UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title
Mechanism of CRISPR–Cas Immunological Memory

Permalink
https://escholarship.org/uc/item/26n3b6f5

Author
Nunez, James Karlo

Publication Date
2016

Peer reviewed|Thesis/dissertation

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

Mechanism of CRISPR–Cas Immunological Memory

by
James Karlo Nuñez

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy
in
Molecular and Cell Biology
in the
Graduate Division
of the
University of California, Berkeley

Committee in Charge:
Professor Jennifer A. Doudna, Chair
Professor Ming C. Hammond
Professor Eva Nogales
Professor David E. Wemmer

Spring 2016

ABSTRACT

Mechanism of CRISPR–Cas Immunological Memory

by
James Karlo Nuñez

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

Professor Jennifer A. Doudna, Chair

Immunological memory, defined as the ability for cells to rapidly respond to previously encountered pathogens, has long been thought to be present solely in eukaryotes. Recently, it was discovered that bacteria and archaea also possess analogous adaptive immune systems called CRISPR–Cas (clustered regularly interspaced short palindromic repeats–CRISPR associated). CRISPRs are distinct genomic loci comprising of repeating 20-50 base pair (bp) sequences that are separated by foreign DNA-derived segments called spacers. Upon viral infection, a ~30 bp fragment of the foreign DNA is acquired into the CRISPR locus as a new spacer, serving as a molecular memory of that specific infection. The CRISPR RNA (crRNA) products of CRISPR loci are used as guides by Cas proteins to silence infectious foreign genetic elements containing sequence complementarity to the crRNA. The research presented in this dissertation elucidates the mechanisms underlying spacer acquisition during CRISPR immunity.

Cas1 and Cas2, the only universally conserved Cas proteins present in CRISPR-harboring organisms, are sufficient to elicit spacer acquisition in *E. coli*. Using a combination of biochemical, bacterial genetics and X-ray crystallography experiments, we reveal the requirement of Cas1 and Cas2 to associate as a protein complex during CRISPR immunity. Cas1 is the catalytic subunit, whereas Cas2 functions as a scaffold within the complex. By determining crystal structures of Cas1–Cas2 bound to a 33 bp DNA from M13 bacteriophage, we uncover Cas1–Cas2 as a molecular ruler that precisely measures the foreign DNA across the entire complex with each DNA end terminating in the catalytic Cas1 active sites. Collectively, our structural data uncover how Cas1–Cas2 captures, measures and prepares foreign DNA for insertion into the CRISPR locus.

A hallmark of CRISPR immunological memory is the addition of foreign DNA spacers into CRISPR loci. By developing an *in vitro* assay to study spacer acquisition, we reveal that purified Cas1–Cas2 can be loaded with 18-50 bp spacer DNA for integration into a plasmid encoding the CRISPR locus. Cas1–Cas2 integrates spacers adjacent to the CRISPR repeats via a nucleophilic reaction of the DNA 3'–OH ends into the target DNA, in a mechanism similar to many retroviral integrases and transposases. Furthermore, the integration host factor (IHF) protein in *E. coli* binds the CRISPR leader to induce sufficient DNA bending required for the integration reaction, explaining the mechanistic basis for leader-proximal spacer acquisition observed *in vivo*. Together, these *in vitro* assays fully recapitulate immunological memory during prokaryotic CRISPR immunity.

ACKNOWLEDGEMENTS

The road to this thesis would not have been possible without the people who were part of the journey with me. I thank Jennifer for giving me freedom and support throughout my graduate career. I am forever indebted to three people who mentored and trained me in graduate school – Philip Kranzusch, Amy Lee and Stephen Floor. I cannot imagine how different my perspective on science would be if I had been mentored by anyone else. Thank you for your endless scientific advice and more importantly, your friendship. I will miss our daily shenanigans in the Pirate Bay and I cannot wait to see the great discoveries that you and your future labs will make! I owe a ton of credit to Alan Engelman for being an integrase guru and for mentoring me solely by email exchange. You guided us to the right path so many times and I hope our work on Cas1–Cas2 is only the beginning of future collaborations. I would not have attended graduate school without the guidance from my previous mentor and still very much one of my science heroes - Catherine Musselman. Thank you for everything!

Many of my favorite memories of graduate school were times spent with the people in the Doudna lab. Kaihong, thank you for managing the lab and instilling a lot of realism onto us. Enbo, I will miss joking around with you at every hour of the day even when you are busy! Megan, I cannot thank you enough for being the best thesis mate and for being there at every step of graduate school, in science and outside the lab! I would have gone insane without you. I thank everyone in the lab for creating an amazing environment. I especially thank those who were part of the rambunctious Doudna lab south side – Philip Kranzusch, Stephen Floor, Kendall Condon, Steve Lin, Yun Bai, Brett Staahl, Romain Rouet, Janice Chen, Lucas Harrington, Ashwin Banfal, Lawrence Bai, Mary Anne Kidwell, Ross Wilson and Addison Wright. The energy you brought to the south side was unforgettable and I will forever treasure your friendship. I had the great pleasure in overlapping with two young scientists who always kept my spirits up with spontaneous singing and nerding out on science and random topics – Kendall Condon and Chantal Guegler. I also had the great honor in becoming close friends with Janice Chen and Lucas Harrington. I wish we had overlapped for more than a year so I can witness the great discoveries you will make in the lab! I extend my acknowledgements to past and present members of the Nogales and Cate labs for contributing to the festive environment on the 7th floor, namely Amy Lee, Jonas Noeske, Annsea Park and Ben LaFrance for being my go-to people on the floor.

There are many people outside the lab who kept me well-rounded and sane. I thank my MCB 2011 classmates for a lot of unforgettable events from house parties, hikes and so many other social events. I thank friends at TAC for sharing my love for exercise and fitness. I am grateful to the many people who I got to share countless hours on the tennis court with. I thank my travel buddies for exploring the world with me: Megan Hochstrasser, David Taylor, Evan Worden, Lucas Brower, Cindy Wang, Kevin Doxzen, Dat Bui and Lucas Harrington. Lastly, I thank Tiffany Pollard for inspiration.

I am incredibly grateful to my family for continually cheering me on throughout my life. Your love and support have always propelled me to pursue my aspirations and I know you will continue to support me in the road ahead. I especially thank my mother and Lola Coring for the sacrifices they made for me and brothers. Every day, I am in awe of your work ethic and perseverance. Finally, I thank Lucas for being an incredible addition to my life this past year. I cannot wait to see what the future has in store for us!

This thesis is dedicated to my dearest mother, Nunem,
and grandmother, Lola Coring.

TABLE OF CONTENTS

Abstract **1**
Acknowledgements ..... **i**
Dedication ..... **ii**
Table of Contents ..... **iii**
List of Figures ..... **vi**
List of Tables ..... **viii**

Chapter 1. Prokaryotic CRISPR–Cas adaptive immunity **1**
1.1 CRISPR–Cas adaptive immunity 2
1.2 CRISPR–Cas diversity 4
1.3 Immunological memory during CRISPR immunity 4
1.3.1 Type I acquisition 5
1.3.2 Self- versus non-self recognition 6
1.3.3 Mechanism of protospacer integration 8
1.4 Aims and implications 9

Chapter 2. Cas1–Cas2 complex formation mediates CRISPR–Cas Immunological memory **10**
2.1 Introduction 11
2.2 Results 11
2.2.1 Cas1 and Cas2 form a specific complex 11
2.2.2 Crystal structure of the Cas1–Cas2 complex 14
2.2.3 Cas1–Cas2 complex formation is required *in vivo* 18
2.2.4 The catalytic activity of Cas2 is dispensable *in vivo* 18
2.2.5 Cas1–Cas2 specifically binds the CRISPR locus 20
2.3 Discussion 21
2.4 Materials and Methods 24
2.4.1 Cas1 and Cas2 protein preparation 24
2.4.2 *In vitro* Cas1–Cas2 complex formation 24
2.4.3 Crystallization and structure determination 24
2.4.4 Isothermal titration calorimetry 25
2.4.5 Spacer acquisition assays 25
2.4.6 Immunoprecipitation assays 26
2.4.7 DNA affinity precipitation assays 26
2.4.8 Structure-based sequence alignments 26
2.5 Accession Code 26
2.6 Acknowledgements 27
2.7 Author Contributions 27

Chapter 3. Mechanism of foreign DNA integration into CRISPR loci **28**
3.1 Introduction 29
3.2 Results 29
3.2.1 Protospacer DNA integration by Cas1–Cas2 29
3.2.2 Integration and disintegration products 32

3.2.3 Integration requires 3′–OH protospacer ends..... 35
3.2.4 Supercoiled DNA and CRISPR locus requirements..... 37
3.2.5 CRISPR repeats provide integration specificity..... 38
3.2.6 Cas1–Cas2 orients protospacers into CRISPR loci..... 40
3.3 Discussion..... 43
3.4 Materials and Methods..... 45
2.4.1 Cas1, Cas2 and DNA protein preparation..... 45
2.4.2 *In vitro* integration assays..... 45
2.4.3 Radiolabeled protospacer integration assays 45
2.4.4 Disintegration assays..... 46
2.4.5 High-throughput sequencing..... 46
2.4.6 Computational analysis..... 47
3.5 Accession Code..... 47
3.6 Acknowledgements..... 47
3.7 Author Contributions..... 47

Chapter 4. Mechanism of foreign DNA capture and length determination..... 48
4.1 Introduction..... 49
4.2 Results..... 49
4.2.1 Cas1–Cas2 prefers protospacers with 3′ overhang ends..... 49
4.2.2 Crystal structure of Cas1–Cas2 bound to foreign DNA..... 51
4.2.3 Coordination of 3′–OH nucleophiles and DNA backbone..... 53
4.2.4 Mechanism of protospacer DNA end separation..... 56
4.2.5 Model of protospacer DNA integration..... 57
4.3 Discussion..... 60
4.4 Materials and Methods..... 61
4.4.1 Cas1, Cas2 and DNA preparation..... 61
4.4.2 *In vivo* spacer acquisition and in vitro integration assays..... 61
4.4.3 Cas1–Cas2 complex formation with protospacer DNA..... 61
4.4.4 Crystallization and structure determination..... 62
4.4.5 Fluorescence polarization..... 62
4.4.6 Cas1 sequence alignment..... 62
4.5 Accession Code..... 63
4.6 Acknowledgements..... 63
4.7 Author Contributions..... 63

Chapter 5. A host factor essential for CRISPR immunological memory..... 64
5.1 Introduction..... 65
5.2 Results..... 65
5.2.1 IHF specifically binds the CRISPR leader..... 65
5.2.2 IHF is required for *in vivo* spacer acquisition..... 68
5.2.3 IHF specifies leader-proximal integration..... 70
5.2.4 Sequence-specific integration into leader-repeat border..... 72
5.2.5 Target DNA bending drives spacer integration..... 74
5.3 Discussion..... 76
5.4 Materials and Methods..... 78

5.4.1 DNA and protein preparation……. 78
5.4.2 Knockout strain construction, growth assays and *in vivo* acquisition assays..... 78
5.4.3 Gel mobility shift assays..... 79
5.4.4 Hydroxyl radical footprinting..... 79
5.4.5 *In vitro* integration assays..... 79
5.4.6 High throughput sequencing..... 80
5.5 Accession Code..... 81
5.6 Acknowledgements..... 81
5.7 Author Contributions..... 82

Chapter 6. Summary and Applications..... 83
6.1 Mechanism of CRISPR–Cas immunological memory..... 84
6.2 Future directions..... 84
6.3 Applications..... 86

Bibliography..... 87

Appendix I. Chemical and biophysical modulation of Cas9 for tunable genome engineering……..... 99
I.1 Abstract…..... 100
I.2 Introduction..... 100
I.3 Split-Cas9 systems for genome engineering..... 102
I.3.1 Intein-inducible split-Cas9..... 103
I.3.2 Rapamycin-inducible split-Cas9..... 104
I.3.3 Optically controlled split-Cas9..... 104
I.4 Optogenetic control of Cas9..... 104
I.5 Chemically inducible Cas9..... 106
I.6 Natural inhibitors of CRISPR systems..... 108
I.7 Summary and perspectives..... 108
I.8 Acknowledgements..... 109

LIST OF FIGURES

Figure 1.1 Prokaryotic defense systems 1
Figure 1.2 CRISPR–Cas adaptive immunity 3
Figure 1.3 Classification of CRISPR–Cas systems 4
Figure 1.4 PAM-dependent spacer acquisition in *E. coli* 5
Figure 1.5 Overall mechanism of spacer acquisition 6
Figure 1.6 Protospacer substrate generation by the RecBCD complex 7
Figure 1.7 Cas1 and Cas2 proteins 8

Figure 2.1 Cas1 and Cas2 associate to form a complex..... 12
Figure 2.2 Reconstitution and crystallization of the Cas1–Cas2 complex..... 14
Figure 2.3 Crystal structure of the Cas1–Cas2 complex..... 16
Figure 2.4 Disruption of complex formation abrogates spacer acquisition..... 17
Figure 2.5 Cas1 is the catalytic subunit within the Cas1–Cas2 complex..... 19
Figure 2.6 Cas1–Cas2 complex specifically recognizes the CRISPR locus..... 20
Figure 2.7 Structure-based Cas1 and Cas2 alignments..... 23

Figure 3.1 The Cas1–Cas2 complex integrates protospacers *in vitro*..... 30
Figure 3.2 Radiolabeled protospacers are integrated into pCRISPR..... 31
Figure 3.3 Assignment of Cas1–Cas2-catalyzed integration products..... 32
Figure 3.4 Relaxed plasmid topoisomers byproducts..... 33
Figure 3.5 Cas1 catalyzes disintegration reactions..... 34
Figure 3.6 Integration requires 3′–OH protospacer ends..... 35
Figure 3.7 Cas1 tyrosine mutants support integration activity..... 36
Figure 3.8 Supercoiled target requirement..... 37
Figure 3.9 Integration occurs into the CRISPR locus..... 39
Figure 3.10 Off-target integration into pUC19..... 40
Figure 3.11 CRISPR repeat mutations abolish integration..... 41
Figure 3.12 Cas1–Cas2 correctly orients the protospacer..... 42
Figure 3.13 Mechanism of spacer integration..... 43

Figure 4.1 Cas1–Cas2 prefers protospacers with 3′ overhang ends..... 50
Figure 4.2 Crystal structure of Cas1–Cas2 bound to protospacer DNA..... 51
Figure 4.3 Omit maps of the protospacer DNA..... 53
Figure 4.4 Active site positioning of nucleophilic 3′–OH..... 54
Figure 4.5 Coordination of protospacer DNA within Cas1–Cas2..... 55
Figure 4.6 Cas1–Cas2 splays the ends of the protospacer DNA..... 56
Figure 4.7 Sequence alignment of Cas1 proteins in Type I CRISPR systems..... 58
Figure 4.8 Model of protospacer DNA integration..... 59

Figure 5.1 IHF specifically binds the CRISPR leader..... 66
Figure 5.2 IHF binds to a consensus sequence in the leader..... 67
Figure 5.3 IHF knockout strain construction..... 68
Figure 5.4 IHF is required for spacer acquisition *in vivo*..... 69
Figure 5.5 IHF directs leader-proximal spacer integration..... 71

Figure 5.6 IHF directs leader-proximal spacer integration..... 73
Figure 5.7 CRISPR DNA bending drives protospacer integration..... 75
Figure 5.8 mechanism of spacer integration in *E. coli*..... 77

Figure 6.1 Mechanism of CRISPR–Cas immunological memory..... 84
Figure 6.2 Schematic of T cell barcoding strategy..... 86

Figure I.1 CRISPR–Cas9 genome editing technology..... 101
Figure I.2 Split-Cas9 strategies for controllable genome editing..... 103
Figure I.3 Cas9-mediated optogenetic editing and transcriptional activation..... 105
Figure I.4 Chemicals used for Cas9-mediated genome editing..... 107

LIST OF TABLES

Table 2.1 Sequence of newly acquired spacers..... 13
Table 2.2 X-ray crystallography data collection and refinement statistics..... 15

Table 4.1 Summary of X-ray crystallography data collection and refinement.....52

Table 5.1 DNA sequences used for cloning, knockout strain construction and integration assay experiments..... 81

Table I.1 Split-Cas9 systems for genome engineering..... 102

CHAPTER 1

Prokaryotic CRISPR–Cas adaptive immunity

A portion of the work presented in this chapter has been published previously as part of the following paper: Wright, A.V., Nuñez, J.K. & Doudna, J.A. Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. *Cell* **164**, 29-44 (2015).

1.1 CRISPR–Cas adaptive immunity

The ability to fight infectious genetic elements is a vital part of life. Like eukaryotes, bacteria and archaea are in constant battle against foreign infections. As a result, prokaryotes have evolved a remarkable repertoire of defense mechanisms to survive infections against viruses and plasmids[1,2]. The host can employ restriction endonucleases for sequence-specific cleavage of foreign DNA or block phage receptors to evade phage recognition on the host surface (**Fig. 1.1**). The host may also elicit an abortive infection response, a form of suicide to protect the population from propagation of viral particles. Although these defense systems are general innate responses to many infections, they do not provide adaptive immunity against each invader. A recent discovery of an adaptive immune system called CRISPR–Cas has uncovered a sophisticated mechanism that bacteria and archaea use to silence foreign nucleic acids, allowing the host to elicit a robust immune response upon re-infection of each invader.

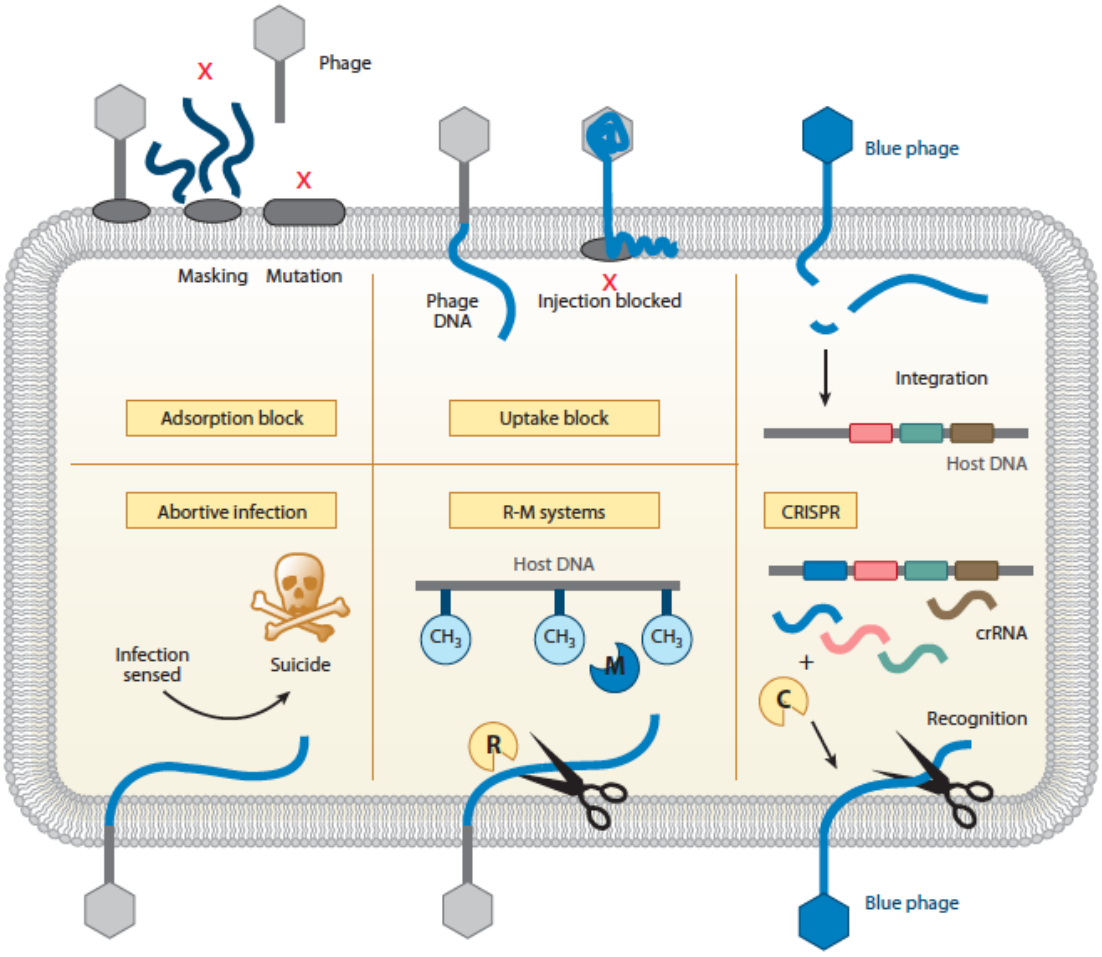


Figure 1.1 | Prokaryotic defense systems. An overview of immune systems found in bacteria and archaea. Adsorption block, uptake block, abortive infection and restriction-modification (R-M) provide general immunity against foreign infections, whereas CRISPR adaptive immunity allows the host to silence each invader using short RNA molecules. The figure is adapted from Westra et al., 2012.

Present in the genomes of approximately half of bacterial species and almost all archaea, the clustered regularly interspaced short palindromic repeats (CRISPR) – CRISPR associated (Cas) immune system relies on short RNAs for sequence-specific cleavage of complementary DNA or RNA[3,4]. CRISPR genomic loci consist of repeat sequences, typically 20-50 base pairs (bp) in length, separated by variable spacer sequences of similar length that frequently match a segment of foreign nucleic acids[5–9]. Directly upstream of these repeat-spacer arrays is an AT-rich leader region that typically spans 100-300 bp[6].

Productive immunity occurs in three stages – adaptation or spacer acquisition, CRISPR RNA (crRNA) processing and interference[4] (**Fig. 1.2**). During spacer acquisition, a segment of foreign DNA is incorporated into the CRISPR locus as a new spacer[3]. Subsequently, the CRISPR locus is transcribed to produce a transcript that becomes processed into short crRNAs[10–13]. Each crRNA consists of one spacer sequence and portions of the flanking repeats. During interference, the crRNAs are used as guides by Cas proteins for targeting foreign nucleic acids with sequence complementarity to the spacer region of the crRNA[14–17].

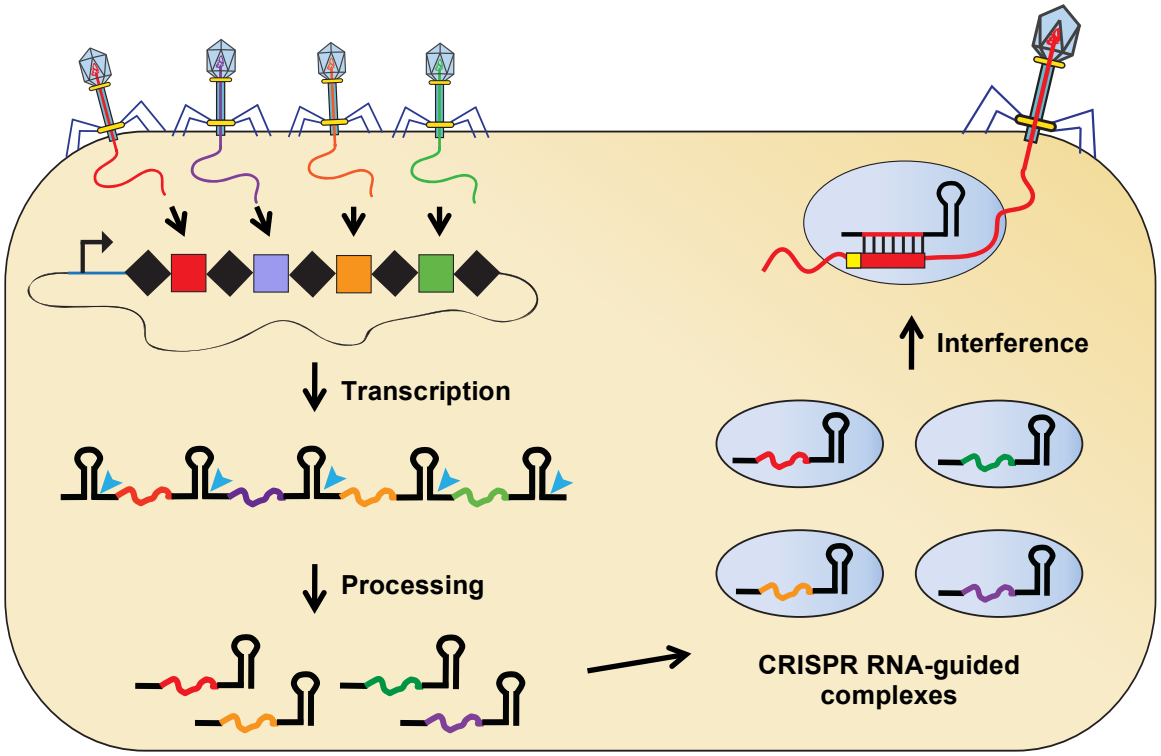


Figure 1.2 | CRISPR–Cas adaptive immunity. Foreign DNA spacers are incorporated into the CRISPR locus to generate memory of each infection. The black diamonds represent the repeats and the squares variable spacer sequences. The CRISPR locus is transcribed into a long precursor crRNA that becomes processed by an endoribonuclease (blue arrows) to generate short crRNA guides. The crRNAs are subsequently loaded into an interference protein complex that searches for complementary foreign nucleic acids for silencing.

1.2 CRISPR–Cas diversity

CRISPR–Cas systems are highly diverse and are classified by the Cas proteins that function in the interference stage[18]. Class 1 CRISPR systems encode multi-subunit protein interference complexes and Class 2 systems employ a single protein for interference[18,19] (**Fig. 1.3**). Class 1 systems are further subdivided into three different Types – I, III and IV – based on distinct proteins in the targeting complexes. Class 2 systems are divided into Types I, V and VI. Cas9 is the signature protein of Type II, Type V encode either Cpf1, C2c1 or C2c3 and Type VI systems contain C2c2. Although most *cas* genes are highly divergent and occur only in certain CRISPR loci, *cas1* and *cas2* are notably conserved across CRISPR systems. Genetic experiments, as well as spacer acquisition assays in *Escherichia coli* K12, demonstrate that Cas1 and Cas2 are the only Cas proteins required for new spacer acquisition into the host CRISPR locus[20–22]. The conserved presence of *cas1* and *cas2* suggest a common mechanism of spacer acquisition across the three CRISPR types. Despite these findings, along with previous biochemical studies identifying Cas1 and Cas2 as metal-dependent nucleases[23–27], the molecular functions of Cas1 and Cas2 during CRISPR–Cas immunity remain elusive.

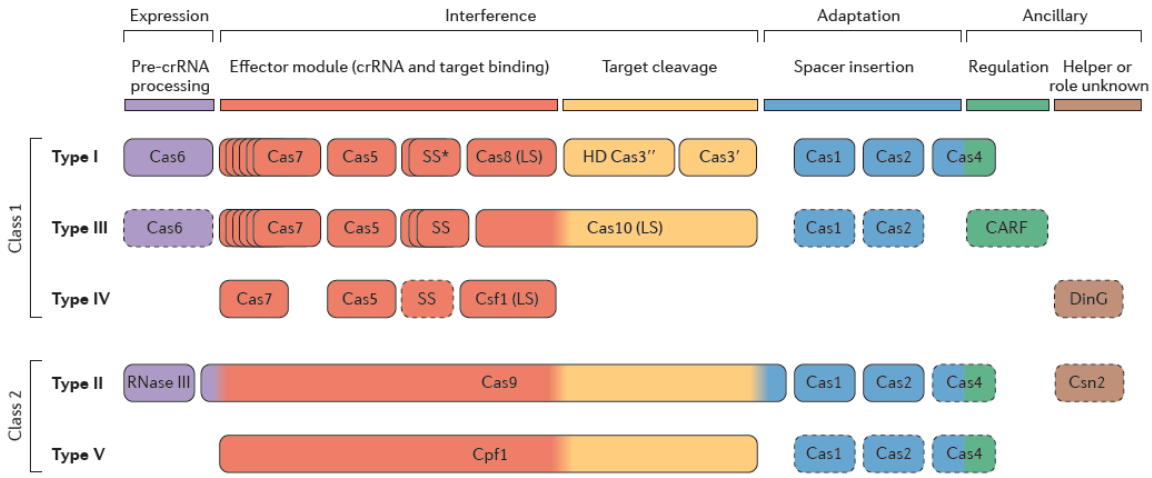


Figure 1.3 | Classification of CRISPR–Cas systems. Class 1 CRISPR systems use multi-subunit protein complexes for interference whereas Class2 systems rely on a single protein for RNA-guided interference. Adaptation or spacer acquisition relies on Cas1 and Cas2 proteins that are present in almost all CRISPR systems. Adapted from Makarova et al., 2015.

1.3 Immunological memory during CRISPR immunity

The spacer acquisition stage of CRISPR–Cas immunity allows the host to gain memory of previous foreign encounters – a hallmark of prokaryotic immunological memory[28,29]. In the *Streptococcus thermophilus* Type II-A system, where acquisition was first detected experimentally, new spacers from bacteriophage DNA are inserted into the leader end of the CRISPR locus, causing duplication of the first repeat to maintain the repeat-spacer architecture[3]. Subsequent studies using the *E. coli* Type I-E system verified that Cas1 and Cas2 mediate spacer acquisition[20–22]. The selection of

new protospacer sequences is nonrandom and, in most systems, depends on the presence of a 2-5 nucleotide protospacer adjacent motif (PAM) found next to the protospacer sequence[30,31] (**Fig. 1.4**). PAM-specific selection of protospacers is critical for immunity, as crRNA-guided interference in most systems depends on the PAM sequence for foreign DNA detection and destruction, which avoids self-targeting at the PAM-free CRISPR locus. Interestingly, spacers originating from the host genome are present in almost 20% of CRISPR-containing organisms, suggesting alternative roles of the CRISPR-Cas machinery in directing other processes such as endogenous gene regulation and genome evolution[32]. Spacer acquisition has been observed experimentally in various systems across Types I-III. This thesis focuses on the mechanism of spacer acquisition in Type I-E systems.

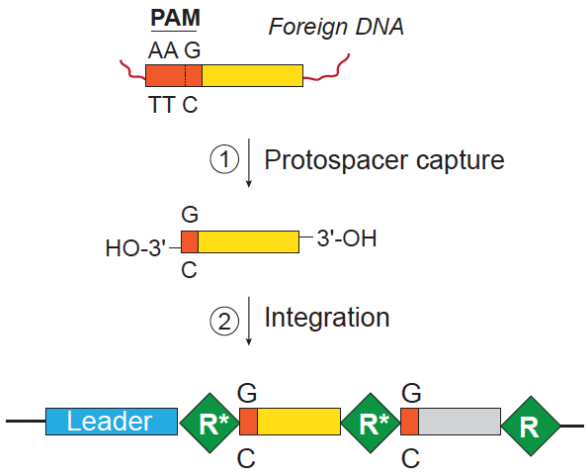


Figure 1.4 | PAM-dependent spacer acquisition in *E. coli*. Spacer acquisition in Type I systems occurs in a PAM-dependent manner. In *E. coli*, the AWG PAM is recognized in foreign DNA and is processed to 33 bp. The first nucleotide of the mature protospacer is 5′-G/3′-C. During spacer integration, the protospacer is oriented with the 5′-G becoming the first nucleotide proximal to the leader. The first repeat becomes duplicated to maintain the repeat-spacer architecture in the CRISPR locus (asterisk).

1.3.1 Type I acquisition

Spacer acquisition in *E. coli* occurs via two mechanisms – naïve and primed. Naïve acquisition initiates upon infection by previously unencountered DNA and relies on the Cas1–Cas2 integrase complex to recognize and acquire new spacers from foreign DNA (**Fig. 1.5**). Overexpression of Cas1 and Cas2 in the absence of other Cas proteins leads to the acquisition of 33 bp spacers at the leader-proximal end of the CRISPR array[20,21]. The PAM of the *E. coli* CRISPR–Cas system was identified as 5′-AWG-3′, with the G becoming the first nucleotide of the integrated spacer[20–22,33–37]. In addition to the PAM, a dinucleotide motif, AA, found at the 3′ end of the protospacer was also shown to be present in a disproportionately large number of spacers[38]. A crystal structure of the Cas1–Cas2 complex bound to an unprocessed protospacer revealed sequence-specific contacts with the 5′-CTT-3′ sequence on the PAM-complementary strand, suggesting that Cas1 recognizes PAM sites on potential protospacers before they are processed for integration[39].

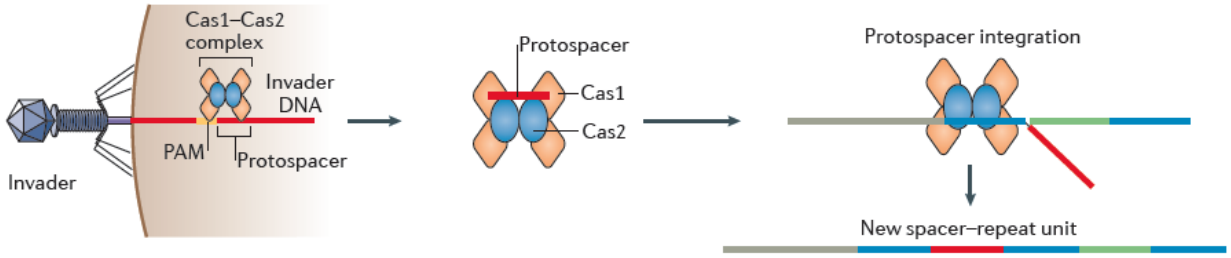


Figure 1.5 | Overall mechanism of spacer acquisition. Cas1 and Cas2 function as a complex to detect foreign DNA adjacent to PAM sequences. Upon capture of foreign DNA protospacers, Cas1–Cas2 catalyzes the integration of new spacers into the CRISPR locus, resulting in the duplication of the first repeat (blue). Adapted from Amitai and Sorek, 2016.

After a spacer is acquired from a new invader, the resulting crRNA assembles with Cas proteins to form Cascade, the interference complex capable of targeting PAM-adjacent DNA sequences matching the spacer sequence of the crRNA[14,40,41]. Upon target binding, the helicase/nuclease Cas3 is recruited to the site and processively degrades the foreign DNA[42–46]. Strikingly, when Cascade encounters a mutant PAM or protospacer that prevents Cas3 degradation, hyperactive spacer acquisition from the targeted plasmid or genome is triggered in a process called priming[21,22,34,47,48]. Priming increases the host's repertoire of functional spacers, allowing the host to adapt to invaders that evade the CRISPR–Cas system by mutation. Cascade is capable of binding escape mutant target sites, and recent single molecule studies showed that the presence of Cas1 and Cas2 allows for the recruitment of Cas3 to these sites[48–50]. The recruited Cas3 can then translocate in either direction, in contrast to the unidirectional movement observed at perfect targets, without degrading the target DNA[50]. Cas1 and Cas2 may accompany the translocating Cas3 and be activated for protospacer selection, allowing for robust acquisition on either side of the target site.

Primed spacer acquisition has also been shown experimentally in the *Pectobacterium atrosepticum* Type I-F system, in which Cas1 and Cas2 are naturally fused as a single polypeptide that associates with Cas1, as well as in the *Haloarcula hispanica* Type I-B system, where naïve acquisition was not experimentally observed[47,48,51,52]. Spacer acquisition in *H. hispanica* also requires Cas4, a 5′→3′ exonuclease found in most Type I subtypes as well as Type II-B and Type V systems, and which might be involved in generating 3′ overhangs on protospacers prior to integration[18,51,53]. Although Cas1 and Cas2 maybe the minimal proteins required for spacer acquisition in some systems, the association of Cas1, Cas2 and the interference machinery allows the host to coordinate robust adaptive immunity in Type I systems.

1.3.2 Self- versus non-self recognition

The mechanism underlying the preference for foreign over self DNA during protospacer selection remained poorly understood until a recent study on spacer acquisition in *E. coli* (**Fig. 1.6**). Naïve spacer acquisition was shown to be highly dependent on DNA replication and foreign –derived spacers were preferred over self-

derived spacers by bout 100- to 1,000-fold[37]. Analysis of the source of self-derived spacers demonstrated that protospacers were acquired largely from genomic loci predicted to frequently generate stalled replication forks and double-stranded DNA breaks[37]. Such harmful dsDNA breaks are repaired by the helicase-nuclease RecBCD complex, which degrades the broken ends until reaching a Chi-site, after which only the 5′ is degraded[54] (**Fig. 1.6**). Due to the lower frequency of Chi sites in foreign DNA, RecBCD is predicted to preferentially degrade plasmids and viral DNA, resulting in the generation of candidate protospacer substrates for Cas1 and Cas2 (ref. 37). RecBCD degrades DNA asymmetrically, yielding single-stranded fragments ranging from tens to hundreds of nucleotides long from one strand and kilobases long from the other[54]. It is unclear how Cas1–Cas2 substrates, which are 33 bp long, are generated from RecBCD products. It is possible that ssDNA products re-anneal to produce partial duplexes, followed by processing to 33 bp by an unknown mechanism prior to integration into the CRISPR locus. Recent crystal structures of Cas1–Cas2 with bound protospacer reveal that the complex defines the length of the duplex region of the protospacer via a ruler mechanism and may cleave the 3′ overhangs to their final length[39,55]. The involvement of a helicase/nuclease in both Type I-E primed and naïve acquisition (Cas3 and RecBCD, respectively), as well as Cas4-containing subtypes, hints at a conserved mechanism for protospacer generation. It is also worth noting that RecBCD is conserved primarily in Gram-negative bacteria, while Gram-positive bacteria and archaea rely on AddAB and HerA-NurA, respectively, to fill a similar role[54,56]. Whether CRISPR–Cas systems in these organisms have evolved to cooperate with these evolutionarily distinct machineries remains to be tested.

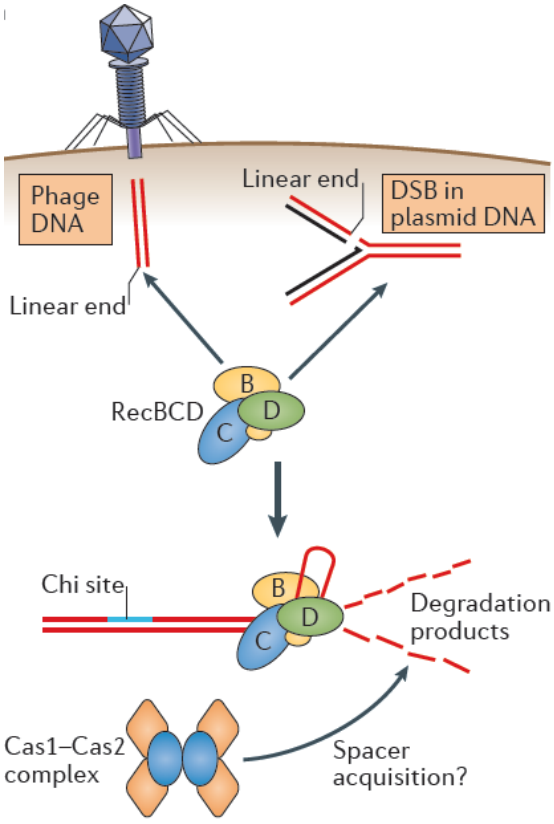


Figure 1.6 | Protospacer substrate generation by the RecBCD complex. It is hypothesized that during naïve spacer acquisition, double-stranded DNA breaks (DSB) in foreign DNA are recognized and processed the RecBCD complex. Degradation products from the processing step are recognized by the Cas1–Cas2 complex as protospacers. Protospacers derived from host genome are underrepresented due to the presence of many Chi sites in the *E. coli* genome, which slow the rate of DNA processing by RecBCD. Thus, protospacers are mostly derived from foreign DNA due to high levels of DNA processing and substrate generation of plasmids and viral genomes. Adapted from Amitai and Sorek, 2016.

1.3.1 Mechanism of protospacer integration

Cas1 and Cas2 play central roles in the acquisition of new spacers, where they function as a complex[35]. Initial studies of Cas1 and Cas2 focused on crystal structures of the individual proteins and their biochemical properties (**Fig. 1.7**). Crystal structures of Cas1 and Cas2, with or without bound protospacer DNA, revealed two copies of a Cas1 dimer bridged by a central Cas2 dimer[35,39,55]. Cas1 functions catalytically, while Cas2 appears to serve a primarily structural role[20,21,35,57].

The first insight into the mechanism of protospacer integration was gained by Southern blot analysis of the genomic CRISPR locus of *E. coli* cells overexpressing Cas1 and Cas2. This revealed integration intermediates consistent with two transesterification reactions, where each strand of the protospacer is integrated into opposite sides of the leader proximal repeat[57]. This integrase-like model was further bolstered by the *in vitro* reconstitution of protospacer integration into a plasmid-encoded CRISPR locus using purified Cas1–Cas2 complex[58]. The integration reaction required double-stranded DNA protospacers with 3′–OH ends that are integrated into plasmid via a direct nucleophilic transesterification reaction, reminiscent of retroviral integrases and DNA transposases[59,60].

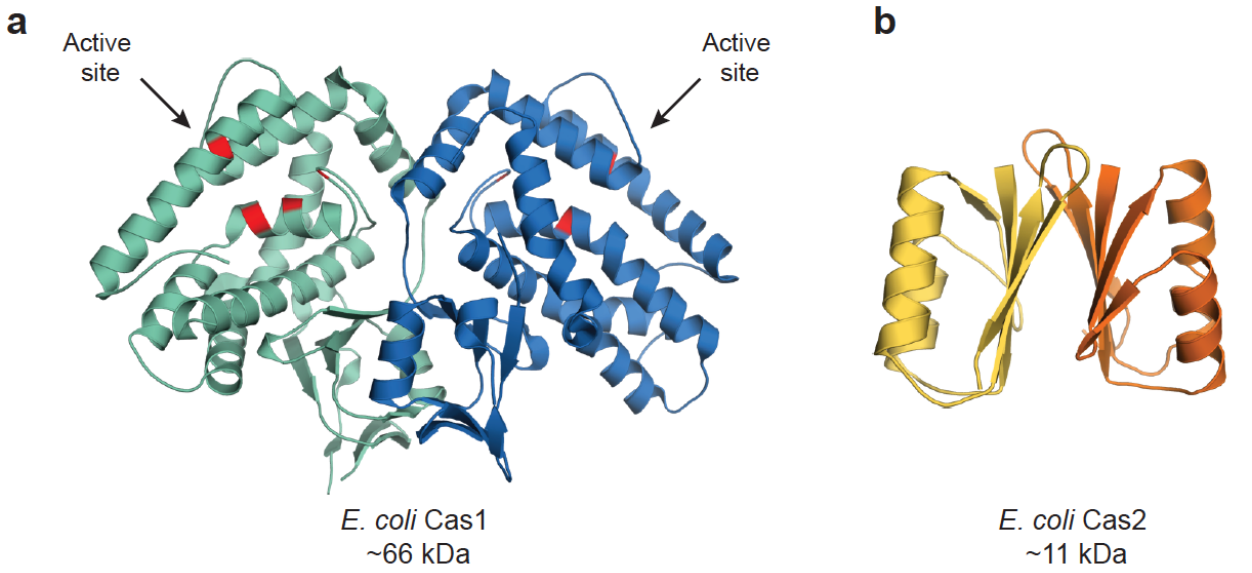


Figure 1.7 | Cas1 and Cas2 proteins. (a) Crystal structure of the *E. coli* Cas1 homodimer (PDB 3NKD). The predicted active sites are highlighted in red. **(b)** Crystal structure of the *E. coli* Cas2 homodimer (PDB 4MAK).

Although deep sequencing of *in vitro* integration products revealed preferential protospacer integration adjacent to the first repeat, confirming that Cas1–Cas2 directly recognize the CRISPR locus, integration also occurred at the borders of every repeat at varying levels[58]. This contrasts with spacer acquisition only occurring at the first repeat in *E. coli in vivo*[20–22]. Chapter 5 of this thesis reveals the involvement of the integration

host factor (IHF) in binding and bending the CRISPR leader, allowing Cas1–Cas2 to integrate adjacent to the first repeat *in vitro.* Target DNA bending is a common requirement in other DNA integration systems, further highlighting similarities in Cas1–Cas2-mediated spacer integration with other integration and transposition systems.

To determine if the Cas1–Cas2 complex has sequence specificity for the leader-repeat sequence, a recent study took advantage of the Cas1-catalyzed disintegration reaction, a reversal of the integration reaction also observed with retroviral integrases and transposases[61]. Disintegration activity was stimulated when using the correct leader-repeat border sequences, highlighting intrinsic sequence-specific recognition by Cas1. Furthermore, disintegration was faster at the leader-repeat junction compared to the repeat distal end. Taken together, protospacer integration likely begins at the leader-repeat junction via sequence-specific recognition by Cas1, followed by a second nucleophilic attack at the leader distal end. This ensures precise duplication of the first repeat, as observed *in vivo*, after repair by host proteins. The integration mechanism is hypothesized to be highly specific, as almost all acquired spacers with a corresponding AAG PAM are oriented with the 5′-G at the leader-proximal end, leading to functional crRNA-dependent targeting by Cascade and Cas3 (ref. 36). A preference for integration in the protospacer orientation was observed *in vitro* when protospacers with a 5′-G were used[58]; however, inclusion of part of the PAM in spacers has only been observed in *E. coli*, raising the question of how Cas1–Cas2 in other systems properly orient the integration reaction.

1.4 Aims and implications

The aim of this thesis is to understand the basic biology of immunological memory during CRISPR–Cas adaptive immunity. Specifically, the functions of Cas1 and Cas2 will be discussed using findings from biochemical, structural and genetic experiments in *E. coli*. Following capture of foreign DNA protospacers, we discuss the mechanism of protospacer integration into CRISPR loci. With mechanistic understanding of Cas1–Cas2-mediated spacer acquisition, we aim to open new avenues for developing novel CRISPR biotechnology applications based on DNA integration.

CHAPTER 2

Cas1–Cas2 complex formation mediates CRISPR–Cas immunological memory

A portion of the work presented in this chapter has been published previously as part of the following paper: Nuñez, J.K., Kranzusch, P.J., Noeske, J., Wright, A.V., Davies, C.W. & Doudna, J.A. Cas1–Cas2 complex formation mediates CRISPR–Cas adaptive immunity. *Nature Structural & Molecular Biology* **21**, 528-534 (2014).

2.1 Introduction

The first stage of CRISPR–Cas adaptive immunity is the capture and acquisition of foreign DNA sequences into the host CRISPR locus, as part of immunological memory formation in bacteria and archaea[3,28,29]. Although most *cas* genes are highly divergent and occur only in certain CRISPR loci, *cas1* and *cas2* are conserved across almost all CRISPR systems[18]. Genetic experiments in *Escherichia coli* K12 demonstrate that Cas1 and Cas2 are the only Cas proteins required for new spacer acquisition[20,21]. The conserved presence of Cas1 and Cas2 suggest a common mechanism of spacer acquisition across all CRISPR–Cas systems. Despite these findings, along with previous biochemical studies identifying Cas1 and Cas2 as metal-dependent nucleases, the molecular functions of Cas1 and Cas2 during CRISPR–Cas immunity remain elusive.

We used biochemical, structural and genetic experiments to investigate the specific functions of Cas1 and Cas2 during CRISPR–Cas immunity. We show that Cas1 and Cas2 form a stable complex *in vitro* and *in vivo* and present the crystal structure of the *E. coli* Cas1–Cas2 complex. We combine an *in vivo* spacer acquisition assay with mutagenesis and immunoprecipitation experiments to show that physical disruption of complex formation abrogates spacer acquisition. While active site mutations in Cas1 inhibit spacer acquisition, the catalytic activity of Cas2 is not required for either Cas1–Cas2 complex formation or spacer acquisition *in vivo*. The Cas1–Cas2 complex is uniquely capable of recognizing the CRISPR leader-repeat sequence, a property not shared by either protein alone. Together, these results provide key functional insights into a Cas1–Cas2 complex that are likely to be shared across all CRISPR systems.

2.2 Results

2.2.1 Cas1 and Cas2 form a specific complex

The *E. coli* K12 (MG1655) strain has two endogenous CRISPR loci, one of which is flanked by eight *cas* genes[62] (**Fig. 2.1a**). In agreement with a previously developed assay[20], when Cas1 and Cas2 from K12 were overexpressed in *E. coli* BL21-AI cells, which lack all Cas-encoding genes, new spacer acquisition was detectable by PCR amplification of the CRISPR locus (**Fig. 2.1b**). We sequenced newly acquired spacers and verified that spacer acquisition in this model system retained accurate acquisition of 33-bp spacers, mostly derived from the foreign plasmid used for protein overexpression (**Table 2.1**). In addition to the insertion of the 33 bp spacer, each acquisition event duplicates the first repeat (28 bp), thereby expanding the parental locus by 61 bp. Although these results demonstrate that spacer acquisition requires only the proteins Cas1 and Cas2, we observed variable protospacer adjacent motif (PAM) sequences adjacent to the protospacer in the foreign DNA. These results support the conclusion that the *E. coli* CRISPR interference machinery, consisting of the Cascade complex and the Cas3 nuclease/helicase, are required for an accurate 'priming' process in which the interference stage is coupled to spacer acquisition to yield strict AAG PAM selection[21,22,33,34].

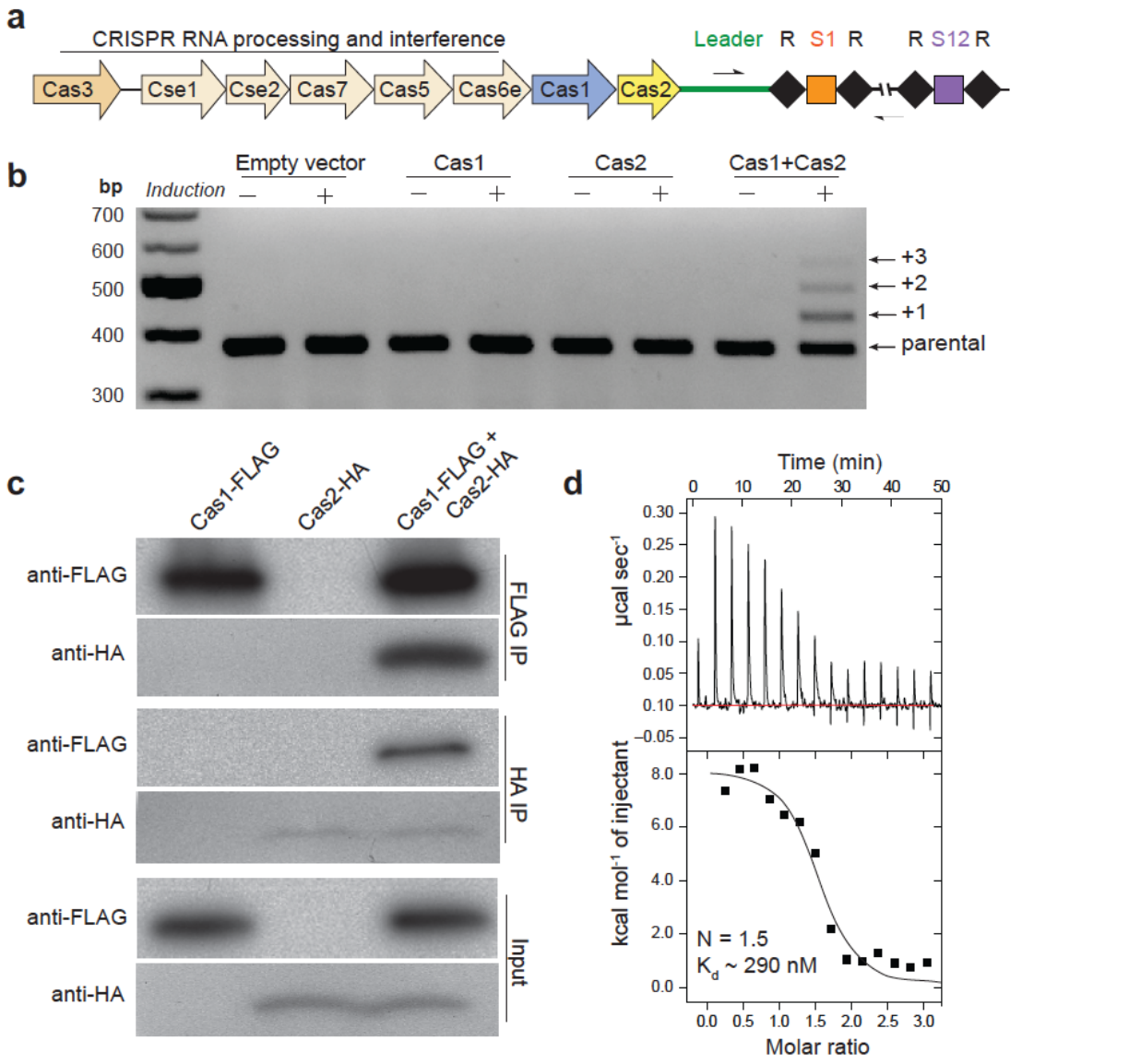


Figure 2.1 | Cas1 and Cas2 associate to form a complex. (**a**) The CRISPR locus of *E. coli* K12 consists of 28-bp repeats (black diamonds) separated by 33-bp spacers (squares). The black arrows flanking the leader and repeat-spacer array represent the primers used for PCR amplification in spacer acquisition assays. (**b**) Agarose gel of the PCR amplified CRISPR-I locus of BL21-AI cells after induced expression of empty vector, Cas1, Cas2 or both Cas1 and Cas2. Distinct bands represent the number of repeat-spacer array additions into the genomic CRISPR locus. (**c**) FLAG and HA immunoprecipitations in lysates overexpressing Cas1 only, Cas2 only or both. (**d**) ITC trace of Cas1 injection into a cell containing Cas2. The reported N (stoichiometry) and K_d values are averages from three independent experiments.

Source	Strand	PAM	Length	Sequence	Location
plasmid	+	ATG	33	GTTAGTCATGCCCCGCGCCCACCGGAAGGAGCT	
plasmid	+	TCG	33	GCTCGAGCTGAAGGAGATATACCATGAGTATGT	*cas1–rbs–cas2*
plasmid	+	AAG	33	GAATGTCATTGCGCTGCCATTCTCCAAATTGCA	
plasmid	-	TTG	33	GTATTTCTCCAGCGGCAAGCACGTCCTCTATAA	*cas1*
plasmid	-	CCG	33	GGAAGCAGTGTGACCGTGTGCTTCTCAAATGCC	
plasmid	-	TGT	33	TTATATCCCGCCGTTAACCACCATCAAACAGGA	*lacI*
plasmid	-	AAG	33	GCATAAACACGAACGCCCGCTTCCCCCACCCAT	*cas1*
plasmid	-	GTG	33	GGAAGCGGCGATGGCGGAGCTGAATTACATTCC	*lacI*
plasmid	+	CAG	33	GCAATGGCATCCTGGTCATCCAGCGGATAGTTA	*lacI*
plasmid	-	ACT	33	TGGAAAGCGGGCAGTGAGCGCAACGCAATTAAT	*lacI*
plasmid	+	TGC	33	CTGAAACCTCAGGCATTTGAGAAGCACACGGTC	
plasmid	-	TTA	33	AATTAAGCTGCGCTAGTAGACGAGTCCATGTGC	
plasmid	-	AAG	33	GATCATGGAGACGCGATCTTTGAGTGGAATGGG	*cas1*
plasmid	+	AAT	33	TGAATCGGCCAACGCGCGGGGAGAGGCGGTTTG	*lacI*
plasmid	-	AAT	33	TGGCAACAGGCTGTGCATCTTCAGGTGGGGCCG	*cas1*
plasmid	+	GCG	33	GGGAAACGGTCTGATAAGAGACACCGGCATACT	*lacI*
plasmid	-	CAC	33	CCCTGGCGCCCAATACGCAAACCGCCTCTCCCC	*lacI*
plasmid	-	GAT	33	TGGCTAATCTGCCTCGTAAGCGCGGAGGTACAT	*cas2*
plasmid	-	CCC	33	CAATACGCAAACCGCCTCTCCCCGCGCGTTGGC	*lacI*
plasmid	-	AAG	34	GTCCTGTCGGGTTTCGCCCGTACTGTCAGATTCA	
plasmid	+	AAG	33	GTCAGCCCCATACGATATAAGTTGTAATTCTCA	
plasmid	+	GGA	33	AGAAATACAACCGCCGGCCCCACCTGAAGATGC	*cas1*
plasmid	+	TAA	33	AACAAATAGCTAGCTCACTCGGTCGCTACGCTC	
plasmid	-	AAG	34	GCGGCGATGGCGGAGCTGAATTACATTCCCAACC	*lacI*
plasmid	+	TGG	33	GCGATGACCTGGCTTCCCCTTAATCCCATTCCA	*cas1*
plasmid	+	GCG	33	GGGAAACGGTCTGATAAGAGACACCGGCATACT	*lacI*
genomic	+	AAG	33	GTCCAGTTGTGCGGGAAGACTTCGAGCATTTGT	*mannonate/altronate dehydratase*
genomic	+	GTT	33	TAATACCGTTGAAATGATGGTCCATATCCATTG	*serine acetyltransferase*
genomic	+	AAT	33	TCTTGATTCCTGAACTGATAGGCTACCTGGCGA	*methyl-directed mismatch repair protein*
genomic	-	GTA	33	ATGCAGGCATGATAGCAAAAATGGCGAGGATGG	*Yick MFS transporter & inhibitor of heme biosynthesis*
genomic	-	AAG	33	GCCATTAATGGCGCAGGCGCTGCATATGGGGGA	*protein of unknown function*
genomic	+	GAT	33	TGGCAGTTTCAATCAGTTCTTGCCGGGTCTTCA	*EnvR DNA binding transcriptional repressor*

Table 2.1 | Sequence of newly acquired spacers. A list of a sample of new spacers acquired into the CRISPR locus of BL21-AI. The PAM is defined as the third base being the first nucleotide of the spacer and the first two nucleotides are the -1 and -2 positions in the DNA source.

With the finding that Cas1 and Cas2 are the only Cas proteins required for spacer acquisition, we tested whether Cas1 and Cas2 form a stable complex *in vivo.* We overexpressed Cas1-FLAG and Cas2-HA fusion protein in BL21-AI cells and conducted immunoprecipitation experiments in cell lysates. Selective elution from FLAG or HA affinity beads with either a 3×-FLAG or an HA peptide resulted in the co-elution of Cas1 and Cas2 (**Fig. 2.1c**). To verify that this interaction is direct, we separately purified the untagged construct of each protein and determined the K_d of the interaction to be ~290 nM, as measured by isothermal titration calorimetry (ITC; **Fig. 2.1d**).

2.2.2 Crystal structure of the Cas1–Cas2 complex

To gain insights into the structural organization of the Cas1–Cas2 complex, we sought to determine the crystal structure of the complex. We purified each protein and reconstituted the complex *in vitro*. Gel filtration chromatography showed the co-migration of both proteins as one peak and SDS-PAGE analysis confirmed the presence of both proteins in the peak fractions (**Fig. 2.2a,b**). The Cas1–Cas2 complex yielded sheet-like crystals that diffracted X-rays to 2.3-Å resolution (**Fig. 2.2c**). We determined the structure by single-wavelength anomalous dispersion (SAD) on selenomethionine-derivatized crystals and refined the resulting model to an R_{work}/R_{free} of 22%/24% (**Table 2.2**).

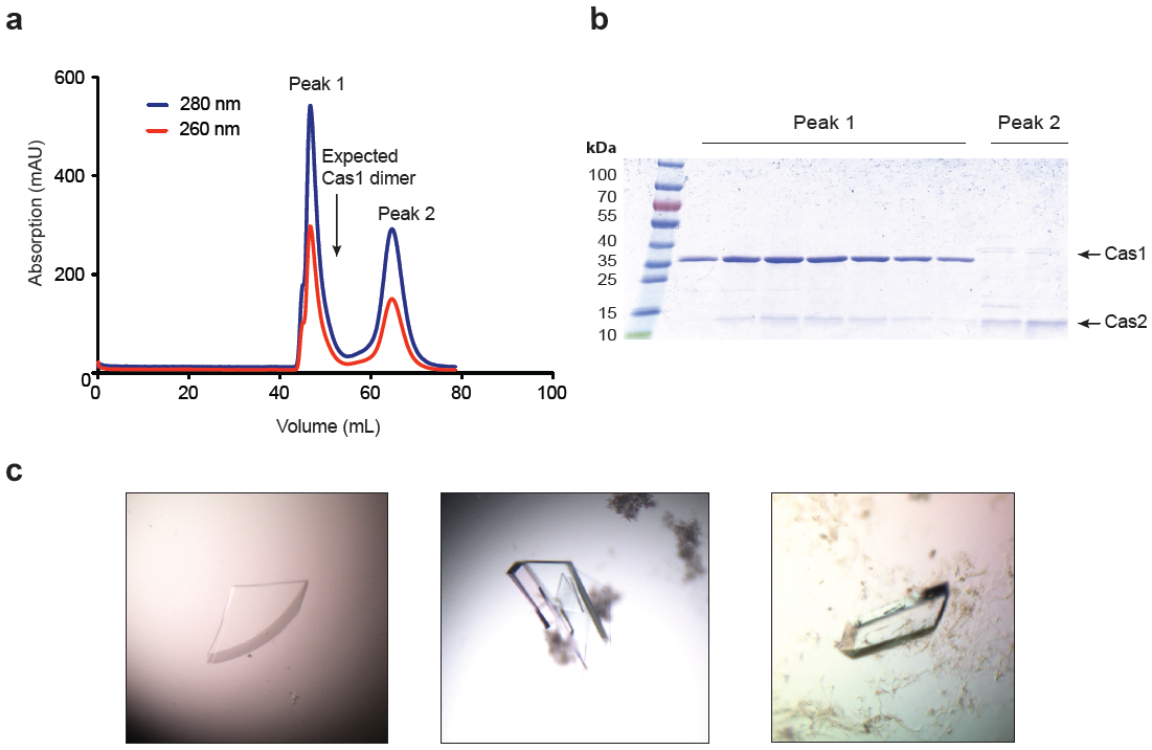


Figure 2.2 | Reconstitution and crystallization of the Cas1–Cas2 complex. (**a**) Gel filtration chromatogram (Superdex 75 16/60) of pre-incubated, separately purified Cas1 and Cas2. The arrow points to the elution peak of Cas1 only. Peak 2 corresponds to Cas2 only. (**b**) Coomassie-stained SDS-PAGE of fractions corresponding to the two peaks shown in (**a**). The arrows point to bands corresponding to Cas1 and Cas2. (**c**) Representative crystals of the Cas1–Cas2 complex.

	Native	Se derivative
Data collection		
Space group	$P2_1$	$P2_1$
Cell dimensions		
a, *b*, *c* (Å)	94.875, 125.70, 99.31	93.70, 127.70, 99.32
α, β, γ (°)	90, 102.74, 90	90, 102.32, 90
Resolution (Å)	62.85–2.3 (2.383–2.3)	48.63–2.89
R_{merge}	0.1301 (1.088)	0.2045 (1.576)
$I / \sigma I$	9.19 (1.00)	10.32 (1.37)
Completeness (%)	97.22 (91.21)	91.87
Redundancy	3.1 (2.9)	7.8 (7.2)
Refinement		
Resolution (Å)	62.85–2.3 (2.383–2.3)	
No. reflections	97,929 (9,271)	
R_{work} / R_{free}	0.225 / 0.245	
No. atoms	10,451	
Protein	9,926	
Water	525	
B factors		
Protein	55.60	
Water	53.90	
r.m.s. deviations		
Bond lengths (Å)	0.003	
Bond angles (°)	0.62	

One crystal was used per data set. Values in parentheses denote highest-resolution shell.

Table 2.2 | X-ray crystallography data collection and refinement statistics.

The overall architecture of the asymmetric unit is a hexameric complex consisting of two Cas1 dimers (Cas1a-b and Cas1c-d) that sandwich one Cas2 dimer (**Fig. 2.3a**). Cas1a and Cas1c make contacts with the Cas2 dimer, whereas Cas1b and Cas1d only make contacts with its dimer pair. The Cas1c-Cas2 protein-protein interface buries a large surface area of ~3,100 Å^2, whereas the Cas1a-Cas2 interface buries an additional 800 Å^2 contributed by the C terminus of Cas1a, as described further below. Similar contacts are present between Cas1a and Cas1c with Cas2 on opposite sides, thus creating a symmetrical complex.

To identify Cas1 and Cas2 conformational changes that occur upon complex formation, we superimposed previously determined structures of apo Cas1 (PDB 3NKD)[23] and Cas2 (PDB 4MAK) from *E. coli* with the Cas1–Cas2 complex structure (**Fig. 2.3b,c**). We observed structural rearrangements in the C-terminal α-helical domains of Cas1b and Cas1d as compared to apo Cas1 (**Fig. 2.3b**); however, it is unclear whether this conformational change results from complex formation or crystallographic packing of another complex next to these subunits. In addition to minor conformational changes present in the canonical βαββαβ ferredoxin fold of Cas2, the C terminus forms two antiparallel β strands (β6-β7) that contact β4 of Cas1 (**Figs 2.3c** and **2.4a**). This region is unresolved in the apo-Cas2 structure, which terminates at the C terminus of β5. Presumably, the β6-β7 region is flexible before complex formation with Cas1.

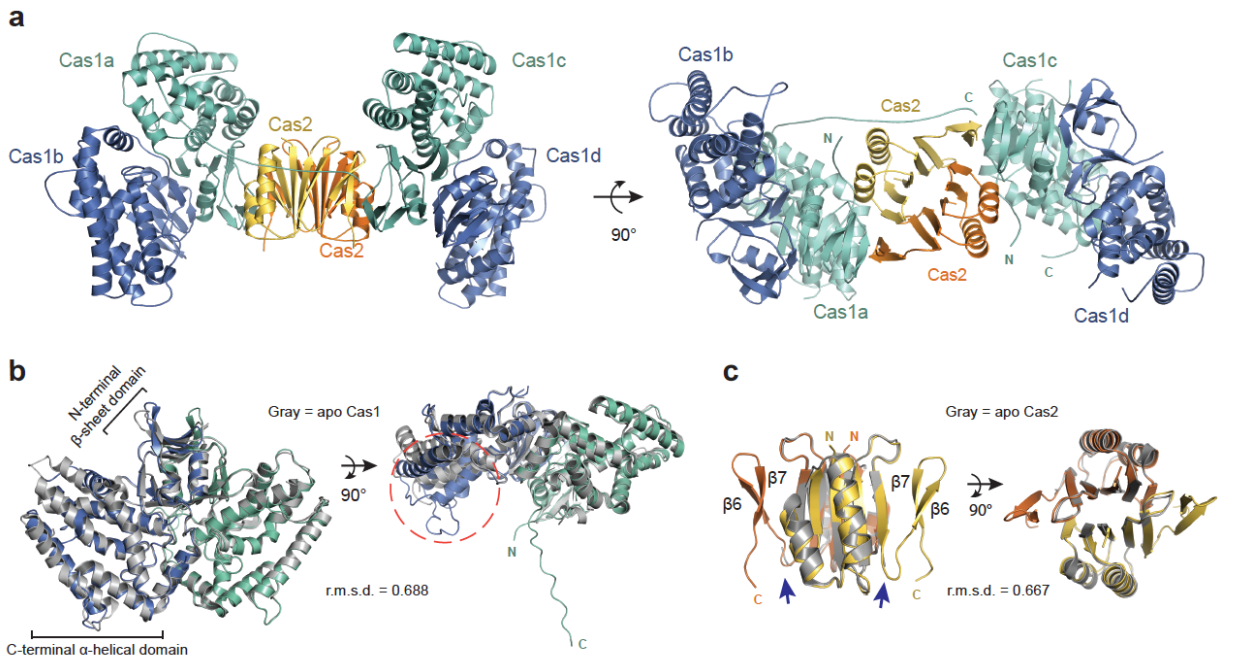


Figure 2.3 | Crystal structure of the Cas1–Cas2 complex. (**a**) Overall structure, consisting of a Cas2 dimer (yellow and orange) and two Cas1 dimers (denoted with suffixes a–d; blue and teal). (**b**) Superposition of the Cas1a–Cas1b dimer with the previously determined *E. coli* Cas1 structure (gray, PDB 3NKD). The dashed orange circle highlights the conformational change observed in the α-helical domain of Cas1a. (**c**) Superposition of the Cas2 dimer in the complex with the previously determined *E. coli* Cas2 structure (gray, PDB 4MAK). The blue arrows point to the last resolved residue in the 4MAK structure. The N and C indicate the termini of each monomer; r.m.s.d. values of the superpositions are indicated.

The interface between Cas1 and Cas2 consists of hydrogen bonding, electrostatic and hydrophobic interactions. We observed extensive electrostatic contacts between three arginine residues (R245, R252 and R256) in α8 of Cas1 with two acidic residues (E65 and D84) of Cas2 (**Fig. 2.4a**). The R252 residue is positioned between E65 and D84 and may sample salt bridge interactions between the two acidic residues, although we only observed continuous density between R252 and E65 at the Cas1a-Cas2 interface. In the same region, backbone hydrogen bond contacts are present between the newly resolved Cas2 β7 C terminus and β4 of Cas1, as discussed further below.

Although Cas1 does not undergo major conformational changes upon Cas2 binding (0.69-Å backbone r.m.s.d.), the proline-rich C-terminal tail of Cas1a is distinctively ordered in only the bound state and is stabilized by hydrophobic and electrostatic contacts (**Fig. 2.4b**). At the middle of the tail, I291 from Cas1 is positioned in a hydrophobic pocket of Cas2 that includes W44 and W60 (**Fig. 2.4b**). The C terminus of Cas1c is likely to span the opposite face of Cas2 to fully encapsulate the dimer, although we did not observe electron density for this tail, owing to crystal packing of an adjacent complex on this face.

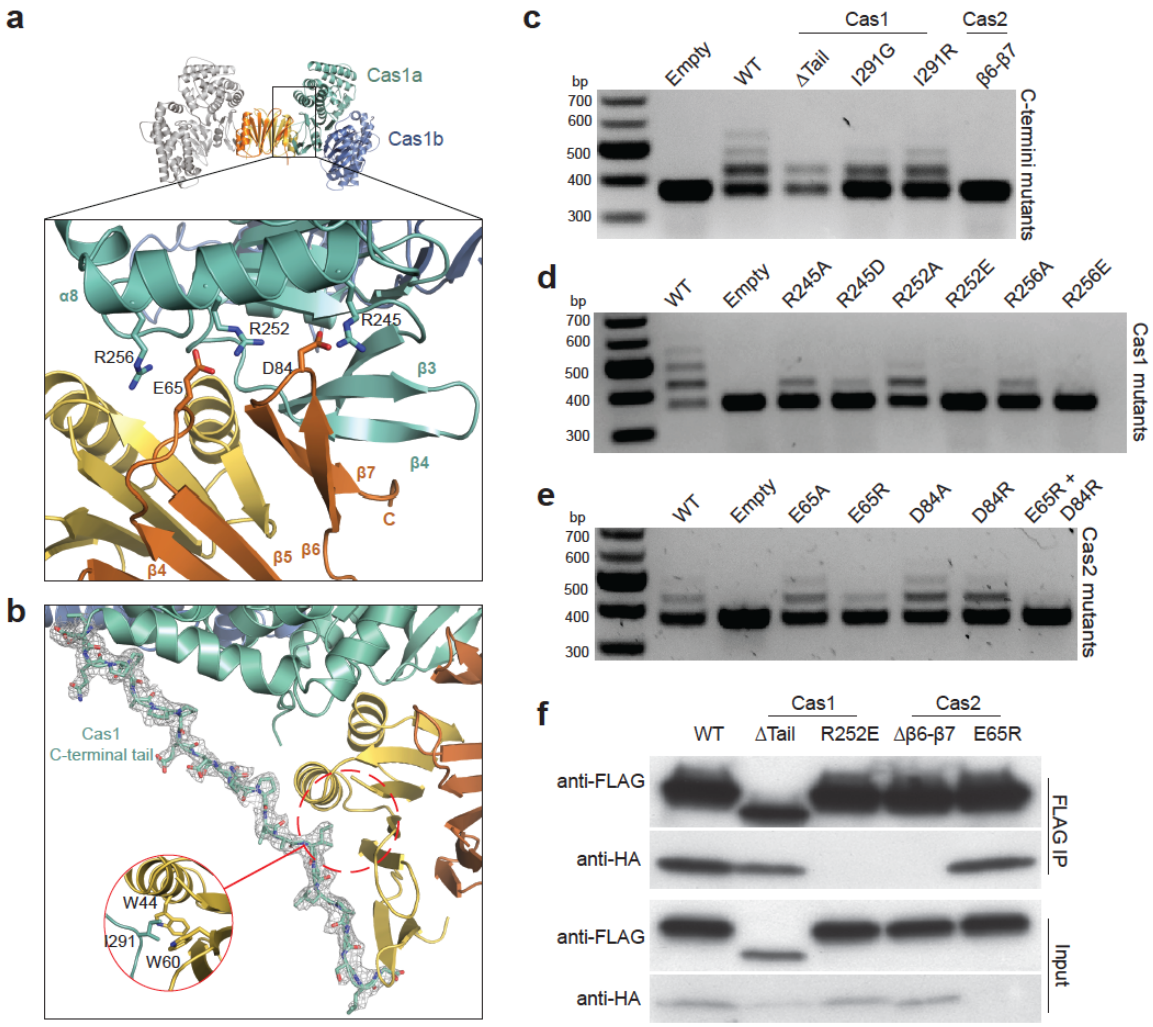


Figure 2.4 | Disruption of complex formation abrogates spacer acquisition. (**a**) A close-up view of the Cas1a–Cas2 protein-protein interface, with annotations for the residues involved in electrostatic interactions. (**b**) View of the ordered C-terminal tail of Cas1a with the electron density mesh contoured at 1.0 sigma. (**c–e**) Agarose gels of *in vivo* acquisition assays with mutations of Cas1 and Cas2 at the C-termini **(c)** and the electrostatic interface (**d,e**). (**f**) Western blot of FLAG immunoprecipitations in BL21-AI cells expressing Cas1-FLAG and Cas2-HA or various mutations of Cas1 and Cas2. Despite the low expression of Cas2 E65R, we still detect its co-elution with Cas1.

2.2.3 Cas1–Cas2 complex formation is required *in vivo*

To determine the function of Cas1–Cas2 complex formation, we conducted spacer acquisition assays in cells expressing Cas1 and Cas2 bearing mutations at the interprotein interfaces. Deletion of the Cas2 C terminus (∆β6-β7) prevented detectable spacer acquisition, suggesting that this structured region is critical for function (**Fig. 2.4c**). In contrast, deletion of the Cas1 tail (∆P282-S305) did not abolish spacer acquisition. Furthermore, mutation of Cas1 I291, which binds in a hydrophobic pocket with W44 and W60 of Cas2, had no effect on spacer acquisition (**Fig. 2.4b,c**). Thus, the C-terminal tail of Cas1 is not essential for spacer acquisition, although it may supplement the critical interactions at the interface described below.

Mutations of residues involved in the electrostatic interactions between subunits have drastic effects on spacer acquisition. Cas1 constructs with mutations at the arginines α8 to alanine (R245A, R252A or R256A) could still acquire spacers. However, constructs with mutations of the same residues to the opposite charge (R245D, R252E or R256E) support little or no detectable spacer acquisition compared to that in wild-type Cas1 (**Fig. 2.4d**). In comparison, single mutations of either of the two acidic Cas2 interface residues E65 or D84 to alanine or arginine had little or no effect on spacer acquisition compared to that in wild-type Cas2. A double mutation of both residues to arginine (E65R and D84R) abolished spacer acquisition *in vivo* (**Fig. 2.4e**). Thus, although mutations in Cas1 at the electrostatic interface are more deleterious than those in Cas2, complementary charges in this interface permit acquisition.

To confirm that the observed *in vivo* effects are due to disruption of Cas1–Cas2 complex formation, we conducted FLAG immunoprecipitation experiments in lysates of BL21-AI cells overexpressing Cas1-FLAG and Cas2-HA mutants. Mutations that had little effect on spacer acquisition (Cas1 ∆tail and Cas2 E65R) did not perturb coprecipitation of Cas1 and Cas2, thus indicating that the mutants are still able to form the complex (**Fig. 2.4f**). In contrast, acquisition-defective mutants (Cas1 R252E and Cas2 ∆β6-β7) could no longer form a stable Cas1–Cas2 complex. This result highlights the importance of the structured Cas2 C terminus and the positioning of R252 between the two Cas2 acidic residues in the electrostatic interaction interface. Together, these findings support the conclusion that Cas1–Cas2 complex formation is required for spacer acquisition *in vivo*.

2.2.4 The catalytic activity of Cas2 is dispensable *in vivo*

Despite the available literature on the biochemical activities of Cas1 and Cas2, the functional roles of these proteins during spacer acquisition are unknown. Cas1 is reported to be a sequence-independent, metal-dependent nuclease that can cleave singe-stranded (ss) DNA, linear and plasmid double-stranded (ds) DNA, ssRNA and various DNA repair intermediates such as Holliday junctions[23,24,63,64]. Three different Cas2 homologs have been found to have metal-dependent nuclease activity with preference for ssRNA[25] or dsDNA[27] or to lack any detectable nuclease activity[26]. The heterohexamic Cas1–Cas2 complex has five potential active sites: one for each Cas1 monomer and one at the Cas2 homodimer interface (**Fig. 2.5a,b**). We conducted spacer

acquisition assays with active site residue mutations in Cas1 and Cas2 to determine whether the nuclease activities of both proteins are required.

Alanine substitution of the conserved Cas1 active site residues abolished spacer acquisition, as previously reported, thus demonstrating the critical role of Cas1 active sites during the CRISPR adaptation stage (**Fig. 2.5c**). Despite the low protein sequence conservation of Cas2 proteins, a conserved acidic residue from each monomer is positioned in the active site to coordinate a metal ion during catalysis in vitro (**Fig. 2.5b**). Surprisingly, Cas2 mutated in the signature catalytic E9 residue to alanine or arginine supported spacer acquisition at frequencies similar to those observed in the presence of wild-type Cas2 (**Fig. 2.5d**). A mutation of this acidic residue has previously been shown to have drastic effects on nucleic acid substrate cleavage *in vitro*[25,27]. Cas2 with an R14A mutation, which has also been shown to be catalytically inactive in the *Sulfolobus solfataricus* Cas2 *in vitro*[25], was still active in acquiring spacers. Of the nearby arginine residues, only the R18A construct had low spacer-acquisition levels. This residue interacts with Cas1 at the interprotein interface, as supported by its continuous density with the Cas1 backbone. These results support the notion that Cas1 is the likely nuclease that catalyzes the integration reaction of spacer substrates, whereas the function of Cas2 during CRISPR–Cas immunity may not be nucleic acid cleavage.

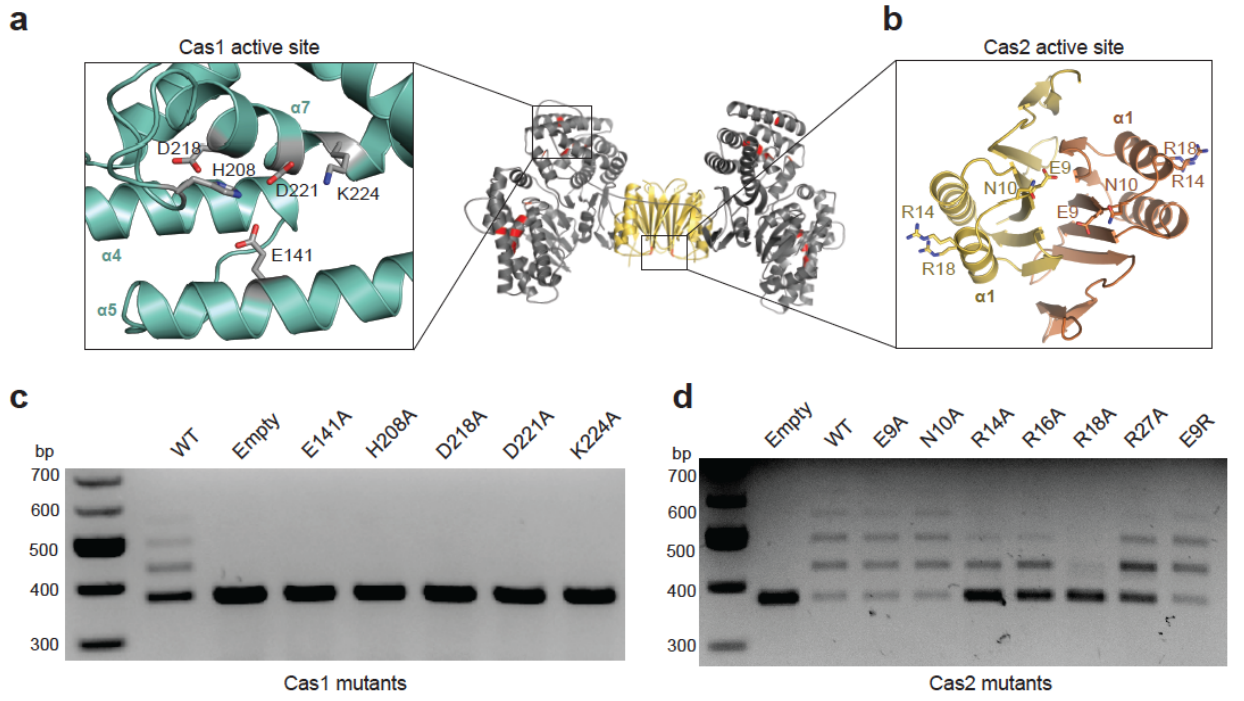


Figure 2.5 | Cas1 is the catalytic subunit within the Cas1–Cas2 complex. (a,b) Close-up views of the Cas1 and Cas2 active sites with stick representations for the conserved residues mutated *in vivo*. In the middle is a general view of the active sites in the complex, highlighted in red. **(c,d)** Acquisition assays of active site residue mutations in Cas1 and Cas2.

2.2.5 Cas1–Cas2 specifically binds the CRISPR locus

The molecular basis for new CRISPR spacer acquisition at the leader-proximal end of the CRISPR locus has been elusive. Although it has been hypothesized that Cas1 or Cas2 might provide such spacer acquisition specificity, previous studies have focused on the individual activities of these two proteins. These studies reported sequence-nonspecific DNA-binding properties of purified Cas1 and Cas2. Our findings that Cas1 and Cas2 form an essential complex that is required for CRISPR spacer acquisition *in vivo* led us to test for CRISPR DNA binding by Cas1 and Cas2. We initially conducted electrophoretic mobility shift assays of purified Cas1 and Cas2, either alone or as a complex, with various DNA substrates. In agreement with previous findings for either protein alone, purified Cas1–Cas2 complex had no sequence-specific DNA binding activity. These results suggest that other host factors may be required to stimulate loading of Cas1 and/or Cas2 on the CRISPR locus.

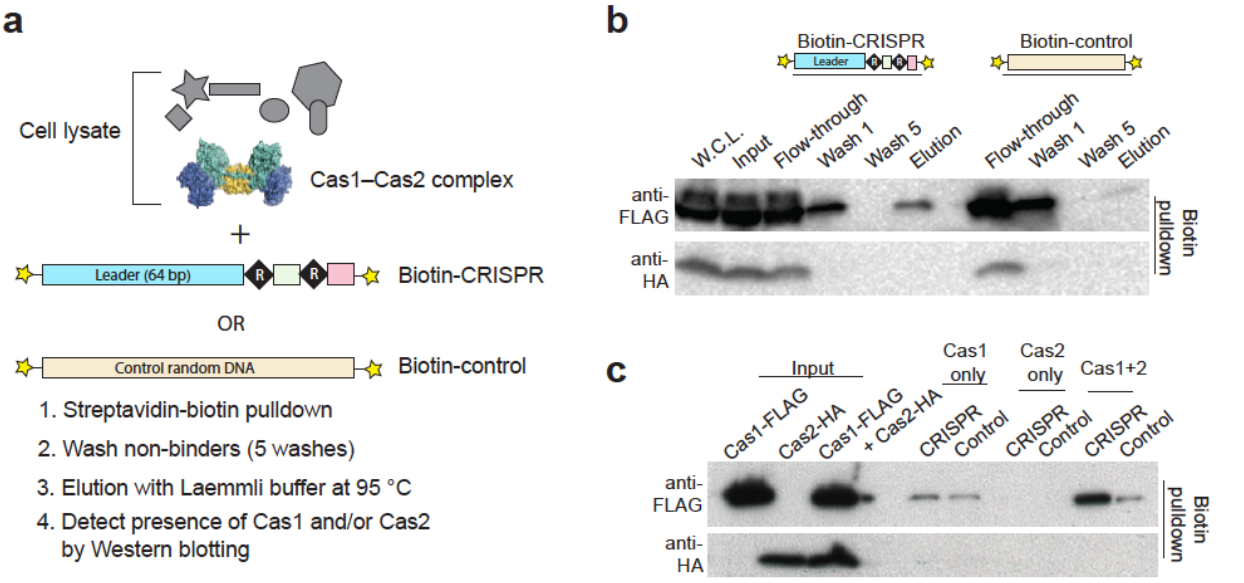


Figure 2.6 | Cas1–Cas2 complex specifically recognizes the CRISPR locus. (**a**) Schematic of the biotinylated DNA affinity precipitation experiments. The stars represent the 5'-biotin labels. (**b**) Western blot of fractions in the biotinylated DNA affinity precipitations. W.C.L. refers to the whole cell lysate. (**c**) Western blot of the DNA affinity precipitations in BL21-AI lysates from overexpression of Cas1-FLAG only, Cas2-HA only or both.

To alternatively probe for CRISPR locus binding specificity, we conducted biotinylated DNA affinity precipitation assays in lysates of BL21-AI cells overexpressing Cas1-FLAG and Cas2-HA (**Fig. 2.6a**). We first tested the ability of Cas1 and/or Cas2 to bind a 186-bp 5′-biotinylated dsDNA containing two CRISPR repeats, two spacers and the minimal 60-bp leader sequence previously shown to be required for spacer acquisition *in vivo*[20]. For a control, we used a DNA of similar length with no CRISPR sequence. After a series of washes to remove nonspecific binders western blot analysis of the elution samples confirmed the preferential binding of Cas1 to the CRISPR DNA

compared to the control DNA (**Fig. 2.6b**). Surprisingly, we did not detect Cas2 in the elution samples; this could be because of the washing conditions removing the weakly bound Cas2. To determine whether Cas2 is required for the preferential binding of Cas1 to the CRISPR DNA, we conducted the affinity precipitation experiment in BL21-AI cell lysates containing overexpressed Cas1 only, Cas2 only or Cas1 and Cas2. Although we did not detect the presence of Cas2 in any of the elution samples, we found that Cas1 lost preference for CRISPR DNA binding in the absence of Cas2 (**Fig. 2.6c**). Thus, Cas1 requires Cas2 for binding to the CRISPR DNA for spacer integration via a yet uncovered mechanism.

2.3 Discussion

The acquisition of new spacer sequences into CRISPR loci during immunological memory formation in bacteria and archaea requires the two conserved CRISPR-associated proteins Cas1 and Cas2. Our findings show that Cas1 and Cas2 assemble into a stable complex whose formation is essential for the incorporation of foreign DNA spacers into the host CRISPR locus *in vivo*. The 2.3-Å crystal structure of the Cas1–Cas2 complex reveals a 2:1 stoichiometry in which a Cas2 dimer binds two Cas1 dimers to form a crablike architecture that specifies the site of integration at the leader end of the CRISPR locus.

Our findings point to likely biochemical functions of Cas1 and Cas2 within the complex. Both proteins have been investigated independently and shown to possess nonspecific nuclease activity *in vitro*[23–25,27,63,64]. On the basis of our active site mutational studies *in vivo* (**Fig. 2.5**), in which catalytically defective Cas1 mutants were incapable of supporting spacer acquisition, Cas1 functions as a *bona fide* nuclease involved in the adaptation stage of CRISPR–Cas immunity. In contrast, the catalytic activity of Cas2 is dispensable for spacer acquisition *in vivo*. Our observation that Cas2 does not coprecipitate with the biotinylated DNA probes suggests that Cas2 may bind weakly to the Cas1–CRISPR DNA complex. Future structural studies of Cas1–Cas2 bound to CRISPR DNA may potentially uncover these interactions. It remains possible that the nuclease activity of Cas2 contributes to a CRISPR-independent process, as suggested by the structural homology between Cas2 and VapDHi toxin of the VapDHI-VapX toxin-antitoxin system in *Haemophilus influenza*[65].

The ability for Cas1 to assemble with Cas2, in addition to its catalytic function, is essential for spacer acquisition. Mutations in either Cas1 or Cas2 that disrupt Cas1–Cas2 complex formation *in vitro* also interfere with spacer acquisition *in vivo*. Furthermore, this functionally critical interaction is conserved across divergent CRISPR systems. Recent experiments have provided evidence for Cas1–Cas2-containing complexes in the Type I-A CRISPR system in the crenarchaeon *Thermoproteus tenax* – in which Cas1 and Cas2 exist as a fusion protein – and in the Type I-F system in the plant pathogen *Pectobacterium atrosepticum*[52,66]. Despite the essential nature of the Cas1–Cas2 interaction in *E. coli*, we note that the observed interprotein interface contacts may not be conserved in other CRISPR–Cas systems. Structural alignment of available Cas1 and Cas2 structures shows poor conservation of the three critical arginines that form salt bridges with E65 and D84 of Cas2 (**Fig. 2.7a,b**). These residues

may vary in other Cas1–Cas2 protein complexes, or they may be replaced by different interactions that ensure Cas1–Cas2 assembly in divergent CRISPR–Cas systems.

Together with previous work, our findings establish that at least two multiprotein complexes are fundamental for a fully functional Type I CRISPR–Cas system: a Cas1–Cas2 spacer acquisition complex and an RNA-guided DNA interference complex. Whether these complexes interact to form a multifunctional supercomplex is not yet known. However, an interesting hint about this possibility comes from *P. atrosepticum*, in which Cas2 exists as an N-terminal fusion with Cas3, the foreign DNA-targeting nuclease recruited by the DNA interference Cascade complex. The Cas2-Cas3 fusion protein in this organism has also been shown to associate with Cas1 (ref. 52). It is thus possible that at least some Type I CRISPR–Cas systems use the Cas1–Cas2 complex not only for new spacer acquisition but also for coupling this process to target recognition and destruction.

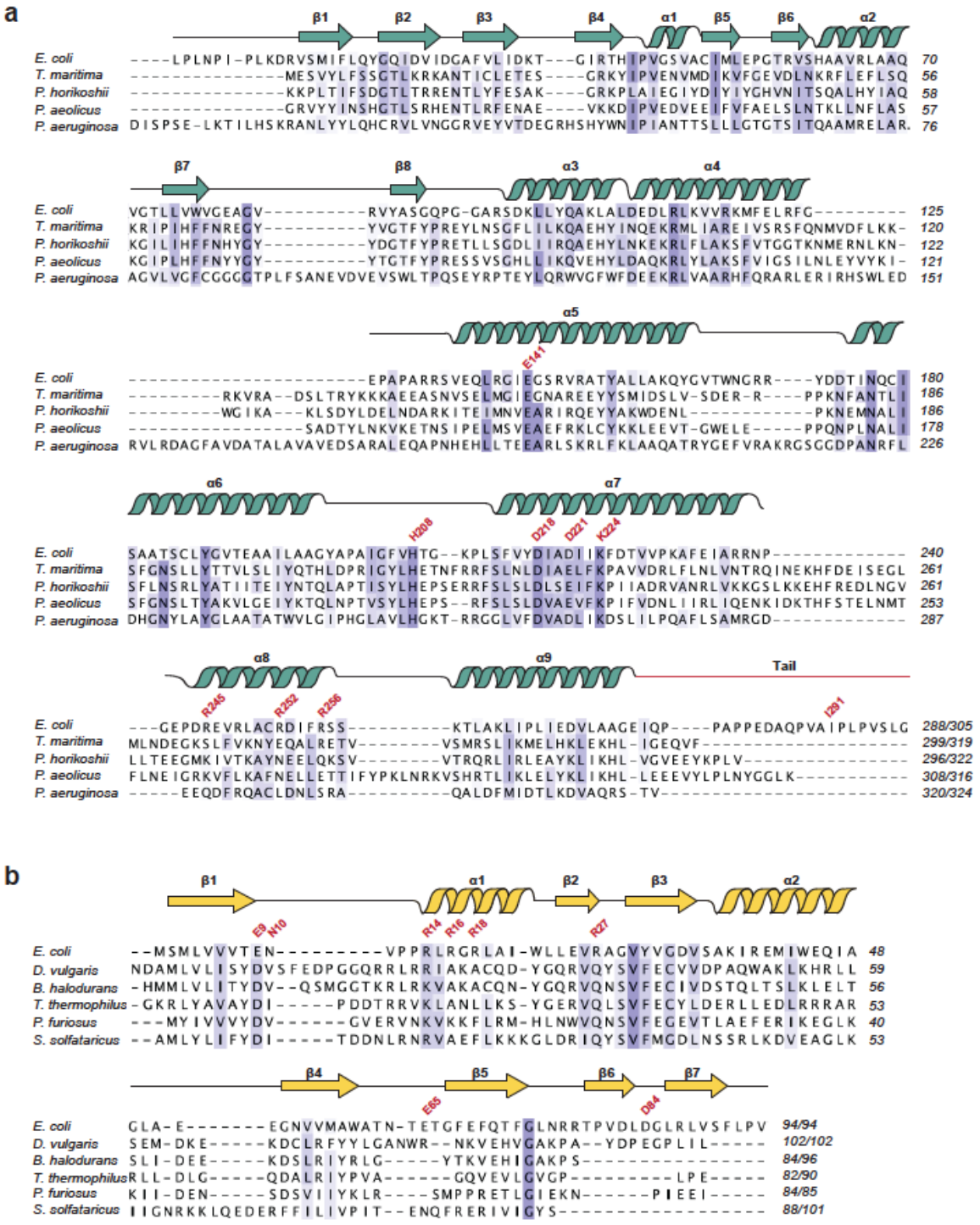


Figure 2.7 | Structure-based Cas1 and Cas2 alignments. The *E. coli* Cas1 (**a**) and (**b**) protein was aligned to its respective homologs for which crystal structures are available, as described in the Methods section. The number annotations at the C-termini refer to the last amino acid residue resolved in the structure, followed by the last residue of the full-length protein. The secondary structure cartoon is for the *E. coli* protein and the annotated residues (red) refer to the ones that were mutated in this study. The BLOSUM62 score conservation threshold is set to 50%, as reflected by the blue colors of the alignment.

2.4 Materials and Methods

2.4.1 Protein purification

The *cas1* and *cas2* genes were PCR amplified from *E. coli* K12 (MG1655) genomic DNA. The ca*s1* gene was cloned into a Gateway compatible expression vector (pHMGWA) containing an N-terminal His_{6x}MBP tag and *cas2* was cloned into pET16b (Novagen) containing a C-terminal MBPHis_{6x} tag. The proteins were purified separately using the same protocol. The constructs were expressed in BL21(DE3) cells, grown to 0.6-0.8 OD_{600}, and induced overnight at 16 °C with 0.5 mM isopropyl-β-D-thiogalactopyranoside (IPTG). The cells were harvested and re-suspended in buffer A (500 mM KCl, 20 mM HEPES-KOH, pH 7.4, 10 mM imidazole, 0.1% Triton X-100, 2 mM TCEP, 0.5 mM phenylmethylsulfonyl fluoride [PMSF], "Complete, EDTA-free" protease inhibitor [Roche] and 10% glycerol). After lysis by sonication, the lysates were cleared by centrifugation and incubated with Ni-NTA affinity resin in batch (QIAGEN). The resin was washed in buffer B (with 500 mM KCl, 20 mM HEPES-KOH, pH 7.4, 10 mM imidazole, 5% glycerol, 1 mM TCEP) and the protein was eluted with buffer B supplemented with 300 mM imidazole. The eluted protein was dialyzed against buffer B at 4 °C in the presence of TEV protease to remove the affinity tags. The protein was concentrated and further purified on tandem MBPTrap HP (GE Healthcare) and Superdex 75 (16/60) size exclusion chromatography column using buffer B in the absence of imidazole. The selenomethionine proteins were overexpressed in minimal media as previously described[67] and subsequently purified using the same purification protocol for the native proteins.

2.4.2 *In vitro* Cas1–Cas2 complex formation

Purified Cas1 and Cas2 were separately dialyzed against 150 mM KCl, 20 mM HEPES-KOH, pH 7.4, 5% glycerol and 1 mM TCEP at 4 °C overnight. The proteins were incubated together at a 1:3 Cas1:Cas2 molar ratio for one hour on ice. The sample was loaded on a Superdex 75 (16/60) size exclusion column and the peak fractions corresponding to the complex were pooled and concentrated for crystallization (**Fig. 2.2a,b**). We note that the molar ratio of Cas1:Cas2 for pre-incubation was chosen to obtain clear separation between the Cas1–Cas2 complex and Cas1 only peaks on gel filtration. There is no difference in the retention time of the complex when the proteins were pre-incubated at a 1:1 ratio.

2.4.3 Crystallization and structure determination

The selenomethionine-derivatized complex was concentrated to ~4 mg mL^{-1} and crystallized by hanging drop vapor diffusion at room temperature in 120 mM calcium acetate, 10% (w/v) PEG 8000 and 50 mM sodium cacodylate, pH 6.50. The crystals were briefly transferred into a drop containing 20% glycerol for cryoprotection and frozen in liquid nitrogen until data collection. The native protein was crystallized in 150 mM NaCl, 6% (w/v) PEG 8000 and 100 mM Tris, pH 8.0. The crystals were frozen in the presence of 20% PEG 8000 as cryoprotectant.

The diffraction data were collected under cryogenic conditions at the Lawrence Berkeley National Laboratory Advanced Light Source (beamline 8.3.1). The selenomethionine-derivative crystals were scanned for fluorescence to determine the selenium absorption edge and diffraction data was collected at the peak wavelength. The data were processed with XDS[68] and SCALA[69]. The crystals belonged to a monoclinic space group ($P\ 2_1$) with four copies of Cas1 and two copies of Cas2 in the asymmetric unit. Autosol was used for experimental phasing and AutoBuild within PHENIX[70] to obtain an initial model. Iterative rounds of model building in Coot[71] and refinement using PHENIX was performed to obtain a selenomethionine-derivative model. The 2.3-Å native model was built using Phaser–Molecular Replacement in PHENIX for phasing, followed by Coot and PHENIX for model building and refinement, respectively.

2.4.4 Isothermal titration calorimetry

ITC experiments were conducted on a MicroCal Auto-iTC200 system (GE Healthcare). Purified Cas1 and Cas2 were separately dialyzed at 4 °C against 200 mM KCl, 20 mM HEPES-KOH, pH 7.5, 5% glycerol and 1 mM TCEP. Cas1 (150 μM) was titrated into the cell containing 15 μM Cas2 with 10-15 3.1 μl injections at 4 °C. The Origin software (OriginLab) was used for baseline correction, integration and curve fitting. The K_d reported is an average of three separate experiments (308, 212 and 351 nM, standard deviation of 58 nM). The calculated N values of each independent run were 1.52, 1.51 and 1.50.

2.4.5 Spacer acquisition assays

The assays were conducted as previously described[5] with slight modifications. Briefly, *cas1* and *cas2* were both cloned into pCDF-1b (Novagen) and transformed into *E. coli* BL21-AI (Invitrogen). After preparation of an overnight culture in LB medium containing 50 μg ml^{-1} streptomycin, a sample was transferred (1:300) into a 10 ml culture containing 0.2% L-arabinose, 0.1 mM IPTG and 50 μg ml^{-1} streptomycin to induce protein expression. After 20–24 hours, a sample of the culture was diluted in water, boiled at 95 °C for 5 min and centrifuged (16,100 × g). A sample of the supernatant was used as template for PCR amplification of the CRISPR locus using the same primers as previously described[5]. The PCR reactions were analyzed on 1.5% agarose gels. For comparison in spacer acquisition efficacy of mutant proteins, the OD_{600} of each culture was measured and the amount of culture obtained for PCR amplification was normalized accordingly. The newly acquired spacers reported in **Table 2.1** were obtained by plating a sample of the culture on LB agar plates and amplifying the CRISPR-I locus of single clones to detect locus expansion. The PCR products of clones with expanded loci were submitted for sequencing. The gene annotations were obtained from the NCBI Basic Local Alignment Search Tool (BLAST)[72] using the *Escherichia coli* BL21 (taxid:469008) genomic sequence. All of the acquisition assays reported in this study have been replicated at least three times.

2.4.6 Immunoprecipitation assays

The Cas1-FLAG and Cas2-HA constructs were both cloned into pCDF-1b and co-expressed in BL21-AI cells for 20–24 hours in spacer acquisition-inducing conditions, as described above. The amount of cells used was normalized to the OD_{600} measurements of the cultures. The cells were pelleted and re-suspended in lysis buffer (150 mM KCl, 50 mM Tris, pH 7.50, 1 mM TCEP, 1% Triton X-100, 0.5 mM PMSF and protease inhibitors). After sonication on ice, the lysates were cleared and rocked for 1.5–2 hours at 4 °C with either anti-FLAG M2 or anti-HA affinity resin (Sigma-Aldrich). The resin was washed five times with 400 mM KCl, 50 mM Tris, pH 7.50, 1 mM TCEP and 1% Triton X-100. The proteins were eluted with either 100 ng μl^{-1} 3×-FLAG peptide (DYKDDDDK) or HA peptide (YPYDVPDYA), synthesized by David King (HHMI, UC Berkeley). The epitope-tagged proteins were detected with monoclonal anti-FLAG M2-peroxidase (HRP) mouse antibody (Sigma-Aldrich A8592, 1:22,000) or HRP-conjugated anti-HA mouse antibody (Cell Signaling 2999S, 1:10,000). All of the FLAG immunoprecipitation experiments reported in this study have been replicated at least three times.

2.4.7 DNA affinity precipitation assays

The 186-bp CRISPR DNA bait was generated by PCR amplification of the BL21-AI CRISPR-I locus using 5' biotin-conjugated forward and reverse primers (synthesized by Integrated Device Technology). The 186-bp control DNA was PCR amplified from the *ori* sequence of the pUC19 vector. The input lysates were prepared as described above for IP assays. The amount of biotinylated DNA probe was normalized to 100 nM and rocked with the lysate at 4 °C for 30 min. Avidin agarose (Pierce, **Fig. 2.6e**) was added to the reaction and rocked for an additional 1.5-2 hours. The samples were washed five times with lysis buffer and the proteins were eluted with Laemmli buffer by boiling at 95 °C for 5 minutes. Western blotting was conducted to detect Cas1-FLAG and Cas2-HA in the samples as described in the immunoprecipitation procedure.

2.4.8 Structure-based sequence alignments

The amino acid sequences of the Cas1 and Cas2 proteins were obtained from the RCSB Protein Data Bank. The alignment was generated by PROMALS3D[73] and the output was analyzed on Jalview[74]. The BLOSUM62 score threshold on Jalview was set to 50% to generate the conservation colors. The PDB IDs of the Cas1 structures are: 3LFX (*Thermotoga maritima*), 3PV9 (*Pyrococcus horikoshii*), 2YZS (*Aquifex aeolicus*) and 3GOD (*Pseudomonas aeruginosa*). The PDB IDs of the Cas2 structures are: 3OQ2 (*Desulfovibrio vulgaris*), 4ES2 (*Bacillus halodurans*), 1ZPW (*Thermus thermophilus*), 2I0X (*Pyrococcus furious*) and 2I8E (*Sulfolobus solfataricus*).

2.5 Accession Code

Coordinates and structure factors for the Cas1–Cas2 complex have been deposited in the Protein Data Bank under accession code 4P6I.

2.6 Acknowledgements

We are grateful to Doudna lab members for providing input on this work. We thank Christopher Davies for performing and analyzing analytical ultracentrifugation experiments, Stephen Floor, Amy Lee, Ho Young Lee, Ross Wilson, Ruiyu Wu and Kaihong Zhou for technical assistance, and the 8.3.1 beamline staff at the Advanced Light Source and Anthony Iavarone (UC Berkeley) for mass spectrometry. We acknowledge David King (UC Berkeley, HHMI) for synthesizing FLAG and HA peptides. The 8.3.1 beamline is supported by UC Office of the President, Multicampus Research Programs and Initiatives grant MR-15-328599 and Program for Breakthrough Biomedical Research, which is partially funded by the Sandler Foundation. This project was funded by a US National Science Foundation grant to J.A.D (No. 1244557). J.K.N. and A.V.W. are supported by National Science Foundation Graduate Research Fellowships and J.K.N. by a UC Berkeley Chancellor's Fellowship. P.J.K. is a Howard Hughes Medical Institute Fellow of the Life Sciences Research Foundation. J.N. is supported by a Long-Term Postdoctoral Fellowship from the Human Frontier Science Program Organization. J.A.D. is an Investigator of the Howard Hughes Medical Institute.

2.7 Author Contributions

These authors contributed to the data presented in this chapter: James K. Nuñez, Philip J. Kranzusch, Jonas Noeske, Addison V. Wright and Jennifer A. Doudna. J.K.N. performed the protein purification, biochemical assays and crystallization experiments. X-ray diffraction data were collected by J.K.N., P.J.K. and J.N., and structure determination was performed by J.K.N and P.J.K. A.V.W. assisted with *in vivo* acquisition and immunoprecipitation assays. J.K.N. and J.A.D. designed the study and wrote the manuscript.

CHAPTER 3

Mechanism of foreign DNA integration into CRISPR loci

A portion of the work presented in this chapter has been published previously as part of the following paper: Nuñez, J.K., Lee, A.S.Y., Engelman, A.N. & Doudna, J.A. Integrase-mediated spacer acquisition during CRISPR–Cas adaptive immunity. *Nature* **519**, 193-198 (2015).

3.1 Introduction

Prokaryotic adaptive immunity relies on clustered regularly interspaced short palindromic repeats (CRISPRs) together with CRISPR associated (Cas) proteins to detect and destroy foreign nucleic acids[3,4]. CRISPR loci contain an AT-rich leader sequence followed by repetitive sequence elements flanking ~30 base pair (bp) spacer segments that are transcribed to produce precursor CRISPR RNAs (pre-crRNAs)[7–9,14]. Spacers are frequently virus- or plasmid-derived, although 'self'-derived spacers from the host chromosome are present in some CRISPR loci[32,75]. After pre-crRNA processing and assembly with Cas proteins, the resulting RNA-guided surveillance complexes target and cleave foreign nucleic acids bearing sequence complementarity to the crRNA spacer sequence[10–12,14–16,76]. How spacer DNA sequences, termed protospacers, are captured and acquired into the host CRISPR locus remains unknown.

Overexpression of Cas1 and Cas1 in *Escherichia coli* leads to the site-specific acquisition of 33 bp protospacers at the leader end of the CRISPR locus[20–22]. Cas1 and Cas2 are the only universally conserved Cas proteins in CRISPR–Cas systems and function as a protein complex *in vivo*[35], suggesting that the Cas1–Cas2 complex might possess DNA recombination activity. Here we show that the purified Cas1–Cas2 complex integrates oligonucleotide DNA substrates into acceptor DNA to yield products similar to those generated by retroviral integrases and transposases. Cas1 is the catalytic subunit and Cas2 substantially increases integration activity. Protospacer DNA with free 3′–OH ends and supercoiled target DNA are required, and integration occurs preferentially at the ends of CRISPR repeats and at sequences adjacent to cruciform structures abutting AT-rich regions, similar to the CRISPR leader sequence. Our results demonstrate the Cas1–Cas2 complex to be the minimal machinery that catalyzes spacer DNA acquisition and explain the significance of CRISPR repeats in providing sequence and structural specificity for Cas1–Cas2-mediated adaptive immunity.

3.2 Results

3.2.1 Protospacer DNA integration by Cas1–Cas2

To test whether the Cas1–Cas2 complex is sufficient to catalyze DNA recombination *in vitro*, assays were conducted using purified Cas1–Cas2 complex, 33 bp protospacer DNA and an acceptor 'target' plasmid consisting of the pUC19 backbone with an inserted CRISPR locus (pCRISPR; **Fig. 3.1a**). Co-incubation of these reagents converted the supercoiled plasmid into three main products: relaxed and linear plasmid species and a fast-migrating species we term Band X (**Fig. 3.1b,c**). Product formation required Cas1, Cas2 and the protospacer DNA (**Fig. 3.1d,e**), and was consistent with previous divalent metal ion-dependent and sequence-nonspecific *in vitro* activity requirements of Cas1 and Cas2. Product DNA migration was not affected by treatment with EDTA, EDTA and phenol–chloroform or proteinase K in the presence of EDTA and detergent, indicating that product DNAs are unlikely to be bound to Cas1 and/or Cas2. Consistent with product DNA resulting from covalent integration of protospacer DNA into the plasmid, the relaxed and linear forms of pCRISPR

radiolabeled in reactions containing ^{32}P-labeled protospacer DNA (**Fig 3.2a,b**). Although Cas1 alone catalyzed a low level of protospacer integration in the presence of Mn^{2+}, the reaction was enhanced substantially by the presence of Cas2 (**Fig. 3.2c**).

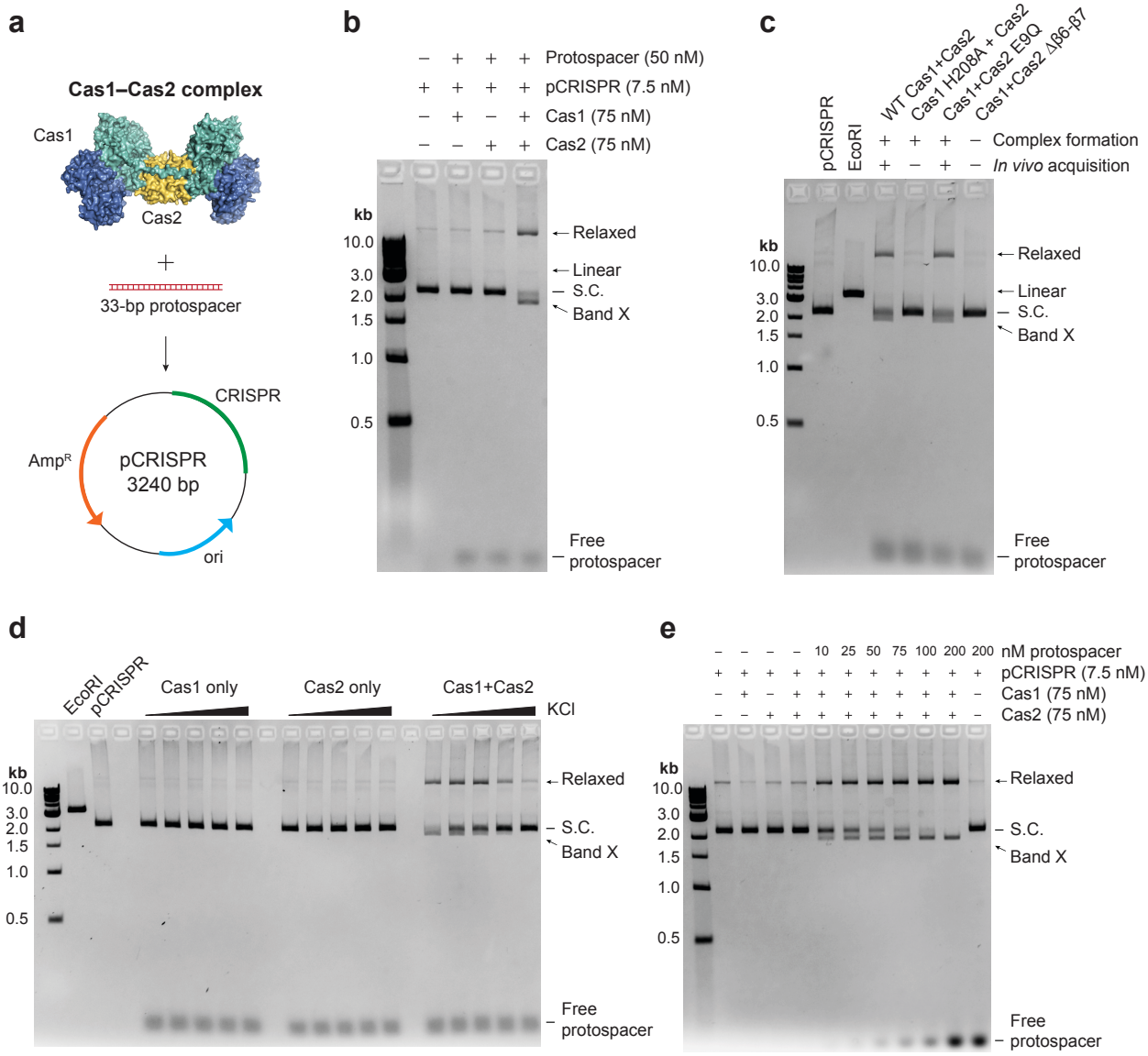


Figure 3.1 | The Cas1–Cas2 complex integrates protospacers *in vitro*. (**a**) Schematic of the *in vitro* integration assay (PDB code 4P6I for Cas1–Cas2). (**b**) The presence of Cas1, Cas2 and a protospacer results in the conversion of the supercoiled pCRISPR into relaxed, linear and Band X products. (**c**) Neither the Cas1 H208A active site mutant nor the complex formation-defective Cas2 Δβ6-β7 deletion mutant support the reaction. The Cas2 E9Q active mutant is as active as the wild type. (**d**) Salt-dependence assay using Cas1 or Cas2 only and Cas1+Cas2. The titration corresponds to 0, 25, 50, 100 and 200 mM KCl, in addition to the salt carried in from the reaction reagents. (**e**) Integration assays with increasing protospacer concentrations.

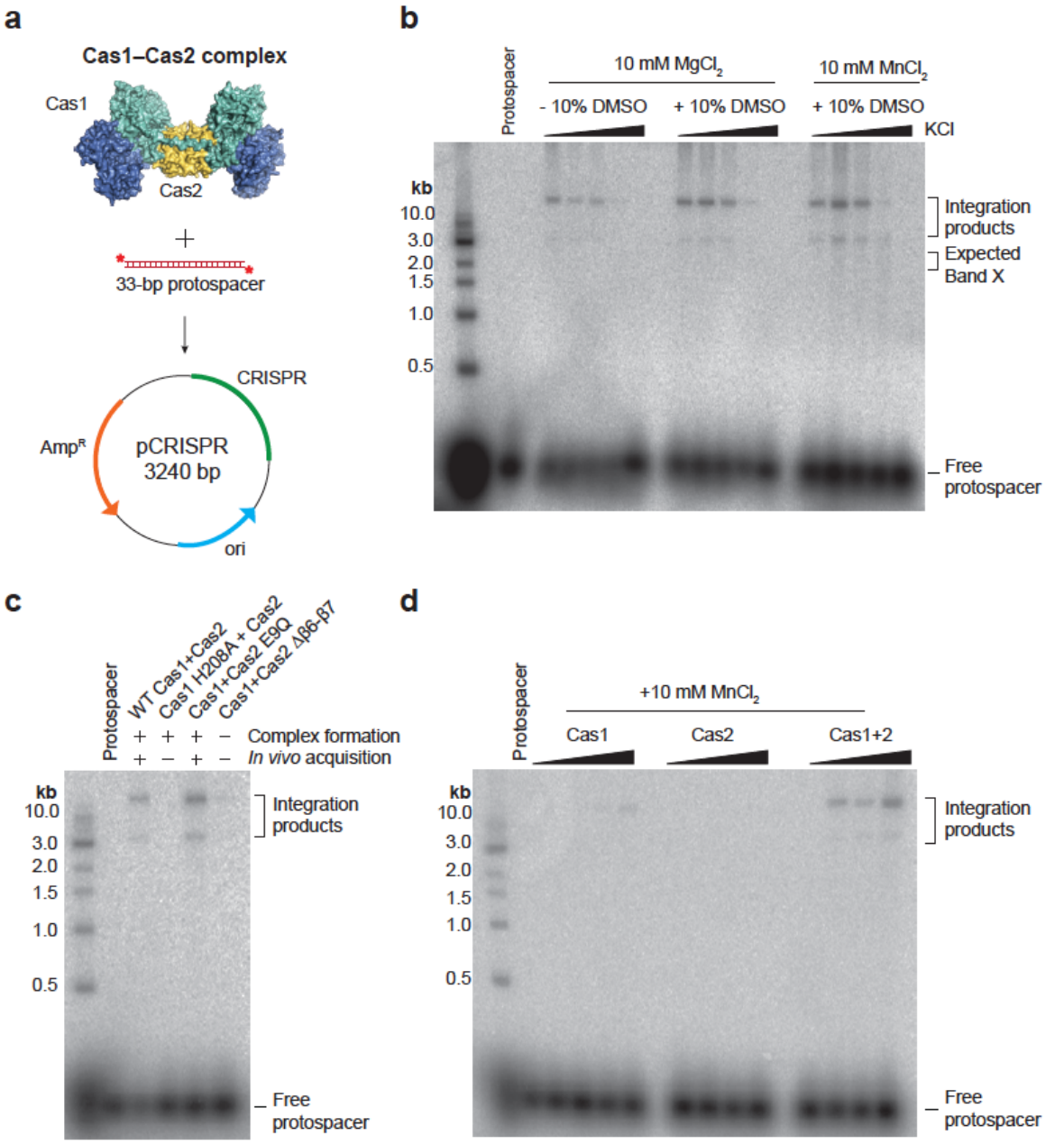


Figure 3.2 | Radiolabeled protospacers are integrated into pCRISPR. (**a**) Schematic of the *in vitro* integration assay using ^{32}P-radiolabeled protospacers. The asterisks denote radiolabels at the 5′ DNA ends. (**b**) Salt- and metal-dependence of radiolabeled protospacer integration into pCRISPR. (**c**) The Cas1 H208A active site mutant and complex-defective Cas2 Δβ6-β7 deletion are defective in integration, whereas the Cas2 E9Q mutant is still active. These results are consistent with *in vivo* spacer acquisition assays. (**d**) Integration assays with only Cas1 results in low levels of integration products in the presence of $MnCl_2$ and product formation is stimulated in the presence of Cas2.

Bacteria expressing Cas1 active site mutants, but not Cas2 active site mutants, are incapable of acquiring new spacers *in vivo*, demonstrating the catalytic role of Cas1 during spacer acquisition[20,21,35]. Consistent with these data, Cas1 active site mutants H208A and D221A were defective for protospacer integration *in vitro*, whereas the Cas2 E9Q active site mutant supported integration. The Cas2 C-terminal (Δβ6-β7) deletion mutant, which is defective for complex formation with Cas1 and spacer acquisition *in vivo*, failed to support Cas1-mediated integrase activity (**Fig. 3.1c** and **Fig. 3.2d**). We conclude that our *in vitro* assay recapitulates the *in vivo* functions of Cas1 and Cas2 during spacer acquisition.

3.2.2 Integration and disintegration products

We next tested whether the reaction products of Cas1–Cas2-mediated DNA integration resemble those formed by the strand transfer activity of retroviral integrases and cut-and-paste transposases[77–80]. These enzymes generate two main products *in vitro* corresponding to half-site and full-site integration events (**Fig. 3.3a**). We observed similar gel mobility of the slowly migrating DNA product generated by Cas1–Cas2 and Nb.BbvCI nickase-digested pCRISPR, consistent with the slow-migrating relaxed DNA species corresponding to half-site products and/or products resulting from full-site integration of one protospacer molecule. Digestion with EcoRI, which cuts pCRISPR once, converted the reaction products to linear DNAs, (**Fig. 3.3b**, compare lane 4 to lane 2, and **Fig. 3.3c**). We therefore conclude that both the relaxed and Band X DNA products comprise unit-sized pCRISPR circles. Thus, the 3.2 kb pCRISPR is not digested to shorter fragments during the reaction.

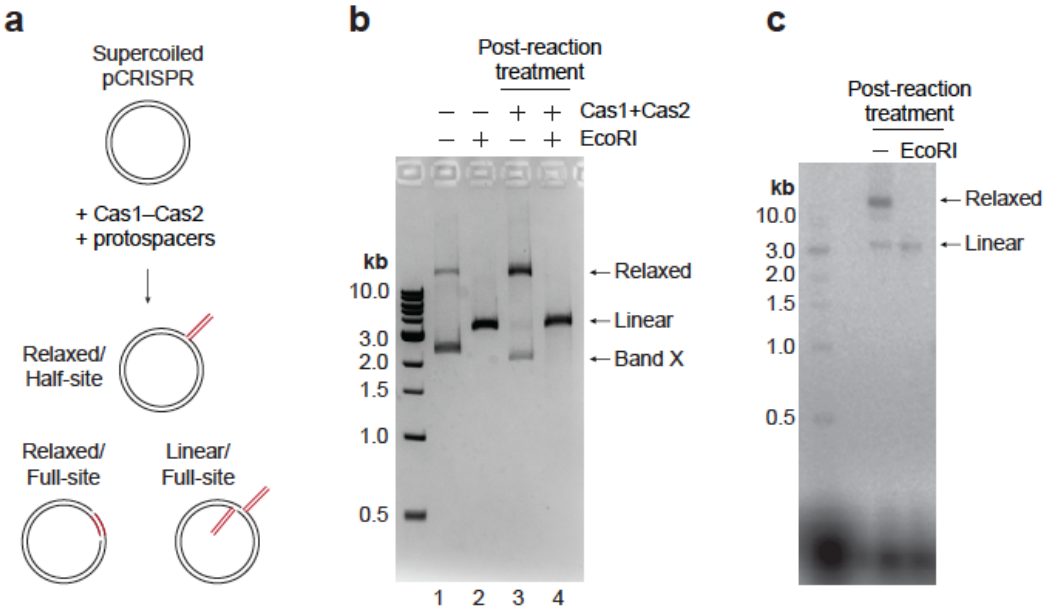


Figure 3.3 | Assignment of Cas1–Cas2-catalyzed integration products. (**a**) Schematic of half-site and full-site integration products. (**b**) Linearization of the integration products (lane 4). Lane 3 is the untreated reaction products. (**c**) Linearization of integration products from radiolabeled protospacer reactions.

We observed that Band X did not become radiolabeled in reactions conducted with ^{32}P-labeled protospacer DNA. A time-course analysis revealed relaxed DNA product formation within the first minute, followed by accumulation of Band X between 10 and 30 min (**Fig. 3.4a**). To determine the properties of Band X, the purified product was analyzed in two different types of agarose gels – one pre-stained with ethidium bromide, similar to the gels presented thus far, and the other stained with ethidium bromide after electrophoresis (post-stained). Although Band X migrated as a single species in the pre-stained gel, a ladder of species that migrated faster than the relaxed products was observed in the post-stained gel (**Fig. 3.4b,c**). These intermediates are reminiscent of plasmid topoisomers[81,82]. The same pre- and post-stained agarose gel analysis was performed on the entire integration reaction, generating similar results to those observed with purified Band X (**Fig. 3.4d,e**).

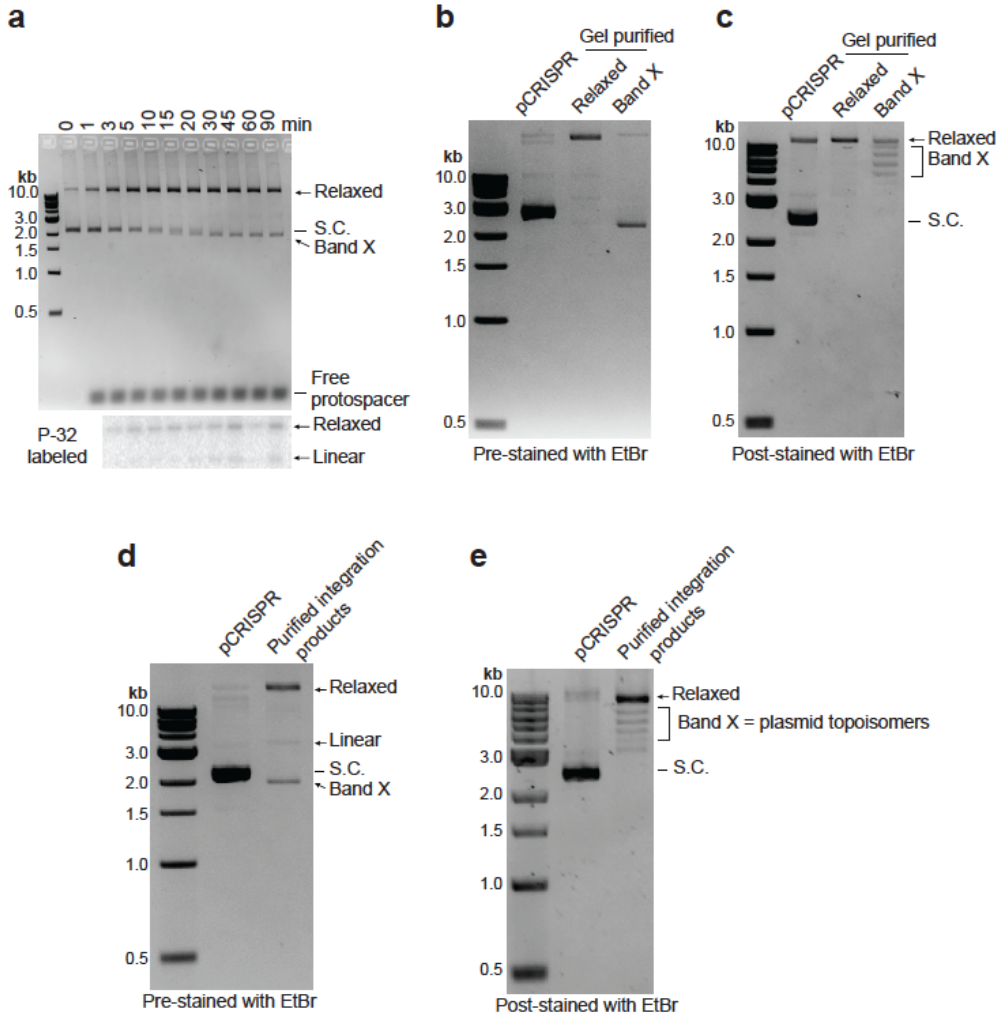


Figure 3.4 | Relaxed plasmid topoisomers byproducts. (**a**) Time course experiments show the initial formation of relaxed products, followed by Band X formation. (**b,c**) Analysis of gel-purified relaxed and Band X on agarose gels pre-stained with ethidium bromide (**b**) or post-stained after electrophoresis (**c**). (**d,e**) Analysis of the entire integration reactions on pre-stained or post-stained agarose gels with ethidium bromide.

We wondered whether Band X arose from protospacer excision from half-site integration products to regenerate pCRISPR in different supercoiled states, analogous to the *in vitro* reversal of integration activity of retroviral integrases and transposases (termed disintegration, step 2 in **Fig. 3.5a**)[83,84]. To test this hypothesis, a synthetic Y-structured DNA intermediate that mimics the half-site integration product into the leader-repeat border of the CRISPR locus was radiolabeled such that the liberated 33 bp protospacer DNA could be detected following disintegration activity (**Fig. 3.5b**). Using this substrate, we observed that Cas1 catalyzed disintegration activity either by itself or in the presence of Cas2 (**Fig. 3.5c,d**). Disintegration activity was confirmed by radiolabeling the 20 nucleotide (nt) target DNA strand containing the leader sequence and monitored the formation of the joined 40 bp leader-repeat DNA (**Fig. 3.5e-g**). Thus, Cas1–Cas2 integration and disintegration activities are similar to those of retroviral integrases and transposases, with Cas1 functioning in integration and disintegration reactions. Disintegration activity of Cas1 has been further studied in a separate study[61].

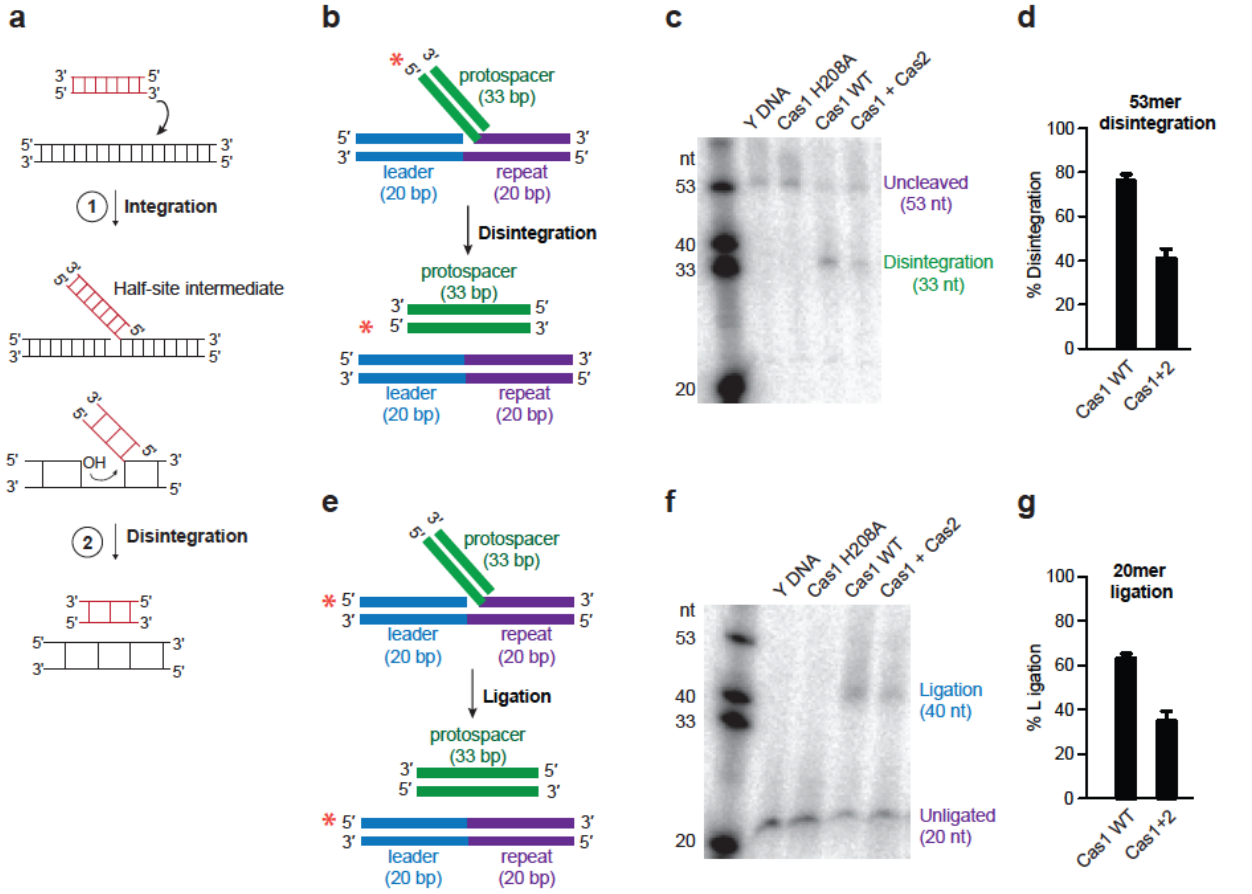


Figure 3.5 | Cas1 catalyzes disintegration reactions. (**a**) Schematic of the proposed integration (step 1) and disintegration (step 2) reactions. (**b**) Diagram of the Y-DNA used to monitor the disintegration reaction. The 5′ end of the protospacer (green) is radiolabeled. (**c,d**) Disintegration of the protospacer as detected by the conversion of the 53 nt substrate to 33 nt. (**e-g**) Similar to **b-d** except the 20 nt leader fragment is radiolabeled and its ligation to the 20 nt repeat DNA is monitored.

3.2.3 Integration requires 3′–OH protospacer ends

We next investigated the DNA protospacer and target DNA requirements for integration. Single-stranded protospacer DNA failed to support the reaction (**Fig. 3.6a**). The Cas1–Cas2 complex accommodated various protospacer lengths *in vitro* despite the strict 33 bp requirement for spacer acquisition *in vivo* (**Fig. 3.6c**), suggesting that protospacer length is pre-determined before integration *in vivo* by an unknown mechanism. The Cas1–Cas2 complex integrated DNA substrates with blunt ends or with 3′-overhangs up to 5 nt in length (**Fig. 3.6d**). In contrast to retroviral integrases, substrates with 5′-overhangs were not integrated (**Fig. 3.6d**).

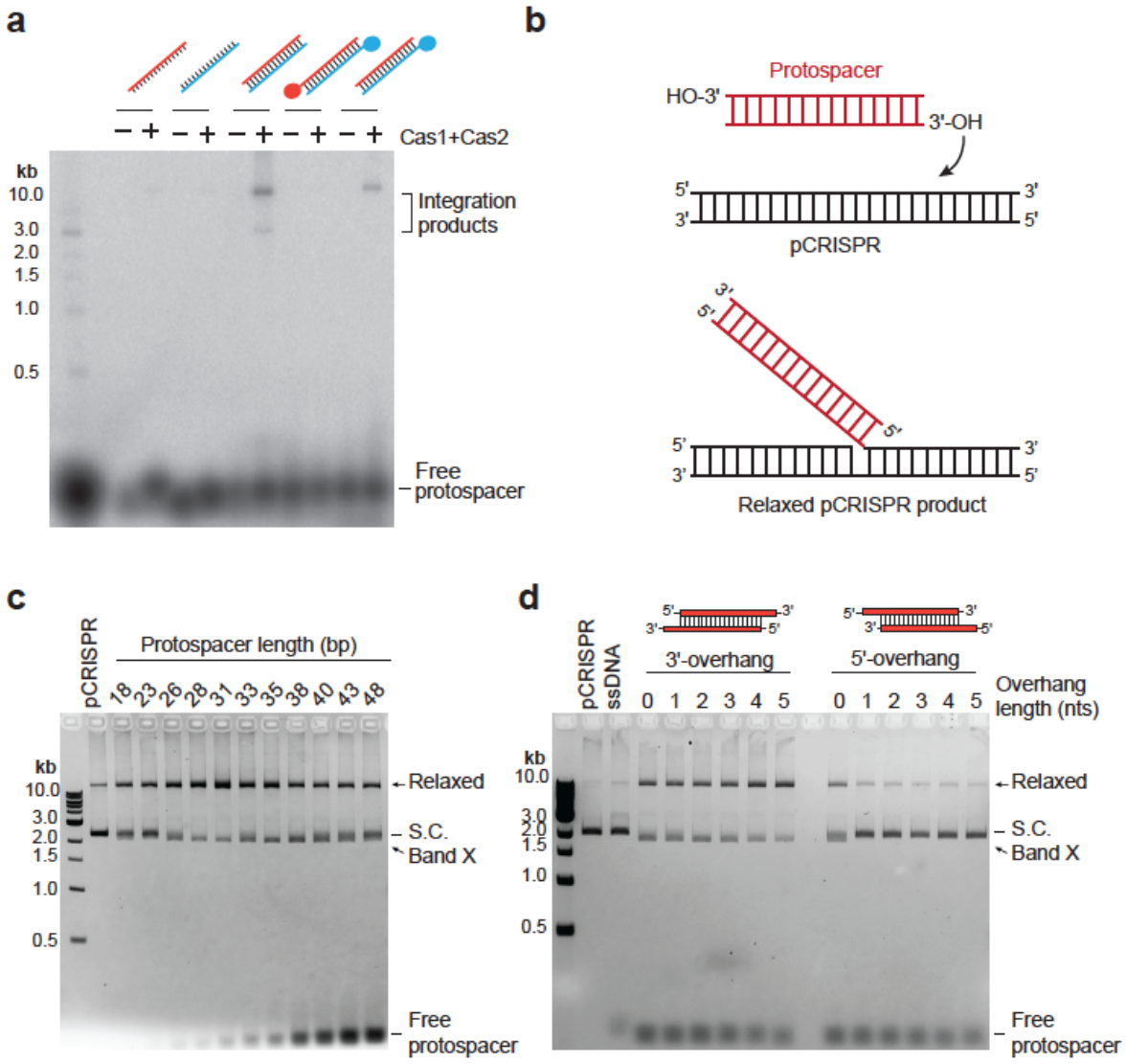


Figure 3.6 | Integration requires 3′–OH protospacer ends. (**a**) Integration assays using single-stranded DNAs and either –OH or –PO_4 at the 3′ ends of radiolabeled protospacers. The red and blue circles denote 3′–PO_4 ends. (b) Schematic of nucleophilic reaction by the 3′–OH end of the protospacer. (**c,d**) Integration assays using the indicated lengths of protospacer DNA (**b**) or varying 5′ or 3′ overhang lengths (**c**).

Retroviral integration and transposition reactions proceed via nucleophilic attack of DNA 3′–OH groups at target DNA phosphodiester bonds[59,60]. We found that phosphorylation of both 3′–ends of the protospacer ablated integration, whereas phosphorylation of only one 3′ end strongly limited integration (**Fig. 3.6a,b**). By analogy to known integrase enzyme mechanisms, DNA integration could proceed by Cas1-catalyzed direct nucleophilic attack of the substrate 3′–OH on the target DNA, or by formation of a Cas1–DNA intermediate, as occurs in the serine and tyrosine families of recombinases[85]. Four tyrosine residues in the vicinity of the Cas1 active site[23,35,63] could be involved in forming such a covalent intermediate (**Fig. 3.7a,b**). Purified Cas1 mutant proteins in which each tyrosine was individually changed to alanine supported protospacer integration *in vitro* at levels comparable to wild-type Cas1–Cas2 (**Fig. 3.7c**). Interestingly, the Cas1 Y188F mutant is completely defective in spacer acquisition *in vivo,* suggesting that Y188 may be involved in other aspects of spacer acquisition such as protospacer capture as explained in Chapter 4 (**Fig. 3.7d**). Together, the integration reaction likely proceeds via direct nucleophilic attack of protospacer 3′–OH ends onto the target DNA phosphodiester bonds, a mechanism previously hypothesized to occur *in vivo*[57] (**Fig. 3.6b**).

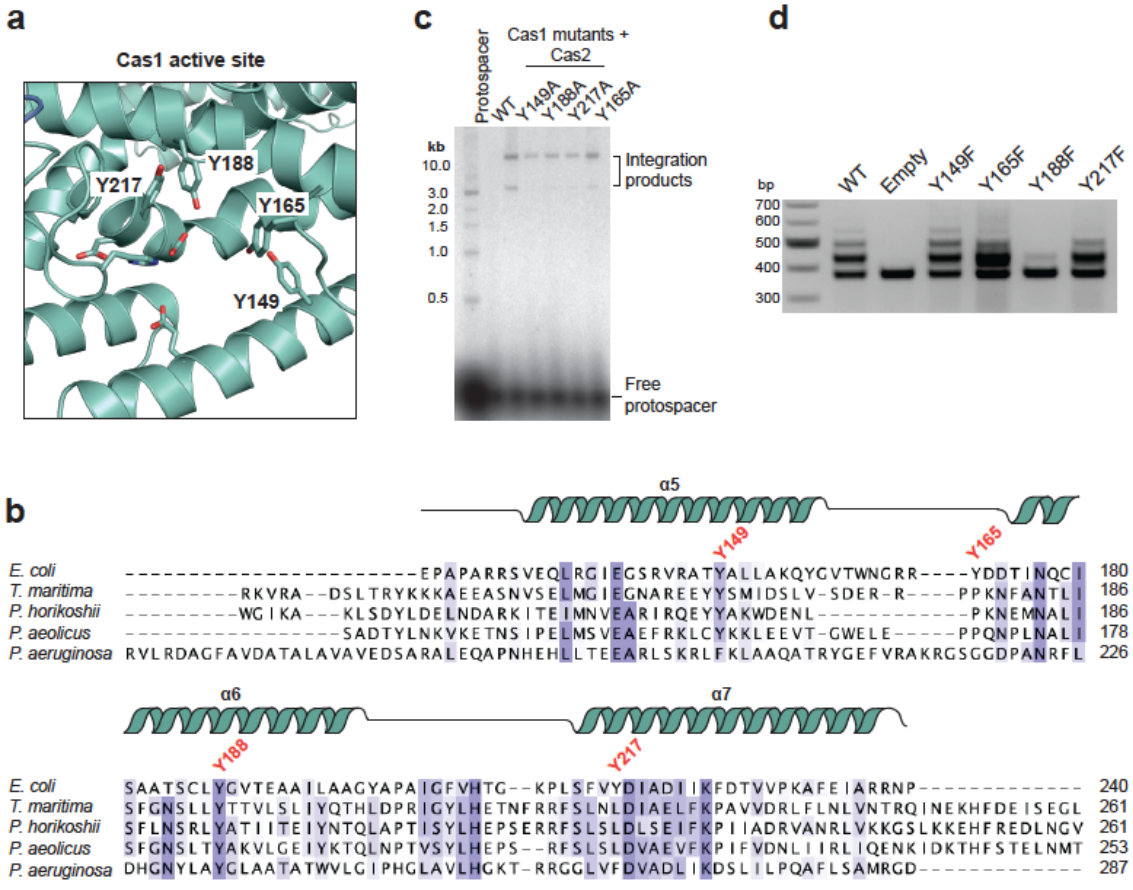


Figure 3.7 | Cas1 tyrosine mutants support integration activity. (**a**) A close-up of the Cas1 active site with the tyrosine residues labeled. (**b**) Structure-based sequence alignment of Cas1 proteins, highlighting the tyrosine residues mutated to alanine in this study. (**c**) Radiolabeled

protospacer integration assay of Cas1 tyrosine mutants complexed with wild-type Cas2. (**d**) *In vivo* spacer acquisition assays of Cas1 tyrosine mutants co-expressed with wild-type Cas2.

3.2.4 Supercoiled DNA and CRISPR locus requirements

Cas1 and Cas2 overexpression leads to site-selective spacer acquisition proximal to the leader end of the CRISPR locus, a result consistent with observations in native populations of CRISPR-containing bacteria[20–22,86]. To determine what drives such site-specific integration, we first tested various forms of the pCRISPR plasmid to determine target DNA requirements. Integration requires target DNA supercoiling, as neither relaxed nor linear pCRISPR, nor the isolated 1 kb CRISPR locus, supported integration (**Fig. 3.8a**).

As a control, we tested supercoiled pUC19 DNA, the parental plasmid of pCRISPR that lacks a CRISPR locus, and were surprised to observe integration products upon incubation with Cas1 and Cas2 in the presence of protospacer DNA (**Fig. 3.8a**). We also detected integration products with supercoiled plasmids that contain different backbone sequences than pUC19. This finding raised two possibilities: either *in vitro* spacer integration is non-specific with respect to target DNA sequence or structures and/or sequence(s) favoring integration are present in the plasmids tested for integration. To determine if integration preferentially occurred at the CRISPR locus of pCRISPR, products of radiolabeled reactions were double-digested to separate the CRISPR locus (960 bp) from the pUC19 plasmid backbone (~2.27 kb). Suggestive of CRISPR-specific integration, the 32P-radiolabel migrated solely with the CRISPR locus fragment (**Fig. 3.8b**). The same result was observed when the experiment was conducted using a target plasmid containing the CRISPR locus and a different backbone sequence (pACYC, **Fig. 3.8c**).

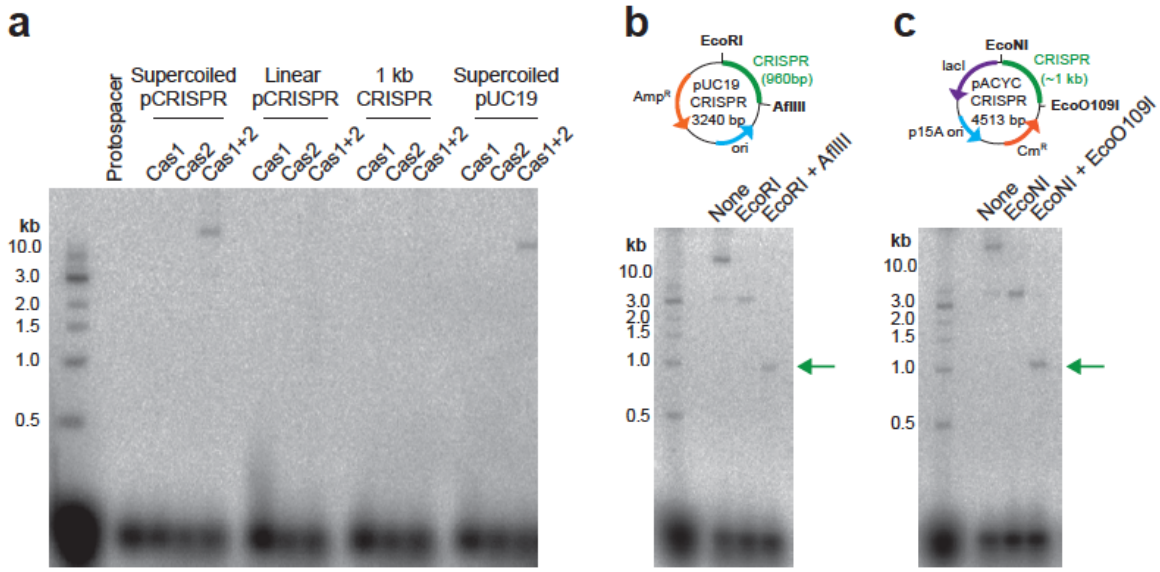


Figure 3.8 | Supercoiled target requirement. (**a**) Comparison of protospacer integration into different DNA targets. (**b,c**) Restriction enzyme digestion of pCRISPR, either in a pUC19 (**b**) or pACYC (**c**) backbone, after the itegration assay detects integration into the CRISPR fragment (green arrows).

3.2.5 CRISPR repeats provide integration specificity

To determine the exact sites of protospacer integration in these reactions, we performed high-throughput sequencing of reaction products that resulted from using either pCRISPR or the parental pUC19 vector as the target of integration (**Fig. 3.9a**). Of the 7,866 protospacer-pCRISPR junctions retrieved, ~71% mapped to the CRISPR locus (**Fig. 3.9b**). Protospacer insertion occurred at the borders of each repeat, with the most preferred site at the first repeat adjacent to the leader (**Fig. 3.9c**). The minus strand of each repeat (the bottom strand in **Fig. 3.9b,c** that runs 5′ to 3′ towards the leader sequence) is also highly preferred, highlighting the role of CRISPR repeats in providing sequence specificity for the Cas1–Cas2 complex. Sequence alignment of the integration sites revealed strong preference for sequences resembling the CRISPR repeat on both strands of pCRISPR, further supporting the selection of CRISPR repeat borders by the Cas1–Cas2 complex (**Fig. 3.9d-f**).

The most frequent integration site in the pUC19 control plasmid mapped to the *amp* resistance gene adjacent to the AT-rich promoter sequence (~8.8% of 5,524 total retrieved junctions, **Fig. 3.10a**). An inverted repeat sequence with a propensity to form a DNA cruciform[87] occurs 9 nt adjacent to this integration site (plus strand sequence: 5′–TTCAATATTATTGAA–3′), suggesting that potential DNA cruciform formation adjacent to AT-rich sequences is important for protospacer integration. Sequence analysis of pUC19 target sites revealed the propensity for a G nucleotide to occur at the –2 and +1 positions of the protospacer insertion site, similar to the preferred pCRISPR sites (**Fig. 3.10b,c**). These observations imply that in addition to sequence, pCRISPR repeat selectivity stems from the unique structural features of these sites, such as their ability to form cruciforms (**Fig. 3.11a**).

To further test the ability for Cas1–Cas2 to recognize the sequence and structural features of the CRISPR repeats, we repeated the high throughput sequencing experiments using pCRISPR target DNA bearing one repeat sequence containing either WT sequence, a mutant that disrupts cruciform formation or a mutant that flips the cruciform stems (**Fig. 11b-e**). As expected, Cas1–Cas2 prefers to integrate adjacent to the only repeat sequence in the WT pCRISPR (**Fig. 11b**). However, pCRISPR target DNAs that with the disrupted cruciform and flipped stem mutants are no longer recognized by Cas1–Cas2. Integration events into these targets occur at many sites along the plasmid, suggesting that Cas1–Cas2 may integrate at non-CRISPR sites when the cognate CRISPR locus is not present (**Fig. 3.11c,d**). Since both mutations disrupt the sequence of the CRISPR repeats, it remains untested whether the loss of specificity is due to the sequence mutations of the CRISPR repeats or the loss of a cruciform structure in the disrupted cruciform pCRISPR target.

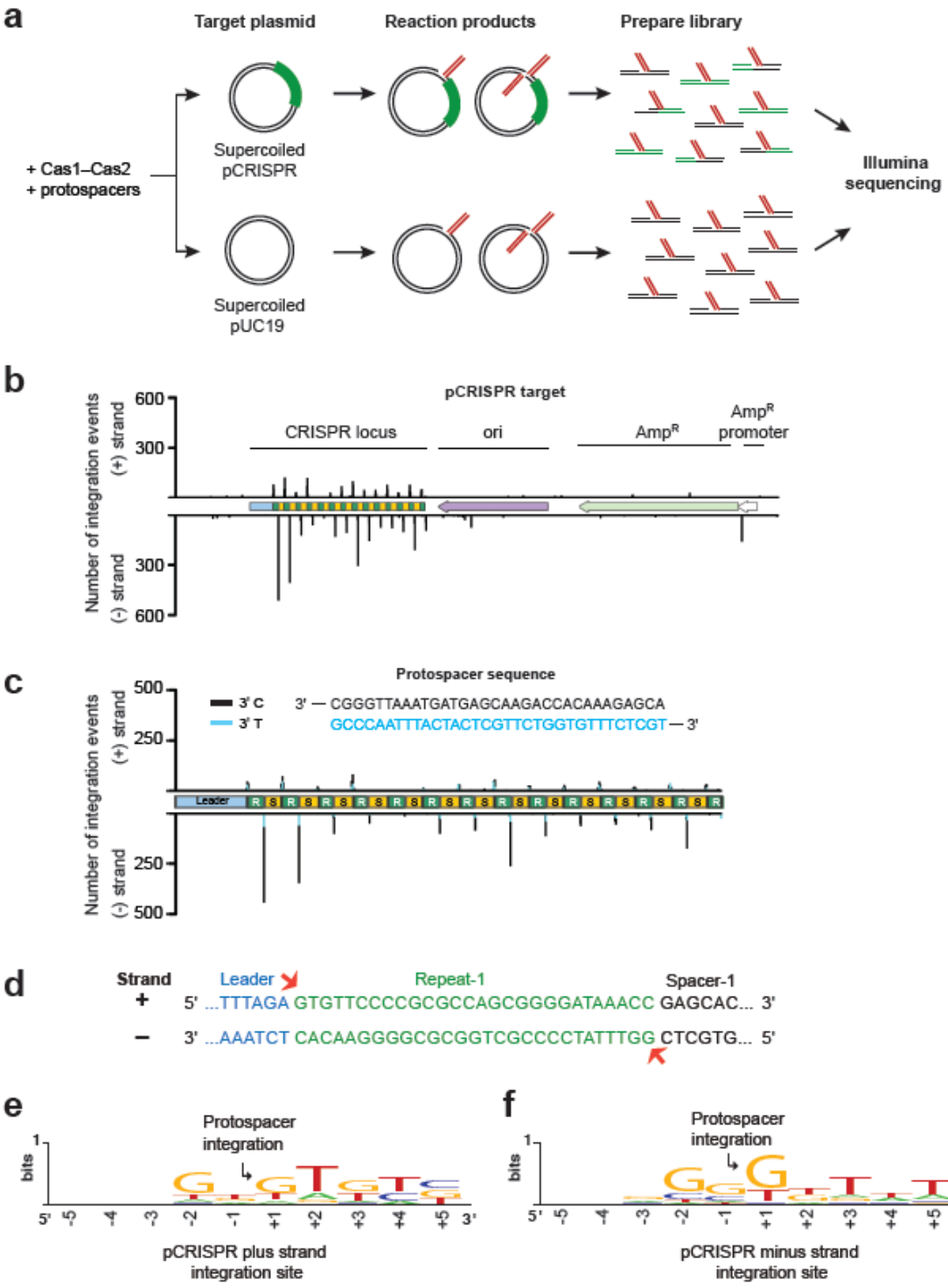


Figure 3.9 | Integration occurs into the CRISPR locus. (**a**) Schematic of the deep sequencing approach to locate the integration sites within pCRISPR and pUC19. (**b**) Integration sites along pCRISPR. (**c**) Integration sites along the ~1 kb CRISPR locus. The cyan peaks are integration sites of the strand bearing the 3′ T and the black peaks for the 3′ C-containing strand. (**d**) Sequence of the CRISPR repeat with the integration sites indicated by the red arrows. (**e,f**) WebLogo plots of the –5 to +5 positions surrounding the integration sites on the plus (**e**) and minus (**f**) strands of pCRISPR.

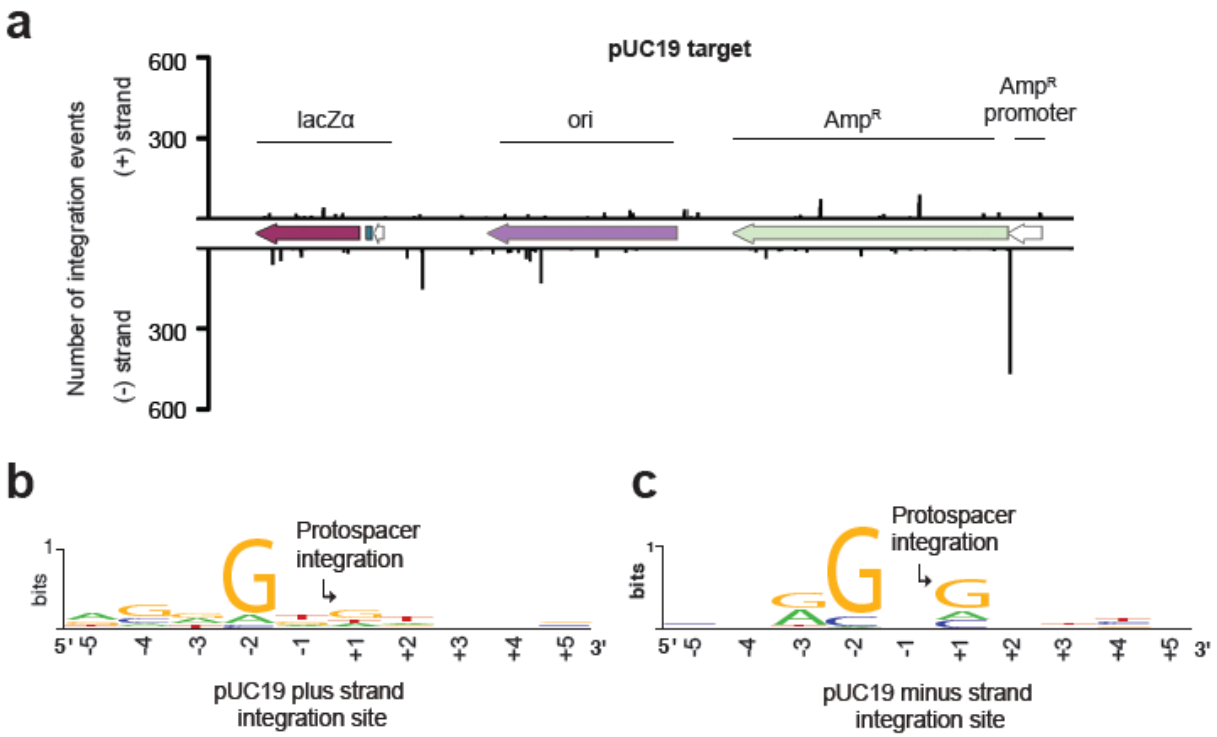


Figure 3.10 | Off-target integration into pUC19. (**a**) Integration sites along the pUC19 vector. (**b**) WebLogo plots of the –5 to +5 positions surrounding the integration sites on the plus (**b**) and minus (**c**) strands of pCRISPR.

3.2.6 Cas1–Cas2 orients protospacers into CRISPR loci

In *E. coli*, newly acquired spacers harbor a 5′ G as the first nucleotide flanking the leader-proximal end of the repeats, which originates from the last nucleotide of the AAG protospacer adjacent motif (PAM) from foreign DNA[20–22,34,36,88] (**Fig. 3.12a**). Such positional specificity is critical for crRNA-guided interference, as a mutation in this position of the corresponding crRNA disrupts PAM binding and subsequent target destruction[30,89,90]. We found that ~73% of all integration events into pCRISPR used the 3′ C end instead of the 3′ T end of the protospacer DNA during integration (**Fig. 3.12b**), and there was a strong preference for this nucleotide to attack the minus strand of the repeat sequence (**Fig. 3.9b,c**). A similar nucleotide bias was observed in the pUC19 target plasmid sequence data (**Fig. 3.10b**). This preference positions the G at the 5′ end of the protospacer substrate as the first nucleotide of the newly integrated spacer in the CRISPR locus (**Fig. 3.13**). When we used protospacer DNAs lacking a 3′ C or bearing 3′ C on both ends, the preference for integration into the minus strand of the CRISPR locus was significantly decreased (**Fig. 3.12c-e**). Thus, the Cas1–Cas2 complex plays a critical role in correctly orienting the C 3′–OH end of protospacer DNA substrates for incorporation within the CRISPR locus.

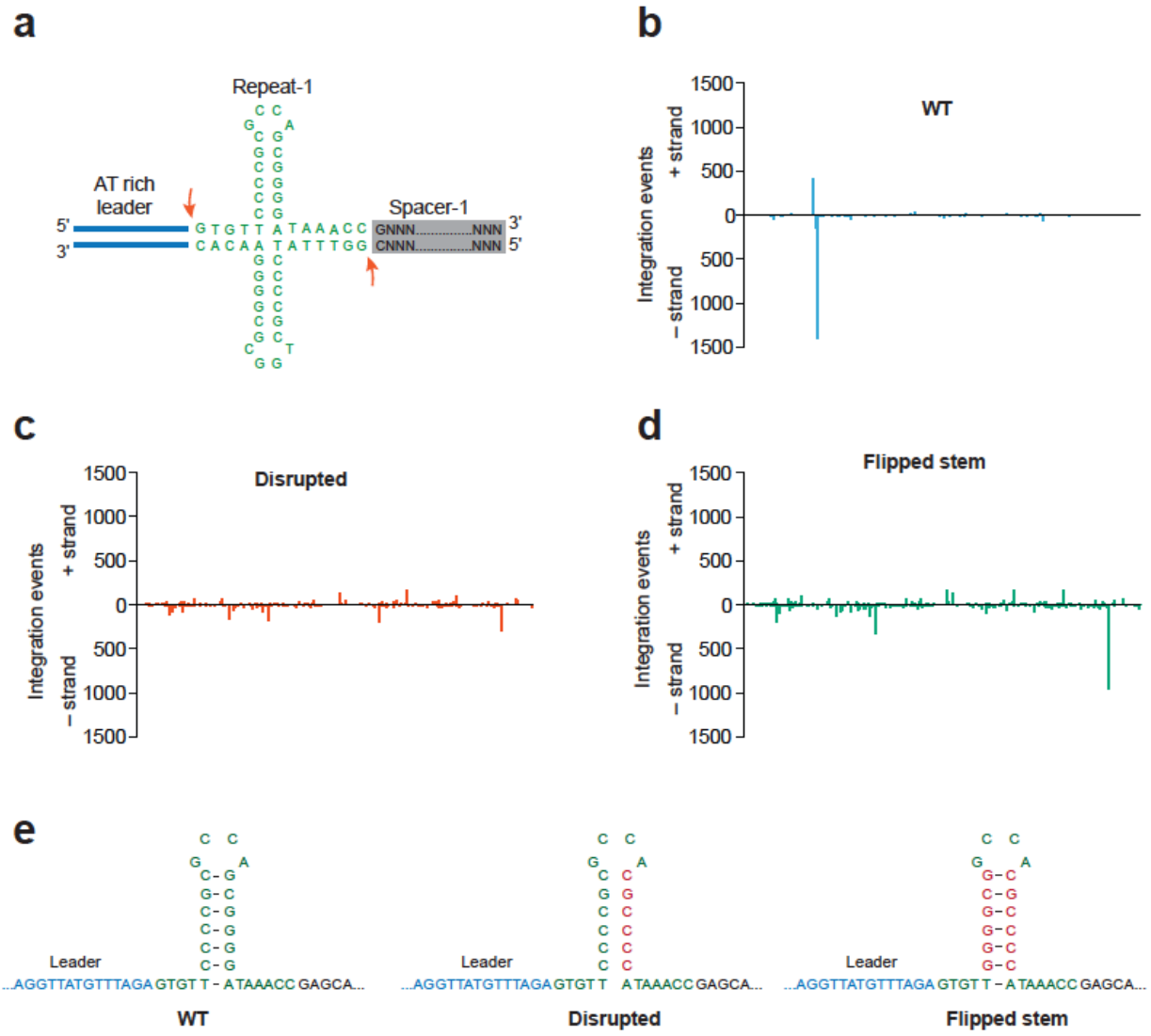


Figure 3.11 | CRISPR repeat mutations abolish integration. (**a**) Schematic of DNA cruciform formation of the CRISPR repeat. The arrows depict the integration sites. (b-d) Integration sites along pCRISPR targets containing one repeat with the WT (**b**), disrupted cruciform stem (**c**) or flipped cruciform stem (**d**) sequence. The sequences of the repeat constructs are shown in (**e**).

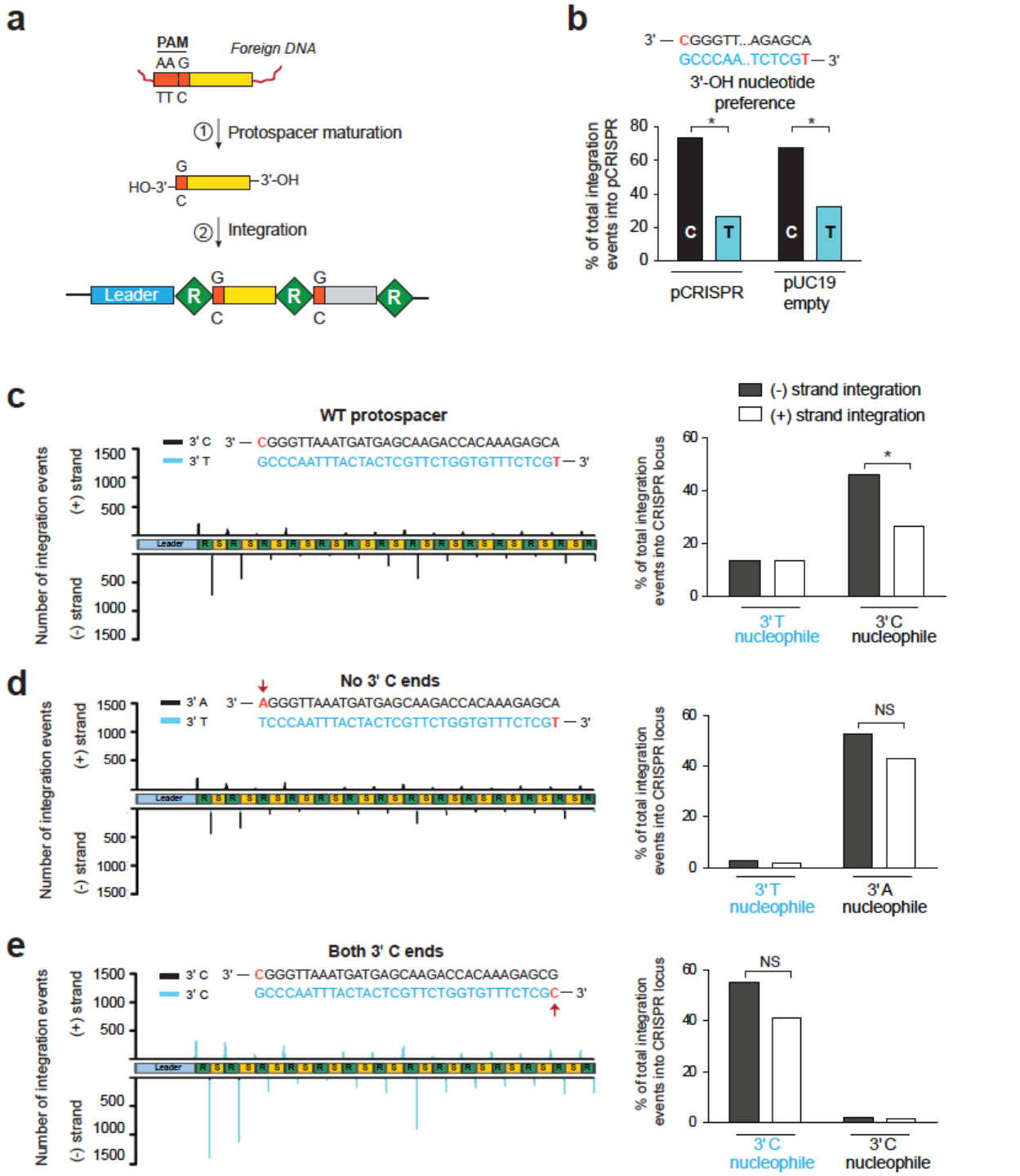


Figure 3.12 | Cas1–Cas2 correctly orients the protospacer. (**a**) Schematic of PAM recognition (AAG) in foreign DNA, followed by protospacer maturation by inclusion of the 5′ G in the 33 bp protospacer. The 5′ G becomes the first nucleotide of every integrated spacer. (**b**) Comparison of C 3′–OH or T 3′–OH selection in total reads from pCRISPR and pUC19 targets (n=7,866 for pCRISPR and n=5,524 for pUC19, chi-square test, *P<0.0001). (**c-e**) Sequencing libraries of reactions using protospacers bearing mutations of the terminal nucleotides with either no C 3′–OH or both containing C 3′–OH. NS=not significant.

3.3 Discussion

The results here explain the mechanistic basis for foreign DNA acquisition during CRISPR–Cas adaptive immunity (**Fig. 3.13**). The Cas1–Cas2 complex catalyzes integration of protospacers at the leader-end of the CRISPR locus and also selects the terminal C 3′–OH as the attacking nucleophile, resulting in the 5′ G on the opposite strand of the protospacer becoming the first nucleotide of the newly integrated spacer. This orientation bias, previously observed *in vivo*, is a key step during immunity for productive downstream foreign DNA targeting by the Cascade complex and Cas3 effector nuclease. Interestingly, the presence of the complete AAG PAM in the protospacer is not required for *in vitro* integration, suggesting that a highly specific selection or processing step occurs *in vivo* to exclude the AA nucleotides from the mature protospacer before integration.

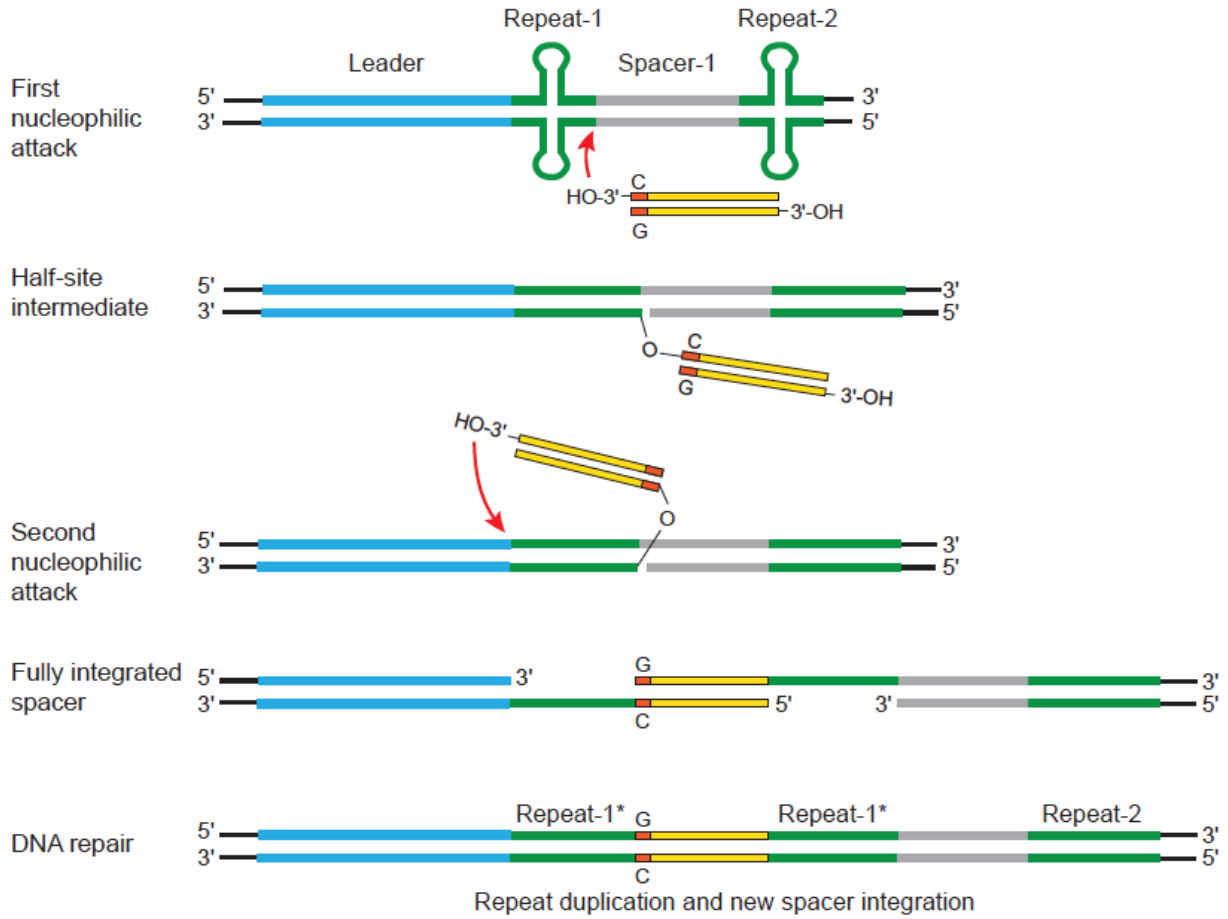


Figure 3.13 | Mechanism of spacer integration. The first nucleophilic attack occurs on the minus strand of the first repeat, distal to the leader, by the C 3′–OH end of the protospacer to produce a half-site intermediate. The second nucleophilic attack occurs on the opposite strand at the leader-repeat border. The resulting single-stranded DNA gaps are repaired by yet uncharacterized mechanisms and the protospacer is fully integrated with the G as the first nucleotide at its 5′ end. The asterisk denotes the duplication of the first repeat, as previously observed *in vivo*.

We propose a two-step integration mechanism in which the C 3′–OH first attacks the minus strand of the CRISPR repeat to produce a half-site intermediate (**Fig. 3.13**). The 3′–OH on the opposite strand of the integrating DNA then attacks the target 28 bp away on the opposite side of the repeat on the plus strand, leading to full integration of the protospacer (**Fig. 3.13**). Our *in vitro* system predominantly traps the first step of this two-step integration mechanism, suggesting that the second nucleophilic attack is greatly accelerated *in vivo* in the presence of cellular factors. This model is consistent with spacer integration intermediates that are observed *in vivo*, In which protospacers are integrated such that staggered cleavage at each end of the repeat generates single-stranded gaps that ensures repeat duplication[57]. The *in vivo* conditions could also promote the high specificity of integration to occur solely downstream of the first repeat of the CRISPR locus in E. coli, instead of at every repeat, as observed in our *in vitro* assay. In Chapter 5, we describe the discovery of a host factor in E. coli, integration host factor (IHF), that mediates leader-specific spacer integration by Cas1–Cas2 and we update our model with the first nucleophilic attack occurring at the leader-end.

CRISPR spacer integration shares mechanistic similarities with retroviral integration and DNA transposition, where the integrase/transposases enzyme uses donor DNA 3′–OH ends to make a staggered cut at the DNA target site, which concurrently joins the donor DNA to target DNA 5′–phosphates. Completion of the integration reaction requires a DNA polymerase to fill in sequence gaps and a DNA ligase to seal the phosphodiester backbone[91]. Similar polymerase and ligase functions are required to complete CRISPR spacer acquisition *in vivo*, although the specific enzymes involved have not yet been identified. Despite these similarities, we note that the Cas1 active site does not harbor the RNase H fold that defines the retroviral integrase enzyme superfamily[92]. This structural difference could explain the unexpected production of different topoisomers of pCRISPR (Band X) *in vitro*, although the physiological significance of Band X production remains unclear.

Our results highlight the fundamental role of repeat sequences at multiple stages of CRISPR–Cas adaptive immunity. In addition to creating structures within nascent CRISPR transcripts that ensure correct RNA processing during crRNA maturation[93], the repeats operate at the DNA level to recruit Cas1–Cas2 complex for sequence- and structure-specific protospacer integration. We envision that this recruitment involves transient DNA cruciform formation within the CRISPR inverted repeats that occurs as a function of target DNA supercoiling[94]. The observation that a preferred non-CRISPR site of Cas1–Cas2-mediated DNA integration is proximal to an inverted repeat adjacent to an AT-rich sequence suggests the fascinating possibility that CRISPR loci arise in naïve genomes through integration events that become self propagating through creation of repetitive sequences with properties that ensure continual recognition and activity by the Cas1–Cas2 integration machinery.

3.4 Materials and Methods

3.4.1 Cas1, Cas2 and DNA preparation

The *cas1* and *cas2* genes from *E. coli* K12 (MG1655) were cloned into expression vectors and the proteins were separately purified as previously described. The proteins were stored in 100 mM KCl, 20 mM HEPES-NaOH, 5% glycerol and 1 mM TCEP at −80 °C before use. Single-stranded DNAs were synthesized (Integrated DNA Technologies). Double-stranded DNA protospacers were annealed in 20 mM HEPES-NaOH, pH 7.5, 25 mM KCl, 10 mM $MgCl_2$ or $MnCl_2$, 1 mM DTT, 10% DMSO by heating at 95 °C for 3 min and slow cooling to room temperature. The sequence of the 33 bp protospacer used in this study was shown to be highly acquired *in vivo* in *E. coli* K12 after M13 bacteriophage infection: strand 1 (5′-GCCCAATTTACTACTCGTTCTGGTGTTTCTCGT-3′) and strand 2 (5′-ACGAGAAACACCAGAACGAGTAGTAAATTGGGC-3′). The pCRISPR target plasmid was constructed by PCR amplifying the *E. coli* BL21-AI genomic CRISPR locus and cloning the fragment into pUC19 using the following primers with the underlines indicating the respective restriction sites used: forward/EcoRI: 5′-ACGTCGAATTCTACCTTTTTAATCAATGG-3′ and reverse/AflIII: 5′-ACGTCACATGTGGTTATATGGTGGTTTATCC-3′. The pACYC CRISPR plasmid was constructed by cloning the CRISPR fragment into a pACYCDuet-1 vector using the EcoNI and AvrII restriction sites.

3.4.2 *In vitro* integration assays

The integration reactions were performed in 20 mM HEPES-NaOH, pH 7.5, 25 mM KCl, 10 mM $MgCl_2$ or $MnCl_2$, 1 mM DTT and 10% DMSO. There was little difference when DMSO was omitted from the reaction, in contrast to its *in vitro* integration enhancement with HIV-1 integrase[95]. All of the reactions were conducted with $MgCl_2$ unless otherwise noted. For reactions with the Cas1–Cas2 complex, separately purified Cas1 and Cas2 were pre-incubated for 20–30 min at 4 °C to allow complex formation. The protospacer DNAs were incubated with the protein(s) for 10–15 min at 4 °C, followed by the addition of the target pCRISPR or pUC19 plasmid DNA. The reactions were conducted at 37 °C for 1 h and quenched with DNA loading buffer containing a final concentration of 50 mM EDTA. The products were analysed on 1.5% agarose gels pre-stained with ethidium bromide. All of the reactions, except those shown in **Fig. 3.1** were conducted with 75 nM protein, 200 nM protospacers and 7.5 nM pCRISPR to clearly visualize band X from pCRISPR. Reactions in **Fig. 3.1** were performed with 50 nM protospacers. Each integration and disintegration assay was performed a minimum of three times.

3.4.3 Radiolabeled protospacer integration assays

Pre-annealed double-stranded protospacer DNA substrates were 5′-radiolabelled using [γ-^{32}P]-ATP (PerkinElmer) and T4 polynucleotide kinase (New England Biolabs). Protospacers with 3′-PO_4 ends were 5′-radiolabelled using T4 polynucleotide kinase with 3′ phosphatase minus activity (New England Biolabs). The reactions were carried out in the same buffer as above. Unless otherwise noted, 200 nM of Cas1–Cas2 was

first incubated with 20 nM protospacers at 4 °C for 10–15 min, followed by the addition of 200 ng (~5 nM) of pCRISPR. The reactions were conducted at 37 °C for 1 h and quenched with 25 mM EDTA and 0.4% SDS. The DNA samples were deproteinized with 30 μg of proteinase K for 1 h at 37 °C and ethanol precipitated. The reactions were analysed on 1.5% agarose gels. After electrophoresis, the gels were dried onto positively charged nylon transfer membrane (GE Healthcare) and imaged using Phosphor Screens (GE Healthcare). The restriction enzyme digest experiments were performed by first conducting the integration reaction, followed by addition of the respective enzymes (New England Biolabs), which were allowed to digest for an additional 1 h at 37 °C.

3.4.4 Disintegration assays

The four single stranded DNA substrates were annealed in buffer (25 mM KCl, 20 mM HEPES-NaOH, pH 7.5, 10 mM $MgCl_2$) to form 20 nM of Y DNA in a stepwise manner: 95 °C for 3 min, 60 °C for 30 min, 37 °C for 20 min and gradual cooling to room temperature. The annealing reactions were analysed on a 15% native polyacrylamide gel to confirm the formation of the Y DNA. The disintegration assay was performed in the integration reaction buffer with 100 nM protein and 5 nM Y DNA at 37 °C for 1 h. The reactions were quenched with formamide buffer and heated at 95 °C before loading on 15% 7M urea-polyacrylamide gels. The sequences of the four strands are as follows: A-20 nt (5′- AATAATAGGTTATGTTTAGA-3′); B-40 nt (5′-CCCGCTGGCGCGGGGAACACTCTAAACATAACCTATTATT-3′); C-53 nt (5′-GCCCAATTTACTACTCGTTCTGGTGTTTCTCGTGTGTTCCCCGCGCCAGCGGG-3′); and D-33 nt (5′- ACGAGAAACACCAGAACGAGTAGTAAATTGGGC-3′).

3.4.5 High-throughput sequencing

The integration reaction was performed with 75 nM Cas1–Cas2, 200 nM protospacer and 7.5 nM pCRISPR or pUC19 in 20 mM HEPES, pH 7.5, 25 mM KCl, 10 mM $MgCl_2$, 10% DMSO and 1 mM DTT. The DNAs were isolated by phenol–chloroform extraction and ethanol precipitation. The excess protospacers were removed using 100K MWCO Amicon Ultra-0.5 ml centrifugal filters. The resulting integration products were digested into smaller DNA fragments using dsDNA Fragmentase (New England Biolabs) for 75 min at 37 °C and quenched at 65 °C for 15 min. Fragments were end repaired using T4 DNA Polymerase (NEB), Klenow (NEB) and T4 PNK (NEB) and A-tailed with Klenow exo (3′ to 5′ exo minus) (NEB). Adapters were ligated onto fragments using T4 DNA ligase (NEB) and cDNA libraries were amplified by PCR using Phusion (NEB). Libraries were sequenced on an Illumina HiSeq2500 on rapid run mode. The oligonucleotides used are: universal adaptor: 5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGA CGCTCTTCCGATC*T-3′ (*phosphorothioate bond); indexed adaptor: 5′-/5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCAC-index-ATCTCGTATGCCGTC TTCTGCTTG-3′); PCR primers: 5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGA-3′, 5′-CAAGCAGAAGACGGCATACGAGAT-3′.

3.4.6 Computational analysis

For preprocessing, 3′ adapters were removed from raw Illumina reads using Cutadapt (http://code.google.com/p/cutadapt/), discarding reads shorter than 15 nt. Reads containing integrated protospacer were selected using Cutadapt, requiring the presence of at least 10 nt of protospacer sequence with no errors. After creating Bowtie[96] indexes from fasta files of the pUC19 empty and pCRISPR plasmid sequences, these reads were mapped to the respective plasmids using Bowtie, allowing up to 2 mismatches and requiring unique mapping. Sequence motif analysis depicted in **Fig. 3.9** and **3.10** were generated using WebLogo, using integration sites that are represented at least ten times in the sequencing data[97].

3.5 Accession Code

Sequencing data presented in **Fig. 3.9** and **3.10** are deposited in Gene Expression Omnibus under accession number GSE64552.

3.6 Acknowledgements

We are grateful to Minyong Chung, Philip Kranzusch and Addison Wright for technical assistance and members of the Doudna laboratory and Jamie Cate for discussions. This project was funded by US National Science Foundation grant no. 1244557 to J.A.D. and by NIH grant AI070042 to A.E. This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 Instrumentation Grants S10RR029668 and S10RR027303. J.K.N. is supported by a US National Science Foundation Graduate Research Fellowship and a UC Berkeley Chancellor's Graduate Fellowship. A.S.Y.L. is supported as an American Cancer Society Postdoctoral Fellow (PF-14-108-01-RMC). J.A.D. is an Investigator of the Howard Hughes Medical Institute and a member of the Center for RNA Systems Biology.

3.7 Author Contributions

These authors contributed to the data presented in this chapter: James K. Nuñez, Amy S.Y. Lee, Tracey L. Hinder, Alan N. Engelman and Jennifer A. Doudna. J.K.N. performed the biochemical experiments. A.S.Y.L. processed and analyzed the high-throughput sequencing data. J.K.N. and T.L.H. prepared and analyzed the high-throughput sequencing data in **Fig. 3.11**. J.K.N., A.S.Y.L., A.N.E. and J.A.D. designed the study and wrote the manuscript.

CHAPTER 4

Mechanism of foreign DNA capture and length determination

A portion of the work presented in this chapter has been previously published as part of the following paper: 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). *These authors contributed equally to this work.

4.1 Introduction

CRISPR loci are defined by repetitive elements that are separated by similarly-sized spacer sequences acquired from foreign DNA during the adaptation stage of CRISPR–Cas adaptive immunity[4,28]. CRISPR transcripts generated from the loci assemble with Cas proteins to detect and cleave foreign nucleic acids bearing sequence complementarity to the spacer segment[3,10–12,14,16,76]. In *E. coli*, expression of the Cas1–Cas2 protein complex triggers acquisition of new 33 bp spacers at the A/T-rich leader end of the CRISPR locus[20–22,35,37]. How the Cas1–Cas2 complex selects 33 bp protospacers of variable sequences and activates the 3′–OH ends for integration remains unknown. As the Cas1–Cas2 complex is sufficient to initiate spacer acquisition and adaptation of the CRISPR–Cas immune system, we hypothesized that the protein complex alone must provide the structural basis for the unknown mechanism of spacer length determination.

Here we present X-ray crystal structures of the *Escherichia coli* Cas1–Cas2 complex bound to cognate 33 nucleotide (nt) protospacer DNA substrates. The protein complex creates a curved binding surface spanning the length of the DNA and splays the ends of the protospacer to allow each terminal nucleophilic 3′–OH to enter a channel leading into the Cas1 active sites. Phosphodiester backbone interactions between the protospacer and the proteins explain the sequence-nonspecific substrate selection observed *in vivo*[7–9]. Our results uncover the structural basis for foreign DNA capture and the mechanism by which Cas1–Cas2 functions as a molecular ruler to dictate the sequence architecture of CRISPR loci.

4.2 Results

4.2.1 Cas1–Cas2 prefers protospacers with 3′ overhang ends

To determine the crystal structure of the Cas1–Cas2 integrase bound to foreign DNA, we first screened for DNA substrates for use in crystallization experiments. We used an *in vitro* integration assay, as described in Chapter 3, to test versions of a 33 bp sequence with constant overall length but different 3′ single-stranded overhang lengths. The protospacer sequence is derived from the M13 bacteriophage genome and is highly acquired into the *E. coli* CRISPR locus after infection[21]. Unexpectedly, protospacers with overhanging 3′ nucleotides are strongly preferred by the Cas1–Cas2 complex over a completely double-stranded 33 bp protospacer (**Fig. 4.1a-c**). Single-stranded DNA and substrates with 5′ overhangs are poor substrates for integration, highlighting the ability of Cas1–Cas2 to select specific DNA substrates prior to integration[58]. The most preferred protospacer DNA for *in vitro* integration consists of five overhanging nucleotides on each 3′ end. To determine the molecular basis of protospacer capture, we assembled Cas1–Cas2 with the preferred protospacer substrate, isolated protospacer-bound Cas1–Cas2 complexes by gel filtration chromatography and set crystallization trays with the DNA-bound complex (**Fig. 4.1d-f**).

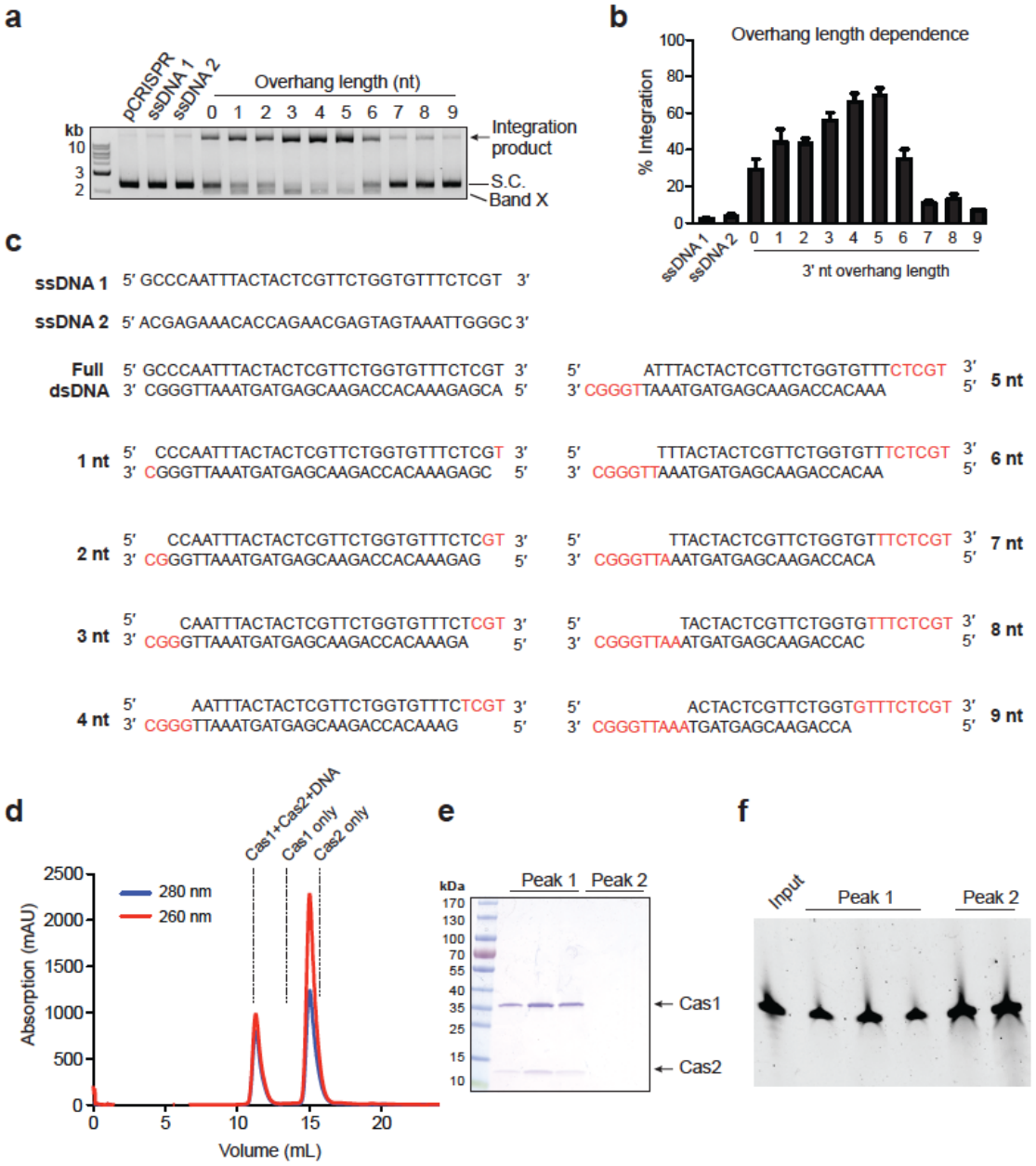


Figure 4.1 | Cas1–Cas2 prefers protospacers with 3′ overhang ends. (**a**) A representative agarose gel of *in vitro* integration reactions using increasing lengths of 3′ single-strand protospacer DNA overhangs. S.C., supercoiled pCRISPR substrate; Band X, relaxed pCRISPR byproduct. (**b**) A plot of percent integration of protospacers ± standard deviation from *in vitro* integration assays shown in (**a**). (**c**) Protospacer sequences for the assays described in (**a**) and (**b**) with the red nucleotides indicating the 3′ overhang regions. (**d**) Gel filtration chromatogram of pre-assembled Cas1–Cas2 complex with protospacer DNA containing five-nucleotide 3′ overhangs. The dotted lines indicate the expected elution volumes of DNA-bound Cas1–Cas2, Cas1 only, and Cas2 only. (**e,f**) The fractions from peak 1 (~12 ml) and peak 2 (~15 ml) were analyzed by Coomassie-stained SDS-PAGE (**e**) and 12% urea-PAGE (**f**) to confirm the presence of Cas1, Cas2 and protospacer DNA.

4.2.2 Crystal structure Cas1–Cas2 bound to foreign DNA

We determined crystal structures of Cas1–Cas2 bound to protospacer DNA in the presence and absence of Mg^{2+} at 3.0 Å and 3.2 Å resolutions, respectively (**Table 4.1**). The structures reveal a hexameric protein architecture comprising four copies of Cas1 and two copies of Cas2, in which the protospacer spans the central Cas2 dimer and terminates within individual Cas1 subunits on each end of the complex (**Fig. 4.2a**). Structural superposition of the Cas1–Cas2 complex with and without bound DNA reveals a DNA-induced change in Cas1 subunit orientation in which each Cas1 dimer rotates ~10° in opposing directions against the central Cas2 hub (**Fig. 4.2b,c**). Cas1–Cas2 protospacer capture positions each single-stranded protospacer 3′ end within a channel leading directly to a Cas1 active site. Simulated annealing omit maps show clear electron density for the double-helical region and the five nucleotide overhangs on each end of the protospacer (**Fig. 4.3a-c**). The constrained protein channel guiding each DNA strand from its double-helical region to the single strand-accomodating Cas1 active site explains the specificity of Cas1–Cas2 for five-nucleotide 3′ overhang substrates. Together, the overall architecture of the complex reveals a ruler mechanism in dictating protospacer length during foreign DNA capture.

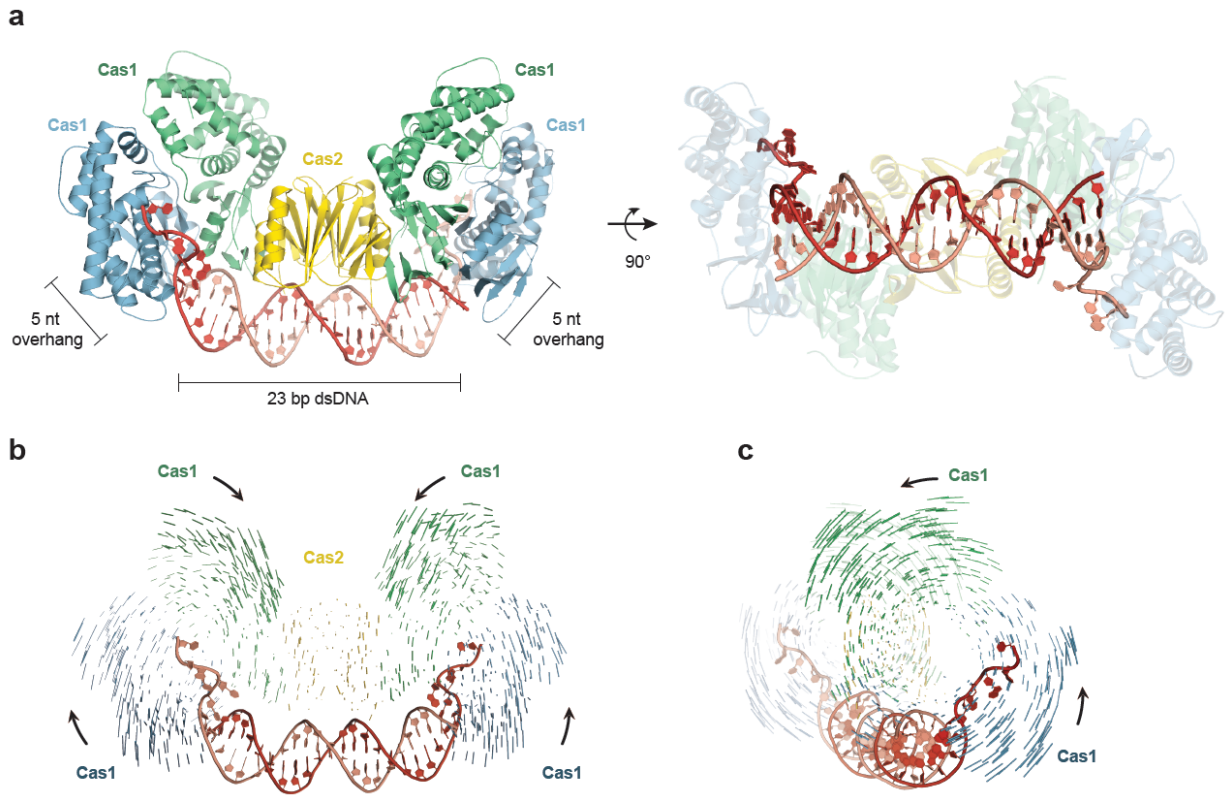


Figure 4.2 | Crystal structure of Cas1–Cas2 bound to protospacer DNA. (**a**) Two views of the overall architecture of Cas1–Cas2 bound to protospacer DNA reveals a ruler mechanism during foreign DNA capture. The line segments indicate the length of the DNA, spanning a total of 33 nucleotides. (**b,c**) Vector lines depicting the conformational changes the Cas1–Cas2 complex undergoes upon protospacer DNA binding compared to the apo complex (PDB 4P6I, Chapter 2). The Cas1 subunits rotate towards the direction of the arrows.

	Without Mg^{2+}	With Mg^{2+}	Splayed substrate
Data collection			
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$
Cell dimensions			
a, b, c (Å)	88.02, 120.01, 196.01	75.66, 165.93, 167.26	88.02, 123.01, 196.01
α, β, γ (°)	90, 90, 90	90, 90, 90	90, 90, 90
Resolution (Å)	49.00–3.20 (3.36 – 3.20)	46.41–2.95 (3.06–2.95)	48.9–3.35 (3.42–3.35)
R_{merge} (%)	30.8 (146)	19.6 (157)	28.5 (126)
R_{pim} (%)	12.8 (61.4)	10.8 (86.3)	21.6 (94.3)
I/σ	6.4 (1.5)	9.8 (1.4)	5.0 (1.3)
$CC_{1/2}$	98.5 (72.4)	99.3 (42.0)	98.3 (72.7)
Completeness (%)	99.8 (99.0)	100 (99.9)	99.6 (97.7)
Redundancy	6.7 (6.6)	7.9 (8.0)	4.1 (4.0)
Wilson B factor ($Å^2$)	63.8	64.0	73.7
Refinement			
Resolution (Å)	49.00–3.20	46.41–2.95	49.00–3.35
No. reflections	35,808 (3,502)	44,960 (4,418)	31,049 (2885)
R_{work}/R_{free}	24.2/27.0	23.0/25.4	23.2/27.4
No. atoms			
Protein	9,375	9,576	9,375
DNA	1,142	1,142	1,165
Metal	0	4	0
Average B-factors ($Å^2$)			
Protein	65.9	66.6	86.6
DNA	76.2	67.2	103.0
Metal		51.6	
R.m.s deviations			
Bond lengths (Å)	0.003	0.003	0.004
Bond angles (°)	0.72	0.75	0.81
Ramachandran statistics (%)			
Favored	96.0	95.0	96.0
Allowed	3.75	4.51	3.58
Outliers	0.25	0.49	0.42

One crystal was used for each structure.
Highest resolution shell is shown in parenthesis.

Table 4.1 | Summary of X-ray crystallography data collection and refinement.

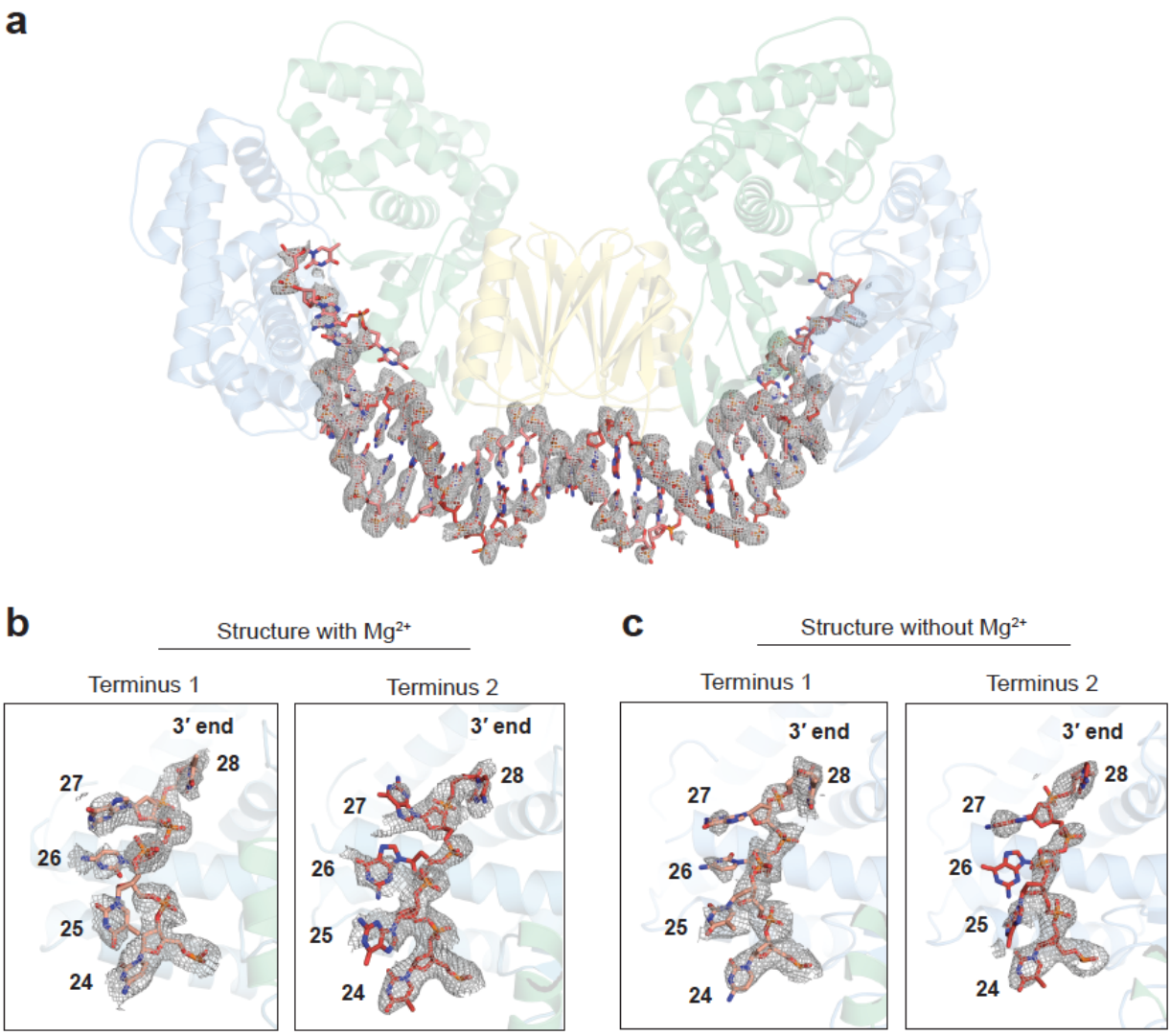


Figure 4.3 | Omit maps of the protospacer DNA. (**a**) Simulated annealing F_O–F_C omit electron density map of the entire protospacer DNA using the 'no Mg^{2+}' map and model. (**b,c**) Simulated annealing F_O–F_C omit electron density maps of the terminal five nucleotides in the active sites of the structures (**a**) with Mg2+ or (**b**) without Mg^{2+} in the crystallization condition. The maps are contoured at 2.0σ.

4.2.3 Coordination of 3′–OH nucleophiles and DNA backbone

Two of the four Cas1 subunits, colored green in Fig. 4.2, are not occupied with the protospacer 3′ ends and are probably non-catalytic, since the 3′–OH nucleophile and the scissile phosphodiester bond of the target DNA must be in the same active site for direct nucleophilic integration (**Fig. 4.4a**). In the catalytic Cas1 subunits, colored blue

in **Fig. 4.2**, the 3′ terminal base is involved in a stacking interaction with Y217 that positions the nucleophilic 3′–OH ends of the protospacer near the conserved metal-binding residues E141, H208 and D221 (**Fig. 4.4b,c**). Although we cannot assign density for Mg^{2+} in the active sites, these three residues have been shown previously to coordinate a Mn^{2+} ion in the active site of Cas1 from *Pseudomonas aeruginosa*[63]. Furthermore, alanine mutations at these positions disrupt *in vivo* spacer acquisition[20,21,35]. Thus, the observed positioning of the 3′–OH nucleophiles and catalytic residues represents the active configuration of the nucleoprotein complex immediately before spacer integration.

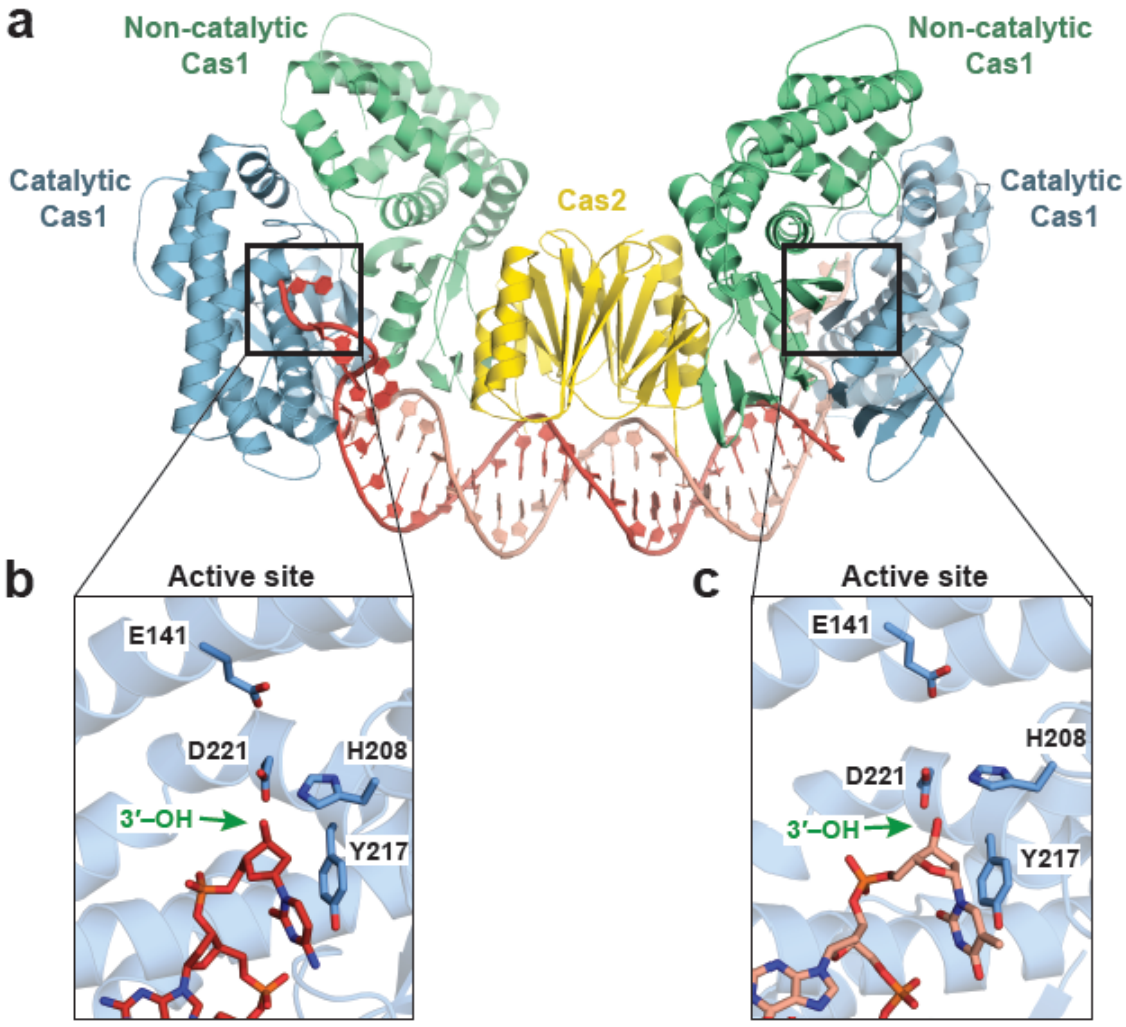


Figure 4.4 | Active site positioning of nucleophilic 3′–OH. (**a**) Assignment of the Cas1 subunits based on the presence of the nucleophilic 3′–OH ends in the active sites. (**b,c**) Stick configurations of the catalytic Cas1 active sites (blue subunits in **a**) that coordinate the nucleophilic 3′–OH ends of the protospacer (green arrow).

The majority of the interactions between Cas1–Cas2 and protospacer DNA involve coordination of the phosphate backbone rather than base-specific contacts, consistent with the variable sequence selection of protospacers that is essential for resistance to diverse foreign sequences[7–9]. Two central regions of the Cas1–Cas2 complex, which we term the 'arginine clamp' and the 'arginine channel', stabilize the protospacer (**Fig. 4.5a-d**). The arginine clamp interacts with the middle of the duplex region where four Arg residues coordinate each DNA strand: Cas1 R41 and Cas2 R16, R77 and R78 (**Fig. 4.5c**). Reverse charge mutations of Cas1 R41 and Cas2 R16 and R78 drastically reduce spacer acquisition *in vivo*, whereas the Cas2 R77E mutant functions similar to wild-type Cas2 (**Fig. 4.5e**). Thus, Cas1 R41, Cas2 R16 and R78 are the key constituents of the arginine clamp. The contribution of Cas2 to protospacer DNA binding supports the previous hypothesis that the main function of Cas2 is to form a non-catalytic scaffold within the Cas1–Cas2 complex[35].

Cas1 residues R66, R84, R245 and R248 line the arginine channel that stabilizes the junction where the duplex region terminates and the single-stranded DNA overhang enters the active site. Reverse charge mutations of each arginine lining the arginine channel disrupt spacer acquisition *in vivo* (**Fig. 4.5e**). In addition, purified Cas1 R59D or R66D proteins complexed with wild-type Cas2 are highly defective in integrating 33-bp duplex or five-nucleotide overhang protospacer substrates *in vitro* (**Fig. 4.5f**). Fluorescence polarization assays demonstrate that the mutant complexes exhibit dramatically reduced affinity for protospacer DNA, highlighting the critical role of this part of the Cas1–Cas2 complex for protospacer capture and complex stability (**Fig. 4.5g**).

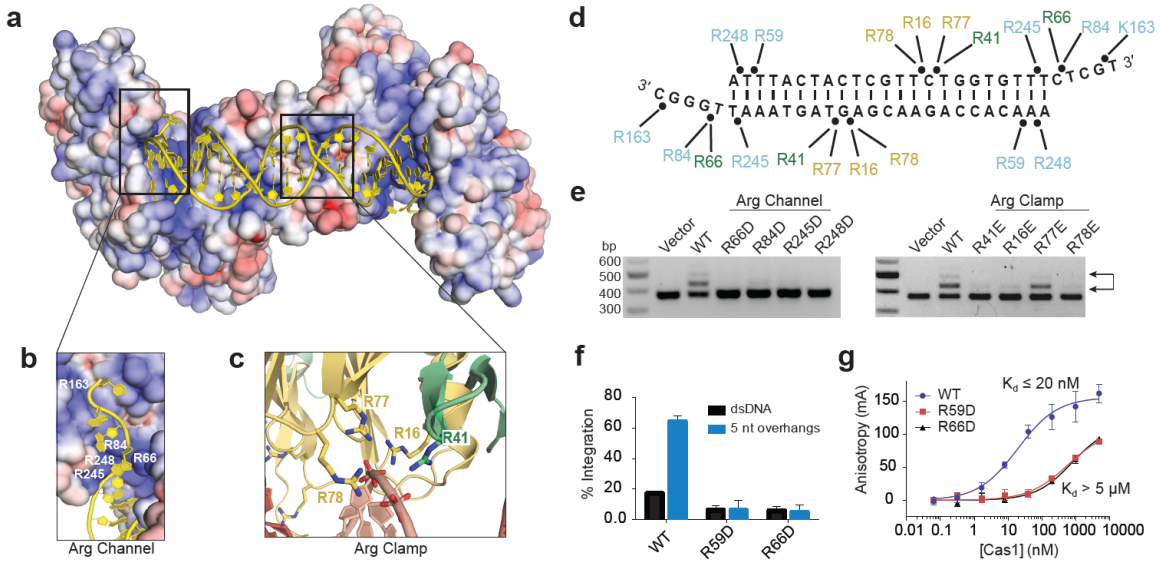


Figure 4.5 | Coordination of protospacer DNA within Cas1–Cas2. (**a**) Electrostatic potential surface representation of the Cas1–Cas2 complex with the protospacer shown in yellow. (**b**) Close up of the arginine channel that stabilizes the ssDNA overhang. (**c**) Stick configuration representation of arginine clamp residues that coordinate the protospacer duplex region. (**d**) Map of amino acid residues that coordinate the protospacer phosphodiester backbone (black dots). Residue colors indicate the Cas1–Cas2 promoters from Fig 4.2a. (**e**) Agarose gels of *in*

vivo spacer acquisition assays of arginine channel and clamp mutant proteins. WT, wild-type. (**f**) Plot of percent *in vitro* integration of either double-stranded DNA (dsDNA; black) or five-nucleotide (nt) overhang (blue) protospacers with wild-type Cas1, Cas1 (R59D) or Cas1 (R66D) complexed with Cas2. (**g**) Fluorescence polarization binding assays of a five-nucleotide overhang protospacer with the same mutants in **f** complexed with Cas2. The calculated relative binding affinities (K_d) are indicated. Error bars represent the standard deviation of three independent experiments.

4.2.4 Mechanism of protospacer DNA end separation

The Cas1–Cas2–DNA crystal structures uncover a protein wedge that terminates the protospacer double-stranded DNA region and allows single-stranded DNA overhangs to enter the arginine channel. A stacking interaction of the 5′ terminal base (adenine 6 in **Fig. 4.6**) with Y22 of Cas1 stabilizes protospacer duplex unwinding, directing each single-stranded 3′ overhang to sharply bend ~90° away from the duplex and into the active site channel (**Fig. 4.6a**). A mutation of Y22 to alanine reduces spacer acquisition *in vivo*, whereas a phenylalanine mutation has near wild-type levels of acquisition, consistent with a specific role for Cas1 Y22 base-stacking in protospacer strand splaying (**Fig. 4.6b**).

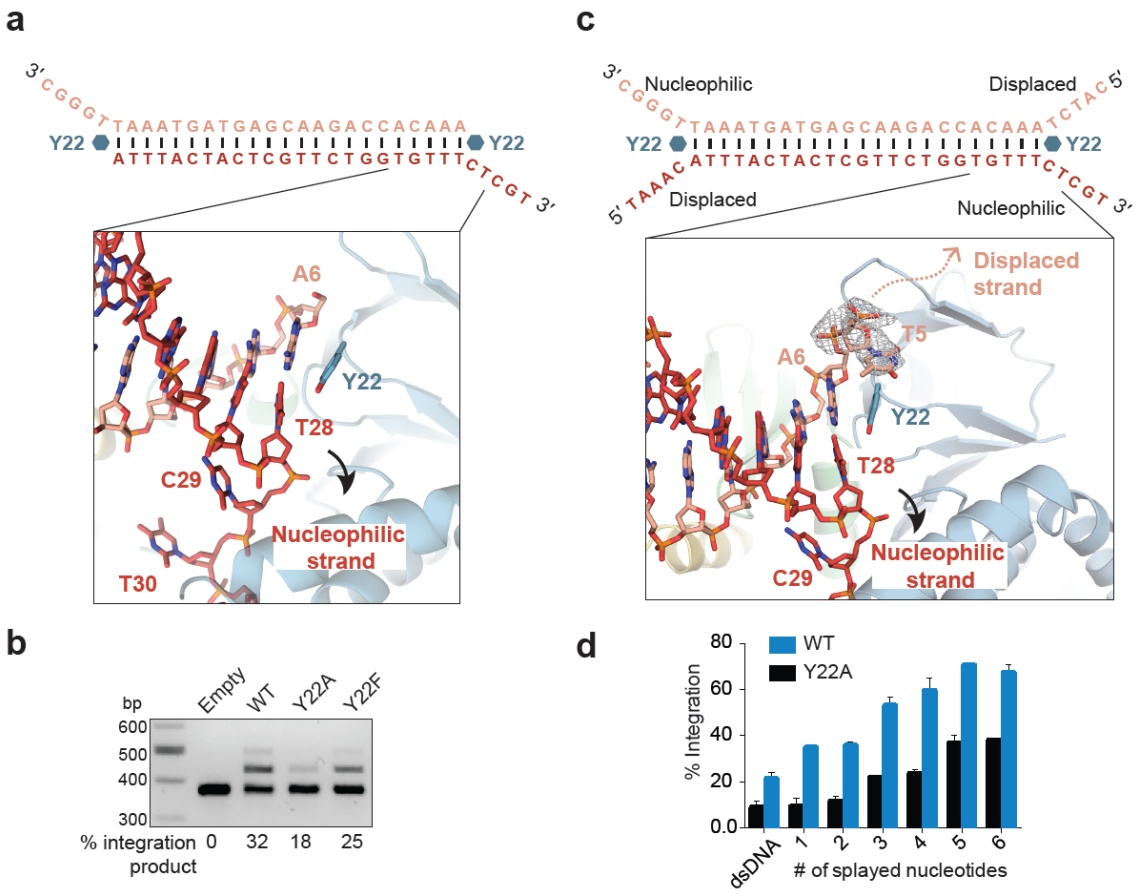


Figure 4.6 | Cas1–Cas2 splays the ends of the protospacer DNA. (**a**) Sequence of the protospacer DNA and a close up of the DNA fork where the dsDNA terminates and the ssDNA

overhang begins. Cas1 Y22 is involved in base stacking with the A6 base. (**b**) Agarose gel of *in vivo* acquisition assay of co-expressed WT Cas1 or the indicated Cas1 mutant with Cas2. Quantification is the mean of three independent experiments ± standard deviation. (**c**) Crystal structure of Cas1–Cas2 bound to splayed DNA ends. The gray mesh shows the $2F_O–F_C$ density contoured at 2.2σ of the first ejected nucleotide of the displaced strand. The arrows indicate the opposite trajectories of each strand. Plot of percent integration of increasing number of splayed nucleotides at the protospacer ends using WT Cas1 (blue) or Cas1 (Y22A) (blue) complexed with Cas2.

The observed stacking interaction raises the possibility that fully duplexed protospacers are separated by Cas1 Y22, thereby displacing the 5′ end of the duplex, which we term of the non-nucleophilic strand, from the nucleophilic strand carrying the 3′–OH. DNA transposases and retroviral integrases also utilize end fraying to isolate the reactive DNA strands for chemistry within enzyme active sites[98–100]. To test this potential activity of Cas1–Cas2, we introduced an increasing number of mismatches at the ends of the 33 bp protospacer to disrupt end base pairing and assayed their potential for *in vitro* integration (**Fig. 4.6d**). Similar to the 3′ overhang substrates, the 4- and 5-nucleotide frayed ends are highly preferred, presumably due to the lower energy required for capture of these substrates compared to perfectly duplexed ends. The complex containing Y22A mutant regains marginal activity with substrates containing 5- or 6-nucleotide splayed ends, suggesting that Y22 steers the non-nucleophilic DNA strand away from the active site (**Fig. 4.6d**). Together, we hypothesize that Cas1 Y22 plays a critical role in measuring the double helical region of protospacers. Sequence alignment of representative Cas1 proteins in Type I CRISPR systems reveals that Y22 is not universally conserved in other bacteria, suggesting that additional or different Cas1 residues may stabilize the splayed ends in other CRISPR–Cas systems (**Fig. 4.7**).

To determine the trajectory of the displaced non-nucleophilic strand after end splaying, we crystallized Cas1–Cas2 with a protospacer with five-nucleotide frayed ends on both sides (**Fig. 4.6c**). The electron density at the fork is similar to the structures described above, except that we observe the first nucleotide of the displaced non-nucleophilic strand pointing in the opposite direction from the nucleophilic single-stranded DNA strand. Clear electron density is not observed for the remaining nucleotides of the displaced strand, indicating that they are not stabilized by the complex.

4.2.5 Model of protospacer DNA integration

An alternative crystal form grown in the presence of Mg^{2+} reveals secondary Cas1–DNA interactions that provide additional insight into the mechanism of Cas1–Cas2–protospacer complex binding to the target genomic CRISPR locus for protospacer integration. In addition to the two Cas1 catalytic active sites carrying the 3′–OH ends of the protospacer, the non-catalytic Cas1 active sites interact with the protospacer DNA from a symmetry mate, revealing a possible coordination of the target DNA during integration (**Fig. 4.8a**). The non-catalytic Cas1 engages the DNA minor groove by contacts with α helix 7, causing a slight kink on the DNA compared to our

alternative crystal form lacking Mg^{2+}. A close-up of the active site shows continuous density for Mg^{2+} with E141, H208, D221 and a phosphate backbone of the presumed target DNA, capturing a snapshot of scissile phosphodiester bond coordination before integration (**Fig. 4.8b**).

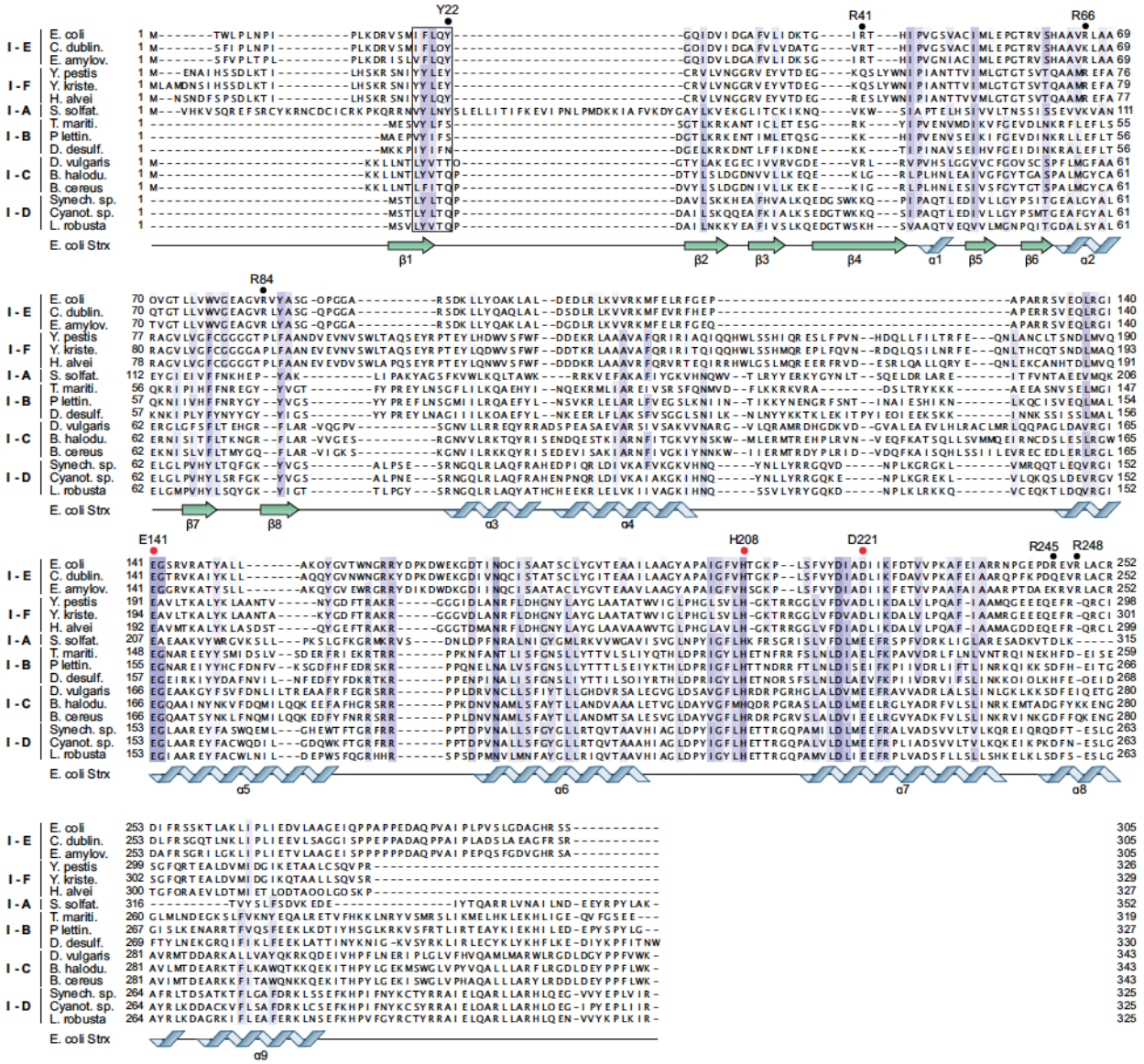


Figure 4.7 | Sequence alignment of Cas1 proteins in Type I CRISPR systems. (**a**) Sequence alignments of Cas1 from representative organisms with Type I CRISPR systems. The *E. coli* sequence is displayed at the top. The dogs indicate the residues described in this study, with the red dots indicating the metal-binding residues. The box highlights the non-universal conservation of the *E. coli* Y22 residue in the β1 region of type I CRISPR systems. The secondary structure representations shown are for the *E. coli* Cas1.

Because integration must occur in the active site that coordinates the 3′–OH of the protospacer DNA, we modeled the protein-DNA interactions from the non-catalytic Cas1 active sites into the catalytic Cas1 ctive sites. This reveals the positioning of the nucleophilic 3′–OH of the protospacer ends for attacking the scissile phosphodiester bond in the modeled DNA (**Fig. 4.8c,d**). Further work will be needed to shed light on how the complex specifically recognizes the leader-repeat region of the CRISPR locus for integration, as recently observed *in vitro*[57,58,61].

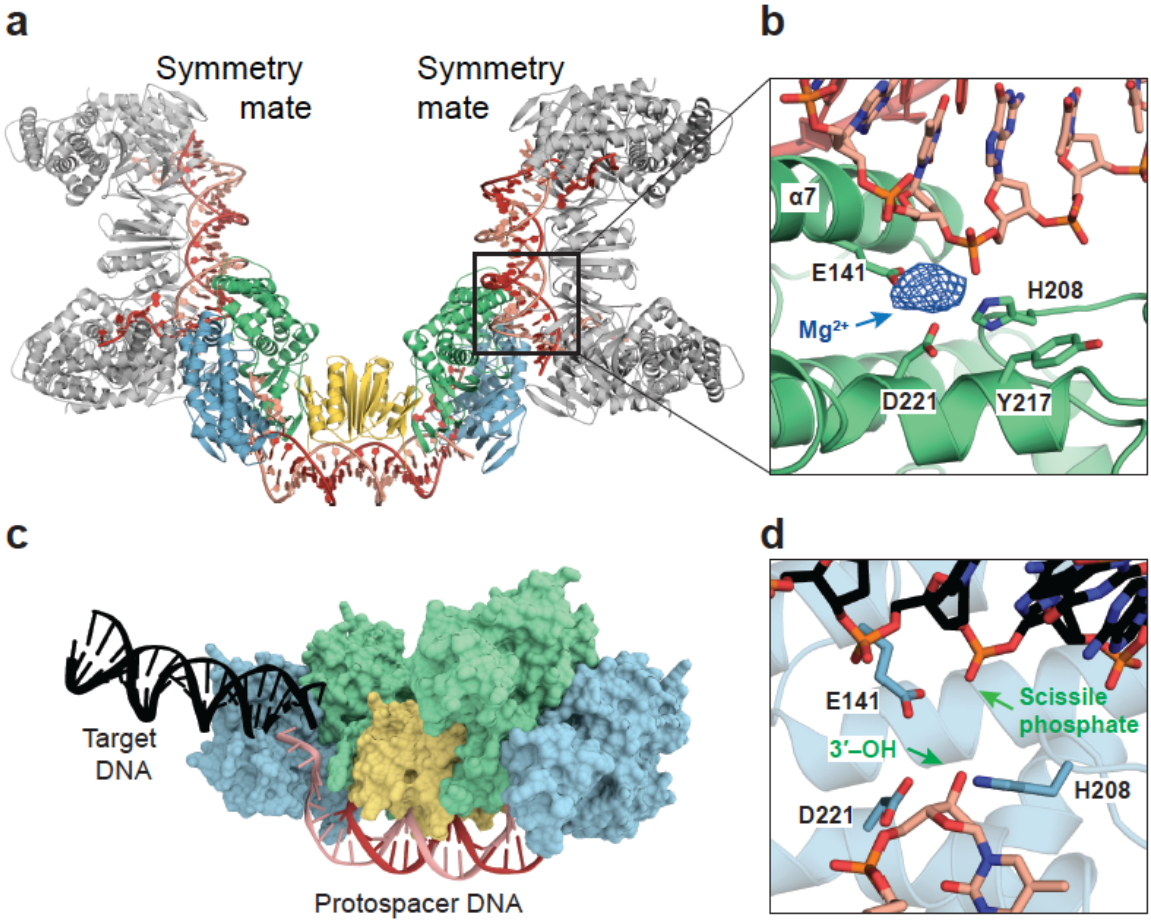


Figure 4.8 | Model of protospacer DNA integration. (**a**) View of crystal packing from a symmetry mate complex (gray) showing coordination of the symmetry DNA along a Cas1 active site. (**b**) A magnified view of the coordination of the phosphodiester backbone with metal-binding residues E141, H208 and D221. The mesh represents a F_O–F_C density for a Mg^{2+} ion, contoured at 2.2σ. (**c,d**) Model of protospacer DNA integration into target DNA (black) and positioning of the scissile phosphate (green arrow) and the 3′–OH nucleophile in the Cas1 active site.

4.3 Discussion

Together, these data explain key aspects of Cas1–Cas2 integrase-mediated acquisition of new foreign DNA protospacers into bacterial genomes. First, we show that the substrates for integration are double-stranded DNA. Importantly, however, optimal substrates include a central 23 bp helical region flanked by five single-stranded nucleotides on each 3′ end. If substrates for CRISPR integration come from single-stranded DNA products of RecBCD, as recently suggested, they must somehow anneal or otherwise become double stranded before Cas1–Cas2 capture[37]. Recently, a crystal structure of Cas1–Cas2 bound to protospacer DNA containing the protospacer adjacent motif (PAM) sequence was reported[39], revealing the intricate base-specific contacts of the -1 and -2 positions of the protospacer by Cas1. Since the reported structure contained a protospacer with PAM sequences on both ends of the DNA, it remains unclear how the Cas1–Cas2 complex searches for protospacers with only one PAM sequence, as observed *in vivo*.

Second, the Cas1–Cas2 integrase architecture specifies the precise length of integrated DNA, ensuring uniformity of spacer lengths within CRISPR loci. The coordination of the DNA by phosphodiester backbone contacts with Cas1 and Cas2 reveal the biological observation that spacers are highly diverse in sequence except for the PAM nucleotides. Unexpectedly, Cas1 has an additional role in splaying the ends of the protospacer DNA to guide the single stranded DNA carrying the 3′–OH into the Cas1 active sites. The Cas1 Y22 residue appears to play a critical functioning in dictating the length of the double helical region to 23 base pairs. Future structural studies of Cas1–Cas2 integrase complexes bound to protospacer DNA from other organisms could uncover the residue(s) responsible for dictating protospacer lengths, as the *E. coli* Cas1 Y22 is not universally conserved.

Finally, the structure-based model of DNA target sequence positioning suggests that in addition to catalyzing the integration reaction, Cas1 plays a role in binding the target CRISPR locus. Since the terminal nucleotides containing the 3′–OH nucleophiles are coordinated similarly in the Cas1 active sites, target binding could possibly disrupt the structural symmetry observed in the crystal structure to coordinate the sequence-specific integration reactions at the leader-end of the CRISPR locus. Insights into target site recognition may offer strategies for altering or enhancing integration site specificity, with implications for use of the Cas1–Cas2 integrase as a genome-modifying technology.

4.4 Materials and Methods

4.4.1 Cas1, Cas2 and DNA preparation

The Cas1 and Cas2 proteins from *E. coli* K12 (MG1655) were cloned and separately purified as described in Chapter 2. Single-stranded DNA oligonucleotides purchased from Integrated DNA Technologies were annealed in 20 mM HEPES-NaOH, pH 7.5, 25 mM KCl, 10 mM $MgCl_2$ by heating at 95 °C for 3 min and slow cooling to room temperature. The pCRISPR DNA target for *in vitro* integration was constructed as previously described[58]. The DNA substrates used for crystallization were gel purified prior to complex formation. The sequences for the five nt overhang substrates used for crystallization are: ssDNA1 (5′-ATTTACTACTCGTTCTGGTGTTTCTCGT-3′) and ssDNA2 (5′-AAACACCAGAACGAGTAGTAAATTGGGC-3′). The sequences for the five-nucleotide splayed substrates are: ssDNA1 (5′-TAAACATTTACTACTCGTTCTGGTGTTTCTCGT-3′) and ssDNA2 (5′-CATCTAAACACCAGAACGAGTAGTAAATTGGGC-3′).

4.4.2 *In vivo* spacer acquisition and *in vitro* integration assays

The *in vivo* acquisition assays were performed as described in Chapter 2. The *in vitro* integration reactions were conducted as previously described in Chapter 3 with slight modifications. After pre-incubation of equimolar Cas1 and Cas2 at 4 °C, 100 nM of the resulting Cas1–Cas2 complex was incubated with 100 nM protospacer DNA for an additional 10–15 min at room temperature. The integration reaction was activated by the addition of 300 ng (~5 nM) pCRISPR, incubated at 37 °C for 1 h and quenched with DNA loading buffer supplemented with EDTA at a final concentration of 20 mM. The reaction products were analyzed on 1.5% agarose gels. Percent integration activity values were determined by quantifying the band intensity of the relaxed pCRISPR product and dividing over the intensity of all bands detected by Image Lab Software (Bio-Rad). We note that the integration activity could be a mixture of half-site and full-site integration products, as described in Chapter 3.

4.4.3 Cas1–Cas2 complex formation with protospacer DNA

Purified Cas1 and Cas2 were incubated with protospacer DNA at equimolar concentrations (50 μM) in Buffer A (500 mM KCl, 20 mM HEPES-NaOH, pH 7.5, 1 mM DTT, 10 mM EDTA), followed by overnight dialysis at 4 °C against Buffer B (100 mM KCl, 20 mM HEPES-NaOH, pH 7.5, 1 mM DTT, 5 mM EDTA). The dialyzed sample was applied on a Superdex 75 10/300 column (GE Healthcare) in Buffer B. Peak fractions were pooled and concentrated to ~3 mg ml^{-1} for crystallization. Optimized crystals were grown by hanging-drop vapor diffusion at room temperature in two different conditions, as described in the text. The Mg^{2+}-containing crystals grew as gem-like morphologies in 50 mM MES, pH 6.1, 10% isopropanol and 20 mM $MgCl_2$. The ‘no Mg^{2+} crystals’ grew as rods in 100 mM sodium citrate tribasic pH 5.6, 200 mM sodium acetate and 8% PEG 8000 (w/v). The crystals were briefly transferred into a drop containing either 25% ethylene glycol (with Mg^{2+} crystals) or 30% glycerol (without Mg^{2+} crystals) for

cryoprotection and frozen in liquid nitrogen. The Cas1–Cas2 complex with a splayed DNA substrate crystallized in the same conditions as the no Mg^{2+} crystals.

4.4.4 Crystallization and structure determination

X-ray diffraction data were collected under cryogenic conditions at beamline 8.3.1 at the Lawrence Berkeley National Laboratory Advanced Light Source. Initial phases were obtained by sequential molecular replacement using individual protein components of the Cas1–Cas2 apo structure (PDB 4P6I) as search models. Following initial placement of two Cas1 dimers and a dimer of Cas2, phases were improved by performing one round of rigid body refinement in PHENIX[70]. The resulting maps showed clear unbiased density for protospacer DNA, and subsequent model building was performed through iterative rounds of building in Coot[71] and refinement in PHENIX with NCS restraints on the protein subunits. The asymmetric unit of the three structures contains one copy of the Cas1–Cas2 complex bound to protospacer DNA. Statistics for the final crystal structures are reported in **Table 4.1**. The final structures are missing clear density for the loop connecting α6 and α7 of Cas1. We assume this loop to be highly disordered as it is also not observed in the apo *E. coli* Cas1 crystal structure (PDB 3NKD) and the apo Cas1–Cas2 complex (PDB 4P6I)[23,35].

4.4.5 Fluorescence polarization

Fluorescence polarization assays were performed in 20 mM HEPES-NaOH, pH 7.5, 25 mM KCl, 5 mM EDTA, 1 μg ml^{-1} BSA and 1 mM DTT. Cas1–Cas2 were complexed and purified over gel filtration for all binding assays. The 3′–fluorescein labeled DNA substrate was added to the protein solution at a final concentration of 5 nM and the DNA–protein mixture was allowed to incubate for 30 min at 22 °C. Measurements were made by excitation at 485 nm and monitoring emission at 535 nm. Data were fit to a binding isotherm to obtain K_d. Each experiment was conducted in triplicates and error bars represent the standard deviation.

4.4.6 Cas1 sequence alignment

The *cas1* sequences were obtained from the National Center for Biotechnology Information (NCBI) Gene Data Bank. A representative *cas1* from each CRISPR Type I subtype were chosen based on previous subtype assignments and the alignment was generated using MAFFT[18,101]. The organisms chosen for the alignment are: *Escherichia coli K-12, Cronobacter dublinensis 582, Erwinia amylovora, Yersinia pestis biovar Antiqua str. B42003004, Yersinia kristensenii, Hafnia alvei, Sulfolobus solfataricus, Thermotoga maritime, Pseudothermotoga lettingae, Deferribacter desulfuricans, Desulfovibrio vulgaris, Bacillus halodurans, Bacillus cereus, Synechocystis sp. PCC 6803, Cyanothece sp. PCC 8802* and *Limnoraphis robusta.*

4.5 Accession Code

Atomic coordinates and structure factors for the reported crystal structures have been deposited at the Protein Data Bank under accession codes 5DS4 (no Mg^{2+}), 5DS5 (with Mg^{2+}) and 5DS6 (splayed DNA).

4.6 Acknowledgements

We thank George Meigs and the 8.3.1 beamline staff at the Advanced Light Source for assistance with data collection and members of the Doudna laboratory for comments and discussions. The 8.3.1 beamline is supported by UC Office of the President, Multicampus Research Programs and Initiatives grant MR-15-328599 and Program for Breakthrough Biomedical Research, which is partially funded by the Sandler Foundation. This project was funded by a US National Science Foundation grant No. 1244557 to J.A.D. and by NIH grant AI070042 to A.N.E. J.K.N. and L.B.H. are supported by US National Science Foundation Graduate Research Fellowships and J.K.N. by a UC Berkeley Chancellor's Graduate Fellowship. P.J.K. is supported as a Howard Hughes Medical Institute Fellow of the Life Sciences Research Foundation. J.A.D. is an Investigator of the Howard Hughes Medical Institute and a member of the Center for RNA Systems Biology.

4.7 Author Contributions

These authors contributed to the data presented in this chapter: James K. Nuñez, Lucas B. Harrington, Philip J. Kranzusch, Alan N. Engelman and Jennifer A. Doudna. J.K.N. and L.B.H. conducted the crystallography, biochemistry and *in vivo* spacer acquisition assays. J.K.N., L.B.H. and P.J.K. collected the X-ray diffraction data and determined the crystal structures. J.K.N., L.B.H., P.J.K., A.N.E. and J.A.D. designed the study, analyzed all data and wrote the manuscript.

CHAPTER 5

A host factor essential for CRISPR immunological memory

A portion of the work presented in this chapter has been published previously as part of the following paper: Nuñez, J.K., Bai, L., Harrington, L.B., Hinder, T.L. & Doudna, J.A. A host factor essential for CRISPR immunological memory. *Molecular Cell* (2015).

5.1 Introduction

Bacteria and archaea use clustered regularly interspaced short palindromic repeats – CRISPR associated protein (CRISPR–Cas) adaptive immune systems to detect and destroy foreign nucleic acids[3]. Within CRISPR loci, ~30-base pair (bp) spacer sequences frequently derived from foreign nucleic acids are flanked by ~25–50 bp repeating sequences in the host genome[7–9]. Directly upstream of the repeat-spacer arrays is an A-T-rich leader sequence that typically spans 100–300 bp[6]. During viral infection or plasmid introduction, cells integrate new spacers captured from foreign DNA into the host CRISPR locus to generate an immunological memory of each invader. The CRISPR locus is subsequently transcribed and processed to produce mature CRISPR RNA (crRNA) guides that assemble with Cas proteins to cleave foreign nucleic acids containing sequence complementarity to the crRNA spacer sequence[4,17].

Although many aspects of crRNA maturation and crRNA-guided interference by Cas proteins have been elucidated, the basic principles underlying new spacer acquisition are only beginning to be understood[29]. In *Escherichia coli*, Cas1 and Cas2, the only Cas proteins found in all CRISPR–Cas systems, together form the integrase complex responsible for capturing 33 bp segments of foreign DNA, or protospacers, as potential spacers[20–22,35,37,39]. During subsequent protospacer integration, the Cas1–Cas2 complex catalyzes direct nucleophilic attack of the 3′–OH ends of the DNA substrate at the leader end of the CRISPR locus[57,58].

In previous work, purified Cas1–Cas2 was found to integrate 33 bp protospacers into a supercoiled plasmid encoding the target CRISPR locus. While these reactions showed the high specificity of Cas1–Cas2 for recognizing the CRISPR locus, integration occurred adjacent to every CRISPR repeat sequence[58]. In contrast, *in vivo* protospacer integration in *E. coli* only occurs adjacent to the first repeat that borders the leader[20–22,33,37]. We hypothesized that a missing factor specifies leader-specific spacer integration by the *E. coli* Cas1–Cas2 integrase. Here we uncover the essential role of the integration host factor (IHF) protein in leader-proximal spacer acquisition during CRISPR-Cas adaptive immunity and the requirement of target DNA bending for Cas1–Cas2-mediated spacer integration. IHF binds to the leader sequence and induces a sharp DNA bend, allowing the Cas1–Cas2 integrase to catalyze the first integration reaction at the leader-repeat border. Together, these results reveal that Cas1–Cas2-mediated spacer integration requires IHF-induced target DNA bending and explain the elusive role of CRISPR leader sequences during spacer acquisition.

5.2 Results

5.2.1 IHF specifically binds the CRISPR leader

In considering possible host factors that could contribute to CRISPR spacer acquisition, IHF stands out as a highly expressed heterodimer belonging to a family of bacterial histone-like architectural proteins. Upon binding to its target sequence, IHF induces a ~160° DNA bend that contributes to many biological processes including DNA replication, transcription, site-specific recombination, and transposition[102–104]. In contrast to other architectural proteins that bind DNA non-sequence specifically, IHF recognizes

specific A-T rich sites containing the consensus sequence 5′–WATCAANNNNTTR–3′, where W is A or T, R is A or G, and N is any nucleotide[104–108]. We noticed that the leader sequence of the *E. coli* BL21 CRISPR-I locus contains a near-cognate IHF binding site and a poly(A)-tract at the 5′ end present in many known IHF binding sites (**Fig. 5.1a**). The IHF recognition sequence lies within the 60 bp leader segment of the CRISPR-I locus required for spacer acquisition in the BL21 strain[20].

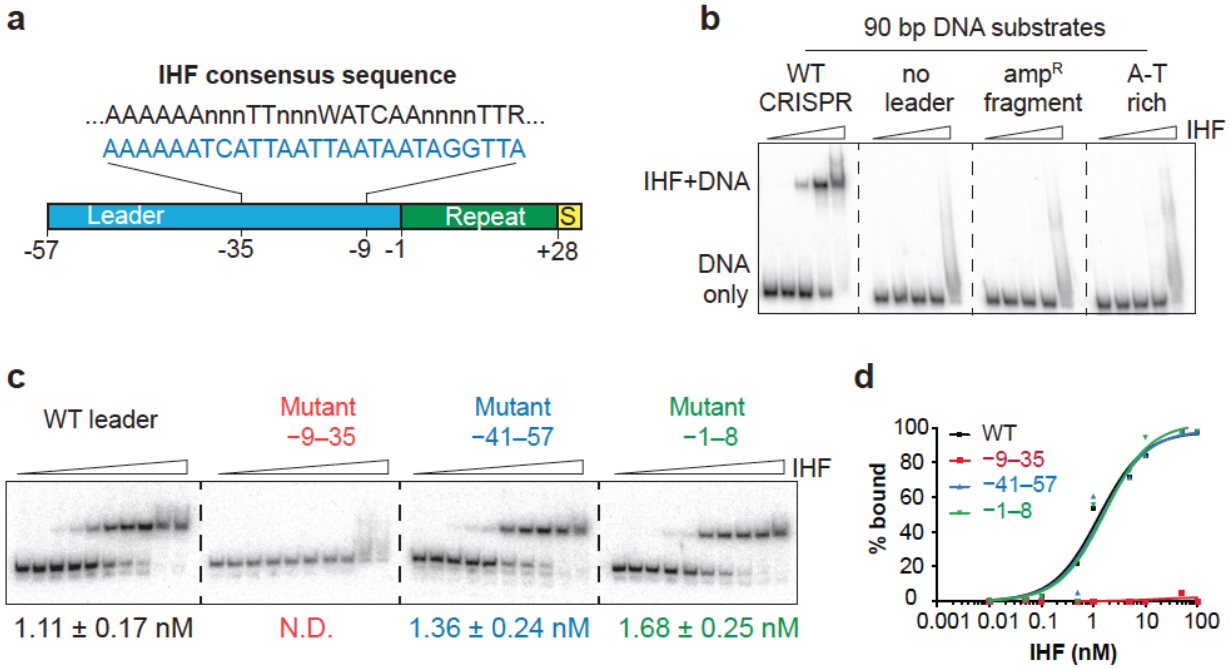


Figure 5.1 | IHF specifically binds the CRISPR leader. (**a**) Diagram of the IHF consensus binding sequence (black) and the IHF binding site in the CRISPR leader (blue). The 28 bp direct repeat is shown in green, followed by a portion of the adjacent spacer (S), shown in yellow. (**b**) Gel mobility shift assays of IHF binding to 90 bp DNA constructs with 5′-phosphate radiolabels. The "no leader" construct contains the repeat sequence and a non-A-T-rich 57 bp DNA sequence that does not resemble the CRISPR leader. The amp^R fragment and A-T rich probes are derived from the pUC19 backbone. The sequences are shown in Table 5.1. (**c,d**) Representative gel shifts and quantification of IHF binding to DNA with various mutations in the CRISPR leader. The measured affinities for each construct are reported as an average from three independent experiments with SD values. N.D. denotes not determined.

To test whether IHF binds to the BL21 CRISPR-I leader sequence, we performed native gel mobility shift assays with purified *E. coli* IHF. These experiments showed that IHF binds with high affinity to a DNA construct containing 57 bp of the leader, the 28 bp repeat, and 5 bp of spacer-1 (**Fig. 5.1b**). Binding was not observed when the leader was replaced with a non-A-T-rich sequence, confirming that IHF does not recognize the CRISPR repeat (**Fig. 5.1b**). In addition, IHF does not bind A-T-rich fragments from either the *amp^R* promoter or the plasmid backbone of pUC19, confirming its sequence specificity for the CRISPR leader (**Fig. 5.1b**). Randomized mutation of the IHF binding sequence (−9–35, **Fig. 5.1a**) completely abolished binding, whereas mutations outside

of the binding site did not affect binding (**Fig. 5.1c,d**). Hydroxyl radical footprinting experiments confirmed that IHF protects the three signature regions on the native DNA corresponding to the three minor grooves bound by IHF, whereas DNA containing the −9–35 mutation did not produce protected fragments (**Fig. 5.2a-c**). Thus, IHF binds specifically to a consensus sequence in the CRISPR-I leader.

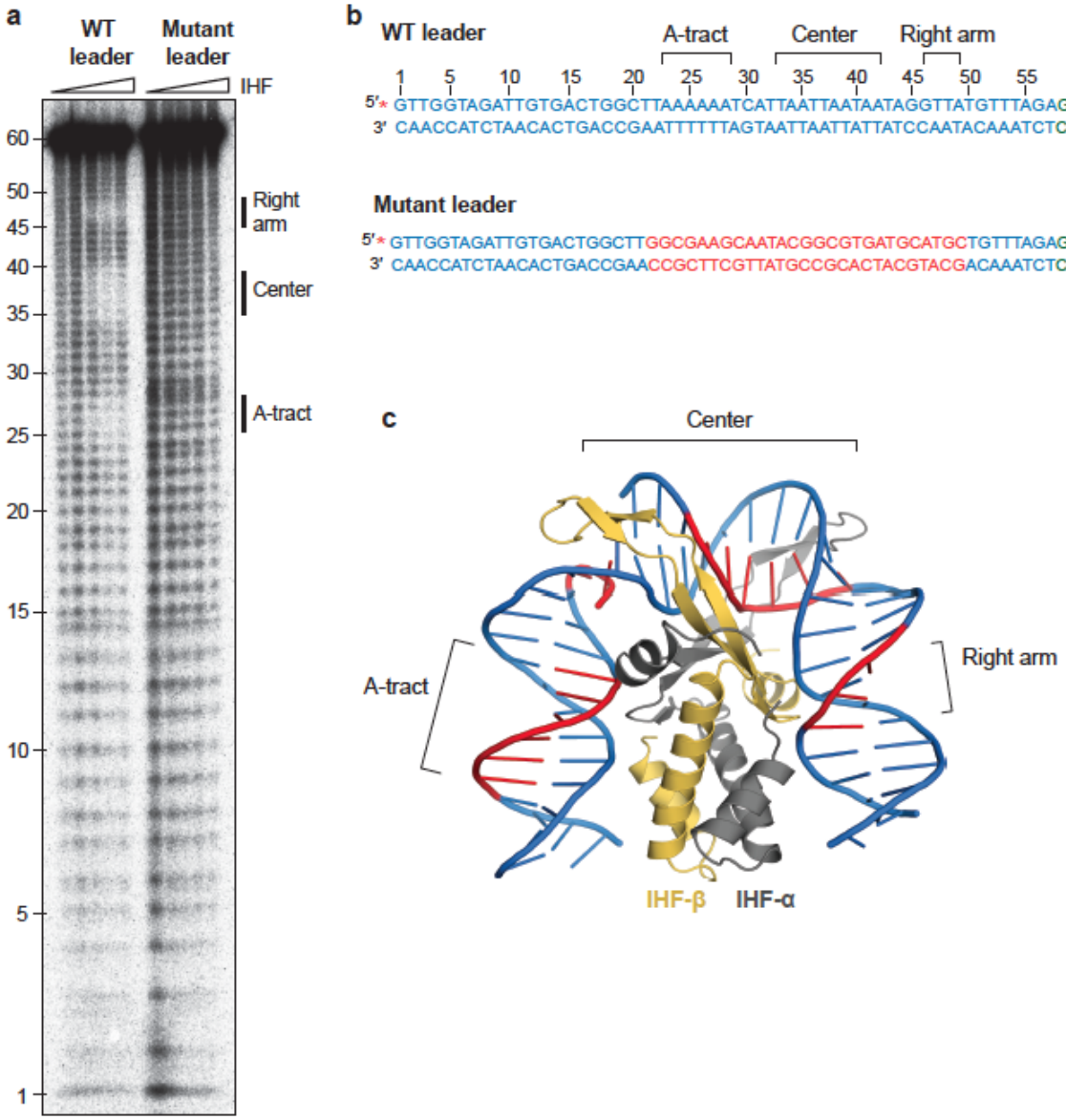


Figure 5.2 | IHF binds to a consensus sequence in the leader. (**a,b**) Hydroxyl radical probing of IHF binding to either the CRISPR leader or a mutant −9–35 sequence that disrupts IHF binding. The sequences used are shown in (b) with mutations indicated in red. The numerical annotation begins at the 5′ radiolabeled end of the top strand. (**c**) Model of the protection sites within a previously determined crystal structure of an IHF-DNA complex (PDB 1IHF). The blue depicts the leader and the red highlights the protected regions on the CRISPR leader by IHF.

5.2.2 IHF is required for *in vivo* spacer acquisition

To determine whether IHF is required for CRISPR spacer acquisition, we constructed IHF knockout (KO) strains of *E. coli* BL21 by introducing nonsense mutations in the open reading frames of the IHF subunits *α* and *β* (**Fig. 5.3a,b**). We constructed individual subunit KO strains since the IHF-αβ heterodimer is the functional oligomer *in vivo*[109]. Consistent with IHF functioning as a nonessential protein, the KO strains exhibited delayed growth phenotypes but otherwise grew to OD levels comparable to the WT strains (**Fig. 5.3c**). PCR amplification of the CRISPR locus after Cas1 and Cas2 expression showed that the KO strains were completely defective in acquiring spacers, even after diluting the cultures into fresh media for another growth cycle (Cycle 2, **Fig. 5.4a**). Spacer acquisition was restored when a plasmid expressing IHF (pIHF) was transformed into each KO strain, highlighting the critical role of IHF in spacer acquisition (**Fig. 5.4a**, last two lanes). High throughput sequencing of the KO strain genomic DNA did not detect reads mapping to spacer acquisition into the CRISPR locus nor elsewhere in the genome, further supporting the high specificity of Cas1–Cas2 to integrate spacers into the CRISPR locus in the presence of IHF (**Fig. 5.4b**). These results also indicate that other architectural proteins in *E. coli*, such as HU, HNS and Fis, cannot substitute for the role of IHF in spacer acquisition.

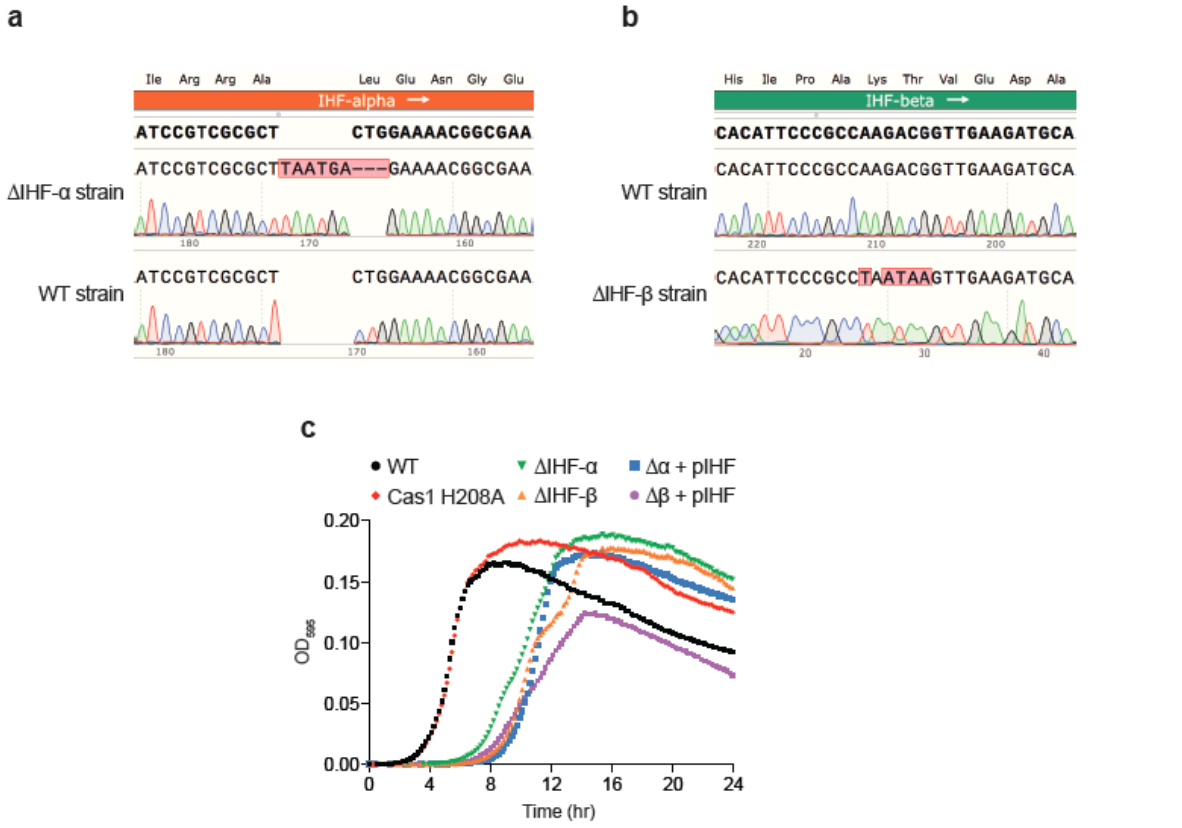


Figure 5.3 | IHF knockout strain construction. (**a,b**) Sequence traces of the (a) Δ*ihf-α* and (b) Δ*ihf-β* open reading frames in *E. coli* BL21 knockout strains used for spacer acquisition assays in Fig. 5.4. The donor DNA used in the Cas9-mediated knockout construction consisted of two adjacent stop codons. The sequence trace of the WT strain is also depicted for comparison. (c) Growth curve of the strains used in spacer acquisition assays in Fig. 5.4. The plots are an average of three independent experiments.

To determine if a mutant IHF binding site in the leader affects spacer acquisition *in vivo*, we tested spacer acquisition into plasmids encoding either the WT leader-repeat-1 sequence or the leader-repeat-1 sequence containing the −9–35 mutation used in **Fig. 5.1c**. PCR amplification of the plasmid encoding the WT CRISPR locus resulted in functional spacer acquisition, whereas the −9–35 leader mutant was completely inactive (**Fig. 5.4c**). Thus, loss of IHF binding to the CRISPR leader inactivates spacer acquisition *in vivo*. Shuffling of the −1–20 region of the leader has also been shown to be inactive for spacer acquisition *in vivo*[20], presumably due to the inability for IHF to bind to the leader and loss of sequence-specific recognition of the leader-repeat border as described below.

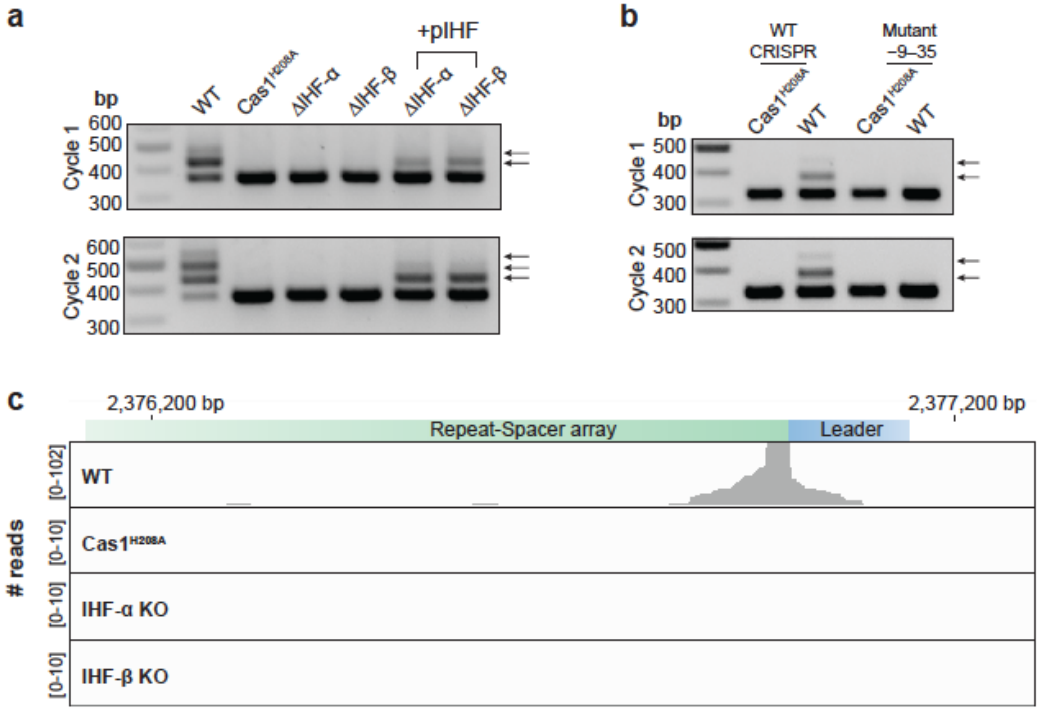


Figure 5.4 | IHF is required for spacer acquisition *in vivo*. (**a**,**b**) PCR amplification of the genomic BL21 CRISPR-I locus after expressing Cas1 and Cas2 in *E. coli* BL21 WT (lane 1) or *ihf α* or *β* knockout strains (lanes 3-6). The catalytically inactive Cas1 H208A and Cas2 were expressed in the WT strain as a negative control (lane 2). Expressing IHF from a plasmid (+pIHF, lanes 5 and 6) in the knockout strains reactivates spacer acquisition. The black arrows point to bands representing expansion of the CRISPR locus. Cycle 2 is a dilution of Cycle 1 into fresh media to allow for another growth cycle. (**b**) Mapped reads of spacer acquisition into the BL21 genomic CRISPR-I locus after expressing WT Cas1+Cas2 or Cas1^{H208A}+Cas2 in WT BL21 cells (top two rows) or WT Cas1+Cas2 in IHF-α or IHF-β BL21 knockout strains (bottom two rows). Each mapped read contains a covalent junction between the genome and the plasmid used for Cas1 and Cas2 overexpression, as described in the Methods. The schematic of the CRISPR locus is drawn above the plots. (**c**) PCR amplification of plasmid DNA encoding either the WT CRISPR-I locus or the CRISPR-I locus with a −9–35 mutation in the leader after expression of Cas1 and Cas2 in BL21 cells. The Cas1^{H208A} lanes represent expression of the Cas1 H208A catalytic mutant with WT Cas2 as a negative control.

5.2.3 IHF specifies leader-proximal integration

We next examined the mechanistic role of IHF in spacer acquisition. Previously, we reconstituted spacer integration *in vitro* by incubating purified Cas1–Cas2 with a 33 bp protospacer and adding a supercoiled target plasmid encoding the CRISPR locus (pCRISPR) to activate the reaction. Protospacer integration occurred adjacent to every repeat in the CRISPR locus, revealing the sequence specificity of Cas1–Cas2[58]. However, *in vivo* studies in *E. coli* BL21 and K12 strains have shown that spacers are acquired strictly adjacent to the first repeat that borders the leader[20–22]. Strikingly, when IHF is added to *in vitro* integrations, the majority of integration events within the CRISPR locus now occur adjacent to the first repeat. This result suggests that IHF directs Cas1–Cas2 to preferentially recognize the leader end of the locus (**Fig. 3a-c**).

Since the cellular role of IHF is DNA bending, we reasoned that IHF-induced bending at the leader-end of the CRISPR locus prepares the target DNA for Cas1–Cas2-mediated protospacer integration. In the presence of IHF, we detected a slight increase in total protospacer integration into pCRISPR *in vitro* (**Fig. 3d,e**). The level of relaxed pCRISPR topoisomers, which are not substrates for integration, substantially decreased in the presence of IHF (Band X, **Fig. 3d**). Thus, IHF binding and bending of relaxed plasmid byproducts likely reactivates these plasmids for integration. To test this possibility, we performed integration assays into linearized pCRISPR using radiolabeled protospacers. Whereas previous experiments, performed in the absence of IHF, did not detect robust integration into linear DNA[58], we now observe a distinct band corresponding to protospacer integration into the linear target in the presence of IHF (**Fig. 3f**). This result is consistent with IHF inducing necessary topological constraints on the CRISPR leader for Cas1–Cas2-mediated protospacer integration.

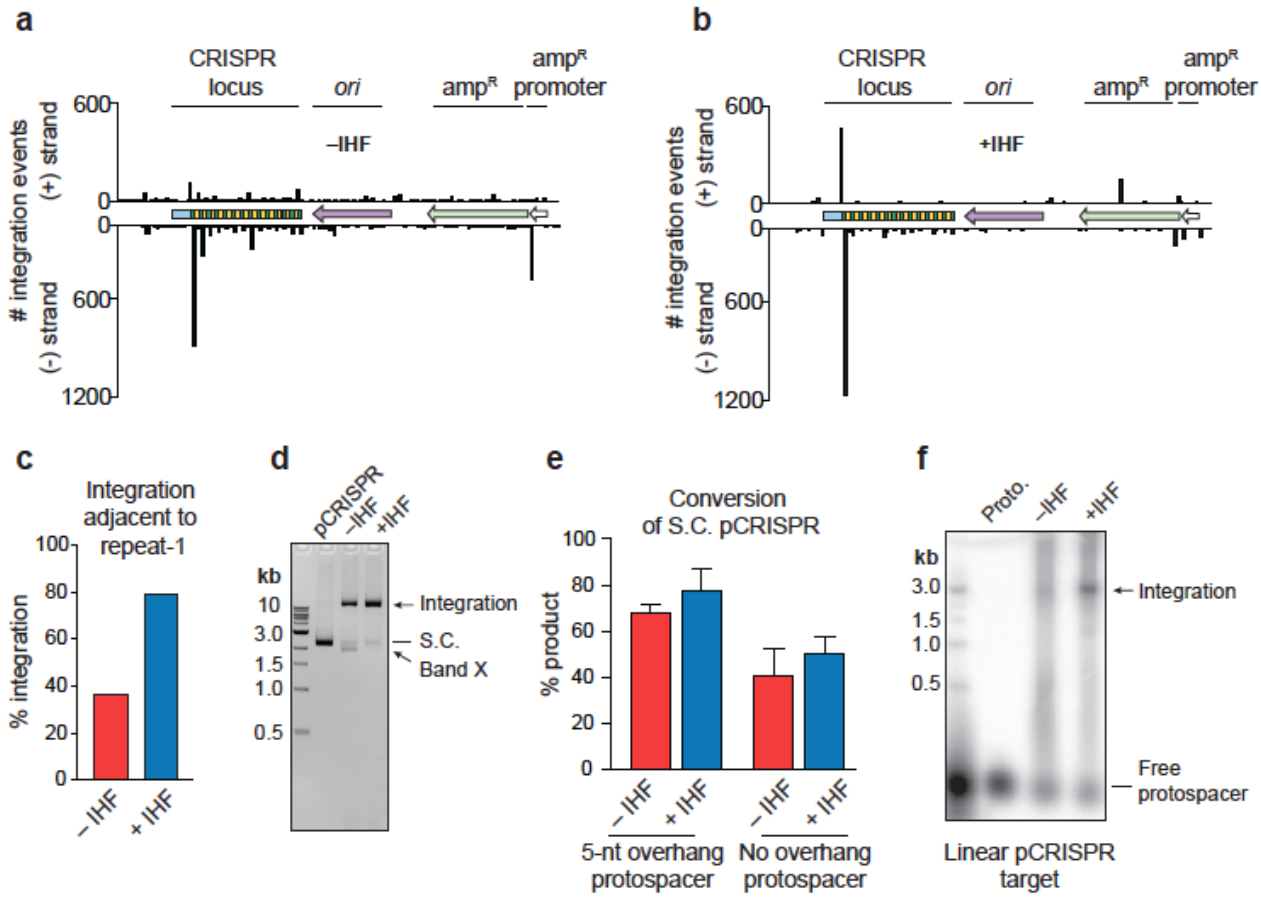


Figure 5.5 | IHF directs leader-proximal spacer integration. (**a,b**) Plots of integration sites along a plasmid encoding the CRISPR locus, as revealed by high throughput sequencing of the *in vitro* integration products in the (a) absence or (b) presence of IHF. The notable regions of the plasmids are annotated above the plots. The CRISPR locus contains the leader (blue), repeats (green), and spacers (yellow). (**c**) Quantification of percent integration events adjacent to repeat-1 over the rest of the CRISPR locus in sequencing libraries with (red) or without (blue) IHF. (**d**) A representative gel of *in vitro* integration reactions using Cas1, Cas2, protospacer DNA and pCRISPR target DNA in the presence or absence of IHF. The arrows point to the integration products and "Band X" side products that are collections of relaxed pCRISPR topoisomers (Nuñez 2015b). S.C. denotes the supercoiled plasmid substrate. (**e**) Quantification of the integration reactions in (**d**) showing the conversion of the supercoiled pCRISPR and Band X species to the integration product in the absence (red) or presence (blue) of IHF. The 5-nt overhang protospacers contain five nucleotide overhangs at the 3′ ends, which are highly preferred by Cas1–Cas2, compared to a 33 bp protospacer with no overhangs (Table 5.1). The values were obtained from three independent experiments. Error bars represent standard deviation. (**f**) Representative gel of *in vitro* integration of 5′ radiolabeled protospacers into a linearized pCRISPR. The arrow points to the integration product.

5.2.4 Sequence-specific integration into leader-repeat border

The observation that IHF mediates leader-specific spacer acquisition motivated a set of experiments to test the order of events leading to spacer integration. To create minimal DNA integration target sites, we designed 90 bp DNA constructs containing 57 bp of the CRISPR-I leader from *E. coli* strain BL21, 28 bp of the repeat, and 5 bp of spacer-1. After incubating the linear target DNA with Cas1–Cas2, IHF and ^{32}P-labeled protospacers, we detected two products corresponding to the expected integration reactions: a 61 nucleotide (nt) product corresponding to integration at the leader-repeat border (top strand) and a 113 nt product from the repeat-spacer integration event (bottom strand; **Fig. 5.6a,b**). The formation of these products is stimulated in the presence IHF (**Fig. 5.6b**). In addition, a shuffled −1–20 leader sequence is only partially active, and the absence of the leader or presence of a leader containing the −9–35 mutant IHF binding site substantially decreases protospacer integration adjacent to the repeat (**Fig. 5.6b**). Thus, efficient integration requires prior IHF engagement of the CRISPR leader and bending of the target DNA.

We next measured the kinetics of product formation to gain insights into the order of integration reactions. We detected faster formation of the integration product at the leader-repeat junction than at the repeat-spacer junction, suggesting that integration first occurs at the leader-end of the repeat (**Fig. 5.6c**). Interestingly, accumulation of the leader-repeat integration product decreases after 5 min, reflecting the occurrence of the reverse reaction (disintegration) of the newly integrated protospacer from the half-site intermediate as previously described[58,61]. The rapid disintegration of the top strand intermediate and slower disintegration of the bottom strand explains the higher accumulation of bottom strand intermediate in previous deep sequencing analyses, which were quenched at 1 hr[58]. The initial rate of integration into the leader-repeat border is ~14 fold higher than that occurring at the repeat-spacer border, demonstrating that the leader-repeat integration reaction occurs first (**Fig. 5.6d-f**; black plots).

Use of a target DNA with sequence mutations at the leader-repeat border resulted in a ~3-fold decrease in the initial integration rate into the top strand but no effect in integration into the bottom strand, highlighting the ability for Cas1–Cas2 to recognize the sequence of the leader-repeat border. (**Fig. 5.6d,e**; red plots). In contrast, mutations at the repeat-spacer border had no effect on either integration event (**Fig. 5.6d,e**; blue plots). We tested *in vivo* spacer acquisition into a plasmid encoding either the WT CRISPR locus, mutations at the leader-repeat border, or mutations at the repeat-spacer border. Consistent with the *in vitro* results, the leader-repeat mutation abolishes spacer acquisition whereas the the repeat-spacer mutation remains active. (**Fig. 5.6f**). Together with disintegration[61] and *in vivo* studies in *E. coli*[33], Cas1–Cas2 relies on sequence specificity at the leader-repeat border but allows degeneracy at the repeat-spacer end.

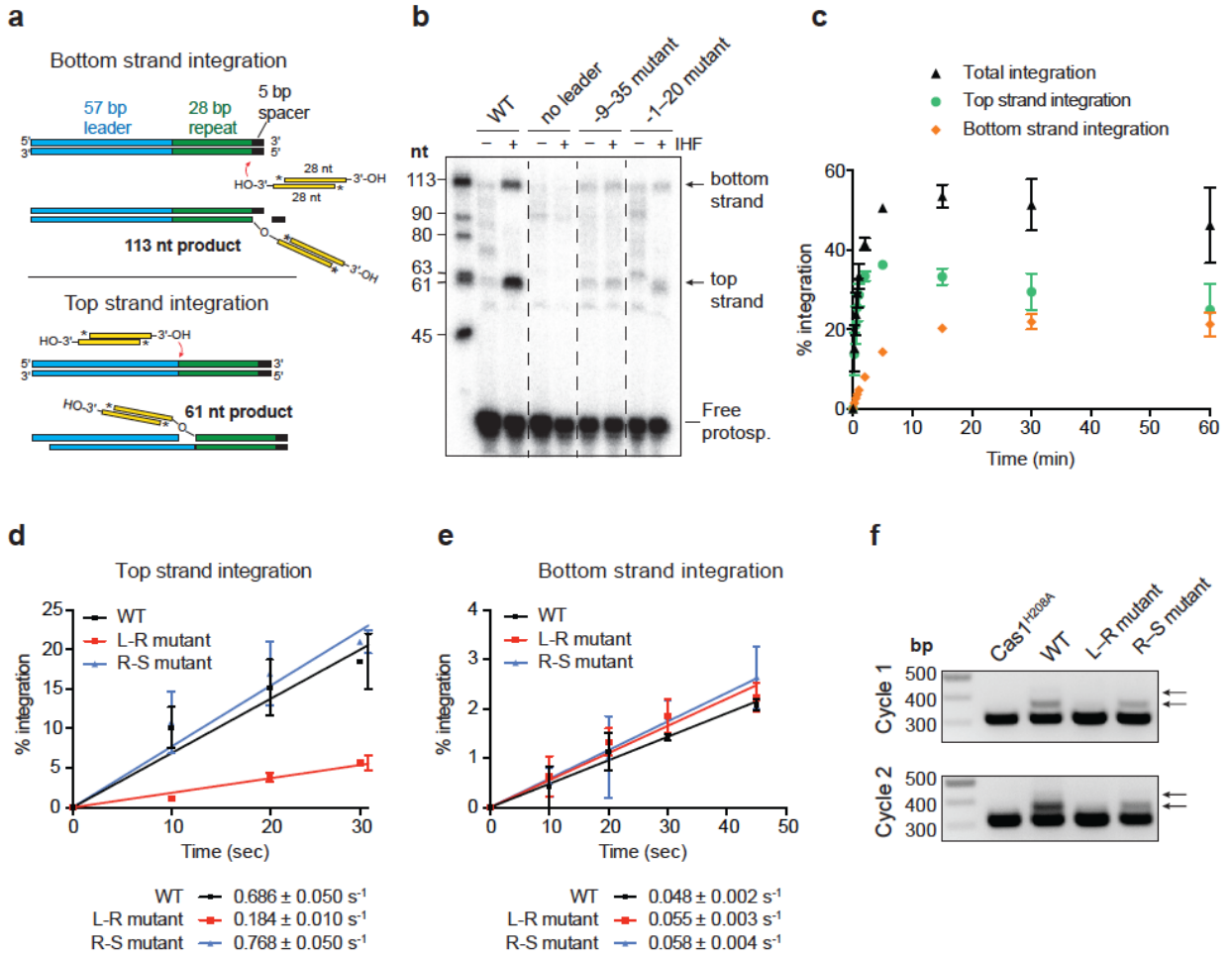


Figure 5.6 | IHF directs leader-proximal spacer integration. (**a**) Cartoon of the two integration events into the target CRISPR DNA. The bottom strand integration corresponds to integration into the repeat (green) - spacer (black) border. The top strand represents integration into the leader (blue) - repeat border. The protospacer used contains 5-nt overhangs on both sides, as shown in Table 5.1. (**b**) Representative gel of *in vitro* integration reactions using various linear CRISPR targets with the indicated mutations in the leader. The arrows point to bands corresponding to the two integration events depicted in (**a**). (**c**) Plot of percent integration of protospacers into the WT CRISPR target DNA over 1 hr. The black plots correspond to the sum of the top (green plots) and bottom (orange plots) integration events. Error bars represent standard deviation. (**d,e**) Plots of the initial rates of product formation within the linear range (0–30 sec) of the integration reaction measured at 25°C. The black plots correspond to reactions using the WT CRISPR DNA. The red plots are from reactions using a target DNA with mutations in the leader-repeat region (L-R), whereas the blue plots are from target DNA containing mutations in the repeat-spacer region (R-S). The quantified average rates are noted below with SD values from three independent experiments. (**f**) *In vivo* spacer acquisition assays into a plasmid encoding either the WT CRISPR locus, mutations at the leader-repeat border (L-R), or mutations at the repeat-spacer border (R-S). The Cas1^{H208A} lane depicts expression of Cas1 H208A catalytic mutant with Cas2 with a plasmid encoding the WT CRISPR locus.

5.2.5 Target DNA bending drives spacer integration

We hypothesized that in the absence of IHF, Cas1–Cas2 requires supercoiled target DNA to gain access to the scissile phosphate bonds for integration. Target DNA deformations have been observed previously in high resolution structures of DNA transposition and retroviral integration enzymes[110–113]. To test whether bending of the CRISPR repeat DNA is sufficient to trigger protospacer integration irrespective of the leader and IHF, we cloned the inactive BL21 CRISPR-II locus into a plasmid that also contains the active CRISPR-I locus and tested for protospacer integration by Cas1–Cas2 in the absence of IHF. In contrast to CRISPR-I, CRISPR-II lacks an A-T-rich leader sequence upstream of the repeats and does not acquire new spacers *in vivo*[20]. Strikingly, the CRISPR-II repeats are active for integration within the plasmid context, similar to the CRISPR–I repeats (**Fig. 5.7a**).

We wondered whether the CRISPR-II locus is inactive *in vivo* due to insufficient DNA bending directly upstream of the repeats. To test this possibility, we replaced the −9–35 bp region of CRISPR-II with the IHF binding site from CRISPR-I and tested for spacer integration *in vitro* in the presence of Cas1–Cas2 and IHF (**Fig. 5.7b**; CRISPR-II*). Whereas the linear native CRISPR-II locus did not support integration adjacent to the repeat with or without IHF, the hybrid construct containing the −9–35 bp IHF binding site gave rise to the expected integration products in the presence of IHF (**Fig. 5.7c**). We further confirmed by gel shift assays that the activation is due to IHF binding (**Fig. 5.7d**; middle lanes). Together, these results reveal that IHF-induced bending of the CRISPR leader sequence is a critical step in spacer acquisition. As discussed further below, we predict that sequence-specific positioning of the integration target sites within the Cas1 active sites requires target DNA deformations that can only be accessed when the DNA is bent either by IHF or plasmid supercoiling.

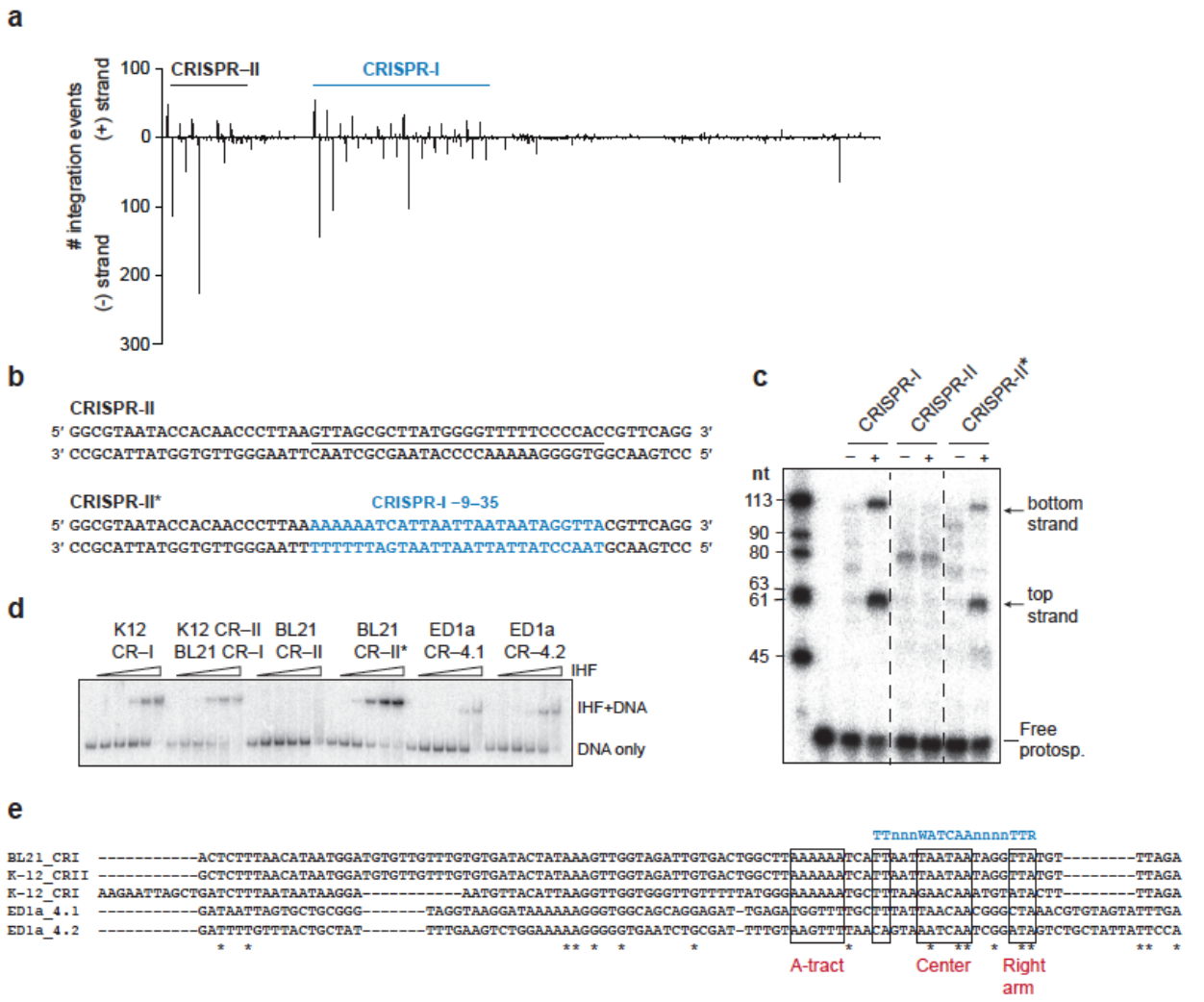


Figure 5.7 | CRISPR DNA bending drives protospacer integration. (**a**) Plot of integration sites along a plasmid encoding the CRISPR-II locus that lacks a leader sequence *in vivo* (purple) and the CRISPR-I locus that contains a leader sequence (green). (**b**) Leader sequence of the native *E. coli* BL21 CRISPR-II and a hybrid CRISPR-II leader with the −9–35 region from CRISPR I (labeled CRISPR-II*). (**c**) Representative gel of *in vitro* integration into CRISPR-I, CRISPR-II, and CRISPR-II* showing the activation of the hybrid CRISPR-II* locus as target for protospacer integration. The reaction was quenched after 15 min at 37°C. (**d**) Gel mobility shift assays of IHF binding to various leader sequences from CRISPR loci in three different *E. coli* strains – K12, BL21, and ED1a. (**e**) Sequence alignment of the leader sequences in (**d**). The boxes highlight the IHF protection sites shown in Fig. 5.2c and the consensus sequence is noted in blue.

5.3 Discussion

A critical step in CRISPR–Cas adaptive immunity is the acquisition of new foreign DNA-derived spacers that provide the host with a genetic record of its encounters with pathogens and plasmids. Our studies show that a host factor, IHF, ensures that new spacers are integrated site-specifically at the leader-proximal end of the CRISPR locus in *E. coli* (**Fig. 5.8**). We discovered the unexpected involvement of IHF in binding to a consensus site in the CRISPR leader sequence that induces DNA bending directly upstream of the first CRISPR repeat. The bent DNA provides the Cas1–Cas2 integrase access to the scissile phosphate at the leader-repeat border, enabling the first integration reaction to occur. Subsequently, the second integration reaction occurs 28 bp downstream at the repeat-spacer-1 border. Consistent with previous data[33,61], the first nucleophilic attack by the protospacer depends on base-specific readout of the leader-repeat sequence by Cas1–Cas2, whereas the second nucleophilic attack by the other end of the protospacer occurs at a site selected strictly by distance rather than sequence (**Fig. 5.8**). We do not expect IHF to participate in protein-protein contacts with Cas1–Cas2, consistent with previous studies showing its primary function as a DNA bending protein[103].

Since IHF is present only in Gram-negative bacteria[102], we predict that other organisms employ different host proteins to provide the necessary deformation of the CRISPR leader for Cas1–Cas2 recognition of the target site. In this study, we focused on the Type I-E CRISPR system of the *E. coli* BL21 strain. Gel mobility shift assays showed that IHF binds to four different CRISPR leader sequences from two other *E. coli* strains – K12 and ED1a – suggesting that IHF-mediated spacer acquisition is conserved in *E. coli* CRISPR systems (**Fig. 5.7d,e**). Alternatively, other CRISPR-harboring organisms may have evolved mechanisms to bypass the requirement for IHF or other DNA bending proteins. Instead, the Cas1–Cas2 in these organisms may access the target DNA sequence by directly inducing target DNA bending. *In vitro* reconstitution of spacer integration in other systems will provide insights into mechanistic divergence of spacer acquisition across different CRISPR pathways.

Target DNA bending during integration is a recurring theme in other mobile DNA processes, such as DNA transposition and retroviral integration, since the bent DNA resembles the integration reaction intermediate[110–116]. In CRISPR spacer acquisition, target DNA bending may induce local deformations (e.g. widening) in the major or minor grooves, allowing Cas1–Cas2 to make base-specific contacts with the leader-repeat border sequences. Alternatively, DNA bending could cause local DNA unpairing within the leader-repeat border that is subsequently recognized by Cas1–Cas2. Future high-resolution structural studies of the Cas1–Cas2-protospacer complex bound to target DNA may reveal how deformations at the integration target site enable its accommodation within the Cas1 active sites.

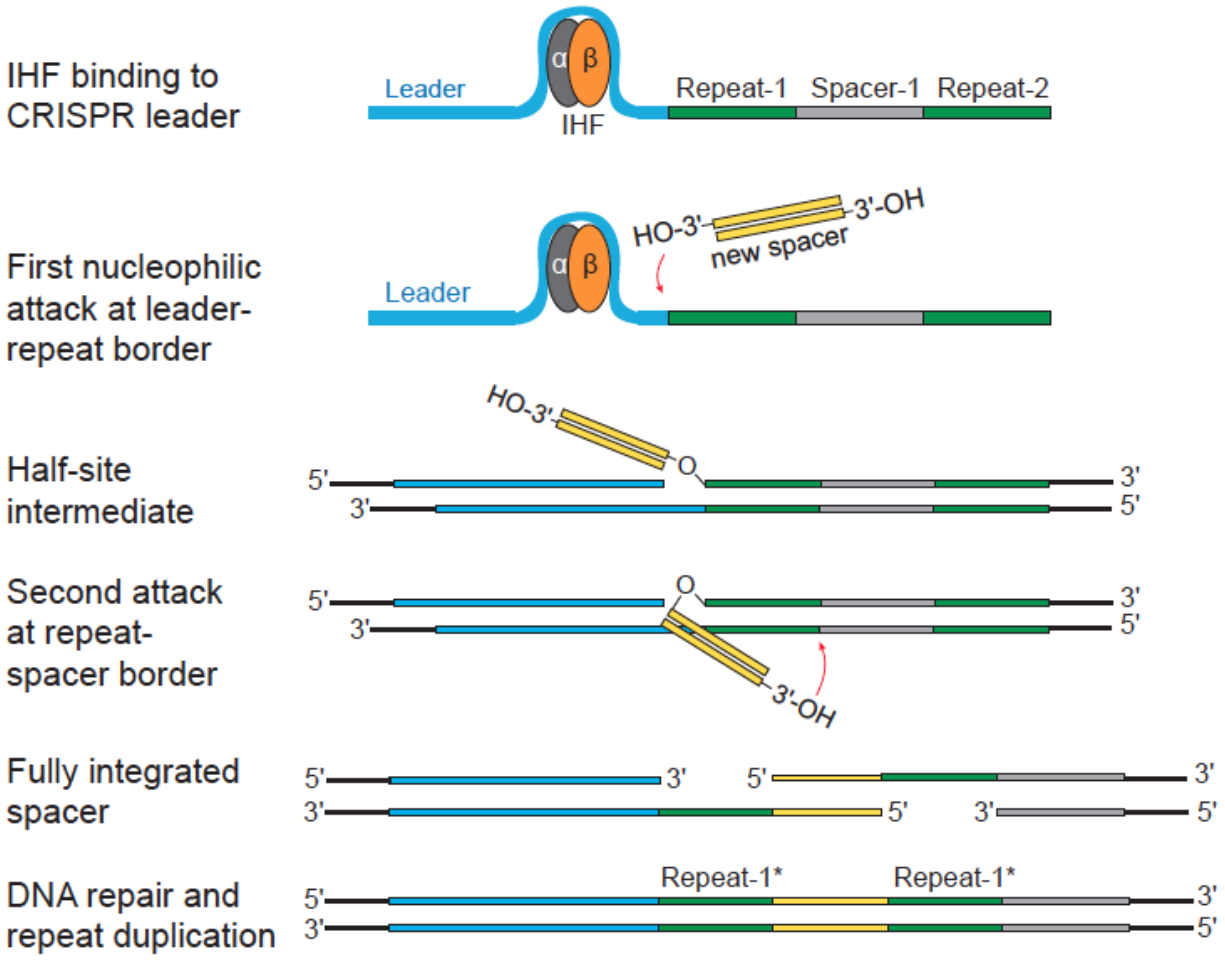


Figure 5.8 | Mechanism of spacer integration in *E. coli*. IHF (gray and orange) binds to a specific site in the leader sequence and induces a sharp DNA bend at its binding site. The bent target DNA allows Cas1–Cas2 with an incoming spacer to recognize the leader-repeat border sequence and catalyzes the first integration attack. The second integration event occurs via a ruler mechanism 28 bp downstream at the repeat-spacer-1 border. Following DNA repair, the new spacer is integrated and duplicates repeat-1 (asterisks).

5.4 Materials and Methods

5.4.1 DNA and protein preparation

Protospacer and target DNA substrates were gel-purified as single strands and annealed in 20 mM HEPES-NaOH, pH 7.5, 25 mM KCl and 5 mM $MgCl_2$. The DNA sequences used in this study are listed in **Table 5.1**. The Cas1 and Cas2 proteins from *E. coli* K12 (MG155) were cloned and purified as previously described[35]. The *ihf-α* and *ihf- β* genes were PCR amplified from *E. coli* (BL21) genomic DNA using the indicated primers in **Table 5.1** and cloned into pACYCDuet-1 (Novagen). The *ihf-α* gene was cloned to contain an N-terminal His_6 tag, followed by a TEV protease cleavage sequence. The proteins were expressed in BL21 (DE3) and grown to 0.6–0.8 OD_{600} in LB media, followed by overnight induction at 16°C with 0.5 mM isopropyl-β-D-thiogalactopyranoside (IPTG). The cells were harvested, resuspended and lysed in buffer A (500 mM KCl, 20 mM HEPES-NaOH, pH 7.4, 10 mM imidazole, 0.1% Triton X-100, 2 mM TCEP, 0.5 mM phenyl-methylsufonyl fluoride (PMSF), Complete EDTA-free protease inhibitor (Roche) and 10% glycerol). The lysates were centrifuged and the soluble fraction was incubated with Ni-NTA affinity resin in batch (Qiagen). The resin was washed in buffer B (500 mM KCl, 20 mM HEPES-NaOH, pH 7.4, 10 mM imidazole, 5% glycerol and 1 mM TCEP) and the bound samples were eluted in buffer B supplemented with 300 mM imidazole. The eluted sample was concentrated and TEV protease was added to remove the affinity tag at 4°C overnight. The IHF heterodimer was separated from the fusion tag on a 5 ml HiTrap Heparin HP column (GE Healthcare), eluting with a linear gradient of 0.25–1 M KCl. The peak fractions were pooled, concentrated and further purified on a Superdex 75 (16/60) size exclusion column (GE Healthcare) with buffer B in the absence of imidazole. Purified IHF was stored in 100 mM KCl, 20 mM HEPES-NaOH, 5% glycerol and 1 mM TCEP at −80°C before use.

5.4.2 Knockout strain construction, growth assays and *in vivo* acquisition assays

The *ihf* knockout strains were generated using Cas9 in combination with λ red recombinase. *E. coli* BL21-AI harboring pCas9 and pKD46 was grown in SOB media supplemented with 34 µg/ml chloramphenicol and 100 µg/ml ampicillin at 30°C [117,118]. When OD_{600} reached 0.5, the cultures were induced with 50mM L-arabinose, allowed to grow for an additional 30 minutes, and cells were made electrocompetent by washing three times with 10% glycerol pre-cooled to 4°C. Freshly prepared electrocompetent cells were transformed with a plasmid encoding a crRNA guide targeting the *ihf* open reading frames and single strand DNA donor templates containing premature stop codons[117]. The donor sequences are shown in **Table 5.1**. Transformed cells were recovered for 2 hours and grown at 30°C on plates containing 15 µg/ml chloramphenicol, 50 µg/ml kanamycin, and 100 µg/ml ampicillin. Resulting colonies were sequenced and the knockout strains were cured of pKD46 and pCas9 by consecutive restreaking on plates at 37°C.

In vivo spacer acquisition assays were performed as previously described in LB media containing 50 µg/ml streptomycin, 0.2% L-arabinose, and 0.25 mM IPTG [20]. For

the rescue experiments, the knock strains were also transformed with a plasmid expressing the *ihf* genes (pIHF) and the cultures were supplemented with 40 μg/ml streptomycin, 25 μg/ml chloramphenicol, 0.2% L-arabinose, and 0.25 mM IPTG. Cycle 1 denotes the first growth cycle (18-24 hr). A 1:600 dilution of Cycle 1 cultures were inoculated into fresh media to allow for another growth cycle (Cycle 2). The plasmid used to express the crRNA was a gift from Luciano Marraffini (Addgene plasmid #42875).

Growth curves were generating by diluting overnight starter cultures 1:600 into 100 μl LB media containing 50 μg/ml streptomycin, 0.2% L-arabinose, and 0.25 mM IPTG in a covered 96 well plate. The rescue experiments also contained 25 μg/ml chloramphenicol. Cultures were grown at 37°C with orbital shaking (218rpm and 3mm amplitude) and OD_{595} measurements were made every 10 minutes.

5.4.3 Gel mobility shift assays

Target DNA substrates were radiolabeled 5′-radiolabeled using [γ-^{32}P]-ATP (PerkinElmer) and T4 polynucleotide kinase (New England Biolabs). Binding assays were performed in buffer containing 20 mM HEPES-NaOH, pH 7.5, 50 mM KCl, 5 mM $MgCl_2$, 5% glycerol and 1 mM DTT in a total volume of 15 μl. Each reaction contains 0.05 nM radiolabeled DNA and 0.05 nM 90 bp A-T-rich competitor DNA. The leader mutants and the A-T-rich competitor sequences are listed in **Table 5.1**. The reactions were incubated at room temperature for 30 min and resolved at 4°C on 8% native polyacrylamide gels containing 0.5× TBE. DNA was visualized by phosphorimaging, quantified with ImageQuant (GE Healthcare) and analyzed with Prism (GraphPad).

5.4.4 Hydroxyl radical footprinting

The indicated radiolabeled single stranded DNA substrates in Fig. 5.2b were annealed with the cold complement. The resulting double stranded DNA was incubated with IHF in buffer containing 20 mM HEPES-NaOH, pH 7.5, 50 mM KCl, 5 mM $MgCl_2$ and 1 mM DTT for 30 min at room temperature. Hydroxyl radicals were generated using the Fenton reaction, as previously described [119]. The cutting reactions were performed at room temperature for 2 min, quenched with 4% glycerol and 0.6 M sodium acetate, followed by ethanol precipitation at −20°C. The dried DNA pellets were resupended in DNA loading dye containing 47.5% formamide and resolved on 12% denaturing polyacrylamide gels containing 7M urea. The reactions in Fig. 5.2a contain 1 nM DNA with increasing IHF concentrations (0, 0.01, 0.1, 1, 10 nM).

5.4.5 *In vitro* integration assays

Integration assays with pCRISPR target DNA were performed as previously described with slight modifications[58]. The reaction buffer contains 20 mM HEPES-NaOH, pH 7.5, 50 mM KCl, 5 mM $MgCl_2$, 10% DMSO, and 1 mM DTT. Cas1–Cas2 complex (100 nM) was incubated with 100 nM protospacer DNA while 50 nM IHF was incubated with 7.5 nM pCRISPR in a separate tube. After 10–15 min incubation at room temperature, the Cas1–Cas2-protospacer complex was added to the IHF-pCRISPR mix

and incubated at 37°C for 1 hr. The reaction was quenched with DNA loading buffer containing a final concentration of 20 mM EDTA and analyzed on 1.5% agarose gels pre-stained with ethidium bromide.

Integration assays with 5′-radiolabeled protospacer DNA were performed using 200 nM Cas1–Cas2 complex, 25 nM IHF, 20 nM protospacer and 10 nM target DNA. Cas1–Cas2 complex was incubated with protospacer DNA to allow complex formation for 10–15 min at 37°C. Concurrently, IHF was incubated with target DNA in a separate tube. To activate the reaction, the Cas1–Cas2-protospacer complex was added to the IHF-target DNA mix and quenched with DNA loading buffer for a final reaction of 12.5 mM EDTA and 47.5% formamide. The samples were heated at 95°C for 5 min and immediately loaded on 8% denaturing polyacrylamide gels. The reactions shown in **Fig. 5.6b** and **5.7c** were quenched after 15 min at 37°C and time course experiments in Fig. 5.6c were performed at 37°C. The time course experiments in **Fig. 5.6c-e** were performed at 25°C to quantify the initial rates. The points were fitted to a linear plot. Percent integration values were calculated using ImageQuant (GE Healthcare) by quantifying the band intensity of either the top or bottom strand band and dividing over the intensity of all bands in the lane. All *in vitro* integration experiments were performed a minimum of three times and error bars represent the standard deviation.

5.4.5 High throughput sequencing

The sequencing data in **Fig. 5.5a,b** were from integration reactions containing 100 nM Cas1–Cas2, 100 nM protospacer DNA with no overhangs, +/- 50 nM IHF and 7.5 nM pCRISPR. The data in **Fig. 5.7a** were from obtained from integration reactions as previously described[58]. The CRISPR-II locus was amplified from *E. coli* (BL21) genomic DNA and cloned into pUC19. The DNA libraries were prepared, sequenced on an Illumina HiSeq2500 and analyzed as previously described[58]. To calculate the percent integration adjacent to repeat-1 +/− IHF in **Fig. 5.5c**, only integration sites that occur more than five times within the CRISPR locus were taken into consideration. In **Fig. 5.4c**, the genomic DNA was subjected for Illumina sequencing after overexpression of Cas1+Cas2 in either WT or the IHF knockout strains to detect spacer acquisition into the CRISPR-I locus. Genomic DNA was isolated using Qiagen DNeasy Blood and Tissue Kit and the libraries were prepared as previously described[58]. Cutadapt (http://code.google.com/p/cutadapt/) was used to remove 3′ adapters from raw Illumina reads, discarding reads shorter than 15 nt. Spacer acquisition events were retrieved by detecting reads containing junctions between the genome and the plasmid used to overexpress Cas1+Cas2. After creating Bowtie2 indexes from fasta files of pCas1+Cas2 and the BL21 genome, the reads were first mapped to pCas1+Cas2 and retrieved reads were mapped to the BL21 genome. The resulting reads were visualized using the Integrative Genomics Viewer software.

DNA	Sequence
Cloning	
ihf-alpha, F	ACGTCGGATCCGGAAAACCTGTATTTTCAGGGCTCCATGGCGCTTACAAAAGCTGAAATGTC
ihf-alpha R	ACGTCGCGGCCGCTTACTCGTCTTTGGGCGAAGCG
ihf-beta, F	ACGTCCTCGAGACCAAGTCAGAATTGATAGAAAGACTTGCC
ihf-beta, R	ACGTCCCTAGGTTAACCGTAAATATTGGCGCGATCGC
Protospacers	
No overhang	5'–GCCCAATTTACTACTCGTTCTGGTGTTTCTCGT–3' 3'–CGGGTTAAATGATGAGCAAGACCACAAAGAGCA–5'
5-nt 3' overhangs	5'–ATTTACTACTCGTTCTGGTGTTTCTCGT–3' 3'–CGGGTTAAATGATGAGCAAGACCACAAA–5'
EMSA probes	
WT Leader	5'–GTTGGTAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGAGTG–3' 3'–CAACCATCTAACACTGACCGAATTTTTTAGTAATTAATTATTATCCAATACAAATCTCAC–5'
−1–8 mutant	5'–GTTGGTAGATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTAGTGTAATTGTG–3' 3'–CAACCATCTAACACTGACCGAATTTTTTAGTAATTAATTATTATCCAATCACATTAACAC–5'
−9–35 mutant	5'–GTTGGTAGATTGTGACTGGCTTGGCGAAGCAATACGGCGTGATGCATGCTGTTTAGAGTG–3' 3'–CAACCATCTAACACTGACCGAACCGCTTCGTTATGCCGCACTACGTACGACAAATCTCAC–5'
−41–57 mutant	5'–TGTATGTTAGGAGGTTCGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGAGTG–3' 3'–ACATACAATCCTCCAAGCCGAATTTTTTAGTAATTAATTATTATCCAATACAAATCTCAC–5'
A-T rich competitor	5'–CTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCACCTAGATCCTTTTAAATTAAA–3' 3'–GACTGCGAGTCACCTTGCTTTTGAGTGCAATTCCCTAAAACCAGTACTCTAATAGTTTTTCCTAGAAGTGGATCTAGGAAAATTTAATTT–5'
No leader	5'–CTACGCACTTCTGGCGAAGCAATACGGCGTGATGCATGCATGCAATCGATCGATCGAGTGTTCCCCGCGCCAGCGGGGATAAACCGAGCA–3' 3'–GATGCGTGAAGACCGCTTCGTTATGCCGCACTACGTACGTACGTTAGCTAGCTAGCTCACAAGGGGCGCGGTCGCCCCTATTTGGCTCGT–5'
amp^R fragment	5'–ATTCAGCTCCGGTTCCCAACGATCAAGGCGAGTTACATGATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCCTCCGATCGT–3' 3'–TAAGTCGAGGCCAAGGGTTGCTAGTTCCGCTCAATGTACTAGGGGGTACAACACGTTTTTTCGCCAATCGAGGAAGCCAGGAGGCTAGCA–5'
KO donors	
Δ*ihf-α* donor	GTTTTTCGAAGAGATCCGTCGCGCTTAATGAGAAAACGGCGAACAGGTGAAACTCTCTGG
Δ*ihf-β* donor	CTTGCCACCCAGCAATCGCACATTCCCGCCTAATAAGTTGAAGATGCAGTAAAAGAGATG

Table 5.1 | DNA sequences used for cloning, knockout strain construction and integration assay experiments.

5.5 Accession Code

The accession number for the sequencing data reported in this chapter is NCBI Gene Expression Omnibus: GSE79992.

5.6 Acknowledgements

We thank Amy Lee and David Burstein for assisting with deep sequencing data analysis and Matt Davis for providing reagents to generate the knockout strains. We are grateful to Alan Engelman (Harvard), Addison Wright, and Doudna lab members for helpful discussions. This project was funded by US National Science Foundation grant No. 1244557 to J.A.D. This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 Instrumentation Grants S10RR029668 and S10RR027303. J.K.N. and L.B.H. are supported by U.S. National Science Foundation Graduate Fellowships and J.K.N. by a UC Berkeley Chancellor's Graduate Fellowship. L.B. acknowledges support from a UC Berkeley SURF Rose Hills Fellowship. J.A.D. is an Investigator of the Howard Hughes Medical Institute. The authors have filed a related patent; J.A.D. is a co-founder and Scientific Advisory Board member of Caribou Biosciences and Intellia Therapeutics and a co-founder of Editas Medicine.

5.7 Author Contributions

These authors contributed to the data presented in this chapter: James K. Nuñez, Lawrence Bai, Lucas B. Harrington, Tracey L. Hinder and Jennifer A. Doudna. J.K.N. and L.B. performed the biochemical experiments. J.K.N., L.B., and T.L.H. analyzed the deep sequencing data. L.B.H. constructed the *ihf* knockout strains. J.K.N. and J.A.D. designed the study and wrote the manuscript.

Chapter 6

Summary and Future Directions

6.1 Mechanism of CRISPR–Cas immunological memory

Adaptive immunity in bacteria and archaea relies on CRISPR–Cas systems for acquiring foreign DNA into the host CRISPR locus for immunological memory. Spacer acquisition occurs in two stages – foreign DNA capture, followed by their integration into the CRISPR locus (**Figure 6.1**). The Cas1–Cas2 integrase complex plays a central role in both stages. During foreign DNA or protospacer capture, Cas1–Cas2 engages the protospacer by splaying the DNA ends to separate the strand carrying the 3′–OH nucleophile from the non-nucleophilic strand. The 3′–OH ends are accommodated within the catalytic Cas1 active sites. The *E. coli* Cas1–Cas2 precisely measures the protospacer from end-to-end to 33 bp, explaining mechanistic basis for the conserved length of spacers in CRISPR loci. Cas2 functions as a protein scaffold within the Cas1–Cas2 complex that stabilizes the DNA phosphodiester backbone.

Following protospacer capture, Cas1–Cas2 catalyzes two integration reactions using the 3′–OH nucleophilic ends of the DNA. The first integration event occurs at the leader-repeat border and the next at the repeat-spacer border. *In vitro*, Cas1–Cas2 integrates the protospacer adjacent to every repeat in a supercoiled plasmid encoding the CRISPR locus. Although this highlights the ability of Cas1–Cas2 for recognizing the CRISPR repeats, *in vivo* spacer acquisition occurs adjacent to the first repeat bordering the leader sequence. Specific integration adjacent to the leader in *E. coli* depends on the integration host factor (IHF) protein to induce a sharp bend on the leader, allowing Cas1–Cas2 to access the targeted sequence for the integration reaction.

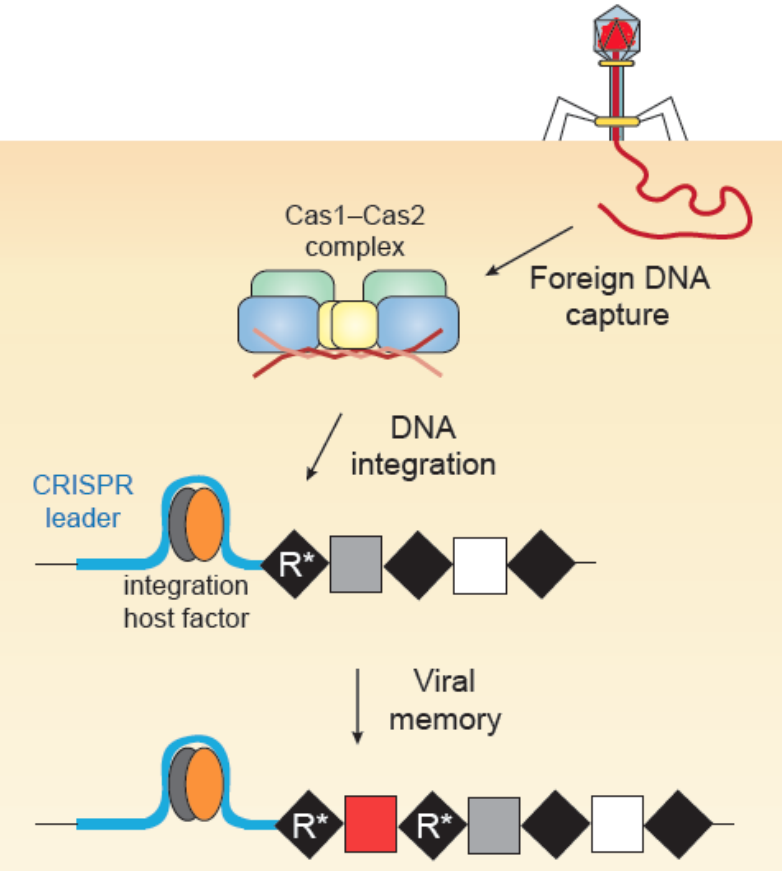


Figure 6.1 | Mechanism of CRISPR–Cas immunological memory. Spacer acquisition initiates with the recognition of foreign DNA by Cas1–Cas2. Upon DNA binding, Cas1–Cas2 splays the DNA ends to position the strand carrying the 3′–OH nucleophile within opposite Cas1 active sites. Upon protospacer engagement, Cas1–Cas2 catalyzes protospacer integration adjacent to the first repeat bordering the leader (asterisk). The IHF protein induces a sharp bend on the CRISPR leader to initiate the integration reaction. After integration and DNA repair, a new spacer is acquired and the first repeat is duplicated.

6.2 Future directions

Key aspects underlying the mechanism of spacer acquisition still remain poorly understood. Protospacer selection in foreign DNA is nonrandom and relies on the presence of a 3-5 nt protospacer adjacent motif (PAM) sequence directly upstream of the protospacer. In *E. coli*, the predominant PAM is AWG during naïve spacer acquisition and strictly AAG during primed spacer acquisition. The G nucleotide becomes the first nucleotide of the mature 33 bp protospacer. A recent crystal structure of Cas1–Cas2 bound to protospacer DNA with a PAM sequence reveals base-specific contacts of the PAM in the Cas1 active sites[39]. However, a crystal structure of the mature 33-bp protospacer also show base-specific contacts when the terminal nucleotide is not a G. Thus, the DNA is symmetrically coordinated on both PAM and non-PAM ends. How Cas1–Cas2 specifically recognizes the PAM on one protospacer end with no sequence specificity on the other end remains unknown. It is possible that upon PAM recognition on one DNA end, Cas1–Cas2 cleaves directly 33 bp downstream to produce a mature protospacer.

The mechanisms underlying primed spacer acquisition are still being investigated. During primed spacer acquisition, active interference of foreign DNA by Cascade and Cas3 results in robust spacer acquisition by Cas1–Cas2 (ref. 21). It is hypothesized that active DNA cleavage by Cas3 produces DNA fragments that are captured by Cas1–Cas2. Single molecule studies have shown the co-migration of Cas1–Cas2 with Cas3 along DNA, suggesting that complex formation results in primed spacer acquisition[50]. During naïve spacer acquisition, DNA cleavage by the RecBCD complex has been shown to directly induce spacer acquisition[37]. Since the products of RecBCD and Cas3 are single stranded DNA, further work is needed on how protospacers become double-stranded since single-stranded DNA are not substrates for DNA integration.

How Cas1–Cas2 specifically recognizes the sequence at the leader-repeat border is unknown. There is no crystal structure of the Cas1–Cas2-DNA complex bound to target DNA, which could reveal the mechanism of target DNA capture. Based on our studies on IHF, we predict the target DNA to be distorted to allow for base-specific readout of the leader-repeat sequence.

The content described in this thesis are on the Type I-E CRISPR–Cas system in *E. coli*, which only requires Cas1 and Cas2 for spacer acquisition. Many CRISPR–Cas systems require additional proteins for spacer acquisition. Many Type I systems and Type II-B require Cas4, whereas Type II-A has Csn2 requires Csn2. The roles of these proteins during spacer acquisition are elusive. It is hypothesized that the exonucleolytic activity of Cas4 is required for protospacer maturation[53], whereas Csn2 has been shown to slide along DNA with no nuclease activity[120,121]. In addition, Type II spacer acquisition also requires Cas9[122,123]. It was shown that Cas9, Cas1, Cas2 and Csn2 form a complex[123], however the mechanistic function of complex formation is not understood.

6.2 Applications

The development of Cas9 for targeted genomic manipulation has resulted in wide interest in further engineering other Cas proteins for biotechnology applications. Recent mechanistic understanding of the Cas1–Cas2 integrase could be potentially useful for a variety of applications. One area of interest is for integrating unique DNA barcodes into mammalian genomes for lineage mapping of cells. Such DNA barcoding technology has been used to study cell dynamics, tracking DNA mutations in cancer cells and stem cell differentiation. Current technologies for DNA barcoding rely on retroviral integration of the barcodes, which is non-sequence specific. Cas1–Cas2 has an advantage of high sequence-specificity of the target while non-sequence specific selection of the barcode substrate. One specific example is to introduce unique barcodes into single naïve T cells to track their differentiation into specific lineages (**Fig. 6.2**). This work will be conducted through a collaboration with the Doudna and Alex Marson's laboratories (UCSF).

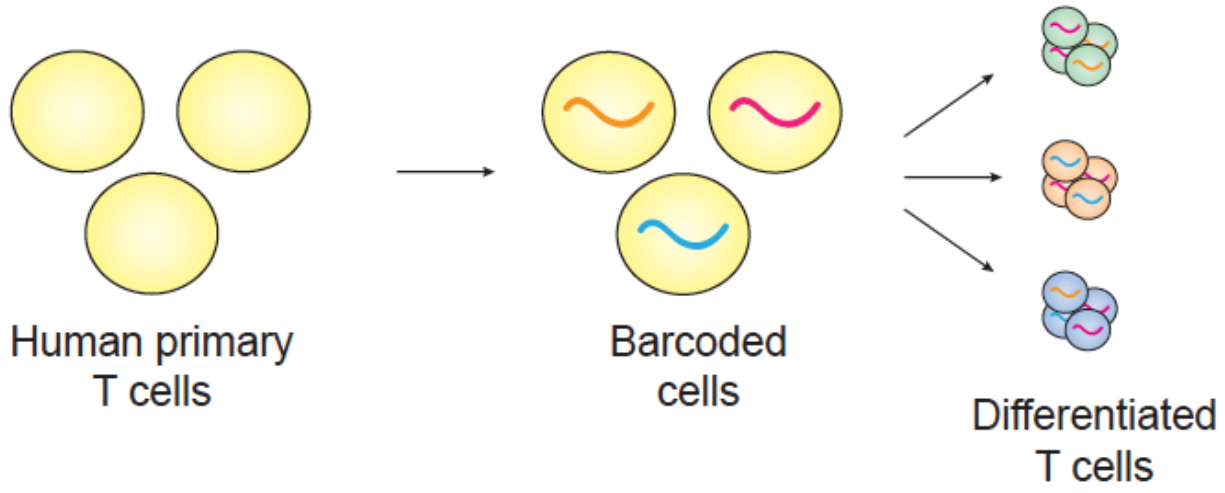


Figure 6.2 | Schematic of T cell barcoding strategy. Primary T cells will be barcoded with unique DNA barcodes using either Cas1–Cas2 or Cas9-mediated methods. The T cells will then be differentiated into various lineages to determine specific factors and events that leads to differentiation. A similar barcoding strategy could also be applied for tracking the differentiation of cancer cells and stem cells.

Bibliography

1. Labrie, S. J., Samson, J. E. & Moineau, S. Bacteriophage resistance mechanisms. *Nat Rev Microbiol* **8,** 317–327 (2010).
2. Westra, E. R. *et al.* The CRISPRs, they are a-changin': how prokaryotes generate adaptive immunity. *Annu Rev Genet* **46,** 311–339 (2012).
3. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315,** 1709–1712 (2007).
4. Marraffini, L. A. CRISPR-Cas immunity in prokaryotes. *Nature* **526,** 55–61 (2015).
5. Mojica, F. J., Diez-Villasenor, C., Soria, E. & Juez, G. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Mol Microbiol* **36,** 244–246 (2000).
6. Jansen, R., Embden, J. D., Gaastra, W. & Schouls, L. M. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* **43,** 1565–1575 (2002).
7. 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).
8. 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).
9. Mojica, F. J., Diez-Villasenor, C., Garcia-Martinez, J. & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* **60,** 174–182 (2005).
10. Carte, J., Wang, R., Li, H., Terns, R. M. & Terns, M. P. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes Dev* **22,** 3489–3496 (2008).
11. 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).
12. Deltcheva, E. *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471,** 602–607 (2011).
13. Charpentier, E., Richter, H., van der Oost, J. & White, M. F. Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. *FEMS Microbiol. Rev.* **39,** 428–441 (2015).
14. Brouns, S. J. *et al.* Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321,** 960–964 (2008).

15. Marraffini, L. A. & Sontheimer, E. J. CRISPR Interference Limits Horizontal Targeting DNA. *Science* **322,** 1843–1845 (2008).

16. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337,** 816–821 (2012).

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

18. Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.* **13,** 722–736 (2015).

19. Shmakov, S. *et al.* Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. *Mol. Cell* **60,** 385–397 (2015).

20. 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).

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

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

23. Babu, M. *et al.* A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. *Mol Microbiol* **79,** 484–502 (2011).

24. Kim, T. Y., Shin, M., Huynh Thi Yen, L. & Kim, J. S. Crystal structure of Cas1 from Archaeoglobus fulgidus and characterization of its nucleolytic activity. *Biochem Biophys Res Commun* **441,** 720–725 (2013).

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

26. 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).

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

28. Heler, R., Marraffini, L. A. & Bikard, D. Adapting to new threats: the generation of memory by CRISPR-Cas immune systems. *Mol Microbiol* **93,** 1–9 (2014).

29. Amitai, G. & Sorek, R. CRISPR-Cas adaptation: new insights into the mechanism of action. *Nat. Rev. Microbiol.* **14,** 67–76 (2016).

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

31. Mojica, F. J. M. *et al.* Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155,** 733–740 (2009).

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

33. Diez-Villaseñor, C. *et al.* 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).

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

35. 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).

36. Shmakov, S. *et al.* Pervasive generation of oppositely oriented spacers during CRISPR adaptation. *Nucleic Acids Res* **42,** 5907–5916 (2014).

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

38. Yosef, I. *et al.* DNA motifs determining the efficiency of adaptation into the Escherichia coli CRISPR array. *Proc Natl Acad Sci U S A* **110,** 14396–14401 (2013).

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

40. Jore, M. M. *et al.* Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nat Struct Mol Biol* **18,** 529–536 (2011).

41. Lintner, N. G. *et al.* Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE). *J Biol Chem* **286,** 21643–21656 (2011).

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

43. Sinkunas, T. *et al.* Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *EMBO J* **30,** 1335–1342 (2011).

44. Westra, E. R. *et al.* CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. *Mol Cell* **46,** 595–605 (2012).

45. Sinkunas, T. *et al.* In vitro reconstitution of Cascade-mediated CRISPR immunity in Streptococcus thermophilus. *EMBO J* **32,** 385–394 (2013).

46. Hochstrasser, M. L. *et al.* CasA mediates Cas3-catalyzed target degradation

during CRISPR RNA-guided interference. *Proc Natl Acad Sci U S A* **111,** 6618–6623 (2014).

47. Li, M., Wang, R., Zhao, D. & Xiang, H. Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process. *Nucleic Acids Res* **42,** 2483–2492 (2013).

48. Richter, C. *et al.* Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. *Nucleic Acids Res* **42,** 8516–8526 (2014).

49. Blosser, T. R. R. *et al.* Two distinct DNA binding modes guide dual roles of a CRISPR-Cas protein complex. *Mol Cell* **58,** 60–70 (2015).

50. Redding, S. *et al.* Surveillance and processing of foreign DNA by the Escherichia coli CRISPR-Cas system. *Cell* **163,** 854–865 (2015).

51. Li, M., Wang, R. & Xiang, H. Haloarcula hispanica CRISPR authenticates PAM of a target sequence to prime discriminative adaptation. *Nucleic Acids Res.* **42,** 7226–7235 (2014).

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

53. Lemak, S. *et al.* Toroidal structure and DNA cleavage by the CRISPR-associated [4Fe-4S] cluster containing Cas4 nuclease SSO0001 from Sulfolobus solfataricus. *J Am Chem Soc* **135,** 17476–17487 (2013).

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

55. 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).

56. Blackwood, J. K. *et al.* End-resection at DNA double-strand breaks in the three domains of life. *Biochem Soc Trans* **41,** 314–320 (2013).

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

58. 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).

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

60. Mizuuchi, K. & Adzuma, K. Inversion of the phosphate chirality at the target site of Mu DNA strand transfer: evidence for a one-step transesterification mechanism.

Cell **66,** 129–140 (1991).

61. 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).

62. Diez-Villasenor, C., Almendros, C., Garcia-Martinez, J. & Mojica, F. J. Diversity of CRISPR loci in Escherichia coli. *Microbiology* **156,** 1351–1361 (2010).

63. Wiedenheft, B. *et al.* Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. *Structure* **17,** 904–912 (2009).

64. Han, D., Lehmann, K. & Krauss, G. SSO1450--a CAS1 protein from Sulfolobus solfataricus P2 with high affinity for RNA and DNA. *FEBS Lett* **583,** 1928–1932 (2009).

65. Makarova, K. S., Anantharaman, V., Aravind, L. & Koonin, E. V. Live virus-free or die: coupling of antivirus immunity and programmed suicide or dormancy in prokaryotes. *Biol Direct* **7,** 40 (2012).

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

67. Van Duyne, G. D., Standaert, R. F., Karplus, P. A., Schreiber, S. L. & Clardy, J. Atomic structures of the human immunophilin FKBP-12 complexes with FK506 and rapamycin. *J Mol Biol* **229,** 105–124 (1993).

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

69. Evans, P. Scaling and assessment of data quality. *Acta Crystallogr D Biol Crystallogr* **62,** 72–82 (2006).

70. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* **66,** 213–221 (2010).

71. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* **60,** 2126–2132 (2004).

72. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215,** 403–410 (1990).

73. Pei, J., Kim, B. H. & Grishin, N. V. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res* **36,** 2295–2300 (2008).

74. Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25,** 1189–1191 (2009).

75. Stern, A., Keren, L., Wurtzel, O., Amitai, G. & Sorek, R. Self-targeting by CRISPR: gene regulation or autoimmunity? *Trends Genet* **26,** 335–340 (2010).

76. Garneau, J. E. *et al.* The CRISPR/Cas bacterial immune system cleaves

bacteriophage and plasmid DNA. *Nature* **468,** 67–71 (2010).

77. Li, M. & Craigie, R. Processing of viral DNA ends channels the HIV-1 integration reaction to concerted integration. *J Biol Chem* **280,** 29334–29339 (2005).

78. Cherepanov, P. LEDGF/p75 interacts with divergent lentiviral integrases and modulates their enzymatic activity in vitro. *Nucleic Acids Res* **35,** 113–124 (2007).

79. Hare, S. *et al.* A novel co-crystal structure affords the design of gain-of-function lentiviral integrase mutants in the presence of modified PSIP1/LEDGF/p75. *PLoS Pathog* **5,** e1000259 (2009).

80. Yang, J. Y., Jayaram, M. & Harshey, R. M. Positional information within the Mu transposase tetramer: catalytic contributions of individual monomers. *Cell* **85,** 447–455 (1996).

81. DiNardo, S., Voelkel, K. A., Sternglanz, R., Reynolds, A. E. & Wright, A. Escherichia coli DNA topoisomerase I mutants have compensatory mutations in DNA gyrase genes. *Cell* **31,** 43–51 (1982).

82. Pruss, G. J., Manes, S. H. & Drlica, K. Escherichia coli DNA topoisomerase I mutants: increased supercoiling is corrected by mutations near gyrase genes. *Cell* **31,** 35–42 (1982).

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

84. Au, T. K., Pathania, S. & Harshey, R. M. True reversal of Mu integration. *EMBO J* **23,** 3408–3420 (2004).

85. Curcio, M. J. & Derbyshire, K. M. The outs and ins of transposition: from mu to kangaroo. *Nat Rev Mol Cell Biol* **4,** 865–877 (2003).

86. Tyson, G. W. & Banfield, J. F. Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Env. Microbiol* **10,** 200–207 (2008).

87. Sheflin, L. G. & Kowalski, D. Altered DNA conformations detected by mung bean nuclease occur in promoter and terminator regions of supercoiled pBR322 DNA. *Nucleic Acids Res* **13,** 6137–6154 (1985).

88. Goren, M. G., Yosef, I., Auster, O. & Qimron, U. Experimental definition of a clustered regularly interspaced short palindromic duplicon in Escherichia coli. *J Mol Biol* **423,** 14–16 (2012).

89. Semenova, E. *et al.* Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc Natl Acad Sci U S A* **108,** 10098–10103 (2011).

90. Westra, E. R. *et al.* Type I-E CRISPR-cas systems discriminate target from non-target DNA through base pairing-independent PAM recognition. *PLoS Genet* **9,** e1003742 (2013).

91. Craigie, R. & Bushman, F. D. HIV DNA integration. *Cold Spring Harb Perspect*

Med **2,** a006890 (2012).

92. Nowotny, M. Retroviral integrase superfamily: the structural perspective. *EMBO Rep* **10,** 144–151 (2009).

93. Hochstrasser, M. L. & Doudna, J. A. Cutting it close: CRISPR-associated endoribonuclease structure and function. *Trends Biochem Sci* **40,** 58–66 (2014).

94. Palecek, E. Local supercoil-stabilized DNA structures. *Crit Rev Biochem Mol Biol* **26,** 151–226 (1991).

95. Engelman, A. & Craigie, R. Efficient magnesium-dependent human immunodeficiency virus type 1 integrase activity. *J Virol* **69,** 5908–5911 (1995).

96. 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).

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

98. Savilahti, H., Rice, P. A. & Mizuuchi, K. The phage Mu transpososome core: DNA requirements for assembly and function. *EMBO J* **14,** 4893–4903 (1995).

99. Scottoline, B. P., Chow, S., Ellison, V. & Brown, P. O. Disruption of the terminal base pairs of retroviral DNA during integration. *Genes Dev* **11,** 371–382 (1997).

100. Katz, R. A., Merkel, G., Andrake, M. D., Roder, H. & Skalka, A. M. Retroviral integrases promote fraying of viral DNA ends. *J Biol Chem* **286,** 25710–25718 (2011).

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

102. Dillon, S. C. & Dorman, C. J. Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. *Nat. Rev. Microbiol.* **8,** 185–195 (2010).

103. Friedman, D. I. Integration host factor: a protein for all reasons. *Cell* **55,** 545–554 (1988).

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

105. Craig, N. L. & Nash, H. A. E. coli integration host factor binds to specific sites in DNA. *Cell* **39,** 707–716 (1984).

106. Yang, C. C. & Nash, H. a. The interaction of E. coli IHF protein with its specific binding sites. *Cell* **57,** 869–880 (1989).

107. Goodrich, J. A., Schwartz, M. L. & McClure, W. R. Searching for and predicting the activity of sites for DNA binding proteins: compilation and analysis of the binding sites for Escherichia coli integration host factor (IHF). *Nucleic Acids Res.* **18,** 4993–5000 (1990).

108. Aeling, K. A. *et al.* Indirect Recognition in Sequence-specific DNA Binding by

Escherichia coli Integration Host Factor: The Role of DNA Deformation Energy. *J. Biol. Chem.* **281,** 39236–39248 (2006).

109. Prieto, A. I. *et al.* Genomic analysis of DNA binding and gene regulation by homologous nucleoid-associated proteins IHF and HU in Escherichia coli K12. *Nucleic Acids Res.* **40,** 3524–3537 (2012).

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

111. Maskell, D. P. *et al.* Structural basis for retroviral integration into nucleosomes. *Nature* **523,** 366–369 (2015).

112. Montaño, S. P., Pigli, Y. Z. & Rice, P. a. The μ transpososome structure sheds light on DDE recombinase evolution. *Nature* **491,** 413–7 (2012).

113. Arias-Palomo, E. & Berger, J. M. An Atypical AAA+ ATPase Assembly Controls Efficient Transposition through DNA Remodeling and Transposase Recruitment. *Cell* **162,** 860–871 (2015).

114. Müller, H. P. & Varmus, H. E. DNA bending creates favored sites for retroviral integration: an explanation for preferred insertion sites in nucleosomes. *EMBO J.* **13,** 4704–14 (1994).

115. Surette, M. G. & Chaconas, G. A protein factor which reduces the negative supercoiling requirement in the Mu DNA strand transfer reaction is Escherichia coli integration host factor. *J. Biol. Chem.* **264,** 3028–3034 (1989).

116. Craigie, R. & Mizuuchi, K. Role of DNA topology in Mu transposition: Mechanism of sensing the relative orientation of two DNA segments. *Cell* **45,** 793–800 (1986).

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

118. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 6640–6645 (2000).

119. Zaychikov, E., Schickor, P., Denissova, L. & Heumann, H. Hydroxyl radical footprinting. *Methods Mol. Biol.* **30,** 21–32 (1994).

120. Nam, K. H., Kurinov, I. & Ke, A. Crystal structure of clustered regularly interspaced short palindromic repeats (CRISPR)-associated Csn2 protein revealed Ca2+-dependent double-stranded DNA binding activity. *J Biol Chem* **286,** 30759–30768 (2011).

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

122. Wei, Y., Terns, R. M. & Terns, M. P. Cas9 function and host genome sampling in

Type II-A CRISPR-Cas adaptation. *Genes Dev* **29,** 356–361 (2015).

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

124. Jinek, M. *et al.* A Programmable Dual-RNA – Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337,** 816–822 (2012).

125. Doudna, J. A. & Charpentier, E. The new frontier of genome engineering with CRISPR-Cas9. *Science* **346,** 1258096–1258096 (2014).

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

127. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339,** 823–6 (2013).

128. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339,** 819–23 (2013).

129. Jinek, M. *et al.* RNA-programmed genome editing in human cells. *Elife* **2,** e00471 (2013).

130. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152,** 1173–1183 (2013).

131. Larson, M. H. *et al.* CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat. Protoc.* **8,** 2180–2196 (2013).

132. Gilbert, L. A. *et al.* CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154,** 442–51 (2013).

133. Gilbert, L. A. *et al.* Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159,** 647–661 (2014).

134. Cheng, A. W. *et al.* Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system. *Cell Res.* **23,** 1163–71 (2013).

135. Maeder, M. L. *et al.* CRISPR RNA-guided activation of endogenous human genes. *Nat. Methods* **10,** 977–9 (2013).

136. Perez-Pinera, P. *et al.* RNA-guided gene activation by CRISPR-Cas9-based transcription factors. *Nat. Methods* **10,** 973–6 (2013).

137. Mali, P. *et al.* CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat. Biotechnol.* **31,** 833–8 (2013).

138. Ran, F. A. *et al.* Double nicking by RNA-guided CRISPR cas9 for enhanced genome editing specificity. *Cell* **154,** 1380–1389 (2013).

139. Fu, Y., Sander, J. D., Reyon, D., Cascio, V. M. & Joung, J. K. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat. Biotechnol.* **32,** 279–84 (2014).

140. Kleinstiver, B. P. *et al.* High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529,** 490–495 (2016).

141. Slaymaker, I. M. *et al.* Rationally engineered Cas9 nucleases with improved specificity. *Science* **351,** 84–88 (2015).

142. Tsai, S. Q. *et al.* Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nat. Biotechnol.* **32,** 569–76 (2014).

143. Guilinger, J. P., Thompson, D. B. & Liu, D. R. Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. *Nat. Biotechnol.* **32,** 577–582 (2014).

144. Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31,** 827–32 (2013).

145. González, F. *et al.* An iCRISPR Platform for Rapid, Multiplexable, and Inducible Genome Editing in Human Pluripotent Stem Cells. *Cell Stem Cell* **15,** 215–226 (2014).

146. Dow, L. E. *et al.* Inducible in vivo genome editing with CRISPR-Cas9. *Nat. Biotechnol.* **33,** 390–394 (2015).

147. Kim, S., Kim, D., Cho, S. W., Kim, J. & Kim, J. S. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Res.* **24,** 1012–1019 (2014).

148. Ramakrishna, S. *et al.* Gene disruption by cell-penetrating peptide-mediated delivery of Cas9 protein and guide RNA. *Genome Res.* **24,** 1020–7 (2014).

149. Lin, S., Staahl, B., Alla, R. K. & Doudna, J. a. Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *Elife* **3,** 1–13 (2014).

150. Zuris, J. A. *et al.* Cationic lipid-mediated delivery of proteins enables efficient protein-based genome editing in vitro and in vivo. *Nat. Biotechnol.* **33,** 73–80 (2014).

151. Jinek, M. *et al.* Structures of Cas9 Endonucleases Reveal RNA-Mediated Conformational Activation. *Science* **343,** 1247997–1247997 (2014).

152. Nishimasu, H. *et al.* Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156,** 935–49 (2014).

153. Wright, A. V *et al.* Rational design of a split-Cas9 enzyme complex. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 2984–9 (2015).

154. Shah, N. H. & Muir, T. W. Inteins: nature's gift to protein chemists. *Chem. Sci.* **5,** 446–461 (2014).

155. Truong, D.-J. J. *et al.* Development of an intein-mediated split–Cas9 system for gene therapy. *Nucleic Acids Res.* **43,** 6450–6458 (2015).

156. Zetsche, B., Volz, S. E. & Zhang, F. A split-Cas9 architecture for inducible

genome editing and transcription modulation. *Nat. Biotechnol.* **33,** 139–142 (2015).

157. Nishimasu, H. *et al.* Crystal Structure of Staphylococcus aureus Cas9. *Cell* **162,** 1113–1126 (2015).

158. Banaszynski, L. & Liu, C. Characterization of the FKBP⊙ Rapamycin⊙ FRB Ternary Complex. *J. Am. …* 4715–4721 (2005).

159. Laplante, M. & Sabatini, D. M. mTOR signaling in growth control and disease. *Cell* **149,** 274–93 (2012).

160. Nihongaki, Y., Kawano, F., Nakajima, T. & Sato, M. Photoactivatable CRISPR-Cas9 for optogenetic genome editing. *Nat. Biotechnol.* **9,** 1–8 (2015).

161. Kawano, F., Suzuki, H., Furuya, A. & Sato, M. Engineered pairs of distinct photoswitches for optogenetic control of cellular proteins. *Nat. Commun.* **6,** 6256 (2015).

162. Konermann, S. *et al.* Optical control of mammalian endogenous transcription and epigenetic states. *Nature* **500,** 472–6 (2013).

163. Kennedy, M. J. *et al.* Rapid blue-light–mediated induction of protein interactions in living cells. *Nat. Methods* **7,** 973–975 (2010).

164. Baker, A. S. & Deiters, A. Optical control of protein function through unnatural amino acid mutagenesis and other optogenetic approaches. *ACS Chem. Biol.* **9,** 1398–1407 (2014).

165. Nihongaki, Y., Yamamoto, S., Kawano, F., Suzuki, H. & Sato, M. CRISPR-Cas9-based Photoactivatable Transcription System. *Chem. Biol.* **22,** 169–174 (2015).

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

167. Hemphill, J., Borchardt, E. K., Brown, K., Asokan, A. & Deiters, A. Optical Control of CRISPR/Cas9 Gene Editing. *J. Am. Chem. Soc.* **137,** 5642–5645 (2015).

168. Chaves, I. *et al.* The cryptochromes: blue light photoreceptors in plants and animals. *Annu. Rev. Plant Biol.* **62,** 335–64 (2011).

169. Gautier, A. *et al.* Genetically encoded photocontrol of protein localization in mammalian cells. *J. Am. Chem. Soc.* **132,** 4086–8 (2010).

170. Chin, J. W. Expanding and reprogramming the genetic code of cells and animals. *Annu. Rev. Biochem.* **83,** 379–408 (2014).

171. Davis, K. M., Pattanayak, V., Thompson, D. B., Zuris, J. A. & Liu, D. R. Small molecule–triggered Cas9 protein with improved genome-editing specificity. *Nat. Chem. Biol.* **11,** 316–8 (2015).

172. Yu, C. *et al.* Small Molecules Enhance CRISPR Genome Editing in Small Molecules Enhance CRISPR Genome Editing in Pluripotent Stem Cells. *Cell Stem Cell* **16,** 142–147 (2015).

173. Chu, V. T. *et al.* Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. *Nat. Biotechnol.* **33,** 543–548 (2015).

174. Bondy-Denomy, J. *et al.* Multiple mechanisms for CRISPR–Cas inhibition by anti-CRISPR proteins. *Nature* **526,** 136–139 (2015).

175. Bondy-Denomy, J., Pawluk, A., Maxwell, K. L. & Davidson, A. R. Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature* **493,** 429–32 (2013).

176. Pawluk, A., Bondy-denomy, J., Cheung, V. H. W., Maxwell, K. L. & Davidson, R. A New Group of Phage Anti-CRISPR Genes Inhibits the Type I-E CRISPR-Cas System of Pseudomonas aeruginosa. **5,** 1–7 (2014).

177. Hendel, A. *et al.* Chemically modified guide RNAs enhance CRISPR-Cas genome editing in human primary cells. *Nat. Biotechnol.* **33,** (2015).

178. Rahdar, M. *et al.* Synthetic CRISPR RNA-Cas9-guided genome editing in human cells. *Proc. Natl. Acad. Sci.* **112,** 110–117 (2015).

Appendix I

Chemical and biophysical modulation of Cas9 for tunable genome engineering

A portion of the work presented in this chapter has been published previously as part of the following paper: Nuñez, J.K., Harrington, L.B. & Doudna, J.A. Chemical and biophysical modulation of Cas9 for tunable genome engineering. *ACS Chemical Biology* **11**, 681-688 (2016).

I.1 Abstract

The application of the CRISPR–Cas9 system for genome engineering has revolutionized the ability to interrogate genomes of mammalian cells. Programming the Cas9 endonuclease to induce DNA breaks at specified sites is achieved by simply modifying the sequence of its cognate guide RNA. Although Cas9-mediated genome editing has been shown to be highly specific, cleavage events at off-target sites have also been reported. Minimizing, and eventually abolishing, unwanted off-target cleavage remains a major goal of the CRISPR–Cas9 technology before its implementation for therapeutic use. Recent efforts have turned to chemical biology and biophysical approaches to engineer inducible genome editing systems for controlling Cas9 activity at the transcriptional and protein levels. Here we review recent advancements to modulate Cas9-mediated genome editing by engineering split-Cas9 constructs, inteins, small molecules, protein-based dimerizing domains and light-inducible systems, highlighting the advantages and limitations of these recently developed technologies.

I.2 Introduction

Cas9 is an RNA-guided DNA endonuclease originating from the Type II CRISPR–Cas system in bacteria[4,18,124–126]. In contrast to previous genome engineering tools such as transcription activator-like effector nucleases (TALENs) and zinc finger nucleases (ZFNs), Cas9 is targeted to sites in the genome through simple DNA-RNA base pairing. For effective target DNA binding and cleavage, Cas9 requires two RNA components – a CRISPR RNA (crRNA) that contains sequence complementarity to the DNA and a transactivating RNA (tracrRNA) that base pairs with a short region of the crRNA[12,124,127–129] (**Fig. I.1a**). The crRNA and tracrRNA can be fused into a single guide RNA (sgRNA) by a linker and retains DNA binding and cleavage activities similar to the dual-RNA system[124,127–129]. In addition to the RNA guide, cleavage of the DNA also requires the presence of specific sequence next to the target DNA, termed the protospacer adjacent motif (PAM)[9,31]. The most widely used Cas9 for genetic manipulations is from the pathogenic bacteria *Streptococcus pyogenes*, which requires a 5′-NGG-3′ PAM directly adjacent to a 20 bp target DNA[124] (**Fig. I.1a**). Upon binding to the correct PAM-containing target, cleavage is triggered on both DNA strands by separate nuclease domains of Cas9. The RuvC domain cleaves the non-target strand, while the HNH domain acts on the target strand[124] (**Fig. I.1b**). Alternatively, catalytically inactive Cas9 (dCas9) has been widely used for site-specific transcriptional activation (CRISPRa) or inactivation (CRISPRi) by either blocking RNA polymerase[130,131] or fusing Cas9 to gene activators or silencers[132–136].

The ability to program Cas9 to bind any DNA sequence containing a PAM by simply modifying guide RNA sequence has been exploited for genome engineering purposes in a wide variety of organisms. In mammalian cells, genomic DNA breaks are repaired by either non-homologous end-joining (NHEJ) that results in an insertion-deletion (indel) mutation or homology-directed repair (HDR) that requires donor DNA to be site-specifically recombined into the site of cleavage (**Fig. I.1c**).

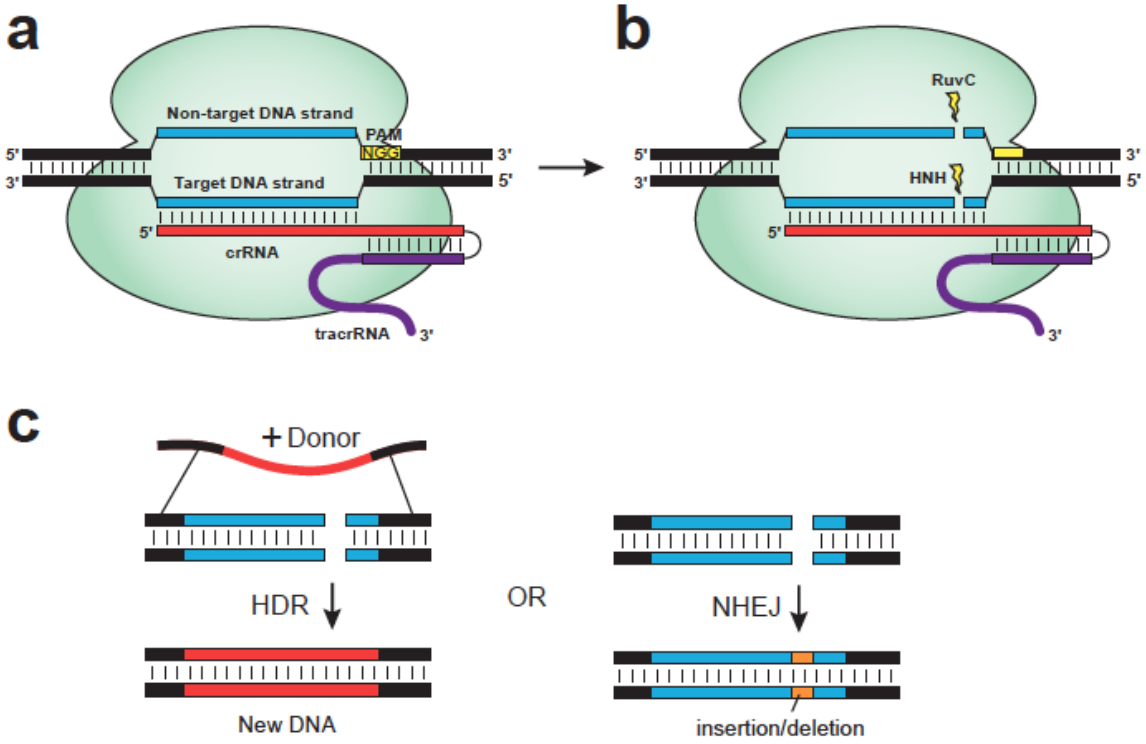


Figure I.1 | CRISPR–Cas9 genome editing technology. (**a**) Cas9 (green) requires a crRNA (orange) and tracrRNA (purple) guide to target DNA sequences with 20 bp complementa«ry to the crRNA. Additionally, a PAM sequence (yellow, 5′-NGG-3′) is required on the non-target strand for target binding and cleavage. (**b**) DNA cleavage occurs in two separate active sites – the RuvC domain cuts the non-target strand, whereas the hNH domain cleaves the target strand. (**c**) Two repair pathways are activated after DNA cleavage. Homology directed repair (HDR) requires a user-defined donor DNA template with flanking sequences matching to the site of cleavage. Non-homologous end joining (NHEJ) stitches the cleavage site that results in an insertion/deletion mutation (orange).

A critical aim of the CRISPR–Cas9 technology is to introduce site-specific DNA breaks at targeted sites while preventing potentially harmful off-target cleavage along the genome. Efforts to decrease off-target cleavage have included using paired Cas9 nickases[137,138], truncated guide RNAs[139], mutations impacting interactions with the target DNA[140,141] and FokI-dCas9 fusion[142,143]. Studies have shown that high expression of Cas9 results in elevated off-target cleavage[139,144], leading to the development of methods to control Cas9 levels at the transcriptional[145,146] and the protein levels by delivering pre-assembled Cas9-guide RNA ribonucleoprotein (RNP) complexes into cells[147–150]. Alternatively, many studies have turned to inducible systems to modulate Cas9 activity in mammalian cells using chemical and optical methods. By controlling the dosage of active Cas9, these systems also give the researcher the ability to control the location and time of Cas9 cleavage in a population of cells. Here we discuss recent chemical biology and biophysical approaches to control Cas9 activity, highlighting studies that utilize split-Cas9, inteins, small molecules, dimerizing domains and light-inducible Cas9 systems. A summary of the systems, along with their advantages and limitations, is provided in **Table I.1**.

Table I.1 | Summary of inducible Cas9 systems.

Type of System	Basic Mechanism	Application	Advantages	Limitations	Ref
Light-inducible	Photocaged lysine blocking active conformation	DNA cleavage	• Does not require additional proteins	• Requires expression of unnatural amino acids • Reduced editing efficiencies	(49)
	dCas9-CIBN dimerization to CRY2 fused to p65 or VP64	Transcriptional activation (CRISPRa)	• Large dynamic range • Low background	• Slow on-rate • Large size of additional proteins	(47,48)
	Magnet based dimerization of split-Cas9	DNA cleavage, interference of transcription (CRISPRi)	• Small size of additional proteins, ideal for AAV vectors • Rapid and tunable on-off rates	• Some background activity	(42)
Chemically inducible	Rapamycin-induced dimerization of FRB and FKBP fused split Cas9	DNA cleavage, Transcriptional activation	• Low background • Small size of additional proteins, ideal for AAV vectors	• Reduced editing efficiencies • Extended induction times required	(38,39)
	4-hydroxytamoxifen induced activation of intein	DNA cleavage	• Validated reduction in off-target cleavage	• Reduced overall editing efficiencies	(53)
	Indirect screen for small molecules	DNA cleavage	• Utilizes wildtype Cas9 • Selectively enhances NHEJ or HDR	• Small improvement in editing efficiencies	(54)
Non-inducible split-Cas9	Rationally split along DNA binding groove	DNA cleavage	• Low levels of self dimerization with shortened guide	• Reduced editing efficiencies • Only shown to work when delivered as RNP	(35)
	Intein fused split-Cas9 lobes excise DnaE upon dimerization	DNA cleavage	• Demonstrated AAV delivery • Validated in multiple cell-types	• Irreversible	(37)

I.3 Split-Cas9 systems for genome engineering

The first available molecular structures of Cas9 revealed two structural lobes separated by a central channel that accommodates the RNA:DNA hybrid[151,152] (**Figure I.2a**). Taking advantage of the bi-lobe architecture of Cas9, studies have engineered "split-Cas9" constructs with the goal of inducing Cas9 activity only when the two separated lobes assemble into a functional unit. In addition to the inducible nature of split-Cas9 systems that will allow for spatiotemporal control of Cas9 activity, a notable motivation is the ability to package the split Cas9 gene fragments into separate viral vectors, which is currently limited to ~4.7 kb for the widely used recombinant adeno-associated virus (rAAV) vectors (*S. pyogenes* Cas9 is ~4.2 kb without guide RNA and necessary promoters).

As proof of principle, the two lobes, termed α-helical or recognition lobe (REC) and nuclease lobe, can be expressed and purified separately[153]. Although guide RNA-dependent DNA cleavage was recapitulated by Wright and Sternberg *et al.* using split-Cas9 *in vitro*, the measured editing efficiency *in vivo* was drastically lower compared to full-length Cas9 when delivered as pre-assembled RNP complexes[153]. Additional studies have employed split-Cas9 systems for genome editing using chemically- and light-induced dimerization domains and inteins to trigger Cas9 complex formation *in vivo*.

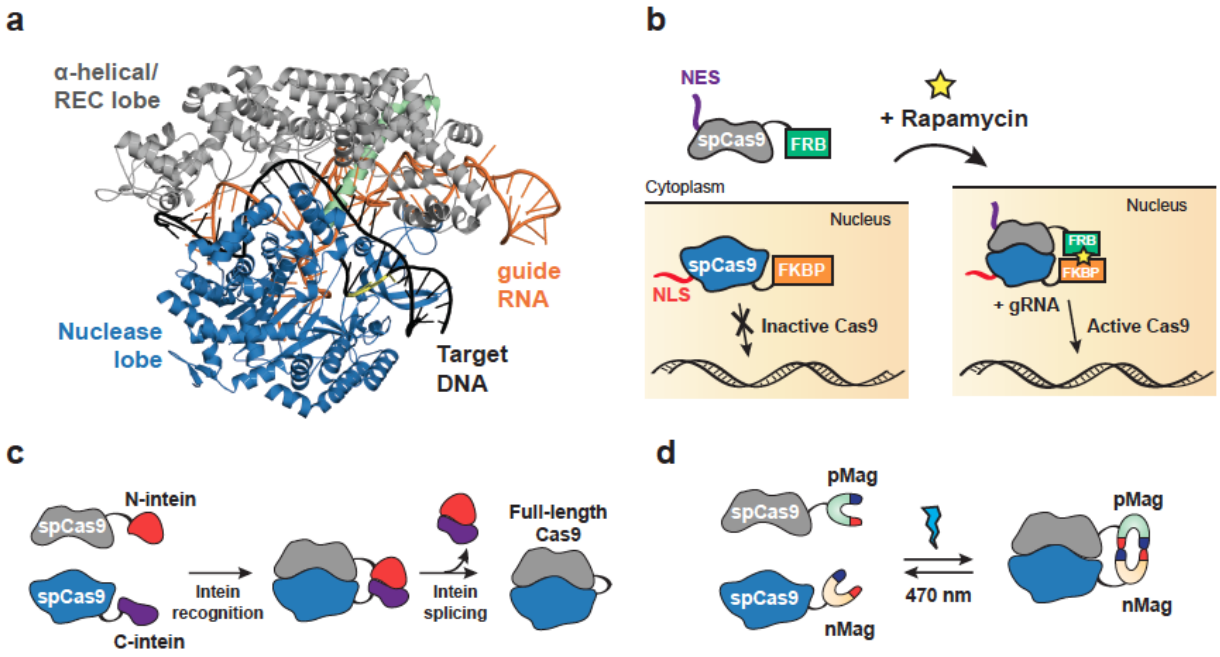


Figure I.2 | Split-Cas9 strategies for controllable genome editing. (**a**) Crystal structure of *S. pyogenes* Cas9 bound to guide RNA (orange) and PAM-containing target DNA (black, PDB 4UN3) reveals a bilobe architecture – alpha helical or REC (gray) and nuclease (blue). (**b**) The intein-based split-Cas9 is composed of two separate Cas9 fragments fused to intein sequences that become spliced out upon dimerization, leading to active Cas9. (**c**) Rapamycin-inducible split-Cas9 consists of FRB and FKRB fragments covalently fused to separate split Cas9 constructs (spCas9, blue and gray). Complex formation and subsequent DNA cleavage is induced in the nucleus upon the addition of rapamycin into cells. (**d**) The photoactivatable split Cas9 system attaches protein "Magnets' that dimerize upon blue light treatment. Reversibility is achieved by turning off blue light treatment.

I.3.1 Intein-inducible split-Cas9

Inteins are naturally occurring proteins with splicing-like properties for excising segments of the primary polypeptide sequence after translation, leading to a final functional protein[154] (**Figure I.2b**). Taking advantage of this system, one study fused each split Cas9 fragment to complementary intein sequences that were spliced out upon dimerization, resulting in a full-length Cas9 (ref. 155). Thus, the intein-split-Cas9 system allows activation of Cas9 only when both intein-fused Cas9 fragments are expressed. Editing efficiencies by NHEJ and HDR pathways using the intein-split-Cas9 system were comparable to WT full-length Cas9, although off-target cleavage rates were not reported. Interestingly, packaging of the fragments into separate AAV vectors also resulted in on-target cleavage, showing proof of principle that Cas9 can be delivered on AAV vectors as split fragments[155]. Since the system is irreversible, a limitation is the inability to control Cas9 activity after intein-mediated excision.

I.3.2 Rapamycin-inducible split-Cas9

As a facile method for inducing Cas9 complex formation *in vivo*, the split polypeptides were expressed in mammalian cells as fusions to either the FK506 binding protein 12 (FKBP) or FKBP rapamycin binding (FRB) domains[156,157] (**Figure I.2c**). Cas9 activation is chemically induced by addition of the immunosuppressant small molecule rapamycin that allows FKBP-FRB dimerization[158]. By adding a nuclear export sequence (NES) to one fragment and a nuclear localization sequence (NLS) to the other, the split-Cas9 fragments assembled in a manner that resulted in NHEJ-mediated editing at 19-29% in the presence of rapamycin without detectable background editing in the absence of rapamycin. Expressing the fragments at low levels by lentiviral expression significantly reduced off-target cleavage frequencies to undetectable levels, although the on-target cleavage was 2-fold lower. A limitation of the system is the weeks-long treatment of cells (~4-6 weeks) with rapamycin that could lead to undesirable effects by disrupting the master regulator protein mammalian target of rapamycin (mTOR)-related pathways[159], as well as self-dimerization of the split fragments upon breakdown of the nuclear envelope during cell division.

I.3.3 Optically controlled split-Cas9

To resolve the issues of irreversibility and/or potentially chemically invasive methods described above, a recent study took advantage of photoinducible dimerizing protein domains called Magnets to engineer a photoactivatable Cas9 (paCas9) system[160]. By fusing each split-Cas9 lobe to complementary Magnet fragments (p and nMagnet), electrostatic interactions between the p and n fragments triggers Magnet dimerization by applying noninvasive blue light (470 nm LED) to cells[161] (**Figure I.2d**). The Magnets, derived from a fungal circadian clock photoreceptor Vivid (VVD), have been engineered to dimerize in the order of seconds while maintaining slow dissociation rates ($t_{1/2}$, 4.7 hr). Genome editing using paCas9 resulted in NHEJ-mediated repair comparable to full length Cas9 while maintaining low editing in the absence of light. HDR efficiencies remained ~2-fold lower compared to WT Cas9 (ref. 160). An added advantage of the paCas9 system over chemically inducible systems is the ability to spatially control Cas9 activation simply by the pattern of light applied onto cells. Furthermore, in contrast to chemically induced Cas9 paCas9-mediated editing was rapidly reversible by moving cells from blue-light treatment to dark conditions. Due to the small size of the Magnet fragments (150 amino acids each), the paCas9 system provides ideal lengths for packaging each split-Cas-Magnet construct into separate rAAV vectors although this was not tested in the study.

I.4 Optogenetic control of Cas9

Light-inducible genomic interrogation had been previously shown to induce site-specific transcriptional activation using protein-based zinc-finger proteins (ZFPs), transcription-activator-like effectors (TALEs)[162] and Cre DNA recombination system[163], however programming these systems to target variable sequences along a genome is laborious[164]. Such optogenetic transcriptional activators are reversible and noninvasive,

in contrast to small molecule-inducible systems such as tetracycline that requires physical removal from cells for reversibility. In addition to paCas9, other optogenetic Cas9-based systems have been developed to induce transcriptional activation[165,166] and genome editing[167]. Targeted transcriptional gene activation by Cas9, termed CRISPRa, is accomplished by fusing catalytically inactive Cas9 (dCas9) to a transcriptional activator domain[132,134–136].

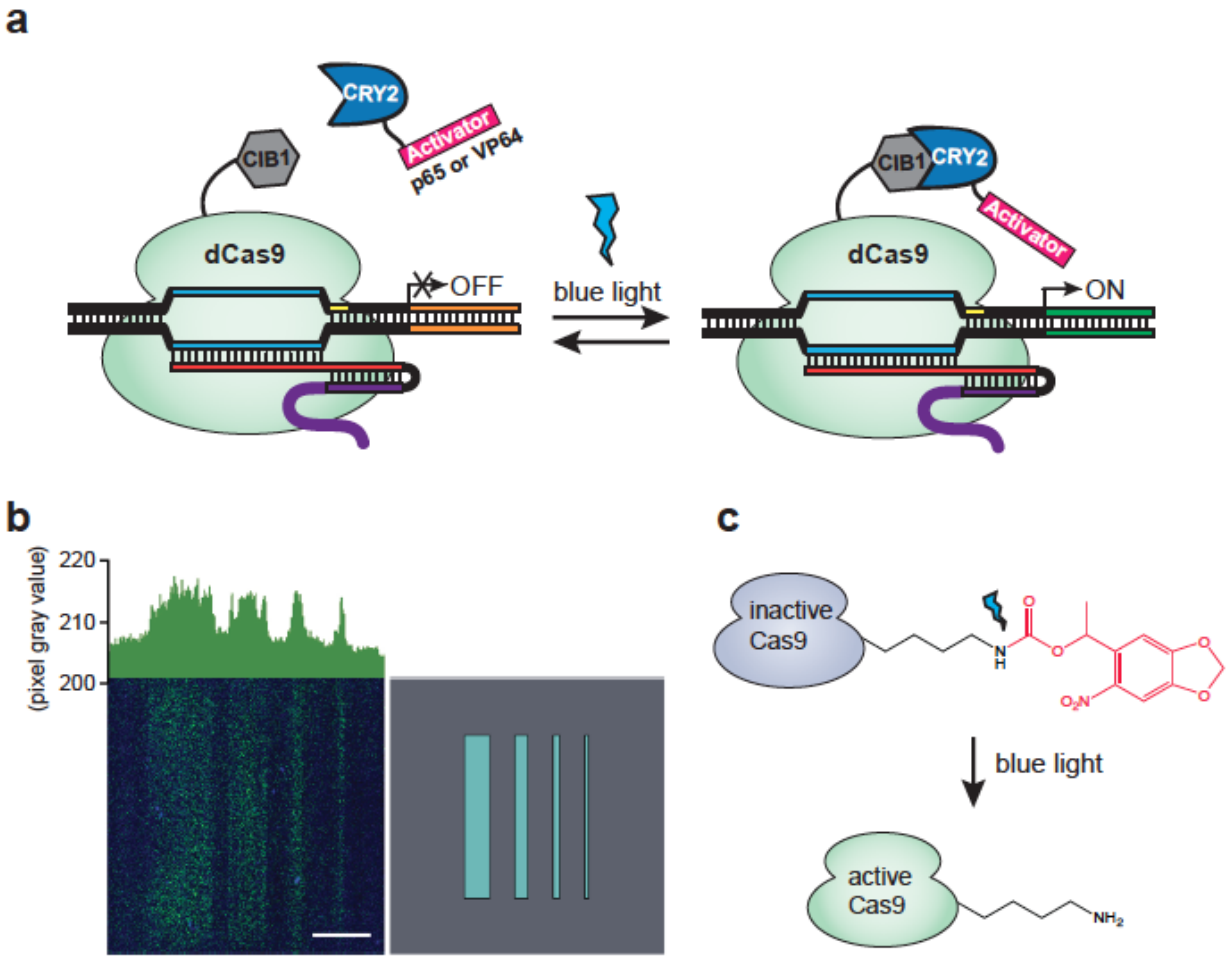


Figure I.3 | Cas9-mediated optogenetic editing and transcriptional activation. (**a**) Transcriptional activation of specified genes is engineered by programming catalytically dead Cas9 (dCas9) fused to CIB1 and separately fusing CRY2 with either P65 or VP64 transcriptional activators. Blue light treatment induces CIB1–CRY2 dimerization, leading to robust expression of the downstream gene). (**b**) Spatiotemporal transcriptional activation is exemplified by applying patterned blue light onto cells harboring light-indicuble Cas9 that results in GFP fluorescence governed by the pattern of light. Scale bar, 2 mm. (**c**) An alternative light-inducible Cas9 system with an engineered photocaged lysine that renders the protein inactive. The protecting group (red) becomes cleaved upon blue light treatment, resulting in fully active Cas9. Panel b is reprinted by permission from ref. 166. Macmillan Publishers Ltd., copyright 2015.

Motivated to optogenetically fine-tune the transcription of specific genes, two studies took advantage of the light-inducible dimerizing proteins CRY2 and CIB1 derived from *Arabidopsis thaliana*[165,166] (**Figure I.3a**). The photosensor CRY2 forms a covalent adduct with the chromophore flavin adenine dinucleotide (FAD) and undergoes a conformational change upon blue light irradiation that leads to dimerization of CRY2 with CIB1[163,164,168]. Anchoring to the desired locus is achieved by fusing CIB1 to dCas9 with multiple guide RNAs targeting the same endogenous promoter. Blue light treatment of cells triggers the dimerization of CIB1 with CRY2 fused to a transcriptional activator domain VP64[165,166] or p65 (ref. 165). This results in increase in mRNA levels over the course of hours (<10 fold increase after 2-3 hr light treatment) and at levels comparable to constitutively active dCas9-CRY2-CIB1 once the maximum signal is reached. Similar to paCas9, spatiotemporal activation can be visualized by the pattern of light applied onto cells (**Figure I.3b**). Additionally, transcriptional activation of multiple genes (multiplexing) can be achieved by simultaneously generating guides targeting different gene promoters[165].

An alternative photoactivable Cas9 system relies on Cas9 engineered to encode an unnatural photocaged lysine that renders the protein catalytically inactive[167]. Ultraviolet (UV) light irradiation cleaves the protecting group on the lysine that subsequently activates Cas9 for editing (**Figure I.3c**). The system requires cells with engineered tRNA synthesises that will acylate tRNAs with a desired unnatural amino acid[169,170]. Photocaged lysines were encoded within Cas9 that were predicted to block guide RNA binding or complex assembly with the target DNA[167]. UV treatment of cells resulted in editing of an endogenous gene at levels higher than without UV treatment albeit at levels lower than WT Cas9. Thus, the system requires further optimizations to achieve editing efficiencies comparable to aforementioned optogenetic systems and irreversibility would remain an issue.

I.5 Chemically inducible Cas9

The use of small molecules to regulate Cas9 activity has been recently explored and holds potential for activating genome editing by targeting Cas9 or other cellular targets that affect editing. **Figure I.4** highlights chemicals that have been used to control Cas9-mediated genome editing. A direct approach to chemically induce Cas9 activity was recently engineered by covalently encoding an intein sequence to Cas9 that disrupts DNA cleavage activity[171]. Excision of the intein is activated by addition of 4-hydroxytamoxifen (4-HT), producing active full-length Cas9. Despite the 2-fold decrease in NHEJ efficiency, presumably due to incomplete intein excision, the ratio of on-target/off-target was 6-fold higher on average and up to 25-fold compared to WT Cas9. Since the 4-HT receptor is derived from the human estrogen receptor ligand-binding domain and has the potential to be activated by other estrogen receptor modulators, a point mutation was engineered on the receptor that prevented undesired activation by β-estradiol[171].

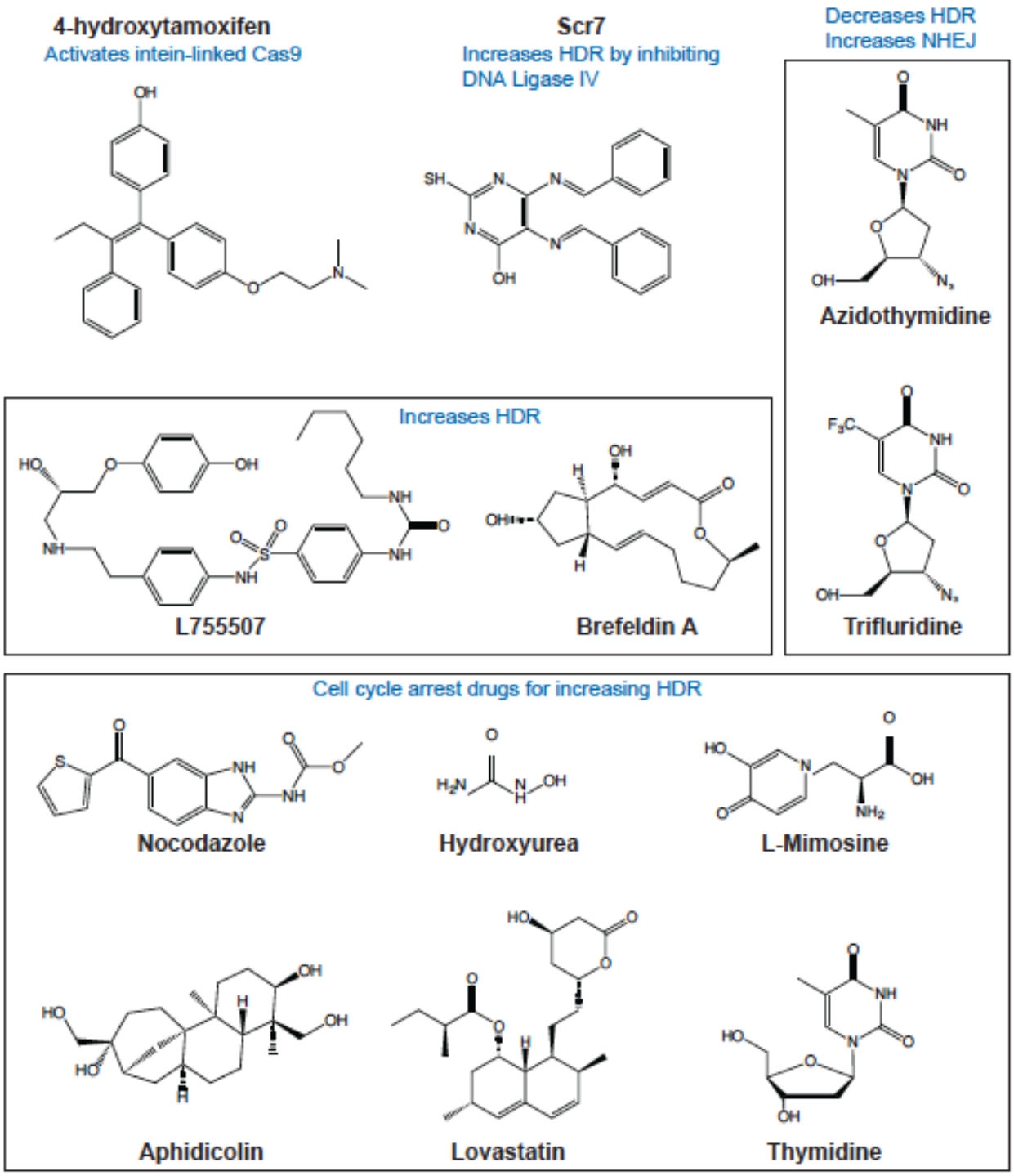


Figure I.4 | Chemicals used for Cas9-mediated genome editing. A panel of small molecules that have been experimentally used to control Cas9-mediated genome editing The blue descriptions highlight the effects of the chemicals on genome editing.

Other studies have relied on small molecules to indirectly regulate Cas9 activity, such as doxycycline-inducible expression of Cas9 (ref. 145, 146) and small molecules that target endogenous cellular processes[149] (**Figure I.4**). With the goal of increasing HDR levels in mouse embryonic stem cells, a screen of ~4,000 small molecules resulted in two compounds with 2-3 fold higher editing efficiencies compared to untreated cells[172]. The two compounds, L75507 and Brefeldin A, are not structurally similar and have been previously shown to target different cellular pathways (**Figure I.4**). The molecules activated HDR in a dose-dependent manner and toxicity was reported to be low after days-long treatment. Interestingly, two thymidine analogs, azidothymidine and trifluridine, decreased HDR by 3-fold while slightly increasing NHEJ levels, suggesting that the choice of small molecule could dictate the repair pathway after Cas9-guided cleavage[172]. In support of this, the small molecule Scr7 has been shown to decrease NHEJ and in turn, boost HDR efficiencies by inhibiting endogenous DNA ligase IV[83,173]. Furthermore, timely delivery of purified guide RNA-Cas9 RNP complexes, coupled with various cell cycle arrest drugs, has also been shown to increase HDR efficiency[149] (**Figure I.4**). Future work to screen other small molecule libraries could lead to more discoveries of promising chemicals that increase HDR or NHEJ while maintaining a high on/off-target ratio and low toxicity to cells.

I.6 Natural inhibitors of CRISPR systems

Although Cas9 activity can be modulated using the methods mentioned above, an unexplored area of research is to identify natural regulators of Cas9 in bacteria. The CRISPR–Cas system in prokaryotes is an effective adaptive immune response against mobile genetic elements[3]. Recently, some phages that infect *Pseudomonas aeruginosa* were found to evolve “anti-CRISPR” proteins that directly bind to different Cas proteins, leading to complete inactivation of the CRISPR–Cas system[174]. Anti-CRISPR proteins have been found to target Types I-E and I-F CRISPR systems[175,176], which possess other effector Cas proteins that function analogously to Cas9. Given the rapid evolutionary arms race between prokaryotes and mobile genetic elements, it is likely that natural mechanisms exist to modulate Cas9 activity perhaps in the form of anti-CRISPR proteins or small molecules. Such modulators could be potentially useful for turning off Cas9 activity for genome engineering purposes, as the aforementioned inducible systems rely on an “on switch” for activating Cas9.

I.7 Summary and perspectives

Using light, chemicals and dimerizing proteins, the utility of Cas9-based DNA targeting and cleavage has been greatly enhanced. However, the rate at which new Cas9 based tools continue to be developed is testament to the fact that researchers have left many avenues unexplored and underdeveloped. For example, most of the systems described above are limited by design to use in cultured mammalian cells. Validating or improving these systems to work at the organism level will be critical for their practical application as tools. Moreover, most engineering of the Cas9 complex has been focused on the protein itself. Modulation of the guide RNA, through riboswitches or other RNA based tools, provides a host of new opportunities to control

the activity of the Cas9 complex in cells. Recent studies have introduced chemical modifications on the guide RNA, such as phosphorothioate backbones and 2′-O methyl ribose substitutions, which show an increase in editing[177,178]. Finally, the power of Cas9 lies in its ability to leverage the innate DNA repair machinery of cells. As a result, the explosive use of Cas9 has highlighted how basic our understanding of these repair pathways in mammalian systems remains poor and unquestionably will spark renewed interest in this area of research. As our understanding of these pathways improves, they will likely provide us more targeted approaches to genetic manipulation and new proteins to selectively enhance or suppress Cas9 based genome engineering. The simple ability to direct a protein complex to targeted regions of the genome has been an unexpected and sudden advancement in biological research, and with further developments, it has the potential to become a tool that moves beyond lab based biology and towards therapeutic cures.

I.8 Acknowledgements

We acknowledge support from National Science Foundation grant No. 1244557 to J.A.D. J.K.N. and L.B.H. are supported by U.S. National Science Foundation Graduate Research Fellowships and J.K.N. by a UC Berkeley Chancellor's Graduate Fellowship. J.A.D. is an Investigator of the Howard Hughes Medical Institute. J.K.N. and L.B.H. wrote this review.