we made two substrates that mimic the predicted half-site intermediates during integration: the pre-PAM processing intermediate (half-site substrate 1) and the post-PAM processing intermediate (half-site substrate 2) (Fig. 4.5b). We tested these half-site substrates with both the WT and dead DEDDh complexes. In the time-course reactions with half-site substrate 1, which contains the unprocessed PAM, we observe a 100-nt band indicating the formation of the expected full-site integration product for the WT complex but not for the dead DEDDh complex (Fig. 4.5c). In the reactions with half-site substrate 2, which already has the PAM removed, we observe the 100-nt full-site integration product for both the WT and dead DEDDh complexes. The full-site integration product generated from the half-site substrate 1 appears to be the same size as the full-site integration products generated from half-site substrate 2 (100 nt). This suggests that in the WT reaction, the PAM in half-site substrate 1 is fully removed prior to full-site integration. The lower intensity of the full-site integration product generated from half-site substrate 1 compared to that generated from half-site substrate 2 may be a result of inefficient PAM removal. We conclude that the DEDDh active site carries out PAM processing and that PAM processing is necessary for full-site integration. Importantly, full-site integration generates a precisely defined product size. The time-course assay shows that the nonspecific DEDDh exonuclease immediately generates a ladder of fragments of variable size from the PAM-containing substrate strand. Yet, only one of these fragment sizes is selected to produce the full-site integration product. We propose that the single-nucleotide precision is a result of the Cas1—Cas2 ruler, which simultaneously defines the spacer size and acts as a gate-keeping mechanism preventing PAM insertion into the CRISPR array. This is supported by the results from the plasmid-based full-site integration assay, which shows that the PAM is absent in every sequence. This represents an alternative mechanism for PAM processing compared to Cas4, which uses a sequence-specific mechanism to endonucleolytically cleave the PAM[30,52].

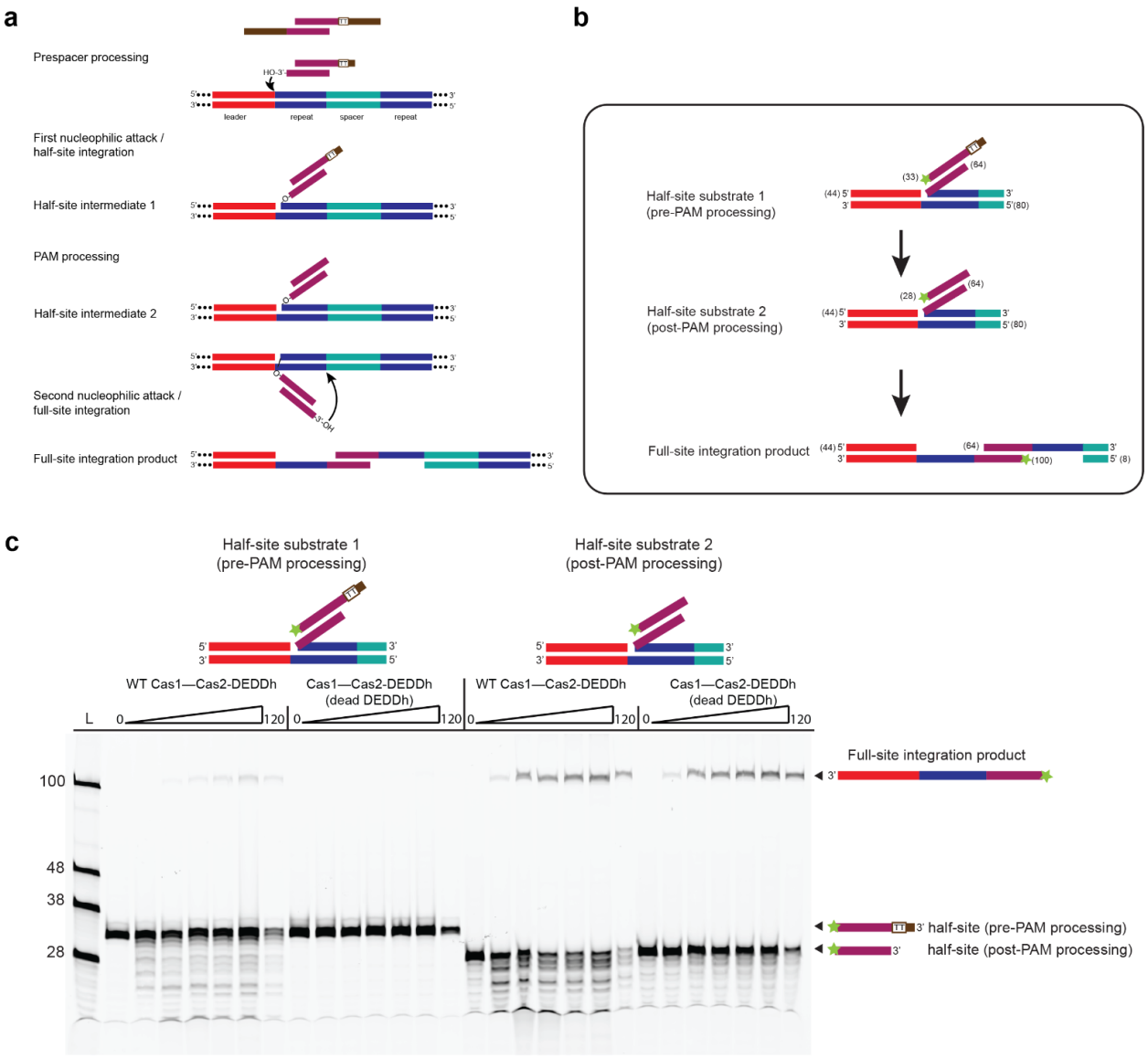


Figure 4.5 | DEDDh is responsible for PAM processing. (a) Schematic of stepwise prespacer processing and integration during spacer acquisition from prespacer. The PAM motif is depicted. **(b)** DNA substrates mimicking the half-site intermediates depicted in **a** and the expected full-site integration product. The lengths of the substrate strands are shown in parentheses by the 5' end of each strand. Star indicates 6-carboxyfluorescein label. **(c)** Integration reactions by WT and mutant Cas1—Cas2-DEDDh (DEDDh-D132A) with substrates depicted in **b**. Labeled substrate strands and full-site integration product strand are indicated.

4.3.4 PAM motif in prespacer results in lower integration efficiency *in vitro*

Despite the progress made in recent years[49], the mechanisms involved in prespacer biogenesis and PAM selection are poorly understood. Elucidating these processes are important to understand how PAM-flanking sequences are preferentially

selected for spacer acquisition to create functional spacers. With our natural exonuclease fusion to the CRISPR integrase, we are in an advantaged position to investigate prespacer selection and subsequent processing and integration. As mentioned earlier, previous studies suggest that the preferred binding substrate for the Cas1—Cas2 integrase is a 23-bp duplex DNA with long single-stranded 3' overhangs with the PAM in the designated position 5 nt away from the duplex[16,49]. A recent study hypothesized that Cas1—Cas2 preferentially binds these prespacer substrates containing the PAM motif over prespacers without the PAM motif and that this increased binding affinity facilitates integration of PAM-flanking sequences[49].

We set out to determine whether our Cas1—Cas2-DEDDh complex preferentially selects prespacers containing the PAM motif for integration. To do so, we generated a small prespacer pool consisting of equal concentrations of two prespacers: one PAM-containing prespacer and one PAM-deficient prespacer (Fig. 4.6a). The prespacers each contain a pair of identifying internal barcodes such that, after processing and integration, we can determine which prespacer substrate an integrated spacer came from. For the first prespacer pool, the 5'-tcg-3' / 5'-gat-3' barcode pair is associated with the PAM-containing prespacer and the 5'-cgt-3' / 5'-atg-3' barcode pair is associated with the PAM-deficient prespacer. We conducted a full-site integration assay with the prespacer pool and sequenced the resulting colonies. Surprisingly, out of 95 sequenced colonies, we observed a significantly higher number (~3-fold more) of integration events from the PAM-deficient prespacer than from the PAM-containing one (Fig. 4.6b). To account for any biases resulting from the internal barcode sequences, we generated a second prespacer pool, where we switched the barcode pairs. After performing the full-site integration assay with the second prespacer pool, we still observe that the PAM-deficient prespacer is significantly favored for integration (Fig. 4.6c). For this pool, there is an even greater difference: we observed ~9-fold more integration events from the PAM-deficient prespacer than from the PAM-containing prespacer. One explanation for the lower integration efficiency from the PAM-containing prespacer *in vitro* is inefficient PAM removal, which is necessary for full-site integration (Fig. 4.5c). The lack of selection for PAM-containing prespacers in this experiment could support a model in which PAM selection *in vivo* happens further upstream during the prespacer biogenesis pathway. We propose that at some early stage, only certain PAM-containing sequences are selected as viable prespacer precursors and channeled to create the canonically preferred 23-bp duplex prespacers.

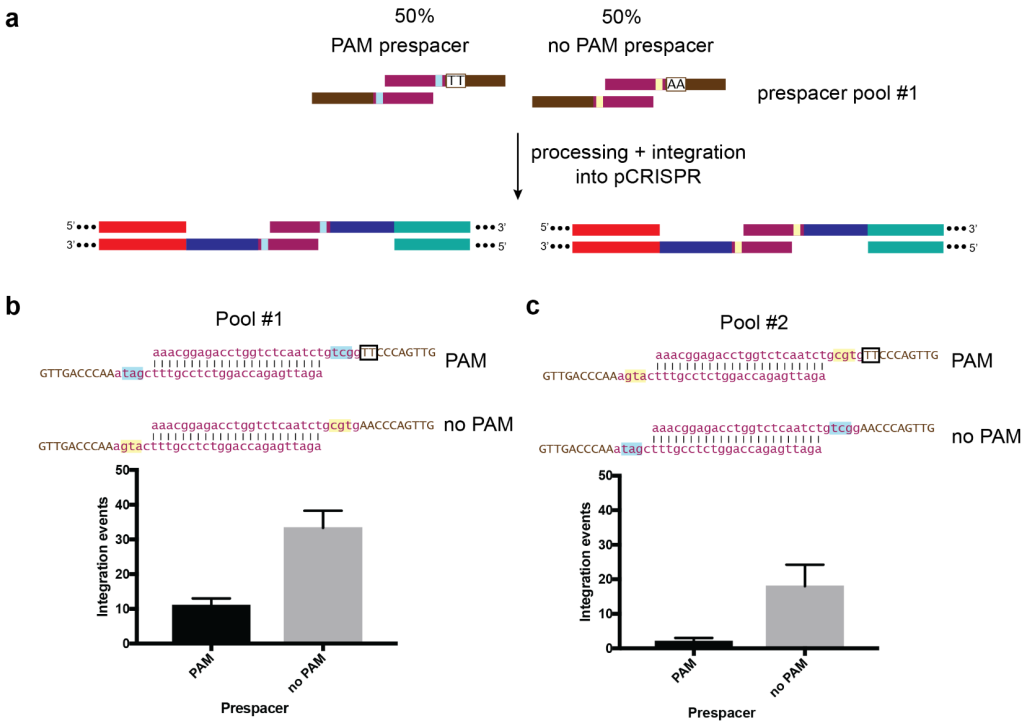


Figure 4.6 | Presence of PAM motif lowers integration efficiency from prespacer *in vitro*. (a) Schematic of prespacer pool #1 used for *in vitro* full-site integration assay. Prespacer pool consists of 50% of prespacers containing PAM and 50% of prespacers without PAM. The prespacer substrates are distinguished by internal 3-nt barcode pairs. The 5'-tcg-3' / 5'-gat-3' barcode pair is colored blue and associated with the PAM-containing prespacer; the 5'-cgt-3' / 5'-atg-3' barcode pair is colored yellow and associated with the PAM-deficient prespacer. **(b)** Number of integration events from each prespacer in pool #1 out of 95 sequenced clones. The mean and standard deviation of three independent replicate experiments are shown. The prespacer substrates are depicted with the barcodes pairs as colored in **a**. TT PAM motif is boxed. The protospacer sequence is shown in magenta and the excess nucleotides that are processed prior to integration are shown in brown. **(c)** Number of integration events from each prespacer in pool #2 out of 95 sequenced clones. The mean and standard deviation of three independent replicate experiments are shown. Prespacers in prespacer pool #2 are depicted: here, the 5'-tcg-3' / 5'-gat-3' barcode is associated with the PAM-deficient prespacer and the 5'-cgt-3' / 5'-atg-3' barcode is associated with the PAM-containing prespacer.

4.3.5 Substrate preferences for selective processing by Cas1—Cas2-DEDDh

To probe the question of when PAM selection occurs during prespacer biogenesis, we set out to investigate Cas1—Cas2-DEDDh's ability to recognize and selectively process potential prespacer precursors. We find that Cas1—Cas2-DEDDh cannot selectively process blunt-end double-stranded DNA (dsDNA) (using either ruler-guided trimming or PAM-dependent trimming) but can carry out PAM-dependent trimming with certain PAM-containing ssDNA substrates. In the first processing assay in this series, we test a 59-bp PAM-containing dsDNA and a 59-bp PAM-deficient dsDNA substrate and their respective constituent 59-nt ssDNA strands (Fig. 4.7a). For both

dsDNA substrates, we only observe nonspecific trimming and there is no evidence of PAM-dependent or ruler-guided trimming. However, with their constituent ssDNA strands, we do observe PAM-dependent trimming in the PAM-containing ssDNA, similarly to what we observed earlier with the 38-nt ssDNA substrates (Fig. 4.3c). All the PAM-deficient ssDNA substrates were nonspecifically trimmed. To investigate dependence on the position of the PAM in ssDNA, we varied the position of the TT motif in the 59-nt ssDNA substrates (Fig. 4.7b). Curiously, for these substrates, we only observe PAM-dependent trimming for two of the three PAM positions tested. We further explored PAM recognition in ssDNA by testing another set of 56-nt ssDNA substrates that have varied PAM positions and only consist of As, Ts, and Cs (Fig. 4.7c). We observed that all of these new substrates were nonspecifically trimmed regardless of whether they contained the PAM motif. Based on these results, we suspect that additional structural features or sequence motifs in the DNA might play a role in facilitating or inhibiting PAM-dependent trimming in ssDNA. Our results indicate that Cas1—Cas2-DEDDh can selectively process certain PAM-containing ssDNA sequences but not others.

We also explored the necessary substrate features for Cas1—Cas2-DEDDh's ruler-guided trimming to the functionally relevant size for integration. Although the 23-bp duplex prespacer appears to be the canonically preferred substrate, it is unclear when the 23-bp duplex is formed during prespacer biogenesis. We investigated Cas1—Cas2-DEDDh's dependence on the 23-bp duplex region of the prespacer for ruler-guided processing by varying the duplex lengths of PAM-containing prespacers from 15-31 bp (Fig. 4.8a). In the processing reactions with the pre-duplexed prespacers, we only observe functional processing with the 23-bp duplex prespacer (Fig. 4.8b). There appears to be varying degrees of specific trimming in some of the other prespacers as well, although the products all seem to be non-functional for integration. The results indicate that there is a strict requirement for the 23-bp duplex region of the prespacer for ruler-guided processing. As with other CRISPR integrases, there is likely a structural basis for this internal ruler. In the *E. coli* Cas1—Cas2 structure, a tyrosine residue acts as a wedge to cap the ends of the 23-bp DNA duplex through a stacking interaction[10,16]. A previous model suggests that this wedge can function to splay longer DNA duplex regions by displacing the 5' end and defining the 23-bp duplex region. If this is true and the Cas1—Cas2-DEDDh is able to measure out 23 bp and splay the ends of longer duplexes, then we would expect it to perform functional ruler-guided trimming on longer DNA duplexes. However, our results suggest there is no tolerance for any variation in duplex size for these prespacers.

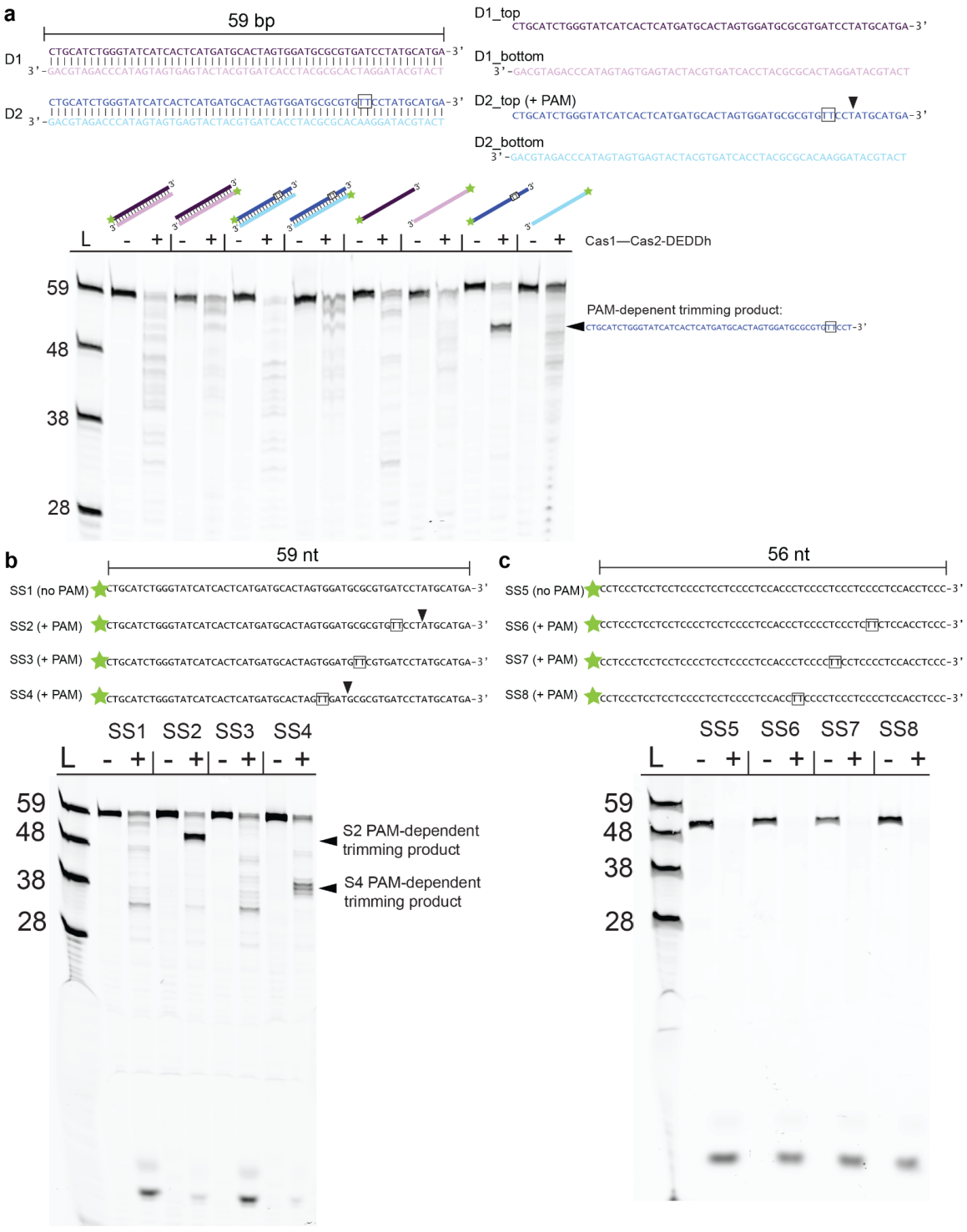


Figure 4.7 | Cas1—Cas2-DEDDh selectively performs PAM-dependent trimming on certain ssDNA substrates but not dsDNA. (a) Processing reactions by Cas1—Cas2-DEDDh with PAM-deficient and PAM-containing dsDNA and ssDNA substrates, shown on top and schematized above the gel. The TT PAM motif is boxed and shown in the relevant substrates. Arrowhead indicates PAM-dependent trimming. **(b)** Processing reactions by Cas1—Cas2-DEDDh with 59-nt ssDNA substrates. The TT PAM motif is boxed and its position is shown in the relevant substrates. Arrowheads indicate PAM-dependent trimming.

(c) Processing reactions by Cas1—Cas2-DEDDh with 56-nt ssDNA substrates with only As, Cs, and Ts. The TT PAM motif is boxed and its position is shown in the relevant substrates.

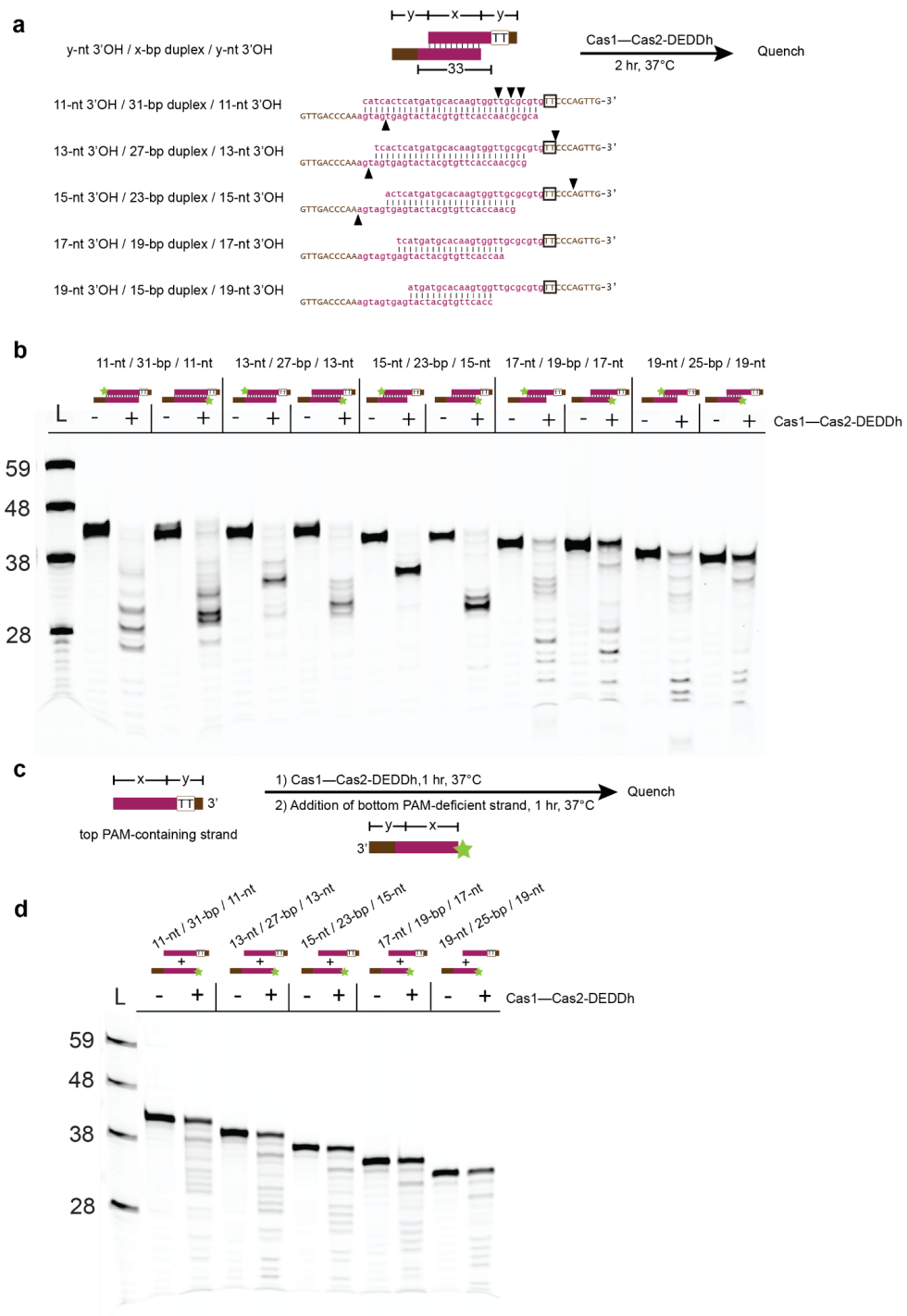


Figure 4.8 | Cas1—Cas2-DEDDh exhibits strict requirement for pre-duplexed 23-bp DNA duplex prespacer for functional ruler-guided trimming. (a) Prespacer substrates with varying duplex lengths used in processing assay. The duplex and overhang lengths are indicated, and the TT PAM motif is boxed. The arrowheads indicate specific processing positions as observed from the processing products

visualized on the denaturing gel in **b**. **(b)** Processing assay with pre-duplexed prespacer substrates shown in **a**. Substrates are schematized. **(c)** Schematic of processing assay with stepwise addition of the top and bottom prespacer strands of the substrates shown in **a**. The unlabeled top PAM-containing strand is incubated first with Cas1—Cas2-DEDDh followed by delayed addition of the labeled bottom PAM-deficient strand. **(d)** Processing assay with stepwise addition of the top and bottom prespacer strands of the substrates shown in **a.** Substrates are schematized.

Some have hypothesized that Cas1—Cas2 first binds to ssDNA prespacer precursors and then finds and facilitates pairing with the complementary partner[49]. We tested this hypothesis by carrying out stepwise addition of the complementary strands of the prespacer substrates from the previous experiment for processing by Cas1—Cas2-DEDDh. We first incubated the Cas1—Cas2-DEDDh complex with the single-stranded PAM-containing strand for one hour and then added the labeled complementary PAM-deficient strand for another hour reaction (Fig. 4.8c). In all of the reactions, even for the prespacer strands with the 23-nt complementary region, the products show only nonspecific processing of the labeled strand (Fig. 4.8d). The results suggest that the Cas1—Cas2-DEDDh complex can only perform ruler-guided trimming when the 23-bp prespacer is pre-duplexed and not after delayed addition of the complementary strand. Together, these findings point to a strict requirement for a pre-duplexed 23-bp duplex prespacer for functional processing by Cas1—Cas2-DEDDh into an integration-viable product.

4.4 Discussion

In CRISPR adaptive immunity, the coordination between the CRISPR integrase and host exonucleases is critical in the generation of viable integration substrates[49,50]. In this study, we elucidate the mechanisms of prespacer processing and integration by a naturally occurring Cas1—Cas2-DEDDh complex, which utilizes an alternative PAM processing mechanism compared to the well-studied Cas4 endonuclease. We further investigate Cas1—Cas2-DEDDh's substrate preferences for selective processing and integration and test previous models for prespacer selection. Through these experiments, we uncover unexpected nuances of the Cas1—Cas2-DEDDh complex's substrate selectivity. Our findings help advance our understanding of how prespacers are processed and selected for spacer acquisition.

In our experiments, we find that Cas1—Cas2-DEDDh processes prespacers (23-bp DNA duplexes with extended single-stranded 3' overhangs) into viable integration substrates in both a ruler-guided and PAM-dependent manner (Fig. 4.9). We identify that the DEDDh active site is responsible for full processing of prespacer substrate, including PAM processing. The Cas1—Cas2 supplies the internal ruler to measure out and protect the protospacer region from trimming. Cas1 also likely contains the binding pocket for PAM identification and initial protection from trimming. The DEDDh exonuclease coordinates closely with Cas1—Cas2 to process prespacers

in a stepwise manner. While the Cas1—Cas2-DEDDh complex immediately processes excess unprotected nucleotides in the 3' overhangs, it delays PAM processing. In other CRISPR systems, delayed PAM processing serves as the basis for determining the orientation of integration by promoting the non-PAM side to integrate first into the integration target site[30,49]. Though the integrated spacers in our *in vitro* integration assay do not exhibit an orientation bias, this can be accounted for by the observed lack of specificity for the leader-repeat border. We hypothesize that the delayed PAM processing by the Cas1—Cas2-DEDDh complex may influence spacer orientation *in vivo*, where host factors likely help provide greater specificity for the integration target site. Our experiments show the Cas1—Cas2-DEDDh complex processes the PAM in the half-site integration intermediate and subsequently carries out full-site integration to produce an exquisitely precise product size, especially in comparison to the large variance in size of the trimming intermediates from the exonuclease. Our experiments suggest that prespacer processing relies on the precise structural coordination between the integrase and exonuclease domains and the kinetic coordination of their enzymatic activities.

The process of PAM selection and subsequent removal is an elegant means by which CRISPR systems achieve robust immunity against foreign invaders while avoiding auto-immunity. Delayed PAM processing appears to be a common theme across different systems[30,49,50]. However, the means by which CRISPR systems identify and process the PAM during adaptation are remarkably diverse. Many CRISPR systems, including several type I, II, and V systems, use the Cas4 endonuclease for PAM identification and processing, which has been characterized in recent studies[30,51,52,55]. Our system represents a group of CRISPR systems, including the *E. coli* type I-E system, that utilize Cas1 for PAM identification and host exonucleases for PAM processing[49,50]. A third mode, and the least understood, exists in type II-A systems, which use Cas9 for PAM identification during adaptation[56,57]. There are key similarities and differences between the mechanisms employed by Cas4-containing systems and those used by the type I-E systems. Both systems initially protect the PAM from trimming by host exonucleases through PAM sequestration. In both systems, PAM processing occurs in the half-site intermediate, likely triggered by conformational changes[30]. The two systems use different mechanisms for PAM processing. Cas4 utilizes a sequence-specific mechanism to process the PAM via endonucleolytic cleavage of the labile bond between the PAM and protospacer. For the type I-E systems, however, we propose that the exonuclease and integrase use an alternative ruler-based gate-keeping mechanism to coordinate full PAM processing. Our results indicate that the non-sequence-specific DEDDh exonuclease is responsible for trimming the PAM in the half-site intermediate after PAM release from Cas1 protection. Despite not having sequence-specificity for PAM processing, the type I-E system still achieves exquisite control in fully processing the PAM for full-site integration. To explain this

observation, we propose that the single-nucleotide precision of the Cas1—Cas2 ruler imposes a gate-keeping mechanism to prevent PAM insertion into the CRISPR array. These different mechanisms may provide an evolutionary advantage in different contexts. In some scenarios, using a distinct sequence-specific enzyme for PAM processing may allow for better precision and kinetic control. However, there may be situations where it is more efficient to package the PAM-binding pocket within the integrase and reuse the same host exonucleases for prespacer trimming and PAM processing.

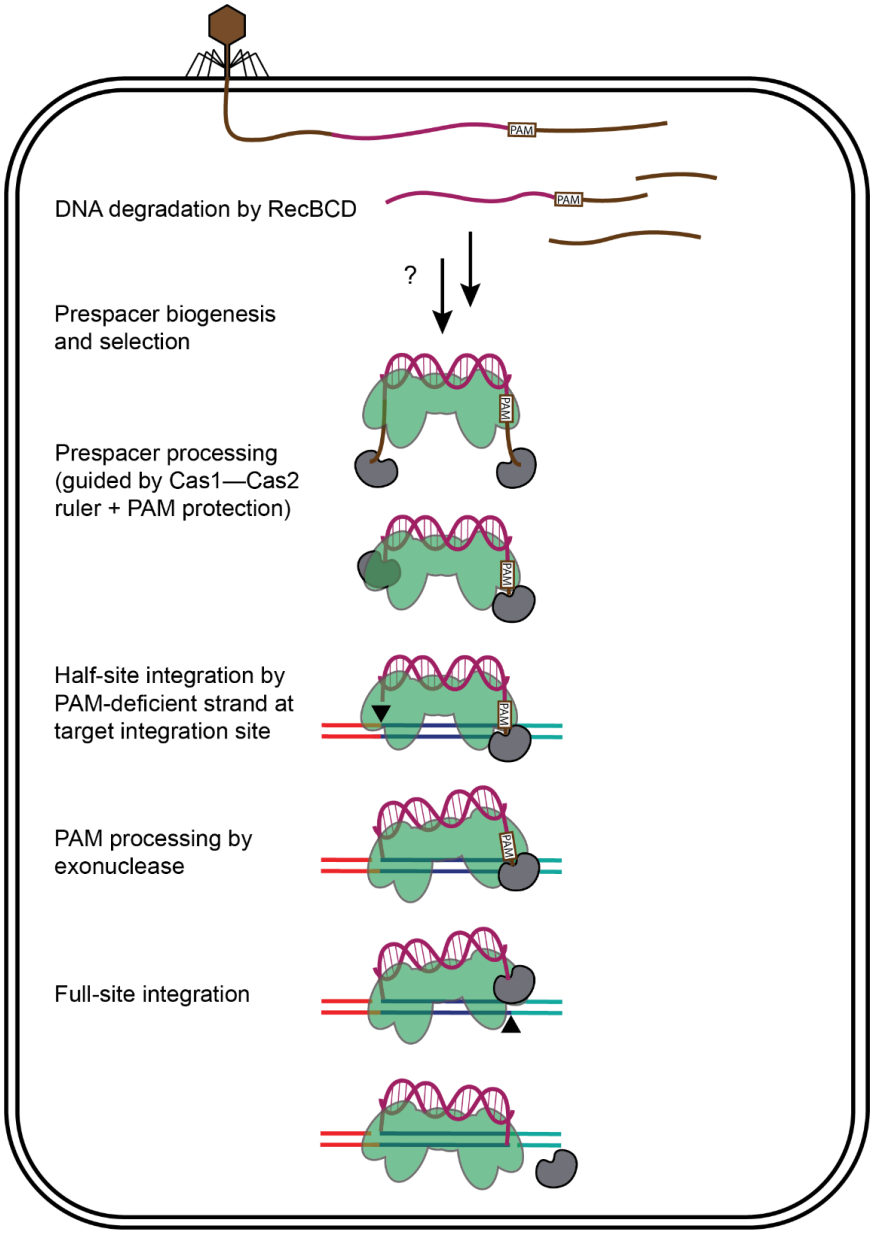


Figure 4.9 | Model of prespacer processing and integration. Prespacers derive from DNA fragments. The details of prespacer biogenesis and selection are still unknown, indicated by the interrogation mark.

Cas1—Cas2 and exonucleases coordinate prespacer processing, using Cas1—Cas2's internal ruler and PAM protection. Full processing of the PAM-deficient prespacer strand allows for half-site integration at the target integration site at the leader-side of the repeat. This is followed by PAM processing by the exonuclease, which is necessary to facilitate full-site integration at the spacer-side of the repeat. DNA is colored: protospacer, magenta; all other foreign DNA, brown; leader, red; repeat, blue; spacer, turquoise. Arrowheads indicate integration.

While we have made progress understanding PAM processing immediately before integration, the upstream steps involving PAM selection and prespacer biogenesis remain a mystery. Unexpectedly, our experimental data suggest lower integration efficiency from PAM-containing prespacers compared to PAM-deficient ones. We hypothesize that PAM selection may occur early during prespacer biogenesis, prior to the formation of the canonical prespacer used in our assay. Given these results, we explored the Cas1—Cas2-DEDDh complex's ability to capture and selectively process potential prespacer precursors. Interestingly, we observe PAM-dependent processing of certain PAM-containing ssDNAs but not others. Our results are suggestive of additional sequence motifs or structural features that are necessary for selection and processing of ssDNA prespacer precursors. We further explored the Cas1—Cas2-DEDDh complex's duplex requirement for ruler-guided processing. Our data reveal a strict requirement for a pre-duplexed prespacer with a 23-bp duplex region and that Cas1—Cas2-DEDDh alone cannot splay longer duplex regions for functional ruler-guided trimming[10]. Stepwise addition of the complementary ssDNA strands of the prespacer also suppressed ruler-guided trimming. While our results suggest Cas1—Cas2-DEDDh may in fact first bind and process PAM-containing ssDNA prespacer precursors, it alone cannot facilitate subsequent capture and binding of the complementary strand to form the duplex prespacer, as a previous study proposed[49]. Our experiments indicate that prespacer biogenesis and selection is much more nuanced than the current models suggest and may involve more unidentified DNA motifs and host factors to aid in the capture of prespacer precursors and duplex formation.

Our work elucidates the mechanisms of a natural CRISPR integrase-exonuclease fusion in prespacer/PAM processing and provides some insight into the enigma of prespacer selection and biogenesis. We find common themes across various CRISPR systems as well as notable differences that play a role in the evolution of CRISPR adaptation. Our results can be further applicable to inform the engineering of new technologies, especially in the realm of molecular recording in DNA, which is challenged by the current lack of understanding of the components necessary for prespacer selection and processing[64,138].

4.5 Materials and methods

4.5.1 Protein purification

The *Megasphaera NM10_related* Cas1 and Cas2-DEDDh genes were codon optimized for *E. coli* expression, ordered as G-blocks, PCR amplified, and cloned separately into a pET-based expression vector with an N-terminal 10xHis-MBP-TEV tag. After transformation into chemically competent Rosetta cells, cells were grown to an OD600 of ~0.6 and induced overnight at 16 °C with 0.5 mM isopropyl-β-D-thiogalactopyranoside (IPTG). Cells were harvested and resuspended in lysis buffer (20 mM HEPES, pH 7.5, 500 mM NaCl, 10 mM imidazole, 0.1% Triton X-100, 1 mM Tris (2-carboxyethyl)phosphine (TCEP), Complete EDTA-free protease inhibitor (Roche), 0.5 mM phenylmethylsulfonyl fluoride (PMSF), and 10% glycerol). After lysis by sonication and clarification of lysate by centrifugation, the supernatant was incubated with Ni-NTA resin (Qiagen). The resin was washed with wash buffer (20 mM HEPES, pH 7.5, 500 mM NaCl, 10 mM imidazole, 1 mM TCEP, and 5% glycerol) and protein was eluted with wash buffer supplemented with 300 mM imidazole. After overnight digestion with TEV protease, the salt concentration was diluted to 300 mM NaCl using ion-exchange buffer A (20 mM HEPES, pH 7.5, 1 mM TCEP, and 5% glycerol) and run through a tandem MBPTrap column (GE Healthcare) and HiTrap heparin HP column (GE Healthcare) to remove the MBP and bind the protein onto the heparin column. The protein was eluted with a gradient from 300 mM to 1 M KCl, concentrated, and purified on a Superdex 200 (16/60) column with storage buffer (20 mM HEPES, pH 7.5, 500 mM KCl, 1 mM TCEP, and 5% glycerol). The same purification protocol was used for Cas1 and Cas2-DEDDh (WT and D132A mutant). Sequences of proteins are shown in Table 4.1.

Protein	Sequence
Cas1	MAGPIIAGKSESSELPRVEDRATFIYIEHAKINRVDSAVTVAEAKGVVRIPAAMIG VLLLGPGTDISHRAVELLGDTGTALVWVGEQGVRYYASGRALARSTRFLVKQA ELVTNERSRLRVARRMYQMRFPTEDVSKLTMQQLRSHEGARVRRKYRELSK KYNVPWKKRVYNPDDFAGGDPINQALSAAHVALYGLVHSVVAALGLSPGLGF VHTGHDRSFIYDVADLYKAEITVPIAFAVAAEAEEGQDIGQLARLRTRDAFVDG KILKRMVKDLQTLLEIPEEGQIEAEPLSLWDDKEKLVPYGVNYSEVTSCP*
Cas2-DEDDh	MPMTVITLKNVPQSLRGDLTRWMQEIATGVYVGNFNSRIREYLWRRVQETMG AGEASMCFAARNELGYDFLTENASRSVIDYDGLPLIFIPKEQSAVSDLPKGFST AAKLHRAHIAGSGKKKEKPIRYVVIDIETDGKDAKRNHILEIGAIRCEDGKETHF TALISGDAVPPSITKLTGITATLLQKEGQEEKKVLTAFREFIGDDDLVGYHVSFDI EFLRQAFKKYGLGYLKNKTHDLLRIVKKEQLFQADYKLETSLQSYGIHKKVPH RALGDAELVKCLAKKLNKF*

Table 4.1 | Sequences of the type I-E *Megasphaera NM10_related* Cas1 and Cas2-DEDDh proteins used in this study.

4.5.2 DNA substrate preparation

To make the target integration plasmid pCRISPR, the leader and the first 3 repeats and spacers of the CRISPR array were ordered as two DNA fragments, which were amplified by PCR and inserted into a pUC19 backbone by Gibson Assembly. DNA oligos used in this study were ordered from Integrated DNA Technologies (IDT). Prespacers and the half-site substrates were formed by heating at 95 °C for 5 min and slow cooling to room temperature in HEPES hybridization buffer (20 mM HEPES, pH 7.5, 25 mM KCl, and 10 mM MgCl2). For the half-site substrate, hybridization was carried out with a 1.5-fold excess of the two shortest strands and 1.25-fold excess of the second-largest strand and purified on a 8% native PAGE gel. Sequences of cloning primers and DNA substrates are shown in Table 4.2.

Name/Description	Sequence	Figure
S1_top	actcatgatgcacaagtggttgcgcgtg	4.1
S1_bottom	gcaaccacttgtgcatcatgagtgatga	4.1
S2_top (Fig. 4.1, 4.2) mixed OH 1 (AA PAM)_top (Fig. 4.3)	actcatgatgcacaagtggttgcgcgtgAACCCAGTTG	4.1, 4.2, 4.3
S2_bottom (Fig. 4.1, 4.2) Bottom strand of prespacers (Fig. 4.3) 15-nt 3'OH / 23-bp duplex / 15-nt 3'OH (bottom) (Fig. 4.8)	gcaaccacttgtgcatcatgagtgatgaAACCCAGTTG	4.1, 4.2, 4.3, 4.8
S3_top	actcatgatgcacaagtggttgcgcgtgACCTAGAAGTAACCCAGTTG	4.1, 4.2
S3_bottom	gcaaccacttgtgcatcatgagtgatgaACCTAGAAGTAACCCAGTTG	4.1, 4.2
polyA OH (AA PAM)_top	actcatgatgcacaagtggttgcgcgtgAAAAAAAAAA	4.3
poly T OH (TT PAM)_top	actcatgatgcacaagtggttgcgcgtgTTTTTTTTTT	4.3
poly C OH (CC PAM)_top	actcatgatgcacaagtggttgcgcgtgCCCCCCCCCC	4.3
poly G OH (GG PAM)_top	actcatgatgcacaagtggttgcgcgtgGGGGGGGGGG	4.3
mixed OH 2 (TT PAM)_top	actcatgatgcacaagtggttgcgcgtgTTTTCAGTTG	4.3
mixed OH 3 (TT PAM)_top (Fig. 4.3) 15-nt 3'OH / 23-bp duplex / 15-nt 3'OH (top) (Fig. 4.8)	actcatgatgcacaagtggttgcgcgtgTTCCCAGTTG	4.3. 4.8
mixed OH 4 (TA PAM)_top	actcatgatgcacaagtggttgcgcgtgTACCCAGTTG	4.3
Full-site prespacer_top (Fig. 4.4b,c) Full-site prespacer (cgt barcode, PAM)_top (Fig. 4.6)	aaacggagacctggtctcaatctgcgtgTTCCCAGTTG	4.4, 4.6
Full-site prespacer_bottom Full-site prespacer (atg barcode)_bottom (Fig. 4.6)	agattgagaccaggtctccgtttcatgaAACCCAGTTG	4.4, 4.6
Half-site substrate bottom strand (spacer-repeat-leader)	tgcgcgtgggatcacccccgctcgtgcgggaaagacagtaatggattcctttatttttcgccctttttacgcttactgac*g*t	4.5
Half-site substrate leader strand	acgtcagtaagcgtaaaagggcgaaaataaaggaatccatta*c*t	4.5
Half-site substrate protospacer-repeat-spacer strand	agattgagaccaggtctccgtttcatgagtctttcccgcacgagcgggggtgatcccacgcg*c*a	4.5
Half-site substrate 1 protospacer + unprocessed PAM strand	aaacggagacctggtctcaatctgcgtgTTCCC	4.5
Half-site substrate 2 protospacer strand	aaacggagacctggtctcaatctgcgtg	4.5
Full-site prespacer_top (with PAM) (Fig. 4.4d) Full-site prespacer (tcg barcode, PAM)_top (Fig. 4.6)	aaacggagacctggtctcaatctgtcggTTCCCAGTTG	4.4, 4.6
Full-site prespacer_top (no PAM) (Fig. 4.4e) Full-site prespacer (tcg barcode, no PAM)_top (Fig. 4.6)	aaacggagacctggtctcaatctgtcggAACCCAGTTG	4.4, 4.6

Full-site prespacer_bottom (Fig. 4.4d,e) Full-site prespacer (gat barcode)_bottom (Fig. 4.6)	agattgagaccaggtctccgtttcgataAACCCAGTTG	4.4, 4.6
Full-site prespacer (cgt barcode, no PAM)_top (Fig. 4.6)	aaacggagacctggtctcaatctgcgtgAACCCAGTTG	4.6
D1_top (Fig. 4.7a) SS1 (Fig. 4.7b)	CTGCATCTGGGTATCATCACTCATGATGCACTAGT GGATGCGCGTGATCCTATGCATGA	4.7
D1_bottom	TCATGCATAGGATCACGCGCATCCACTAGTGCATC ATGAGTGATGATACCCAGATGCAG	4.7
D2_top (Fig. 4.7b) SS2 (Fig. 4.7b)	CTGCATCTGGGTATCATCACTCATGATGCACTAGT GGATGCGCGTGTTCCTATGCATGA	4.7
D2_bottom	TCATGCATAGGAACACGCGCATCCACTAGTGCATC ATGAGTGATGATACCCAGATGCAG	4.7
SS3	CTGCATCTGGGTATCATCACTCATGATGCACTAGT GGATGTTCGTGATCCTATGCATGA	4.7
SS4	CTGCATCTGGGTATCATCACTCATGATGCACTAGTT GATGCGCGTGATCCTATGCATGA	4.7
11-nt 3'OH / 31-bp duplex / 11-nt 3'OH (top)	catcactcatgatgcacaagtggttgcgcgtgTTCCCAGTTG	4.8
11-nt 3'OH / 31-bp duplex / 11-nt 3'OH (bottom)	acgcgcaaccacttgtgcatcatgagtgatgaAACCCAGTTG	4.8
13-nt 3'OH / 27-bp duplex / 13-nt 3'OH (top)	tcactcatgatgcacaagtggttgcgcgtgTTCCCAGTTG	4.8
13-nt 3'OH / 27-bp duplex / 13-nt 3'OH (bottom)	gcgcaaccacttgtgcatcatgagtgatgaAACCCAGTTG	4.8
17-nt 3'OH / 19-bp duplex / 17-nt 3'OH (top)	tcatgatgcacaagtggttgcgcgtgTTCCCAGTTG	4.8
17-nt 3'OH / 19-bp duplex / 17-nt 3'OH (bottom)	aaccacttgtgcatcatgagtgatgaAACCCAGTTG	4.8
19-nt 3'OH / 15-bp duplex / 19-nt 3'OH (top)	atgatgcacaagtggttgcgcgtgTTCCCAGTTG	4.8
19-nt 3'OH / 15-bp duplex / 19-nt 3'OH (bottom)	ccacttgtgcatcatgagtgatgaAACCCAGTTG	4.8
Chloramphenicol selection cassette PCR forward primer	GGCCGGTCTCCAGATtgatcggcacgtaagaggttc	4.4
Chloramphenicol selection cassette PCR reverse primer	GGCCTGGTCTCAAAACattctcaccaataaaaaacgcccg	4.4

Table 4.2 | DNA oligonucleotides used in this study. Sequences are written 5' to 3'. Asterisk indicates phosphorothioate bond.

4.5.3 Processing assays

Processing assays were conducted in integration buffer (20 mM HEPES, pH 7.5, 125 mM KCl, 10 mM $MgCl_2$, 1 mM DTT, 0.01% Nonidet P-40, and 10% DMSO). Cas1 (4 μM) and Cas2-DEDDh (2 μM) were pre-complexed for 30 min. at 4 °C before addition of fluorescent DNA substrate (312.5 nM) and reacting for 2 hours at 37 °C. Reaction was

quenched by addition of 2 vol. quench buffer (95% formamide, 30 mM EDTA, 0.2% SDS, and 400 µg/mL heparin) and heating at 95 °C for 4 min., before analysis on a 14% urea-PAGE gel. Reactions were visualized by Typhoon FLA gel imaging scanner and quantification of intensities was performed with ImageQuantTL 8.2. The percent processing activity is quantified as the ratio of the final product bands' intensity to the total intensity of all bands in the lane.

4.5.4 Ligation assays with pCRISPR integration target plasmid

Ligation assays were conducted in integration buffer (20 mM HEPES, pH 7.5, 125 mM KCl, 10 mM $MgCl_2$, 1 mM DTT, 0.01% Nonidet P-40, and 10% DMSO). Cas1 (4 µM) and Cas2-DEDDh (2 µM) were pre-complexed for 30 min. at 4 °C before addition of DNA substrate (312.5 nM) and integration target pCRISPR (20 ng/mL, ~10 nM) and reacting for 2 hours at 37 °C. Reaction was quenched with 0.4% SDS and 25 mM EDTA, treated with proteinase K for 15 min at room temperature, and then treated with 3.4% SDS. Reactions were analyzed on a 1.5% agarose gel and visualized by Typhoon FLA gel imaging scanner.

4.5.5 Full-site integration assays

Integration assays (50 uL reactions) were conducted in integration buffer (20 mM HEPES, pH 7.5, 125 mM KCl, 10 mM $MgCl_2$, 1 mM DTT, 0.01% Nonidet P-40, and 10% DMSO). Cas1 (4 µM) and Cas2-DEDDh (2 µM) were pre-complexed for 30 min. at 4 °C before addition of DNA substrate containing BsaI cut sites (312.5 nM) and reacting for 15 minutes, followed by the addition of the integration target pCRISPR (20 ng/mL, ~10 nM) and incubating for 2 hours at 37 °C. Products were purified with a DNA Clean and Concentrator 5 kit (Zymo Research) and eluted with 6 µl water. A gap-filling reaction (20 µl total, 37 C for 30 min) was conducted with the purified integration products as described by a previous study[58]: 6 µl purified acquisition reaction, 6.5 µl water, 2 µl 10x Taq DNA Ligase buffer (NEB), 2 uL dNTP Solution Mix (10 mM stock, NEB), 2 µl Taq DNA ligase (80 U, NEB), and 1 µl T4 DNA Polymerase (1 U, NEB). Gap-filling reactions purified using the Zymo Research kit and eluted with 6 uL water. A Golden Gate-compatible chloramphenicol selection cassette was generated by PCR with primers encoding BasI cut sites and purified using the Qiagen MinElute PCR Purification Kit. The sequences of primers used are shown in Table 4.2. A Golden Gate cloning reaction was performed with the purified, gap-filled integration products and chloramphenicol selection cassette following a standard BsaI assembly protocol. The products were purified using the Zymo Research kit and eluted with 6 uL water and 1 uL was electroporated into DH10B cells (NEB). Electroporated cells were recovered in 975 uL of LB and plated on LB agar containing carbenicillin (100 µg/ml) and

chloramphenicol (25 µg/ml). Of the surviving colonies, 95 were sequenced using Sanger Sequencing and sequences were analyzed using SnapGene Version 5.0.8.

4.5.6 CRISPR locus bioinformatic analysis

Cas1-DEDDh loci from metagenomic data were identified by determining genomes that contained a CRISPR locus via CRISPRDetect, and coding sequences within 5kb of the array were extracted[139]. A DEDDh HMM model was built from BLAST searches against the NCBI nr database that were manually verified[140]. The coding sequences were searched against the DEDDh model using hmmsearch with E < 1 × 10−5[140]. Matches that also contained credible hits to Cas1 and neighboring other Cas proteins were shortlisted for this work. A preliminary Cas1-DEDDh model was computed using AlphaFold2 to aid in structure-building[141].

4.6 Acknowledgements

We thank J. Zhou for her technical help in preparing materials. J.Y.W. is supported by the US National Science Foundation Graduate Fellowship and previously by the Berkeley Graduate Fellowship. B.A.-S. is supported by the US National Science Foundation Graduate Fellowship. Correspondence and requests for materials should be addressed to J.A.D. (doudna@berkeley.edu).

4.7 Author contributions

J.Y.W. conceived of and designed experiments with input from J.A.D. J.Y.W., O.T., and G.L. performed biochemistry experiments. P.S. conducted cryo-EM data collection and processed cryo-EM data. B.A.-S. identified the CRISPR-Cas locus of interest and recommended it for this study, and computed preliminary structural models.

Bibliography

1. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).

2. Koonin, E. V. & Makarova, K. S. Origins and evolution of CRISPR-Cas systems. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **374**, 20180087 (2019).

3. Hille, F. *et al.* The biology of CRISPR-Cas: backward and forward. *Cell* **172**, 1239–1259 (2018).

4. Shmakov, S. *et al.* Discovery and functional characterization of diverse class 2 CRISPR-Cas systems. *Mol. Cell* **60**, 385–397 (2015).

5. Grissa, I., Vergnaud, G. & Pourcel, C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8**, 172 (2007).

6. McGinn, J. & Marraffini, L. A. Molecular mechanisms of CRISPR–Cas spacer acquisition. *Nat. Rev. Microbiol.* **17**, 7–12 (2019).

7. Gleditzsch, D. *et al.* PAM identification by CRISPR-Cas effector complexes: diversified mechanisms and structures. *RNA Biol.* **16**, 504–517 (2019).

8. Behler, J. & Hess, W. R. Approaches to study CRISPR RNA biogenesis and the key players involved. *Methods* **172**, 12–26 (2020).

9. Makarova, K. S. *et al.* Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.* **18**, 67–83 (2019).

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

11. Xiao, Y., Luo, M., Dolan, A. E., Liao, M. & Ke, A. Structure basis for RNA-guided DNA degradation by Cascade and Cas3. *Science* **361**, (2018).

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

13. Swarts, D. C., van der Oost, J. & Jinek, M. Structural basis for guide RNA processing and seed-dependent DNA targeting by CRISPR-Cas12a. *Mol. Cell* **66**, 221–233.e4 (2017).

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

15. Hickman, A. B., Kailasan, S., Genzor, P., Haase, A. D. & Dyda, F. Casposase structure and the mechanistic link between DNA transposition and spacer acquisition by CRISPR-Cas. *eLife* **9**, (2020).

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

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

18. Wright, A. V. *et al.* Structures of the CRISPR genome integration complex. *Science* **357**, 1113–1118 (2017).

19. Hickman, A. B. & Dyda, F. DNA transposition at work. *Chem. Rev.* **116**, 12758–12784 (2016).

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

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

22. Datsenko, K. A. *et al.* Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat. Commun.* **3**, 945 (2012).

23. Arslan, Z., Hermanns, V., Wurm, R., Wagner, R. & Pul, Ü. Detection and characterization of spacer integration intermediates in type IE CRISPR--Cas system. *Nucleic Acids Res.* **42**, 7884–7893 (2014).

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

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

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

27. Nuñez, J. K. *et al.* Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. *Nat. Struct. Mol. Biol.* **21**, 528–534 (2014).

28. Sasnauskas, G. & Siksnys, V. CRISPR adaptation from a structural perspective. *Current Opinion in Structural Biology* **65**, 17–25 (2020).

29. Wilkinson, M. *et al.* Structure of the DNA-bound spacer capture complex of a type II CRISPR-Cas system. *Mol. Cell* **75**, 90–101.e5 (2019).

30. Hu, C. *et al.* Mechanism for Cas4-assisted directional spacer acquisition in CRISPR–Cas. *Nature* **598**, 515–520 (2021).

31. Burstein, D. *et al.* New CRISPR–Cas systems from uncultivated microbes. *Nature* **542**, 237–241 (2016).

32. Wright, A. V. *et al.* A functional mini-integrase in a two-protein type V-C CRISPR system. *Mol. Cell* **73**, 727–737.e3 (2019).

33. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Classification and nomenclature of CRISPR-Cas systems: where from here? *The CRISPR Journal* **1**, 325–336 (2018).

34. Santiago-Frangos, A., Buyukyoruk, M., Wiegand, T., Krishna, P. & Wiedenheft, B.

Distribution and phasing of sequence motifs that facilitate CRISPR adaptation. *Curr. Biol.* **31**, 3515–3524.e6 (2021).

35. Koonin, E. V. & Makarova, K. S. CRISPR-Cas: evolution of an RNA-based adaptive immunity system in prokaryotes. *RNA Biol.* **10**, 679–686 (2013).

36. Bertelsen, M. B. *et al.* Structural basis for toxin inhibition in the VapXD toxin-antitoxin system. *Structure* **29**, 139–150.e3 (2021).

37. Kwon, A.-R. *et al.* Structural and biochemical characterization of HP0315 from Helicobacter pylori as a VapD protein with an endoribonuclease activity. *Nucleic Acids Res.* **40**, 4216–4228 (2012).

38. Ka, D., Kim, D., Baek, G. & Bae, E. Structural and functional characterization of Streptococcus pyogenes Cas2 protein under different pH conditions. *Biochem. Biophys. Res. Commun.* **451**, 152–157 (2014).

39. Stamos, J. L., Lentzsch, A. M. & Lambowitz, A. M. Structure of a thermostable group II intron reverse transcriptase with template-primer and its functional and evolutionary implications. *Mol. Cell* **68**, 926–939.e4 (2017).

40. Wang, J. Y. *et al.* Structural coordination between active sites of a CRISPR reverse transcriptase-integrase complex. *Nat. Commun.* **12**, 2571 (2021).

41. Silas, S. *et al.* Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. *Science* **351**, aad4234 (2016).

42. Koonin, E. V. & Makarova, K. S. Mobile genetic elements and evolution of CRISPR-Cas systems: all the way there and back. *Genome Biol. Evol.* **9**, 2812–2825 (2017).

43. Silas, S. *et al.* On the origin of reverse transcriptase-using CRISPR-Cas systems and their hyperdiverse, enigmatic spacer repertoire. *MBio* **8**, (2017).

44. Mohr, G. *et al.* A reverse transcriptase-Cas1 fusion protein contains a Cas6 domain required for both CRISPR RNA biogenesis and RNA spacer acquisition. *Mol. Cell* **72**, 700–714.e8 (2018).

45. Nussenzweig, P. M. & Marraffini, L. A. Molecular mechanisms of CRISPR-Cas immunity in bacteria. *Annual Review of Genetics* **54**, 93–120 (2020).

46. Wu, C. *et al.* Mechanisms of spacer acquisition by sequential assembly of the adaptation module in Synechocystis. *Nucleic Acids Res.* **49**, 2973–2984 (2021).

47. Levy, A. *et al.* CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* **520**, 505–510 (2015).

48. Wigley, D. B. Bacterial DNA repair: recent insights into the mechanism of RecBCD, AddAB and AdnAB. *Nat. Rev. Microbiol.* **11**, 9–13 (2013).

49. Kim, S. *et al.* Selective loading and processing of prespacers for precise CRISPR adaptation. *Nature* **579**, 141–145 (2020).

50. Ramachandran, A., Summerville, L., Learn, B. A., DeBell, L. & Bailey, S. Processing and integration of functionally oriented prespacers in the Escherichia coli CRISPR system

depends on bacterial host exonucleases. *J. Biol. Chem.* **295**, 3403–3414 (2020).

51. Kieper, S. N. *et al.* Cas4 facilitates PAM-compatible spacer selection during CRISPR adaptation. *Cell Rep.* **22**, 3377–3384 (2018).

52. Lee, H., Dhingra, Y. & Sashital, D. G. The Cas4-Cas1-Cas2 complex mediates precise prespacer processing during CRISPR adaptation. *Elife* **8**, (2019).

53. Musharova, O. *et al.* Prespacers formed during primed adaptation associate with the Cas1–Cas2 adaptation complex and the Cas3 interference nuclease–helicase. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).

54. Drabavicius, G. *et al.* DnaQ exonuclease-like domain of Cas2 promotes spacer integration in a type I-E CRISPR-Cas system. *EMBO Rep.* **19**, (2018).

55. Lee, H., Zhou, Y., Taylor, D. W. & Sashital, D. G. Cas4-dependent prespacer processing ensures high-fidelity programming of CRISPR arrays. *Mol. Cell* **70**, 48–59.e5 (2018).

56. Heler, R. *et al.* Cas9 specifies functional viral targets during CRISPR–Cas adaptation. *Nature* **519**, 199–202 (2015).

57. Heler, R. *et al.* Mutations in Cas9 enhance the rate of acquisition of viral spacer sequences during the CRISPR-Cas immune response. *Mol. Cell* **65**, 168–175 (2017).

58. Jakhanwal, S. *et al.* A CRISPR-Cas9–integrase complex generates precise DNA fragments for genome integration. *Nucleic Acids Res.* **49**, 3546–3556 (2021).

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

60. Dillard, K. E. *et al.* Assembly and translocation of a CRISPR-Cas primed acquisition complex. *Cell* **175**, 934–946.e15 (2018).

61. Xue, C., Whitis, N. R. & Sashital, D. G. Conformational control of Cascade interference and priming activities in CRISPR immunity. *Mol. Cell* **64**, 826–834 (2016).

62. Nicholson, T. J. *et al.* Bioinformatic evidence of widespread priming in type I and II CRISPR-Cas systems. *RNA Biol.* **16**, 566–576 (2019).

63. Nussenzweig, P. M., McGinn, J. & Marraffini, L. A. Cas9 cleavage of viral genomes primes the acquisition of new immunological memories. *Cell Host Microbe* **26**, 515–526.e6 (2019).

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

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

66. Schmidt, F., Cherepkova, M. Y. & Platt, R. J. Transcriptional recording by CRISPR spacer acquisition from RNA. *Nature* **562**, 380–385 (2018).

67. Munck, C., Sheth, R. U., Freedberg, D. E. & Wang, H. H. Recording mobile DNA in the gut microbiota using an Escherichia coli CRISPR-Cas spacer acquisition platform. *Nature Communications* **11**, 1–11 (2020).

68. Wright, A. V., Nuñez, J. K. & Doudna, J. A. Biology and applications of CRISPR systems: harnessing nature's toolbox for genome engineering. *Cell* **164**, 29–44 (2016).

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

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

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

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

73. Brouns, S. J. J. *et al.* Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960–964 (2008).

74. Garneau, J. E. *et al.* The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67–71 (2010).

75. Hale, C. R. *et al.* RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* **139**, 945–956 (2009).

76. Wei, Y., Chesne, M. T., Terns, R. M. & Terns, M. P. Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in Streptococcus thermophilus. *Nucleic Acids Res.* **43**, 1749–1758 (2015).

77. Nuñez, J. K., Bai, L., Harrington, L. B., Hinder, T. L. & Doudna, J. A. CRISPR immunological memory requires a host factor for specificity. *Mol. Cell* **62**, 824–833 (2016).

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

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

80. Fagerlund RD, Wilkinson ME, Klykov O, Barendregt A, Pearce FG, Kieper SN, Maxwell HW, Capolupo A, Heck AJ, Krause KL, Bostina M, Scheltema RA, Staals RHJ, Fineran PC. Spacer capture and integration by a type I-F Cas1–Cas2-3 CRISPR adaptation complex. *PNAS* **114**, E5122–E5128 (2017).

81. Rollie, C., Graham, S. & Rouillon, C. Prespacer processing and specific integration in a Type IA CRISPR system. *Nucleic acids* **46**, 1007–1020 (2018).

82. Krupovic, M. & Koonin, E. V. Self-synthesizing transposons: unexpected key players in the evolution of viruses and defense systems. *Curr. Opin. Microbiol.* **31**, 25–33 (2016).

83. Chen, I.-M. A. *et al.* IMG/M: integrated genome and metagenome comparative data

analysis system. *Nucleic Acids Res.* **45**, D507–D516 (2017).

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

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

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

87. Deveau, H. *et al.* Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *J. Bacteriol.* **190**, 1390–1400 (2008).

88. Krupovic, M., Béguin, P. & Koonin, E. V. Casposons: mobile genetic elements that gave rise to the CRISPR-Cas adaptation machinery. *Current Opinion in Microbiology* **38**, 36–43 (2017).

89. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

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

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

92. Mével, M. *et al.* Novel neutral imidazole-lipophosphoramides for transfection assays. *Chem. Commun.* 3124–3126 (2008).

93. Liu, T. Y., Iavarone, A. T. & Doudna, J. A. RNA and DNA targeting by a reconstituted Thermus thermophilus type III-A CRISPR-Cas system. *PLoS One* **12**, e0170552 (2017).

94. Marraffini, L. A. CRISPR-Cas immunity in prokaryotes. *Nature* **526**, 55–61 (2015).

95. Kojima, K. K. & Kanehisa, M. Systematic survey for novel types of prokaryotic retroelements based on gene neighborhood and protein architecture. *Mol. Biol. Evol.* **25**, 1395–1404 (2008).

96. Simon, D. M. & Zimmerly, S. A diversity of uncharacterized reverse transcriptases in bacteria. *Nucleic Acids Res.* **36**, 7219–7229 (2008).

97. Toro, N. & Nisa-Martínez, R. Comprehensive phylogenetic analysis of bacterial reverse transcriptases. *PLoS One* **9**, e114083 (2014).

98. Toro, N., Mestre, M. R., Martínez-Abarca, F. & González-Delgado, A. Recruitment of reverse transcriptase-Cas1 fusion proteins by type VI-A CRISPR-Cas systems. *Frontiers in Microbiology* **10**, 2160 (2019).

99. Toro, N., Martínez-Abarca, F. & González-Delgado, A. The reverse transcriptases associated with CRISPR-Cas systems. *Sci. Rep.* **7**, 7089 (2017).

100. González-Delgado, A., Mestre, M. R., Martínez-Abarca, F. & Toro, N. Spacer acquisition from RNA mediated by a natural reverse transcriptase-Cas1 fusion protein associated with a type III-D CRISPR–Cas system in Vibrio vulnificus. *Nucleic Acids Research* **47**, 10202–10211 (2019).

101. Hale, C. R. *et al.* Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. *Molecular Cell* **45**, 292–302 (2012).

102. Tamulaitis, G. *et al.* Programmable RNA shredding by the type III-A CRISPR-Cas system of Streptococcus thermophilus. *Molecular Cell* **56**, 506–517 (2014).

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

104. Peng, W., Feng, M., Feng, X., Liang, Y. X. & She, Q. An archaeal CRISPR type III-B system exhibiting distinctive RNA targeting features and mediating dual RNA and DNA interference. *Nucleic Acids Research* **43**, 406–417 (2015).

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

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

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

108. Beloglazova, N. *et al.* A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. *Journal of Biological Chemistry* **283**, 20361–20371 (2008).

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

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

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

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

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

114. Shao, Y. *et al.* A non-stem-loop CRISPR RNA is processed by dual binding Cas6. *Structure*

24, 547–554 (2016).

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

116. Lambowitz, A. M. & Zimmerly, S. Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harb. Perspect. Biol.* **3**, a003616 (2011).

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

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

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

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

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

122. Rohou, A. & Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *Journal of Structural Biology* **192**, 216–221 (2015).

123. Tang, G. *et al.* EMAN2: An extensible image processing suite for electron microscopy. *Journal of Structural Biology* **157**, 38–46 (2007).

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

125. Punjani, A., Zhang, H. & Fleet, D. J. Non-uniform refinement: adaptive regularization improves single-particle cryo-EM reconstruction. *Nature Methods* **17**, 1214–1221 (2020).

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

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

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

129. Punjani, A. & Fleet, D. J. 3D variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM. *J. Struct. Biol.* **213**, 107702.

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

131. Liebschner, D. *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr D Struct Biol* **75**, 861–877 (2019).

132. Williams, C. J. *et al.* MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* **27**, 293–315 (2018).

133. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).

134. Goddard, T. D. *et al.* UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14–25 (2018).

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

136. Shiimori, M., Garrett, S. C., Graveley, B. R. & Terns, M. P. Cas4 nucleases define the PAM, length, and orientation of DNA fragments integrated at CRISPR loci. *Mol. Cell* **70**, 814–824.e6 (2018).

137. Makarova, K. S. & Koonin, E. V. Annotation and classification of CRISPR-Cas systems. *Methods in Molecular Biology* 47–75 (2015).

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

139. Biswas, A., Staals, R. H. J., Morales, S. E., Fineran, P. C. & Brown, C. M. CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics* **17**, 356 (2016).

140. Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019).

141. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).