CRISPR's little helpers

CRISPR-Cas Proteins involved in PAM selection

Kieper, S.N.

DOI
10.4233/uuid:caeb7b8e-1d0c-4a10-8af3-cb6662267243

Publication date
2021

Document Version
Final published version

Citation (APA)
Kieper, S. N. (2021). *CRISPR's little helpers: CRISPR-Cas Proteins involved in PAM selection*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:caeb7b8e-1d0c-4a10-8af3-cb6662267243

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

CASIMIR PHD SERIES 2021-14
ISBN 978-90-8593-480-6

CRISPR'S LITTLE HELPERS: CRISPR-CAS PROTEINS INVOLVED IN PAM SELECTION SEBASTIAN KIEPER

CRISPR'S LITTLE HELPERS:
CRISPR-CAS PROTEINS INVOLVED IN PAM SELECTION

SEBASTIAN KIEPER

SK '21

INVITATION

TO THE PUBLIC DEFENSE OF MY PHD THESIS

CRISPR'S LITTLE HELPERS: CRISPR-CAS PROTEINS INVOLVED IN PAM SELECTION

WEDNESDAY, 16TH JUNE 2021

SENAATZAAL
AULA CONGRESCENTRUM
TU DELFT
MEKELWEG 5
DELFT

09.30 - 09.45
LAYMAN'S TALK

10.00 - 11.00
PUBLIC DEFENSE

S.N.KIEPER@GMAIL.COM

CRISPR's little helpers:
CRISPR-Cas Proteins involved in PAM selection

CRISPR's little helpers:
CRISPR-Cas Proteins involved in PAM selection

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology,
by the authority of the Rector Magnificus, Prof. Dr. ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates,
to be defended publicly on
Wednesday, 16th of June 2021 at 10 am.

by

Sebastian Niklas KIEPER
Master of Science in Biotechnology
Wageningen University & Research, Netherlands
Born in Hanover, Germany

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus	chairperson
Dr. ir. S.J.J. Brouns	Delft University of Technology, promotor
Dr. C. Joo	Delft University of Technology, promotor

Independent members:

Prof. dr. P.A.S. Daran-Lapujade	Delft University of Technology
Prof. dr. P.C. Fineran	University of Otago
Prof. dr. M.F. White	University of St Andrews
Dr. R.H.J. Staals	Wageningen University
Dr. J. Lebbink	Erasmus University Medical Center

Reserve:

Prof. dr. ir. S.J. Tans	Delft University of Technology

This work is part of the research programme of the Foundation for Fundamental Research on Matter (FOM), which is part of the Dutch Research Council (NWO).

Keywords: CRISPR Adaptation, Spacer Integration, PAM Selection, Cas4
Printed by: Gildeprint

Front & Back:
Cryo-SEM of high pressure frozen and freeze-fractured *Synechocystis* sp. 6803 cells (SE Micrograph inspiration from van de Meene et al., 2006).
Black & White Oil painting by S.N. Kieper.

Casimir PhD Series 2021-14
ISBN 978-90-8593-480-6
An electronic version of this dissertation is available at http://repository.tudelft.nl

To my grandmother Rosina Kieper and my daughter Matilda

Contents

1

General Introduction

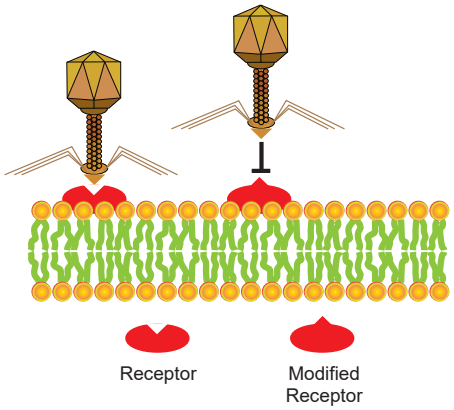
1

1.1 They will never become friends – The arms race between prokaryotes and bacteriophages

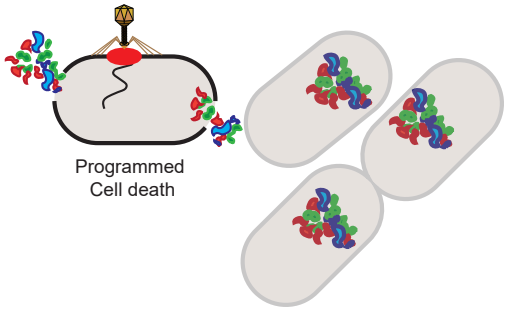
The interaction between predators and prey is a strong driver of the evolution of both entities. One of the oldest examples of this co-evolution is the evolutionary arms race between prokaryotes and their viruses (bacteriophages) which forces both, host and invader, to constantly adapt and evolve. Considering that prokaryotes are outnumbered 10-fold by their pathogens [1-3] the needs for strong lines of defense become apparent. Similar to eukaryotic defense mechanisms, prokaryotes evolved an arsenal of innate and adaptive immunity systems that can control, decrease and eliminate viral infections (Fig. 1.1) [4, 5]. While innate immunity confers virus protection relying on non-specific defense mechanisms, adaptive immune systems elicit a pathogen-specific response based on previous encounters with the invader [4, 6]. Innate immunity represents the first line of defense either interfering with phage adsorption or phage replication. Phages adapted to their host require specific receptors in order to adsorb to the cell and inject their genetic cargo, hence host surface modifications can prevent phage uptake (Fig. 1.1A) [7-9]. If phage uptake has taken place, abortive infection mechanisms can initiate programmed cell death in order to prevent phage replication and to contain the infection (Fig. 1.1B) [4, 10]. Furthermore, phage replication can be suppressed by restriction-modification systems that target and cleave specific sequences of the invading DNA elements (Fig. 1.1C) [11, 12].

It has long been thought that adaptive immunity is exclusive for eukaryotic organisms. Eukaryotic adaptive immunity relies on highly specialized cells and processes that respond to an initial exposure to an antigen [13]. This initial exposure creates an immunological memory that boosts and enhances the response to subsequent infections. The ability of prokaryotes to elicit an adaptive immune response was only recognized in the early 2000s when the significance of short repeating palindromic sequences in prokaryotic genomes was understood [14, 15]. The presence of those repeating sequences in the genome of *Escherichia coli* was firstly described by a Japanese group [16] but the authors did not immediately recognize the significance of this observation. Only little later, Spanish scientist Prof. Francisco Mojica ob-

A Surface Modifications



B Abortive Infection



C Restriction-Modification Systems

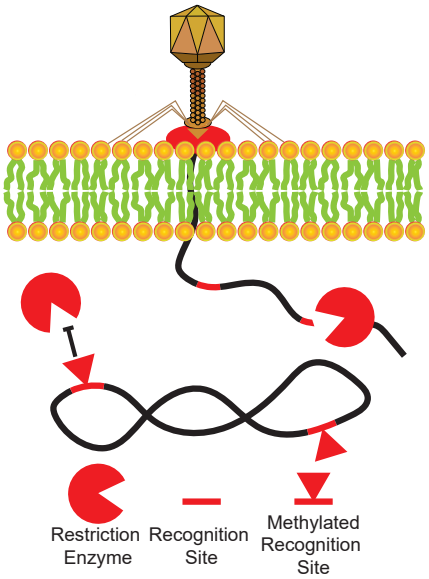


Figure 1.1 - Innate prokaryotic defense mechanisms. **A** Surface modifications such as modification of the phage receptor can inhibit or abolish phage adsorption to the cell. **B** By inducing programmed cell death, an infected cell can prevent phage replication. This altruistic defense mechanism sacrifices the individual cell in order to protect the population from infection. **C** Restriction modification systems rely on restriction nucleases that recognize and cleave unmodified invader DNA but leave modified (methylated) genomic host DNA intact.

served near-perfect repeating sequences of 30 basepairs interspaced by unique sequences of roughly 36 basepairs in his subject of study *Haloferax mediterranei* [17]. Prof. Mojica subsequently discovered similar structures in *H. mediterranei*-related as well as in more distant species of halophilic archaea, which in combination with the previously reported presence of such repeats, made him realize that there must be some biological significance to those structures [18]. Naming of those curious structures went through a whole series of evolution, ranging from short regularly spaced repeats (SRSRs; [17]) to spacers interspersed direct repeats (SPIDRs; [19]) and large cluster of tandem repeats (LCTRs; [20]). Eventually the terminology merged into Clustered Regularly Interspaced Short Palindromic Repeats with its well-known acronym CRISPR [21]. The same publication that introduced the new acronym CRISPR also described the presence of certain genes in the direct vicinity of those loci which were thereby named CRISPR-associated (*cas*) genes [21]. The domains that were commonly found in those Cas proteins suggested a role in DNA repair or DNA metabolism, however, direct evidence for this hypothesis was lacking. The final hint for the biological function of CRISPR came from the comprehensive analysis of the spacers interspersing the repeats, showing that those unique sequences were frequently derived from mobile genetic elements (MGEs) [22-24]. The observation that the presence of those MGE-derived spacers coincided with immunity of the carrier against those MGEs eventually lead to the understanding that CRISPR spacers might be involved in guiding a defense system.

Indeed, in 2007 the final proof of this hypothesis came from challenging *Streptococcus thermophilus* with a phage from which spacers had been incorporated into the bacterial CRISPR loci [25]. As a consequence, those *Streptococcus* strains displayed a phage resistant phenotype that was reliant on the presence of the CRISPR locus as well as the *cas* genes, demonstrating that CRISPR-Cas constituted an adaptive immunity system in prokaryotes [25]. Only shortly after this key finding, Brouns et al. provided the mechanistic details that explain

CRISPR-mediated phage immunity: A Cas protein complex coined Cascade (CRISPR-associated complex for antiviral defense) that matures CRISPR transcripts into short CRISPR RNA fragments (crRNA) and subsequently uses those crRNA molecules as guides to interfere with virus proliferation [26]. This overall concept of using catalogued MGE-derived sequences (spacers) to synthesize a RNA transcript that eventually guides Cas proteins towards an invader is a core feature of all CRISPR-Cas systems known to date, although the exact way of executing this immunity step varies and is the basis for the broad classification of CRISPR-Cas systems.

1.2 Diversity and classification of CRISPR-Cas systems

CRISPR-Cas systems not only appear to be widely distributed among prokaryotes (approximately 47% of bacteria and archaea contain CRISPR-Cas loci [27]) but also vary greatly in their Cas protein components. Multiple criteria are applied in order to classify CRISPR-Cas systems into two classes (Class 1 and Class 2), six types as well as currently 19 subtypes (see Table 1.1) [27, 28]. Systems that employ protein complexes composed of several subunits to elicit interference belong to Class 1 systems while systems using a large multidomain protein are classified as Class 2 systems [27]. The two classes are further divided into types (type I to type VI) depending on the presence of type specific unique signature *cas* genes. These signatures consist of *cas3* for type I, *cas10* for type III, *cas9* for type II, *csf1* (large subunit, *cas8*-like) for type IV, *cas12* for type V, and *cas13* for type VI [27, 29, 30]. Differentiation of CRISPR subtypes presents a more complex matter since only a limited number of subtypes contain defined diagnostic signature genes. For example, CRISPR type II-A is identified by the presence of *csn2* while type II-B systems are assigned by the presence of *cas4*. Subtypes that cannot be readily identified by signature genes are defined through their specific CRISPR locus organization and comparison of conserved genes. However, this approach of assigning subtypes suffers from certain ambiguities, leading to a growing number of CRISPR-Cas variants that cannot be readily classified [31]. Despite the astounding diversity of CRISPR-Cas systems, basic functional principles of the molecular mechanism are shared across the different systems.

Class	Type	Signature gene	pre-crRNA Processing	Target	Self vs Non-self discrimination	Effector components
Class 1	type I	*cas3*	Cas6, Cas5d	DNA	PAM	Cascade, Cas3, crRNA
	type III	*cas10*	Cas6 + RNase E	DNA + RNA	Repeat	Cmr/Csm, Cas10, crRNA
	type IV	*csf1*	Cas6	DNA/RNA ?	?	?
Class 2	type II	*cas9*	RNase III	DNA	PAM	Cas9 + crRNA + tracrRNA
	type V	*cas12*	Cas12	DNA	PAM	Cas12 + crRNA
	type VI	*cas13*	Cas13	RNA	PFS	Cas13 + crRNA

Table 1.1 - Overview of Class1 and Class2 CRISPR-Cas systems. The corresponding CRISPR-Cas types belonging to Class1 and Class2 systems are indicated with their cognate signature genes, their crRNA biogenesis pathway, the characterized target, the mechanism of self vs non-self discrimination as well as the involved effector components

1.3 Molecular mechanism of CRISPR-Cas adaptive immunity

The adaptive and inheritable nature of CRISPR-Cas mediated defense relies on the integration of virus derived fragments into the bacterial genome [32, 33]. This memory function allows an immunized bacterial strain to pass on immunity to future generations and ensures long-term protection. The full molecular mechanism is divided into three distinct stages (Fig. 2) that consist of the acquisition of viral fragments in a process called adaptation (for full review of the adaptation stage see Chapter 2 of this thesis), the transcription and processing of the acquired information during the expression stage and finally the assembly of matured transcripts (crRNA) with the effector proteins to initiate the interference stage.

1.3.1 CRISPR-Cas adaptation

Exemplary for the importance of the first stage of CRISPR-Cas immunity is the strong conservation of the protein responsible for the acquisition of CRISPR spacers, the Cas1 integrase protein [29, 30]. In the type I-E system of *E. coli* the naïve adaptation stage (naïve referring to acquisition of spacers from an invader that has not been encountered previously) exclusively requires the Cas1 integrase and the Cas2 protein [34]. The Cas1-Cas2 heterohexameric adaptation complex forms through electrostatic and hydrophobic interactions leading to the assembly of two Cas1 dimers and one Cas2 dimer [35, 36]. Overall, the initial immunization requires the identification of invading genetic material, the processing into spacer precursors as well as the integration as a novel spacer into the CRISPR array (Fig 1.2 - Stage I). A genome wide study analyzing the origin of newly acquired spacers showed that spacers are largely derived from plasmid DNA, despite the excess of chromosomal DNA in the cell [37]. This observation was explained by the adaptation complex deriving spacers from DNA degradation intermediates that arise during the repair of double-stranded DNA breaks (DSBs). In the context of the *E. coli* type I-E system, the RecBCD machinery is involved in partial degradation of DSB affected DNA until encountering a Chi site [38, 39]. The resulting degradation products then serve as a pool of spacer precursors for the adaptation complex. The uneven distribution of those Chi sites in genomic DNA and plasmid DNA therefore provides an explanation for the preferen-

1

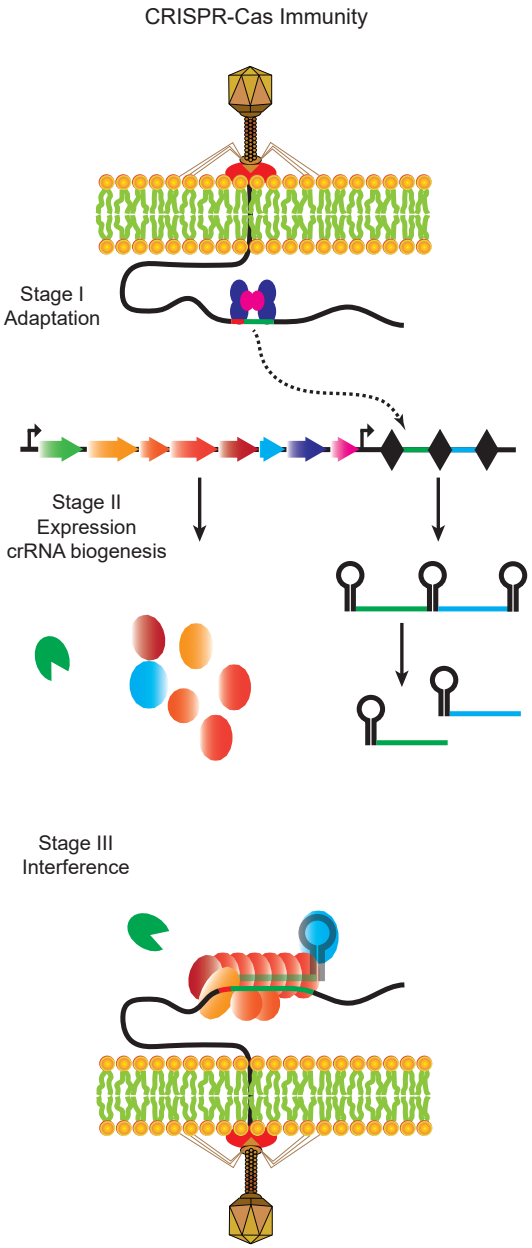


Figure 1.2 – Adaptive prokaryotic immunity conferred by CRISPR-Cas follows three stages.
1. Adaptation - Acquisition of foreign genetic material 2. Expression of Cas proteins and crRNA biogenesis and 3. Interference complex assembly, target recognition and target degradation.

tial uptake of foreign DNA [37]. A recent single-molecule study shed light on the exact mechanism by which the Cas1-Cas2 complex selects prespacers from this pool of fragments [40]. Suitable prespacers are selected based on the presence of a 3 bp protospacer adjacent motif (PAM) which is also required for the interference stage of CRISPR immunity [40, 41]. After capture by the Cas1-Cas2 complex, host factor nucleases process the overhangs present on the spacer precursor to the consensus length while integration into the CRISPR locus occurs in a step-wise manner [40, 42]. In Cas4 containing CRISPR systems this last processing step of the precursor is likely to be executed by either the Cas4 protein alone or a combination of host factor nucleases and Cas4. For a more detailed overview of CRISPR adaptation and the role Cas4 plays in this process see Chapter 2, Chapter 3 and Chapter 4 of this thesis.

1.3.2 Transcription and processing of CRISPR RNA

The genetic information that is acquired in the adaptation stage forms the core of CRISPR-mediated immunity. In order to make use of this genetic memory, the CRISPR array is transcribed yielding a long pre crRNA molecule containing the palindromic repeats as well as the viral fragments integrated previously (Fig. 1.2 - Stage II) [43-46]. Importantly, the leader sequence located upstream of the CRISPR array contains the promoter sequence that ultimately drives transcription [47]. Due to the palindromic nature of the repeats, the pre crRNA adopts a secondary stem loop structure which is required for recognition by the cognate processing endoribonuclease factor [48, 49]. Depending on the CRISPR system, three different mechanisms have been discovered that result in the generation of mature crRNA (for review see [50]. Two of the three mechanisms rely on processing by proteins from the Cas5 or Cas6 endoribonuclease superfamily. The Cas6 protein is the core processing subunit that binds the stem loop and cleaves the pre-crRNA within the repeat sequences. This maturation step yields a mature crRNA molecule that, in case of the type I-E system, contains the 32 nucleotide (nt) spacer flanked by a 8 nt 5' handle and a 21 nt 3' stem loop structure [46]. Another common processing factor is the Cas5d endoribonuclease that some CRISPR systems utilize which lack the *cas6* gene [50]. The Cas5d protein similarly recognizes and cleaves specific features of the repeat sequence, resulting in the cognate spacer flanked by parts of the repeat [51-53]. CRISPR systems that lack

both Cas6 and Cas5d processing factors either rely on processing by the non-Cas host factor RNase III [54], RNase E [55] or on processing by the CRISPR effector protein (e.g. Cas12 [56]).

Interestingly, Cas6 remains bound to the processed crRNA and act as the initiator of the Cascade complex assembly [46, 57, 58]. In case of the type I-E Cascade, the binding of Cas6 to the stem loop of the crRNA provides a docking point for the Cas7 protein forming the helical backbone of Cascade [59]. Following the Cas7 backbone assembly, the Cas5 subunit caps the 5' handle of the crRNA and acts as a binding site for the large subunit of the complex, the Cse1 protein [59, 60]. Lastly, the belly of the complex is formed by two Cse2 subunits [59, 60]. The assembled Cascade complex subsequently patrols the cell in order to initiate the interference stage upon binding a dsDNA molecule with complementarity to the crRNA.

1.3.3 CRISPR Interference

The CRISPR interference stage comprises the binding and subsequent cleavage of dsDNA that is complementary to the crRNA loaded in the Cascade complex (Fig 1.2 - Stage III). One of the major hurdles to overcome is the vast amount of DNA present in the cell that needs to be screened for the target DNA sequence (called protospacer). In order to prevent recognition and cleavage of host genomic DNA that otherwise would result in an autoimmunity response, the Cascade surveillance complex initially probes potential targets for the presence of a trinucleotide sequence called PAM. The PAM allows for the discrimination between actual protospacers present in invading viral DNA (the sequence from which the spacer is derived) and the spacers located in the CRISPR array. In the type I-E system the task of probing for the PAM is executed by the large Cas8e subunit of Cascade that via three structural features senses this trinucleotide motif by interacting with the minor groove of the DNA strand. The probing via minor groove interactions potentially allows for more promiscuous PAM recognition [61]. In contrast, the type II Cas9 protein senses the PAM through conserved arginine residues in the C-terminal that engage with the major groove, resulting in more stringent PAM recognition requirements [62]. Upon recognition of the respective PAM, in both type I and type II systems, the interference stage proceeds by initiation of an R-loop structure in which the dsDNA is uni-directionally unwound and the non-target strand displaced [63-66]. Interestingly, the decreased

cleavage activity with respect to mutated PAMs is more likely caused by altered R-loop formation kinetics rather than stability of the R-loop [67]. Unwinding occurs from the PAM proximal end of the protospacer, allowing for crRNA:protospacer hybridization from PAM proximal to PAM distal end of the protospacer. R-loop propagation occurs simultaneously with the crRNA:protospacer hybridization, leading to abortion of R-loop formation when mismatches between crRNA and target strand are encountered. Upon completion of R-loop formation, the type I-E Cascade undergoes a conformational change resulting in a locked state of the effector protein-DNA complex. This locking of the bound R-loop licenses DNA degradation by recruitment of the trans-acting Cas3 protein (in case of type I systems). The Cas9 endonuclease employs a similar yet different mechanism in order to induce DNA cleavage. Full hybridization between crRNA and protospacer leads to conformational changes of the catalytic HNH and RuvC domains of Cas9, eventually positioning the active sites such that each of the nuclease domains cleave one of the DNA strands [68]. This cleavage of the invading DNA ultimately abolishes further propagation of the targeted virus, interfering with the infection.

1.4 The type I-D CRISPR-Cas system

Until recently only little attention was paid to the CRISPR-Cas type I-D system that unites unique features of both type I and type III CRISPR systems (Fig 1.3) [29]. The feature that makes the type I-D system stand out from all other type I systems is that it contains the *cas10d* gene, a variant of the type III signature gene *cas10* [29]. In contrast to type III systems which encode Cas10 proteins involved in secondary messenger production, the PALM domain associated with this messenger production is inactivated in the type I-D Cas10d variant [69]. Interestingly, the *cas10d* gene contains an internal translation site, resulting in an additional small complex subunit (Cas11d) directly derived from the large Cas10d subunit transcript. The stoichiometry of the type I-D Cascade strongly resembles that of other type I systems, however, the overall Cascade architecture is more closely related to type III systems [69]. In contrast to type III systems, the type I-D system degrades both ssDNA and dsDNA rather than RNA [70]. Interestingly, the ssDNA degradation pattern resembles the RNA cleavage of type III systems, highlighting the hybrid type I and type III nature of the type I-D system [70] . Whereas target degradation in type

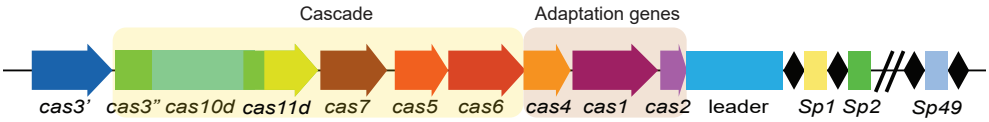


Figure 1.3 - Overview of the CRISPR-Cas type I-D locus as found in the cyanobacterium *Synechocystis* sp. 6803. The two functionally distinct modules (the interference Cascade module and the adaptation module) are highlighted in yellow and tan, respectively.

I systems relies on the recruitment of the nuclease-helicase Cas3, in the type I-D system this effector protein is split into two functionally distinct units [71]. The helicase subunit Cas3' is encoded separately, while the HD nuclease domain Cas3" is fused to Cas10d (Fig. 1.3) [71]. This unique arrangement has sparked the idea to utilize the type I-D system as a genome editing tool in plants and mammalian cells, harnessing the cleavage activity of the HD nuclease domain within the Cas10d subunit [72].

CRISPR adaptation of the type I-D system is mediated by two separate Cas4-Cas1 and Cas1-Cas2 complexes (see Chapter 4 of this thesis). The adaptation proteins are encoded in the adaptation module consisting of the *cas4, cas1* and *cas2* genes (Fig. 1.3). While spacer integration in the type I-D system only requires *cas1* and *cas2*, the selection of spacers that correspond to the consensus GTN PAM [73] is strongly dependent on the presence of *cas4* (see Chapter 3 of this thesis). Taken together, the type I-D system offers unique insights into CRISPR-Cas evolution and diversity as well as into the significance of the Cas4 protein in the context of CRISPR adaptation.

1.5 Thesis Outline

Chapter 2 on page 23: "CRISPR-Cas: Adapting to Change"

In **Chapter 2** we discuss how bacteria and archaea keep their genetic CRISPR memory updated. To keep up with the ever-changing pool of bacteriophages that result from a constant evolutionary arms race, numerous variations of the CRISPR adaptation theme have evolved. We review the current advances in our understanding of naïve and primed CRISPR adaptation. Furthermore, we highlight the involvement of different interference, adaptation and accessory proteins and provide a mechanistic overview of how an updated CRISPR immune status is maintained.

Chapter 3 on page 51: "Cas4 Facilitates PAM-Compatible Spacer Selection during CRISPR Adaptation"

In **Chapter 3** we investigate the role of the CRISPR-Cas accessory protein Cas4 from the cyanobacterial CRISPR-Cas type I-D system. By analyzing spacer sequences that were acquired in the presence and absence of Cas4, we demonstrate that Cas4 is crucially important for the acquisition of spacers conferring CRISPR immunity. Although the Cas1 and Cas2 adaptation proteins are sufficient for the integration of novel spacers, only spacers that were acquired in the presence of Cas4 correspond to the type I-D consensus GTN PAM. Our work explains the strong conservation of Cas4 during the evolution of CRISPR-Cas systems by directly contributing to functional anti-phage immunity.

Chapter 4 on page 77: "Cas4-Cas1 is a PAM-processing factor mediating half-site spacer integration during CRISPR adaptation"

In **Chapter 4** we biochemically reconstitute the type I-D adaptation module and elucidate the mechanism by which Cas4 contributes to functional spacer selection. We show that the Cas4 protein strongly interacts with the Cas1 integrase, forming a distinctive Cas4-Cas1 integration complex. This complex sequence specifically recognizes, processes and integrates prespacer substrates containing the type I-D PAM sequence. Additionally, we find a Cas1-Cas2 complex that aids

in the processing and integration of the non-PAM sites of prespacers. Taken together, our work results in a model that sheds light on the Cas4-dependendent spacer acquisition mechanism which ensures the integration of interference-proficient spacers.

Chapter 5 on page 111: "Conserved motifs in the CRISPR leader sequence control spacer acquisition levels in Type I-D CRISPR-Cas systems"

In **Chapter 5** we provide insights into conserved motifs within the type I-D leader sequence. We assess spacer integration efficiency with sequentially truncated leader sequences and find that spacer integration is significantly reduced when certain motifs are not included in the leader. By creating alignments with other type I-D leader sequences, we identify three conserved motifs that each contribute to the efficiency of spacer integration. In line with earlier leader characterization studies, we suggest that the identified motifs serve as recognition signals for the adaptation proteins. Guiding the adaptation proteins towards the CRISPR array facilitates spacer integration at the correct integration site and therefore allows for faster and more efficient CRISPR immunization.

Chapter 6 on page 129: "Cas3-derived target DNA degradation fragments fuel primed CRISPR adaptation"

In **Chapter 6** we demonstrate how the Cas3 helicase-nuclease protein connects the interference and adaptation stage, resulting in a positive feedback loop called primed adaptation. When the Cascade complex identifies a target sequence, it recruits the Cas3 effector protein for target degradation. Cas3 processes target DNA into short fragments enriched for thymine-stretches in their 3' overhangs. The Cas1-Cas2 integration complex captures Cas3-derived degradation fragments followed by further processing and integration into the CRISPR array. This work highlights how primed CRISPR adaptation is enhanced by the sequence specificity of Cas3 and Cas1-Cas2. The combined activities of effector and adaptation proteins increases the propensity of functional spacer integration, boosting the immune response against already catalogued invaders.

Appendix on page 180: "CRISPR-Cas Reduced to A Minimum"

In the Appendix we highlight the discoveries of Wright et al. (2019) and Edraki et al. (2019) which demonstrate that the architecture of CRISPR-Cas immunity can be condensed without losing functionality. Wright et al. provide insights into CRISPR adaptation solely relying on the Cas1 integrase protein while Edraki et al. characterize a small Cas9 variant with less stringent PAM requirements.

1

REFERENCES

1. Bergh, O., K.Y. Børsheim, G. Bratbak, and M. Heldal, High abundance of viruses found in aquatic environments. Nature, 1989. 340(6233): p. 467-8.
2. Chibani-Chennoufi, S., A. Bruttin, M.-L. Dillmann, and H. Brüssow, Phage-Host Interaction: an Ecological Perspective. Journal of Bacteriology, 2004. 186(12): p. 3677.
3. Weinbauer, M.G., Ecology of prokaryotic viruses. FEMS Microbiol Rev, 2004. 28(2): p. 127-81.
4. Labrie, S.J., J.E. Samson, and S. Moineau, Bacteriophage resistance mechanisms. Nat Rev Microbiol, 2010. 8(5): p. 317-27.
5. Dy, R.L., C. Richter, G.P. Salmond, and P.C. Fineran, Remarkable Mechanisms in Microbes to Resist Phage Infections. Annu Rev Virol, 2014. 1(1): p. 307-31.
6. Westra, E.R., D.C. Swarts, R.H. Staals, M.M. Jore, S.J. Brouns, and J. van der Oost, The CRISPRs, they are a-changin': how prokaryotes generate adaptive immunity. Annu Rev Genet, 2012. 46: p. 311-39.
7. Hyman, P. and S.T. Abedon, Bacteriophage host range and bacterial resistance. Adv Appl Microbiol, 2010. 70: p. 217-48.
8. Rakhuba, D.V., E.I. Kolomiets, E.S. Dey, and G.I. Novik, Bacteriophage receptors, mechanisms of phage adsorption and penetration into host cell. Pol J Microbiol, 2010. 59(3): p. 145-55.
9. Hyman, P., Phage Receptor, in Reference Module in Life Sciences. 2017, Elsevier.
10. Chopin, M.C., A. Chopin, and E. Bidnenko, Phage abortive infection in lactococci: variations on a theme. Curr Opin Microbiol, 2005. 8(4): p. 473-9.
11. Bickle, T.A. and D.H. Krüger, Biology of DNA restriction. Microbiol Rev, 1993. 57(2): p. 434-50.
12. Vasu, K. and V. Nagaraja, Diverse functions of restriction-modification systems in addition to cellular defense. Microbiol Mol Biol Rev, 2013. 77(1): p. 53-72.
13. Bonilla, F.A. and H.C. Oettgen, Adaptive immunity. J Allergy Clin Immunol, 2010. 125(2 Suppl 2): p. S33-40.
14. Ishino, Y., M. Krupovic, and P. Forterre, History of CRISPR-Cas from Encounter with a Mysterious Repeated Sequence to Genome Editing Technology. Journal of Bacteriology, 2018. 200(7): p. e00580-17.
15. Mojica, F.J.M. and F. Rodriguez-Valera, The discovery of CRISPR in archaea and bacteria. The FEBS Journal, 2016. 283(17): p. 3162-3169.
16. Ishino, Y., H. Shinagawa, K. Makino, M. Amemura, and A. Nakata, Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. Journal of bacteriology, 1987. 169(12): p. 5429-5433.
17. Mojica, F.J., G. Juez, and F. Rodríguez-Valera, Transcription at different salinities of Haloferax mediterranei sequences adjacent to partially modified PstI sites. Molecular microbiology, 1993. 9(3): p. 613-621.
18. Mojica, F.J., C. Ferrer, G. Juez, and F. Rodríguez-Valera, Long stretches of short tandem repeats are present in the largest replicons of the Archaea Haloferax mediterranei and Haloferax volcanii and could be involved in replicon partitioning. Molecular microbiology, 1995. 17(1): p. 85-93.
19. Jansen, R., J.D.A. van Embden, W. Gaastra, and L.M. Schouls, Identification of a Novel Family of Sequence Repeats among Prokaryotes. OMICS: A Journal of Integrative Biology, 2002. 6(1): p. 23-33.

20. She, Q., R.K. Singh, F. Confalonieri, Y. Zivanovic, G. Allard, M.J. Awayez, C.C.Y. Chan-Weiher, I.G. Clausen, B.A. Curtis, A. De Moors, G. Erauso, C. Fletcher, P.M.K. Gordon, I. Heikamp-de Jong, A.C. Jeffries, C.J. Kozera, N. Medina, X. Peng, H.P. Thi-Ngoc, P. Redder, M.E. Schenk, C. Theriault, N. Tolstrup, R.L. Charlebois, W.F. Doolittle, M. Duguet, T. Gaasterland, R.A. Garrett, M.A. Ragan, C.W. Sensen, and J. Van der Oost, The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. Proceedings of the National Academy of Sciences, 2001. 98(14): p. 7835.

21. Jansen, R., J.D.A.V. Embden, W. Gaastra, and L.M. Schouls, Identification of genes that are associated with DNA repeats in prokaryotes. Molecular microbiology, 2002. 43(6): p. 1565-75.

22. Mojica, F.J.M., C.s. Díez-Villaseñor, J.s. García-Martínez, and E. Soria, Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. Journal of Molecular Evolution, 2005. 60(2): p. 174-182.

23. Bolotin, A., B. Quinquis, A. Sorokin, and S.D. Ehrlich, Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology, 2005. 151(8): p. 2551-2561.

24. Pourcel, C., G. Salvignol, and G. Vergnaud, CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology, 2005. 151(3): p. 653-663.

25. Barrangou, R., C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D.A. Romero, and P. Horvath, CRISPR provides acquired resistance against viruses in prokaryotes. Science, 2007. 315(5819): p. 1709-12.

26. Brouns, S.J.J., M.M. Jore, M. Lundgren, E.R. Westra, R.J.H. Slijkhuis, A.P.L. Snijders, M.J. Dickman, K.S. Makarova, E.V. Koonin, and J. van der Oost, Small CRISPR RNAs guide antiviral defense in prokaryotes. Science (New York, N.Y.), 2008. 321(5891): p. 960-964.

27. Makarova, K.S., Y.I. Wolf, O.S. Alkhnbashi, F. Costa, S.A. Shah, S.J. Saunders, R. Barrangou, S.J.J. Brouns, E. Charpentier, D.H. Haft, P. Horvath, S. Moineau, F.J.M. Mojica, R.M. Terns, M.P. Terns, M.F. White, A.F. Yakunin, R.A. Garrett, J. van der Oost, R. Backofen, and E.V. Koonin, An updated evolutionary classification of CRISPR–Cas systems. Nature Reviews Microbiology, 2015. 13(11): p. 722-736.

28. Shmakov, S., O.O. Abudayyeh, K.S. Makarova, Y.I. Wolf, J.S. Gootenberg, E. Semenova, L. Minakhin, J. Joung, S. Konermann, K. Severinov, F. Zhang, and E.V. Koonin, Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. Mol Cell, 2015. 60(3): p. 385-97.

29. Makarova, K.S., D.H. Haft, R. Barrangou, S.J.J. Brouns, E. Charpentier, P. Horvath, S. Moineau, F.J.M. Mojica, Y.I. Wolf, A.F. Yakunin, J. van der Oost, and E.V. Koonin, Evolution and classification of the CRISPR–Cas systems. Nature Reviews Microbiology, 2011. 9(6): p. 467-477.

30. Koonin, E.V., K.S. Makarova, and F. Zhang, Diversity, classification and evolution of CRISPR-Cas systems. Current Opinion in Microbiology, 2017. 37: p. 67-78.

31. Makarova, K.S., Y.I. Wolf, and E.V. Koonin, Classification and Nomenclature of CRISPR-Cas Systems: Where from Here? The CRISPR Journal, 2018. 1(5): p. 325-336.

32. Barrangou, R., C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D.a. Romero, and P. Horvath, CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. Science, 2007. 315(5819): p. 1709-1712.

33. Fineran, P.C. and E. Charpentier, Memory of viral infections by CRISPR-Cas adaptive immune systems: Acquisition of new information. Virology, 2012. 434(2): p. 202-209.

34. Yosef, I., M.G. Goren, and U. Qimron, Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. Nucleic Acids Research, 2012. 40(12): p. 5569-5576.

35. Nuñez, J.K., P.J. Kranzusch, J. Noeske, A.V. Wright, C.W. Davies, and J.A. Doudna, Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive

immunity. Nature Structural & Molecular Biology, 2014. 21(6): p. 528-534.

36. Nuñez, J.K., A.S.Y. Lee, A. Engelman, and J.A. Doudna, Integrase-mediated spacer acquisition during CRISPR–Cas adaptive immunity. Nature, 2015. 519(7542): p. 193-198.

37. Levy, A., M.G. Goren, I. Yosef, O. Auster, M. Manor, G. Amitai, R. Edgar, U. Qimron, and R. Sorek, CRISPR adaptation biases explain preference for acquisition of foreign DNA. Nature, 2015. 520(7548): p. 505-510.

38. Taylor, A.F. and G.R. Smith, RecBCD enzyme is altered upon cutting DNA at a chi recombination hotspot. Proceedings of the National Academy of Sciences, 1992. 89(12): p. 5226.

39. Amundsen, S.K., A.F. Taylor, M. Reddy, and G.R. Smith, Intersubunit signaling in RecBCD enzyme, a complex protein machine regulated by Chi hot spots. Genes & development, 2007. 21(24): p. 3296-3307.

40. Kim, S., L. Loeff, S. Colombo, S. Jergic, S.J.J. Brouns, and C. Joo, Selective loading and processing of prespacers for precise CRISPR adaptation. Nature, 2020.

41. Shah, S.A., S. Erdmann, F.J. Mojica, and R.A. Garrett, Protospacer recognition motifs: mixed identities and functional diversity. RNA Biol, 2013. 10(5): p. 891-9.

42. Ramachandran, A., L. Summerville, B.A. Learn, L. DeBell, and S. Bailey, Processing and integration of functionally oriented prespacers in the *Escherichia coli* CRISPR system depends on bacterial host exonucleases. The Journal of biological chemistry, 2020. 295(11): p. 3403-3414.

43. Lillestøl, R.K., P. Redder, R.A. Garrett, and K. Brügger, A putative viral defence mechanism in archaeal cells. Archaea (Vancouver, B.C.), 2006. 2(1): p. 59-72.

44. Tang, T.H., N. Polacek, M. Zywicki, H. Huber, K. Brugger, R. Garrett, J.P. Bachellerie, and A. Hüttenhofer, Identification of novel non-coding RNAs as potential antisense regulators in the archaeon *Sulfolobus solfataricus*. Mol Microbiol, 2005. 55(2): p. 469-81.

45. Tang, T.H., J.P. Bachellerie, T. Rozhdestvensky, M.L. Bortolin, H. Huber, M. Drungowski, T. Elge, J. Brosius, and A. Hüttenhofer, Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus. Proc Natl Acad Sci U S A, 2002. 99(11): p. 7536-41.

46. Brouns, S.J.J., M.M. Jore, M. Lundgren, E.R. Westra, R.J.H. Slijkhuis, A.P.L. Snijders, M.J. Dickman, K.S. Makarova, E.V. Koonin, and J. van der Oost, Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. Science, 2008. 321(5891): p. 960-964.

47. Alkhnbashi, O.S., S.A. Shah, R.A. Garrett, S.J. Saunders, F. Costa, and R. Backofen, Characterizing leader sequences of CRISPR loci. Bioinformatics, 2016. 32(17): p. i576-i585.

48. Jansen, R., J.D.A.V. Embden, W. Gaastra, and L.M. Schouls, Identification of genes that are associated with DNA repeats in prokaryotes. Molecular Microbiology, 2002. 43(6): p. 1565-1575.

49. Kunin, V., R. Sorek, and P. Hugenholtz, Evolutionary conservation of sequence and secondary structures in CRISPR repeats. Genome Biology, 2007. 8(4): p. R61-R61.

50. Li, H., Structural Principles of CRISPR RNA Processing. Structure (London, England : 1993), 2015. 23(1): p. 13-20.

51. Garside, E.L., M.J. Schellenberg, E.M. Gesner, J.B. Bonanno, J.M. Sauder, S.K. Burley, S.C. Almo, G. Mehta, and A.M. MacMillan, Cas5d processes pre-crRNA and is a member of a larger family of CRISPR RNA endonucleases. Rna, 2012. 18(11): p. 2020-8.

52. Koo, Y., D. Ka, E.J. Kim, N. Suh, and E. Bae, Conservation and variability in the structure and function of the Cas5d endoribonuclease in the CRISPR-mediated microbial immune system. J Mol Biol, 2013. 425(20): p. 3799-810.

53. Nam, K.H., C. Haitjema, X. Liu, F. Ding, H. Wang, M.P. DeLisa, and A. Ke, Cas5d protein processes pre-crRNA and assembles into a *cas*cade-like interference complex in subtype I-C/Dvulg CRISPR-Cas system. Structure, 2012. 20(9): p. 1574-84.

54. Deltcheva, E., K. Chylinski, C.M. Sharma, K. Gonzales, Y. Chao, Z.A. Pirzada, M.R. Eckert, J. Vogel, and E. Charpentier, CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature, 2011. 471(7340): p. 602-7.

55. Behler, J., K. Sharma, V. Reimann, A. Wilde, H. Urlaub, and W.R. Hess, The host-encoded RNase E endonuclease as the crRNA maturation enzyme in a CRISPR–Cas subtype III-Bv system. Nature Microbiology, 2018. 3(3): p. 367-377.

56. Fonfara, I., H. Richter, M. Bratovič, A. Le Rhun, and E. Charpentier, The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. Nature, 2016. 532(7600): p. 517-521.

57. Jore, M.M., M. Lundgren, E. van Duijn, J.B. Bultema, E.R. Westra, S.P. Waghmare, B. Wiedenheft, U. Pul, R. Wurm, R. Wagner, M.R. Beijer, A. Barendregt, K. Zhou, A.P.L. Snijders, M.J. Dickman, J.A. Doudna, E.J. Boekema, A.J.R. Heck, J. van der Oost, and S.J.J. Brouns, Structural basis for CRISPR RNA-guided DNA recognition by Cascade. Nature Structural & Molecular Biology, 2011. 18(5): p. 529-536.

58. Wiedenheft, B., G.C. Lander, K. Zhou, M.M. Jore, S.J.J. Brouns, J. van der Oost, J.A. Doudna, and E. Nogales, Structures of the RNA-guided surveillance complex from a bacterial immune system. Nature, 2011. 477(7365): p. 486-489.

59. Jackson, R.N., S.M. Golden, P.B.G. van Erp, J. Carter, E.R. Westra, S.J.J. Brouns, J. van der Oost, T.C. Terwilliger, R.J. Read, and B. Wiedenheft, Crystal structure of the CRISPR RNA-guided surveillance complex from *Escherichia coli*. Science, 2014. 345(6203): p. 1473-1479.

60. Zhao, H., G. Sheng, J. Wang, M. Wang, G. Bunkoczi, W. Gong, Z. Wei, and Y. Wang, Crystal structure of the RNA-guided immune surveillance Cascade complex in *Escherichia coli*. Nature, 2014. 515(7525): p. 147-50.

61. Hayes, R.P., Y. Xiao, F. Ding, P.B.G. van Erp, K. Rajashankar, S. Bailey, B. Wiedenheft, and A. Ke, Structural basis for promiscuous PAM recognition in type I–E Cascade from *E. coli*. Nature, 2016. 530(7591): p. 499-503.

62. Anders, C., O. Niewoehner, A. Duerst, and M. Jinek, Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. Nature, 2014. 513(7519): p. 569-573.

63. Sinkunas, T., G. Gasiunas, S.P. Waghmare, M.J. Dickman, R. Barrangou, P. Horvath, and V. Siksnys, In vitro reconstitution of Cascade-mediated CRISPR immunity in *Streptococcus thermophilus*. The EMBO Journal, 2013. 32(3): p. 385-394.

64. Westra, E.R., P.B.G. van Erp, T. Künne, S.P. Wong, R.H.J. Staals, C.L.C. Seegers, S. Bollen, M.M. Jore, E. Semenova, K. Severinov, W.M. de Vos, R.T. Dame, R. de Vries, S.J.J. Brouns, and J. van der Oost, CRISPR Immunity Relies on the Consecutive Binding and Degradation of Negatively Supercoiled Invader DNA by Cascade and Cas3. Molecular Cell, 2012. 46(5): p. 595-605.

65. Gasiunas, G., R. Barrangou, P. Horvath, and V. Siksnys, Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proceedings of the National Academy of Sciences, 2012. 109(39): p. E2579.

66. Jinek, M., K. Chylinski, I. Fonfara, M. Hauer, J.A. Doudna, and E. Charpentier, A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. Science, 2012. 337(6096): p. 816.

67. Szczelkun, M.D., M.S. Tikhomirova, T. Sinkunas, G. Gasiunas, T. Karvelis, P. Pschera, V. Siksnys, and R. Seidel, Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. Proceedings of the National Academy of Sciences, 2014. 111(27): p. 9798-9803.

68. Sternberg, S.H., B. LaFrance, M. Kaplan, and J.A. Doudna, Conformational control of DNA target cleavage by CRISPR-Cas9. Nature, 2015. 527(7576): p. 110-113.

69. McBride, T.M., E.A. Schwartz, A. Kumar, D.W. Taylor, P.C. Fineran, and R.D. Fagerlund, Diverse CRISPR-Cas Complexes Require Independent Translation of Small and Large Subunits from a Single Gene. Mol Cell, 2020. 80(6): p. 971-979.e7.

70. Lin, J., A. Fuglsang, A.L. Kjeldsen, K. Sun, Y. Bhoobalan-Chitty, and X. Peng, DNA targeting by subtype I-D CRISPR–Cas shows type I and type III features. Nucleic Acids Research, 2020. 48(18): p. 10470-10478.

71. Makarova, K.S., L. Aravind, Y.I. Wolf, and E.V. Koonin, Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. Biology Direct, 2011. 6(1): p. 38.

72. Osakabe, K., N. Wada, T. Miyaji, E. Murakami, K. Marui, R. Ueta, R. Hashimoto, C. Abe-Hara, B. Kong, K. Yano, and Y. Osakabe, Genome editing in plants using CRISPR type I-D nuclease. Commun Biol, 2020. 3(1): p. 648.

73. Shah, S.A., S. Erdmann, F.J.M. Mojica, and R.A. Garrett, Protospacer recognition motifs. RNA Biology, 2013. 10(5): p. 891-899.

2

CRISPR-Cas: Adapting to change

Science

2017 Apr 7;356(6333).

Simon A. Jackson[1†], Rebecca E. McKenzie[2†], Robert D. Fagerlund[1], Sebastian N. Kieper[2], Peter C. Fineran[1,3*] and Stan J.J. Brouns[2,4*]

1. Department of Microbiology and Immunology, University of Otago, Post Office Box 56, Dunedin 9054, New Zealand.

2. Department of Bionanoscience, Kavli Institute of Nanoscience, Delft University of Technology, Van der Maasweg 9, 2629 HZ Delft, Netherlands.

3. Bio-Protection Research Centre, University of Otago, Post Office Box 56, Dunedin 9054, New Zealand.

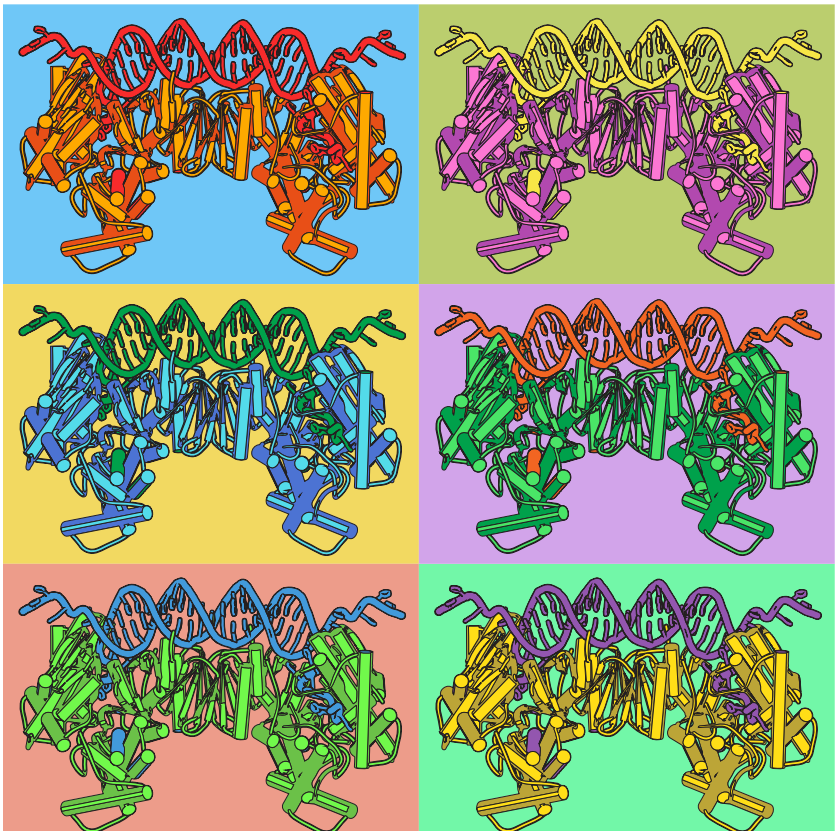
4. Laboratory of Microbiology, Wageningen University, Wageningen, Netherlands.

†These Authors contributed equally

* Corresponding Authors

2.1 Abstract

Bacteria and archaea are engaged in a constant arms race to defend against the ever-present threats of viruses and invasion by mobile genetic elements. The most flexible weapons in the prokaryotic defense arsenal are the CRISPR-Cas adaptive immune systems, which are capable of selective identification and neutralization of foreign elements. CRISPR-Cas systems rely on stored genetic memories to facilitate target recognition. Thus, to keep pace with a changing pool of hostile invaders, the CRISPR memory banks must be regularly updated by the addition of new information, through a process termed adaptation. In this review, we outline the recent advances in our understanding of the molecular mechanisms governing adaptation and highlight the diversity between systems.



2.2 Adaptive immunity in prokaryotes

Bacteria and archaea are constantly threatened by phage infection and invasion by mobile genetic elements (MGEs) through conjugation and transformation. In response, a defense arsenal has evolved, including various 'innate' mechanisms and the CRISPR-Cas adaptive immune systems [1-3]. CRISPR-Cas systems are widely distributed, present in 50% and 87% of complete bacterial and archaeal genomes, respectively, and are classified into two major classes consisting of 6 types according to their Cas proteins [4, 5]. CRISPR-Cas systems function as RNA-guided nucleases that provide sequence-specific defense against invading MGEs [6, 7]. Their repurposing, particularly Cas9, has stimulated a biotechnological revolution in genome editing that has resulted in breakthroughs across many biological fields [8]. In native hosts, the advantage conferred by CRISPR-Cas systems over innate defenses lies in the ability to update their resistance repertoire in response to infection (termed CRISPR adaptation). Adaptation is achieved by incorporating short DNA fragments from MGEs into CRISPR arrays to form memory units termed spacers, which are subsequently transcribed and processed to CRISPR RNAs (crRNAs) (Fig. 2.1). Cas proteins associate with crRNAs to form crRNA-effector complexes, which seek and destroy invading MGEs. Thus, adaptation of CRISPR arrays is a crucial process required to ensure persistent CRISPR-Cas defense [9, 10].

Adaptation in nature appears widespread, highlighting the dynamic interaction between hosts and invaders [11-13]. When a prokaryotic community undergoes CRISPR adaptation, individual cells acquire different, and often multiple spacers. This population diversity increases defense by limiting the reproductive success of MGE variants that evade recognition through genetic mutations (escape mutants) [14]. The CRISPR polymorphisms resulting from adaptation enable differentiation of species subtypes, including economically and clinically relevant isolates, and allow tracking of pathogen outbreaks [15, 16].

Typically, new spacers are inserted at one end of the array in a position closest to the promoter driving CRISPR transcription – termed the leader (Fig. 2.1) [6, 17-19]. This polarization of the CRISPR records provides a chronological account of the battle between phages and bacteria, analyses of which can provide insights into phage-host co-oc-

currences, evolution and ecology [20, 21]. Moreover, spacer integration at the leader end enhances defense against recently encountered MGEs, potentially due to elevated crRNA abundance [22]. However, in some systems, the repeats themselves contain internal promoters, which might make leader-proximal spacer integration less important [23]. CRISPR arrays typically contain 10-30 spacers, but some species contain arrays with over 500 spacers [24]. Spacers that may no longer be under evolutionary selection can be lost via recombination between CRISPR repeats [11, 25].

Early bioinformatic studies showed many spacers were of foreign origin, hinting that CRISPR loci would form the memory of an immune system [15, 26-28]. Subsequent confirmation of this link between spacers and resistance to phage and MGEs was gained experimentally [6, 7, 29]. Despite the elegance of memory-directed defense, CRISPR adaptation is not without complications. Paradoxically, the spacers required for defense must be added to CRISPRs during exposure to MGEs [30, 31]. In addition, the inadvertent acquisition of spacers from host DNA must be avoided because this will result in cytotoxic self-targeting – akin to autoimmunity [32, 33]. Recently, significant progress has been made toward understanding the molecular mechanisms governing how, when and why CRISPR spacers are acquired. Here, we review these studies and highlight the insights they shed on both the function and evolution of CRISPR-Cas systems.

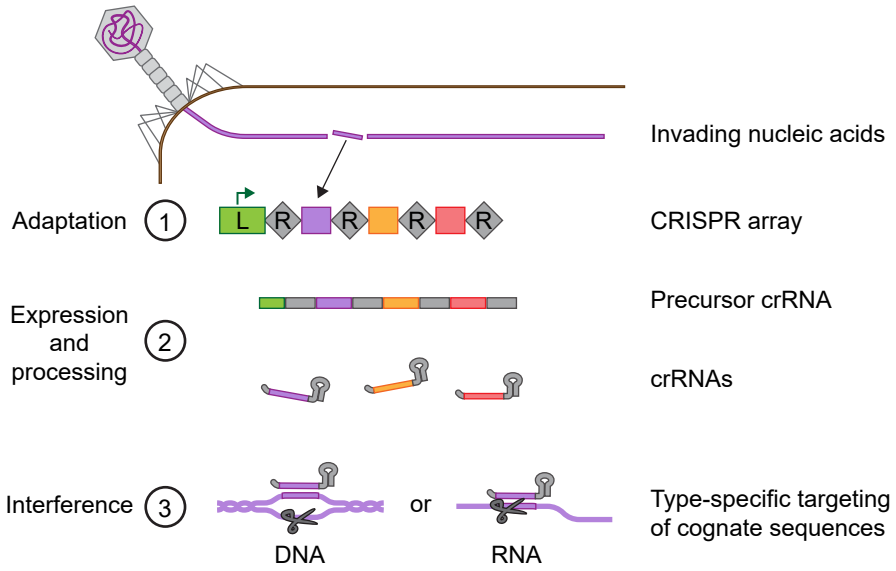


Figure 2.1: CRISPR-Cas adaptation and defense. A simplified schematic of CRISPR-Cas defense, which consists of an array of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and CRISPR-associated (Cas) proteins encoded by *cas* genes (omitted for clarity). CRISPR-Cas defense consists of three defined stages **1** Adaptation, the creation of memory of prior infections formed via the insertion of small foreign DNA sequences into the leader (L) end of the CRISPR array, where they are stored as spacers (colored squares) between duplicated repeats (R). **2** Expression and CRISPR-RNA (crRNA) biogenesis, the transcription and processing of the array into small guide RNA sequences. **3** Interference, degradation of the target foreign invader by sequence-specific binding and cleavage.

2.3 Molecular mechanism of adaptation

At the forefront of adaptation are Cas1 and Cas2 proteins, which form a Cas14-Cas22 complex [34, 35] (hereafter Cas1-Cas2) – the 'workhorse' of spacer integration (Fig. 2.2). Illustrative of their key roles in spacer integration, the *cas1* and *cas2* genes are associated with nearly all CRISPR-Cas systems [4]. Cas1-Cas2-mediated spacer integration prefers dsDNA substrates and proceeds via a mechanism resembling retroviral integration [36, 37]. In addition to Cas1-Cas2, a single repeat, at least part of the leader sequence [17, 18, 22, 38], and additional host factors for repair of the insertion sites (e.g. DNA polymerase) are required [39]. Spacer integration requires three main processes: 1) substrate capture 2) recognition of the CRISPR locus and 3) integration within the array.

2.3.1 Substrate capture

During substrate capture, Cas1-Cas2 is loaded with an integration-compatible pre-spacer, which is thought to be partially duplexed DNA. In the Cas1-Cas2:pre-spacer complex, each single-stranded 3'OH end of the pre-spacer DNA extends into a single active subunit of each Cas1 dimer [40] located either side of a central Cas2 dimer [41, 42] (Fig. 2.2). The branch points of the splayed DNA are stabilized by a Cas1 wedge, which acts as a molecular ruler to control spacer length. Although it is likely that Cas1-Cas2 rulers exist and measure different spacer sizes in all systems, the mechanism has only been demonstrated in the *Escherichia coli* type I-E system, where two tyrosine residues bookend the core 23 nt dsDNA region [41, 42]. Details of how pre-spacer substrates are produced from foreign DNA is discussed later.

2.3.2 Recognition of the CRISPR locus

Prior to integration, the substrate-bound Cas1-Cas2 complex must locate the CRISPR leader-repeat sequence. Adaptation complexes of several systems display intrinsic affinity for the leader-repeat region *in vitro* [36, 43], yet this is not always wholly sufficient to provide the specificity observed *in vivo*. For the type I-E system, leader-repeat recognition is assisted by the integration host factor (IHF) heterodimer, which binds in the leader [44]. IHF binds DNA in a sequence-specific manner and induces ~120° DNA bending, providing a cue to accu-

rately localize Cas1-Cas2 to the leader-repeat junction [44, 45]. A conserved leader motif upstream of the IHF pivot is proposed to stabilize the Cas1-Cas2-leader-repeat interaction and increase adaptation efficiency, supporting bipartite binding of the adaptation complex to DNA sites either side of bound IHF [45].

IHF is absent in many prokaryotes, including archaea and gram-positive bacteria, suggesting other leader-proximal integration mechanisms exist. Indeed, type II-A Cas1-Cas2 from *Streptococcus pyogenes* catalyzed leader-proximal integration *in vitro*, at a level of precision comparable to the type I-E system with IHF [43, 44]. Hence, type II-A systems may rely solely on intrinsic sequence specificity for the leader-repeat. A short leader-anchoring site (LAS) adjacent to the first repeat and ~6 bp of this repeat were essential for adaptation [22, 38, 43] and are conserved in systems with similar repeats. Placement of an additional LAS in front of a non-leader repeat resulted in adaptation at both sites [38], whereas LAS deletion caused ectopic integration at a downstream repeat adjacent to a spacer containing a LAS-like sequence [22]. Taken together, this shows specific sequences upstream of CRISPR arrays direct leader-polarized spacer integration, both via direct Cas1-Cas2 recognition and assisted by host proteins, such as IHF.

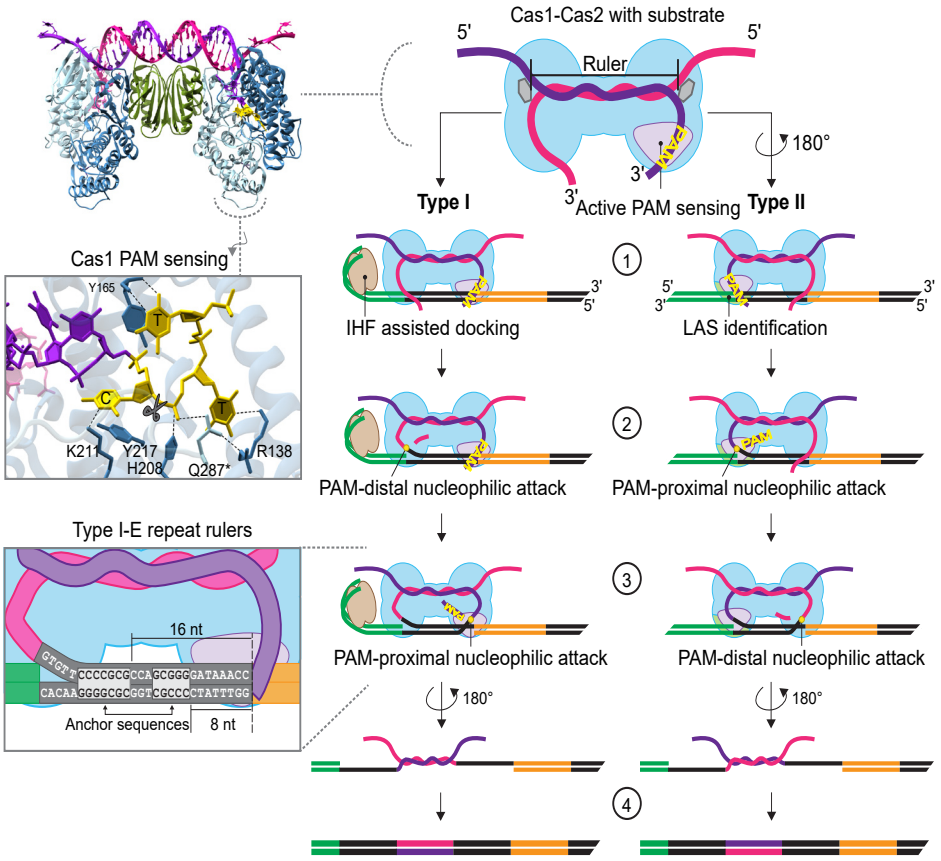


Figure 2.2: Cas1-Cas2-mediated spacer acquisition. The substrate loaded Cas1-Cas2 protein complex (*E. coli* type I-E structure shown top left; PDB 5DQZ) with the active PAM sensing domain highlighted (light purple) and a partially duplexed DNA pre-spacer substrate (strands are purple and pink) [41, 42]. The Cas1 PAM sensing insert shows the canonical type I-E PAM (CTT), residue-specific interactions (a residue from the non-catalyic Cas1 monomer is annotated with *), and site of PAM processing (scissors). The ruler mechanism determining spacer length for the type I-E systems uses two conserved tyrosine residues (grey hexagons). Spacer integration proceeds as follows: **1** the Cas1-Cas2:pre-spacer complex binds the leader (green) and first repeat (black). **2** The first nucleophillic attack occurs at the leader-repeat junction and gives rise to a half-site intermediate. **3** The second nucleophillic attack occurs at the repeat-spacer (orange) boundary resulting in full site integration. The type I-E repeat is magnified (lower left) to indicate the inverted repeats within its sequence and highlight the anchoring sites of the molecular rulers that determine the point of integration. **4** Host DNA repair enzymes fill the intergration site. For additional details, see the text.

2.3.3 Integration into the CRISPR array

In almost all types of CRISPR-Cas systems, the presence of a short sequence motif in the target nucleic acid adjacent to where the crRNA basepairs is essential for interference (the target-strand that the crRNA pairs to is known as the protospacer) (Fig. 2.3) [46]. This sequence motif is termed a protospacer adjacent motif (PAM) and is a key feature for spacer selection during adaptation [17, 27, 47, 48]. Acquisition of interference-proficient spacers requires processing of the pre-spacer substrate at a specific position relative to a PAM and also integration into the CRISPR array in the correct orientation. The active site of each Cas1 monomer contains a PAM sensing domain [41, 42] and the presence of a PAM within the pre-spacer substrate ensures integration in the appropriate orientation [49-51]. Accordingly, PAM proximal processing, resulting in complete or partial (in the case of type I-E) removal of the PAM, is likely to occur after Cas1-Cas2 orients and docks at the leader-repeat. In contrast, if complete processing occurred before docking to the CRISPR locus, then the PAM directionality cue would be lost. Cas1-mediated processing of the pre-spacer creates two 3'OH ends required for nucleophilic attack on each strand of the leader-proximal repeat [36, 37, 52]. The initial nucleophilic attack most likely occurs at the leader-repeat junction and forms a half-site intermediate, then a second attack at the existing repeat-spacer junction generates the full-site integration product (Fig. 2.2). The precise order of the pre-spacer processing and integration steps remains to be fully determined, yet considerable progress toward elucidating the reaction mechanisms has been made.

Following the first nucleophilic attack, Cas1-Cas2 employs molecular rulers that harness the intrinsic sequence-specificity of the complex to define the site of the second attack and ensure accurate repeat length duplication. CRISPR repeats are often semi-palindromic, containing two short inverted repeat (IR) elements, but the location of these can vary [53]. In type I-B and I-E systems, the IRs occur close to the center of the repeat (Fig. 2.2) and are important for adaptation [54, 55]. In the type I-E system, both IRs act as anchors for the Cas1-Cas2 complex, positioning the active site for the second attack at the repeat-spacer boundary [54]. However, in the type I-B system from Haloarcula hispanica, only the first IR was essential for integration, and thus a single molecular ruler directed by an anchor between the IRs was proposed [55]. In contrast, in the type II-A systems of *Strep-*

tococcus thermophilus and *S. pyogenes* the IRs are located distally within the repeats, suggesting these short sequences may directly position the nucleophilic attacks without molecular rulers [38, 43]. Although further work is required to determine how the spacer integration events are directed in different CRISPR-Cas systems, it seems likely the conserved leader-repeat regions at the beginning of CRISPR arrays maintain recognizable sequences to ensure Cas1-Cas2 localizes appropriately and spacer insertion and repeat duplication is of the correct length.

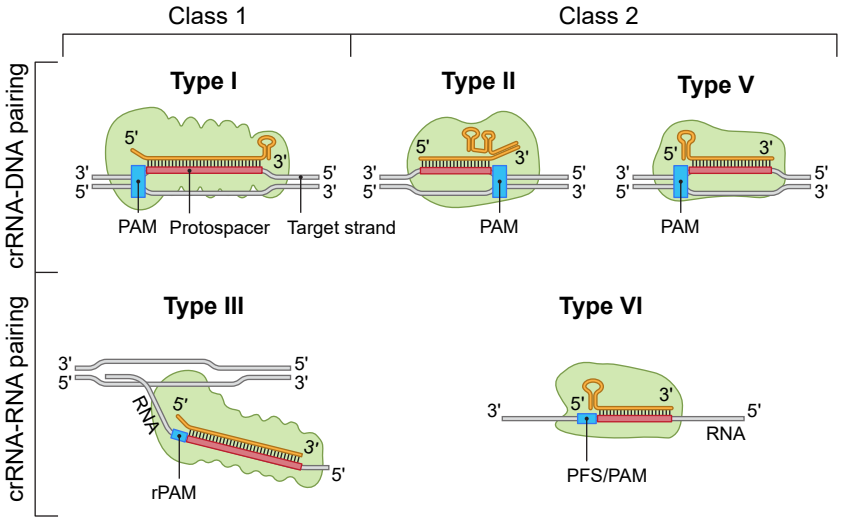


Figure 2.3: Target interactions and the PAMs of different CRISPR-Cas types. DNA targets are recognized by the crRNA-effector complexes of types I, II and V, resulting in formation of an R-loop with the non-target strand displaced. The target strand contains the protospacer (red), which is complementary to the spacer (crRNA, orange) sequence. The protospacer adjacent motif (PAM, blue) is located at either the 3' end of the protospacer (type I and type V) or the 5' end (type II). The PAM assignment is consistent with target-centric nomenclature [46]. Type III and VI recognize RNA targets, with type III exhibiting transcription-dependent DNA targeting. Some type III systems require an RNA-based PAM (rPAM). Type VI systems exhibit a protospacer flanking sequence (PFS) specificity, which is analogous to a PAM.

2.4 Production of spacers from foreign DNA

2.4.1 Naive adaptation

Acquisition of spacers from MGEs that are not already catalogued in host CRISPRs is termed naïve adaptation [56] (Fig. 2.4). To facilitate naïve adaptation, pre-spacer substrates are generated from foreign material and loaded onto Cas1-Cas2. Currently, the main known source of these precursors is the host RecBCD complex [57]. Stalled replication forks that occur during DNA replication can result in double strand breaks (DSBs), which are repaired via RecBCD-mediated unwinding and degradation of the dsDNA ends back to the nearest Chi sites [58]. During this process, RecBCD produces ssDNA fragments that are proposed to anneal, forming substrates suitable for use by Cas1-Cas2 [57]. Loading of substrates into Cas1-Cas2 is likely enhanced by interaction between Cas1 and RecBCD [59], positioning the adaptation machinery adjacent to the site of substrate generation. The increased number of active origins of replication and the paucity of Chi sites on MGEs, versus the host chromosome, biases naïve adaptation toward foreign DNA. Furthermore, RecBCD recognizes unprotected dsDNA ends, which are commonly present in phage genomes upon injection or prior to packaging, thereby providing an additional phage-specific source of naïve adaptation substrates [57, 60].

Despite the clear role of RecBCD in substrate generation, naïve adaptation also occurs in its absence, albeit with reduced bias toward foreign DNA [57]. Events other than DSBs might also stimulate naïve adaptation, such as R-loops that prime plasmid replication [61], lagging ends of incoming conjugative elements [62], and even CRISPR-Cas mediated spacer integration events themselves [51, 57]. Furthermore, it is unknown whether all CRISPR-Cas systems display an intrinsic adaptation bias towards foreign DNA. Complicating results, spacer acquisition from the host genome in native systems could be underestimated because the resulting self-targeting means these genotypes are typically lethal [32, 33, 51, 63]. For example, in the *S. thermophilus* type II-A system, adaptation appears biased toward MGEs, yet nuclease-deficient Cas9 (dCas9) failed to discriminate between acquisition from host versus foreign DNA [63] and it is unknown whether the adaptation was reliant on DNA break repair. Further studies in a range of

host systems are required to clarify how diverse CRISPR-Cas systems balance the requirement for naïve adaptation from MGEs against the risk of self-acquisition events.

2.4.2 crRNA-directed adaptation (Priming)

Mutations in the target PAM or protospacer sequences can abrogate immunity, allowing MGEs to escape CRISPR-Cas defenses [47, 64, 65]. Furthermore, the immunological effectiveness of individual spacers varies: often several target-specific spacers are required to both mount an effective defense [66, 67] and prevent proliferation of MGE escape mutants [13, 14]. Thus, CRISPR-Cas systems need to adapt faster than the foreign element can evade targeting. Indeed, type I systems have evolved a mechanism known as primed adaptation (priming) to facilitate rapid CRISPR adaptation [68, 69], even against highly divergent invaders [65] (Fig. 2.4). In contrast to naïve adaptation, priming utilizes target recognition by crRNAs from pre-existing spacers to direct spacer acquisition toward invaders whose proliferation exceeds the existing defense capabilities. This often occurs with MGE escape mutants, but also when the CRISPR-Cas expression level is insufficient to provide immunity – even with spacers perfectly targeting the MGE

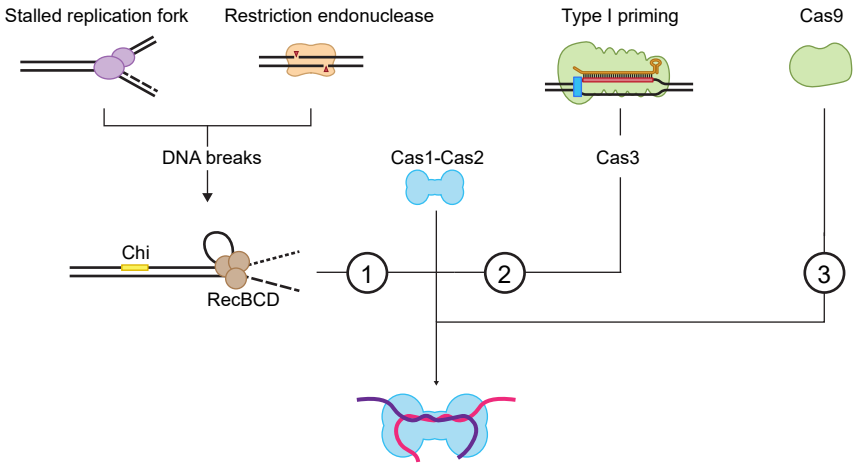


Figure 2.4: Cas1-Cas2 substrate production pathways. **1** Naïve generation of substrates by RecBCD activity on DNA ends resulting from DSBs from stalled replication forks, innate defenses such as restriction endonuclease activity or from the ends of phage genomes (not shown). **2** Primed substrate production in type I systems. **3** Cas9-dependent spacer selection in type II systems. For details, see the text.

[65, 68-72].

Priming begins with target recognition by crRNA-effector complexes. Therefore, factors that influence target recognition (i.e. the formation and stability of the R-loop – see Fig. 2.3), including PAM sensing and crRNA:target complementarity, affect the efficiency of primed adaptation [64, 65, 67, 73-80]. Furthermore, these same factors influence conformational rearrangements in the target-bound crRNA-effector complex, coalescing to favor either interference or priming [67, 74, 75, 78, 81]. In type I-E systems, the Cas8e (Cse1) subunit of Cascade can adopt one of two conformational modes [78, 81], which may promote either direct or Cas1-Cas2-stimulated recruitment of the effector Cas3 nuclease [74, 75, 81].

Cas3, found in all type I systems, exhibits 3' to 5' helicase and endonuclease activity that nicks, unwinds and degrades target DNA [82-85]. *In vitro* activity of the type I-E Cas3 produces ssDNA fragments of ~30-100 nucleotides that are enriched for PAMs in their 3' ends, which anneal to provide partially duplexed pre-spacer substrates [73]. The spatial positioning of Cas1-Cas2 during primed substrate generation has not been clearly established, although Cas1-Cas2-facilitated recruitment of Cas3 would imply the adaptation machinery is localized close to the site of substrate production [74, 81]. In support of this, Cas3 in type I-F systems is fused to the C-terminus of Cas2 and forms a Cas1-Cas2-3 complex [35] that couples the adaptation machinery directly to the source of substrate generation during primed adaptation [51, 86].

Despite different crRNA-effector:target interactions favoring distinct Cas3 recruitment modes, primed adaptation can occur from both escape mutants and interference-proficient targets [51, 68, 69, 87]. When target copy-number influences are excluded for type I-E and type I-F systems, interference-proficient targets promote stronger spacer acquisition than escape targets [51, 87]. This provides a positive feedback loop, reinforcing immunity against recurrent threats even in the absence of escapees [51, 69]. However, because target interference rapidly destroys the invader, more spacer acquisition is provoked by escape mutants where replication of the MGE outpaces its destruction. Over time, the prolonged presence of the invader, combined with the priming-centric target recognition mode, results in higher net production of pre-spacer substrates from escape mutants [51, 72, 73, 87].

Because priming initiates with site-specific target recognition (i.e. targeting a 'priming' protospacer), Cas1-Cas2 compatible substrates are subsequently produced from MGEs with locational biases (Fig. 2.5). Mapping the MGE sequence positions and strands targeted by newly acquired spacers (i.e. their corresponding protospacers) revealed subtype-specific patterns and has provided much of our insight into the priming mechanisms [50, 51, 68, 69, 86, 88, 89]. In type I-E systems, new protospacers map to the same strand [50, 69] as the priming protospacer (Fig. 2.5). For type I-B priming, Cas3 is predicted to load onto either strand at the priming protospacer, resulting in a bidirectional distribution of new protospacers [88]. For type I-F priming, the first new protospacer typically maps to the strand opposite the priming protospacer, in a direction consistent with Cas3 loading and helicase activity on the non-target strand. Furthermore, once the first spacer is acquired, two targets in the MGE will be recognized and substrate production can be driven from both locations [51, 86] (Fig. 2.5). However, in a head-to-head contest interference-proficient targets dominate, thus, subsequent spacers (i.e. the second and beyond) generally result from targeting by the first new spacer and are typically located back towards the original priming protospacer [51] (Fig. 2.5). The dominance of the first new spacer also holds true for type I-E [69, 87] and likely all other systems that display priming. However, these are generalized models and many questions remain unresolved, such as the mechanisms resulting in strand selection and why some spacer sequences are more highly acquired from MGEs than others. Further analyses of priming in different systems, particularly the order of new spacers acquired, will greatly inform our understanding of primed Cas1-Cas2 substrate production.

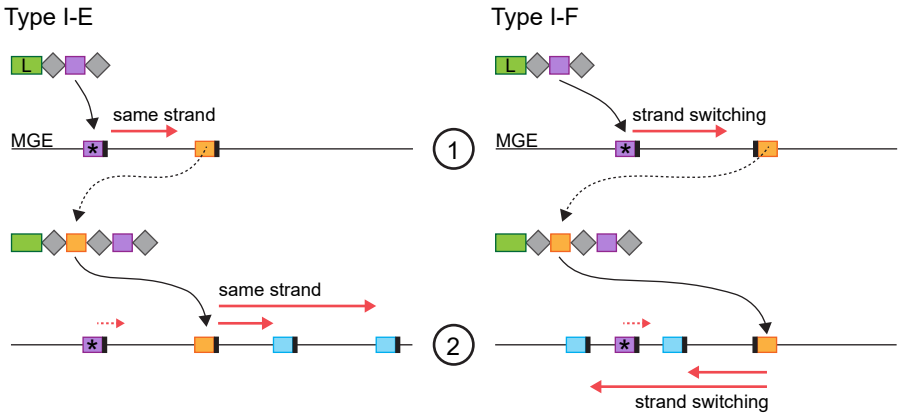


Figure 2.5: Primed adaptation from a multi-copy MGE by type I-E and I-F CRISPR-Cas systems. **1** An existing spacer (purple) with homology to an MGE sequence that has escaped interference (the 'priming' protospacer denoted with an asterisk) directs target recognition – the PAM adjacent to the protospacer is shown in black (PAMs at the right or left of protospacers indicate the strand each protospacer is on). The crRNA-effector complex recruits Cas3 and the 3' to 5' helicase activity (illustrated by the red arrow) results in the acquisition of a new spacer that maps to a protospacer (orange) from a site distal to the initial priming location. **2** The new interference-proficient spacer directs targeting of the MGE and recruitment of Cas3. Hence, subsequent spacers (mapping to blue protospacers) typically originate from Cas3 activity (red arrows) beginning at this location. See text for details.

2.4.3 Cas protein-assisted production of spacers

Given the apparent advantages conferred by priming in type I systems, mechanisms to utilize existing spacers to direct adaptation are likely to exist in other CRISPR-Cas types. For example, DNA breaks induced by interference activity of class 2 CRISPR-Cas effector complexes could trigger host DNA repair mechanisms (e.g. RecBCD), thereby providing substrates for Cas1-Cas2. In agreement with a generalized DNA break-stimulated adaptation model, restriction enzyme activity stimulated RecBCD-facilitated adaptation [57]. This may also partially account for the enhanced adaptation observed during phage infection of a host possessing an innate defense restriction-modification system [31], but whether this was RecBCD-dependent is unknown. For CRISPR-Cas-induced DNA breaks, spacer acquisition would be preceded by target recognition, hence the resulting adaptation could be considered related to 'priming' [90]. Although direct evidence to support this concept is lacking, adaptation in type II-A systems requires Cas1-Cas2,

Cas9, a tracrRNA and Csn2 [63, 90]. In support of a role for Cas9 in substrate generation, the PAM-sensing domain of Cas9 enhances the acquisition of spacers with compatible PAMs [90]. However, Cas9 nuclease activity is dispensable [63] and existing spacers are not strictly necessary [90], suggesting that PAM interactions of Cas9 could be sufficient to select appropriate new spacers. Some Cas9 variants can also function with non-CRISPR RNAs and tracrRNA [91], raising the possibility that host or MGE-derived RNAs might direct promiscuous Cas9 activity, resulting in DNA breaks, or replication fork stalling and trigger spacer integration.

2.5 Roles of accessory Cas proteins in adaptation

Although Cas1 and Cas2 play a central role in adaptation, type-specific variations in *cas* gene clusters occur. In many systems, Cas1-Cas2 is assisted by accessory Cas proteins, which are often mutually exclusive and type-specific [4]. For example, in the *S. thermophilus* type II-A system, deletion of *csn2* impaired the acquisition of spacers from invading phages [6]. Csn2 assembles into ring-shaped homo-tetramers with a calcium-stabilized central channel [92, 93] that binds cooperatively to the free ends of linear dsDNA and can translocate by rotation-coupled movement [94, 95]. Given that substrate-loaded type II-A Cas1-Cas2 is capable of full-site spacer integration *in vitro* [43], Csn2 is likely to play an earlier role in either pre-spacer substrate production, selection or processing. Potentially, Csn2 binding to the free ends of dsDNA provides a cue to direct nucleases necessary for substrate generation [94].

Cas4, another ring-forming accessory protein, is found in type I, II-B and V systems [4]. Confirming its role in adaptation, Cas4 is necessary for type I-B priming in *H. hispanica* [88] and interacts with a Cas1-2 fusion protein in the *Thermoproteus tenax* type I-A system [96]. Fusions between Cas4 and Cas1 are found in several systems, supporting a functional association with adaptation. Cas4 contains a RecB-like domain and four conserved cysteine residues, which are presumably involved in the coordination of an iron-sulfur cluster [97]. However, Cas4 proteins appear to be functionally diverse with some possessing uni- or bi-directional exonuclease activity [97, 98], while others exhibit ssDNA endonuclease activity and unwinding activity on

dsDNA [98]. Due to its nuclease activity, Cas4 is hypothesized to trim pre-spacer substrates and aid adaptation by generating 3' overhangs in the duplex pre-spacer substrate.

To provide immunity, type III systems require spacers complementary to RNA transcribed from MGEs (Fig. 2.3) [99, 100]. Some bacterial type III systems contain fusions of Cas1 with reverse transcriptase domains (RTs), which provide a mechanism to integrate spacers from RNA substrates [101]. The RT-Cas1 fusion from *M. mediterranea* can integrate RNA precursors into an array, which are subsequently reverse transcribed to generate DNA spacers [101]. However, integration of DNA-derived spacers also occurs, indicating that the RNA derived-spacer route is not exclusive [101]. Hence, the integrase activity of RT-Cas1-Cas2 is extended by the reverse transcriptase activity, enabling enhanced build-up of immunity against highly transcribed DNA MGEs and potentially from RNA-based invaders.

Despite evidence that accessory Cas proteins are involved in spacer acquisition, their roles mostly remain elusive. Furthermore, other host proteins may also be required for pre-spacer substrate production. For example, RecG is required for efficient primed adaptation in type I-E and I-F systems, but its precise role remains speculative [39, 102]. Additionally, it remains enigmatic why some CRISPR-Cas systems appear to require accessory proteins, whilst closely related types do not. For example, type II-C systems lack *cas4* or *csn2* that assist in type II-A and II-B adaptation, respectively. These type-specific differences exemplify the diversity that has arisen during evolution of CRISPR-Cas systems.

2.6 Evolution of adaptation

The expanding knowledge of spacer integration has led to a promising theory for the evolutionary origin of CRISPR-Cas systems [103]. Casposons are transposon-like elements typified by the presence of Cas1 homologs, casposases, which catalyze site-specific DNA integration and result in the duplication of repeat sites analogous to CRISPR adaptation [104, 105]. It is proposed that ancestral innate defenses gained DNA integration functionality from casposases, seeding the genesis of prokaryotic adaptive immunity [106]. The innate ancestor remains to be determined, but is likely to be a nuclease-based system. Co-occurrence of casposon-derived terminal inverted repeats and cas-

posases in the absence of full casposons might represent an intermediate of the CRISPR signature repeat-spacer-repeat structures [107]. However, the evolutionary journey from the innate immunity-casposase hybrid to full adaptive immunity remains unclear. Nevertheless, comparative genomics indicate that all known CRISPR-Cas systems evolved from a single ancestor [4, 5].

The more compact class 2 CRISPR-Cas systems likely evolved from class 1 ancestors, through acquisition of genes encoding new single-subunit effector proteins and loss of additional *cas* genes [5]. Evolution of CRISPR-Cas types would have required stringent co-evolution of the adaptation machinery, leader-repeat sequences [108], crRNA processing mechanisms and effector complex function. However, despite the subsequent divergence of CRISPR-Cas systems into several types, Cas1-Cas2 remains the workhorse of spacer acquisition, central to the success of CRISPR-Cas systems [4, 5]. As long as spacers can be acquired from MGEs, unique effector machineries capable of utilizing the information stored in CRISPRs will continue to evolve.

Mechanisms to generate Cas1-Cas2 compatible substrates, such as primed adaptation might have arisen because naïve acquisition is an inefficient and undirected process, potentially leading to high rates of lethal self-targeting spacers. However, despite the apparent advantages of primed adaptation, it was recently reported that promiscuous binding of crRNA-effector complexes to the host genome results in a basal level of self-priming, the extent of which is likely underrepresented due to the lethality of such events [51]. Host *cas* gene regulation mechanisms have arisen to balance the likelihood of self-acquisition events against the requirement to adapt to new threats, for example, when the risk of phage infection or HGT is high [109, 110]. Alternatively, it has been proposed that selective acquisition of self-targeting spacers could provide benefits such as invoking altruistic cell death [111], rapid genome evolution [33], regulation of host processes [112, 113], or even preventing the uptake of other CRISPR-Cas systems [114].

2.7 OUTLOOK

The past four years has seen rapid progress to understand the adaptation phase of CRISPR-Cas immunity. Despite this progress, many facets of CRISPR adaptation require further attention. Synergy between innate defense systems and adaptation is relatively unexplored,

but two roles can be envisioned; DNA breaks [57] stimulating generation of substrates for spacer acquisition (Fig. 4) or stalling of infection to 'buy time' for adaptation [31, 115, 116]. Analogously, it remains to be determined whether interference by CRISPR-Cas systems other than type I can also stimulate primed adaptation. If not, the benefits of priming might provide an explanation for why type I systems are more prevalent than other types.

It is also unclear why many CRISPR-Cas systems have multiple arrays used by a single set of Cas proteins, rather than a solo array. Given that Cas1-Cas2 is directed to leader-repeat junctions during integration, multiple arrays might provide additional integration sites, increasing adaptation efficiency. In addition, parallel CRISPR arrays should increase crRNA production from recently acquired spacers (i.e. due to polarization) [22]. Whereas some strains have multiple CRISPR arrays belonging to the same type, other hosts have several types of CRISPR-Cas systems simultaneously [117]. The benefits of harboring multiple CRISPR-Cas systems are not entirely clear, but can result in spacers used by different system to extend targeting to both RNA and DNA [118]. From an adaptation perspective, multiple systems might enable a wider PAM repertoire to be sampled during spacer selection. Additional systems in a single host could also be a response to defy phage- and MGE-encoded anti-CRISPR proteins, which can inhibit both interference and primed adaptation [119-121], or may allow some systems to function in defense, while others perform non-canonical roles in gene regulation [113].

While Cas effector nucleases (e.g. Cas9) have been harnessed for many biotechnological applications, the use of repurposed CRISPR-Cas adaptation machinery has yet to be widely exploited. The sequence-specific integrase activity holds promise in synthetic biology, such as for the insertion of specific sequences (or barcodes) to mark and track cells in a population. In *E. coli* the feasibility of such an approach is evident [49], but transition to eukaryotic systems will provide the greatest utility where lineage tracking and cell fate could be followed, as has been performed with Cas9 [122]. The elements required for leader-specific integration must be carefully considered for the introduction of CRISPR-Cas adaptation into eukaryotic cells, as unintended ectopic integrations could be problematic given the larger eukaryotic sequence space. Ultimately, our understanding of adaptation in prokaryotes may lead to applications where entire CRISPR sys-

tems are transplanted into eukaryotic cells to prevent viral invaders. As we begin to comprehend adaptation in more detail the opportunities to repurpose other parts of these remarkable prokaryotic immune systems is increasingly becoming reality.

References

1. R. L. Dy, C. Richter, G. P. Salmond, P. C. Fineran, Remarkable Mechanisms in Microbes to Resist Phage Infections. Annu Rev Virol 1, 307-331 (2014).
2. J. E. Samson, A. H. Magadán, M. Sabri, S. Moineau, Revenge of the phages: defeating bacterial defences. Nat Rev Microbiol 11, 675-687 (2013).
3. L. A. Marraffini, CRISPR-Cas immunity in prokaryotes. Nature 526, 55-61 (2015).
4. K. S. Makarova et al., An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol 13, 722-736 (2015).
5. P. Mohanraju et al., Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. Science 353, aad5147 (2016).
6. R. Barrangou et al., CRISPR provides acquired resistance against viruses in prokaryotes. Science 315, 1709-1712 (2007).
7. S. J. Brouns et al., Small CRISPR RNAs guide antiviral defense in prokaryotes. Science 321, 960-964 (2008).
8. A. V. Wright, J. K. Nuñez, J. A. Doudna, Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. Cell 164, 29-44 (2016).
9. G. Amitai, R. Sorek, CRISPR-Cas adaptation: insights into the mechanism of action. Nat Rev Microbiol 14, 67-76 (2016).
10. S. H. Sternberg, H. Richter, E. Charpentier, U. Qimron, Adaptation in CRISPR-Cas Systems. Mol Cell 61, 797-808 (2016).
11. M. J. Lopez-Sanchez et al., The highly dynamic CRISPR1 system of Streptococcus agalactiae controls the diversity of its mobilome. Mol Microbiol 85, 1057-1071 (2012).
12. G. W. Tyson, J. F. Banfield, Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. Environ Microbiol 10, 200-207 (2008).
13. A. F. Andersson, J. F. Banfield, Virus population dynamics and acquired virus resistance in natural microbial communities. Science 320, 1047-1050 (2008).
14. S. van Houte et al., The diversity-generating benefits of a prokaryotic adaptive immune system. Nature 532, 385-388 (2016).
15. C. Pourcel, G. Salvignol, G. Vergnaud, 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).
16. F. Liu et al., Novel virulence gene and clustered regularly interspaced short palindromic repeat (CRISPR) multilocus sequence typing scheme for subtyping of the major serovars of *Salmonella enterica* subsp. enterica. Appl Environ Microbiol 77, 1946-1956 (2011).
17. I. Yosef, M. G. Goren, U. Qimron, Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. Nucleic Acids Res 40, 5569-5576 (2012).
18. C. Díez-Villaseñor, N. M. Guzmán, C. Almendros, J. García-Martínez, F. J. Mojica, 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).
19. S. Erdmann, R. A. Garrett, Selective and hyperactive uptake of foreign DNA by adaptive immune systems of an archaeon via two distinct mechanisms. Mol Microbiol 85, 1044-1056 (2012).
20. C. L. Sun, B. C. Thomas, R. Barrangou, J. F. Banfield, Metagenomic reconstructions of bacterial CRISPR loci constrain population histories. ISME J 10, 858-870 (2016).
21. D. Paez-Espino et al., Uncovering Earth's virome. Nature 536, 425-430 (2016).

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

23. Y. Zhang et al., Processing-independent CRISPR RNAs limit natural transformation in Neisseria meningitidis. Mol Cell 50, 488-503 (2013).

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

25. P. Horvath et al., Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. J Bacteriol 190, 1401-1412 (2008).

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

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

28. K. S. Makarova, N. V. Grishin, S. A. Shabalina, Y. I. Wolf, E. V. Koonin, A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. Biol Direct 1, 7 (2006).

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

30. S. T. Abedon, Facilitation of CRISPR adaptation. Bacteriophage 1, 179-181 (2011).

31. A. P. Hynes, M. Villion, S. Moineau, Adaptation in bacterial CRISPR-Cas immunity can be driven by defective phages. Nat Commun 5, 4399 (2014).

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

33. R. B. Vercoe et al., Cytotoxic chromosomal targeting by CRISPR/Cas systems can reshape bacterial genomes and expel or remodel pathogenicity islands. PLoS Genet 9, e1003454 (2013).

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

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

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

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

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

39. I. Ivančić-Baće, S. D. Cass, S. J. Wearne, E. L. Bolt, Different genome stability proteins underpin primed and naive adaptation in *E. coli* CRISPR-Cas immunity. Nucleic Acids Res 43, 10821-10830 (2015).

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

41. J. Wang et al., Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems. Cell 163, 840-853 (2015).

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

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

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

45. K. N. Yoganand, R. Sivathanu, S. Nimkar, B. Anand, Asymmetric positioning of Cas1-2 complex and Integration Host Factor induced DNA bending guide the unidirectional homing of protospacer in CRISPR-Cas type I-E system. Nucleic Acids Res, (2016).

46. R. T. Leenay, C. L. Beisel, Deciphering, communicating, and engineering the CRISPR PAM. J Mol Biol, (2016).

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

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

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

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

51. R. H. Staals et al., Interference-driven spacer acquisition is dominant over naive and primed adaptation in a native CRISPR-Cas system. Nat Commun 7, 12853 (2016).

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

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

54. M. G. Goren et al., Repeat Size Determination by Two Molecular Rulers in the Type I-E CRISPR Array. Cell Rep 16, 2811-2818 (2016).

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

56. P. C. Fineran, E. Charpentier, Memory of viral infections by CRISPR-Cas adaptive immune systems: acquisition of new information. Virology 434, 202-209 (2012).

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

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

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

60. L. W. Enquist, A. Skalka, Replication of bacteriophage lambda DNA dependent on the function of host and viral genes. I. Interaction of red, gam and rec. J Mol Biol 75, 185-212 (1973).

61. J. Gowrishankar, J. K. Leela, K. Anupama, R-loops in bacterial transcription: their causes and consequences. Transcription 4, 153-157 (2013).

62. E. R. Westra et al., CRISPR-Cas systems preferentially target the leading regions of MOBF conjugative plasmids. RNA Biol 10, 749-761 (2013).

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

64. E. Semenova 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).

65. P. C. Fineran et al., Degenerate target sites mediate rapid primed CRISPR adaptation. Proc Natl Acad Sci U S A 111, E1629-1638 (2014).

66. D. Paez-Espino et al., Strong bias in the bacterial CRISPR elements that confer immunity to phage. Nat Commun 4, 1430 (2013).

67. C. Xue et al., CRISPR interference and priming varies with individual spacer sequences. Nucleic Acids Res 43, 10831-10847 (2015).

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

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

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

71. A. G. Patterson, J. T. Chang, C. Taylor, P. C. Fineran, Regulation of the Type I-F CRISPR-Cas system by CRP-cAMP and GalM controls spacer acquisition and interference. Nucleic Acids Res 43, 6038-6048 (2015).

72. K. Severinov, I. Ispolatov, E. Semenova, The Influence of Copy-Number of Targeted Extrachromosomal Genetic Elements on the Outcome of CRISPR-Cas Defense. Front Mol Biosci 3, 45 (2016).

73. T. Künne et al., Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation. Mol Cell 63, 852-864 (2016).

74. S. Redding et al., Surveillance and Processing of Foreign DNA by the *Escherichia coli* CRISPR-Cas System. Cell 163, 854-865 (2015).

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

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

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

78. R. P. Hayes et al., Structural basis for promiscuous PAM recognition in type I-E Cascade from *E. coli*. Nature 530, 499-503 (2016).

79. P. B. van Erp et al., Mechanism of CRISPR-RNA guided recognition of DNA targets in *Escherichia coli*. Nucleic Acids Res 43, 8381-8391 (2015).

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

81. C. Xue, N. R. Whitis, D. G. Sashital, Conformational Control of Cascade Interference and Priming Activities in CRISPR Immunity. Mol Cell 64, 826-834 (2016).

82. T. Sinkunas 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).

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

84. E. R. Westra 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).

85. Y. Huo et al., Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. Nat Struct Mol Biol 21, 771-777 (2014).

86. C. Richter 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).

87. E. Semenova et al., Highly efficient primed spacer acquisition from targets destroyed by the *Escherichia coli* type I-E CRISPR-Cas interfering complex. Proc Natl Acad Sci U S A 113, 7626-7631 (2016).

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

89. C. Rao et al., Active and adaptive Legionella CRISPR-Cas reveals a recurrent challenge to the pathogen. Cell Microbiol 18, 1319-1338 (2016).

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

91. T. R. Sampson, S. D. Saroj, A. C. Llewellyn, Y. L. Tzeng, D. S. Weiss, A CRISPR/Cas system mediates bacterial innate immune evasion and virulence. Nature 497, 254-257 (2013).

92. K. H. Nam, I. Kurinov, A. Ke, 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).

93. P. Ellinger et al., The crystal structure of the CRISPR-associated protein Csn2 from Streptococcus agalactiae. J Struct Biol 178, 350-362 (2012).

94. Z. Arslan et al., Double-strand DNA end-binding and sliding of the toroidal CRISPR-associated protein Csn2. Nucleic Acids Res 41, 6347-6359 (2013).

95. K. H. Lee et al., Identification, structural, and biochemical characterization of a group of large Csn2 proteins involved in CRISPR-mediated bacterial immunity. Proteins 80, 2573-2582 (2012).

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

97. J. Zhang, T. Kasciukovic, M. F. White, The CRISPR associated protein Cas4 Is a 5' to 3' DNA exonuclease with an iron-sulfur cluster. PLoS One 7, e47232 (2012).

98. S. Lemak 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).

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

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

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

102. G. E. Heussler, J. L. Miller, C. E. Price, A. J. Collins, G. A. O'Toole, Requirements for Pseudomonas aeruginosa Type I-F CRISPR-Cas Adaptation Determined Using a Biofilm Enrichment Assay. J Bacteriol 198, 3080-3090 (2016).

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

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

105. P. Beguin, N. Charpin, E. V. Koonin, P. Forterre, M. Krupovic, Casposon integration shows strong target site preference and recapitulates protospacer integration by CRISPR-Cas systems. Nucleic Acids Res, (2016).

106. E. V. Koonin, M. Krupovic, Evolution of adaptive immunity from transposable elements combined with innate immune systems. Nat Rev Genet 16, 184-192 (2015).

107. M. Krupovic, S. Shmakov, K. S. Makarova, P. Forterre, E. V. Koonin, Recent Mobility of Casposons, Self-Synthesizing Transposons at the Origin of the CRISPR-Cas Immunity. Genome Biol Evol 8, 375-386 (2016).

108. O. S. Alkhnbashi et al., Characterizing leader sequences of CRISPR loci. Bioinformatics 32, i576-i585 (2016).

109. A. G. Patterson et al., Quorum Sensing Controls Adaptive Immunity through the Regulation of Multiple CRISPR-Cas Systems. Mol Cell, (2016).

110. N. M. Høyland-Kroghsbo et al., Quorum sensing controls the Pseudomonas aeruginosa CRISPR-Cas adaptive immune system. Proc Natl Acad Sci U S A, (2016).

111. E. V. Koonin, F. Zhang, Coupling immunity and programmed cell suicide in prokaryotes: Life-or-death choices. Bioessays, (2016).

112. R. Li et al., Type I CRISPR-Cas targets endogenous genes and regulates virulence to evade mammalian host immunity. Cell Res 26, 1273-1287 (2016).

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

114. C. Almendros, N. M. Guzman, J. Garcia-Martinez, F. J. Mojica, Anti-*cas* spacers in orphan CRISPR4 arrays prevent uptake of active CRISPR-Cas I-F systems. Nat Microbiol 1, 16081 (2016).

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

116. M. E. Dupuis, M. Villion, A. H. Magadán, S. Moineau, CRISPR-Cas and restriction-modification systems are compatible and increase phage resistance. Nat Commun 4, 2087 (2013).

117. R. H. J. Staals, S. J. J. Brouns, in CRISPR-Cas Systems: RNA-mediated Adaptive Immunity in Bacteria and Archaea, R. Barrangou, J. van der Oost, Eds. (Springer Berlin Heidelberg, Berlin, Heidelberg, 2013), pp. 145-169.

118. J. Elmore, T. Deighan, J. Westpheling, R. M. Terns, M. P. Terns, DNA targeting by the type I-G and type I-A CRISPR-Cas systems of Pyrococcus furiosus. Nucleic Acids Res 43, 10353-10363 (2015).

119. J. Bondy-Denomy, A. Pawluk, K. L. Maxwell, A. R. Davidson, Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. Nature 493, 429-432 (2013).

120. A. Pawluk et al., Inactivation of CRISPR-Cas systems by anti-CRISPR proteins in diverse bacterial species. Nat Microbiol 1, 16085 (2016).

121. D. Vorontsova et al., Foreign DNA acquisition by the I-F CRISPR-Cas system requires all components of the interference machinery. Nucleic Acids Res 43, 10848-10860 (2015).

122. S. D. Perli, C. H. Cui, T. K. Lu, Continuous genetic recording with self-targeting CRISPR-Cas in human cells. Science 353, (2016).

3

Cas4 Facilitates PAM-Compatible Spacer Selection during CRISPR Adaptation

Cell Reports
2018 Mar 27; 22(13): 3377–3384.

Sebastian N. Kieper[1†], Cristóbal Almendros[1†], Juliane Behler[2], Rebecca E. McKenzie[1], Franklin L. Nobrega[1], Anna C. Haagsma[1], Jochem N.A. Vink[1], Wolfgang R. Hess[2,3] and Stan J.J. Brouns[1,4]

1. Kavli Institute of Nanoscience, Department of Bionanoscience, Delft University of Technology, Van der Maasweg 9, 2629 HZ Delft, the Netherlands.

2. Genetics and Experimental Bioinformatics, Faculty of Biology, University of Freiburg, Schänzlestraße 1, 79104 Freiburg, Germany.

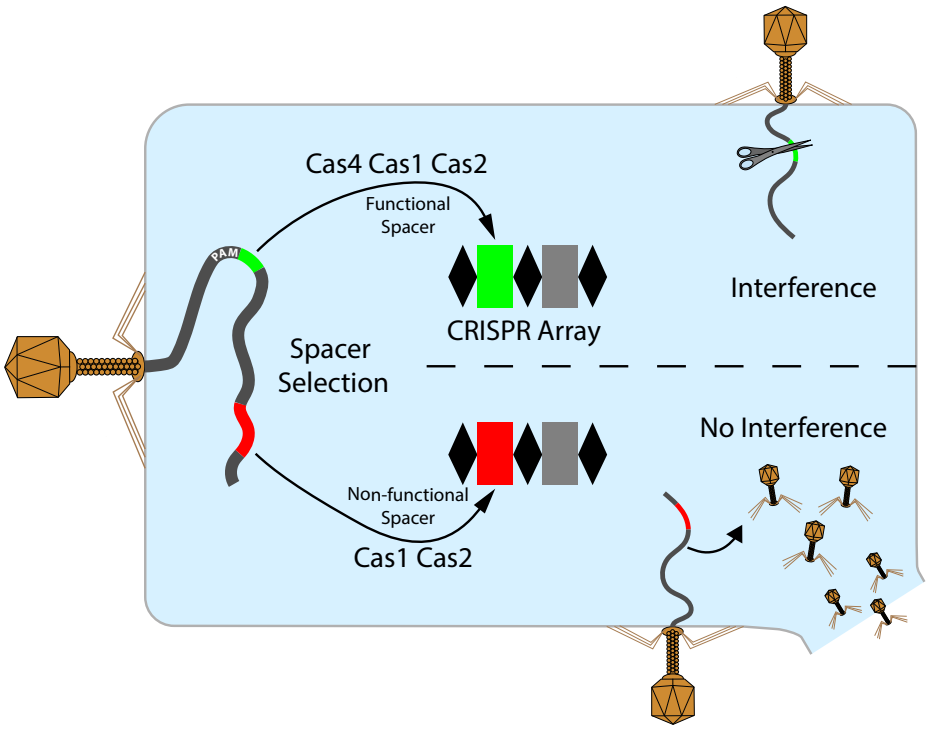
3. Genetics and Experimental Bioinformatics, Faculty of Biology, University of Freiburg, Schänzlestraße 1, 79104 Freiburg, Germany; Freiburg Institute for Advanced Studies, University of Freiburg, Albertstr. 19, 79104 Freiburg, Germany.

4. Laboratory of Microbiology, Wageningen University, Stippeneng 4, 6708 WE Wageningen, the Netherlands.

†These Authors contributed equally

3.1 ABSTRACT

CRISPR-Cas systems adapt their immunological memory against their invaders by integrating short DNA fragments into clustered regularly interspaced short palindromic repeat (CRISPR) loci. While Cas1 and Cas2 make up the core machinery of the CRISPR integration process, various class I and II CRISPR-Cas systems encode Cas4 proteins for which the role is unknown. Here, we introduced the CRISPR adaptation genes *cas1*, *cas2*, and *cas4* from the type I-D CRISPR-Cas system of *Synechocystis* sp. 6803 into *Escherichia coli* and observed that *cas4* is strictly required for the selection of targets with protospacer adjacent motifs (PAMs) conferring I-D CRISPR interference in the native host *Synechocystis*. We propose a model in which Cas4 assists the CRISPR adaptation complex Cas1-2 by providing DNA substrates tailored for the correct PAM. Introducing functional spacers that target DNA sequences with the correct PAM is key to successful CRISPR interference, providing a better chance of surviving infection by mobile genetic elements.



52

3.2 Introduction

Microbes require updating their adaptive immune repertoire to keep up with ever changing mobile genetic elements (MGEs) such as bacteriophages and conjugative plasmids. The prokaryotic CRISPR-Cas system is an adaptive immune system that uses clustered regularly interspaced short palindromic repeats (CRISPRs) and their associated proteins (Cas) [1-3]. In a process termed CRISPR adaptation microbes integrate short sequences from MGEs into their CRISPR array [4-6]. This array then becomes a source for small RNAs (i.e. crRNA) that guide Cas nuclease complexes to their target [7-9].

CRISPR-Cas systems are grouped into two major classes that each hold several types and multiple subtypes, and remarkably all encode *cas1* and *cas2* genes [10]. In the type I-E system of *E. coli cas1* and *cas2* are necessary and sufficient to mediate expansion of the CRISPR array [11]. However, next to *cas1* and *cas2* a number of other *cas* genes in Class I and II systems have been directly linked to the spacer integration process, suggesting that CRISPR adaptation across different systems has different requirements [10]. Type II-B (Cas9), type V (Cas12a), and most type I (I-A, I-B, I-C, I-D, I-U) CRISPR-Cas systems contain *cas4* genes in conserved gene clusters with *cas1* and *cas2* genes, while in some systems *cas4* is fused with *cas1* (I-B, I-U, V-B) [12]. Deletion of *cas4* from the I-A type abrogated CRISPR adaptation in a *Sulfolobus islandicus* strain overexpressing *csa3*, a regulator of *cas* gene expression [13], while deletion of *cas4* in type I-B revealed that *cas4* is essential for CRISPR adaptation against HHPV-2 in *Haloarcula hispanica* [14]. Additionally, interaction between the Cas1/2 fusion protein, Csa1 and Cas4 of the archaeal type I-A system was found *in vitro* [15]. These findings suggest a strong functional association of Cas4 and the Cas1 and Cas2 adaptation proteins. Despite the conservation of the *cas4* gene among these highly diverse CRISPR-Cas systems, a functional role for Cas4 has not been shown *in vivo*. Early biochemical studies have found different Cas4 proteins as monomers, dimers and decamers and to contain either [2Fe-2S] or [4Fe-4S] iron-sulfur clusters [16-18]. Furthermore, Cas4 proteins were shown to be active nucleases with catalytic domains belonging to the PD-DEXK phosphodiesterase superfamily [12]. It was suggested that the observed catalytic activities play a role in either the generation or the processing of spacer precursors, i.e DNA substrates that are

used by Cas1 and Cas2 to form spacers [16-18]. Recently, Rollie et al. showed *in vitro* that Cas4 cleaves 3' overhangs of prespacer substrates containing protospacer adjacent motifs (PAM) [19].

Obtaining new spacers that target an invading DNA sequence with a correct PAM is central to the success of CRISPR adaptation. The PAM is a short sequence motif [20-22] that is required for crRNA-effector complexes such as Cascade, Cas9 and Cas12a to find their target DNA and avoid targeting host CRISPR arrays [8]. Only when a new spacer has been selected from a target adjacent to a PAM, CRISPR interference can efficiently take place. In type II systems Cas9 assists Cas1-2 to select PAM-compliant spacers [23, 24], but it remains unknown what other factors also contribute to PAM selection.

Here we have determined the biological role of Cas4 by employing *in vivo* spacer acquisition assays in a heterologous *E. coli* host. We show that the type I-D adaptation proteins Cas1 and Cas2 from the cyanobacterium *Synechocystis* sp. 6803 are necessary and sufficient to integrate spacers into the CRISPR array. However, providing *cas4* results in a significant enrichment of new spacers with PAM motifs that support CRISPR interference in the type I-D CRISPR-Cas system of the native host *Synechocystis*. Altogether our results demonstrate that Cas4 enhances functional memory formation, which increases the chance of surviving infections by MGEs.

3.3 Experimental Procedures

3.3.1 Bacterial strains and growth conditions

E. coli strains DH5α, BW25113 (WT), JW2788 (BW25113 Δ*recB*), JW2790 (BW25113 Δ*recC*) and JW2787 (BW25113 Δ*recD*) were grown in Lysogeny Broth (LB) at 37˚C and continuous shaking at 180 rpm or grown on LB agar plates (LBA) containing 1.5% (wt/vol) agar. *Synechocystis* sp. 6803 was cultivated as described previously [44]. When required, the media were supplemented with 100 μg ml^{-1} ampicillin, 50 μg ml^{-1} spectinomycin, 25 μg ml^{-1} chloramphenicol, 7.5 μg ml^{-1} gentamicin (see Table S1 for plasmids and corresponding selection markers).

3.3.2 Plasmid construction and transformation

Plasmids used in this study are listed in Table S1. All cloning steps were performed in *E. coli* DH5α. Primers described in Table S2 were used for PCR amplification of the type I-D CRISPR-Cas locus (*cas4*, *cas1*, *cas2* and leader-repeat-spacer1) from *Synechocystis* cell material using the Q5 high-fidelity Polymerase (New England Biolabs). PCR amplicons were subsequently cloned into Berkeley MacroLab LIC vectors (http://macrolab.berkeley.edu/) using either ligation-independent cloning (LIC), or into the pACYCDuet-1 vector system (Novagen (EMD Millipore) using conventional restriction-ligation cloning. The $cas4^{D76A}$ mutant [18] was obtained using a PCR-based mutagenesis using primers listed in Table S2. The conjugative plasmid pVZ322 used in the interference study was obtained by fusing the 5' PAM (GTA, GTT, GTG, GTC, and AGC)-protospacer1 3' sequence in-frame with a gentamicin resistance cassette upstream of its stop codon using inverse PCR using primers listed in Table S2. The gentamicin resistance cassette with and without the PAM-protospacer sequence (pT and pNT respectively) was then assembled with the linearized pVZ322 backbone. All plasmids were verified by Sanger-sequencing (Macrogen Europe, Amsterdam, The Netherlands and GATC Biotech, Konstanz, Germany). Bacterial transformations were either carried out by electroporation (2.5 kV, 25 mF, 200 V) using a ECM 630 electroporator (BTX Harvard Apparatus) or using chemically competent cells prepared according to manufacturer's manual (Mix&Go, Zymo research). Electrocompetent cells were prepared following a protocol adapted

from [45]. Transformants were selected on LBA supplemented with appropriate antibiotics.

3.3.3 *IN VIVO* SPACER ACQUISITION ASSAY

E. coli BW25113 and *E. coli* mutant strains JW2788 (BW25113 Δ*recB*), JW2790 (BW25113 Δ*recC*) and JW2787 (BW25113 Δ*recD*) were transformed with pCas1-2, pCRISPR and either pCas4, pCas4^{D76A} or the pEmp control plasmid (Table S1). Cultures were inoculated from single colonies and passaged once after 24 hours of growth at 37˚C and continuous shaking at 180 rpm. 200 µL of cells were harvested by centrifugation and resuspended in 50 µL of MilliQ water. Subsequently, 2 µL of cell suspension was subjected to spacer detection PCR using a forward primer annealing in the 3' end of the CRISPR repeat of pCRISPR but mismatching the first nucleotide of spacer 1 (degenerated primer mix [23]) and a reverse primer annealing in the vector backbone (Table S2). When higher sensitivity was required, amplicons of expanded pCRISPR arrays were separated from parental pCRISPR array amplicons using the BluePippin automated agarose-electrophoresis system (3% agarose gel cassette, SageScience). The extracted expanded CRISPR array amplicons were then subjected to an additional PCR reaction using the same degenerated primer mix but a different reverse primer matching spacer 1.

3.3.4 NEXT GENERATION SEQUENCING AND STATISTICAL ANALYSIS

After validation of PCR amplicons by gel electrophoresis and clean up with the GeneJET PCR Purification kit (Thermo Fisher Scientific) the samples were analyzed using Invitrogen Qubit fluorometric quantification. Samples were prepared for sequencing with the Nextera XT DNA Library Preparation Kit (Illumina) and each library individually barcoded with the Nextera XT Index Kit v2 SetA (Illumina). Libraries were pooled equally and spiked with ~5% of the PhiX control library (Illumina) to artificially increase the genetic diversity before sequencing on a Nano flowcell (2 x 250 base paired-end) with an Illumina MiSeq. Image analysis, base calling, de-multiplexing and data quality assessments were performed on the MiSeq instrument. FASTAQ files generated by the MiSeq were analyzed by pairing and merging the reads using Geneious 9.0.5 and subsequently extracting newly

acquired spacers by identifying the 3' end of the degenerate primer and the 5' end of the single repeat present in the parental pCRISPR. Unique spacer sequences were mapped to the chromosome and the replicons carried by the corresponding strains with the BLAST-function of Geneious 9.0.5.

3.3.5 Statistical tests

To infer the likelihood of finding a certain distribution of PAMs we used a binomial test, where we estimated the likelihood of the observed frequency of each PAM in the case of a randomly distributed PAM-pool (likelihood per PAM: 1/16). As we performed the test multiple times on the same data set, we used a Bonferroni correction to decrease the probability of a type I error. Spacer size preference was tested by using a bootstrapping resampling method with replacement. 10.000 bootstrap resamples were generated from each observed dataset (each of similar size to the observed dataset). The statistical mode of a certain spacer size within these resamples represented the likelihood of observing this mode in the observed dataset.

3.3.6 *Synechocystis* interference assay

Synechocystis sp. 6803 contains on its megaplasmid pSYSA a type I-D and two type III CRISPR-Cas systems (III-D and III-B) [44]. *Synechocystis* I-D interference assays were performed as described previously [44] with a *Synechocystis* sp. 6803 derivative strain with 16 instead of 49 spacers (spacer 1-14 and 48-49 retained) in its I-D CRISPR array. Conjugation assays were performed using the self-replicating conjugative vector pVZ322 and the gentamicin resistance cassette for selection. Target plasmids with a number of different PAMs were constructed containing the target of spacer 1 of the I-D CRISPR array. Plasmids were conjugated into *Synechocystis* by triparental mating as described [44]. Briefly, overnight cultures of the helper strain *E. coli* J53/RP4 and the donor strain *E. coli* DH5α with the plasmid of interest were diluted and incubated for 2.5 h at 37°C with shaking at 180 rpm. For each conjugation, an OD_{600} of 7.0 of the plasmid-bearing and helper cultures were harvested, resuspended in LB and combined. The mixed culture was incubated for 1 h at 30°C without shaking. In parallel, a *Synechocystis* culture with an OD_{750} of 1.0 was harvested and combined with the mixed culture of the plasmid-bearing and helper

culture. The pellet was resuspended and placed on a sterile filter. After overnight incubation at 30°C, the filter was rinsed and 30 µL of the resulting cell suspension were plated on BG11 agar plates containing 7.5 µg mL^{-1} gentamicin. Transconjugants were counted after further incubation at 30 °C for 2 weeks. Mean values of conjugation efficiency and corresponding standard errors were calculated by dividing the number of transconjugants obtained with the target plasmids (pT) by the number of transconjugants obtained with the non-target control plasmid (pNT). Experiments were performed in biological triplicates and in parallel with the control plasmid.

3.4 Results

3.4.1 The Cas1-Cas2 complex integrates spacers independently of Cas4

To determine the minimal requirements for spacer acquisition in the type I-D system, we cloned *cas4*, *cas1* and *cas2* genes from *Synechocystis* (Fig. 3.1A) into T7-based expression vectors. A minimal CRISPR array with one repeat was obtained by cloning the full-leader sequence followed by the first repeat and the leader proximal spacer (Sp1) of the type I-D system into the pACYC-Duet1 vector system. The resulting plasmids were co-transformed into a WT *E. coli* K12 strain (BW25113) devoid of T7 RNA polymerase. This setup ensured constitutive and low expression levels of the adaptation genes from *Synechocystis*. We first tested the ability of the cells to integrate new spacers into the minimalized CRISPR array either in the presence or absence of the *cas4-1-2* genes. With a sensitive PCR approach (Fig. 3.1B), spacer acquisition was readily detectable in the presence of *cas1*-2 regardless of the presence of the *cas4* gene (Fig. 3.1C). Further, deletion of either *cas1* or *cas2* abolished spacer integration, indicating that the combination of *cas1* and *cas2* is necessary and sufficient to mediate the integration of new spacers (Fig. 3.1C). The detection of expanded CRISPR arrays in *E. coli* K12 demonstrates that spacers were acquired even though *E. coli* is not the natural host of the type I-D system. Consequently, the type I-D adaptation module does not rely on any specific host factors only present in *Synechocystis*.

3.4.2 Cas4 enhances spacer acquisition in the absence of the RecBCD complex

Next, we were interested in knowing whether the observed integration was dependent on the presence of host factors in our heterologous expression system. Since the *E. coli* strain used is not the natural host of the type I-D system, CRISPR adaptation by I-D Cas1-2 does not rely on cyanobacterial factors that are only present in *Synechocystis*. For the type I-E system of *E. coli* a role for the RecBCD complex has been proposed in generating spacer precursors during double stranded DNA break repair at stalled replication forks [25, 26]. The *Synechocystis* genome contains cyanobacterial orthologues of *E. coli* RecB and RecD, but RecC appears to be absent [27]. Hence we sought

to assess spacer integration in *E. coli* Δ*recB*, Δ*recC* and Δ*recD* mutant backgrounds from the KEIO collection [28]. While we observed no difference in spacer acquisition frequencies for pCas1-2 in *recB* and *recC* deletion mutants, integration of spacers in the *recD* mutant was greatly reduced (Fig. 3.1D), but could still be detected with the sensitive spacer detection approach (Fig. 3.S1B). Interestingly, when we supplied *cas4* in the *recB* and *recC* deletion backgrounds we observed a relative increase of array expansion (Fig. 3.1D). The results demonstrate that Cas1-2 is the core requirement for type I-D adaptation as has been found for type I-E [11, 29]. The presence of Cas4 seems to facilitate uptake of spacers in the absence of RecB or RecC, which is consistent with competing pathways for the generation of spacer precursors.

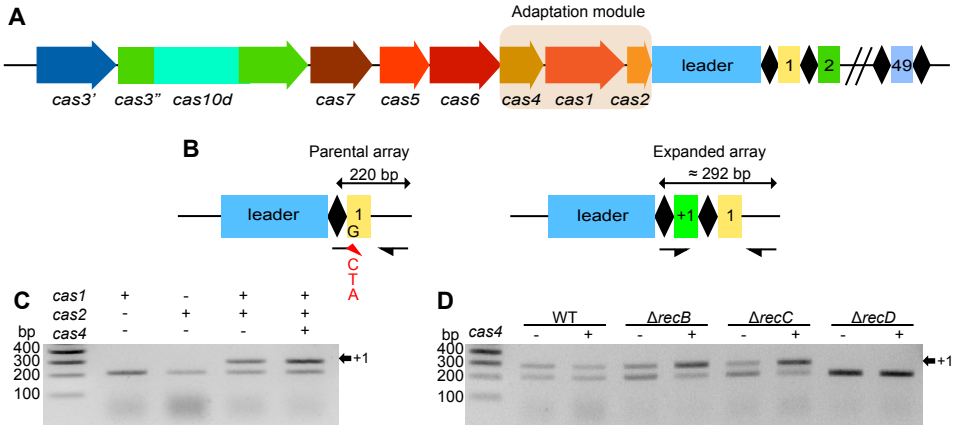


Figure 3.1: A Overview of the type I-D CRISPR-Cas locus as found on the *Synechocystis* sp. 6803 pSYSA megaplasmid. The putative adaptation module consisting of *cas4-cas1-cas2* is highlighted in light brown. Downstream of *cas2* is the leader sequence in blue, followed by the I-D array consisting of 37 bp repeats interspaced by 49 spacers with a statistical mode length of 35 bp (Fig. S2C). **B** Degenerate primer PCR used for the detection of spacer acquisition [23]. The 3' end of the forward primer mix mismatches the first 5' nucleotide of spacer 1 (indicated in red). Spacer integration restores complementarity allowing for efficient amplification. For more sensitive detection the amplicon of expanded arrays was extracted and subjected to a second round of PCR (see Fig. S1A). **C** Co-expression of *cas1* and *cas2* is necessary and sufficient for the integration of new spacers. **D** Assessing spacer integration in WT *E. coli* K12 and different *recBCD* mutant backgrounds in the presence or absence of *cas4*. The presence of *cas4* enhances spacer integration in the Δ*recB* and Δ*recC* genotypes, while spacer integration is below the detection limit of this PCR (described in B) in the Δ*recD* mutant regardless of the presence of *cas4*.

3.4.3 Cas4 influences spacer length

To understand the nature and origin of newly acquired spacers in the presence or absence of *cas4*, we subjected amplicons of expanded I-D arrays to next generation sequencing. Analysis of novel spacers in the absence of *cas4* revealed that spacers of 36 bp length were incorporated most frequently (Fig. 3.2A). This length deviates by one nucleotide from the spacer length found in the native CRISPR array of *Synechocystis* in which the statistical mode of spacer length is 35 bp (Fig. 3.S2C). Interestingly, when *cas4* was supplied to the system, the mode of spacer length was restored to 35 bp. To assess if Cas4 activity was responsible for the change in spacer length, we created an active site mutant in the RecB-domain by substituting a divalent metal-ion binding aspartic acid for alanine (i.e. D76 corresponding to D99 in Sso0001) [18]. When this mutant was introduced in strains containing pCas1-2, the same spacer length mode was observed as when *cas4* was absent, showing that the catalytic activity of Cas4 influences spacer length. Furthermore, it suggests that Cas4 is involved in processing spacer precursors (i.e. prespacers) before they are integrated into the CRISPR array.

3.4.4 New spacers are mostly genome-derived

Next, we mapped the unique spacer sequences to the *E. coli* BW25113 genome as well as to the plasmids harbored by the cells. Approximately 60% of the spacers that were acquired in the absence of the *cas4* mapped to the genome (Fig. 3.2B). We observed increased numbers of spacers targeting the *lacI* gene, which is present both on the plasmid and the genome. Spacers were also preferentially acquired from the chromosomal replication terminus terC (Fig. 3.S3). The enrichment of spacers at the replication terminus is similar to what has been observed previously for type I-E [26], and suggests that the I-D Cas1-2 adaptation complex can use DNA degradation products from RecBCD as substrates for new spacers. When we supplied wild type or mutant *cas4*, spacer acquisition from the genome further increased to 85% and 90%, respectively. However, the preferential uptake of spacers from *terC* was lost (Fig. 3.S3). We observed no orientation bias of the newly integrated spacers for either strand of the genome (Tab. S3). Although *E. coli* is not the native host of the I-D CRISPR system, the results are consistent with the notion that the adaptation proteins of I-D use prespacer substrates from abundant DNA sources in the cell,

in this case the genome.

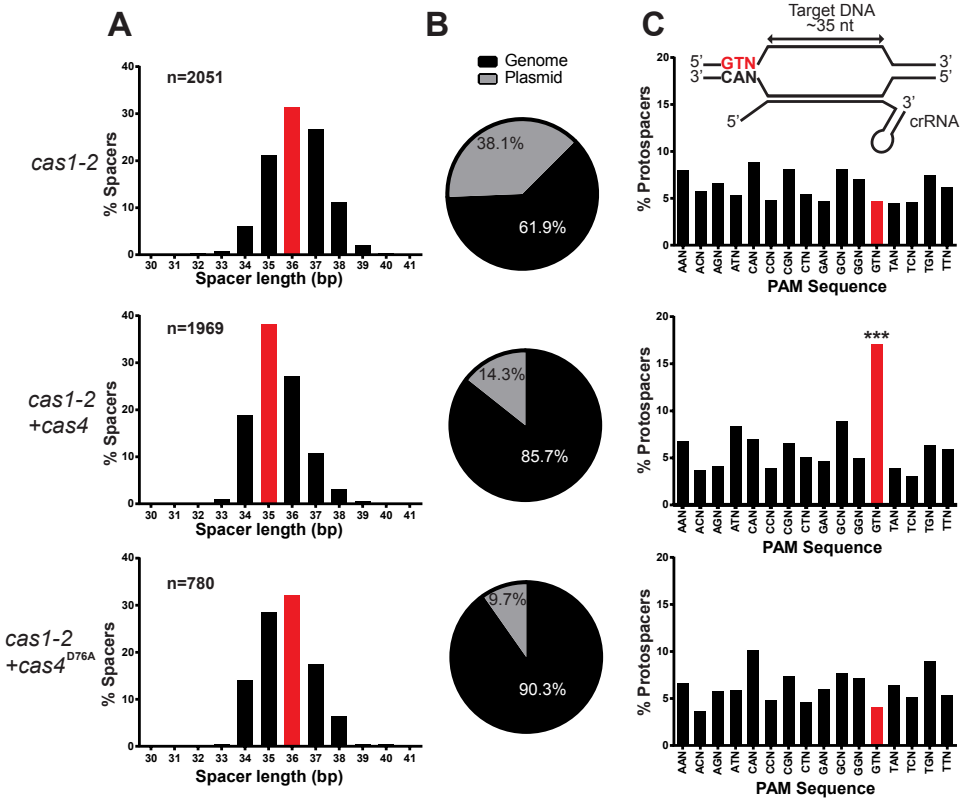


Figure 3.2 Analysis of spacers acquired by type I-D CRISPR adaptation
A Spacer length distribution in cells harbouring different combinations of *cas* genes. The variation in spacer size does not depend on the presence of *cas4*, but is solely dependent on the Cas1-2 adaptation complex. The presence of *cas4* restores the statistical mode of spacer length as found in the native I-D array (35 bp; Fig. S2C). Statistical mode of spacer length is indicated with red bars. **B** Origin of newly acquired spacers. The type I-D adaptation proteins acquire mostly from genomic DNA. **C** Percentage of protospacers with different PAMs. The presence of *cas4* significantly increases the incorporation of spacers that match protospacers with the consensus GTN PAM, while no significant enrichment of PAMs is observed for spacers acquired by Cas1-2 alone or in conjunction with the Cas4^{D76A} mutant. n = number of analyzed spacer sequences, significance level $\alpha = 0.001$.

3.4.5 Cas4 facilitates selection of spacers with a specific PAM

In order to determine which PAMs had been selected during spacer acquisition we mapped the unique spacers to their targets and retrieved their flanking sequences. This revealed that in the absence of *cas4* no particular sequence motifs were enriched in the flanking regions of the target. Interestingly, when we analyzed upstream flanking sequences of targets from spacers acquired in the presence of *cas4*, we observed that spacers with GTN PAMs were significantly enriched (Fig. 3.2C). This GTN PAM matched the previously predicted PAM for I-D systems [22]. When we introduced *cas4*^{D76A} the enrichment of GTN PAMs was no longer observed, indicating that the metal-ion coordinating residue, which is likely important for catalytic activity of Cas4, is also essential for PAM selection. Cas1-2 alone displays no inherent PAM selection preference. In order to assess whether inactivation of the *recB* gene would reduce background levels of spacers derived from RecBCD products, we subjected the expanded arrays from the *recB* mutant to high-throughput sequencing. Although this genetic background did not abolish background spacer integration, the presence of *cas4* further increased GTN PAM-compliant spacers (Fig. 3.S2A-B and Tab. S3).

3.4.6 GTN is a functional PAM in the native type I-D host *Synechocystis*

To test whether the GTN PAM enriched in the presence of *cas4* licenses CRISPR interference in *Synechocystis*, we performed interference assays using a conjugative plasmid containing a protospacer matching spacer 1 of the type I-D array. The protospacer was flanked by one of the four GTN PAMs (GTA, GTC, GTG or GTT) and carried gentamicin resistance for selection. Compared to a non-target control plasmid, we observed a dramatic reduction in the numbers of transconjugants with each of the four possible GTN PAMs (i.e. no transconjugants for GTC, GTG and GTT, and 1 for the GTA PAM) (Fig. 3.3). In contrast, plasmids containing protospacers flanked by AGC PAMs resulted in the same conjugation efficiency as found for the non-target control. We conclude that the type I-D system in *Synechocystis* is active and provides efficient CRISPR interference with GTN PAMs.

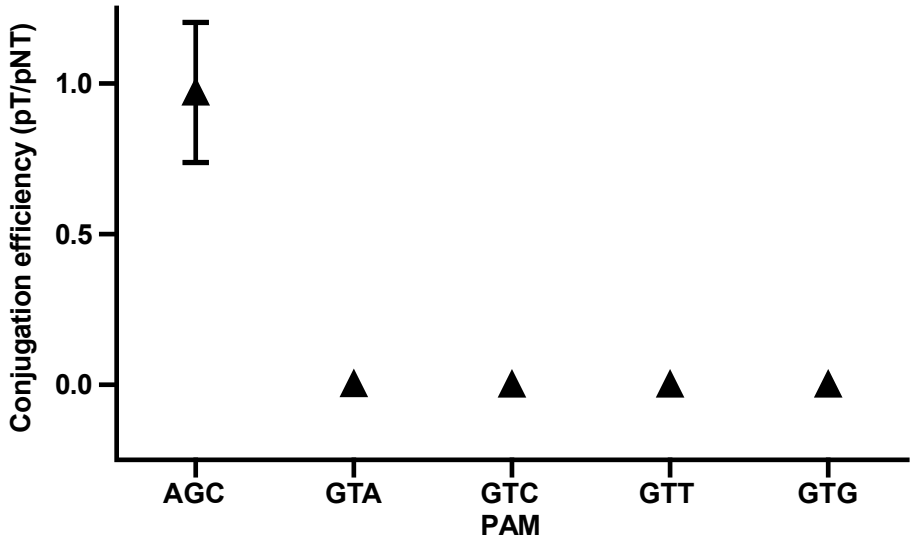


Figure 3.3: Conjugation efficiency of target plasmids (pT) carrying either AGC-protospacer1 or GTN-protospacer1 in *Synechocystis* sp. 6803. The AGC-PAM does not license interference while plasmids containing the observed GTN PAM are efficiently cleared. Data points represent the mean ± SE (n=3).

3.5 Discussion

Microbes face a number of challenges when they update their CRISPR memory. Firstly, how can they select new spacers from invading elements, while preventing to sample their own genome? Secondly, how can they maintain a balance between spacer uptake and turnover? Thirdly, how can they select spacers that give functional CRISPR interference? Here we have addressed the last question, and show that the highly ubiquitous Cas4 protein present in type I, II and V systems helps to integrate spacers targeting DNA sequences with PAMs that support CRISPR interference. Because crRNA-effector complexes such as Cascade, Cas9 and Cas12a are critically reliant on PAMs to find their target DNA and to avoid host CRISPR arrays, the selection of PAM-compliant spacers enhances the success rate of CRISPR interference, and promotes clearance of invader DNA from cells [30]. Apart from influencing the PAM, we found that the statistical mode of spacer length was shifted by 1 nucleotide to shorter spacers, suggesting a role for Cas4 in processing spacer substrates before or during integration. The variable spacer size itself is dictated only by Cas1-2 from type I-D. While structural constraints of the Cas1-2 complex from type I-E, and presumably also of the Cas1-2/3 complex from type I-F act like a molecular ruler that predetermines a fixed spacer length of predominantly 32 nt [31-34], the integration complex of the type I-D system likely displays plasticity and enables incorporation of spacers that vary in size by 5 or 6 nucleotides. This spacer size variation is not only observed in type I-D but also in type I-B [35] and many CRISPR systems containing *cas4* genes. Our data is consistent with a model in which the nuclease activities of Cas4 tailor prespacer substrates for the Cas1-2 adaptation machinery during the integration of new spacers.

The *cas4* gene has long been implicated in CRISPR adaptation. Many *cas4* genes have been found adjacent to *cas1* and *cas2* and in some *cas*es fusions between *cas4* and *cas1* have been observed [10, 12]. The *cas4* gene was shown to be essential for CRISPR adaptation in the type I-B system of *Haloarcula* [14]. Interestingly, a Campylobacter bacteriophage containing a *cas4* gene promoted acquisition of self-targeting spacers in the Campylobacter type II-C system [36], which is in line with our finding that Cas4 promotes the integration of spacer from abundant DNA populations in the cell. The Cas4 protein has been observed in a complex with a Cas1-Cas2 fusion protein and Csa1 in the

Sulfolobus type I-A system and this complex was coined *Cascis* after CRISPR-associated complex for integration of spacers [15]. Although different catalytic activities have been assigned [12, 16-18], the biological role of Cas4 has remained elusive. Only recently it was shown that Cas4 nuclease activity participates in PAM-dependent cleavage of 3' overhangs of prespacers [19]. This sequence specific cleavage is in line with the findings presented in this study in which Cas4-derived spacers are shorter and enriched in functional GTN PAMs.

While Cas4 may aid the generation of spacers with the correct PAM in a number of CRISPR-Cas systems, some other Cas proteins have been found to influence PAM selection as well. The crRNA-guided effector complex Cas9 present in type II-A systems is required for spacer acquisition (Wei et al., 2015) and helps to select new spacers with a correct PAM [23, 24]. Next to Cas1-2, the integration of new spacers in type II-A requires the toroidal DNA binding protein Csn2, a protein known to interact with Cas1 [23, 37].

Other ways to improve taking up spacers with the correct PAM include primed CRISPR adaptation [38, 39], which appears to be a general feature of type I systems only [14, 40-42]. In contrast to naïve spacer acquisition, primed CRISPR adaptation uses pre-existing spacer matches to trigger updates of the CRISPR memory against that target. Apart from Cas1-2, the priming process requires the presence of a crRNA-effector complex (e.g. Cascade) and the DNA nuclease Cas3. It seems that the frequency of acquiring functional spacers during priming is much higher than during naïve spacer acquisition (Jackson et al., 2017). This can be partly explained by considering that functional spacers confer a selective advantage to the host when the interference machinery is present. On the molecular level the increased frequency of functional spacers during priming may be explained by the observation that the Cas3 nuclease cleaves target DNA in a PAM-compatible manner to fuel the Cas1-2 adaptation machinery with suitable DNA substrates for integration [43].

Taken together, a picture has emerged that it is important for microbes to acquire functional instead of randomly selected spacers in their CRISPR arrays, and that there are a variety of ways in which CRISPR systems can accomplish this. The conserved component Cas4, which is present in about half of all CRISPR subtypes, appears to be a Cas protein dedicated to the task of facilitating the integration of func-

tional spacers during CRISPR adaptation.

REFERENCES

1. Jansen, R., et al., Identification of genes that are associated with DNA repeats in prokaryotes. Molecular microbiology, 2002. 43(6): p. 1565-75.
2. Mojica, F.J., et al., Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J Mol Evol, 2005. 60(2): p. 174-82.
3. Barrangou, R., et al., CRISPR provides acquired resistance against viruses in prokaryotes. Science, 2007. 315(5819): p. 1709-12.
4. Sternberg, S.H., et al., Adaptation in CRISPR-Cas Systems. Mol Cell, 2016. 61(6): p. 797-808.
5. Jackson, S.A., et al., CRISPR-Cas: Adapting to change. Science, 2017. 356(6333).
6. Amitai, G. and R. Sorek, CRISPR-Cas adaptation: insights into the mechanism of action. Nat Rev Microbiol, 2016. 14(2): p. 67-76.
7. Marraffini, L.A., CRISPR-Cas immunity in prokaryotes. Nature, 2015. 526: p. 55.
8. Mohanraju, P., et al., Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. Science, 2016. 353(6299): p. aad5147.
9. van der Oost, J., et al., Unravelling the structural and mechanistic basis of CRISPR-Cas systems. Nat Rev Microbiol, 2014. 12(7): p. 479-92.
10. Koonin, E.V., K.S. Makarova, and F. Zhang, Diversity, classification and evolution of CRISPR-Cas systems. Curr Opin Microbiol, 2017. 37: p. 67-78.
11. Yosef, I., M.G. Goren, and U. Qimron, Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. Nucleic Acids Res, 2012. 40(12): p. 5569-76.
12. Hudaiberdiev, S., et al., Phylogenomics of Cas4 family nucleases. BMC Evolutionary Biology, 2017. 17: p. 232.
13. Liu, T., et al., Coupling transcriptional activation of CRISPR-Cas system and DNA repair genes by Csa3a in Sulfolobus islandicus. Nucleic Acids Res, 2017. 45(15): p. 8978-8992.
14. Li, M., et al., Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process. Nucleic Acids Res, 2014. 42(4): p. 2483-92.
15. Plagens, A., et al., Characterization of the CRISPR/Cas subtype I-A system of the hyperthermophilic crenarchaeon Thermoproteus tenax. J Bacteriol, 2012. 194(10): p. 2491-500.
16. 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, 2013. 135(46): p. 17476-87.
17. Lemak, S., et al., The CRISPR-associated Cas4 protein Pcal_0546 from Pyrobaculum calidifontis contains a [2Fe-2S] cluster: crystal structure and nuclease activity. Nucleic Acids Res, 2014. 42(17): p. 11144-55.
18. Zhang, J., T. Kasciukovic, and M.F. White, The CRISPR associated protein Cas4 Is a 5' to 3' DNA exonuclease with an iron-sulfur cluster. PLoS One, 2012. 7(10): p. e47232.
19. Rollie, C., et al., Prespacer processing and specific integration in a Type I-A CRISPR system. Nucleic Acids Research, 2017: p. gkx1232-gkx1232.
20. Deveau, H., et al., Phage Response to CRISPR-Encoded Resistance in *Streptococcus thermophilus*. Journal of Bacteriology, 2008. 190(4): p. 1390-1400.
21. Mojica, F.J.M., et al., Short motif sequences determine the targets of the prokaryotic

CRISPR defence system. Microbiology, 2009. 155(3): p. 733-740.

22. Shah, S.A., et al., Protospacer recognition motifs: mixed identities and functional diversity. RNA Biol, 2013. 10(5): p. 891-9.

23. Heler, R., et al., Cas9 specifies functional viral targets during CRISPR-Cas adaptation. Nature, 2015. 519(7542): p. 199-202.

24. Wei, Y., R.M. Terns, and M.P. Terns, Cas9 function and host genome sampling in Type II-A CRISPR–Cas adaptation. Genes & Development, 2015. 29(4): p. 356-361.

25. Ivancic-Bace, I., et al., Different genome stability proteins underpin primed and naive adaptation in *E. coli* CRISPR-Cas immunity. Nucleic Acids Res, 2015. 43(22): p. 10821-30.

26. Levy, A., et al., CRISPR adaptation biases explain preference for acquisition of foreign DNA. Nature, 2015. 520(7548): p. 505-510.

27. Cassier-Chauvat, C., T. Veaudor, and F. Chauvat, Comparative Genomics of DNA Recombination and Repair in Cyanobacteria: Biotechnological Implications. Frontiers in Microbiology, 2016. 7: p. 1809.

28. Baba, T., et al., Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. Molecular Systems Biology, 2006. 2: p. 2006.0008-2006.0008.

29. Nunez, J.K., et al., Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. Nat Struct Mol Biol, 2014. 21(6): p. 528-34.

30. Mohanraju, P., et al., Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. Science, 2016. 353(6299).

31. Nunez, J.K., et al., Foreign DNA capture during CRISPR-Cas adaptive immunity. Nature, 2015. 527(7579): p. 535-8.

32. Wang, J., et al., Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems. Cell, 2015. 163(4): p. 840-53.

33. Fagerlund, R.D., et al., Spacer capture and integration by a type I-F Cas1-Cas2-3 CRISPR adaptation complex. Proc Natl Acad Sci U S A, 2017. 114(26): p. E5122-E5128.

34. Rollins, M.F., et al., Cas1 and the Csy complex are opposing regulators of Cas2/3 nuclease activity. Proc Natl Acad Sci U S A, 2017. 114(26): p. E5113-E5121.

35. Li, M., et al., The spacer size of I-B CRISPR is modulated by the terminal sequence of the protospacer. Nucleic Acids Research, 2017. 45(8): p. 4642-4654.

36. Hooton, S.P.T. and I.F. Connerton, *Campylobacter jejuni* acquire new host-derived CRISPR spacers when in association with bacteriophages harboring a CRISPR-like Cas4 protein. Frontiers in Microbiology, 2014. 5: p. 744.

37. Ka, D., et al., Crystal Structure of *Streptococcus pyogenes* Cas1 and Its Interaction with Csn2 in the Type II CRISPR-Cas System. Structure, 2016. 24(1): p. 70-79.

38. Datsenko, K.A., et al., Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. Nat Commun, 2012. 3: p. 945.

39. Swarts, D.C., et al., CRISPR Interference Directs Strand Specific Spacer Acquisition. PLoS ONE, 2012. 7(4): p. e35888.

40. Staals, R.H.J., et al., Interference-driven spacer acquisition is dominant over naive and primed adaptation in a native CRISPR–Cas system. Nature Communications, 2016. 7: p. 12853.

41. Rao, C., et al., Active and adaptive Legionella CRISPR-Cas reveals a recurrent challenge to the pathogen. Cellular Microbiology, 2016. 18(10): p. 1319-1338.

42. van Houte, S., et al., The diversity-generating benefits of a prokaryotic adaptive immune system. Nature, 2016. 532(7599): p. 385-8.

43. Künne, T., et al., Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation. Molecular Cell, 2016. 63(5): p. 852-864.

44. Scholz, I., et al., CRISPR-Cas systems in the cyanobacterium *Synechocystis* sp. sp. 6803 exhibit distinct processing pathways involving at least two Cas6 and a Cmr2 protein. PLoS One, 2013. 8(2): p. e56470.

45. Gonzales, M.F., et al., Rapid Protocol for Preparation of Electrocompetent *Escherichia coli* and Vibrio cholerae. Journal of Visualized Experiments : JoVE, 2013(80): p. 50684.

SUPPLEMENTARY
SUPPLEMENTARY FIGURES

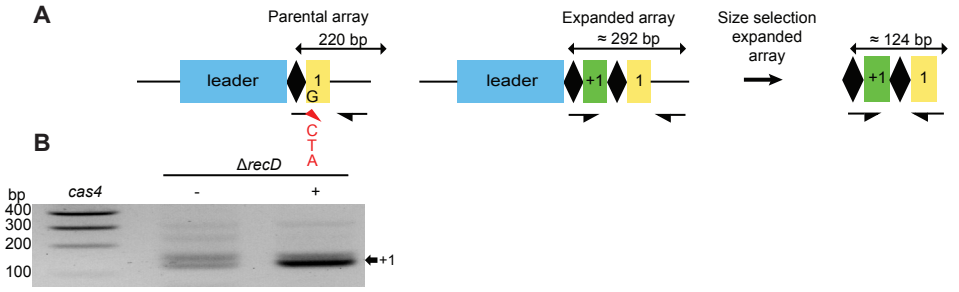


Figure 3.S1 A, related to Fig. 1 D . High sensitivity spacer detection PCR. Following the first round of amplification, the hypothetical amplicon of expanded arrays of ~292 bp (not visible on gel in Fig. 3.1D) is size selected using the BluePippin system (3% agarose cassette, SageScience). The extracted band is then subjected to a second PCR using the same forward degenerate primer mix [23] with a reverse primer annealing in spacer1 of the parental array. B By applying the PCR approach described in **A** spacer integration is observed in the *E. coli* K12 Δ*recD* background in the presence or absence of *cas4*. Band intensity is not a quantitative measure for integration efficiency, but rather a binary result (yes/no), because PCR product input resulting from automatized size selection cannot be normalized. The amplicon corresponding to expanded arrays is indicated with a black arrow.

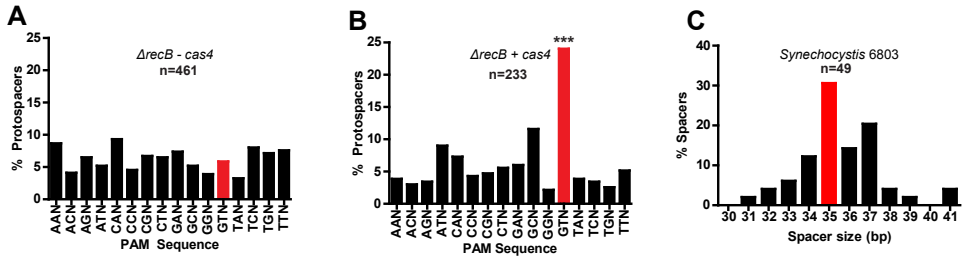


Figure 3.S2 related to Fig. 2. A Percentage of protospacers adjacent to indicated PAM matched by spacers acquired in **A** Δ*recB* without *cas4* and **B** with *cas4*. **C** Spacer size distribution in the native type I-D host *Synechocystis*. n= number of analyzed spacers, significance level α = 0.001.

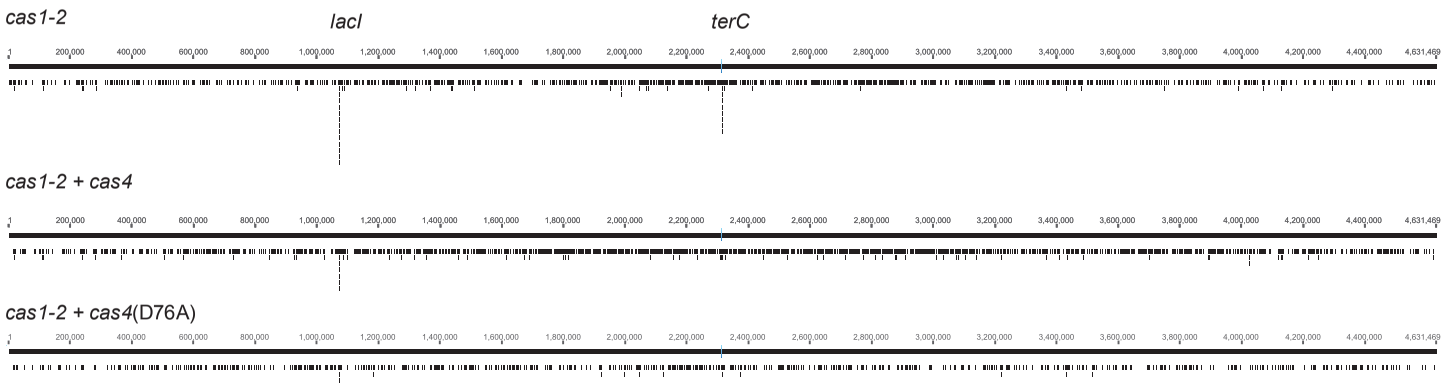


Figure 3.S3 related to New spacers are mainly genome derived. Origin of spacers acquired from the WT *E. coli* K12 genome. In the absence of *cas4* the Cas1-2 integration complex acquires spacers with preference for the lacI sequence and terC site. The preferential uptake of spacers derived from the *terC* site is lost when supplying Cas4 WT or Cas4$^{D76A.}$

SUPPLEMENTARY TABLES

Table 3.S1 – Plasmids used in this study

Name in this study	Name	Insert	Vector	Resistance	Source
pCas2	pTU084	*Synechocystis* sp. 6803 Type I-D *cas2* (deltaCas1)	pET-T7	Amp	This study
pCas1	pTU085	*Synechocystis* sp. 6803 Type I-D *cas1* (deltaCas2)	pET-T7	Amp	This study
pCas4^{D76A}	pTU086	*Synechocystis* sp. 6803 Type I-D *cas4* (D76A)	pET-T7	Spec	This study
pCas4	pTU130	*Synechocystis* sp. 6803 Type I-D *cas4*	pET-T7	Spec	This study
pCRISPR	pTU134	*Synechocystis* sp. 6803 Type I-D Leader-R-S1	pACYCDuet1	Cm	This study
pCas1-2	pTU70	*Synechocystis* sp. 6803 Type I-D *cas1-cas2*	pET-T7	Amp	This study
pEmp	pTU116	NA	pET-T7	Spec	Addgene Plasmid #48329
pVZ322	pNT	NA	pVZ322	Gent	This study
pVZ322 pT-GTA	pT-GTA	GTA-protospacer1-GentR	pVZ322	Gent	This study
pVZ322 pT-GTT	pT-GTT	GTT-protospacer1-GentR	pVZ322	Gent	This study
pVZ322 pT-GTC	pT-GTC	GTC-protospacer1-GentR	pVZ322	Gent	This study
pVZ322 pT-GTG	pT-GTG	GTG-protospacer1-GentR	pVZ322	Gent	This study

Table 3.S2 Oligonucleotides used in this study

Name	Sequence	Description
BG7615	TTATGGAGTTGGGATCTTATTAGATAATAATACTACCAGGTTTTTCTGGTTG	*cas2* rv
BG8223	TACTTCCAATCCAATGCAATGGATGATTATTTACCTTTAGC	*cas4* fw
BG8224	TTATCCACTTCCAATGTTATTATTTAGTAAGTTTTTTTAATTCTTTCGG	*cas4* rv
BN015	CGTCCATGGGAAGTCATTCTTCAAATTTTGGC	leader fw
BN016	TACAAGCTTAGGCATTGAAAGCGACC	Sp1 Rv (for degenerate PCR)
BN114	TTTAAGAAGGAGATATAGATCATGTCTACACTTTACTTGACTCAACC	*cas1* fw
BN135	GCTGCAGTGGAAGAAAGTG	Type I-D *cas4* mutagenesis D76A Fw
BN136	AATAATCCCTTTAACTTTTAGGCG	Type I-D *cas4* mutagenesis D76A Rv

BN143	GCGATCGGGACTGAAACT	Degenerated Fw1
BN144	GCGATCGGGACTGAAACA	Degenerated Fw2
BN145	GCGATCGGGACTGAAACC	Degenerated Fw3
BN156	AGGCATTGAAAGCGACC	Degenerate PCR Rv (internal. Sp1)
BN172	AGATCTGCCATATGTATATCTCCTTC	pAcyc backbone Rv (for degenerate PCR)
BN212	GATCTATATCTCCTTCTTAAAGTTAAAC	*cas1* deletion Rv
BN213	CAGTTATCAGTTGTGTTTTGAC	*cas1* deletion Fw
BN214	TTTTTAGTCGTCAAAACACAAC	*cas2* deletion Rv
BN215	TAATAAGATCCCAACTCCATAAG	*cas2* deletion Fw
GentaR_ pUC19_fwd	CGGTGATGACGGTGAGATTCCATTTTTACACTGATGAATGTTCCGTTGCG	Gentamicin resistance *cas*sette with overlaps to pUC19
GentaR_ pUC19_rev	CGCCTTTGAGTGAGCTCCCGGCATTCGCTGCGCT	Gentamicin resistance *cas*sette with overlaps to pUC19
CRISPR1_ GTG_S1_ fwd	GTGGATTGTTGTGCCCCTGGCGGTCGCTTTCAATGCCTTTAACAATTCGTTCAAGCCGAGATC	GTG PAM motif spacer 1
CRISPR1_ GTG_S1_ rev	AAGGCATTGAAAGCGACCGCCAGGGGCACAACAATCCACGGTGGCGGTACTTGGGTC	GTG PAM motif spacer 1
CRISPR1_ GTA_S1_ fwd	GTAGATTGTTGTGCCCCTGGCGGTCGCTTTCAATGCCTTTAACAATTCGTTCAAGCCGAGATC	GTA PAM motif spacer 1
CRISPR1_ GTA_S1_rev	AAGGCATTGAAAGCGACCGCCAGGGGCACAACAATCTACGGTGGCGGTACTTGGGTC	GTA PAM motif spacer 1
CRISPR1_ GTT_S1_ fwd	GTTGATTGTTGTGCCCCTGGCGGTCGCTTTCAATGCCTTTAACAATTCGTTCAAGCCGAGATC	GTT PAM motif spacer 1
CRISPR1_ GTT_S1_rev	AAGGCATTGAAAGCGACCGCCAGGGGCACAACAATCAACGGTGGCGGTACTTGGGTC	GTT PAM motif spacer 1
CRISPR1_ GTC_S1_ fwd	GTCGATTGTTGTGCCCCTGGCGGTCGCTTTCAATGCCTTTAACAATTCGTTCAAGCCGAGATC	GTC PAM motif spacer 1
CRISPR1_ GTC_S1_rev	AAGGCATTGAAAGCGACCGCCAGGGGCACAACAATCGACGGTGGCGGTACTTGGGTC	GTC PAM motif spacer 1
CRISPR1_ AGC_S1_ fwd	AGCGATTGTTGTGCCCCTGGCGGTCGCTTTCAATGCCTTTAACAATTCGTTCAAGCCGAGATC	AGC mock-PAM motif spacer 1
CRISPR1_ AGC_S1_rev	AAGGCATTGAAAGCGACCGCCAGGGGCACAACAATCGCTGGTGGCGGTACTTGGGTC	AGC mock-PAM motif spacer 1

GentaR_ pVZ322_ fwd	TCTGCTCCTGCAGGTCGACTGATTCCATTTTTACACTGATGAATGTTCCGTTGCGCTGCCC	Gentamicin resistance *cas*sette with overlaps to pVZ322
GentaR_ pVZ322_rev	CCCGGCATTCGCTGCGCTTATGGCAGAGCA	Gentamicin resistance *cas*sette with overlaps to pVZ322

Table 3.S3 Spacer mapping

		Genome		pCas4/pEmp		pCas1-2		pCRISPR		% Protospacers		
Strain	Cas4	Fw	Rv	Fw	Rv	Fw	Rv	Fw	Rv	G	P	% Protospacers with GTN PAM
WT	+	852	835	89	98	10	11	44	30	85.7	14.3	17.1
WT	-	642	627	262	247	27	26	123	97	61.9	38.1	4.7
WT	D76A	351	353	25	26	2	1	10	12	90.3	9.7	4.1
ΔrecB	+	49	51	46	47	4	6	16	14	42.9	57.1	24
ΔrecB	-	125	138	51	73	8	13	25	28	57.0	43.0	5.9

Table 3.S3 Spacer analysis of unique spacers. Total number of unique spacers acquired from the *E. coli* K12 genome, the *cas4* expression plasmid and corresponding empty vector control, the Cas1-Cas2 expression plasmid and the minimalized type I-D array plasmid in different strains and either presence or absence of *cas4*. Strand orientation indicated with Forward (Fw) or Reverse (Rv). The two last columns represent the percentage of protospacers that match the genome (G) or plasmids (P) and spacers matching a protospacer with the GTN PAM.

4

Cas4-Cas1 is a PAM-processing factor mediating half-site spacer integration during CRISPR adaptation

This Chapter has been Accepted for Publication in The CRISPR Journal

Sebastian N. Kieper[1,2], Cristóbal Almendros[1,2], Anna C. Haagsma[1,2], Arjan Barendregt[3,4], Albert J.R. Heck[3,4], Stan J.J. Brouns[1,2§]

1. Department of Bionanoscience, Delft University of Technology, Delft, Netherlands
2. Kavli Institute of Nanoscience, Delft, Netherlands.
3. Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research, Utrecht Institute of Pharmaceutical Sciences, Utrecht University, Utrecht, Netherlands.
4. Netherlands Proteomics Center, Utrecht, Netherlands.

4.1 ABSTRACT

The immunization of bacteria and archaea against invading viruses via CRISPR adaptation is critically reliant on the efficient capture, accurate processing and integration of CRISPR spacers into the host genome. The adaptation proteins Cas1 and Cas2 are sufficient for successful spacer acquisition in some CRISPR-Cas systems. However, many CRISPR-Cas systems additionally require the Cas4 protein for efficient adaptation. Cas4 has been implied in selection and processing of spacer precursors, but the detailed mechanistic understanding of how Cas4 contributes to CRISPR adaptation is lacking. Here we biochemically reconstitute the CRISPR-Cas type I-D adaptation system and show two functionally distinct adaptation complexes, Cas4-Cas1 and Cas1-Cas2. The Cas4-Cas1 complex recognizes and cleaves PAM sequences in 3' overhangs in a sequence-specific manner, while the Cas1-Cas2 complex defines the cleavage of non-PAM sites via host factor nucleases. Both sub-complexes are capable of mediating half-site integration, facilitating the integration of processed spacers in the correct, interference-proficient orientation. We provide a model in which an asymmetric adaptation complex differentially acts on PAM and non-PAM containing overhangs, providing cues for the correct orientation of spacer integration.

4.2 Introduction

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and their associated genes (*cas*) provide adaptive and inheritable immunity against mobile genetic elements (MGEs) in bacteria and archaea [1]. The CRISPR array is composed of palindromic repeats interspersed by sequences derived from MGEs and serves as a template for the biogenesis of CRISPR RNAs (crRNAs) [2, 3]. Cas proteins subsequently assemble around the crRNA to form effector complexes that mediate the recognition and destruction of invading MGEs that have been recorded in the bacterial genome during previous infections [4]. Therefore, the main requirement for the establishment of immunity is the memorization of foreign genetic material in a step called CRISPR adaptation [5, 6]. The core machinery responsible for adaptation is composed of the Cas1 and Cas2 proteins that assemble into the adaptation complex [7-9]. Among the first identified *cas* genes were the adaptation genes *cas1* and *cas2*, the *cas3* gene encoding the nuclease-helicase Cas3 and the *cas4* gene encoding a protein, the function of which has been unknown until recently [10, 11]. The *cas4* gene is widespread among several sub-types of type I, type II and type V systems and therefore present in the majority of CRISPR-Cas systems [12]. Predictions of Cas4 function existed early on, based on the frequent colocalization of the CRISPR adaptation genes *cas1* and *cas2* and the *cas4* gene. This co-localization suggested that Cas4 could be contributing to the adaptation stage and the early studies indeed found supporting evidence for this hypothesis: Adaptation of the type I-A system of *Sulfolobus islandicus* was severely impaired upon deletion of *cas4* [13]. Similarly, deleting *cas4* from the type I-B system of *Haloarcula hispanica* abrogated CRISPR adaptation against HHPV-2 [14]. Biochemical evidence was provided by Plagens et al. showing a protein-protein interaction *in vitro* between Cas4 and the type I-A adaptation fusion protein Cas1/2 and Csa1 demonstrating that Cas4 directly interacts with the adaptation machinery [15]. Recently, several studies defined the role of Cas4 in more detail, finding that the presence of Cas4 increases the fidelity of spacer integration. Specifically, the Cas4 protein of a cyanobacterial type I-D CRISPR-Cas system facilitated the integration of interference-proficient spacers that carry the consensus PAM of the type I-D CRISPR-Cas system. Spacers acquired in the presence of Cas4 displayed shorter lengths compared to those acquired in the absence of Cas4 or in the presence of catalytically inactive variant of the protein

[16]. Additionally, two Cas4 variants (Cas4-1 and Cas4-2) encoded in the type I-A system of *Pyrococcus furiosus* were shown to define the upstream protospacer adjacent motif (PAM) and a downstream NW motif *in vivo* [17]. Deletion of *cas4-1* and *cas4-2* resulted in incorrect processing of prespacers with respect to up- and downstream motifs, random orientation of the integrated spacer as well as large deviations from the consensus spacer length [17]. Previously, biochemical studies of the Cas4 protein, containing an iron-sulfur cluster and a RecB domain, found several nuclease activities, demonstrating endo- and exonuclease activities [18-20]. The requirement of Cas4 for prespacer processing was therefore in accordance with the previously described biochemical activities. Indeed, Lee et al. provided the first mechanistic details of how Cas4 proteins ensure PAM-processing and correct spacer orientation [21, 22]. It was shown that Cas4 tightly interacts with the Cas1 integrase, forming a heterohexameric complex composed of two Cas1 dimers and two Cas4 subunits [21]. This complex would interact with double-stranded prespacer substrates and endonucleolytically cleave PAM sequences in long 3' overhangs, ensuring that only PAM-processed spacers would be eventually integrated into the CRISPR array [21]. Interestingly, the authors did not see interaction of the Cas4-Cas1 complex with Cas2 in their initial experiments, suggesting the possibility that a prespacer is required for assembly of the full complex. After supplying a dsDNA substrate to the three adaptation proteins, Lee et al. could demonstrate the assembly of the full Cas4-Cas1-Cas2 complex [22]. This complex was shown to assemble in a mixture of symmetric and asymmetric architectures as shown by negative-staining Electron Microscopy, in which the asymmetric complex would contain only a Cas4 monomer associated with one of the Cas1 dimers [22]. The authors suggested that this asymmetry might aid in the differential processing of prespacer substrates in which the Cas4 containing half of the complex would interact with the PAM-containing overhang. This hypothesis is supported by the findings in the type I-A system, in which two independent Cas4 homologs are dedicated processing factors for the PAM- and NW motif containing prespacer overhangs [17]. However, how this asymmetric processing is orchestrated in CRISPR systems containing only a single *cas4* gene is currently unknown. In this work we provide mechanistic insights of an asymmetric complex, specifically how the Cas4-Cas1 complex is able to recognize and sequence specifically process the PAM sequence of

the type I-D CRISPR-Cas system. Previously we have shown that the type I-D Cas4 protein facilitates the integration of PAM-compliant spacers *in vivo* [16]. We demonstrate that Cas4 strongly interacts with the Cas1 integrase forming a heteromeric $Cas4_1$-$Cas1_2$ complex. This heteromeric complex does not require the Cas2 protein for processing and half-site integration of PAM-containing prespacer substrates. The catalytic activity of Cas4 is required for prespacer cleavage and is crucially dependent on the presence of Cas1 in order to recognize and process the PAM overhang. We show that this Cas4-Cas1 complex does not cleave the non-PAM containing overhang. Processing of the non-PAM containing overhang potentially relies on the Cas1-Cas2 complex and likely requires host-factor nucleases. We provide a model in which an asymmetric adaptation complex differentially acts on PAM and non-PAM containing overhangs, providing cues for the correct orientation of spacer integration. This correct PAM processing as well as a functional orientation explains the importance and hence the strong conservation of the *cas4* gene, increasing the integration of interference-proficient spacers.

4.3 MATERIAL & METHODS

4.3.1 BACTERIAL STRAINS AND GROWTH CONDITIONS

E. coli strains DH5α and BL21 were grown in Lysogeny Broth (LB) at 37°C and continuous shaking at 180 rpm or grown on LB agar plates (LBA) containing 1.5% (wt/vol) agar. When required, the media were supplemented with 100 μg ml^{-1} ampicillin, 50 μg ml^{-1} spectinomycin, 25 μg ml^{-1} chloramphenicol (see Table S1 for plasmids and corresponding selection markers).

4.3.2 PLASMID CONSTRUCTION AND TRANSFORMATION

Plasmids used in this study are listed in Table S1. All cloning steps were performed in *E. coli* DH5α. Primers described in Table S2 were used for PCR amplification of the type I-D CRISPR-Cas locus (*cas4*, *cas1*, *cas2* and leader-repeat-spacer1) from *Synechocystis* cell material using the Q5 high-fidelity Polymerase (New England Biolabs). PCR amplicons were subsequently cloned into Berkeley MacroLab LIC vectors (https://qb3.berkeley.edu/facility/qb3-macrolab/) using either ligation-independent cloning (LIC), or into the pACYCDuet-1 vector system (Novagen (EMD Millipore) using conventional restriction-ligation cloning. The $cas4^{D76A+K91A}$ mutant was obtained using a PCR-based mutagenesis of $pCas4^{D76A}$ using primers listed in Table S2. All plasmids were verified by Sanger-sequencing (Macrogen Europe, Amsterdam, The Netherlands). Bacterial transformations were either carried out by electroporation (2.5 kV, 25 mF, 200 V) using a ECM 630 electroporator (BTX Harvard Apparatus) or using chemically competent cells prepared according to manufacturer's manual (Mix&Go, Zymo research). Electrocompetent cells were prepared following a protocol adapted from [23]. Transformants were selected on LBA supplemented with appropriate antibiotics.

4.3.3 PROTEIN EXPRESSION AND PURIFICATION

Plasmid encoded *cas* genes were either co-expressed or expressed individually in *E. coli* BL21 AI cells (Invitrogen). Pre-cultures were grown from individual colonies and used for inoculation pre-warmed (37°C) LB medium at an initial OD_{600}=0.05. Protein expression was induced at OD_{600}=0.5 by addition of IPTG and L-arabinose preceded by a 30-minute cold-shock. Cultures were subsequently grown over-

night at 20°C and continuous shaking. Cells were harvested by centrifugation (10 min, 4°C, 2400xg) and subsequently resuspended in lysis buffer (50 mM HEPES pH 7.5, 300 mM KCl, 5% Glycerol, 1 mM DTT, 25 mM Imidazole, 0.1% Triton-X 100) supplemented with cOmplete™, EDTA-free Protease Inhibitor Cocktail (Roche). Cells were lysed by two passages through a CF1 cell disruptor (Constant Systems Ltd.) equilibrated with lysis buffer at a constant pressure of 1 kbar. Lysates were cleared by centrifugation (45 min, 4°C, 25000xg) and filtered through 0.45 µm filter. Protein was bound in batch to HIS-Select (Sigma Aldrich) IMAC resin for 30 min at 4°C and rotary shaking. IMAC resin was then loaded onto Pierce gravity-flow columns (Thermo Scientific) and washed with 10 CV wash buffer (50 mM HEPES pH 7.5, 300 mM KCl, 5% Glycerol, 1 mM DTT, 50 mM Imidazole). Proteins were subsequently block eluted in 0.5 ml elution buffer (50 mM HEPES pH 7.5, 300 mM KCl, 5% Glycerol, 1 mM DTT, 250 mM Imidazole). Protein concentration and purity was determined by NanoDrop A280 spectroscopy and SDS PAGE analysis. Protein elution fractions were pooled and subjected to size exclusion chromatography using Superdex 200 10/300 GL (GE Healthcare) column with 0.5 ml/min flow rate using elution buffer as mobile phase. Cas1-Cas2 complex IMAC elution fractions used for integration assays were prepared for ion-exchange chromatography by adjusting the KCL concentration to 30 mM and subsequently loaded onto HiTrap Heparin HP column (GE Healthcare). Cas1-Cas2 complexes were eluted by gradually increasing KCL concentration to 1 M. Resulting fractions were analyzed by SDS-PAGE and appropriate fractions pooled, snap frozen and stored at -80°C.

4.3.4 Native Mass Spectrometry

Cas4-Cas1 and Cas1-Cas2 complexes were buffer exchanged into 500 mM ammonium acetate (pH 7.5) using seven sequential steps on a centrifugal filter with a molecular weight cut-off of 10 kDa (Sartorius) at 4°C. MS measurements were performed in positive mode by directly infusing the individual complexes at a concentration of 1 µM using an LCT electrospray time-of-flight (Waters, United Kingdom) adjusted for optimal performance in high mass detection [24, 25]. The needles used for electrospray were prepared in house from borosilicate capillaries (Kwik-Fil, World Precision Instruments, Sarasota, FL) on a P97 puller (SutterInstruments, Novato, USA) and gold coated by using Ed-

wards Scancoat Six Pirani 501 Sputter Coater (Edwards Laboratories, Milpitas, USA). During measurement the capillary voltage was kept at 1200V, cone voltage between 80-150V and the source pressure was increased to ≈8mbar. Exact mass measurements of the individual Cas proteins were acquired under denaturing conditions by adding formic acid to a final concentration of 5%. All spectra were mass calibrated by using an aqueous solution of cesium iodide (25 mg/ml). Mass spectra were accumulated, averaged, smoothened and centered, using the software MassLynx 4.1 (Waters, United Kingdom).

4.3.5 NUCLEASE ASSAYS

Oligo-nucleotide sequences used in this study are indicated in Table S2. Oligo-nucleotides with C6-Amino modifications on the 5' terminus were obtained from ELLA Biotech (Planegg, Germany). Cy5 or Cy3 (GE Healthcare) labelling of 5' termini was done in 100 mM Sodium-bicarbonate buffer as described by [26]. Unlabeled oligo-nucleotides were obtained from Integrated DNA Technologies (IDT). Nuclease assays were performed in buffer R (5 mM HEPES pH 7.5, 100 mM Sodium-Glutamate supplemented with 2 mM MnCl2 and 10 mM MgCl2. Annealed and Cy5 and Cy3 labelled oligo-nucleotides (Cy3-BN1829+Cy5-BN1830) were added to a final concentration of 125 nM and purified protein complexes to a final concentration of 500 nM. Reactions were incubated for 1 hour at 30°C after which reactions were quenched by addition of Proteinase K (Thermo Fischer) and incubation for 1 hour at 37°C. The resulting products were analyzed on denaturing PAGE (10% acrylamide, 8M Urea) and analyzed with Amersham Typhoon fluorescence gel scanner (GE Healthcare).

4.3.6 *IN VITRO* SPACER INTEGRATION ASSAYS

Oligo-nucleotide integrations with either Cy5 labeled or unlabeled oligo-nucleotides were performed by pre-incubating indicated protein complexes (500 nM) with oligo-nucleotides (250 nM) on ice for 15 min. Following pre-incubation, either linear CRISPR substrate (obtained by Q5 high-fidelity PCR from pCRISPR using primers BN015+BN1398 or supercoiled pCRISPR were added to a final concentration of 7.5 nM. Reaction mixtures were incubated at 30°C for 1 hour after which reactions were quenched by addition of Proteinase K (Thermo Fischer) and incubation for 1 hour at 37°C. Reactions were run on 1% native agarose gels for 45 min and gels subsequently

stained with SYBR gold (Sigma Aldrich). Gels were scanned for Cy5 and SYBR gold using Amersham Typhoon fluorescence gel scanner (GE Healthcare). For PCR analysis of *in vitro* integration, unlabeled oligo-nucleotides were used in the reaction. Open-circular plasmid DNA was gel isolated and DNA purified using Zymoclean gel recovery kit (ZymoResearch) after which integration was assessed by PCR using primers BN1711+BN1713 (leader distal integration; correct spacer orientation), BN1711+BN1714 (leader distal integration; incorrect spacer orientation), BN1712+BN1713 (leader proximal integration; incorrect spacer orientation) and BN1712+BN1714 (leader proximal integration; correct spacer orientation).

4.3.7 Next generation sequencing and statistical analysis

After validation of PCR amplicons by gel electrophoresis and clean up with the GeneJET PCR Purification kit (Thermo Fisher Scientific) the samples were analyzed using Qubit fluorometric quantification (Invitrogen). Samples were prepared for sequencing with the Nextera XT DNA Library Preparation Kit (Illumina) and each library individually barcoded with the Nextera XT Index Kit v2 SetA (Illumina). Libraries were pooled equally and spiked with ~5% of the PhiX control library (Illumina) to artificially increase the genetic diversity before sequencing on a Nano flowcell (250 nt paired-end) with an Illumina MiSeq. Image analysis, base calling, de-multiplexing and data quality assessments were performed on the MiSeq instrument. FASTAQ files generated by the MiSeq were analyzed by pairing and merging the reads using Geneious 9.0.5 and subsequently extracting the oligo-nucleotide sequences used in the *in vitro* integration assay. Overhang processing was analyzed by annotating the primers used for amplification and comparing the overhangs post-integration to the initial oligo-nucleotide sequence.

4.3.8 *In vivo* spacer integration assays

E. coli BL21 AI cells were co-transformed with either pCas1-Cas2 and pEmpty or pCas1-2 and pCas4 (wild-type Cas4 or Cas4$^{D76A+K91A}$). One transformant for each combination was grown in LB at 37°C and continuous shaking (180 rpm) to OD_{600}=0.3 and made electrocompetent after which pCRISPR was transformed. For each treatment three individual colonies were grown in SOB medium (LB supplement-

ed with 10 mM $MgSO_4$ and 10 mM $MgCl_2$) at 37°C and continuous shaking (180 rpm) to OD_{600}=0.3 after which protein expression was induced by addition of 0.2% L-arabinose and 0.5 mM IPTG. Induced cultures were grown for additional 2 hours at 37°C and continuous shaking (180 rpm). Cells were made electrocompetent and annealed pre-spacer oligo-nucleotides (BN1763+BN1768) electroporated at a final concentration of 1 µM. After 30 min recovery cells were harvested and plasmid DNA extracted using GeneJET plasmid miniprep kit (Thermo Scientific). Extracted plasmid DNA was normalized to 0.5 ng μl^{-1} and subsequently 2 µl used in half-site integration PCRs using primers BN1711+BN1713 (leader distal integration; correct spacer orientation), BN1711+BN1714 (leader distal integration; incorrect spacer orientation), BN1712+BN1713 (leader proximal integration; incorrect spacer orientation) and BN1712+BN1714 (leader proximal integration; correct spacer orientation). PCR amplicons were validated by agarose gel electrophoresis and purified with the GeneJET PCR Purification kit (Thermo Scientific). Purified PCR amplicons were subjected to MiSeq sequencing (Illumina).

4.4 Results

4.4.1 PAM-containing overhang processing depends on orientation of spacer integration

We have previously demonstrated that the type I-D Cas4 protein facilitates the integration of PAM-compatible spacers *in vivo* [16]. In these experiments we looked at the total pool of spacers that were acquired from cytosolic DNA, obscuring the detailed mechanism that governs the processing of prespacer substrates. In order to obtain more detailed insights into processing of PAM and non-PAM containing substrates (Table S3 & S4) and how spacer orientation affects overhang processing, we electroporated an idealized prespacer substrate into *E. coli* cells overexpressing either Cas1-Cas2 or Cas4-Cas1-Cas2 proteins (Fig. 1A&B). Cas4 was either expressed as the wild-type protein or the catalytically inactive mutant (D76A, K91A). In addition to the adaptation genes, the cells were carrying a plasmid containing the type I-D leader and a single repeat (pCRISPR). By employing half-site integration PCRs followed by high-throughput sequencing, we analyzed the 3' overhangs after processing and integration *in vivo* (Fig. 1C). This approach allowed us to differentiate between correct and incorrect spacer orientation, as well as correct and incorrect PAM processing of their 3' end (Fig. 1D&E).

We observed that prespacer overhangs were trimmed by at least 5 nt in all cases, regardless of the presence or absence of Cas4. However, processing of PAM and non-PAM containing overhangs differed depending on the orientation in which the spacer was integrated. In particular, spacers integrated in the correct orientation (Fig. 1D) were more precisely processed in the PAM-containing overhang in the presence of Cas4. Although cells expressing only Cas1-Cas2 or a combination of Cas1-Cas2 with a catalytically inactive Cas4 double mutant (D76A, K91A) also displayed 30% to 35% of correct processing of the PAM overhang, their spacer size distributions were typically broader and shifted towards longer overhang lengths. Analyzing the non-PAM containing overhangs of correctly oriented spacers did not display any differences between the conditions (with and without Cas4), suggesting that prespacer overhangs without PAM are not processed by Cas4, but rather by endogenous *E. coli* nucleases. Spacers integrated in the incorrect orientation (Fig. 1E) showed similar 3' overhangs un-

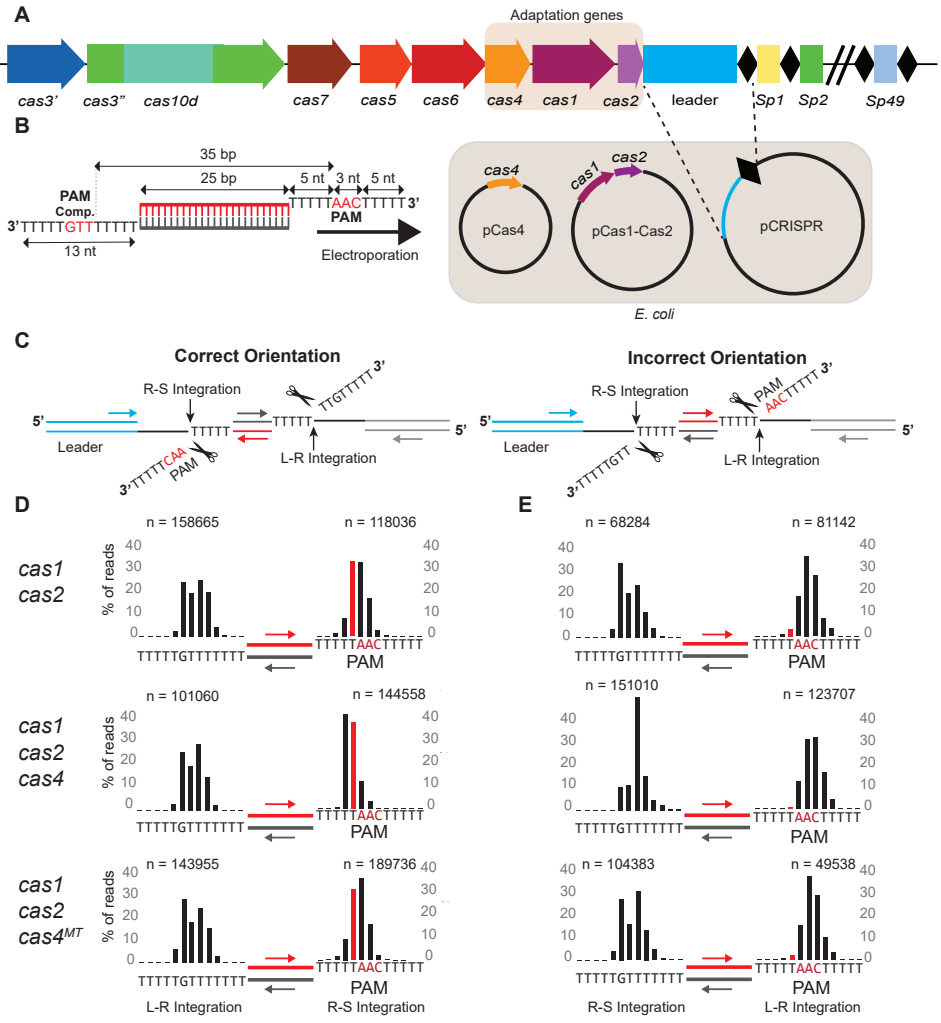


Figure 4.1 – Pre-spacer processing and half-site integration in vivo
A Genetic organization of the type I-D CRISPR-locus. Genes constituting the interference machinery are located upstream of the adaptation complex. The adaptation complex consisting of *cas4*, *cas1* and *cas2* is highlighted in purple. Downstream of *cas2* is the leader sequence followed by the type I-D array.
B Experimental design of spacer processing in vivo assay. Idealized pre-spacer substrates were electroporated into *E. coli* cells carrying the plasmid encoded minimalized type I-D array and expressing the adaptation genes *cas1* and *cas2*. The Cas4 protein was either omitted or co-expressed as the wild-type or D76A+K91A mutant protein. Half-site integration was assessed by PCR as depicted in C.
C PCR scheme allowing differentation of spacer orientation and integration site. PCR amplicons of correct and incorrect spacer orientations and Leader-Repeat (L-R) or

Repeat-Spacer (R-S) integration site were subjected to high-throughput sequencing. **D-E** Overhang processing resulting from high-throughput sequencing of half-site integration PCR. Integration was assessed in correct (**D**) or incorrect (**E**) spacer orientation in either the Cas1-Cas2, Cas1-Cas2 and Cas4 wild-type or Cas4 mutant (MT) background. n = number of integration events.

der all conditions. We observed most accurate processing when Cas4 was present in its active form. In those samples the presence of Cas4 led to an increase in the shortening of overhangs, with a predominant overhang length of 6 nucleotides. In the *E. coli* model system, host factor nucleases potentially act on both PAM and non-PAM containing 3' overhangs that remain unprotected by the core Cas1-Cas2 complex holding the prespacer. Cas4 does not specifically cleave non-PAM containing overhangs, but requires the presence of the PAM in order to engage in sequence-specific processing.

4.4.2 Cas4 forms a strong heteromeric complex with Cas1

In order to assess whether the overhang processing connected to the presence of Cas4 was a result of Cas4 specifically interacting with the Cas1-Cas2 integration complex, we first investigated the formation of a Cas1-Cas2 complex. The Cas1 protein was N-terminally His6-tagged and co-expressed with Cas2 in *E. coli* BL21-AI cells. After the initial nickel-affinity pull-down from cleared cell lysate, the elution fraction was subjected to size exclusion chromatography (SEC), which resulted in one peak containing aggregated protein and another peak species (Fig. 2A). This peak contained three proteins (Fig. 2A) for which the tagged-Cas1, untagged Cas1 (likely due to proteolytic cleavage of the tag by endogenous *E. coli* proteases) and Cas2 protein identity was confirmed by mass spectrometry. Next, we co-expressed the His6-tagged Cas1 with untagged Cas4 and observed strong co-purification of both proteins (Fig. 2B). In order to verify the Cas4-Cas1 interaction, a reverse tagging strategy was used (His6-tagged Cas4 co-expressed with untagged Cas1), which again confirmed the presence of a Cas4-Cas1 complex (Fig. S1). When tagged-Cas1 and Cas2 were co-expressed along with Cas4, we observed a strong co-purification of Cas4 and Cas1 that abolished formation of the Cas1-Cas2 complex since Cas2 eluted separately as a low molecular weight species. This fraction also contained minor amounts of Cas1 and Cas4 that did not assemble into higher order complexes. Our results demonstrate that under these

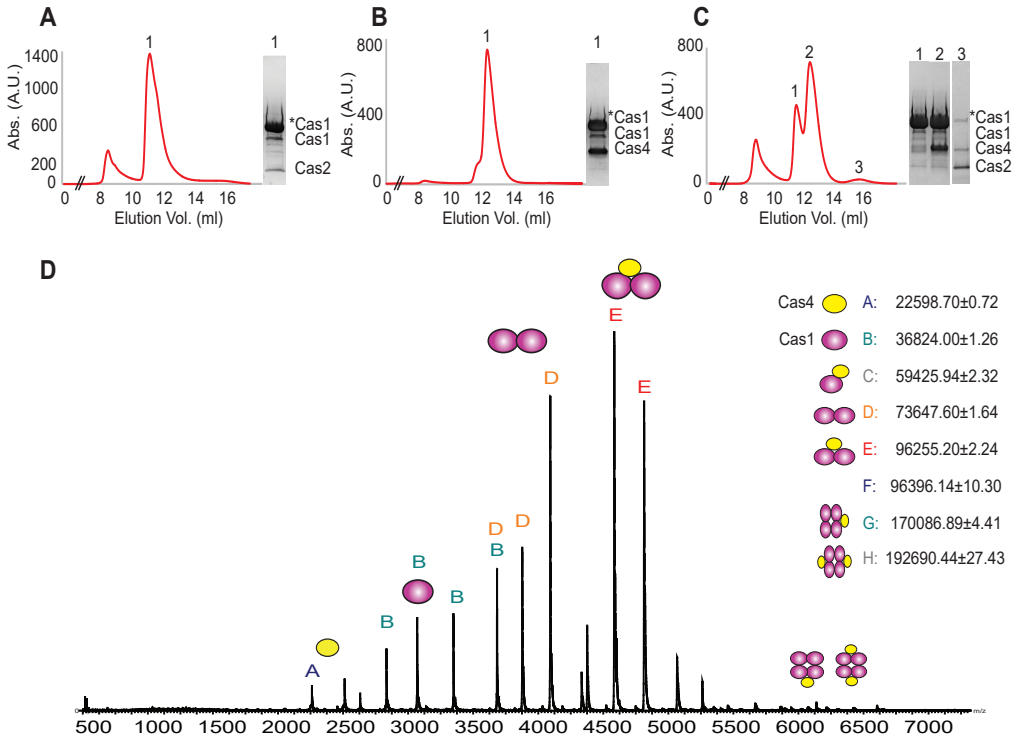


Figure 4.2 - Size exclusion chromatograms and SDS-PAGE analysis of peak fractions (indicated by numbering; non-numbered peaks contain protein aggregates).
A N-terminally tagged *Cas1 associates with untagged Cas2 in the absence of Cas4
B Complex formation of N-terminally tagged *Cas1 and untagged Cas4.
C Co-expression of *Cas1, Cas2 and Cas4. *Cas1 elutes separately (peak 1) from the Cas4-Cas1 complex (peak 2). Cas2 together with dissociated Cas1 and Cas4 elutes as a low molecular weight peak (peak 3).
D Native Mass Spectrometry of Cas4-Cas1 complex as shown in B (native spectrum obtained after removal of His-SUMO tag from Cas1 by TEV protease cleavage). Cas4 monomers assemble with Cas1 dimers into a Cas41-Cas12 complex. Cas1 dimers are also frequently observed. Free monomers of Cas1 and Cas4 are less frequent in the measured sample

conditions Cas1 can form complexes with Cas4 or Cas2, and that these complexes appear to be mutually exclusive. In the presence of both Cas4 and Cas2, Cas1 strongly favors the interaction with Cas4 over the interaction with Cas2.

4.4.3 Cas4 associates with Cas1 in a 1:2 ratio

Next, we determined the stoichiometry of the formed Cas4-Cas1 complex. Previously, Lee et al. demonstrated that the heteromeric complex consists of two Cas1 dimers that each associate with a single Cas4 monomer [21]. To gain insight into the composition of the untagged Cas4-Cas1 complex (Fig. S2) native protein mass spectrometry analysis was performed [25]. The mass spectrum (Fig. 2D) revealed a distribution of different complex species, with the most abundant mass-over-charge (m/z) peaks consisting of either Cas1 dimers (73.6 ± 1.6 kDa) or the Cas4-Cas1 complex consisting of a single Cas1 dimer and a Cas4 monomer resulting in a Cas41-Cas12 complex of 96.3 kDa (Fig. 2D). Even though we observed co-purification of Cas1 and Cas2 in the SEC analysis, the native mass spectrum of the Cas1-Cas2 complex resulted in mainly Cas1 dimers (Fig. S3) with Cas2 likely being lost during the native MS sample preparation.

4.4.4 The Cas4-Cas1 complex sequence specifically processes PAM-containing 3' overhangs

The acquisition of functional spacers not only requires appropriate prespacer selection, but also PAM-compliant processing. We have previously shown that the presence of Cas4 in addition to the core adaptation proteins Cas1 and Cas2 significantly increases the integration of spacers with a correctly processed PAM *in vivo* [16]. Due to the strong interaction of Cas4 and Cas1 we aimed to test whether PAM processing is mediated only by Cas4, or if the heteromeric Cas4-Cas1 complex is required. In order to address this question, we performed prespacer cleavage assays with a dual-labelled model prespacer (Fig. 3A). This model prespacer consisted of a 25 bp duplex flanked by 13 nucleotide 3' overhangs on each side and fluorescent labels at their 5' ends. The top strand was labelled with Cy3 and did not contain a PAM sequence in its 3' overhang, while the bottom strand was labelled with Cy5 and contained the I-D consensus PAM. We found that neither free Cas4 nor Cas1-Cas2 was able to catalyze 3' overhang cleavage. However,

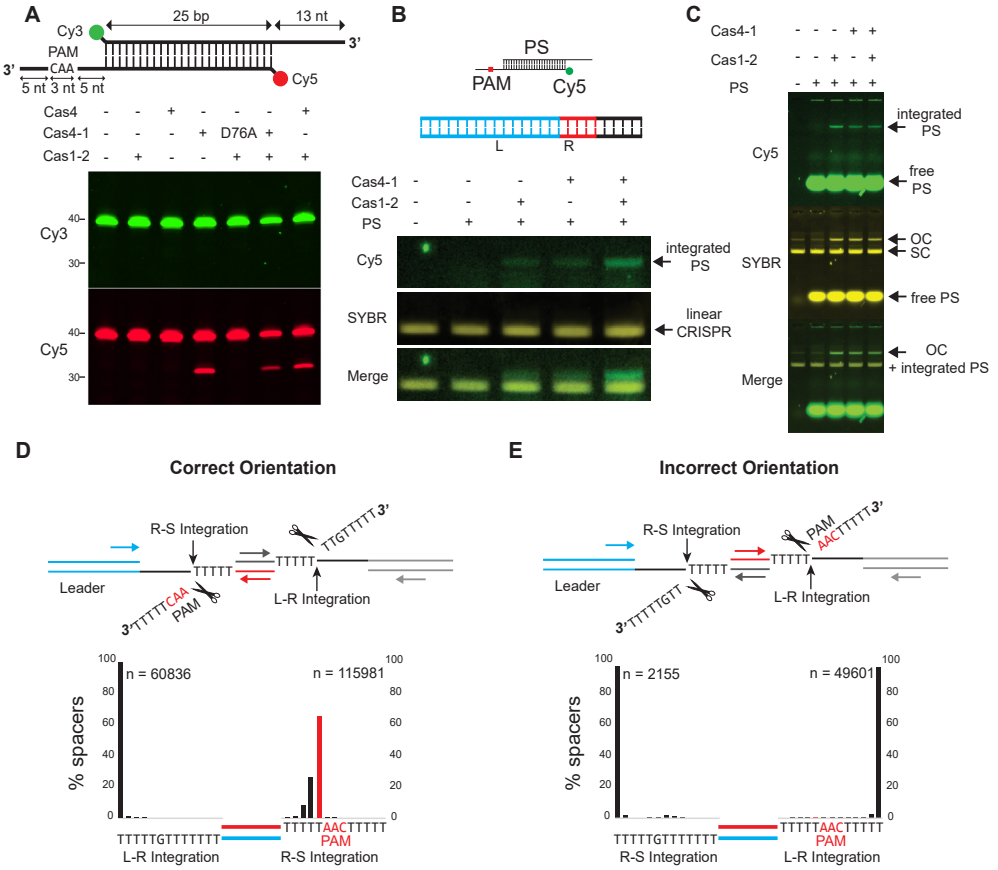


Figure 4.3 – *In vitro* pre-spacer cleavage and integration A Pre-spacer model substrate containing a 25 bp duplex region flanked by 13 nt 3' overhangs incubated with different protein combinations. The non-PAM strand is 5' Cy3 labelled and the PAM-containing strand 5' Cy5 labelled. Cleavage of PAM containing 3' overhang results in Cy5 labelled fragment of 30 nt. Protein samples consist of co-purified Cas4-Cas1 and Cas4D76A-Cas1 complexes, combined Cas4-Cas1 and Cas1-Cas2 complex and individually purified Cas4 in the presence of Cas1-Cas2.
B Integration of labelled pre-spacer into linear CRISPR DNA consisting of the type I-D leader sequence and a single Repeat. Labelled spacer imaged via Cy5, total DNA via SYBR gold stain. Merge of Cy5 and SYBR gold channels indicates integration of Cy5 labelled pre-spacer resulting in a higher molecular weight band.
C Labelled pre-spacer integration into pCRISPR DNA. Similar to B, both adaption complexes facilitate integration into plasmid encoded CRISPR locus. Integration reaction is accompanied by nicking of supercoiled (SC) plasmid DNA, resulting in formation of open-circular (OC) plasmid conformation. Merge image of Cy5 and SYBR channels shows co-localization of OC plasmid species and Cy5 labelled spacer substrate.
D-E High-throughput sequencing of Cas4-Cas1 integrated pre-spacers. Reaction was performed similar to assay shown in **C** using unlabeled pre-spacer DNA. OC plasmids were gel extracted followed by PCRs specific for the leader-repeat (L-R) and repeat-spacer (R-S) integration as well as correct and incorrect spacer orientation. Bar graphs represent the percentage of spacers with specific overhang length depending on integration site and orientation (D-correct; E-incorrect).

the addition of the Cas4-Cas1 complex resulted in a defined band corresponding to processing of the PAM sequence within the PAM-containing overhang (Fig. 3A). This result suggests that PAM recognition is mediated by the interactions within the Cas4-Cas1 complex, where Cas4 acts as the catalytic subunit of the complex.

We have previously shown that mutating D76 in the conserved RecB domain of Cas4 abolished integration of PAM-proficient spacers *in vivo* [16]. Using the D76A mutant in our *in vitro* cleavage assay fully abolished processing activity of the Cas4-Cas1 complex, demonstrating that the RecB domain of Cas4 is indeed the catalytically active site required for PAM processing. Interestingly, although Cas4 did not show processing activity on its own, combining Cas1-Cas2 and Cas4 fully restored processing most likely due to de novo assembly of the Cas4-Cas1 complex. This finding is in line with our co-purification experiments in which Cas4 outcompetes Cas2 for binding with Cas1. The addition of both, the Cas1-Cas2 and the Cas4-Cas1 complex, resulted in processing of the PAM overhang as observed with the Cas4-Cas1 complex alone or combination of Cas4 and Cas1-2 complex. All conditions that showed cleavage of the substrate resulted in a single defined band, suggesting that cleavage occurred via an endonuclease mechanism which is in line with previous studies [21]. The processing of the non-PAM containing overhang was not observed in any of the conditions, indicating that the processing of the non-PAM site presumably relies on host factor nucleases such as DnaQ-like exonucleases or Exonuclease T as recently found in in the I-E system [27, 28]. Our results demonstrate that sequence specific Cas4 activity requires the presence of Cas1 and that the Cas4-Cas1 complex is the core processing complex that sequence-specifically recognizes and processes the PAM sequence before integration.

4.4.5 The Cas4-Cas1 complex integrates new spacers into both linear and supercoiled DNA

Next, we tested whether the Cas4-Cas1 complex not only processes prespacer substrates but also catalyzes their integration into the CRISPR array. We performed adaptation assays using supercoiled plasmid DNA containing the type I-D leader and a single repeat (pCRISPR; Fig. 1A) as well as linear CRISPR array substrates generated by PCR (Fig. 3B). Both linear and plasmid CRISPR loci were incubated with

a Cy5-labelled prespacer, the Cas4-Cas1 and Cas1-Cas2 complexes. Intriguingly, we observed coupling of the labelled prespacer by the Cas4-Cas1 complex to both CRISPR substrates, showing that this sub complex is proficient in catalyzing at least half-site spacer integration (Fig. 3B&C). Spacer integration into plasmid DNA resulted in the formation of open-circular (OC) plasmid conformations. Merging the Cy5 signal of the prespacer and the plasmid DNA signal confirmed that the prespacer was indeed coupled to the OC form of the plasmid. The Cas1-Cas2 complex was able to integrate the prespacer into both linear and supercoiled arrays similar to the Cas4-Cas1 complex. Our observation demonstrates that at least two different sub-complexes exist, which are both capable of catalyzing half-site spacer integration. Taken together, based on the selective PAM-overhang processing of the Cas4-Cas1 complex, we hypothesize that the Cas4-Cas1 complex processes and integrates the PAM containing overhang and the Cas1-Cas2 complex the non-PAM containing overhang.

4.4.6 Correct spacer orientation requires overhang processing prior to integration

In order to analyze the accuracy of spacer integration by the Cas4-Cas1 complex in more detail, OC plasmid resulting from the integration reaction was gel purified and subjected to half-site integration PCRs as described previously (Fig. 1C). PCR products were subjected to Illumina MiSeq sequencing and prespacer sequences were extracted (Table S5). This approach allowed us to assess 3' overhang processing before integration at the leader-proximal or leader-distal integration site. Interestingly, PAM-containing overhangs only showed sequence specific processing when the spacer was correctly oriented with respect to the PAM (Fig. 3D). The Cas4-Cas1 complex cleaved 65% of correctly oriented spacers exactly downstream of the PAM, however, we also observed incorrect removal of a single nucleotide in 25% of sequences and removal of 2 or more nucleotides in 10% of the sequences. Surprisingly, incorrectly oriented spacers did not show any processing of the PAM-containing overhang (Fig. 3E), indicating that integration in the correct orientation is preceded by the processing of the overhang. As predicted from the bulk cleavage assays, we did not observe any processing of the non-PAM containing overhangs regardless of the spacer orientation. Our data show that integration of new spacers in the correct orientation by Cas4-Cas1 requires PAM recognition and

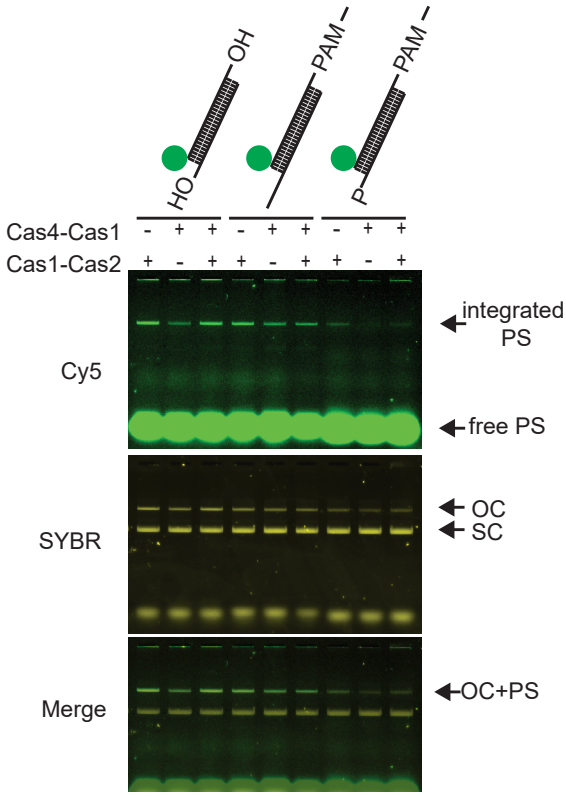


Figure 4.4 - Integration activity with respect to 3' overhang requirements. Pre-spacer substrates were 5' Cy5 labelled in order to follow coupling to pCRISPR. Phosphorylated (P) 3' overhangs were used to block integration of one of the DNA strands.

processing before a spacer can be integrated.

4.4.7 Spacer integration preferentially initiates with the non-PAM overhang

In type I CRISPR-Cas systems spacer integration initiates by first integrating the non-PAM end of the spacer at the leader repeat junction and proceeding with the coupling of the PAM-end of the spacer at the Repeat-Spacer boundary [29-31]. In the Type I-E CRISPR-Cas system that is lacking Cas4, directionality of spacer integration is dictated by the prespacer processing kinetics [27, 28]. We therefore hypothesized that the prespacer processing of our Cas4 containing system could influence the orientation of the integrated spacer. To test the effect of the processed and unprocessed prespacer overhangs on integra-

tion, we assayed spacer integration using the Cy5-labelled prespacer substrates. Prespacers were either fully processed (5 nt 3' overhangs), with an unprocessed non-PAM overhang (13 nt) or with a processed, but integration-deficient [31, 32] 3' phosphorylated non-PAM overhang (Fig. 4.4). The fully processed substrate was efficiently coupled by Cas1-Cas2, Cas4-Cas1 as well as the combination of both. Similarly, the prespacer with an unprocessed non-PAM overhang was coupled efficiently in all three treatments. However, when the processed non-PAM overhang was blocked for integration by 3'-phosphorylation, neither of the protein complexes was able to efficiently couple the spacer, indicating that coupling of the PAM-overhang requires prior integration of the non-PAM overhang. Altogether, this mechanism ensures that integration of the PAM site of the spacer is halted until integration of the non-PAM site has occurred, resulting in the correct orientation of the spacer with respect to the PAM.

4.5 DISCUSSION

Although Cas4 proteins have been recognized as part of the core *cas* gene machinery almost two decades ago [11], its role in acquiring PAM-compatible spacers has been revealed only in the recent years [16, 17, 21, 22, 33]. Here we provide a new mechanistic understanding of how Cas4-dependent PAM selection is achieved during CRISPR adaptation, and specifically how asymmetry of the adaptation complex drives the selection, processing and integration of PAM-compatible spacers. We present a model in which two independent subcomplexes, Cas4-Cas1 and Cas1-Cas2, selectively process the two 3' overhangs of a prespacer (Fig. 5). The interaction of Cas4 with the Cas1 integrase protein is central to the recognition and processing of PAM-containing prespacer substrates. Formation of this Cas4-Cas1 complex is mutually exclusive with formation of the Cas1-Cas2 complex, which may suggest distinct roles of both subcomplexes. We found that the Cas4-Cas1 subcomplex displays prespacer cleavage activity only on PAM-containing 3' overhangs. Cas4-Cas1 removes the PAM via endonuclease cleavage while Cas1-Cas2 defines overhang trimming likely through host factor nucleases. Subsequently, Cas1-Cas2 initiates coupling of the non-PAM overhang to the leader-repeat junction followed by integration of the processed PAM-site at the repeat-spacer junction.

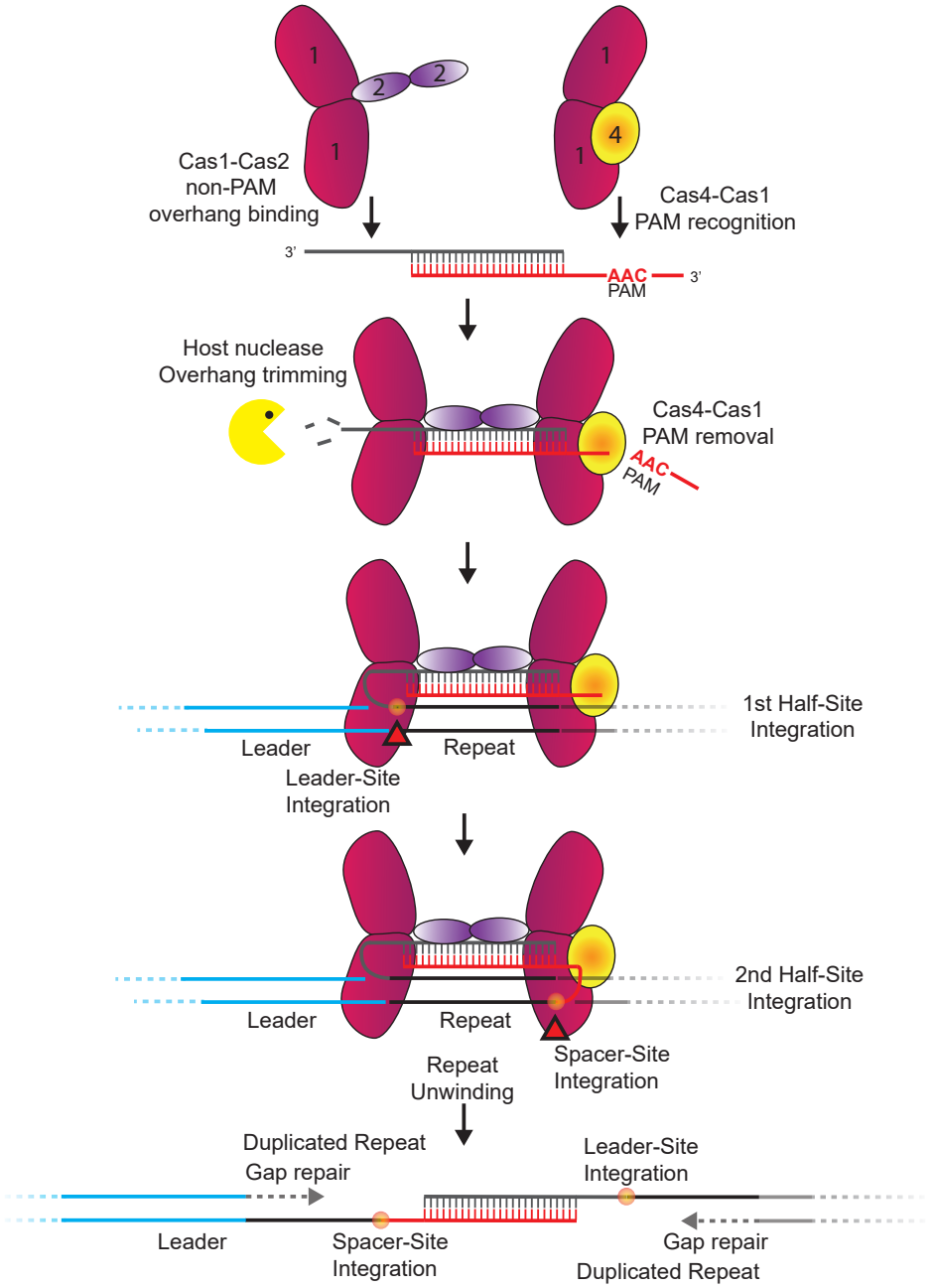


Figure 4.5 – Model of Cas4-Cas1 and Cas1-Cas2 assisted spacer selection, processing and integration. Prespacers with long PAM- and non-PAM containing 3' overhangs are bound by Cas4-Cas1 (PAM overhang) and Cas1-Cas2 (non-PAM overhang). Following processing by host-factor nucleases, the non-PAM site of the spacer is integrated at the leader-repeat site (first half-site integration). Subsequently, the second half-site integration (spacer-site integration) of the PAM-site occurs,

likely orchestrated by release of the Cas4-processed overhang into the integrase site of Cas1. Unwinding of the repeat followed by gap-repair completes repeat duplication and full-site spacer integration.

Our findings are consistent with previous studies that established the existence of two mutually exclusive Cas4-Cas1 and Cas1-Cas2 complexes in type I-C CRISPR-Cas systems [21, 22], and expand our understanding of the roles of these subcomplexes. Moreover, the RecB-domain mediated activity of Cas4 is dependent on the presence of Cas1, since Cas4 alone is not able to recognize and process the PAM sequence. This observation suggests that the Cas4-Cas1 interaction is essential for sequence specific recognition of the PAM. It remains to be determined whether the PAM sequence recognition domain is located within Cas4 or Cas1. Interestingly, PAM selection in the type I-E system is mediated by the C-terminal tail of Cas1 [27], however, this C-terminal proportion is not conserved in the type I-D Cas1 protein. Future structural and biochemical studies will have to address how PAM selection is achieved.

Cas4-Cas1 did not display activity on non-PAM containing overhangs *in vitro*, however, processing of the non-PAM overhang was observed in our *in vivo* setup, suggesting that processing involves other non-Cas proteins. This finding is in line with the Cas4-deficient type I-E system in which host factors such as the ExoT and DnaQ-like exonucleases are required for processing both, PAM-containing and non-PAM overhangs [27, 28]. We propose that, in analogy to the *E. coli* type I-E system, 3'-5' exonucleases act as trimming factors for non-PAM 3' overhangs in the native *Synechocystis* sp. 6803 host. The Cas1 protein of the type I-E system recognizes and protects the PAM from premature trimming, causing a delayed processing of the PAM end that ensures correct orientation [27]. Upon activation, the Cas4-Cas1 complex sequence specifically removes the type I-D PAM, although incorrect processing was observed *in vivo* and *in vitro* that would result in single-nucleotide slipped spacers. Recently, it was observed in the type I-F system that slipped spacers increase primed adaptation which enhances the spacer diversity of the population [34]. Our results suggest the possibility that such erroneous PAM processing could promote the integration of slipped spacers and by extension, primed adaptation as found in other type I CRISPR-Cas systems [1].

Cryo-EM structures of the type I-C Cas4-Cas1-Cas2 complex revealed that the complex might undergo a conformational change (e.g. causing dissociation of Cas4 from the complex) in order to allow for Cas1 mediated integration of the PAM-end site of the spacer. Lee et al. showed that 50% of their Cas4-Cas1-Cas2 complex structures lacked the Cas4 density on one site of the complex, resulting in an asymmetric complex [22]. Our observation of two integrase complexes (Cas4-Cas1 and Cas1-Cas2) that are independently capable of at least half-site spacer integration points towards a similar asymmetrical organization of the full adaptation complex, in which Cas4-Cas1 is involved in PAM-site and Cas1-Cas2 in non-PAM site integration. By testing asymmetric spacer precursors, we demonstrate that integrase activity of the type I-D Cas4-Cas1 complex is potentially halted until integration of the non-PAM overhang has occurred. We propose a model for the type I-D system that relies on a delayed PAM-site integration by Cas4 in order to result in a correctly oriented spacer. In summary, we propose a mechanism in which two functionally independent complexes, Cas4-Cas1 and Cas1-Cas2, sequentially process and integrate prespacer substrates. This mechanism ensures correct spacer orientation as well as correct PAM-processing, thereby resulting in interference-proficient CRISPR adaptation.

REFERENCES

1. Nussenzweig, P.M. and L.A. Marraffini, Molecular Mechanisms of CRISPR-Cas Immunity in Bacteria. Annu Rev Genet, 2020. 54: p. 93-120. DOI: 10.1146/annurev-genet-022120-112523.

2. Barrangou, R., CRISPR-Cas systems and RNA-guided interference. WIREs RNA, 2013. 4(3): p. 267-278. DOI: 10.1002/wrna.1159.

3. van der Oost, J., E.R. Westra, R.N. Jackson, et al., Unravelling the structural and mechanistic basis of CRISPR–Cas systems. Nature Reviews Microbiology, 2014. 12(7): p. 479-492. DOI: 10.1038/nrmicro3279.

4. Brouns, S.J.J., M.M. Jore, M. Lundgren, et al., Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. Science, 2008. 321(5891): p. 960-964. DOI: 10.1126/science.1159689.

5. Jackson, S.A., R.E. McKenzie, R.D. Fagerlund, et al., CRISPR-Cas: Adapting to change. Science, 2017. 356(6333). DOI: 10.1126/science.aal5056.

6. McGinn, J. and L.A. Marraffini, Molecular mechanisms of CRISPR–Cas spacer acquisition. Nature Reviews Microbiology, 2019. 17(1): p. 7-12. DOI: 10.1038/s41579-018-0071-7.

7. Nuñez, J.K., P.J. Kranzusch, J. Noeske, et al., Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. Nature Structural & Molecular Biology, 2014. 21(6): p. 528-534. DOI: 10.1038/nsmb.2820.

8. Nuñez, J.K., L.B. Harrington, P.J. Kranzusch, et al., Foreign DNA capture during CRISPR–Cas adaptive immunity. Nature, 2015. 527(7579): p. 535-538. DOI: 10.1038/nature15760.

9. Yosef, I., M.G. Goren, and U. Qimron, Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. Nucleic Acids Research, 2012. 40(12): p. 5569-5576. DOI: 10.1093/nar/gks216.

10. Hou, Z. and Y. Zhang, Insights into a Mysterious CRISPR Adaptation Factor, Cas4. Mol Cell, 2018. 70(5): p. 757-758. DOI: 10.1016/j.molcel.2018.05.028.

11. Jansen, R., J.D.A.V. Embden, W. Gaastra, et al., Identification of genes that are associated with DNA repeats in prokaryotes. Molecular Microbiology, 2002. 43(6): p. 1565-1575. DOI: 10.1046/j.1365-2958.2002.02839.x.

12. Hudaiberdiev, S., S. Shmakov, Y.I. Wolf, et al., Phylogenomics of Cas4 family nucleases. BMC Evolutionary Biology, 2017. 17(1): p. 232-232. DOI: 10.1186/s12862-017-1081-1.

13. Liu, T., Z. Liu, Q. Ye, et al., Coupling transcriptional activation of CRISPR–Cas system and DNA repair genes by Csa3a in Sulfolobus islandicus. Nucleic Acids Research, 2017. 45(15): p. 8978-8992. DOI: 10.1093/nar/gkx612.

14. Li, M., R. Wang, D. Zhao, et al., Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process. Nucleic Acids Research, 2014. 42(4): p. 2483-2492. DOI: 10.1093/nar/gkt1154.

15. Plagens, A., B. Tjaden, A. Hagemann, et al., Characterization of the CRISPR/Cas Subtype I-A System of the Hyperthermophilic Crenarchaeon Thermoproteus tenax. Journal of Bacteriology, 2012. 194(10): p. 2491-2500. DOI: 10.1128/JB.00206-12.

16. Kieper, S.N., C. Almendros, J. Behler, et al., Cas4 Facilitates PAM-Compatible Spacer Selection during CRISPR Adaptation. Cell Reports, 2018. 22(13): p. 3377-3384. DOI: 10.1016/j.celrep.2018.02.103.

17. Shiimori, M., S.C. Garrett, B.R. Graveley, et al., Cas4 Nucleases Define the PAM, Length, and Orientation of DNA Fragments Integrated at CRISPR Loci. Molecular cell, 2018. 70(5): p. 814-824.e6. DOI: 10.1016/j.molcel.2018.05.002.

18. Lemak, S., N. Beloglazova, B. Nocek, et al., Toroidal Structure and DNA Cleavage by

the CRISPR-Associated [4Fe-4S] Cluster Containing Cas4 Nuclease SSO0001 from *Sulfolobus solfataricus*. Journal of the American Chemical Society, 2013. 135(46): p. 17476-17487. DOI: 10.1021/ja408729b.

19. Lemak, S., B. Nocek, N. Beloglazova, et al., The CRISPR-associated Cas4 protein Pcal_0546 from Pyrobaculum calidifontis contains a [2Fe-2S] cluster: crystal structure and nuclease activity. Nucleic Acids Research, 2014. 42(17): p. 11144-11155. DOI: 10.1093/nar/gku797.

20. Zhang, J., T. Kasciukovic, and M.F. White, The CRISPR Associated Protein Cas4 Is a 5′ to 3′ DNA Exonuclease with an Iron-Sulfur Cluster. PLoS ONE, 2012. 7(10): p. e47232-e47232. DOI: 10.1371/journal.pone.0047232.

21. Lee, H., Y. Zhou, D.W. Taylor, et al., Cas4-Dependent Prespacer Processing Ensures High-Fidelity Programming of CRISPR Arrays. Molecular Cell, 2018. 70(1): p. 48-59.e5. DOI: 10.1016/j.molcel.2018.03.003.

22. Lee, H., Y. Dhingra, and D.G. Sashital, The Cas4-Cas1-Cas2 complex mediates precise prespacer processing during CRISPR adaptation. eLife, 2019. 8(3): p. 1-84. DOI: 10.7554/eLife.44248.

23. Gonzales, M.F., T. Brooks, S.U. Pukatzki, et al., Rapid Protocol for Preparation of Electrocompetent *Escherichia coli* and Vibrio cholerae. Journal of Visualized Experiments : JoVE, 2013(80): p. 50684. DOI: 10.3791/50684.

24. Tahallah, N., M. Pinkse, C.S. Maier, et al., The effect of the source pressure on the abundance of ions of noncovalent protein assemblies in an electrospray ionization orthogonal time-of-flight instrument. Rapid Commun Mass Spectrom, 2001. 15(8): p. 596-601. DOI: 10.1002/rcm.275.

25. van den Heuvel, R.H., E. van Duijn, H. Mazon, et al., Improving the performance of a quadrupole time-of-flight instrument for macromolecular mass spectrometry. Anal Chem, 2006. 78(21): p. 7473-83. DOI: 10.1021/ac061039a.

26. Joo, C. and T. Ha, Labeling DNA (or RNA) for single-molecule FRET. Cold Spring Harb Protoc, 2012. 2012(9): p. 1005-8. DOI: 10.1101/pdb.prot071027.

27. Kim, S., L. Loeff, S. Colombo, et al., Selective loading and processing of prespacers for precise CRISPR adaptation. Nature, 2020. DOI: 10.1038/s41586-020-2018-1.

28. Ramachandran, A., L. Summerville, B.A. Learn, et al., Processing and integration of functionally oriented prespacers in the *Escherichia coli* CRISPR system depends on bacterial host exonucleases. The Journal of biological chemistry, 2020. 295(11): p. 3403-3414. DOI: 10.1074/jbc.RA119.012196.

29. Arslan, Z., V. Hermanns, R. Wurm, et al., Detection and characterization of spacer integration intermediates in type I-E CRISPR–Cas system. Nucleic Acids Research, 2014. 42(12): p. 7884-7893. DOI: 10.1093/nar/gku510.

30. Nuñez, J.K., A.S.Y. Lee, A. Engelman, et al., Integrase-mediated spacer acquisition during CRISPR–Cas adaptive immunity. Nature, 2015. 519(7542): p. 193-198. DOI: 10.1038/nature14237.

31. Rollie, C., S. Schneider, A.S. Brinkmann, et al., Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition. eLife, 2015. 4: p. e08716. DOI: 10.7554/eLife.08716.

32. Fagerlund, R.D., M.E. Wilkinson, O. Klykov, et al., Spacer capture and integration by a type I-F Cas1-Cas2-3 CRISPR adaptation complex. Proceedings of the National Academy of Sciences of the United States of America, 2017. 114(26): p. E5122-E5128. DOI: 10.1073/pnas.1618421114.

33. Zhang, Z., S. Pan, T. Liu, et al., Cas4 Nucleases Can Effect Specific Integration of CRISPR Spacers. Journal of Bacteriology, 2019. 201(12). DOI: 10.1128/JB.00747-18.

34. Jackson, S.A., N. Birkholz, L.M. Malone, et al., Imprecise Spacer Acquisition Generates CRISPR-Cas Immune Diversity through Primed Adaptation. Cell Host and Microbe, 2019. 25(2): p. 250-260.e4. DOI: 10.1016/j.chom.2018.12.014.

SUPPLEMENTARY

SUPPLEMENTARY FIGURES

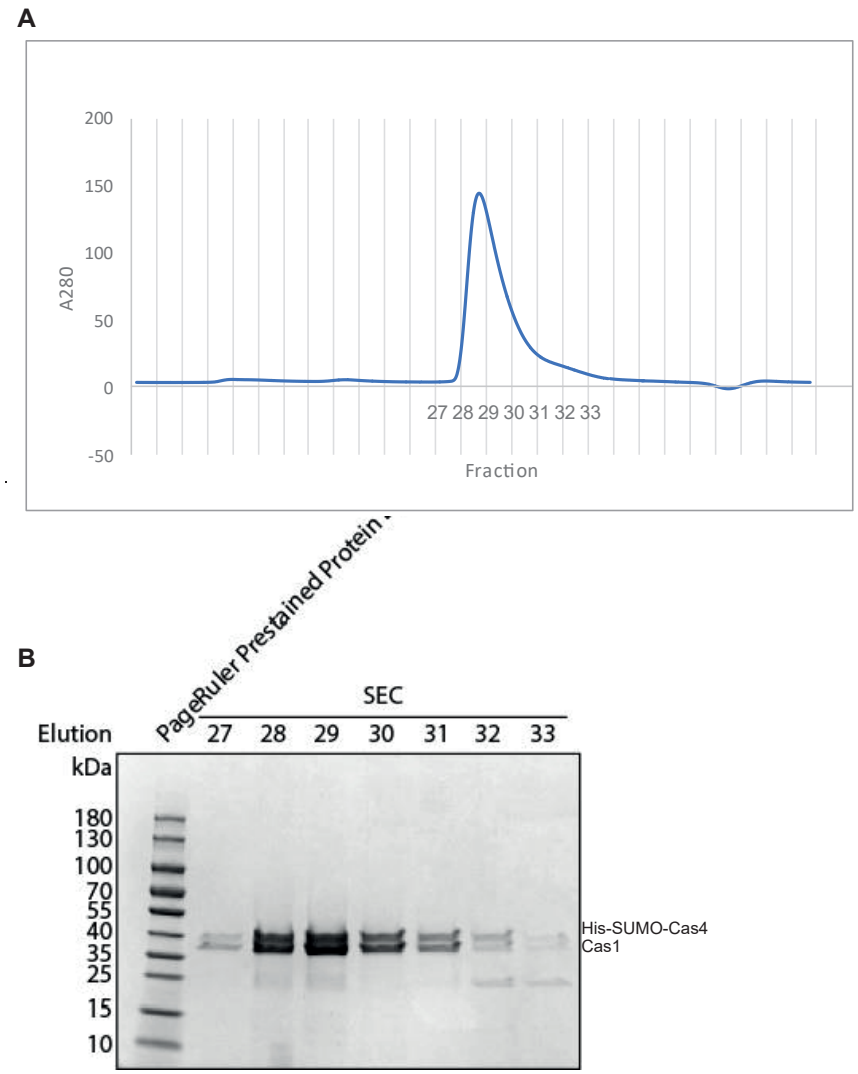


Figure 4.S1 - Related to Fig 2B - Co-purification of His-SUMO-Cas4 and untagged Cas1. A Size exclusion chromatogram of His6-SUMO-Cas4 and untagged Cas1 B SDS PAGE analysis of SEC purification.

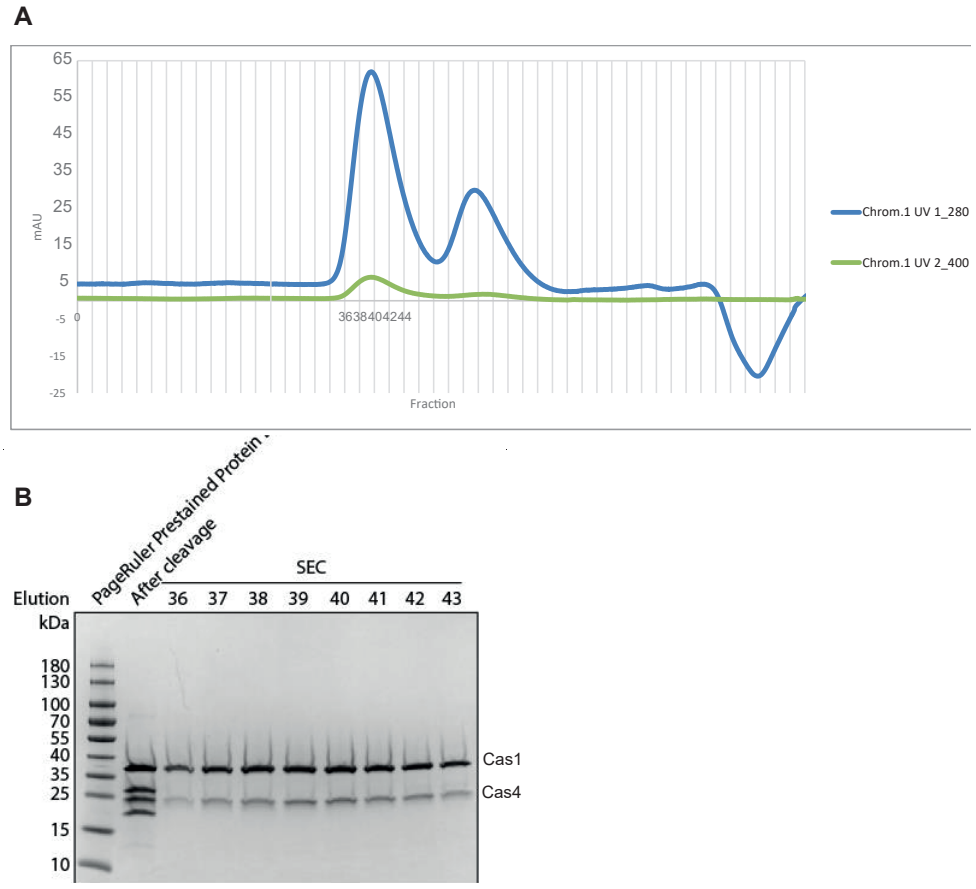


Figure 4.S2 - Related to Fig 2D - Co-purification of untagged Cas4 and untagged Cas1 after TEV protease cleavage of His-SUMO-TEV tag. **A** Size exclusion chromatogram of Cas4 and Cas1 (A280 - total protein; A400 - Cas4 FeS-Cluster) **B** SDS PAGE analysis of SEC purification.

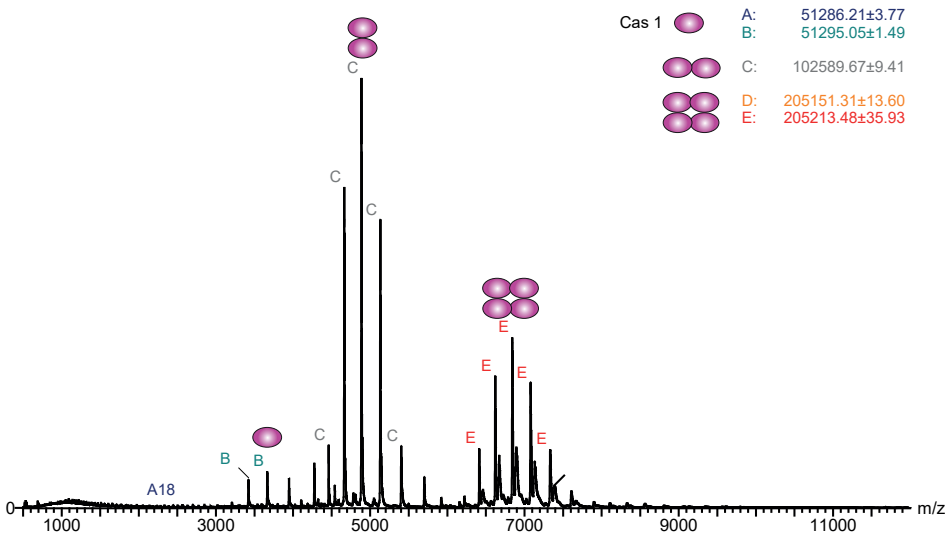


Figure 4.S3 - Related to Fig 2 - Native Mass Spectrometry of Cas1-Cas2 complex as shown in 2A.

Table 4.S1 - Plasmids used in this study

Name in this study	Name	Insert	Vector	Resistance	Source
pCas2	pTU084	*Synechocystis* sp. 6803 Type I-D *cas2* (deltaCas1)	pET-T7	Amp	Kieper et al. (2018)
pCas1	pTU085	*Synechocystis* sp. 6803 Type I-D *cas1* (deltaCas2)	pET-T7	Amp	Kieper et al. (2018)
pCas1	pTU092	*Synechocystis* sp. 6803 Type I-D *cas1*	pET-T7	Spec	This study
pCas4^{D76A}	pTU086	*Synechocystis* sp. 6803 Type I-D *cas4* (D76A)	pET-T7	Spec	Kieper et al. (2018)
pCas4$^{D76A+K91A}$	pTU411	*Synechocystis* sp. 6803 Type I-D *cas4*(D76A+K91A)	pET-T7	Spec	This study
pCas4	pTU130	*Synechocystis* sp. 6803 Type I-D *cas4*	pET-T7	Spec	Kieper et al. (2018)
pCRISPR	pTU134	*Synechocystis* sp. 6803 Type I-D Leader-R-S1	pACYCDuet1	Cm	Kieper et al. (2018)
pCas1-2	pTU70	*Synechocystis* sp. 6803 Type I-D *cas1-cas2*	pET-T7	Amp	Kieper et al. (2018)
pEmp	pTU116	NA	pET-T7	Spec	Addgene Plasmid #48329

Table S2 - Oligonucleotides used in this study

Name	Sequence	Description
BN015	CGTCCATGGGAAGTCATTCTTCAAATTTTGGC	Leader Fw
BN277	GTGGAATACGCAAAAGGC	*cas4* mutagenesis K91A Fw
BN278	AGGAATTAATAAGCCATCACTTTC	*cas4* mutagenesis K91A Rv
BN1398	GCTAGTTATTGCTCAGCGG	pCRISPR bb Rv
BN1711	GGAAGGTTTGCCAAAGTC	Leader Distal Half-Site Integration
BN1712	CTGTTCGACTTAAGCATTATGC	Leader Proximal Half-Site Integration
BN1713	ATCGACACCACCACG	OligoSpecific Primer Fw (PAM overhang)
BN1714	CGTGGTGGTGTCGAT	OligoSpecific Primer Rv (non-PAM overhang)
BN1763	CTACCATCGACACCACCACGCTGGCTTTTTAACTTTTT	25 nt duplex 13nt PAM 3’ ovhng
BN1768	GCCAGCGTGGTGGTGTCGATGGTAGTTTTTTTGTTTTT	25 nt duplex 13nt RvC PAM 3’ ovhng
BN1829	CTACCATCGACACCACCACGCTGGCTTTTTTTGTTTTT	25 nt duplex 13nt RvC PAM 3’ ovhng (5’ C6-Amino)
BN1830	GCCAGCGTGGTGGTGTCGATGGTAGTTTTTAACTTTTT	25 nt duplex 13nt PAM 3’ ovhng (5’ C6-Amino)

PAM Containing Overhang		Overhang					P	A	M	5 nucleotides overhang					No Ovhng	Total #Reads
Correct Orientation	3'	T	T	T	T	T	C	A	A	T	T	T	T	T		
Cas1-Cas2	# Reads	16	14	49	185	678	3545	20702	40292	40528	9833	2040	129	21	4	118036
	% Reads	0,01%	0,01%	0,04%	0,16%	0,57%	3,00%	17,54%	34,14%	34,34%	8,33%	1,73%	0,11%	0,02%	0,00%	100,00%
Cas1-Cas2 + Cas4 WT	# Reads	23	13	19	65	89	289	5033	17871	57166	62610	1244	129	7	0	144558
	% Reads	0,02%	0,01%	0,01%	0,04%	0,06%	0,20%	3,48%	12,36%	39,55%	43,31%	0,86%	0,09%	0,00%	0,00%	100,00%
Cas1-Cas2 + Cas4 MT	# Reads	0	0	45	188	608	3860	30918	70787	61343	18212	3341	384	40	10	189736
	% Reads	0,00%	0,00%	0,02%	0,10%	0,32%	2,03%	16,30%	37,31%	32,33%	9,60%	1,76%	0,20%	0,02%	0,01%	100,00%
Incorrect Orientation	3'	T	T	T	T	T	C	A	A	T	T	T	T	T	No Ovhng	Total #Reads
Cas1-Cas2	# Reads	12	13	21	26	2536	23397	13905	16490	8117	3039	406	122	126	74	68284
	% Reads	0,02%	0,02%	0,03%	0,04%	3,71%	34,26%	20,36%	24,15%	11,89%	4,45%	0,59%	0,18%	0,18%	0,11%	100,00%
Cas1-Cas2 + Cas4 WT	# Reads	5	30	10	18	1299	16217	17559	78643	24033	7533	3812	796	701	354	151010
	% Reads	0,00%	0,02%	0,01%	0,01%	0,86%	10,74%	11,63%	52,08%	15,91%	4,99%	2,52%	0,53%	0,46%	0,23%	100,00%
Cas1-Cas2 + Cas4 MT	# Reads	6	39	29	46	4028	29272	17079	32676	14425	4969	866	304	397	247	104383
	% Reads	0,01%	0,04%	0,03%	0,04%	3,86%	28,04%	16,36%	31,30%	13,82%	4,76%	0,83%	0,29%	0,38%	0,24%	100,00%

Table S3 - PAM-overhang processing *in vivo*

Table S4 - non-PAM-overhang processing *in vivo*

non PAM Containing Overhang							13 nucleotide non-PAM Ovrhang									
Correct Orientation	3'	T	T	T	T	T	C	A	A	T	T	T	T	T	No Ovhng	Total #Reads
Cas1-Cas2	# Reads	68	120	91	125	3994	39770	31509	41266	32478	7136	1029	280	393	406	158665
	% Reads	0,04%	0,08%	0,06%	0,08%	2,52%	25,07%	19,86%	26,01%	20,47%	4,50%	0,65%	0,18%	0,25%	0,26%	100,00%
Cas1-Cas2 + Cas4 WT	# Reads	107	132	102	118	3183	26914	20742	30875	15563	2592	375	96	153	108	101060
	% Reads	0,11%	0,13%	0,10%	0,12%	3,15%	26,63%	20,52%	30,55%	15,40%	2,56%	0,37%	0,09%	0,15%	0,11%	100,00%
Cas1-Cas2 + Cas4 MT	# Reads	153	300	297	480	9193	41659	27013	36170	23057	4392	441	202	296	302	143955
	% Reads	0,11%	0,21%	0,21%	0,33%	6,39%	28,94%	18,76%	25,13%	16,02%	3,05%	0,31%	0,14%	0,21%	0,21%	100,00%
Incorrect Orientation	3'														No Ovhng	Total #Reads
Cas1-Cas2	# Reads	19	21	36	211	804	2694	14834	29823	22610	6961	2847	205	46	31	81142
	% Reads	0,02%	0,03%	0,04%	0,26%	0,99%	3,32%	18,28%	36,75%	27,86%	8,58%	3,51%	0,25%	0,06%	0,04%	100,00%
Cas1-Cas2 + Cas4 WT	# Reads	29	16	38	140	468	973	15462	39326	40434	20859	5443	419	65	35	123707
	% Reads	0,02%	0,01%	0,03%	0,11%	0,38%	0,79%	12,50%	31,79%	32,69%	16,86%	4,40%	0,34%	0,05%	0,03%	100,00%
Cas1-Cas2 + Cas4 MT	# Reads	10	10	15	74	261	918	7748	18843	14596	5152	1735	138	26	12	49538
	% Reads	0,02%	0,02%	0,03%	0,15%	0,53%	1,85%	15,64%	38,04%	29,46%	10,40%	3,50%	0,28%	0,05%	0,02%	100,00%

Table S5 - (non)PAM-overhang processing *in vitro*

		Overhang					P	A	M	5 nucleotides overhang						
PAM Containing Overhang	3’	T	T	T	T	T	C	A	A	T	T	T	T	T	No Ovhng	Total #Reads
Correct Orientation	# Reads	0	0	0	0	0	0	7	32	74954	29599	9587	1440	273	89	115981
	% Reads	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,01%	0,03%	64,63%	25,52%	8,27%	1,24%	0,24%	0,08%	100,00%
Incorrect Orientation	# Reads	2038	41	0	0	6	4	43	18	5	0	0	0	0	0	2155
	% Reads	94,57%	1,90%	0,00%	0,00%	0,28%	0,19%	2,00%	0,84%	0,23%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%

		13 nucleotide non-PAM Ovrhang														
non PAM Containing Overhang	3’	T	T	T	T	T	G	T	T	T	T	T	T	T	No Ovhng	Total #Reads
Correct Orientation	# Reads	60062	574	19	5	37	17	17	80	17	1	7	0	0	0	60836
	% Reads	98,73%	0,94%	0,03%	0,01%	0,06%	0,03%	0,03%	0,13%	0,03%	0,00%	0,01%	0,00%	0,00%	0,00%	100,00%
Incorrect Orientation	# Reads	47434	1114	68	75	31	92	98	301	375	11	1	1	0	0	49601
	% Reads	95,63%	2,25%	0,14%	0,15%	0,06%	0,19%	0,20%	0,61%	0,76%	0,02%	0,00%	0,00%	0,00%	0,00%	100,00%

4

5

Conserved motifs in the CRISPR leader sequence control spacer acquisition levels in Type I-D CRISPR-Cas systems

FEMS Microbiology Letters
2019 Jun; 366(11): fnz129.

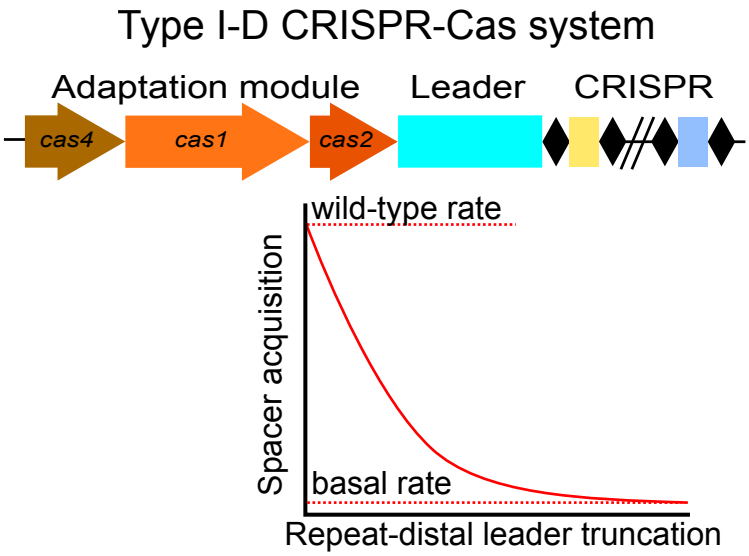
Sebastian N. Kieper[1], Cristóbal Almendros[1], Stan J.J. Brouns[1,2]

1. Kavli Institute of Nanoscience, Department of Bionanoscience, Delft University of Technology, Van der Maasweg 9, 2629 HZ Delft, The Netherlands.

2. Laboratory of Microbiology, Wageningen University, Stippeneng 4, 6708 WE Wageningen, The Netherlands.

5.1 Abstract

Integrating short DNA fragments at the correct leader-repeat junction is key to successful CRISPR-Cas memory formation. The Cas1-2 proteins are responsible to carry out this process. However, the CRISPR adaptation process additionally requires a DNA element adjacent to the CRISPR array, called leader, to facilitate efficient localization of the correct integration site. In this work, we introduced the core CRISPR adaptation genes *cas1* and *cas2* from the Type I-D CRISPR-Cas system of *Synechocystis* sp. 6803 into *Escherichia coli* and assessed spacer integration efficiency. Truncation of the leader resulted in a significant reduction of spacer acquisition levels and revealed the importance of different conserved regions for CRISPR adaptation rates. We found three conserved sequence motifs in the leader of I-D CRISPR arrays that each affected spacer acquisition rates, including an integration anchoring site. Our findings support the model in which the leader sequence is an integral part of type I-D adaptation in *Synechocystis* sp. acting as a localization signal for the adaptation complex to drive CRISPR adaptation at the first repeat of the CRISPR array.



5.2 INTRODUCTION

Mobile genetic elements (MGEs) such as bacteriophages and conjugative plasmids exert an evolutionary pressure on prokaryotes, demanding bacterial and archaeal cells to frequently update their immunological lines of defense. Prokaryotes evolved an adaptive immune system that relies on the use of clustered regularly interspaced short palindromic repeats (CRISPRs) and their associated proteins (Cas) in order to specifically recognize and destroy predatory elements. Target recognition is mediated by the synthesis of small RNAs (i.e. crRNA), derived from CRISPR arrays, that guide Cas nuclease complexes towards the invading MGE [1-4]. The adaptive immune response is created in a step termed CRISPR adaptation in which short MGE-derived sequences are inserted between the repeats giving rise to new "spacers" [5-7]. Spacer acquisition is carried out by the adaptation proteins Cas1 and Cas2, which are universally encoded in the vast majority of all types and subtypes of the two major classes of CRISPR-Cas systems [8, 9]. However, beyond *cas1* and *cas2*, the region adjacent to the CRISPR array (an A-T rich sequence termed leader [10]) as well as the repeat sequence itself are required to guide the integration event towards the correct location [9, 11]. The leader sequence contains the promoter necessary to drive transcription of the CRISPR, but importantly also encodes sequences that are recognized by the Cas1-2 complex and other cellular factors. This includes the Integration Host Factor (IHF) which determines the appropriate integration site at the leader-repeat junction in I-E CRISPR-Cas systems [12]. Localizing the correct integration site is a prerequisite for functional interference and helps to increase the immune diversity which limits the emergence of escape phage mutants [13]. Leader encoded adaptation signals likely co-evolved with their cognate adaptation proteins in order to support spacer acquisition rates that aid in establishing an efficient immune response while at the same time limiting the potential costs connected to high acquisition rates (e.g. autoimmunity) [14, 15]. In the type I-E system, those adaptation signals are found in the sequence 60 bp upstream of the first repeat that ensure efficient spacer integration [9], while the type I-A system requires at least 400 bp of the leader for detectable levels of acquisition [16]. The Cas1-2 complex of the type II-A system relies on intrinsic specificity for a short leader-anchoring site adjacent to the first repeat as well as the repeat itself which both are required and sufficient for catalysis of leader proximal spacer in-

tegration [17-20]. This large variation in leader length, sequence conservation and host factor requirements is exemplary for the broad diversity of CRISPR-Cas systems and provides insights in how different adaptation modules are optimized towards their respective CRISPR array. Here, we focus on the spacer acquisition rates of a cyanobacterial type I-D CRISPR-Cas system and find that the presence of several conserved sequences in the CRISPR leader enhances the efficiency of spacer integration. By employing sensitive *in vivo* spacer acquisition assays in a heterologous *E. coli* host we demonstrate that spacers can be acquired even in the complete absence of the leader. However, efficient spacer uptake requires the conserved 5' region of the leader. Our results underline the importance of the leader sequence as a non-protein factor that controls the levels of CRISPR adaptation, and suggest interaction of the leader sequence with the Cas1-2 adaptation machinery itself.

5.3 Material & Methods

Bacterial strains and growth conditions

E. coli DH5α and BW25113 strains were grown in Lysogeny Broth (LB) at 37˚C and continuous shaking at 180 rpm or grown on LB agar plates (LBA) containing 1.5% (wt/vol) agar. When required, the media were supplemented with 100 μg ml^{-1} ampicillin and 25 μg ml^{-1} chloramphenicol (see Table 5.S1 for plasmids and corresponding selection markers).

5.3.1 Plasmid construction and transformation

Plasmids used in this study are listed in Table 5.S1. All cloning steps were performed in *E. coli* DH5α. Primers described in Table 5.S2 were used for PCR amplification of the type I-D CRISPR locus (leader-repeat-spacer1) from *Synechocystis* sp. 6803 cell material using the Q5 high-fidelity Polymerase (New England Biolabs). PCR amplicons were subsequently cloned into the pACYCDuet-1 vector system (Novagen (EMD Millipore) using restriction-ligation cloning. The pCRISPR leader mutants were obtained by PCR-based mutagenesis using primers listed in Table 5.S2. All plasmids were verified by Sanger-sequencing (Macrogen Europe, Amsterdam, The Netherlands). Bacterial transformations were either carried out by electroporation (200 Ω, 25 μF,

2.5 kV) using a ECM 630 electroporator (BTX Harvard Apparatus) or using chemically competent cells prepared according to manufacturer's manual (Mix&Go, Zymo research). Electrocompetent cells were prepared following a protocol adapted from [21]. Transformants were selected on LBA supplemented with appropriate antibiotics.

5.3.2 *IN VIVO* SPACER ACQUISITION ASSAY

E. coli BW25113 was transformed with pCas1-2 and pCRISPR with varying lengths of the leader sequence (Table 5.S1). Cultures were inoculated from single colonies and passaged once after 24 hours of growth at 37˚C and continuous shaking at 180 rpm. 200 µL of cells cultured for 48 hours were harvested by centrifugation and resuspended in 50 µL of MilliQ water. Subsequently, 2 µL of cell suspension was subjected to spacer detection PCR using a forward primer annealing in the 3' end of the CRISPR repeat of pCRISPR but mismatching the first nucleotide of spacer 1 (degenerated primer mix, BN143+BN144+BN145) [22] and a reverse primer annealing in the vector backbone (BN172) (Table 5.S2). PCR products were separated on 2% agarose gels and were densitometrically quantified using ImageLab 4.0 (BioRad). Statistical analysis was done using GraphPad Prism 4 to perform one-way ANOVA followed by Dunnett's multiple comparison test. When a higher sensitivity was required, amplicons of expanded pCRISPR arrays were BluePippin (SageScience) size selected and subjected to a second PCR reaction as described previously [23, 24].

5.3.3 SEQUENCING OF ACQUIRED SPACERS

BluePippin extracted and re-amplified expanded CRISPR array amplicons were cloned in the pGemT-easy vector (Promega) and Sanger sequenced (Macrogen Europe, Amsterdam, The Netherlands). Using the Geneious 9.0.5 motif search function, the type I-D repeats were annotated in the sequencing reads and the newly acquired spacers extracted. The origin of newly acquired spacers was determined by nucleotide BLAST search against pCas1-2, pCRISPR and the *E. coli* BW25113 genome.

5.4 Results

5.4.1 The leader displays a high degree of conservation

The Cas1-2 adaptation complex is the central element mediating adaptation in almost all CRISPR-Cas systems. It has been proposed that the Cas1 protein co-evolves with its cognate leader as well as the repeat sequence, hence we hypothesized that type I-D Cas1 proteins would recognize conserved motifs within their cognate leader sequences [15, 25]. First, the Cas1 protein of the CRISPR-Cas type I-D system of *Synechocystis* sp. 6803 was used in a BLASTP-search to identify related Cas1 proteins in a variety of different species. Interestingly, most Cas1 proteins that were found were derived from cyanobacterial type I-D systems (Fig. 5.1). Next, we retrieved the leader sequences (defined as the A-T rich adjacent upstream sequence of the CRISPR array [26]) from type I-D systems containing a Cas1 ortholog with at least 60% sequence identity. Below this conservation threshold value we noticed that Cas1 orthologs were more divergent (sequence identity < 40%), and were excluded from the analysis. The 25 selected I-D leader sequences ranged from 202 to 220 bp which represents considerably longer leaders than described for the *E. coli* type I-E system which are typically shorter than 100 bp [9]. We then performed MAFFT alignment of the leaders [27, 28] and identified 3 regions with more than 4 consecutive nucleotides that were highly conserved across all the 25 leader sequences (Fig. 5.1; motifs I-II-III). Interestingly, we found a high degree of conservation at the repeat distal end (II+III) of the leader, while the repeat proximal region displayed more variability with only one conserved motif (I). Altogether, the high conservation of those motifs in the leader sequence suggests that those regions are important for the correct localization of the leader of the CRISPR array, and could serve as recognition signals for the Cas1-2 adaptation complex or host factors to ensure spacer integration at the leader-repeat junction.

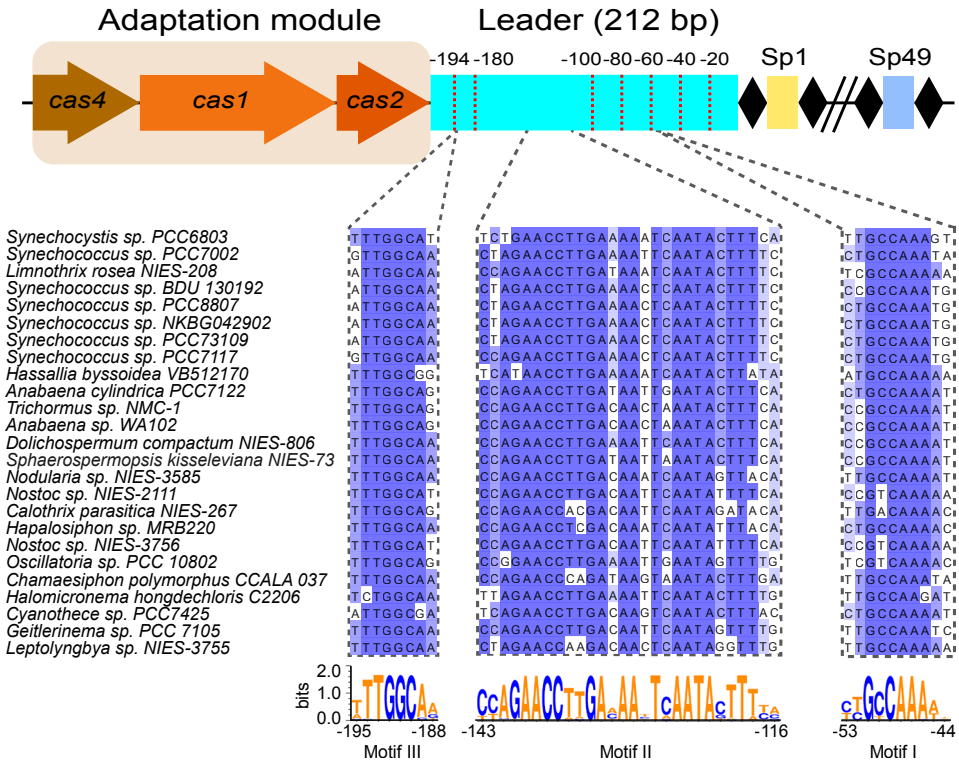


Figure 5.1 – Type I-D arrangement of the adaptation module and the CRISPR array. Downstream of *cas2* is the 212 bp leader sequence. Conserved regions obtained from MAFFT alignments of 25 leaders reveal conserved motifs predominantly at the repeat distal end with increasing sequence variability at the repeat proximal end. Sequence conservation is summarized in Weblogo3 depictions [29]. Leader truncations from the repeat-distal end for experimental investigation of conserved motifs are indicated with red dashed lines.

5.4.2 Leader motifs stimulate spacer acquisition

To get experimental insight into the previously identified conserved regions, we systematically shortened the leader from the repeat-distal end while leaving the repeat-proximal leader intact. The different CRISPR leader-repeat-spacer1 plasmids were transformed into *E. coli* K12 cells containing only Cas1 and Cas2. The *cas4* gene was omitted because we showed previously that the Cas1-2 adaptation proteins are necessary and sufficient to mediate the acquisition of new spacers [23]. After 48 hours of growth, spacer acquisition was assessed by a degenerate primer PCR [24] and acquisition efficiency was quantified from three independent assays based on the relative difference between the band intensity of the expanded CRISPR amplicon compared to the non-expanded CRISPR array (Xue et al., 2015) (Fig. 5.2A, Fig. 5.S2). We observed decreasing adaptation efficiencies depending on the presence or absence of the repeat-distal motifs (Fig. 5.2A). The highest rate of spacer acquisition was obtained with at least 194 bp of the full-leader sequence (212 and 194 constructs) containing conserved motifs II and III. However, further repeat distal truncations of the leader led to significantly impaired spacer uptake (Fig. 5.2A). Expansion of the CRISPR array is readily detectable with PCR up to a leader length of 60 bp (preserving only motif I) although with a relative reduction compared to leaders containing motif II and III. Spacer integration with leaders shorter than 60 bp is below the detection limit of the first PCR and can only be detected using a more sensitive second round of PCR (Fig. 5.2B) as described by McKenzie et al. (2019). With this method, we were able to detect spacer integration even in the absence of the leader. The sequence analysis of spacers that were acquired in the absence of the leader (0 Leader) revealed that the detected integration event gave rise to a single unique spacer (Fig. 5.S1). This very low spacer diversity indicates that the Cas1-2 adaptation complex is able to integrate spacers at the leader-repeat junction even in absence of the leader sequence, albeit at drastically reduced rates.

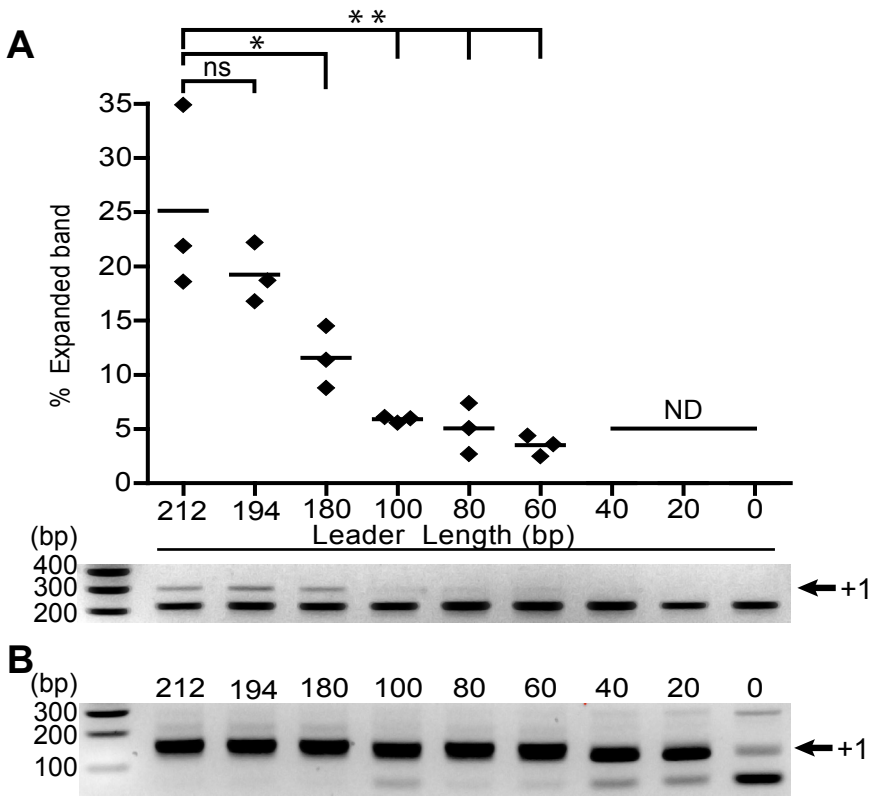


Figure 5.2 – PCR-based detection of spacer acquisition at variable leader length. A – Quantification of expanded CRISPR array band intensity (n = 3). CRISPR adaptation is negatively affected by deletion of motif III that is present in the 212 bp (wild-type) and 194 bp leader (*P < 0.05). Removal of motif II (located in the segment between 180 bp and 100 bp) and motif III significantly reduced (**P < 0.01) acquisition rates close to the detection limit of this PCR. Leaders shorter than 60 bp do not support detectable acquisition (ND). Spacer acquisition rates of the 194 bp leader are not significantly different (ns) from the full 212 bp leader. Statistical significance was calculated using Dunnett's multiple comparisons test. **B** – Second round of PCR enables the detection of spacer acquisition with leaders shorter than 60 bp or absent leader sequences.

5.5 Discussion

During phage infection the integration of novel spacers at the correct site as well as at an appropriate rate is crucial for prokaryotic survival. Recently, it was demonstrated that Cas1-2 can integrate spacers into non-CRISPR genomic regions, however, those non-canonical integration events do not lead to functional spacers that confer CRISPR resistance against sampled invaders [30]. Therefore, since only acquisitions in the CRISPR array provide the most efficient immune response, Cas1-2 must recognize the correct insertion site. Moreover, spacer integration occurs in a polarized manner at the leader proximal end of the array creating a chronological library of past infections that provides higher levels of protection from the most recently integrated spacer [17]. Specificity of the integration reaction towards the cognate CRISPR array might thus be one of the rate limiting factors for rapid and efficient immunization. Here, we demonstrated the importance of conserved leader sequences for naïve acquisition in a minimal I-D CRISPR-Cas system. The alignment of leader sequences from different type I-D systems revealed a conserved region at the repeat distal end as well as a short conserved motif approximately 50 bp upstream of the first repeat, suggesting involvement of those regions in CRISPR array recognition, potentially by the adaptation complex. By systematically truncating the leader from the repeat distal end while leaving downstream sequences intact, we disrupted those leader regions and quantified spacer integration by a semi-quantitative PCR method [31]. Strikingly, we were able to detect spacer acquisition *in vivo* even in the complete absence of the leader sequence by using a sensitive detection method. However, the efficiency of spacer integration is drastically reduced in the absence of the leader. Sequencing of the integration event revealed that only a single unique spacer was acquired. In the absence of the leader the type I-D adaptation complex displays baseline adaptation levels, but this low efficiency event only marginally contributes to protection of the population. In contrast, including at least 60 bp upstream of the I-D repeat increased acquisition rates to detectable levels, demonstrating that motif I (5'-GCCAAA-3') facilitates spacer integration. However, the maximum acquisition rate was only restored when the full leader was provided. Similar results have been obtained *in vitro* for a *Sulfolobus* type I-A CRISPR-Cas system that requires at least 400 bp of the leader for detectable acquisition and the full 531 bp leader for maximum adaptation levels [16]. Furthermore, in a type

I-A system of a related *Sulfolobus* strain a ~ 20 bp deletion within the leader sequence is associated with decreased spacer uptake [32, 33]. Our findings are consistent with the observation that deletions of particular leader sequences result in decreased acquisition rates, although future studies are needed to address whether this is caused by the loss of a specific motif, an accumulating effect of deleting several motifs or because a certain spacing between e.g. motif III and the repeat is required. In the type I-E system integration host factor (IHF) binds a conserved leader motif called IHF-binding site (IBS) and induces a 120° bend that brings another conserved motif, the 5'-TTGGT-3' Integrase Anchoring Site (IAS) in proximity to the leader-repeat junction that increases acquisition efficiency by presumably stabilizing the Cas1-2-leader-repeat interaction [12, 34]. Interestingly, motif III (5'-TTGGC-3') in the type I-D leader strongly resembles the IAS described previously. It is plausible that the type I-D Cas1-2 adaptation complex, analogous to the type I-E complex, can recognize this motif to be correctly positioned to integrate novel spacers. However, the *E. coli* IHF protein is absent from *Synechocystis* sp. 6803 suggesting that other DNA-binding host factors could be involved in recognizing the conserved region II in the type I-D leader. Overall, our work highlights the importance of the leader sequence for the adaptation stage in the type I-D CRISPR-Cas system. Through evolutionary selection of specific sequences in the leader that likely interact with the adaptation proteins, the integration of new spacers into CRISPR arrays occurs accurately at the first repeat of the CRISPR array improving the chances of prokaryotes to survive predatory invasion.

References

1. Marraffini, L.A., CRISPR-Cas immunity in prokaryotes. Nature, 2015. 526: p. 55.
2. van der Oost, J., et al., Unravelling the structural and mechanistic basis of CRISPR-Cas systems. Nat Rev Microbiol, 2014. 12(7): p. 479-92.
3. Barrangou, R., et al., CRISPR provides acquired resistance against viruses in prokaryotes. Science, 2007. 315(5819): p. 1709-12.
4. Brouns, S.J.J., et al., Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. Science, 2008. 321(5891): p. 960-964.
5. Sternberg, S.H., et al., Adaptation in CRISPR-Cas Systems. Mol Cell, 2016. 61(6): p. 797-808.
6. Jackson, S.A., et al., CRISPR-Cas: Adapting to change. Science, 2017. 356(6333).
7. Amitai, G. and R. Sorek, CRISPR-Cas adaptation: insights into the mechanism of action. Nat Rev Microbiol, 2016. 14(2): p. 67-76.
8. Koonin, E.V., K.S. Makarova, and F. Zhang, Diversity, classification and evolution of CRISPR-Cas systems. Curr Opin Microbiol, 2017. 37: p. 67-78.
9. Yosef, I., M.G. Goren, and U. Qimron, Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. Nucleic Acids Res, 2012. 40(12): p. 5569-76.
10. McGinn, J. and L.A. Marraffini, Molecular mechanisms of CRISPR–Cas spacer acquisition. Nature Reviews Microbiology, 2019. 17(1): p. 7-12.
11. Goren, M.G., et al., Repeat Size Determination by Two Molecular Rulers in the Type I-E CRISPR Array. Cell reports, 2016. 16(11): p. 2811-2818.
12. Nuñez, James K., et al., CRISPR Immunological Memory Requires a Host Factor for Specificity. Molecular Cell, 2016. 62(6): p. 824-833.
13. van Houte, S., et al., The diversity-generating benefits of a prokaryotic adaptive immune system. Nature, 2016. 532(7599): p. 385-388.
14. Bradde, S., T. Mora, and A.M. Walczak, Cost and benefits of clustered regularly interspaced short palindromic repeats spacer acquisition. Philosophical Transactions of the Royal Society B: Biological Sciences, 2019. 374(1772): p. 20180095.
15. Shah, S.A. and R.A. Garrett, CRISPR/Cas and Cmr modules, mobility and evolution of adaptive immune systems. Research in Microbiology, 2011. 162(1): p. 27-38.
16. Rollie, C., et al., Prespacer processing and specific integration in a Type I-A CRISPR system. Nucleic Acids Research, 2017: p. gkx1232-gkx1232.
17. McGinn, J. and L.A. Marraffini, CRISPR-Cas systems optimize their immune response by specifying the site of spacer integration. Molecular cell, 2016. 64(3): p. 616-623.
18. Wei, Y., et al., Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in *Streptococcus thermophilus*. Nucleic Acids Research, 2015. 43(3): p. 1749-1758.
19. Wright, A.V. and J.A. Doudna, Protecting genome integrity during CRISPR immune adaptation. Nature Structural &Amp; Molecular Biology, 2016. 23: p. 876.
20. Xiao, Y., et al., How type II CRISPR–Cas establish immunity through Cas1–Cas2-mediated spacer integration. Nature, 2017. 550: p. 137.
21. Gonzales, M.F., et al., Rapid Protocol for Preparation of Electrocompetent *Escherichia coli* and *Vibrio cholerae*. Journal of Visualized Experiments : JoVE, 2013(80): p. 50684.
22. Heler, R., et al., Cas9 specifies functional viral targets during CRISPR-Cas adaptation.

Nature, 2015. 519(7542): p. 199-202.

23. Kieper, S.N., et al., Cas4 Facilitates PAM-Compatible Spacer Selection during CRISPR Adaptation. Cell reports, 2018. 22(13): p. 3377-3384.

24. McKenzie, R.E., Almendros, C., Vink, J.N.A., Brouns, S.J.J. , Using CAPTURE to detect spacer acquisition in native CRISPR arrays. Nat Protoc, 2019.

25. Alkhnbashi, O.S., et al., Characterizing leader sequences of CRISPR loci. Bioinformatics, 2016. 32(17): p. i576-i585.

26. Jansen, R., et al., Identification of genes that are associated with DNA repeats in prokaryotes. Molecular microbiology, 2002. 43(6): p. 1565-75.

27. Katoh, K., et al., MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic acids research, 2002. 30(14): p. 3059-3066.

28. Katoh, K. and D.M. Standley, A simple method to control over-alignment in the MAFFT multiple sequence alignment program. Bioinformatics (Oxford, England), 2016. 32(13): p. 1933-1942.

29. Crooks, G.E., et al., WebLogo: a sequence logo generator. Genome research, 2004. 14(6): p. 1188-1190.

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

31. Xue, C., et al., CRISPR interference and priming varies with individual spacer sequences. Nucleic acids research, 2015. 43(22): p. 10831-10847.

32. Erdmann, S. and R.A. Garrett, Selective and hyperactive uptake of foreign DNA by adaptive immune systems of an archaeon via two distinct mechanisms. Molecular microbiology, 2012. 85(6): p. 1044-1056.

33. Garrett, R.A., et al., CRISPR-Cas Adaptive Immune Systems of the Sulfolobales: Unravelling Their Complexity and Diversity. Life (Basel, Switzerland), 2015. 5(1): p. 783-817.

34. Yoganand, K.N.R., et al., Asymmetric positioning of Cas1–2 complex and Integration Host Factor induced DNA bending guide the unidirectional homing of protospacer in CRISPR-Cas type I-E system. Nucleic Acids Research, 2017. 45(1): p. 367-381.

SUPPLEMENTARY
SUPPLEMENTARY FIGURES

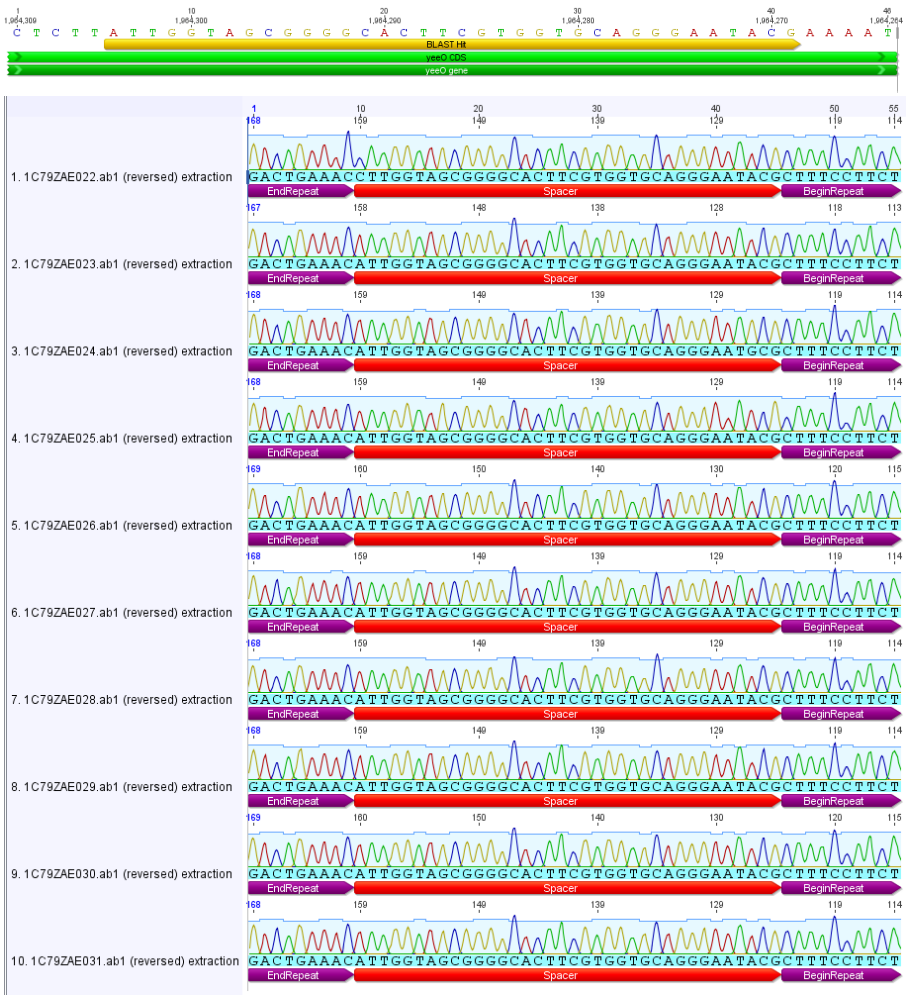


Figure 5.S1 – Extracted sequences of spacers acquired in the absence of the leader. Spacer1 contains a point mutation at position 1 (A to C). Spacers map to the *yeeO* gene of *E. coli* BW25113.

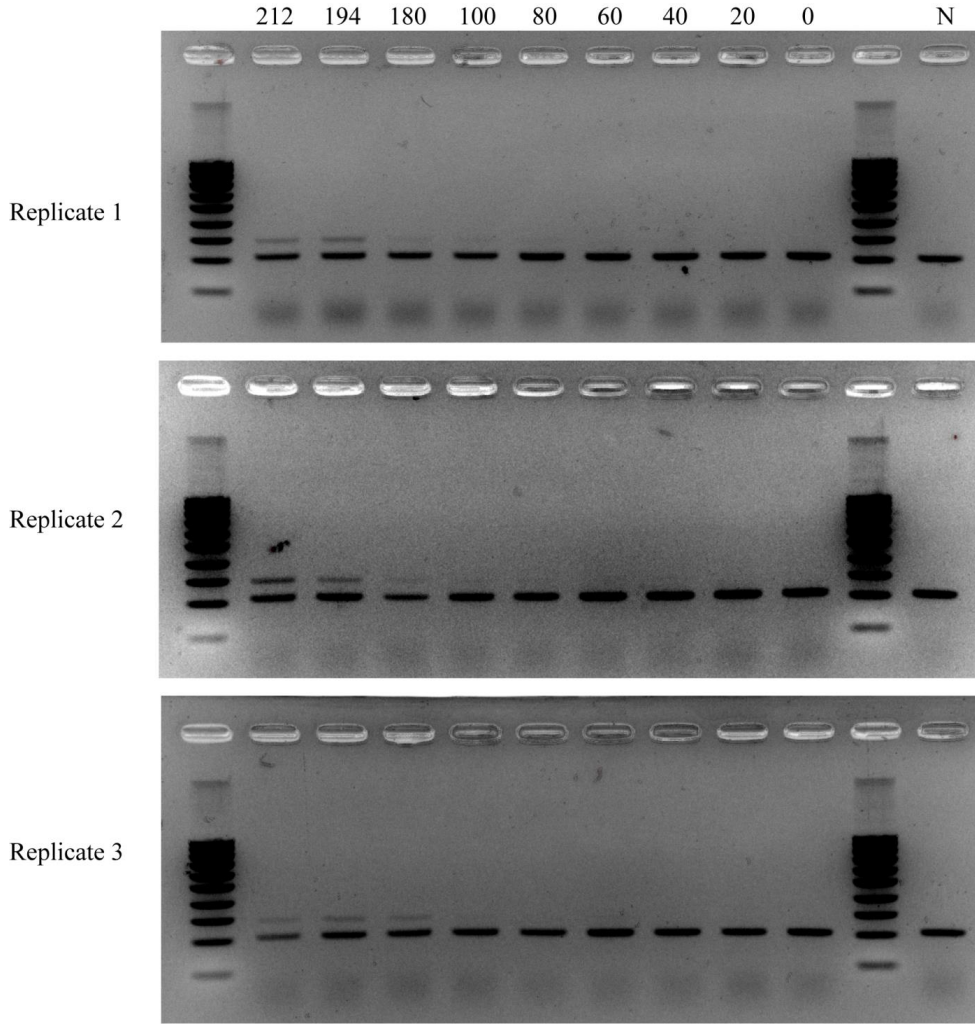


Figure 5.S2 – Three independent acquisition assays for quantification of spacer integration with respect to leader length (related to Fig 2A). Negative control (N) strain omitted pCas1-2 (empty backbone).

Supplementary Tables

Table 5.S1. Plasmids used in this study, Related to Figure 2

Name in this study	Name	Insert	Vector	Resistance	Source
pCas1-2	pTU70	*Synechocystis* sp. 6803 Type I-D *cas1-cas2*	pET-T7	Amp	Kieper et al. 2018
pCRISPR (212L)	pTU134	*Synechocystis* sp. 6803 Type I-D Leader-R-S1	pACYCDuet1	Cm	Kieper et al. 2018
pCRISPR(194L)	pTU080	*Synechocystis* sp. 6803 Type I-D 194 bp Leader-R-S1	pACYCDuet1	Cm	This study
pCRISPR(180L)	pTU081	*Synechocystis* sp. 6803 Type I-D 180 bp Leader-R-S1	pACYCDuet1	Cm	This study
pCRISPR(100L)	pTU096	*Synechocystis* sp. 6803 Type I-D 100 bp Leader-R-S1	pACYCDuet1	Cm	This study
pCRISPR(80L)	pTU095	*Synechocystis* sp. 6803 Type I-D 80 bp Leader-R-S1	pACYCDuet1	Cm	This study
pCRISPR(60L)	pTU094	*Synechocystis* sp. 6803 Type I-D 60 bp Leader-R-S1	pACYCDuet1	Cm	This study
pCRISPR(40L)	pTU093	*Synechocystis* sp. 6803 Type I-D 40 bp Leader-R-S1	pACYCDuet1	Cm	This study
pCRISPR(20L)	pTU082	*Synechocystis* sp. 6803 Type I-D 20 bp Leader-R-S1	pACYCDuet1	Cm	This study
pCRISPR(0L)	pTU083	*Synechocystis* sp. 6803 Type I-D 0 bp Leader-R-S1	pACYCDuet1	Cm	This study

Table 5.S2. Primers used in this study, Related to Figure 2

Name	Sequence 5' – 3'	Description
BN143	GCGATCGGGACTGAAACT	I-D RepeatPrimer (3' mismatches Sp1 5')
BN144	GCGATCGGGACTGAAACA	I-D RepeatPrimer (3' mismatches Sp1 5')
BN145	GCGATCGGGACTGAAACC	I-D RepeatPrimer (3' mismatches Sp1 5')
BN156	AGGCATTGAAAGCGACC	SP1 Rv
BN172	AGATCTGCCATATGTATATCTCCTTC	pACYC backbone primer
BN157	CCATGGTATATCTCCTTATTAAAG	pACYC MCS Rv for Leader deletion
BN240	TTGGCATACCTATAGCG	Type I-D pACYC - 194bp Leader-R-S1
BN216	GCGAGGGCCTTTTCC	Type I-D pACYC - 180bp Leader-R-S1
BN158	TACTATTTTGAAGGTCTGGC	Type I-D pACYC - 100bp Leader-R-S1
BG8224	TGATTTTGGAAAGATATTCTGG	Type I-D pACYC - 80bp Leader-R-S1
BN015	GGAAGGTTTGCCAAAG	Type I-D pACYC - 60bp Leader-R-S1
BN016	CTTTCCTCCACTTTCCC	Type I-D pACYC - 40bp Leader-R-S1
BN114	AAGGGGTCGGAGG	Type I-D pACYC - 20bp Leader-R-S1
BN135	CTTTCCTTCTACTAATCCCG	Type I-D pACYC - 0bp Leader-R-S1
BN136	AATAATCCCTTTAACTTTTAGGCG	Type I-D *cas4* mutagenesis D76A Rv
BN143	GCGATCGGGACTGAAACT	Degenerated Fw1

6

Cas3-derived target DNA degradation fragments fuel primed CRISPR adaptation

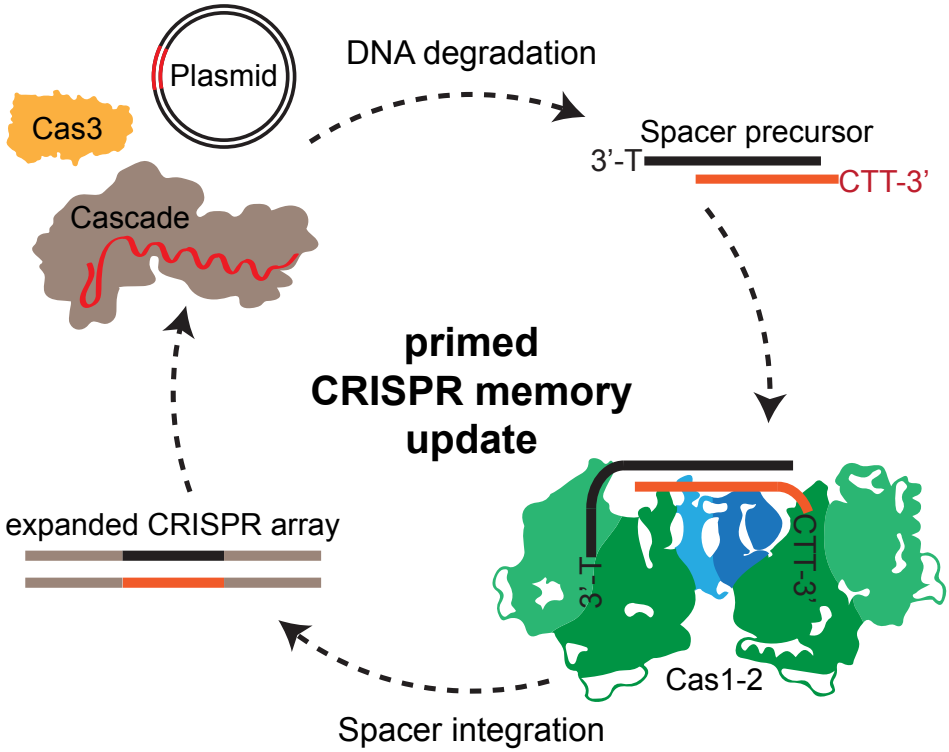
Molecular Cell
2016 Sep 1;63(5):852-64.

Tim Künne[1], Sebastian N. Kieper[1,2], Jasper W. Bannenberg[1], Anne I.M. Vogel[1], Willem R. Miellet[1] , Misha Klein[2], Martin Depken[2], Maria Suarez-Diez [3], Stan J.J. Brouns[1,2]

1. Laboratory of Microbiology, Wageningen University, 6708 WE Wageningen, the Netherlands.
2. Kavli Institute of Nanoscience and Department of BioNanoscience, Delft University of Technology, 2629 HZ, Delft, the Netherlands.
3. Laboratory of Systems and Synthetic Biology, Wageningen University, 6708 WE Wageningen, the Netherlands.

6.1 ABSTRACT

Prokaryotes use a mechanism called priming to update their CRISPR immunological memory to rapidly counter revisiting, mutated viruses and plasmids. Here we have determined how new spacers are produced and selected for integration into the CRISPR array during priming. We show that Cas3 couples CRISPR interference to adaptation by producing DNA breakdown products that fuel the spacer integration process in a two-step, PAM-associated manner. The helicase-nuclease Cas3 pre-processes target DNA into fragments of around 30-100 nt enriched for thymine-stretches in their 3' ends. The Cas1-2 complex further processes these fragments and integrates them sequence specifically into CRISPR repeats by coupling of a 3' cytosine of the fragment. Our results highlight that the selection of PAM-compliant spacers during priming is enhanced by the combined sequence specificities of Cas3 and the Cas1-2 complex leading to an increased propensity of integrating functional CTT-containing spacers.



6.2 Introduction

Priming is a mechanism by which immune systems provide an improved immune response to parasite exposure. In vertebrates, priming of adaptive immunity can occur upon first contact of a T or B cell with a specific antigen and causes epigenetic changes as well as cell differentiation into effector T or B cells, producing high levels of antibodies [1]. More recently, immune priming has been observed in invertebrates, where it provides increased resistance to previously encountered pathogens [2, 3]. In plants, priming refers to a state in which the plant can activate its defense responses more rapidly and strongly when challenged by pathogenic microbes, insects, or environmental stress [4]. In microbes, priming is a mechanism in which cells can update their immunological memory to provide protection against previously encountered but slightly changed viruses or conjugative plasmids [5-9]. Microbial adaptive immune systems do this by integrating short fragments of invader DNA sequences (called spacers) into Clusters of Regularly Interspaced Short Palindromic Repeats (CRISPR). These spacers are transcribed and processed into small CRISPR RNAs and guide Cas (CRISPR-associated) surveillance complexes such as Cascade, Cas9, Cpf1, Csm and Cmr to their DNA or RNA target sequences, resulting in target cleavage and neutralization of the invading threat [10-14].

For many years, the acquisition of new spacers was the least understood process in CRISPR-Cas defense, but recent advances have begun to change this [15-18]. In the Type I-E system of *E. coli*, Cas1 and Cas2 form a complex that binds, processes and integrates DNA fragments into the CRISPR array to form spacers [19-23]. Apart from priming, spacers can also be acquired in a naïve manner. During naïve acquisition the host acquires spacers from an invading DNA element that has not been catalogued in the CRISPR array yet. This process is dependent on DNA replication of the invading DNA element [24] and requires only *cas1* and *cas2* genes [25]. In type I CRISPR-Cas systems, primed acquisition makes use of pre-existing spacers that partially match an invading DNA element. Therefore, primed acquisition of spacers is important to rapidly counter invaders that escape immunity by mutating their target site [5, 26-29]. Priming allows new spacers from such an 'escaper' to be rapidly acquired, leading to renewed immunity. Priming is especially advantageous for a host because the process quickly

generates a population of bacteria with different spacers against the same virus, efficiently driving the virus extinct [30]. In addition to Cas1-2, all remaining Cas proteins are required for priming, including the crRNA effector complex Cascade and the nuclease-helicase Cas3 [5, 7]. Despite knowing the genetic requirements for priming, the exact role of these proteins during priming remains unknown. Several models that explain parts of the priming process have been proposed.

In the Cascade-sliding model, Cascade moves along the DNA until a PAM is encountered, which marks the DNA for acquisition of a new spacer [5]. A second model was proposed in which a Cas1:Cas2-3 complex translocates away from the primed protospacer marked by the crRNA-effector complex until a new PAM is encountered [7]. This new site is then used to acquire a new spacer from. Recently, supporting evidence for this hypothesis has been obtained. Single molecule studies have suggested that Cascade bound to a priming protospacer recruits Cas1-2, which in turn recruit a nuclease inactive Cas3 [31]. A complex of Cas1-3 may then translocate along the DNA to select new spacers. While these models describe the biochemistry and movement of the proteins involved in priming, it has remained unknown how actual DNA fragments from an invading element are obtained to drive the priming process. We have previously put forward a model in which we propose that DNA breakdown products of Cas3 provide the positive feedback needed to fuel the priming process [8]. Similar models were proposed for priming in I-B and I-F systems [6, 9]. In line with that hypothesis, it has recently been suggested that during naïve acquisition spacer precursors are generated during DNA repair at double stranded breaks [24]. These breaks are frequently formed at stalled replication forks during DNA replication and are repaired by the RecBCD complex. RecBCD unwinds the DNA strands with its helicase activity, while degrading the subsequent single stranded stretches using exonuclease activity. The resulting DNA oligomers have been proposed to form precursors for Cas1-2 to produce new spacers. Similar to RecBCD, Cas3 is also a nuclease-helicase that degrades dsDNA by unwinding, with the difference that Cas3 has been shown to degrade one strand at a time [32-36]. This leads to the hypothesis that Cas3 also produces substrates for Cas1-2 mediated spacer acquisition during priming.

Here we have tested that hypothesis and prove that plasmid degradation products produced by Cas3 are bound by the Cas1-2 complex,

processed into new spacers and integrated into the CRISPR array. The cleavage frequency and cleavage specificity of Cas3 facilitate the production of functional spacer precursor molecules that meet all requirements of new spacers. To achieve this, Cas3 produces fragments that are in the range of the length of a spacer (30-100 nt). Furthermore, the cleavage specificity of Cas3 leads to an enrichment of PAM sequences in the 3' end of these fragments, which enhances the selection of productive spacer precursors by Cas1-2. Our results demonstrate that the DNA degradation fragments produced by Cas3 are the direct link between CRISPR interference and adaptation that make the priming mechanism so robust.

6.3 MATERIALS AND METHODS

6.3.1 BACTERIAL STRAINS AND GROWTH CONDITIONS

Escherichia coli strain KD263 was obtained from (Shmakov et al., 2014). *E. coli* strains were grown at 37 °C in Luria Broth (LB; 5 g L^{-1} NaCl, 5 g L^{-1} yeast extract, and 10 g L^{-1} tryptone) at 180 rpm or on LB-agar plates containing 1.5% (wt/vol) agar. When required, medium was supplemented with the following: ampicillin (Amp; 100 μg mL^{-1}), chloramphenicol (Cm; 34 μg mL^{-1}), or kanamycin (Km; 50 μg mL^{-1}). Bacterial growth was measured at 600 nm (OD600).

6.3.2 MOLECULAR BIOLOGY AND DNA SEQUENCING

All oligonucleotides are listed in Table 6.S1. All plasmids are listed in Table 6.S2. All strains and plasmids were confirmed by PCR and sequencing (GATC-Biotech). Plasmids were prepared using GeneJET Plasmid Miniprep Kits (Thermo Scientific). DNA from PCR and agarose gels was purified using the DNA Clean and Concentrator and Gel DNA Recovery Kit (Zymo Research). The library of pGFPuv sp8 mutants was available from a previous study [27]. pMAT MBP-Cas3 was a kind gift from Scott Bailey lab [34].

6.3.3 TRANSFORMATION ASSAY

Transformation assays were carried out in *E. coli* KD263. Cells were grown to OD600 ~0.4, induced with 0.2% L-arabinose and 0.5 mM IPTG and allowed to grow for 1h. Cells were then made chemically competent for heat shock transformation using the RuCl2 method. Cells were co-transformed with 10 ng target plasmid (pWUR836-868, KanR) and 10 ng control plasmid (pWUR835, AmpR) simultaneously [44]. Dilutions of transformants were then plated on LBA plates with Amp and LBA plates with Kan. The transformation efficiency of mutated target plasmids was normalized against the transformation efficiency of the control plasmid.

6.3.4 PLASMID LOSS ASSAY

E. coli KD263 cells were transformed with the target plasmids (pWUR836-868) by heat shock. Individual colonies were picked in triplicate and grown overnight in 5 ml LB supplemented with 2%

glucose to repress *cas* gene expression. The next day, cultures were transferred 1:100 into induced medium (0.2% L-Arabinose, 0.5 mM IPTG) and plasmid loss was monitored. Samples were taken every hour until 5h, and then again at 24h and 48h. Dilutions were plated on non-selective plates and plasmid loss was counted based on loss of fluorescence using a Syngene G-box imager. Plasmid-free colonies were screened for spacer integration by colony PCR using DreamTaq Green DNA polymerase (Thermo Scientific). Acquisition of spacers was detected by PCR using primers BG5301 and BG5302. PCR products were visualized on 2% agarose gels and stained with SYBR-safe (Invitrogen). PCR products were sequenced using Sanger sequencing at GATC (Konstantz, Germany) using primer BG5301.

6.3.5 EMSA assays

Purified Cascade complex with spacer8 crRNA was incubated with plasmid at a range of molar ratios (1:1-100:1, Cascade:DNA) in buffer A (20 mM HEPES pH7.5, 75 mM NaCl, 1 mM DTT) for 30 min. Reactions were run on 1% native agarose gels for 18h at 22 mA in 8 mM sodium-borate buffer. Gels were post stained with SYBR Safe (Invitrogen). Shifted (Cascade bound DNA) and unshifted (free DNA) bands were quantified using the GeneTools software (Syngene) and total Cascade concentration (X) was plotted against the fraction of bound DNA (Y). The curves were fitted with the following formula: Y = (amplitude * X)/(Kd + X) [64]. The amplitude is the maximum fraction of bound DNA. Since the amplitude is not always 1, we cannot directly compare Kd values, instead the 'affinity ratio' was calculated as: amplitude/Kd (i.e. normalizing the Kd against the variable amplitude).

6.3.6 Cas3 DNA degradation assays

Cas3 DNA degradation activity was routinely tested by incubating 500 nM Cas3 with 4 nM M13mp8 single stranded circular DNA in buffer R (5 mM HEPES, pH8, 60 mM KCl) supplemented with 100 μM Ni2+ at 37 °C for 1 h. Plasmid-based assays were performed by incubating 70 nM Cas3 with 70 nM Cascade, 3.5 nM plasmid DNA in buffer R (+ 10 μM CoCl2, 10 mM MgCl2, 2 mM ATP) at 37 °C for 10-60 minutes unless indicated otherwise. For quantifying Cas3 activity, assays were run at normal conditions and samples were taken at 0 min, 1 min, 10 min and 30 min. Samples were immediately quenched with

6x DNA loading dye (Thermo scientific) on ice. Samples were run on agarose gels and supercoiled plasmid bands were quantified using the GeneTools software (Syngene). The DNA degradation was plotted (X: time [min]; Y: Intact Plasmid [%]) and the initial activity of Cas3 [%/min] calculated from the initial slope of the curve.

6.3.7 Protein purification

All proteins were expressed in *E. coli* Bl21-AI cells. Cascade was purified as described earlier [65]. MBP-Cas3 was purified as described in [34]. The Cas1-2 complex was purified as follows. The Cas1-2 operon was PCR amplified with primers BG4556/7 and cloned into pET52b (SmaI/SacI) to make pWUR871. The Cas1-2 complex was purified using the N-terminal StrepII tag on Cas1. Briefly, cells were grown to an OD600 of 0.4, cooled on ice for 30 minutes and induced with 0.5 mM IPTG and 0.2% l-arabinose. Protein was expressed at 20 °C overnight. Cells were collected by centrifugation and lysed in buffer L (20 mM HEPES pH 7.5, 75 mM NaCl, 1 mM DTT, 5% glycerol, 0.1% Triton X100) using a Stansted pressure cell homogenizer. The lysate was cleared by centrifugation and filtration. The cleared lysate was incubated with Strep-tactin beads (IBA) for 30 minutes at 4 °C and loaded into a gravity column. The column was washed with buffer A (20 mM HEPES pH 7.5, 300 mM NaCl, 1 mM DTT, 5% glycerol) and the proteins eluted in buffer B (20 mM HEPES pH 7.5, 75 mM NaCl, 1 mM DTT, 5% glycerol, 2.5 mM biotin). The presence and purity of the Cas1-2 complex was checked via Tris-tricine SDS PAGE (10-20%). The final complex was snap frozen in liquid nitrogen and stored at -80 °C.

6.3.8 Degradation product analysis

To test if Cas3 produces single- or double-stranded DNA products, the reaction products of the plasmid-based assay were incubated with dsDNase (Thermo Scientific) according to manufacturer's protocol. dsDNase exclusively degrades double-stranded DNA. Products were run on a 5% denaturing PAGE gel and visualized using Sybr-Gold (Thermo Scientific). To determine the phosphorylation state of the degradation products, the products were 32P labelled with T4 PNK (Thermo) using the forward and exchange reaction according to the manufacturer's protocol. Labelled DNA was run on an 8% PAGE gel and visualized using a phosphor imaging screen (GE healthcare) and a

Personal molecular imager (Bio-Rad).

Statistical testing against the null hypothesis. We used a version of the empirical bootstrap method [66] to test our data against the null hypothesis that observed behaviors (D±P±) do not correlate with a particular sequence property. To establish the confidence with which the null hypothesis can be disregarded, we construct randomized mock behavioral groups by repeatedly (105 times, resulting in an accuracy in the significance intervals of about) drawing a random selection (allowing repetitions) of sequences from the complete set of 31 protospacers (including the bona fide spacer). The average property of interest is then calculated for the generated mock behavioral groups, giving histograms showing the distribution over the mock sets. The above procedure is performed for the total number of effective mismatches, and the number of mutations within segment 1, and the number of mutations on position 3 within all segments combined.

6.3.9 *In vitro* acquisition assay

Two types of assays were performed. 1) Cas3 plasmid DNA degradation assays were carried out as described above, the reaction products were incubated with Cas1-2 and pWUR869 in buffer R for 60 min. 2) Target plasmid, Cascade, Cas3, Cas1-2 and pWUR869 were incubated in buffer R for 60 min. Component concentrations for assay 1 and 2 were as follows: 70 nM Cascade, 70 nM Cas3, 300 nM Cas1-2, 3.5 nM target plasmid, 5 nM pWUR869 (pCRISPR). Reaction products of both assays were run on a 1.8% TAE-agarose gel. To verify half-site integration of spacers in the CRISPR array as described in [21], nicked pWUR869 was isolated from gel and analyzed by PCR. PCR was performed with forward primer BG5301 (site2) or BG7522 (site1) and reverse primers BG7415/6 (control) or BG6713-15 (3 hotspots) or BG7215/6 (fw/rv of hotspot3). These primers match spacers that are frequently incorporated in vivo [27]. To verify and analyze integration, PCR products were cloned into a pGEMT-easy vector (Promega) and individual clones were sequenced.

6.3.10 NGS library construction

Plasmid degradation assays were performed as previously described. Three different targets were chosen: bona fide target plasmid

(pWUR836) or M4 target plasmid (pWUR853) with 0.13 mM ATP and the m13mp8 assay as described above. Degradation fragments were processed for Illumina MiSeq sequencing as follows. Degradation products were gel purified using the Zymoclean Gel DNA Recovery Kit (Zymo Research), cutting out DNA up to ~500bp. DNA was then poly-A tailed with TdT (Invitrogen) according to manufacturer's protocol (approximately 100 nt tails). Tailed DNA was purified using the DNA Clean and Concentrator Kit (Zymo Research). Subsequently, tailed products were 5' phosphorylated with T4-PNK (Thermo Scientific). Next, the DNA was heated to 95°C to separate DNA strands and a barcoded ssDNA adapter (BG6170/4/6) was ligated to the 5' end of the products. Unincorporated adapters were removed using the DNA Clean and Concentrator Kit (Zymo Research). PCR amplification was performed with BG6179 and BG6180. A second round of PCR amplification was performed with BG6179 and BG6183/7/9 (barcoded). PCR products were purified and sent to the Imagif, Centre for Molecular Genetics, Centre National de la Recherche Scientifique, France for sequencing (paired-end, 2x250nt). Based on the procedure outlined above, a fraction of degradation fragments smaller than 50 nucleotides was purified with lower yields during the initial agarose gel extraction, and could be less populated in the size distribution shown in Fig 3B/S6A.

6.3.11 NGS Data analysis

Sequencing data was deposited at the European Nucleotide Archive under the accession number PRJEB13999. Samples were de-multiplexed using their barcodes. All pair-end reads were mapped to their originating sequences (pWUR836/853, m13mp8) using BLAST and allowing for up to one mismatch. Reads for which both ends could not be aligned to the reference sequence were discarded. For the cleavage sites, distinct start/end positions were analyzed independently (see Table 6.S4 and Table 6.S5 for details). For the duplets a sliding window around the cut point was used. For the duplets the following positions were considered: (-2,-1), (-1,1) and (1,2). In this notation the cut point is between -1 and 1, positive positions are inside the considered fragment and negative positions are outside. Enrichment analysis was performed using a hypergeometric probability distribution to model the background probability density associated to the originating sequence. R packages stats (R-Development-Core-Team, 2008) and gg-

plot2 [67] were used for these computations and to generate corresponding graphics.

6.4 Results

Previous studies have shown that direct interference in Type I CRISPR-Cas systems (i.e. the breakdown of Cascade-flagged invading DNA by Cas3) is relatively sensitive to mutations in the PAM and seed sequence of the protospacer [28, 29, 37, 38]. Priming on the other hand is an extremely robust process capable of dealing with highly mutated targets with up to 13 mutations. Priming is influenced by a complex combination of the number of mutations in a target, the position of these mutations, and the nucleotide identity of the mutation. Furthermore, the degree of tolerance of mutations in a protospacer during interference and priming depends on the spacer choice [29].

6.4.1 Timing of plasmid loss and spacer acquisition reveals distinct underlying processes

In order to find the molecular explanation for why some mutants with equal numbers of mutations show priming while others do not, we performed detailed analysis of a selected set of target mutants obtained previously [27]. From the available list we chose the bona fide target (WT) and 30 mutants carrying an interference permissive PAM (i.e. 5'-CTT-3'). The mutants had between 2 and 5 effective mutations (i.e. mutations outside the kinked positions, 6, 12, 18, 24, 30 ([27, 39-41])) (Figure 6.S1). We used *E. coli* strain KD263 with inducible expression of *cas3* and *cascade-cas1-2* genes [42] to test both direct interference and priming in a plasmid loss setup. Plasmid loss curves of individual mutants (Figure 6.S2) showed four distinct behaviors that led us to classify these target mutants into four groups: mutants capable of only direct interference (D+P-), mutants capable of direct interference and priming (D+P+), mutants capable of only priming (D-P+), and mutants incapable of both direct interference and priming (D-P-) (Figure 6.1A, B). As expected, rapid plasmid loss was observed for the bona fide target, but also for five mutant targets. These target variants (D+P-) showed plasmid loss within 2 hours post induction (hpi), reaching complete loss after 3 hpi (Figure 6.1B bottom left cluster), and did not incorporate new spacers. The D+P+ group of mutants showed a slower decrease in plasmid abundance (starting ~3 hpi) and this decrease was accompanied by incorporation of new spacers 4 hpi

(Figure 6.1B bottom right cluster). The D-P+ group of mutants showed more strongly delayed plasmid loss (>5 hpi), and this loss was preceded or directly accompanied by spacer acquisition (Figure 6.1B top right cluster). Therefore, these mutants could not be cleared from the cells by direct interference initially, but after primed spacer acquisition the plasmid was rapidly lost. No spacer incorporation was observed for D-P- targets and these variants did not show any plasmid loss within 48 hpi, similar to a non-target plasmid (Figure 6.1B top left cluster). This group exemplifies that no naïve acquisition had occurred within 48 h in our experimental setup and that all spacer integration events observed in P+ groups were due to priming. To validate that spacer acquisition occurred by priming, we sequenced the newly incorporated spacers for a representative set of clones, especially including mutants with late acquisition. We did indeed observe the 9:1 strand bias of new spacers that is typical for priming [5, 8, 43]. Taken together, we found that priming is facilitated by slow or delayed direct interference (D+P+), but that it does not strictly require direct interference as exemplified by the D-P+ group.

6.4.2 MODERATE DIRECT INTERFERENCE ACTIVITY FACILITATES THE PRIMING PROCESS

To verify that rapid plasmid loss indeed results from direct interference, we performed plasmid transformation assays of the target plasmid set into *E. coli* KD263 and compared the transformation efficiency to a co-transformed control plasmid [44]. While the bona fide target plasmid exhibited a relative transformation efficiency that was 512x lower than the control plasmid (1/512), also mutants with up to two effective mutations gave rise to strongly decreased transformation efficiencies (1/16 to 1/512) (Fig. 6.1C). This means that these target variants still triggered an efficient direct interference response. Triple mutants showed a range of relative transformation efficiencies from full direct interference (i.e. 1/512) to no direct interference (~1), suggesting a dominant role for the position of the mutations in the protospacer. Mutants with 4 or 5 effective mutations transformed as efficient as the reference plasmid and displayed no direct interference. When we mapped the classification of all the mutants onto the relative transformation efficiency data, the same trend was observed that target variants with the highest direct interference showed no priming. Instead, intermediate levels of direct interference lead to rapid spacer

acquisition, while low levels or the absence of direct interference lead to delayed spacer acquisition. This also confirms that late plasmid loss in the D-P+ group is indeed not caused by direct interference with the original spacer, but by primed spacer acquisition followed by direct interference.

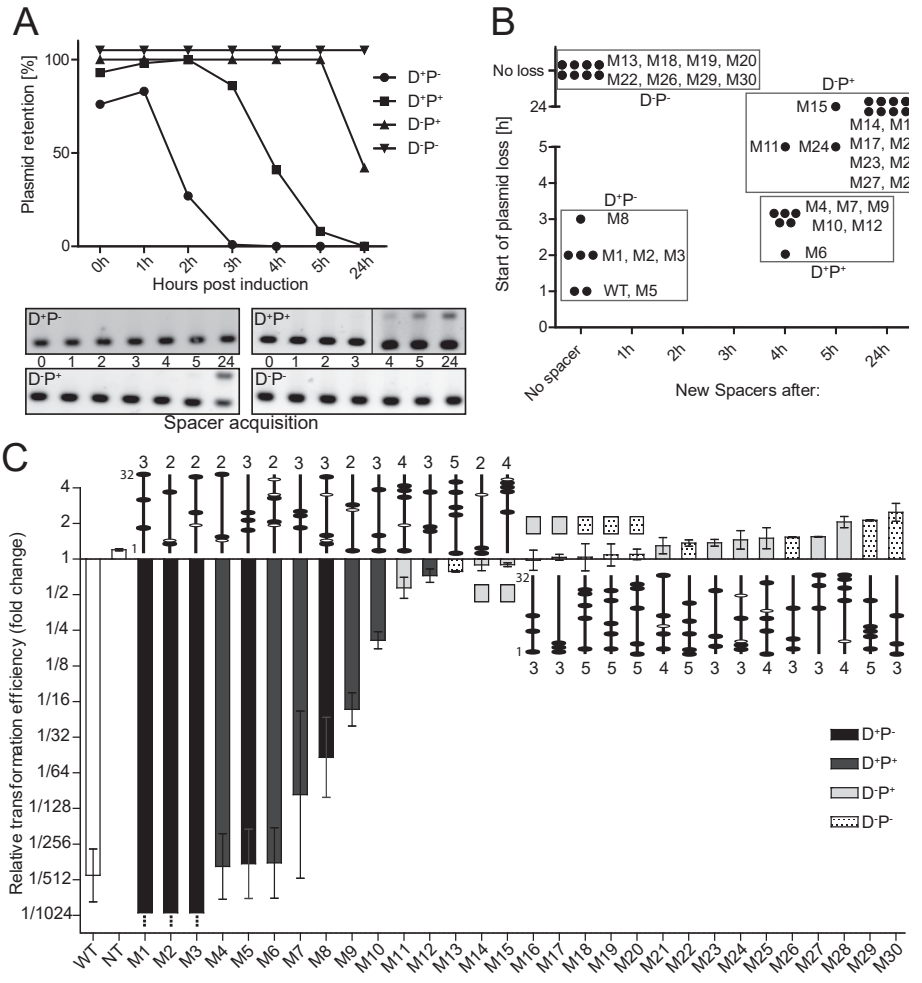


Figure 6.1: Plasmid loss and transformation assay.
Plasmid loss was assessed by plating cells and scoring for the GFP signal at various time points after induction of *cas* genes. Individual assays can be seen in Fig. S2. The bona fide target is abbreviated as WT. **A** Example curves and CRISPR PCR of four different types of plasmid behaviors that were observed: Rapid plasmid loss without spacer integration (D+P-), delayed plasmid loss and spacer integration (D+P+), strongly delayed plasmid loss and spacer integration (D-P+), and no plasmid loss with no spacer integration (D-P-). **B** Summary of plasmid behavior of all mutants, showing timing of first plasmid loss and time of first observable spacer integration. **C** The relative transformation efficiency is plotted for all mutant plasmids (fold change compared to co-transformed non-target plasmid, log2 scale). Bars are color coded based on plasmid behavior classification. Error bars represent the standard error of the mean of triplicate experiments. The positions of mutations are indicated schematically for each mutant (Pos1: Bottom, Pos32: Top). Open ovals represent mutations on positions 6, 12, 18, 24, 30. Closed ovals represent mutations outside of those positions (effective mutations). The amount of effective mutations is indicated above or below the schematic. For a more detailed overview of the mutations, see Fig. S1.

6.4.3 Pairing at the middle position of each segment is important for direct interference

The average number of effective mutations in a protospacer increases gradually over the groups D+P-, D+P+, D-P+, and D-P- (Fig. 6.S1). While D+P- and D+P+ had either 2 or 3 effective mutations, the D-P+ mutants had 3 or 4 mutations and the D-P- mutants carried 3 or 5 effective mutations in the protospacer. In order to quantify how significant the shifts in the average number of mutations are, we used empirical bootstrapping to test against the hypothesis that the classification does not depend on the number of mutations. Our analysis showed that the D+P- and D+P+ groups have significantly fewer mutations than would be expected if the classification did not correlate with the number of mutations (>95% and >68% confidence respectively), while D-P- has significantly more mutations (>95% confidence) (Fig. 6.S3A). We next looked in detail at the number of mutations in each segment, and the position of mutations in each five-nucleotide segment. As has been observed for the seed sequence [28, 38], this showed a significantly lower than average number of mutations in segment 1 for D+P- and D+P+ groups (both 95% confidence, Fig. 6.S3B). Surprisingly, the analysis also revealed that groups showing direct interference (D+P-, D+P+) had no mutations at the third position of each segment (significantly lower than expected, 95% confidence), whereas D-P+ and D-P- groups were enriched for mutations at this position (>68% and >95% confidence respectively, Fig. 6.S3C). This observation therefore suggests that pairing of the middle nucleotide of the segment is somehow important for direct interference. The third nucleotide of each segment could represent a tipping point in the directional pairing of the crRNA to the DNA. This may occur during canonical, PAM-dependent target DNA binding, which leads to R-loop locking, efficient Cas3 recruitment and target DNA degradation [33, 45, 46].

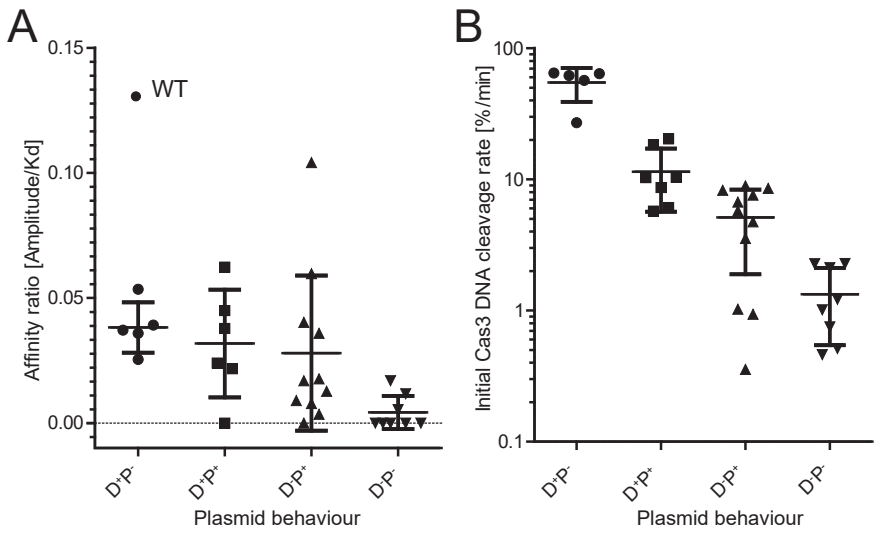


Figure 6.2: EMSA and Cas3 activity assay. A Electrophoretic mobility shift assay (EMSA) of the mutant plasmid set. The affinity ratio (Amplitude/Kd) is plotted for each mutant (see Table S3 for more details). Mutants are separated by the previously made plasmid behavior classification. The mean and standard deviation for each group are indicated. The bona fide target is abbreviated as WT. **B** Cas3 DNA degradation activity assay of mutant plasmid set. The initial Cas3 DNA cleavage rate [%/min] is plotted for each mutant. Mutants are classified according to previously identified plasmid behavior. The mean and standard deviation for each group is indicated. Individual gels for all activity assays can be found in Figure S4.

6.4.4 CASCADE-PLASMID BINDING IS REQUIRED FOR INTERFERENCE AND PRIMING

To determine the biochemical basis of priming, we first asked the question what determines if a mutant target can prime or not, and we hypothesized that the affinity of Cascade for a target plasmid would determine its fate. To test this, we performed plasmid based mobility shift assays with purified Cascade complexes [47]. While the bona fide target and most of the mutant targets were bound to completion at increasing Cascade concentrations, some mutant target plasmids were only partially bound (Table 6.S3), as has been observed before [48]. By calculating an affinity ratio (Amplitude/Kd) and using it as an index for the binding strength, we were able to directly compare the binding properties of all target mutants (Fig. 6.2A). The results show that the bona fide target plasmid had the highest affinity ratio (0.31 nM-1), while the mutants cover a range of ratios ranging from very weak binding (>0.008 nM-1) to almost the same levels as the bona fide

target (<0.1 nM-1). D-P- mutants all cluster together with low ratios (<0.02 nM-1), and 5 out of 8 show no measurable Cascade binding. This suggests that a minimal level of target plasmid binding by Cascade is required for both direct interference and priming. However, the affinity ratio alone does not predict direct interference and/or priming behavior of a target plasmid.

6.4.5 Cas3 DNA cleavage activity determines plasmid fate

Next, we analyzed if the catalytic rate of target DNA degradation by Cas3 would be related to direct interference and priming. Target DNA degradation is required for direct interference and might be required for priming as well, since all *cas* genes are required for priming in *E. coli* [5]. To test this, we performed Cas3 activity assays with the same panel of target plasmids (Fig. 6.2B, Fig. 6.S4). This showed that there is a strong dependence between plasmid fate and Cas3 activity. Mutants capable of only direct interference (D+P-) display 5 to 10 times higher activity than priming mutant classes (D+P+, D-P+), while stable mutants (D-P-) show the lowest Cas3 activity. Furthermore, D+P+ mutants show a higher average activity than D-P+ mutants, although there is overlap between the two groups. The difference between the Cascade affinity and the Cas3 activity plots shows that Cas3 activity is not a simple reflection of Cascade affinity, but is likely influenced by other factors such as conformational differences or the dynamics of Cascade binding. Taken together, there is a link between the Cas3 activity on a target, and target plasmid fate. Direct interference requires the highest Cas3 activity, while priming requires a level of target degradation and occurs at a broad range of intermediate or low Cas3 activities. Finally, it is striking that higher Cas3 activities seem to result in faster priming (D+P+ vs D-P+), while very high Cas3 activities (D+P-) do not lead to priming.

6.4.6 Cas3 produces degradation fragments of near-spacer length

After establishing a connection between plasmid degradation (direct interference) and primed spacer acquisition, we sought to analyze whether the degradation fragments created by Cas3 could serve as spacer precursors. To this end, we performed Cascade-mediated plas-

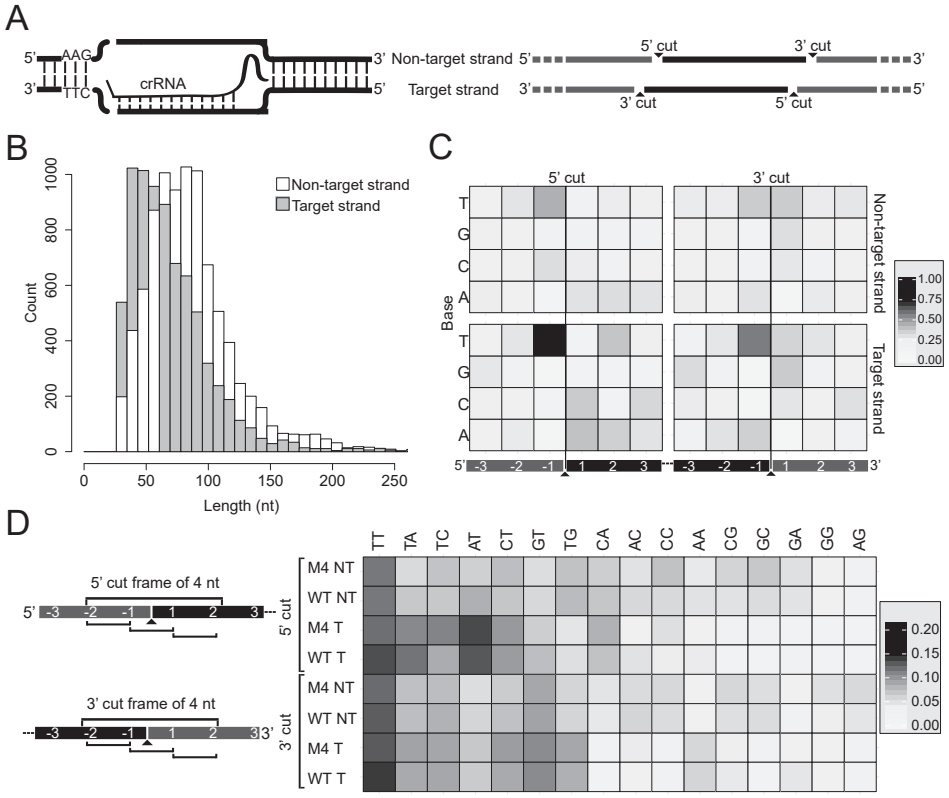


Figure 6.3: Next generation sequencing analysis of Cas3 DNA degradation products. A Left: Schematic of R-loop formed by binding of Cascade to dsDNA target. Right: Schematic showing the four distinct Cas3 cleavage sites in dsDNA target. **B** Length distribution of Cas3 DNA degradation fragments of M4 target. **C** Heat map of nucleotide frequencies around cleavage sites. The cleavage site is between position -1 and 1. Positions indicated in black are on the fragments, positions indicated in grey are outside of fragments. **D** Heat map of dinucleotide frequencies around cleavage sites. Abundance of dinucleotides was measured in a shifting frame within 4 nucleotides around the cleavage sites.

mid degradation assays with Cas3 and plasmids containing the bona fide target or M4 target. Agarose gel electrophoresis showed that both target plasmids were degraded into similar sized products smaller than 300 nt. Further biochemical analysis of the products revealed that the products were of double stranded nature and contained phosphates at their 5' end (Fig. 6.S5A, B). Based on the unidirectional unwinding and single stranded DNA cleavage mechanism of Cas3 [32-36], we had expected to find single stranded DNA. However, it appeared that complementary fragments had re-annealed to form duplexes, most likely generating annealed products with both 3' and 5' overhangs.

In order to determine the exact cleavage patterns of target plasmids by Cas3, we isolated DNA cleavage products from gel and sequenced them using the Illumina MiSeq platform. Analysis of the length of the DNA degradation products from the bona fide and M4 target revealed that the majority of fragments from the target strand had a size of around 30-70 nt (Fig. 6.3B, Fig. 6.S6A). The non-target strand displayed a shifted distribution with most fragments being 60-100 nt long. Instead of cleaving the target DNA randomly, Cas3 produces fragments with a distinct length profile. Furthermore, the length of the main fraction, especially in the target strand, is close to the length of a spacer molecule (i.e. 32/33 nucleotides), supporting the idea that these fragments might be used as spacer precursor molecules.

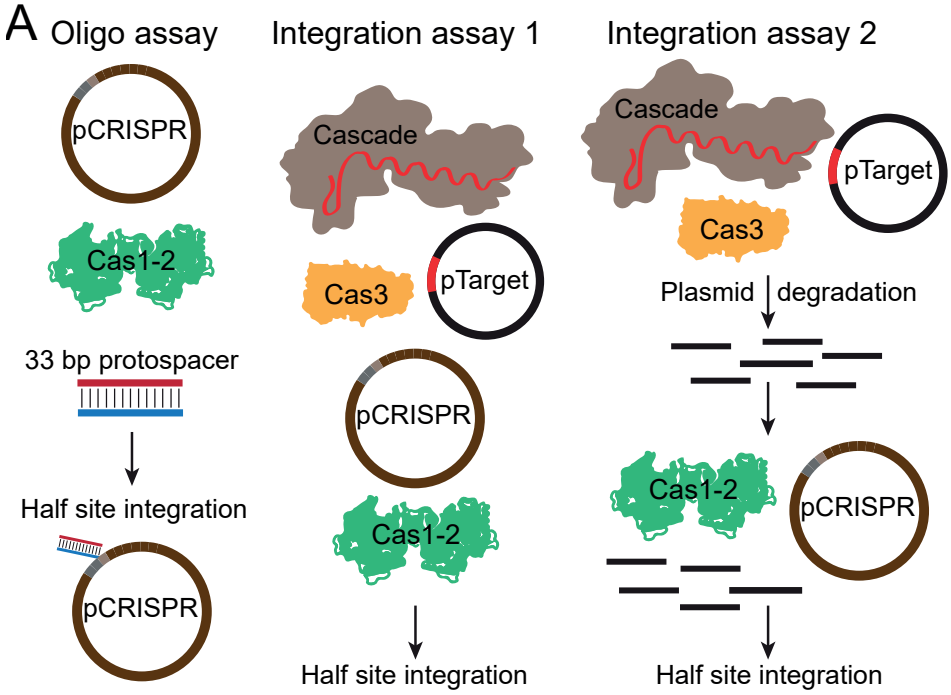
6.4.7 Cas3 cleavage is sequence specific for thymine stretches

In order to see if Cas3 cleaves the target DNA in a sequence specific manner, we analyzed the region encompassing the cleavage site. This revealed a preference for Cas3 to cleave in thymine-rich sequences for both the bona fide and the M4 target, preferably cleaving 3' of a T nucleotide (Fig. 6.3C, D and Fig. 6.S6B). The same pattern was also observed for single stranded m13mp8 DNA cleaved in the absence of Cascade, indicating that T-dependent cleavage specificity is an inherent feature of the HD domain of Cas3. The cleavage specificity of Cas3 leaves one or multiple T nucleotides on the 3' ends of DNA degradation products. This enriches the 3' ends of the fragments for NTT sequences, including the PAM sequence CTT. A considerable proportion of degradation fragments therefore satisfies the requirement of Cas1-2

for having CTT sequences in the 3' ends of spacer precursors in order for these to be correctly integrated into the CRISPR array [23, 49]. Interestingly, C/T-associated cleavage has previously been shown for *Streptococcus thermophilus* Cas3 cleaving oligo nucleotides [35], suggesting that this cleavage specificity may be common for HD-domains of Cas3 proteins.

6.4.8 Cas1-2 integrate Cas3-derived degradation fragments

To find out if Cas3 degradation products can indeed serve as spacer precursors, we reconstituted spacer integration in vitro using purified Cas proteins. Two types of spacer integration assays were performed (Fig. 6.4A): the first assay used all Cas proteins simultaneously (Cascade, Cas3, Cas1-2) to degrade a target plasmid and integrate the resulting fragments into a plasmid carrying a leader and single CRISPR repeat (pCRISPR). The second assay used DNA degradation products from a separate Cascade-Cas3 reaction. These products were incubated with Cas1-2 and pCRISPR, as described [21]. We noticed a pronounced Cas1-2-dependent shift of the degradation fragments in the gel, suggesting the fragments are bound by Cas1-2 (Fig. 6.4B, left panel). Interestingly, when Cas1-2 was present in the reaction we observed twice as much nicking of plasmid pCRISPR, suggesting half site integration of DNA fragments into pCRISPR had occurred (Fig. 6.4B, right panel) [21]. The same pCRISPR nicking activity was observed using purified Cas3 degradation products (integration assay 2) indicating the integration reaction was not dependent on Cascade or Cas3.


A
Oligo assay
Integration assay 1
Integration assay 2
pCRISPR
Cas1-2
33 bp protospacer
Half site integration
pCRISPR
Cascade
Cas3
pTarget
pCRISPR
Cas1-2
Half site integration
Cascade
pTarget
Cas3
Plasmid degradation
Cas1-2
pCRISPR
Half site integration

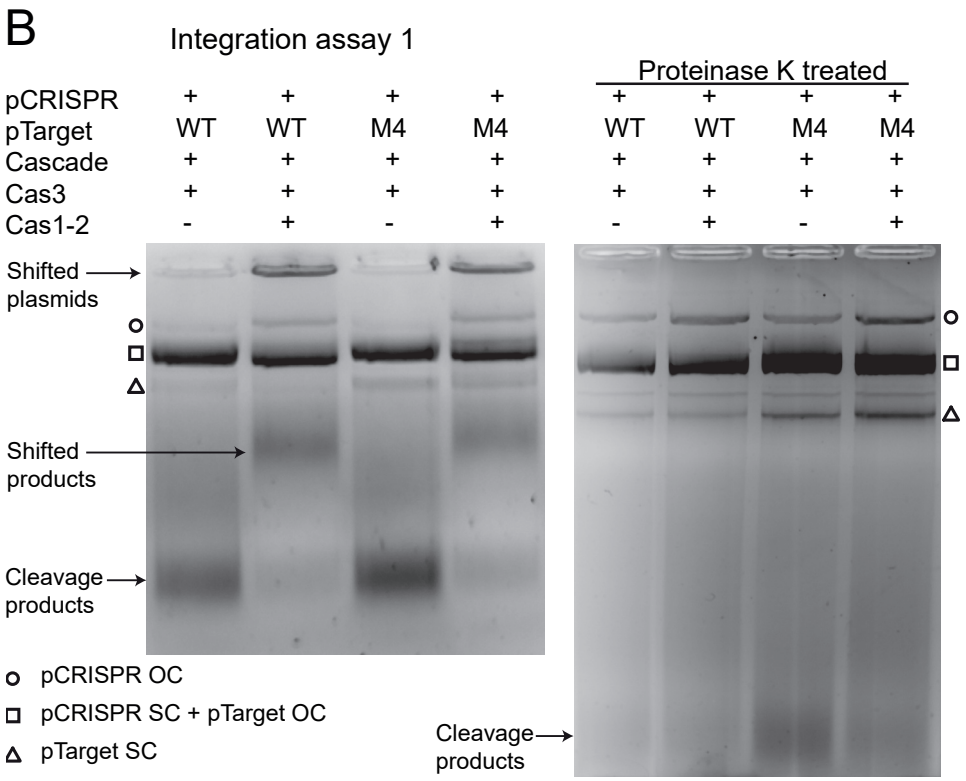

B
Integration assay 1
Proteinase K treated
pCRISPR + + + + + + + +
pTarget WT WT M4 M4 WT WT M4 M4
Cascade + + + + + + + +
Cas3 + + + + + + + +
Cas1-2 - + - + - + - +
Shifted plasmids
Shifted products
Cleavage products
pCRISPR OC
pCRISPR SC + pTarget OC
pTarget SC
Cleavage products

Figure 6.4: *In vitro* spacer acquisition assays. A Illustration of the three types of assays performed. In the oligo assay, pCRISPR is incubated with Cas1-2 and a spacer oligo (BG7415/6), leading to half site integration. In assay 1, pTarget and pCRISPR are incubated with Cascade, Cas3 and Cas1-2 for simultaneous degradation of pTarget and half site integration into pCRISPR. In assay 2, pTarget is incubated with Cascade and Cas3 and the resulting DNA degradation products are then separately incubated with pCRISPR and Cas1-2. **B** Gel electrophoresis of integration assay 1. The bona fide target is abbreviated as WT. Left gel, untreated; right gel, Proteinase K treated. Cas1-2 presence causes upwards shift of DNA. Original plasmids are supercoiled (SC), half site integration causes nicking of pCRISPR, resulting in the open circular conformation (OC).

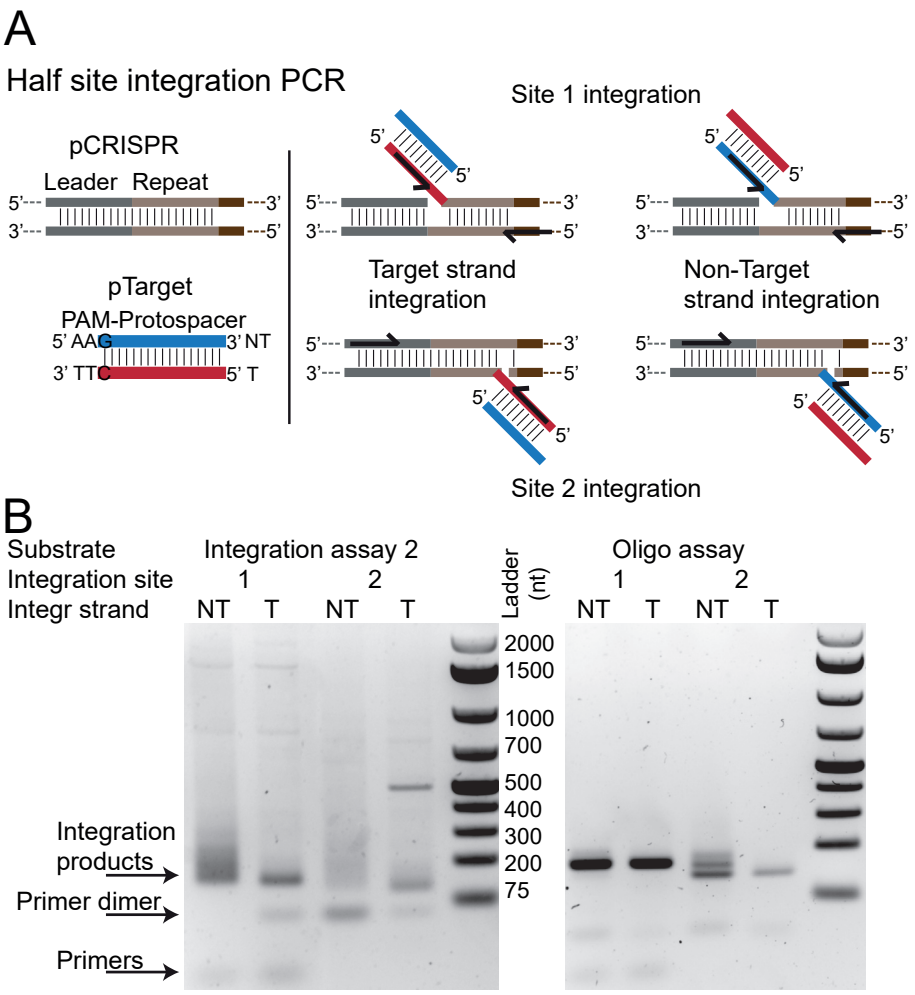


Figure 6.5: Half site integration PCR. A Illustration of the half site integration PCR. Primer sets are chosen to show integration into site 1 (leader-proximal repeat end) and site 2 (leader distal repeat end), and to see both possible orientations of the integrated spacer. Primer sequences were chosen based on frequently incorporated spacers (hotspots) *in vivo* [27]. **B** Gel electrophoresis of half site integration PCR based on integration assay 2 (left) and oligo assay (right). PCR products representing integrations are indicated with an arrow. PCR products were specific to reactions containing all components. Lower running PCR products are primer dimers (verified by sequencing).

To verify that spacer half-site integration had taken place and not just pCRISPR nicking, we gel-isolated the nicked pCRISPR band for PCR analysis. Since we did not know the sequence of the integrated fragments, we selected three primer pairs that would amplify frequently incorporated spacers from the plasmid in vivo [27]. Two of the three tested primers gave a PCR product of the expected size and we chose one of the primers for more detailed analysis. It has previously been shown that the first half-site integration may occur at the boundary of the leader and repeat in the sense strand (i.e. site 1), or at the penultimate base of the repeat in the antisense strand (i.e. site 2) [21, 22]. Furthermore, fragments can be integrated in two different orientations. We performed PCR amplification reactions to test for all four different situations (Fig. 6.5A). This showed that integration of Cas3-derived degradation products occurs sequence specifically at both site 1 and site 2, and in both orientations (Fig. 6.5B).

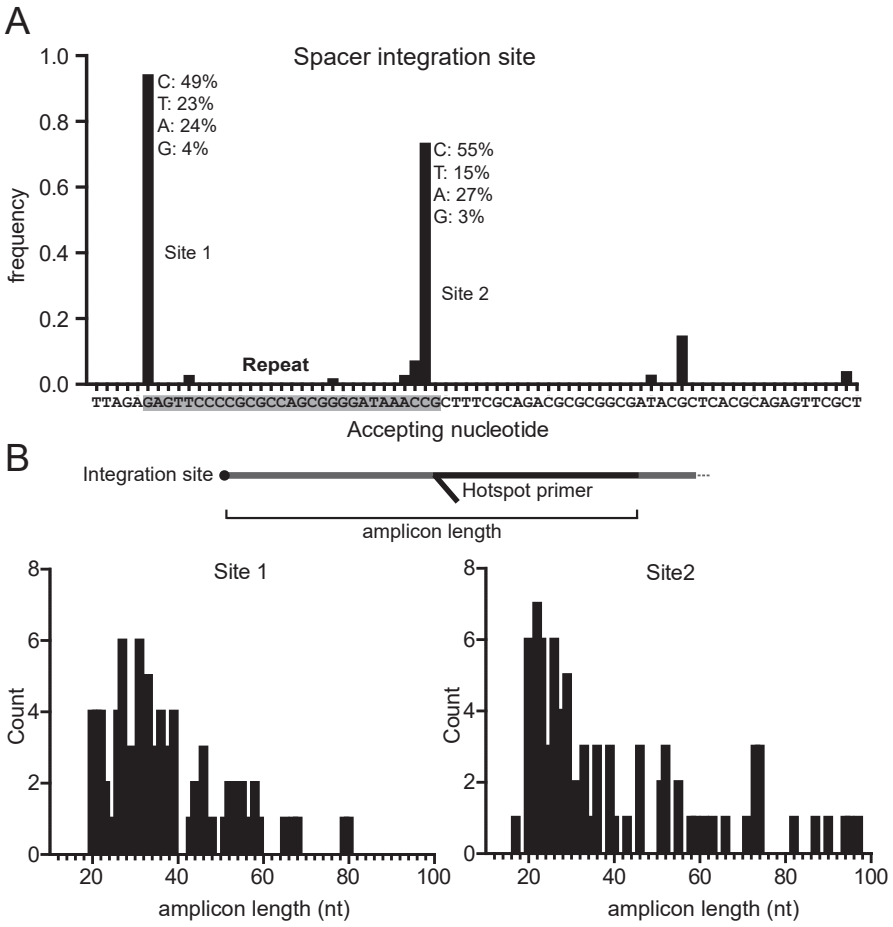
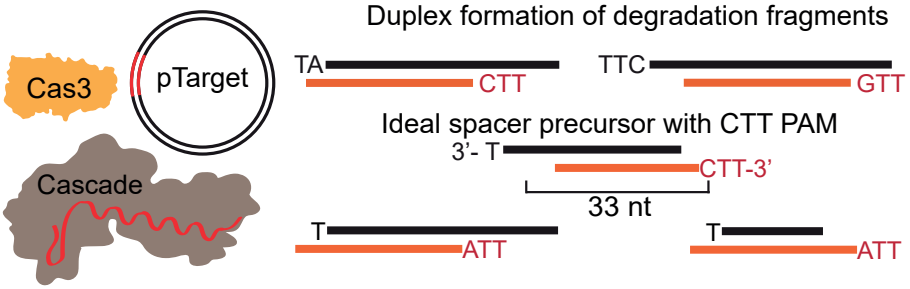


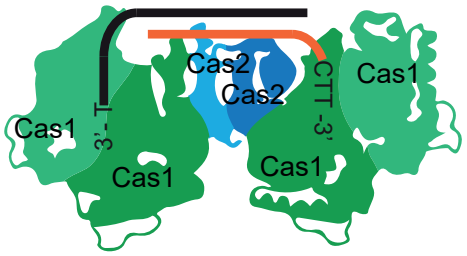
Figure 6: Sequencing analysis of spacer integration. A Frequencies of exact integration locations for integration at site 1 (grey bars) and site 2 (black bars) as determined by sequencing. X-axis gives the backbone nucleotide to which the spacer is coupled. Frequencies of coupled spacer nucleotides are indicated for the 2 canonical insertion locations. **B** Top: Schematic of integrated fragment and method of length determination. Bottom: Length of the integration amplicon for site 1 and site 2.

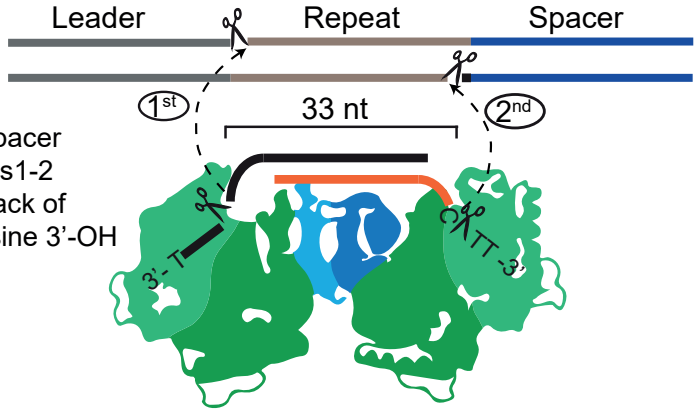
6.4.9 Integration of fragments in the repeat is nucleotide and position specific

In order to obtain more insight into the accuracy of integration, we sequenced 48 clones for each of the four primer sets. The results confirm that fragments from the target and non-target strands are integrated at both site 1 and site 2 of the repeat. Integration is very specific to the correct positions in the repeat. At site 1, 94% of the integrated fragments were coupled correctly to the first nucleotide of the sense strand of the repeat, while at site 2, 73% of integrated fragments were coupled correctly to the penultimate nucleotide of the antisense strand of the repeat, replacing the last nucleotide of the repeat in the process (Fig. 6.6A). In line with previous findings [21, 22], both integration sites show a preference for coupling incoming C nucleotides; 49% and 55% for site 1 and site 2 respectively (Fig. 6.6A). Considering that Cas3 DNA degradation fragments have T nucleotides on their 3' ends, this suggests that precursors have been pre-processed by Cas1-2 before integration, as has been demonstrated for artificial substrates [23]. The majority of the integration amplicons had a length of only 20 to 40 nucleotides (Fig. 6.6B), indicating that the integration reaction prefers short to long substrates. Altogether, we show that the integration of PAM-containing spacers in the repeat during priming is enhanced by the combined sequence specificities of two Cas enzymes: (1) Cas3 which leaves thymines in the 3'-end of DNA fragments, enriching the fragment ends for CTT, and (2) Cas1-2 which prefer CTT carrying substrates and process and couple the 3' cytosine specifically to both integration sites of the repeat.



Binding of a spacer precursor containing a CTT PAM by Cas1-2





Processing of spacer precursor by Cas1-2 nucleophillic attack of Repeat by cytosine 3'-OH

Stable spacer integration intermediate
Gap fill in and repair

New PAM compatible spacer

Figure 6.7: Model of primed spacer acquisition. Cleavage of a targeted plasmid during direct interference by Cascade and Cas3. Cleavage products are near-spacer length and reanneal to form duplexes with 5' and/or 3' overhangs. The fragments are enriched for NTT sequences on their 3' ends. A fraction of the duplexes fulfils spacer precursor requirements: 3' overhangs, CTT at one 3' end and a 33 nt distance between the C and the opposite 3' overhang. Cas1-2 binds spacer precursors with a preference for ideal duplexes as described above [23, 50]. The precursor is processed by Cas1-2 to a length of 33 nt with 3' cytosine. In parallel to processing, 3' ends of the precursor perform a Cas1-2 catalyzed nucleophilic attack on the two integration sites of the repeat [21, 22]. Integration at the leader-repeat junction occurs first [51], subsequently the PAM derived 3' cytosine is integrated to assure correct orientation and production of a functional spacer. A Stable spacer integration intermediate is formed [19]. The gaps are filled in and repaired by the endogenous DNA repair systems, including DNA polymerase I [52].

6.5 Discussion

A remaining gap in our understanding of Type I CRISPR-Cas mechanisms is how new spacers are selected and processed before being incorporated into the CRISPR array. In this work we demonstrate that Cas3 produces spacer precursors for primed adaptation of the CRISPR array. These spacer precursors are 30-100 nt long partially double stranded DNA molecules formed by fragmentation of the target DNA. Cas3 DNA degradation fragments fulfill all criteria for spacer precursors that can be deduced from recent studies of the Cas1-2 complex (Fig. 6.7). Ideal spacer precursors in *E. coli* are partially double stranded duplexes of at least 35 nucleotides containing splayed single stranded 3' ends with a CTT PAM sequence on one of the 3' overhangs [22, 23, 49, 50]. We have shown that Cas3 DNA degradation products are mainly double stranded *in vitro*. This is most likely due to re-annealing of the single stranded products that are produced by the nuclease-helicase activity of Cas3. It is possible that *in vivo* other proteins are involved in the formation of duplexes after degradation. In fact, it has been shown that Cas1 from *Sulfolobus solfataricus* can facilitate the annealing of oligonucleotides [53]. These re-annealed duplexes likely contain a mix of 3' and 5' overhangs, because the two DNA strands of the target are degraded independently. This also results in slightly shorter fragments for the target strand. Despite these differences in fragment size, both strands are cleaved by Cas3 with the same specificity, enriching the 3' ends of the fragments for stretches of thymines. Contrary to the CTT requirements for spacer integration, it is known that Cascade tolerates five different PAM sequences (i.e. CTT, CTA, CCT, CTC, CAT) for direct interference [27, 54]. Howev-

er, the vast majority of new spacers (97%) resulting from primed acquisition carry CTT PAM sequences [42]. This further supports the idea that spacer precursors with CTT-ends are selected non-randomly by the Cas1-2 complex from pools of Cas3 breakdown fragments and further trimmed to a 3' C [23]. These are then coupled to the repeat by nucleophilic attack of the 3'-OH [20, 22]. The T-dependent target DNA cleavage specificity of Cas3 further enhances the production of precursors that fit the requirements of new spacers by creating a pool of DNA fragments with the correct size and correct 3' ends. The interference phase of CRISPR immunity is therefore effectively coupled to the adaptation phase, providing positive feedback about the presence of an invader.

It was previously reported that a dinucleotide motif (AA) at the 3' end of a spacer increases the efficiency of naïve spacer acquisition [55]. We did not observe this motif at the expected distance from the end in the Cas3 DNA degradation fragments, suggesting that Cas3 does not take the AA motif into account when generating spacer precursors.

We found that the integration reaction is very precise for the two correct integration sites in the repeat (site 1 and site 2), and we observed that the integrated fragments most often were the result of a 3' cytosine coupling reaction. *In vivo*, however, only the integration of a CTT-containing fragment at site 2 would lead to a functional spacer targeting a protospacer with PAM (Fig. 6.7), while half site integrations initiating at site 1 would result in 'flipped' spacers [42]. Using a selective PCR strategy, we detected primed spacer acquisition events at both integration sites, and we identified that DNA fragments from both the target and non-target strand of the plasmid could be used for integration. In Type I-E CRISPR-Cas systems, primed spacer acquisitions display a typical 9:1 strand bias for the acquisition of spacers targeting the same strand of DNA as the spacer causing priming [5, 8]. This suggests that *in vivo*, other factors might be involved in further increasing the accuracy of functional spacer integration. This includes the formation of supercomplexes between various Cas proteins (i.e. Cascade, Cas3, Cas1-2) [7, 31, 56], and the involvement of non-Cas host proteins such as PriA, RecG and IHF [51, 52]. IHF ensures that the first integration event takes place at the leader-proximal end of the repeat (site 1) and would be involved in ensuring that the PAM cytosine gets integrated at the leader-distal end (site 2). Supercomplex formation during precursor generation may lead to the selection of

fragments from the target strand containing a CTT PAM at the 3' end. Although the length of the observed integration amplicons is centered around 20-40 nt, we also find amplicons of up to 100 nt. *In vivo*, *E. coli* integrates fragments of 33 nt length. We speculate that trimming of the precursor to 33 nt length occurs after half-site integration and before formation of the stable integration intermediate (Fig. 6.7). Despite the mechanisms that lower erroneous integration of new spacers, it is likely that natural selection of functional spacers *in vivo* also plays a role in the spacers that end up being part of the first population of bacteria following a priming event.

It was surprising that that the bona fide target and several D+P- mutants did not show priming despite providing Cas3 degradation products. Furthermore, the degradation fragments of the bona fide target were very similar to the fragments of the M4 target (D+P+), which cannot explain the difference in priming behavior. We propose that these targets are degraded and cured from the cell too rapidly, giving the acquisition machinery insufficient time to generate new spacers. However, a low level of spacer integration might be taking place at undetectable levels even for the bona fide target, as has been observed previously [8, 29]. In this case, cells with additional spacers do not have a selective growth advantage over cells without new spacers as the plasmid is already effectively cleared from cells without new spacers. Mutant targets with intermediate levels of direct interference however, are replicated and subject to interferencfe over a longer time period, thereby providing more precursors, more time for spacer acquisition to occur, and therefore a greater selective growth advantage. Low levels of direct interference lead to a slow priming response due to the scarcity of spacer precursor molecules. While this paper was under review, another study showed that perfectly matching protospacers with canonical PAMs can indeed stimulate priming and that plasmid targeting is the stimulating factor [57]. In line with our findings, the authors further propose that priming is usually not observed with fully matching protospacers because these targets are degraded too rapidly.

6.5.1 Cut-paste spacer acquisition

We have shown that priming reuses target DNA breakdown products as precursors for new spacers, providing support for a cut and paste mechanism of spacer selection [23]. Compatible models have recently been proposed for naïve spacer acquisition [24]. It was shown

that CRISPR adaptation is linked to double stranded DNA breaks that form at stalled DNA replication forks. Invading genetic elements often go through a phase of active DNA replication when they enter a host cell, and a replication dependent mechanism therefore helps the host to primarily select spacers from the invading element. The RecBCD complex is key in this process as it repairs double stranded breaks by first chewing back the ends of the DNA creating fragments of tens to thousands of nucleotides [15]. These fragments are thought to reanneal and serve as precursors for new spacers. Other studies have shown the direct involvement of crRNA-effector complexes in spacer selection. In the Type I-F CRISPR-Cas system of Pseudomonas aeruginosa the Csy complex is required for naïve spacer acquisition [9]. Also, Cas9 in Type II systems has a direct role in spacer acquisition [58, 59]. Both systems incorporate spacers very specifically from canonical PAM sites, suggesting that the Csy complex and Cas9 are directly involved in PAM recognition during spacer sampling.

6.5.2 MUTATIONS IN THE PROTOSPACER

In this study we have focused on the effect of mutations in the protospacer on direct interference and priming, while maintaining the dominant interference permissive PAM CTT. Apart from underscoring the importance of the number of mutations and existence of a seed sequence [28, 29, 37, 38], we uncover that for direct interference pairing of the middle nucleotide in each 5-nucleotide segment of the protospacer is disproportionately important, and may represent a tipping point in the binding of a target. None of the 30 mutants showing direct interference carried mutations at these middle positions. Also, in a previously obtained list of approximately 3,300 triple mutants showing direct interference [27], mutations at this position were underrepresented (Fig. 6.S3D). This suggests that pairing at the middle position of each segment may be important for continuation of the directional zipping process. This process starts at the PAM and leads to the formation of a canonical locked R-loop, which is required for Cas3 recruitment and target DNA degradation [28, 31, 45, 46, 60, 61]. We stress that we have used variants with CTT PAMs only, which can be engaged by Cascade in the canonical PAM-dependent binding mode [28, 31, 45, 46, 60, 62], and can also trigger priming. It has become clear, however, that targets with mutations in the PAM display a broad spectrum of distinct characteristics depending on the chosen PAM, including a

range of efficiencies of direct interference [63] and the reluctance to trigger efficient Cas3 target DNA degradation [29, 31, 34, 45, 46, 48]. In many cases these PAMs still support the priming process [5, 27, 29]. Targets with highly disfavored PAMs [62] are likely engaged in the non-canonical PAM-independent binding mode [45] and may require recruitment and translocation events of Cas1-2 and Cas3 proteins to initiate the target degradation needed to acquire new spacers.

6.7 Conclusion

The findings presented here showcase the intricate PAM-interplay of all Cas proteins in type I systems to update the CRISPR memory when receiving positive feedback about the presence of an invader. The robustness of priming is achieved by three components that co-evolved to work with PAM sequences: Cas3 producing spacer precursors enriched for correct PAM ends, Cas1-2 selecting PAM-compliant spacer precursors and Cascade efficiently recognizing targets with PAMs. This process stimulates the buildup of multiple spacers against an invader, preventing the formation of escape mutants [5, 7, 8]. When the original spacer triggers sufficiently strong interference, priming acquisition does not frequently occur. This prevents the unnecessary buildup of spacers and keeps the CRISPR array from getting too long. Any subsequent reduction in effectivity of the immune response by further mutations of the invader will in turn allow priming acquisition, restoring immunity.

REFERENCES

1. Bevington, S.L., P. Cauchy, J. Piper, E. Bertrand, N. Lalli, R.C. Jarvis, L.N. Gilding, S. Ott, C. Bonifer, and P.N. Cockerill, Inducible chromatin priming is associated with the establishment of immunological memory in T cells. The EMBO Journal, 2016. 35(5): p. 515-535.
2. Kurtz, J. and K. Franz, Innate defence: evidence for memory in invertebrate immunity. Nature, 2003. 425(6953): p. 37-8.
3. Schmid-Hempel, P., EVOLUTIONARY ECOLOGY OF INSECT IMMUNE DEFENSES. Annual Review of Entomology, 2004. 50(1): p. 529-551.
4. Conrath, U., G.J.M. Beckers, C.J.G. Langenbach, and M.R. Jaskiewicz, Priming for Enhanced Defense. Annual Review of Phytopathology, 2015. 53(1): p. 97-119.
5. Datsenko, K.a., K. Pougach, A. Tikhonov, B.L. Wanner, K. Severinov, and E. Semenova, Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. Nature Communications, 2012. 3(1): p. 945-945.
6. Li, M., R. Wang, D. Zhao, and H. Xiang, Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process. Nucleic Acids Research, 2014. 42(4): p. 2483-2492.
7. Richter, C., R.L. Dy, R.E. McKenzie, B.N.J. Watson, C. Taylor, J.T. Chang, M.B. McNeil, R.H.J. Staals, and P.C. Fineran, Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. Nucleic Acids Research, 2014. 42(13): p. 8516-8526.
8. Swarts, D.C., C. Mosterd, M.W.J. van Passel, and S.J.J. Brouns, CRISPR Interference Directs Strand Specific Spacer Acquisition. PLoS ONE, 2012. 7(4): p. e35888-e35888.
9. Vorontsova, D., K.A. Datsenko, S. Medvedeva, J. Bondy-Denomy, E.E. Savitskaya, K. Pougach, M. Logacheva, B. Wiedenheft, A.R. Davidson, K. Severinov, and E. Semenova, Foreign DNA acquisition by the I-F CRISPR–Cas system requires all components of the interference machinery. Nucleic Acids Research, 2015. 43(22): p. 10848-10860.
10. Carter, J. and B. Wiedenheft, SnapShot: CRISPR-RNA-Guided Adaptive Immune Systems. Cell, 2015. 163(1): p. 260-260.e1.
11. Charpentier, E., H. Richter, J. van der Oost, and M.F. White, Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. FEMS Microbiol Rev, 2015. 39(3): p. 428-41.
12. Makarova, K.S., Y.I. Wolf, O.S. Alkhnbashi, F. Costa, S.A. Shah, S.J. Saunders, R. Barrangou, S.J.J. Brouns, E. Charpentier, D.H. Haft, P. Horvath, S. Moineau, F.J.M. Mojica, R.M. Terns, M.P. Terns, M.F. White, A.F. Yakunin, R.A. Garrett, J. van der Oost, R. Backofen, and E.V. Koonin, An updated evolutionary classification of CRISPR–Cas systems. Nature Reviews Microbiology, 2015. 13(11): p. 722-736.
13. Marraffini, L.A., CRISPR-Cas immunity in prokaryotes. Nature, 2015. 526(7571): p. 55-61.
14. Reeks, J., J.H. Naismith, and M.F. White, CRISPR interference: a structural perspective. The Biochemical journal, 2013. 453(2): p. 155-166.
15. Amitai, G. and R. Sorek, CRISPR–Cas adaptation: insights into the mechanism of action. Nature Reviews Microbiology, 2016. 14(2): p. 67-76.
16. Fineran, P.C. and E. Charpentier, Memory of viral infections by CRISPR-Cas adaptive immune systems: Acquisition of new information. Virology, 2012. 434(2): p. 202-209.
17. Heler, R., L.A. Marraffini, and D. Bikard, Adapting to new threats: the generation of memory by CRISPR-Cas immune systems. Molecular microbiology, 2014. 93(1): p. 1-9.
18. Sternberg, S.H., H. Richter, E. Charpentier, and U. Qimron, Adaptation in CRISPR-Cas Systems. Molecular Cell, 2016. 61(6): p. 797-808.

19. Arslan, Z., V. Hermanns, R. Wurm, R. Wagner, and Ü. Pul, Detection and characterization of spacer integration intermediates in type I-E CRISPR–Cas system. Nucleic Acids Research, 2014. 42(12): p. 7884-7893.

20. Nuñez, J.K., P.J. Kranzusch, J. Noeske, A.V. Wright, C.W. Davies, and J.A. Doudna, Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. Nature Structural & Molecular Biology, 2014. 21(6): p. 528-534.

21. Nuñez, J.K., A.S.Y. Lee, A. Engelman, and J.A. Doudna, Integrase-mediated spacer acquisition during CRISPR–Cas adaptive immunity. Nature, 2015. 519(7542): p. 193-198.

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

23. Wang, J., J. Li, H. Zhao, G. Sheng, M. Wang, M. Yin, and Y. Wang, Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems. Cell, 2015. 163(4): p. 840-853.

24. Levy, A., M.G. Goren, I. Yosef, O. Auster, M. Manor, G. Amitai, R. Edgar, U. Qimron, and R. Sorek, CRISPR adaptation biases explain preference for acquisition of foreign DNA. Nature, 2015. 520(7548): p. 505-510.

25. Yosef, I., M.G. Goren, and U. Qimron, Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. Nucleic Acids Research, 2012. 40(12): p. 5569-5576.

26. Cady, K.C., J. Bondy-Denomy, G.E. Heussler, A.R. Davidson, and G.A. O'Toole, The CRISPR/Cas Adaptive Immune System of Pseudomonas aeruginosa Mediates Resistance to Naturally Occurring and Engineered Phages. Journal of Bacteriology, 2012. 194(21): p. 5728-5738.

27. Fineran, P.C., M.J.H. Gerritzen, M. Suarez-Diez, T. Kunne, J. Boekhorst, S.A.F.T. van Hijum, R.H.J. Staals, and S.J.J. Brouns, Degenerate target sites mediate rapid primed CRISPR adaptation. Proceedings of the National Academy of Sciences, 2014. 111(16): p. E1629-E1638.

28. Semenova, E., M.M. Jore, K.A. Datsenko, A. Semenova, E.R. Westra, B. Wanner, J. van der Oost, S.J.J. Brouns, and K. Severinov, Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proceedings of the National Academy of Sciences, 2011. 108(25): p. 10098.

29. Xue, C., A.S. Seetharam, O. Musharova, K. Severinov, S.J. J. Brouns, A.J. Severin, and D.G. Sashital, CRISPR interference and priming varies with individual spacer sequences. Nucleic Acids Research, 2015. 43(22): p. 10831-10847.

30. van Houte, S., A.K.E. Ekroth, J.M. Broniewski, H. Chabas, B. Ashby, J. Bondy-Denomy, S. Gandon, M. Boots, S. Paterson, A. Buckling, and E.R. Westra, The diversity-generating benefits of a prokaryotic adaptive immune system. Nature, 2016. 532(7599): p. 385-388.

31. Redding, S., S.H. Sternberg, M. Marshall, B. Gibb, P. Bhat, C.K. Guegler, B. Wiedenheft, J.A. Doudna, and E.C. Greene, Surveillance and Processing of Foreign DNA by the *Escherichia coli* CRISPR-Cas System. Cell, 2015. 163(4): p. 854-865.

32. Gong, B., M. Shin, J. Sun, C.-H. Jung, E.L. Bolt, J. van der Oost, and J.-S. Kim, Molecular insights into DNA interference by CRISPR-associated nuclease-helicase Cas3. Proceedings of the National Academy of Sciences, 2014. 111(46): p. 16359-16364.

33. Huo, Y., K.H. Nam, F. Ding, H. Lee, L. Wu, Y. Xiao, M.D. Farchione, Jr., S. Zhou, K. Rajashankar, I. Kurinov, R. Zhang, and A. Ke, Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. Nature structural & molecular biology, 2014. 21(9): p. 771-777.

34. Mulepati, S. and S. Bailey, In vitro reconstitution of an *Escherichia coli* RNA-guided immune system reveals unidirectional, ATP-dependent degradation of DNA target. J Biol Chem, 2013. 288(31): p. 22184-92.

35. Sinkunas, T., G. Gasiunas, S.P. Waghmare, M.J. Dickman, R. Barrangou, P. Horvath, and V. Siksnys, In vitro reconstitution of Cascade-mediated CRISPR immunity in *Streptococcus thermophilus*. The EMBO Journal, 2013. 32(3): p. 385-394.

36. Westra, E.R., P.B.G. van Erp, T. Künne, S.P. Wong, R.H.J. Staals, C.L.C. Seegers, S. Bollen, M.M. Jore, E. Semenova, K. Severinov, W.M. de Vos, R.T. Dame, R. de Vries, S.J.J. Brouns, and J. van der Oost, CRISPR Immunity Relies on the Consecutive Binding and Degradation of Negatively Supercoiled Invader DNA by Cascade and Cas3. Molecular Cell, 2012. 46(5): p. 595-605.

37. Künne, T., D.C. Swarts, and S.J. Brouns, Planting the seed: target recognition of short guide RNAs. Trends Microbiol, 2014. 22(2): p. 74-83.

38. Wiedenheft, B., E. van Duijn, J.B. Bultema, S.P. Waghmare, K. Zhou, A. Barendregt, W. Westphal, A.J.R. Heck, E.J. Boekema, M.J. Dickman, and J.A. Doudna, RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. Proceedings of the National Academy of Sciences, 2011. 108(25): p. 10092-10097.

39. Jackson, R.N., S.M. Golden, P.B.G. van Erp, J. Carter, E.R. Westra, S.J.J. Brouns, J. van der Oost, T.C. Terwilliger, R.J. Read, and B. Wiedenheft, Crystal structure of the CRISPR RNA-guided surveillance complex from *Escherichia coli*. Science, 2014. 345(6203): p. 1473-1479.

40. Mulepati, S., A. Héroux, and S. Bailey, Structural biology. Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. Science, 2014. 345(6203): p. 1479-84.

41. Zhao, H., G. Sheng, J. Wang, M. Wang, G. Bunkoczi, W. Gong, Z. Wei, and Y. Wang, Crystal structure of the RNA-guided immune surveillance Cascade complex in *Escherichia coli*. Nature, 2014. 515(7525): p. 147-50.

42. Shmakov, S., E. Savitskaya, E. Semenova, M.D. Logacheva, K.A. Datsenko, and K. Severinov, Pervasive generation of oppositely oriented spacers during CRISPR adaptation. Nucleic Acids Res, 2014. 42(9): p. 5907-16.

43. Savitskaya, E., E. Semenova, V. Dedkov, A. Metlitskaya, and K. Severinov, High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in *E. coli*. RNA Biol, 2013. 10(5): p. 716-25.

44. Almendros, C. and F.J. Mojica, Exploring CRISPR Interference by Transformation with Plasmid Mixtures: Identification of Target Interference Motifs in *Escherichia coli*. Methods Mol Biol, 2015. 1311: p. 161-70.

45. Blosser, T.R., L. Loeff, E.R. Westra, M. Vlot, T. Künne, M. Sobota, C. Dekker, S.J.J. Brouns, and C. Joo, Two Distinct DNA Binding Modes Guide Dual Roles of a CRISPR-Cas Protein Complex. Molecular Cell, 2015. 58(1): p. 60-70.

46. Rutkauskas, M., T. Sinkunas, I. Songailiene, M.S. Tikhomirova, V. Siksnys, and R. Seidel, Directional R-Loop Formation by the CRISPR-Cas Surveillance Complex Cascade Provides Efficient Off-Target Site Rejection. Cell Rep, 2015. 10(9): p. 1534-1543.

47. Künne, T., E.R. Westra, and S.J. Brouns, Electrophoretic Mobility Shift Assay of DNA and CRISPR-Cas Ribonucleoprotein Complexes. Methods Mol Biol, 2015. 1311: p. 171-84.

48. Hochstrasser, M.L., D.W. Taylor, P. Bhat, C.K. Guegler, S.H. Sternberg, E. Nogales, and J.A. Doudna, CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. Proceedings of the National Academy of Sciences, 2014. 111(18): p. 6618-6623.

49. Shipman, S.L., J. Nivala, J.D. Macklis, and G.M. Church, Molecular recordings by directed CRISPR spacer acquisition. Science, 2016. 353(6298): p. aaf1175-aaf1175.

50. Nuñez, J.K., L.B. Harrington, P.J. Kranzusch, A.N. Engelman, and J.A. Doudna, Foreign DNA capture during CRISPR–Cas adaptive immunity. Nature, 2015. 527(7579): p. 535-538.

51. Nuñez, James K., L. Bai, Lu*cas* B. Harrington, Tracey L. Hinder, and Jennifer A. Doudna, CRISPR Immunological Memory Requires a Host Factor for Specificity. Molecular Cell, 2016. 62(6): p. 824-833.

52. Ivančić-Baće, I., S.D. Cass, S.J. Wearne, and E.L. Bolt, Different genome stability proteins underpin primed and naïve adaptation in *E. coli* CRISPR-Cas immunity. Nucleic acids research, 2015. 43(22): p. 10821-30.

53. Han, D. and G. Krauss, Characterization of the endonuclease SSO2001 from *Sulfolobus solfataricus* P2. FEBS Lett, 2009. 583(4): p. 771-6.

54. Leenay, R.T., K.R. Maksimchuk, R.A. Slotkowski, R.N. Agrawal, A.A. Gomaa, A.E. Briner, R. Barrangou, and C.L. Beisel, Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. Molecular Cell, 2016. 62(1): p. 137-147.

55. Yosef, I., D. Shitrit, M.G. Goren, D. Burstein, T. Pupko, and U. Qimron, DNA motifs determining the efficiency of adaptation into the *Escherichia coli* CRISPR array. Proceedings of the National Academy of Sciences, 2013. 110(35): p. 14396.

56. Plagens, A., B. Tjaden, A. Hagemann, L. Randau, and R. Hensel, Characterization of the CRISPR/Cas Subtype I-A System of the Hyperthermophilic Crenarchaeon Thermoproteus tenax. Journal of Bacteriology, 2012. 194(10): p. 2491-2500.

57. Semenova, E., E. Savitskaya, O. Musharova, A. Strotskaya, D. Vorontsova, K.A. Datsenko, M.D. Logacheva, and K. Severinov, Highly efficient primed spacer acquisition from targets destroyed by the *Escherichia coli* type I-E CRISPR-Cas interfering complex. Proceedings of the National Academy of Sciences, 2016. 113(27): p. 7626-7631.

58. Heler, R., P. Samai, J.W. Modell, C. Weiner, G.W. Goldberg, D. Bikard, and L.A. Marraffini, Cas9 specifies functional viral targets during CRISPR–Cas adaptation. Nature, 2015. 519(7542): p. 199-202.

59. Wei, Y., R.M. Terns, and M.P. Terns, Cas9 function and host genome sampling in Type II-A CRISPR–Cas adaptation. Genes & Development, 2015. 29(4): p. 356-361.

60. Sashital, D.G., B. Wiedenheft, and J.A. Doudna, Mechanism of Foreign DNA Selection in a Bacterial Adaptive Immune System. Molecular Cell, 2012. 46(5): p. 606-615.

61. Szczelkun, M.D., M.S. Tikhomirova, T. Sinkunas, G. Gasiunas, T. Karvelis, P. Pschera, V. Siksnys, and R. Seidel, Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. Proceedings of the National Academy of Sciences, 2014. 111(27): p. 9798-9803.

62. Hayes, R.P., Y. Xiao, F. Ding, P.B.G. van Erp, K. Rajashankar, S. Bailey, B. Wiedenheft, and A. Ke, Structural basis for promiscuous PAM recognition in type I–E Cascade from *E. coli*. Nature, 2016. 530(7591): p. 499-503.

63. Westra, E.R., E. Semenova, K.A. Datsenko, R.N. Jackson, B. Wiedenheft, K. Severinov, and S.J. Brouns, Type I-E CRISPR-*cas* systems discriminate target from non-target DNA through base pairing-independent PAM recognition. PLoS Genet, 2013. 9(9): p. e1003742.

64. van Erp, P.B.G., R.N. Jackson, J. Carter, S.M. Golden, S. Bailey, and B. Wiedenheft, Mechanism of CRISPR-RNA guided recognition of DNA targets in *Escherichia coli*. Nucleic Acids Research, 2015. 43(17): p. 8381-8391.

65. Jore, M.M., M. Lundgren, E. van Duijn, J.B. Bultema, E.R. Westra, S.P. Waghmare, B. Wiedenheft, U. Pul, R. Wurm, R. Wagner, M.R. Beijer, A. Barendregt, K. Zhou, A.P.L. Snijders, M.J. Dickman, J.A. Doudna, E.J. Boekema, A.J.R. Heck, J. van der Oost, and S.J.J. Brouns, Structural basis for CRISPR RNA-guided DNA recognition by Cascade. Nature Structural & Molecular Biology, 2011. 18(5): p. 529-536.

66. Dekking, F., C. Kraaikamp, H. Lopuhaä, and L. Meester, A modern introduction to probability and statistics. Understanding why and how. 2005.

67. Wilkinson, L., ggplot2: Elegant Graphics for Data Analysis by H. WICKHAM. Biometrics, 2011. 67: p. 678-679.

68. Brouns, S.J.J., M.M. Jore, M. Lundgren, E.R. Westra, R.J.H. Slijkhuis, A.P.L. Snijders, M.J. Dickman, K.S. Makarova, E.V. Koonin, and J. van der Oost, Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. Science, 2008. 321(5891): p. 960-964.

SUPPLEMENTARY
SUPPLEMENTARY FIGURES

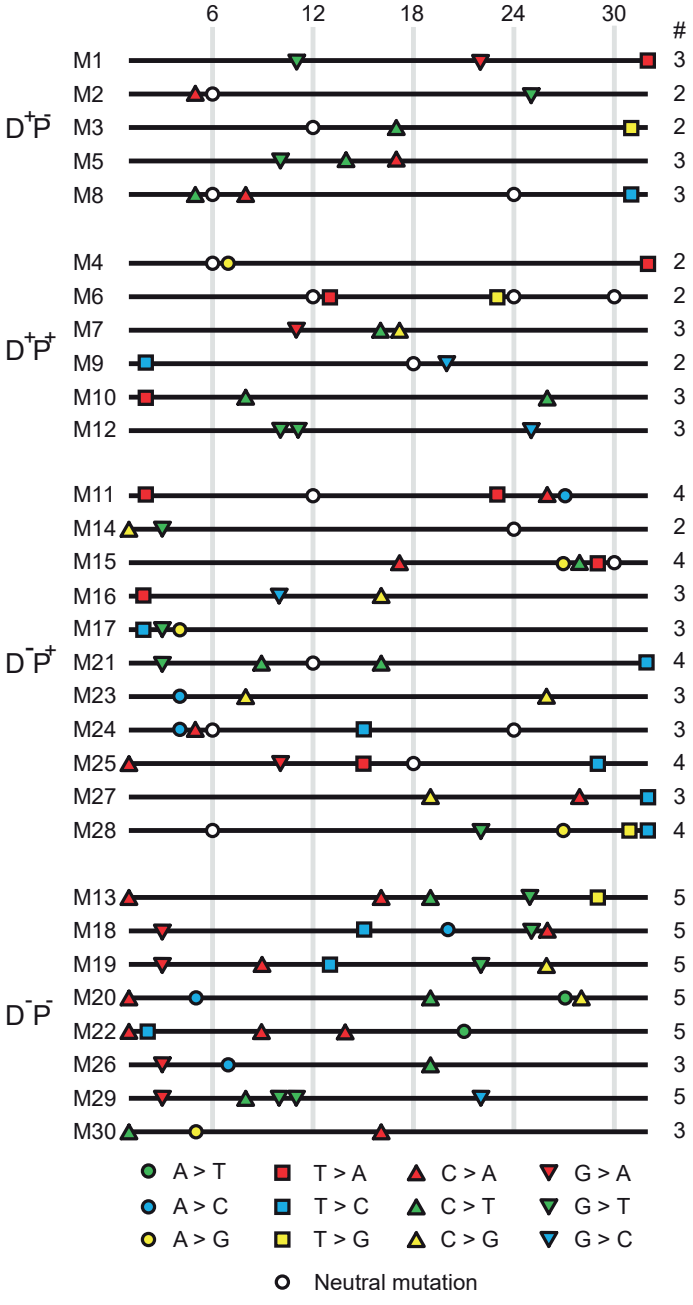
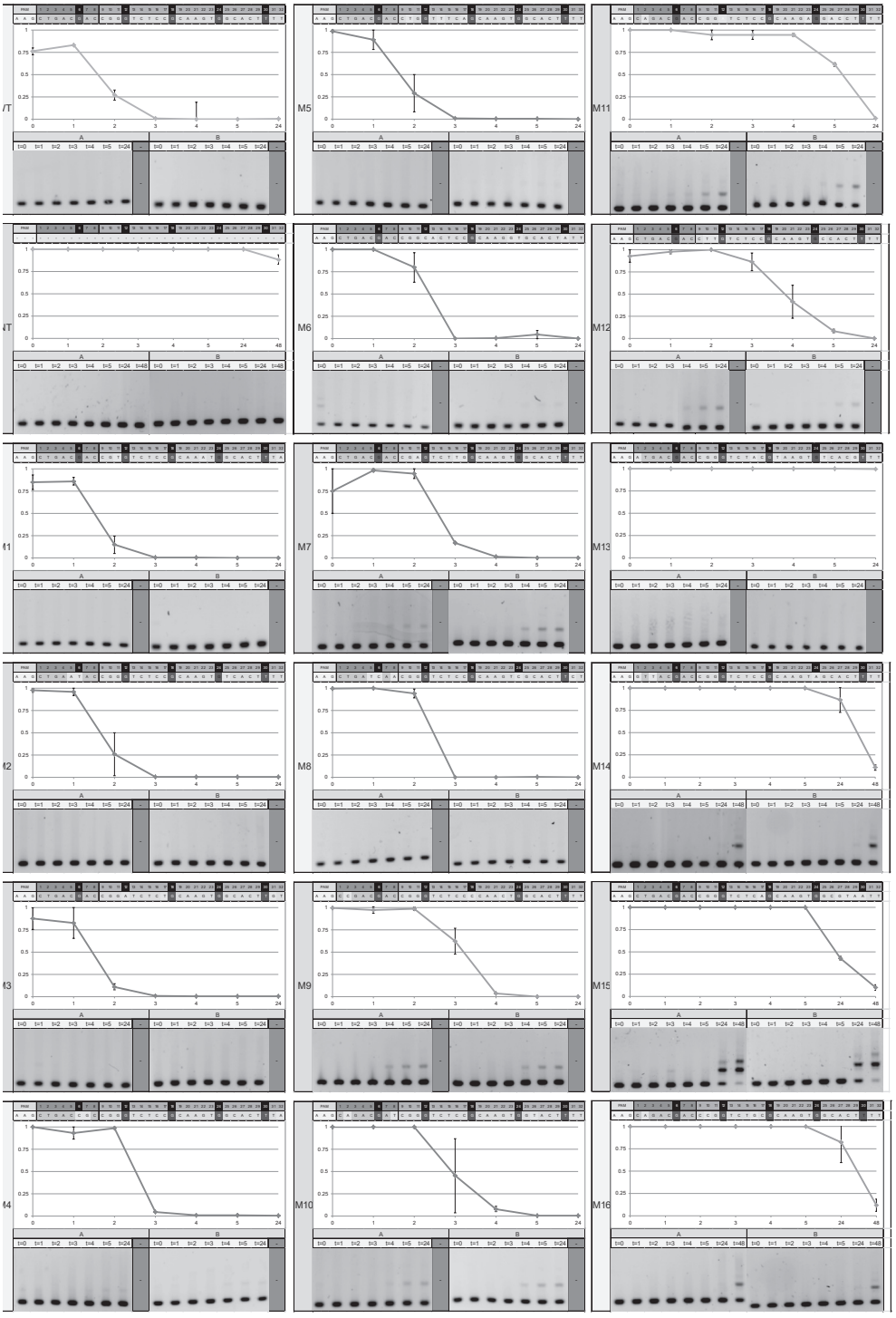


Figure 6.S1. Related to Figure 6.1: Overview of Protospacer8 mutants. 30 mutants of protospacer8 containing either 3 or 5 total mutations were used throughout the study. Mutations on positions 6, 12, 18, 24, 30 (empty circles) are not participating in base-pairing and are therefore not considered as effective mutations. Types of mutations are indicated by colored symbols. Mutants are separated into categories based on their behavior in plasmid loss assays (see also Figure 1B).


M5
M11
M6
M12
M7
M13
M8
M14
M9
M15
M10
M16

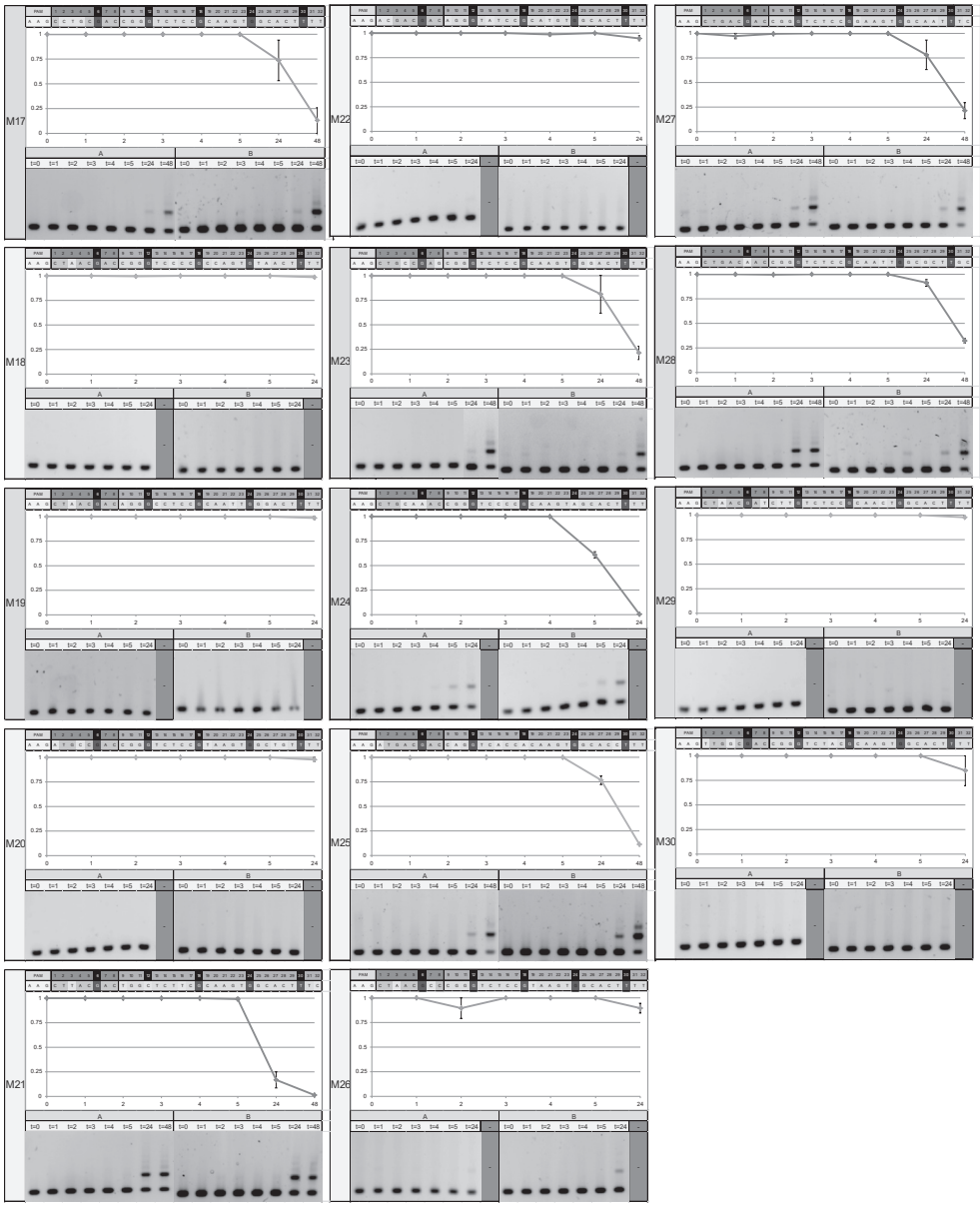


Figure 6.S2. Related to Figure 6.1: Individual plasmid loss assays. Panels for each plasmid mutant with (top to bottom): Sequence with indicated mutations, plasmid loss curves from 0 h to 24 h or 48 h, duplicate of CRISPR PCR showing spacer acquisition. The bottom bands in the PCR gels represent the unextended array, higher bands represent the array with an extra spacer. Error bars in plasmid loss graphs represent the standard deviation of replicate experiments. The bona fide target is abbreviated as WT.

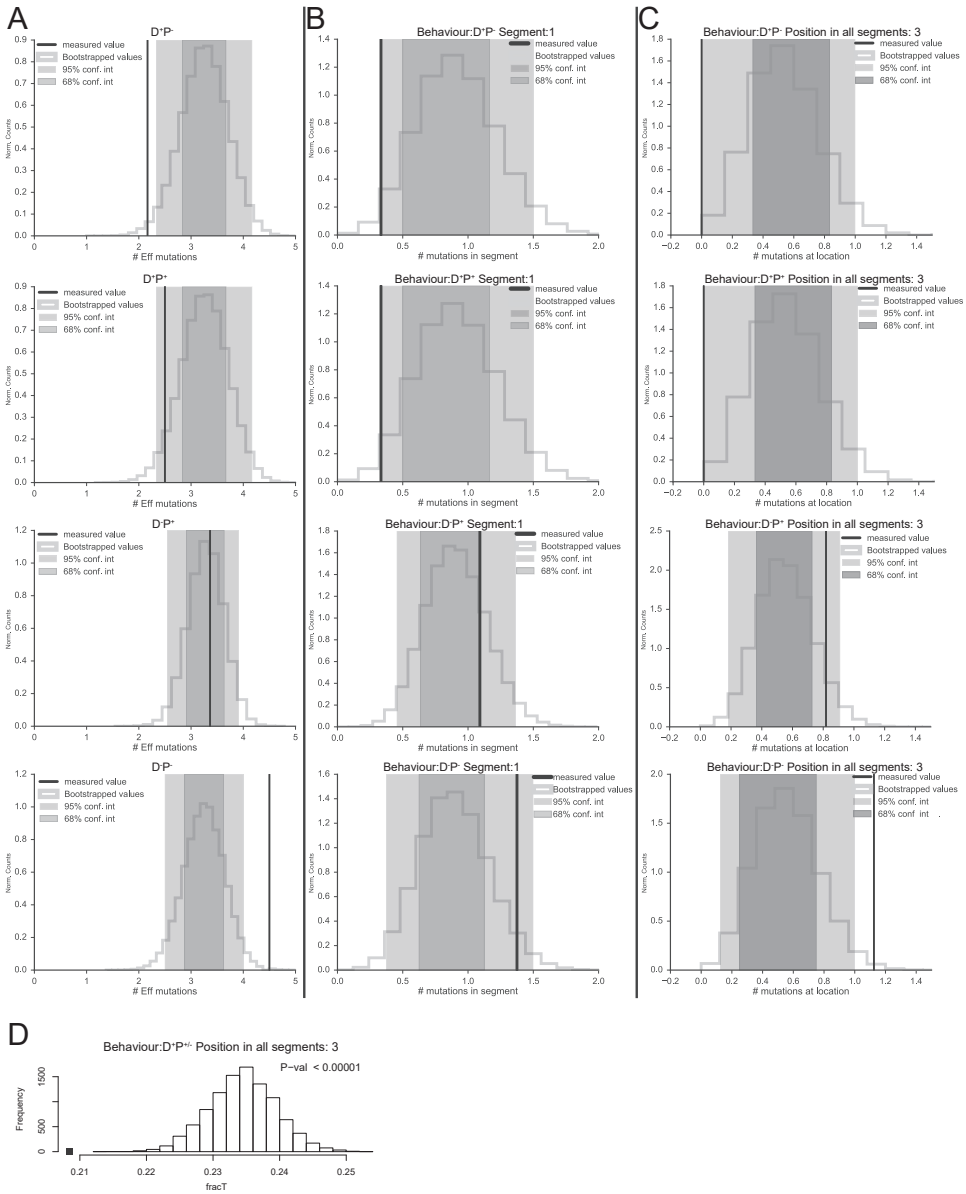


Figure 6.S3. Related to Experimental Procedures, Statistical testing: Statistical pattern analysis of 30 mutants set. Three properties were analyzed separately for each group of plasmid behavior. The average of each behavioral group is indicated by the yellow vertical line. To test if the plasmid behavior depends on a certain property, for each property a distribution was made based on empirical bootstrapping of the whole set of 30 mutants (blue line). The 95% and 68% confidence intervals of each distribution are indicated by the light and dark grey boxes respectively. **A** Average number of effective mutations. **B** Average number of mutations in segment 1. **C** Av-

erage number of mutations on position 3 within all segments combined. **D** Average number of mutations on position 3 within all segments combined but the analysis was performed on a previously published large dataset [27]. From this dataset, mutants with 3 mutations (all canonical PAM) were analyzed. The average of the direct interference group is indicated by the red square.

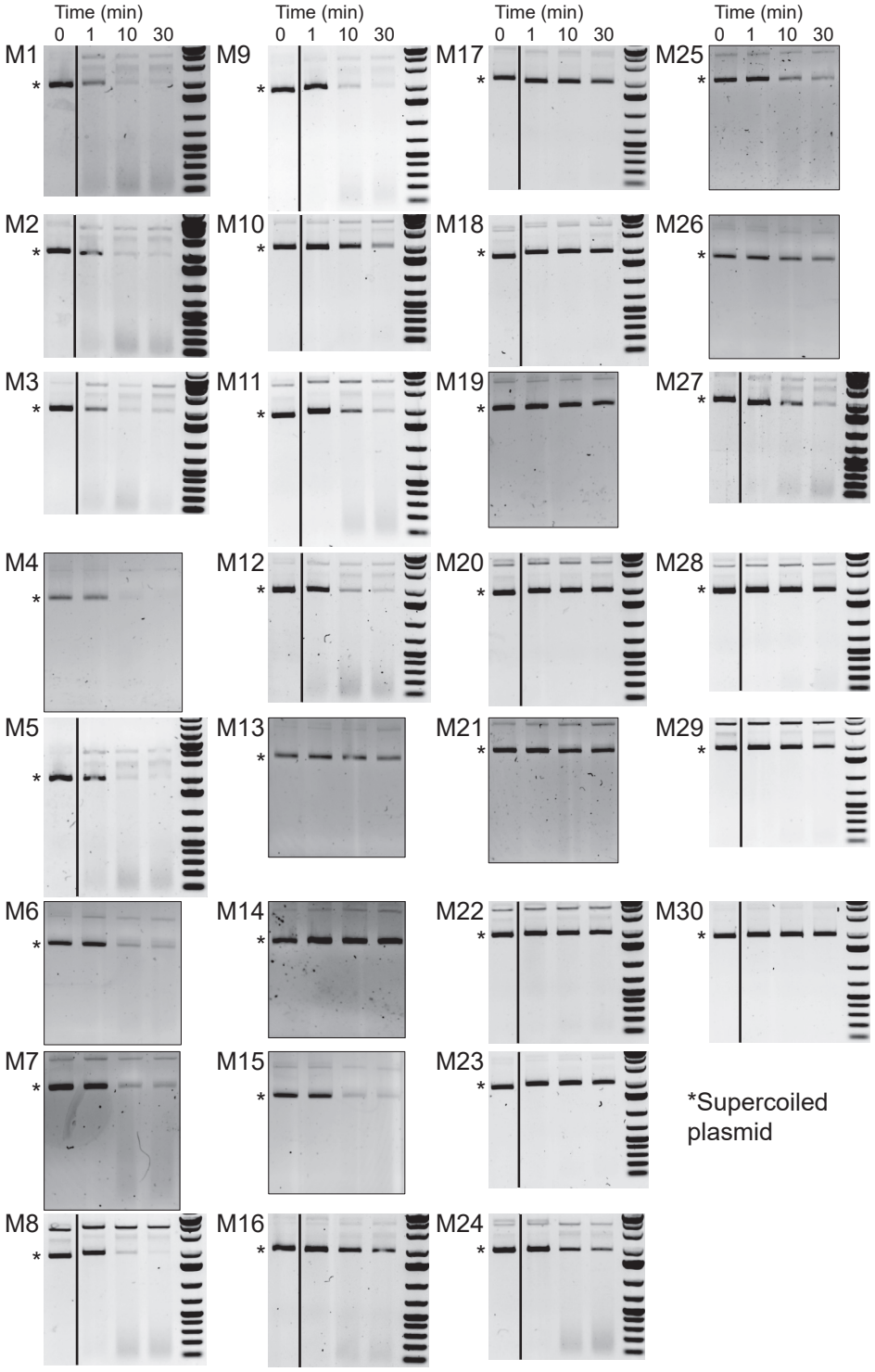

Time (min)
0 1 10 30
M1
M2
M3
M4
M5
M6
M7
M8
M9
M10
M11
M12
M13
M14
M15
M16
M17
M18
M19
M20
M21
M22
M23
M24
M25
M26
M27
M28
M29
M30
*Supercoiled plasmid

Figure 6.S4. Related to Figure 6.2B. Representative gels of Cas3 activity assays. Individual gels for each mutant showing Cas3 plasmid degradation reactions at time points 0, 1, 10, 30 minutes. Vertical black lines indicate removal of 3 gel lanes with irrelevant samples. Supercoiled plasmid is indicated with an asterisk, gel lanes above are linearized and nicked plasmids, which are not considered in quantification.

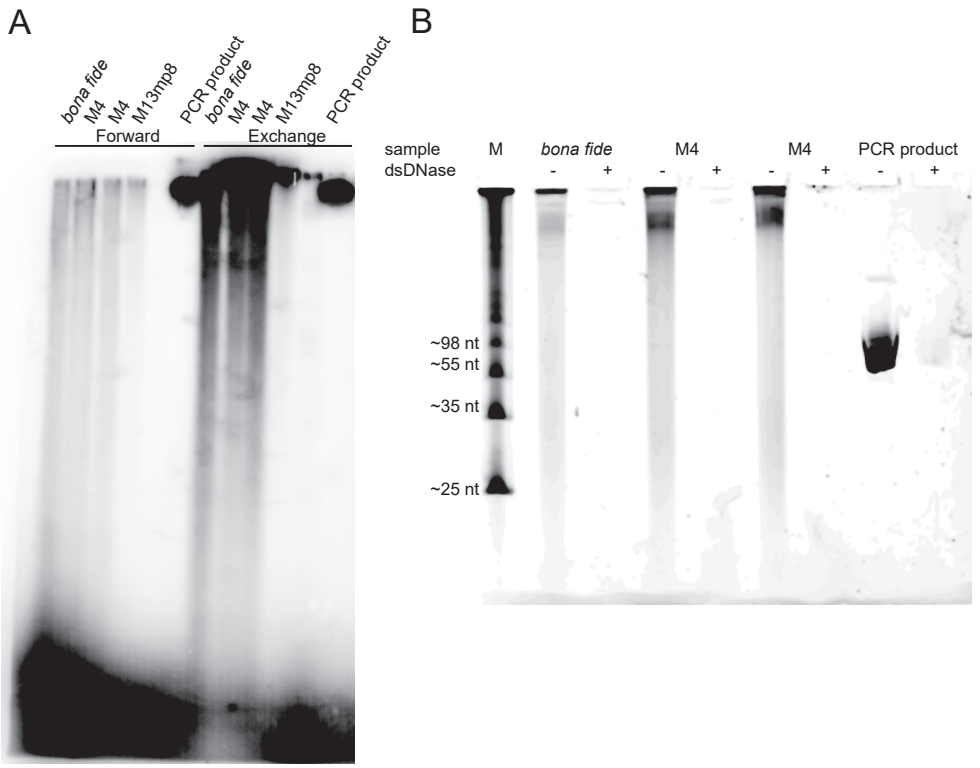


Figure 6.S5. Related to Figure 6.3: Biochemical analysis of Cas3 DNA degradation fragments. A) 32P PNK labeling of degradation fragments from bona fide target plasmid, M4 target plasmid and m13mp8 single stranded plasmid. Forward reaction can only label non-phosphorylated 5'ends, exchange reaction can label both phosphorylated and non-phosphorylated 5' ends. Non-phosphorylated PCR product for reference. B) dsDNase incubation with degradation fragments of bona fide target plasmid and M4 target plasmid. dsDNase is a double stranded DNA specific endonuclease with no activity on single stranded DNA.

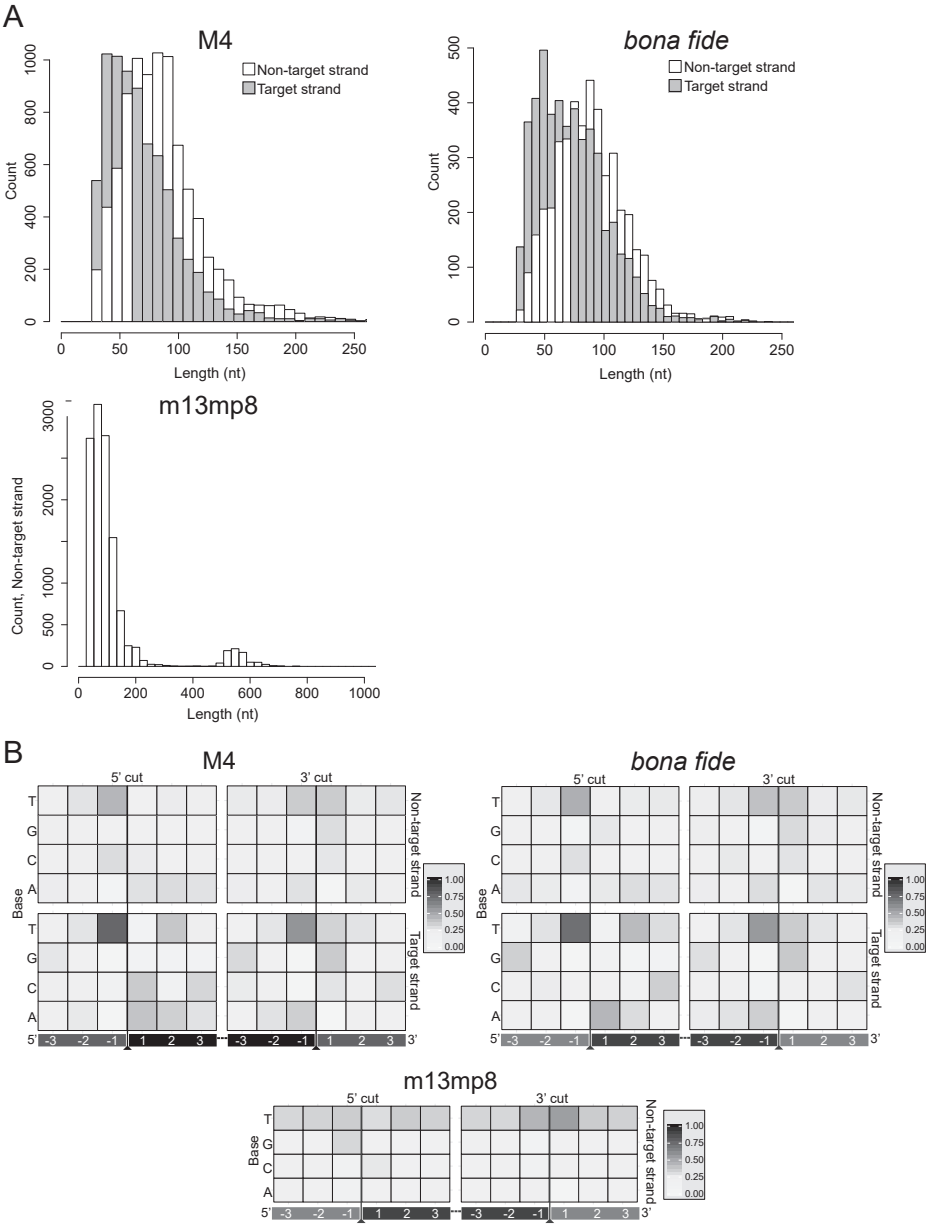

A
M4
Non-target strand
Target strand
Count
Length (nt)
bona fide
Non-target strand
Target strand
Count
Length (nt)
m13mp8
Count, Non-target strand
Length (nt)
B
M4
5' cut
3' cut
Non-target strand
Target strand
Base
T
G
C
A
5'
3'
bona fide
5' cut
3' cut
Non-target strand
Target strand
Base
m13mp8
5' cut
3' cut
Non-target strand
Base

Figure 6.S6. Related to Figure 6.3: Next generation sequencing analysis of Cas3 DNA degradation products. A) Length distribution bar charts for Cas3 DNA degradation products of bona fide target plasmid, M4 target plasmid and m13mp8 single stranded plasmid. B) Heat maps of nucleotide frequencies around cleavage sites for bona fide target plasmid, M4 target plasmid and m13mp8 single stranded plasmid. 5' and 3' cut sites are displayed separately for both target and non-target strand. The cleavage site is between position -1 and 1. Positions indicated in black are on the fragments, positions indicated in grey are outside of fragments.

SUPPLEMENTARY TABLES

Table 6.S1. Related to Figures 6.1-6.6: Oligo nucleotides used in this study

Name	Sequence	Description
BG4556	ATCCCGGGATGACCTGGCTTCCCCTT	Cas1 fw (SmaI)
BG4557	AGTGAGCTCTCAAACAGGTAAAAAAGACACC	Cas2 rv (SacI)
BG5301	AAGGTTGGTGGGTTGTTTTTATGG	CRISPR leader forward primer
BG5302	GGATCGTCACCCTCAGCAGCG	M13_g8 spacer reverse primer
BG6170	CACTCTTTCCCTACACGACGCTCTTCCGATCTGCCTAA	NGS PE 5'Adapter 3
BG6174	CACTCTTTCCCTACACGACGCTCTTCCGATCTGATCTG	NGS PE 5'Adapter 7
BG6176	CACTCTTTCCCTACACGACGCTCTTCCGATCTCTGATC	NGS PE 5'Adapter 9
BG6179	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGC	NGS PE 5'Adapter extension primer
BG6180	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC TTTTTTTTTTTTTTTTTTTTTTTTTTTTVN	NGS PE 3' Tail primer 1
BG6183	CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTCAGACGTGTG	NGS PE 3' Tail primer 2.3
BG6187	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTG	NGS PE 3' Tail primer 2.7
BG6189	CAAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGAGTTCAGACGTGTG	NGS PE 3' Tail primer 2.9
BG6713	GCTCTGCTGAAGCCAGTT	Reverse S437 hot spot pBR322
BG6714	GATCCTCTAGAGTCGACCT	Reverse S429 hot spot bb
BG6715	GCTAGTTGAACGGATCCAT	Reverse S416 hot spot GFP
BG7213	CGCTGCTGCGAAATTTGAAC	pWUR477 single repeat fw
BG7214	AACTCTGCGTGAGCGTATCG	pWUR477 single repeat rv
BG7215	ATCCGTTCAACTAGCAGACC	GFP hotspot nested forward
BG7216	GGTCTGCTAGTTGAACGGAT	GFP hotspot nested reverse
BG7415	CAATTTACTACTCGTTCTGGTGTTTCTCGTCAGGG	Protospacer 35 forward
BG7416	ACGAGAAACACCAGAACGAGTAGTAAATTGGGCTT	Protospacer 35 reverse
BG7522	CTGCGCTAGTAGACGAGTC	pWUR477 behind array reverse

Table 6.S2. Related to Figures 1-6: Plasmids used in this study

Plasmid	Description (positions of all mutations)	Name in paper	source
pWUR835	pGFP-UV Amp	-	[27]
pWUR836	pGFP-UV Km protospacer8 WT	pTarget bona fide	[27]
pWUR837	pGFP-UV Km protospacer8 mutant pos. 1, 3, 24	pTarget M14	[27]
pWUR838	pGFP-UV Km protospacer8 mutant pos. 10, 11, 25	pTarget M12	[27]
pWUR839	pGFP-UV Km protospacer8 mutant pos. 1, 4, 16	pTarget M30	[27]
pWUR840	pGFP-UV Km protospacer8 mutant pos. 2, 3, 4	pTarget M17	[27]
pWUR841	pGFP-UV Km protospacer8 mutant pos. 3, 7, 19	pTarget M26	[27]
pWUR842	pGFP-UV Km protospacer8 mutant pos. 4, 8, 26	pTarget M23	[27]
pWUR843	pGFP-UV Km protospacer8 mutant pos. 2, 10, 16	pTarget M16	[27]
pWUR844	pGFP-UV Km protospacer8 mutant pos. 2, 18, 22	pTarget M9	[27]
pWUR845	pGFP-UV Km protospacer8 mutant pos. 10, 14, 17	pTarget M5	[27]
pWUR846	pGFP-UV Km protospacer8 mutant pos. 11, 16, 17	pTarget M7	[27]
pWUR847	pGFP-UV Km protospacer8 mutant pos. 11, 22, 32	pTarget M1	[27]
pWUR848	pGFP-UV Km protospacer8 mutant pos. 5, 6, 25	pTarget M2	[27]
pWUR850	pGFP-UV Km protospacer8 mutant pos. 2, 8, 26	pTarget M10	[27]
pWUR851	pGFP-UV Km protospacer8 mutant pos. 19, 27, 32	pTarget M27	[27]
pWUR852	pGFP-UV Km protospacer8 mutant pos. 12, 17, 31	pTarget M3	[27]
pWUR853	pGFP-UV Km protospacer8 mutant pos. 6, 7, 32	pTarget M4	[27]
pWUR854	pGFP-UV Km protospacer8 mutant pos. 1, 10, 15, 18, 29	pTarget M25	[27]
pWUR855	pGFP-UV Km protospacer8 mutant pos. 1, 16, 19, 25, 29	pTarget M13	[27]
pWUR856	pGFP-UV Km protospacer8 mutant pos. 1, 4, 19, 27, 28	pTarget M20	[27]
pWUR857	pGFP-UV Km protospacer8 mutant pos. 2, 12, 23, 26, 27	pTarget M11	[27]
pWUR859	pGFP-UV Km protospacer8 mutant pos. 3, 8, 10, 11, 22	pTarget M29	[27]
pWUR860	pGFP-UV Km protospacer8 mutant pos. 3, 15, 20, 25, 26	pTarget M18	[27]
pWUR859	pGFP-UV Km protospacer8 mutant pos. 3, 9, 13, 22, 26	pTarget M19	[27]
pWUR860	pGFP-UV Km protospacer8 mutant pos. 5, 6, 8, 24, 31	pTarget M8	[27]
pWUR861	pGFP-UV Km protospacer8 mutant pos. 4, 5, 6, 15, 24	pTarget M24	[27]

pWUR862	pGFP-UV Km protospacer8 mutant pos. 1, 2, 9, 14, 21	pTarget M22	[27]
pWUR863	pGFP-UV Km protospacer8 mutant pos. 6, 22, 27, 31, 32	pTarget M28	[27]
pWUR864	pGFP-UV Km protospacer8 mutant pos. 12, 13, 23, 24, 30	pTarget M6	[27]
pWUR866	pGFP-UV Km protospacer8 mutant pos. 3, 9, 12, 16, 32	pTarget M21	[27]
pWUR867	pGFP-UV Km protospacer8 mutant pos. 17, 27, 28, 29, 30	pTarget M15	[27]
pWUR868	pGFP-UV Km non-target	pTarget NT	[27]
pWUR748	pMAT11-MBP-Cas3		[34]
pWUR868	pACYC poly spacer8 CRISPR array		This study
pWUR514	cse2 with Strep-tag II (N-term)-*cas7-cas5-cas6*e in pET52b		[65]
pWUR408	cse1 in pRSF-1b, no tags		[68]
pWUR477	pACYC with artificial CRISPR array		[68]
pWUR872	pWUR477 with only one repeat	pCRISPR	This study
pWUR871	Cas1-Cas2 operon with Strep-tag II (N-term) in pET52b		This study

Table 6.S3. Related to Figure 2A: EMSA data from regression analysis

Plasmid	Amplitude	Kd (nM)	Amplitude/Kd
bona fide (WT)	1.0 ± 0.01	7.6 ± 0.8	1.31E-01
M1	0.85 ± 0.01	23.6 ± 2.0	3.59E-02
M2	0.92 ± 0.04	23.6 ± 4.6	3.92E-02
M3	0.99 ± 0.02	18.5 ± 2.7	5.35E-02
M4	1.02 ± 0.04	16.4 ± 3.34	6.23E-02
M5	0.87 ± 0.03	34.3 ± 5.3	2.54E-02
M6	0.0	--	0.00E+00
M7	0.69 ± 0.01	31.6 ± 2.7	2.17E-02
M8	0.65 ± 0.01	17.4 ± 2.0	3.71E-02
M9	0.94 ± 0.03	24.8 ± 4.7	3.78E-02
M10	1.05 ± 0.05	23.4 ± 5.3	4.50E-02
M11	0.39 ± 0.02	22.1 ± 6.0	1.77E-02
M12	0.0	--	0.00E+00
M13	0.0	--	0.00E+00
M14	1.2 ± 0.13	360 ± 79.4	3.46E-03
M15	0.46 ± 0.01	4.4 ± 0.4	1.04E-01
M16	0.78 ± 0.02	46.3 ± 6.7	1.69E-02
M17	1.19 ± 0.02	152.6 ± 10.0	7.79E-03
M18	0.0	--	0.00E+00
M19	0.0	--	0.00E+00
M20	0.0	--	0.00E+00
M21	0.0	--	0.00E+00
M22	0.94 ± 0.01	55.9 ± 2.7	1.69E-02
M23	0.69 ± 0.02	54.1 ± 5.3	1.27E-02
M24	0.9 ± 0.03	22.4 ± 4.0	4.03E-02
M25	0.31 ± 0.01	34.6 ± 6.0	9.02E-03
M26	0.93 ± 0.03	79.4 ± 8.7	1.17E-02
M27	0.74 ± 0.02	20.7 ± 2.7	3.59E-02
M28	1.04 ± 0.04	17.4 ± 3.3	5.97E-02
M29	0.4 ± 0.02	74.2 ± 18.0	5.40E-03
M30	0.0	--	0.00E+00

Table 6.S4. Related to Figure 3: NGS data processing and mapping

Sample name	Total number of reads	Reads mapping to NT strand	Reads mapping to NT strand (%)	Reads mapping to T strand	Reads mapping to T strand (%)
bona fide (WT)	215218	57217	26.6	158001	73.4
M4	101327	23334	23	77993	77
M13mp8	46205	46109	>0.99	96	<0.01

Table 6.S5. Related to Figure 3: NGS data processing for cleavage sites

	Non-target strand (NT)			Target strand (T)		
Sample name	# Distinct Fragments	# Distinct Start	# Distinct End	# Distinct Fragments	# Distinct Start	# Distinct End
bona fide (WT)	8777	1381	1479	7448	1318	1151
M4	4432	971	1076	4784	1029	920
M13mp8	12243	3737	2620			

Appendix

CRISPR-Cas systems reduced to a minimum

Molecular Cell
2019 Feb 21;73(4):641-642.

Cristóbal Almendros[1], Sebastian N. Kieper[1], Stan J.J. Brouns[1,2]

1. Kavli Institute of Nanoscience, Department of Bionanoscience, Delft University of Technology, Van der Maasweg 9, 2629 HZ Delft, the Netherlands.

2. Laboratory of Microbiology, Wageningen University, Stippeneng 4, 6708 WE Wageningen, the Netherlands.

Summary

In this issue of Molecular Cell, Wright et al. (2019) report complete spacer integration by a Cas1 mini-integrase and Edraki et al. (2019) describe accurate genome editing by a small Cas9 ortholog with less stringent PAM requirements.

Preview

Even though CRISPR-Cas systems were identified more than a decade ago, scientists continue to find new variations that shed more light on the mechanism and evolution of these fascinating prokaryotic immune systems. CRISPR loci are composed of clusters of direct repeats separated by unique sequences (spacers) that make up the genetic memory of prokaryotes [1]. The spacers encode the CRISPR RNAs (crRNAs) ultimately guiding Cas effector complexes towards the catalogued mobile genetic element (MGE) [2]. Maintaining immunity is therefore critically reliant on a core step, termed adaptation or spacer acquisition, and involves the appropriate selection and subsequent integration of invader-derived DNA fragments into the CRISPR array. It has been well-established that two proteins, Cas1 and Cas2, form a heterohexameric complex that catalyzes a two-step reaction resulting in integration of the invader DNA fragment into the CRISPR array.

In this issue of Molecular Cell, Wright et al. (2019) report the exception to that rule. The authors identify two type V CRISPR-Cas systems in metagenomes from a mouse and beetle gut that both lack the *cas2* gene. Interestingly, the spacer length found in both systems is considerably smaller (18-20 bp) than in all other CRISPR-Cas systems harboring *cas2*. Wright et al. present compelling *in vitro* evidence that Cas1 proteins from both systems support spacer integration using purified Cas1 and plasmid DNA containing the cognate type V CRISPR array. The type V-C Cas1 (Cas1c) displayed intrinsic preference for prespacers of just 18 bp in length, which is almost half the size of most spacers in CRISPR systems encoding Cas2. Deep sequencing of integration events revealed a surprisingly high frequency of off-target spacer integration into non-CRISPR loci, which could indicate the requirement of a yet unknown host factor or sequence element for specificity in the natural host. The authors further back their findings by employing plasmids reporting successful integration of new spacers [3] and made the interesting observation that Cas1 intrinsically deter-

mines spacer orientation and defines the location of integration depending on the protospacer sequence. Size-exclusion chromatography and mass spectrometry revealed the formation of a tetrameric Cas1 assembly in association with mimics of spacer integration intermediates.

Interestingly, phylogenetic analysis has found that Cas1c and Cas1d are related to stand-alone *cas1* genes, supporting the hypothesis that those minimal integrases are closely related to the ancestral CRISPR-associated Cas1 and Casposons [4]. Therefore, the exciting evolutionary scenario emerges that primitive *cas*-gene clusters accommodated a *cas2* gene to increase spacer size by widening the span of the Cas1 tetrameric 'butterfly-like' assembly of the adaptation complex. Possibly the selection of longer spacers led to larger crRNAs with higher target specificity.

This novel adaptation complex constituted by only one protein raises new research questions. How does Cas1c accomplish accurate spacer integration in CRISPR arrays in its native host? Can these short spacers provide CRISPR immunity? In addition, how does this mini integrase select and process spacers with a correct protospacer adjacent motif (PAM) [5]. The PAM is a sequence required for target recognition and recently a Cas12c variant has been described with less stringent PAM requirements (i.e. 5'-TN-3') [6]. Although it remains to be determined, this less restrictive PAM could mean that there is no PAM selection during spacer integration, as 25% of newly integrated spacers would already confer functional CRISPR immunity.

Besides insights into prokaryotic immunity, the identification of novel Cas effector proteins with less restrictive PAM requirements is important for expanding the biotechnological CRISPR-Cas toolbox. Moreover, viral delivery of genome editing nucleases into target cells would benefit from compact yet accurate effector proteins. In this issue of Molecular Cell, Edraki et al. (2019) report a type II-C nuclease effector protein Cas9 (Nme2Cas9) from *Neisseria meningitides* with a high potential for genome editing applications due to its compact size and minimal PAM sequence requirements.

Although *Streptococcus pyogenes* Cas9 (SpyCas9) is currently the most popular genome editing platform, its gene size of 4.2 kb is close to the 4.5 kb packaging limit of recombinant adeno-associated virus delivery vectors (rAAV). More compact Cas9 orthologues exist, but their relatively complex PAM requirements restrict applications by limiting

suitable targeting sites. The Nme2Cas9 orthologue described by Edraki et al. combines less restrictive dinucleotide PAM requirements with a packaging-compatible compact size of 3.2 kb enabling viral delivery in a single virus particle. The authors demonstrate the potential of Nme2Cas9 to induce non-homologous end joining (NHEJ) repair and Nme2Cas9D16A (HNH nickase) for homology-directed repair (HDR) events in human cells based on a fluorescent readout resulting from open reading frame restoration. With an optimal crRNA spacer size of 22-24 nt Nme2Cas9 is able to edit different mammalian cell types. Moreover, Nme2Cas9 activity can be controlled through the use of anti-CRISPR proteins (Acrs) from four Acr-families [7]. Despite the clear advantage of less stringent PAM requirements of the Nme2Cas9 protein, this feature also raises the question of whether the less stringent PAM requirements would decrease its on-target specificity, and could therefore contribute to undesirable off-target editing events. The authors specifically address this potential issue by performing GUIDE-seq analysis (genome-wide unbiased identification of double-stranded breaks enabled by sequencing) [8] and reveal highly accurate genome editing by Nme2Cas9.

Finally, in order to test the ability of Nme2Cas9 to carry out genome editing in a living multicellular organism, an all-in-one AAV vector was designed and delivered into adult mice. The gene encoding Pcsk9 was chosen as a suitable candidate for in vivo genome editing since it is a regulator of circulating cholesterol homeostasis [9]. Mice infected with the specific Pcsk9-crRNA showed a reduction of cholesterol levels, suggesting that the edit was successful in sufficient numbers of cells to see a phenotypic effect. Furthermore, to demonstrate the ex vivo editing potential, zygotes were modified by disrupting the pathway leading to the production of melanin. The reimplantation of modified zygotes into pseudopregnant females resulted in a population of albino pups confirming that the genome edit was again successful, establishing this Cas9 variant as a new genome editing platform.

CRISPR-Cas systems have evolved by incorporating and diversifying their genes. This stunning diversity of prokaryotic defense systems enables us to discover more Cas proteins with limitless potential in molecular applications and biotechnology. The quest for new CRISPR-Cas systems will continue to expand our knowledge of the bacteriophage/host arms race with all its implications for fundamental and applied sciences.

REFERENCES

1. Jackson, S.A., R.E. McKenzie, R.D. Fagerlund, S.N. Kieper, P.C. Fineran, and S.J. Brouns, CRISPR-Cas: Adapting to change. Science, 2017. 356(6333).

2. Mohanraju, P., K.S. Makarova, B. Zetsche, F. Zhang, E.V. Koonin, and J. van der Oost, Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. Science, 2016. 353(6299): p. aad5147-aad5147.

3. Diez-Villasenor, C., N.M. Guzman, C. Almendros, J. Garcia-Martinez, and F.J. Mojica, CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of *Escherichia coli*. RNA Biol, 2013. 10(5): p. 792-802.

4. Makarova, K.S., Y.I. Wolf, and E.V. Koonin, Classification and Nomenclature of CRISPR-Cas Systems: Where from Here? The CRISPR Journal, 2018. 1(5): p. 325-336.

5. Mojica, F.J.M., C. Díez-Villaseñor, J. García-Martínez, and C. Almendros, Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology, 2009. 155(3): p. 733-740.

6. Yan, W.X., P. Hunnewell, L.E. Alfonse, J.M. Carte, E. Keston-Smith, S. Sothiselvam, A.J. Garrity, S. Chong, K.S. Makarova, E.V. Koonin, D.R. Cheng, and D.A. Scott, Functionally diverse type V CRISPR-Cas systems. Science, 2019. 363(6422): p. 88-91.

7. Lee, J., A. Mir, A. Edraki, B. Garcia, N. Amrani, H.E. Lou, I. Gainetdinov, A. Pawluk, R. Ibraheim, X.D. Gao, P. Liu, A.R. Davidson, K.L. Maxwell, and E.J. Sontheimer, Potent Cas9 Inhibition in Bacterial and Human Cells by AcrIIC4 and AcrIIC5 Anti-CRISPR Proteins. mBio, 2018. 9(6).

8. Tsai, S.Q., Z. Zheng, N.T. Nguyen, M. Liebers, V.V. Topkar, V. Thapar, N. Wyvekens, C. Khayter, A.J. Iafrate, L.P. Le, M.J. Aryee, and J.K. Joung, GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. Nat Biotechnol, 2015. 33(2): p. 187-197.

9. Rashid, S., D.E. Curtis, R. Garuti, N.N. Anderson, Y. Bashmakov, Y.K. Ho, R.E. Hammer, Y.A. Moon, and J.D. Horton, Decreased plasma cholesterol and hypersensitivity to statins in mice lacking Pcsk9. Proc Natl Acad Sci U S A, 2005. 102(15): p. 5374-9.

Summary - CRISPR's little Helpers: CRISPR-Cas Proteins Involved in PAM selection

For millennia, humanity has been plagued by pathogenic bacteria. Until the advent of antibiotic treatments, seemingly harmless bacterial infections could have fatal consequences. However, in the microcosm that these single celled organisms inhabit, the line between being the invader or being invaded is a thin line. Bacteria and archaea are constantly targeted by their viruses (bacteriophages – from Greek "to devour"-bacteria). Without mechanisms in place to protect the prokaryotic cell from infection, bacteriophages would drive whole species to almost extinction. This thesis presents the work in which we applied techniques of molecular biology and biochemistry to investigate the mechanism certain bacterial species use to develop immunity against bacteriophages.

In **Chapter 1** we introduce the adaptive bacterial immune system that utilizes Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) in connection with CRISPR-associated (Cas) proteins. Prokaryotes can memorize bacteriophages or other mobile genetic elements (MGEs) that invade their cells by integrating short pieces of the MGE into the CRISPR arrays that are located in the prokaryotic genome. This is the first of the three stages of CRISPR-Cas immunity and is called adaptation. During adaptation, specialized Cas proteins select fragments of the MGE, process the fragments according to certain specifications and integrate them into the CRISPR locus. Upon integration into the genomic CRISPR locus, this short piece of phage DNA becomes a so-called spacer and together with the remaining CRISPR array, serves as a template for the second stage – expression. During expression, the CRISPR-array is transcribed into a long precursor RNA composed of the palindromic repeats and spacer sequences. This precursor is cut into smaller units containing parts of the repeat and the full spacer sequence, the CRISPR-RNA (crRNA). The mature crRNA is incorporated into Cas protein complexes that patrol the cell and try to match the sequence of the crRNA with potential invader DNA. This last stage is termed interference. When crRNA-Cas complexes find a DNA sequence that has been catalogued in the CRISPR array, the complex binds the recognized DNA and causes its destruction.

The specific CRISPR-Cas mechanism may vary between differ-

ent classes, types and subtypes, however, the three stages are overall shared among all of them.

For efficient interference to occur it is crucial that invading DNA is quickly and selectively memorized. In **Chapter 2** of this thesis we pay special attention to the first stage of CRISPR-Cas immunity and review the current understanding of the molecular mechanisms that govern the adaptation process. The workhorse of CRISPR-adaptation consists of a protein complex composed of the Cas1 and Cas2 proteins. This adaptation complex forms a structure that resembles the shape of a butterfly. With its wing-like structure, the complex' main purpose is the capture, the processing and the correct integration of invading DNA into the CRISPR array. In the large majority of CRISPR-Cas systems a short sequence motif adjacent to the sequence where the crRNA base-pairs is crucial for interference to occur (the target strand to which the crRNA hybridizes is termed protospacer). This short sequence motif is called a protospacer adjacent motif (PAM). The adaptation complex must be able to capture DNA sequences that contain PAM sequences in order to generate interference-proficient spacers. Next to that, the captured sequence must be processed accurately with respect to the position of the PAM and also integrated in a correct orientation. ~~Not an easy task for the adaptation complex.~~ Although in some systems Cas1 and Cas2 alone are sufficient to execute this described function by being able to recognize the PAM, many systems require additional integration factors that aid in this process.

In **Chapter 3** of this thesis we investigate one of those additional integration factors. The *cas4* gene, albeit among the first *cas* genes discovered, remained mysterious in its function. Prior research indicated that the Cas4 protein might directly contribute to adaptation by interacting with the Cas1 protein and indeed, deletion of the *cas4* gene abrogated adaptation in some CRISPR-Cas systems. We set out to gain a deeper understanding of what the exact role of Cas4 in adaptation might be. We took the CRISPR-Cas type I-D adaptation module of the cyanobacterium *Synechocystis* sp. 6803 consisting of the *cas1, cas2* and *cas4* genes and transferred it into our *Escherichia coli* model system. In addition, we supplied a minimalized CRISPR locus into which the adaptation module could integrate novel spacers. We used a tailormade polymerase chain reaction (PCR) approach for the detection of newly integrated spacers. Contrary to our initial expectations, the Cas4 protein was not necessary for acquiring new spacers.

~~Cas1 and Cas2 seemed to do just fine without Cas4.~~ However, when we analyzed the sequences of spacers that were acquired either in the presence or absence of Cas4, we found a clear difference between the two conditions. Spacers that were acquired without Cas4 were selected randomly and would not support the interference stage. In contrast, the spacers that were taken up in the presence of Cas4 were selected according to the PAM requirements of the type I-D system and hence would convey protection to the cell. Altogether these initial findings would explain why the cas4 gene is so universally conserved among many different CRISPR-Cas systems: Cas4 facilitates PAM-compatible spacer selection during CRISPR adaptation.

These observations were helpful in understanding why Cas4 is present in so many diverse systems, but did not explain how Cas4 contributes to adaptation mechanistically. In **Chapter 4** of this thesis we biochemically reconstitute the adaptation complex of the type I-D system. We overexpressed the *cas4* gene together with *cas1* and *cas2* and started fishing for Cas4 to see which other proteins might interact with it. Interestingly, but not surprisingly, we found that the core adaptation protein Cas1 strongly interacts with Cas4. When we made Cas1 with it having the choice between its usual interaction partner Cas2 or the Cas4 protein, Cas1 would choose to interact with Cas4. This finding leads to the understanding that two mutually exclusive adaptation complexes exist, the Cas4-Cas1 and the Cas1-Cas2 complex. We purified both complexes and conducted spacer integration experiments *in vitro* in order to understand their distinct functions. Cas1-Cas2 and Cas4-Cas1 complexes were both able to integrate new spacers into CRISPR loci, which is in line with our *in vivo* findings described in Chapter 3. However, sequence specific recognition and processing of PAM motifs within single-stranded DNA overhangs of spacer precursors was only accomplished by the Cas4-Cas1 complex. While PAM-containing spacer precursors were integrated without additional processing by the Cas1-Cas2 complex, the PAM sequence was accurately removed prior to integration by the Cas4-Cas1 complex. Our combined findings result in a model in which an asymmetric adaptation complex differentially acts on PAM and non-PAM containing overhangs, providing cues for the correct orientation of spacer integration.

In **Chapter 5** of this thesis we investigate sequence motifs located in the so-called *leader* sequence. The *leader* is located upstream of the

CRISPR array and contains promoter elements that drive transcription of the CRISPR. The previous chapters focus on the capture and processing of prespacers, but do not explain how the adaptation machinery is able to localize the correct integration site. Spacers are being integrated in chronological order at the leader-site of the CRISPR array and hence represent a history of viral infections. This feature makes sure that the immune response is directed against bacteriophages that circulated recently. By using alignments between different type I-D *leader* sequences, we discover three conserved sequence motifs that are crucial for efficient adaptation. We progressively delete parts of the *leader* and subsequently quantify spacer integration activity with those shortened leader sequences. The first 30 basepairs of the *leader* contain a conserved region of which deletion results in a substantial reduction of spacer integration activity. Interestingly, the region of which deletion induces the strongest reduction of spacer integration contains a sequence motif resembling the so-called Integrase Anchoring Site (IAS). Previous studies have suggested that this IAS stabilizes the interaction of the Cas1-Cas2 adaptation complex with the CRISPR array and thereby increases the efficiency of spacer integration. Our study leads to a compatible model in which conserved sequences within the type I-D *leader* sequence act as landmarks for the adaptation complex and thereby provide cues for the correct integration site.

While we until here investigated the involvement of the Cas4 protein in naïve adaptation (naïve meaning that the invader being memorized has not been catalogued in CRISPR arrays before), in this **Chapter 6** we demonstrate how interference and adaptation are interconnected in an adaptation pathway called priming. In primed adaptation the invader from which spacers are being acquired was already catalogued in past infections. However, through mutations in the sequence of the invader, the interference stage cannot take place as efficiently as with full complementarity between crRNA and protospacer. In this case primed adaptation is a powerful mechanism that allows restoration of efficient interference. We investigated the Type I-E CRISPR-Cas system of *Escherichia coli* and found that the Cas3 protein is the key player that couples interference to adaptation. The Cas3 protein is recruited by the type I-E interference complex and is responsible for the degradation of invading DNA through its nuclease-helicase activity. We observed that the rate at which Cas3 degraded target DNA depended on the mutations present in the protospacer and the PAM.

Full matches between crRNA and protospacer and a consensus PAM resulted in fast degradation of target DNA and did not trigger priming. However, some protospacer mutations slowed down Cas3 activity and thereby created short DNA fragments that lead to efficient primed adaptation. We analyzed the DNA fragments that resulted from Cas3 mediated degradation and found that they were of appropriate length to serve as spacer precursors. Importantly, they were enriched for the type I-E specific PAM sequence and hence well-suited substrates for capture and integration by the Cas1-Cas2 adaptation complex. Altogether this work demonstrates how certain CRISPR-Cas systems can maintain immunity by updating their memory through an intricate feedback loop that couples the efficiency of invader destruction to the uptake of novel spacer units.

The results of this thesis hold significance for the fundamental understanding of how CRISPR-Cas immunity is established on a molecular level. Cas proteins involved in adaptation and interference pathways co-evolved in order to protect the prokaryotic organism from viral predation by either creating antiviral memory *de novo* or by reinforcing and strengthening pre-existing immunity. Our experimental insights elucidate the role of the Cas4 protein as a dedicated adaptation co-factor that enables selection, processing and integration of spacers that support efficient invader destruction. Furthermore, the sequence context in which the CRISPR is embedded aids the spacer integration process by providing cues for the correct integration site. These features evolved as a result of the predatory pressure that the organisms carrying the defense system are constantly exposed to. However, also MGEs evolve in response to being targeted by CRISPR-Cas immunity. By mutating CRISPR-Cas target sites, viruses can escape from the interference pathway. Our findings of how the Cas3 helicase-nuclease boosts a memory update highlights the ingenious strategies that bacterial and archaeal survival relies on. Altogether this thesis provides insights into a microcosm in which the hunter quickly becomes the hunted and only a few nucleotide differences can make the difference between winning or losing the race against entropy.

SAMENVATTING – CRISPR'S KNECHTJES – CRISPR-CAS EIWITTEN BETROKKEN BIJ PAM SELECTIE

Al millennia lang wordt de mensheid geplaagd door pathogene eencellige organismen. Tot de komst van antibiotische behandelingen konden schijnbaar onschadelijke bacteriële infecties fatale gevolgen hebben. In de microkosmos, de plaats waar eencellige organismen floreren, is de grens tussen aanvaller en slachtoffer minder duidelijk. Bacteriën en archaea zijn alsmaar doelwit van hun virussen (de zogenaamde bacteriofagen, Grieks voor 'bacterie-eter'). Zonder mechanismen om zichzelf te beschermen tegen infectie zouden de bacteriofagen hele soorten eencelligen met uitsterven bedreigen. In dit proefschrift hebben we technieken uit de moleculaire biologie en biochemie gebruikt om het mechanisme te onderzoeken dat bepaalde bacteriesoorten gebruiken om immuniteit tegen bacteriofagen te ontwikkelen.

In **Hoofdstuk 1** introduceren we het adaptieve bacteriële immuunsysteem dat gebruikmaakt van Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) tezamen met CRISPR-associated (Cas) eiwitten. Prokaryoten kunnen geheugen opslaan van bacteriofagen of andere mobiele genetische elementen (MGE's) die hun cellen binnendringen. Ze doen dat door korte stukjes van de MGE te integreren in de CRISPR-arrays die zich in het prokaryotische genoom bevinden. Dit is de eerste van de drie fasen van CRISPR-Cas-immuniteit en wordt "adaptatie" genoemd. Tijdens dit proces selecteren gespecialiseerde Cas-eiwitten fragmenten van de MGE die ze vervolgens volgens bepaalde specificaties verwerken en integreren in de CRISPR locus. Na integratie in de CRISPR locus wordt dit korte stukje faag DNA een "spacer" genoemd. De spacer wordt dan tezamen met de rest van de CRISPR-array gebruikt voor de volgende fase – expressie. Tijdens de expressie wordt de CRISPR array getranscribeerd in een lang precursor-RNA, waarvan de sequentie bestaat uit palindromen herhalingen en spacers. Deze precursor wordt in kleinere eenheden gesneden, het zogenaamde CRISPR-RNA (crRNA), die een gedeelte van de herhaling en een volledige spacer bevatten. Het crRNA wordt opgenomen in Cas eiwitcomplexen om vervolgens de cel te patrouilleren. Als ze een stuk DNA in de cel vinden dat een match heeft met het crRNA betekend dit dat er mogelijk een indringer aanwezig is die

eerder is gecatalogiseerd in de CRISPR array. Dit luidt de laatste fase in: interferentie. Hier wordt de door de crRNA-Cas complex herkende DNA-sequentie vernietigt om de infectie te stoppen. Het specifieke CRISPR-Cas mechanisme kan variëren tussen verschillende klassen, typen en subtypen, maar de drie fasen (adaptatie, expressie, interferentie) zijn universeel.

Voor efficiënte interferentie is het cruciaal dat het binnendringende DNA snel en selectief wordt onthouden. In **Hoofdstuk 2** van dit proefschrift besteden we aandacht aan de eerste fase van CRISPR-Cas immuniteit en bespreken we de huidige kennis van de moleculaire mechanismen die het adaptatieproces sturen. Het werkpaard van CRISPR adaptatie bestaat uit een eiwitcomplex dat is samengesteld uit de eiwitten Cas1 en Cas2. De structuur van dit complex heeft de vorm van een vlinder; het hoofddoel van de vlindervleugels is het vangen, verwerken en correct integreren van binnendringend DNA in de CRISPR array. In de overgrote meerderheid van de CRISPR-Cas systemen is er, grenzend aan het basenparende crRNA, een korte sequentie die cruciaal is voor het optreden van interferentie (de streng waarmee het crRNA hybridiseert wordt protospacer genoemd). Dit motief is de protospacer adjacent motif (PAM). Het adaptatiecomplex moet DNA sequenties vangen die PAMs bevatten, wil het spacers genereren die in staat zijn tot interferentie. Daarnaast moet de sequentie nauwkeurig worden verwerkt met betrekking tot de positie van de PAM en in de correcte oriëntatie geïntegreerd worden. Sommige systemen hebben aan Cas1 en Cas2 voldoende om deze functie te vervullen, doordat ze de PAM kunnen herkennen. Veel andere systemen hebben echter aanvullende integratiefactoren nodig die helpen bij dit proces.

In **Hoofdstuk 3** van dit proefschrift onderzoeken we een van die aanvullende integratiefactoren. Het *cas4* gen, hoewel een van de eerste cas-genen die werd ontdekt, is lang mysterieus gebleven. Eerder onderzoek gaf aan dat het Cas4 eiwit mogelijk direct bijdraagt aan adaptatie door interactie met het Cas1 eiwit. Inderdaad, het verwijderen van het *cas4* gen heft de adaptatie in sommige CRISPR-Cas systemen op. We wilden een beter begrip krijgen van wat de exacte rol van Cas4 zou kunnen zijn tijdens adaptatie. Hiervoor hebben we de CRISPR-Cas type I-D adaptatiemodule van de cyanobacterium *Synechocystis* sp. 6803, bestaande uit de *cas1, cas2* en *cas4*, overgebracht naar ons *Escherichia coli* modelsysteem. Bovendien hebben we een geminimaliseerde CRISPR locus ingebouwd waarin het adaptatiecomplex

nieuwe spacers zou kunnen integreren. Voor de detectie van nieuw geïntegreerde spacers gebruikten we een op maat gemaakt polymerasekettingreactie (PCR) methode. In tegenstelling tot onze aanvankelijke verwachtingen was het Cas4 eiwit niet nodig om nieuwe spacers te verwerven; ~~Cas1 en Cas2 leken het prima te doen zonder Cas4~~. Toen we echter de sequenties van spacers analyseerden, verkregen in de aanwezigheid of afwezigheid van Cas4, vonden we een duidelijk verschil: spacers die werden verkregen zonder Cas4 werden willekeurig geselecteerd zonder dat ze functioneel konden zijn in het interferentie stadium. De spacer die werden opgenomen in de aanwezigheid van Cas4 werden daarentegen wel geselecteerd volgens de PAM vereisten van het type I-D-systeem en konden daarom bescherming bieden voor de cel. Deze eerste bevindingen verklaren waarom het *cas4* gen zo universeel geconserveerd is in veel verschillende CRISPR-Cas systemen: Cas4 faciliteert de selectie van PAM compatibele spacers tijdens CRISPR adaptatie.

Hoewel deze observaties hielpen om te begrijpen waarom Cas4 in zoveel verschillende systemen aanwezig is, het verklaarde niet hoe Cas4 mechanistisch bijdraagt aan adaptatie. In **Hoofdstuk 4** van dit proefschrift reconstrueren het adaptatiecomplex van het type I-D systeem middels biochemische technieken. We brachten het Cas4 eiwit samen met Cas1 en Cas2 tot expressie en keken welke eiwitten meegetrokken werden door Cas4, wat zou duiden op interactie. Hoewel interessant, het was niet verrassend dat bleek dat het adaptatie eiwit Cas1 sterk interacteert met Cas4. Toen we Cas1 de keuze gaven met zijn gebruikelijke interactiepartner Cas2 of het Cas4 eiwit te interacteren, koos Cas1 het Cas4. Deze bevinding leidde tot het inzicht dat er twee adaptatiecomplexen bestaan: het Cas4-Cas1 en het Cas1-Cas2 complex. We hebben beiden complexen gezuiverd en hebben in vitro spacer-integratie experimenten uitgevoerd om hun verschillende functies te begrijpen. Cas1-Cas2- en Cas4-Cas1-complexen waren beide in staat om nieuwe spacers te integreren in de CRISPR locus, wat in lijn is met onze in vivo bevindingen beschreven in Hoofdstuk 3. Sequentie specifieke herkenning en verwerking van PAM motieven in enkelstrengs DNA uiteindes van pre-spacers werd alleen bereikt door het Cas4-Cas1 complex. De PAM-bevattende pre-spacers werden geïntegreerd zonder aanvullende verwerking door het Cas1-Cas2 complex, maar de PAM sequentie werd nauwkeurig verwijderd voorafgaand aan integratie door het Cas4-Cas1-complex. Deze bevindingen resulter-

en in een model waarin een asymmetrisch adaptatiecomplex anders werkt op PAM uiteindes vergelijken met uiteindes zonder PAM, wat aanwijzingen geeft voor de juiste oriëntatie van de spacer integratie.

In **Hoofdstuk 5** van dit proefschrift onderzoeken we sequentiemotieven die zich in de zogenaamde leadersequentie bevinden. De leadersequentie bevindt zich voor de CRISPR array en bevat promotorelementen die de transcriptie van de CRISPRs aansturen. De voorgaande hoofdstukken zijn gericht op het opvangen en verwerken van pre-spacers, maar gaat niet in op hoe de adaptatiemachine in staat is om de juiste integratielocatie te vinden. Door verschillende type I-D leadersequenties te vergelijken ontdekten we drie geconserveerde sequentiemotieven die cruciaal zijn voor efficiënte adaptatie. We verwijderden delen van de leader en kwantificeerden vervolgens de activiteit van de spacer integratie met de verkorte leadersequenties. De eerste 30 basenparen van de leader bevatten een geconserveerd gebied waarin deletie resulteert in een aanzienlijke vermindering van de activiteit van de spacer integratie. Interessant is dat het gebied waarin deletie de sterkste reductie van spacer integratie induceert een sequentiemotief bevat dat lijkt op de zogenaamde Integrase Anchoring Site (IAS). Eerdere studies hebben gesuggereerd dat deze IAS de interactie van het Cas1-Cas2-adaptatiecomplex met de CRISPR array stabiliseert en daardoor de efficiëntie van spacer integratie verhoogt. Onze studie leidt tot een compatibel model waarin geconserveerde sequenties binnen de type I-D leadersequentie fungeren als oriëntatiepunten voor het aanpassingscomplex en daardoor aanwijzingen geven voor de juiste integratieplaats.

Tot nu toe hebben we de betrokkenheid van het Cas4 eiwit bij naïeve adaptatie onderzocht (naïef betekent dat de indringer die wordt onthouden nog niet eerder is gecatalogiseerd in de CRISPR-arrays). In **Hoofdstuk 6** laten we zien hoe interferentie en adaptatie met elkaar verbonden zijn in een adaptatieroute die priming wordt genoemd. Tijdens priming is de indringer waarvan spacers worden verkregen al in eerdere infecties gecatalogiseerd. Door mutaties in de sequentie van de protospacer kan het interferentiestadium echter niet meer zo efficiënt plaatsvinden als bij volledige complementariteit tussen crRNA en protospacer. In dit geval is primed adaptatie een krachtig mechanisme dat herstel van efficiënte interferentie mogelijk maakt. We onderzochten het Type I-E CRISPR-Cas-systeem van *Escherichia coli* en ontdekten dat het Cas3 eiwit de belangrijkste speler is die interferentie koppelt

aan adaptatie. Het Cas3-eiwit wordt gerekruteerd door het type I-E-interferentiecomplex en is verantwoordelijk voor de afbraak van binnendringend DNA door zijn nuclease-helicase activiteit. We vonden dat de snelheid waarmee Cas3 het DNA afbrak afhing van de mutaties die aanwezig waren in de protospacer en de PAM. Volledige complementariteit tussen crRNA en protospacer met een consensus PAM resulteerde in snelle afbraak van het binnendringende DNA en veroorzaakte geen priming. Sommige mutaties in de protospacer vertraagden de Cas3 activiteit echter en creëerden daardoor korte DNA fragmenten die efficiënte primed adaptatie mogelijk maakten. We analyseerden de DNA fragmenten die het resultaat waren van afbraak door Cas3 en ontdekten dat ze de juiste lengte hadden om als pre-spacers te dienen. Een belangrijke vondst was dat ze verrijkt bleken met de type I-E specifieke PAM sequentie en dus goed geschikte substraten waren voor opname en integratie door het Cas1-Cas2 adapatatiecomplex. Dit werk laat zien hoe bepaalde CRISPR-Cas systemen immuniteit kunnen behouden door hun geheugen bij te werken, via een ingewikkelde terugkoppeling die de efficiëntie van vernietiging van indringers koppelt aan de opname van nieuwe spacers.

De resultaten van dit proefschrift zijn van belang voor het fundamentele begrip van hoe CRISPR-Cas immuniteit op moleculair niveau tot stand komt. Cas eiwitten die betrokken zijn bij adaptatie- en interferentie evolueerden samen om de prokaryoot te beschermen tegen virale infecties door ofwel *de novo* een antiviraal geheugen te creëren, of door de reeds bestaande immuniteit te versterken. Onze experimentele inzichten verduidelijken de rol van het Cas4 eiwit als een speciale adaptatie co-factor die selectie, verwerking en integratie van spacers mogelijk maakt ter ondersteuning van efficiënte vernietiging van indringers. Bovendien wordt het spacer integratieproces geassisteerd door de sequentiecontext waarin de CRISPR is ingebed, doordat deze aanwijzingen geeft voor de juiste integratieplaats. Deze kenmerken zijn geëvolueerd als gevolg van de roofzuchtige druk waaraan de organismen die het afweersysteem dragen voortdurend worden blootgesteld. Maar ook de aanvallers zelf evolueren door de aanvallen van CRISPR-Cas immuniteit: door de sequenties die het CRISPR-Cas systeem aanvalt te muteren kunnen virussen ontsnappen aan de interferentie. Onze bevindingen over de manier waarop de Cas3 helicase-nuclease activiteit het geheugen onderhoudt onderstrepen de ingenieuze strategieën die bacteriën en archaea gebruiken om te overleven.

Zusammenfassung – Des CRISPR's kleine Helferlein: Beteiligung von CRISPR-Cas Proteinen an der PAM-Selektion

Über die Jahrtausende wurde die Menschheit von pathogenen einzelligen Organismen geplagt. Bis zum Aufkommen von Antibiotikatherapien konnte eine scheinbar harmlose bakterielle Infektion schnell tödliche Folgen haben. In dem Mikrokosmos, den diese Organismen bewohnen, ist die Linie zwischen Angreifer und Angegriffenem jedoch eine sehr dünne Linie. Bakterien und Archaeen werden ununterbrochen von ihren Viren gejagt (Bakteriophagen – von altgriechisch baktérion ‚Stäbchen' und phageín ‚fressen'). Ohne geeignete Mechanismen, um die prokaryotische Zelle vor Virusinfektionen zu schützen, würde es Bakteriophagen gelingen, ganze Spezies auszurotten. Diese Dissertation beschreibt die Arbeit, in der wir Techniken der molekularen Biologie und Biochemie angewandt haben, um die Mechanismen zu studieren, welche bestimmte bakterielle Spezies gebrauchen, um Immunität gegen Bakteriophagen zu entwickeln.

In **Kapitel 1** dieser Dissertation stellen wir das adaptive bakterielle Immunsystem vor, welches auf gebündelten regelmäßig unterbrochenen kurzen palindromischen Wiederholungen (Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) sowie den damit assoziierten Cas (CRISPR-associated) Proteinen beruht. Prokaryoten können sich Bakteriophagen oder andere mobile genetische Elemente (MGE), die in ihre Zellen eindringen, merken, indem sie ein kurzes Stück der DNA des Eindringlings in das CRISPR Array in ihrem eigenen Erbgut (Genom) integrieren. Dieses Erschaffen eines Immungedächtnisses ist die erste Stufe der CRISPR-Immunität und wird Adaptation genannt. Während der Adaptationsstufe selektieren spezialisierte Cas Proteine einzelne Fragmente des MGEs, bearbeiten das Fragment nach bestimmten Kriterien und integrieren es in den CRISPR-locus. Nach der Integrierung in den genomischen CRISPR-locus, wird dieses kurze Stück der Phagen DNA eine sogenannte Trenn-DNA-Sequenz (Spacer), welche zusammen mit dem übrigen CRISPR-locus die Grundlage für die zweite Stufe bildet – die Expression. Während der Expressionsstufe wird das CRISPR-Array in ein langes Vorläufer-RNA Molekül (prä-crRNA) transkribiert, welches die Sequenzwiederholungen (Repeats) sowie die Trenn-DNA-Sequenzen (Spacer) enthält. Diese Vorläufer-RNA wird in kleinere Einheiten zerstückelt, welche

Teile der Sequenzwiederholungen sowie die vollständige Trenn-DNA-Sequenz enthält, die sogenannte CRISPR-RNA (crRNA). Diese gereifte crRNA wird von Cas Protein-Komplexen aufgenommen, die daraufhin die Zelle patrouillieren und versuchen, die crRNA mit potenzieller Angreifer DNA abzugleichen. Diese letzte Stufe wird Interferenz genannt. Wenn crRNA-Cas Komplexe eine DNA Sequenz aufspüren, die vorher im CRISPR-array katalogisiert wurde, bindet der Komplex das DNA Molekül und leitet dessen Zerstörung ein, welches den Eindringling unschädlich macht. Die spezifischen CRISPR-Cas Mechanismen unterscheiden sich zwischen den einzelnen Klassen, Typen und Subtypen, die drei Stufen sind jedoch ihnen allen gemein.

Damit die Interferenz effizient ablaufen kann, ist es essenziell, dass einfallende DNA schnell und selektiv in das Immungedächtnis aufgenommen wird. In **Kapitel 2** dieser Dissertation widmen wir uns darum vornehmlich der ersten Stufe der CRISPR-Immunität und besprechen den derzeitigen Wissensstand über die molekularen Mechanismen, welche der Adaptationsstufe unterliegen. Das Arbeitstier der Adaptation ist ein Komplex, bestehend aus den Cas1 und Cas2 Proteinen. Dieser Adaptionskomplex nimmt eine Struktur an, welche der Form eines Schmetterlings ähnelt. Mit seiner Flügeln ähnelnden Struktur, ist die Hauptaufgabe dieses Proteinkomplexes die Aufnahme, die Bearbeitung sowie der korrekte Einbau von angreifender DNA in das CRISPR-Array. In der großen Mehrheit von CRISPR-Cas Systemen ist ein kurzes Sequenzmotiv direkt neben der durch die crRNA zu bindenden DNA Sequenz maßgeblich für den Erfolg der Interferenz (die DNA Zielsequenz, an welche sich die crRNA anlagert, wird Protospacer genannt). Dieses kurze Sequenzmotiv heißt Protospacer-benachbartes-Motiv (Protospacer adjacent motif – PAM). Der Adaptationskomplex muss also in der Lage sein, DNA Sequenzen einzufangen, welche ein PAM Motiv enthalten, damit die resultierende Spacer-Sequenz den Kriterien der Interferenz entspricht. Hinzukommend muss die eingefangene Sequenz hinsichtlich des PAM Motivs akkurat prozessiert sowie in der korrekten Orientierung integriert werden. ~~Keine einfache Aufgabe für den Adaptationskomplex.~~ Obwohl in einigen CRISPR-Cas Systemen die Cas1 und Cas2 Proteine ausreichend sind, um diese Aufgabe zu erfüllen, erfordern die meisten anderen CRISPR Systeme zusätzliche Integrationsfaktoren.

In **Kapitel 3** dieser Dissertation untersuchen wir einen dieser zusätzlichen Integrationsfaktoren. Das *cas4* Gen, das, obwohl es den

ersten beschriebenen *cas* Genen angehört, lange Zeit mysteriös erschien hinsichtlich seiner Funktion. Vorhergehende Forschung ließ vermuten, dass das Cas4 Protein direkt zur Adaptation beiträgt, indem es mit dem Cas1 Protein interagiert. Und tatsächlich, das Entfernen des *cas4* Gens hatte das Ausbleiben der Adaptation in manchen CRISPR-Cas Systemen zur Folge. Wir nahmen uns also vor, zu einem tieferen Verständnis der exakten Rolle des Cas4 Proteins zu gelangen. Dazu übertrugen wir das CRISPR-Cas Typ I-D Adaptationsmodul des Cyanobakteriums *Synechocystis* sp. 6803, bestehend aus den *cas1*, *cas2* und *cas4* Genen, in unser *Escherichia coli* Modellsystem. In Ergänzung stellten wir ein minimalisiertes CRISPR-Array zur Verfügung, in welches das Adaptationsmodul neue Spacer einbauen konnte. Um den Einbau neuer Spacers zu verfolgen, benutzten wir eine maßgeschneiderte Polymerase-Kettenreaktion (Polymerase Chain Reaction – PCR). Entgegen unserer ursprünglichen Erwartungen, war das Cas4 Protein nicht notwendig für den Einbau neuer Spacer-Sequenzen. ~~Cas1 und Cas2 kamen scheinbar auch ohne Cas4 gut alleine klar.~~ Sobald wir allerdings die Sequenzen analysierten, welche entweder in der An- oder Abwesenheit von Cas4 integriert wurden, sahen wir einen deutlichen Unterschied zwischen den beiden Versuchsaufbauten. Spacers, die in der Abwesenheit von Cas4 aufgenommen wurden, waren zufällig ausgewählt und würden keinerlei Schutz durch Immunität bieten. Im Gegensatz dazu waren Spacers, die in der Anwesenheit von Cas4 eingebaut wurden, nach den PAM-Kriterien des Typ I-D CRISPR-Cas Systems ausgewählt und dazu geeignet, die Zelle vor Virusreplikation zu beschützen. Zusammen genommen erklären diese anfänglichen Beobachtungen, weshalb das *cas4* Gen in vielen verschiedenen CRISPR-Cas Systemen so universell konserviert ist: Cas4 ermöglicht PAM-Kompatible Spacer Selektion während der CRISPR Adaptationsstufe.

Diese Beobachtungen waren hilfreich, um zu verstehen, weshalb Cas4 in so vielen diversen CRISPR-Cas Systemen vorkommt; die mechanistischen und molekularen Grundlagen hingegen blieben unklar. In **Kapitel 4** dieser Dissertation haben wir deshalb den Typ I-D Adaptationskomplex biochemisch rekonstituiert. Wir überexpressierten Cas4 zusammen mit den Cas1 und Cas2 Proteinen und begannen, nach dem Cas4 Protein zu fischen, um zu schauen, welche der Proteine mit Cas4 interagieren. Interessanterweise, aber nicht sonderlich überraschend, fanden wir, dass das Kernadaptationsprotein Cas1 stark an das Cas4

Protein bindet. Wenn wir Cas1 vor die Wahl stellten, entweder mit seinem regulären Interaktionspartner Cas2 oder mit Cas4 zu interagieren, entschied sich Cas1 immer für das Cas4 Protein. Durch diese Beobachtung gelangten wir zu der Erkenntnis, dass in der Zelle zwei sich gegenseitig ausschließende Proteinkomplexe existieren, nämlich der Cas1-Cas2 und der Cas4-Cas1 Komplex. Wir reinigten beide Komplexe auf und testeten sie *in vitro*, um ihre spezifischen Funktionen zu verstehen. Beide Komplexe waren in der Lage, neue Spacer in CRISPR-loci zu integrieren, was sich mit den Beobachtungen aus Kapitel 3 deckt. Das sequenzspezifische Erkennen und Prozessieren von PAM Motiven in einzelsträngigen Überhangen von Spacer-Vorläufern gelang jedoch nur und ausschließlich dem Cas4-Cas1 Komplex. Während PAM-enthaltende Spacer-Vorläufer ohne weitere Veränderung durch den Cas1-Cas2 Komplex integriert wurden, wurde die PAM Sequenz akkurat durch den Cas4-Cas1 Komplex entfernt, bevor der Spacer integriert wurde. Unsere kombinierten Beobachtungen resultieren in einem Modell, in dem ein asymmetrischer Adaptationskomplex PAM und nicht-PAM enthaltende DNA Überhänge differenziell bearbeitet und dadurch die korrekte Orientierung des Spacer Einbaus bestimmt.

In **Kapitel 5** dieser Dissertation untersuchen wir Sequenzmotive, die sich in der sogenannten leader-Sequenz befinden. Die *leader*-Sequenz befindet sich strangaufwärts des CRISPR-Arrays und enthält unter anderem die Promoterelemente, die die Transkribierung des CRISPR antreiben. Die vorherigen Kapitel fokussieren auf das Einfangen und Prozessieren von prä-Spacern, sie erklären jedoch nicht, wie die Adaptationsmaschinerie in der Lage ist, den korrekten Integrationsort zu bestimmen. Spacer werden in chronologischer Reihenfolge, immer am Beginn des Arrays, integriert und bilden deshalb die Reihenfolge der jüngsten viralen Infektionen ab. Dies stellt sicher, dass die Interferenz vornehmlich gegen aktuell zirkulierende Bakteriophagen gerichtet ist. Durch das Erstellen von Sequenzalignments von verschiedenen Typ I-D *leader*-Sequenzen entdeckten wir drei konservierte Sequenzmotive, welche unerlässlich sind für effiziente Adaptation. Wir entfernten nach und nach Teile der *leader*-Sequenz und quantifizierten die Integrationsaktivität mit diesen gekürzten *leader*-Sequenzen. Die ersten 30 Basenpaare des *Leaders* enthalten konservierte Regionen, dessen Entfernung in einer substanziellen Verminderung des Spacer Einbaus resultierte. Interessanterweise enthält die Region, dessen Kürzung die stärkste Reduktion der Spacer Integration verursachte, ein Sequenz-

motiv, welches der sogenannten Integrase Anker-Seite ähnelt (Integrase Anchoring Site (IAS)). Frühere Publikationen beschrieben, dass die Integrase Anker-Seite vermutlich die Interaktion des Cas1-Cas2 Adaptationskomplexes mit der CRISPR-Sequenz befördert und somit die Effizienz der Spacer Integration signifikant steigert. Unsere Studie gelangt zu einem kompatiblen Modell, in welchem die konservierten Sequenzmotive des Typ I-D *Leaders* als Meilensteine für den Adaptationskomplex dienen und damit den korrekten Integrationsort kennzeichnen.

Bis zu diesem Punkt haben wir das Engagement des Cas4 Proteins in naiver Adaptation beschrieben (naiv in diesem Kontext meint, dass der einzufangende Angreifer nicht vorhergehend in einem CRISPR Array katalogisiert wurde). In **Kapitel 6** dieser Dissertation veranschaulichen wir, wie die Interferenz und die Adaptation miteinander verwoben sind. Diese Route der Adaptation wird als primed Adaptation bezeichnet. In der primed Adaptation war der jetzt angreifende Virus während früherer Infektionen bereits in das Immungedächtnis aufgenommen worden. Durch Mutationen in der DNA Sequenz des Angreifers kann der Angreifer der Immunantwort der Zelle entkommen, da die vollständige Übereinstimmung zwischen crRNA und Zielsequenz nicht mehr gegeben ist. In diesem Falle ist die primed Adaptation ein mächtiger und wirkungsvoller Mechanismus, um die Interferenz wiederherzustellen. Wir untersuchten das Typ I-E CRISPR-Cas System des Darmbakteriums *Escherichia coli* und fanden heraus, dass das Cas3 Protein der Schlüsselspieler ist, der die Interferenz und Adaptation zusammen bringt. Das Cas3 Protein wird durch den Typ I-E Interferenzkomplex rekrutiert und ist, durch seine Nuklease-Helikase Aktivität, für die Zerstörung von angreifender DNA verantwortlich.

Wir beobachteten, dass die Rate, mit der Cas3 Ziel-DNA zersetzte, stark abhängig war von Mutationen in der Protospacer oder PAM Sequenz. Eine volle Übereinstimmung zwischen crRNA und Protospacer sowie die Anwesenheit einer korrekten PAM hatte zur Folge, dass die Ziel-DNA schnell und vollständig zersetzt wurde. Diese schnelle Zerstörung resultierte nicht in primed Adaptation, welche durch die Effektivität der Abwehr auch nicht notwendig war. Manche Mutationen in der Sequenz des Protospacers oder der PAM reduzierten die Aktivität des Cas3 Proteins jedoch signifikant, was in der Erschaffung von kurzen DNA Fragmenten resultierte, welche zu effizienter primed Adaptation führten. Diese kurzen DNA Fragmente analysierten wir und

folgerten, dass sie eine geeignete Länge hätten, um als Vorläufermoleküle für Spacers in Frage zu kommen. Interessanterweise waren die Fragmente angereichert für die Typ I-E spezifische PAM Sequenz und waren daher optimale Substrate für das Einfangen und Integrieren durch den Cas1-Cas2 Adaptationskomplex. Diese Arbeit demonstriert, wie bestimmte CRISPR-Cas Systeme ihre Immunität aufrechterhalten können, indem sie auf clevere Weise die Effizienz der Zerstörung von Eindringlingen an den Einbau von neuen Spacer Einheiten koppeln.

Die in dieser Dissertation präsentierten Ergebnisse enthalten signifikante Einsichten, wie CRISPR-Cas Immunität auf der molekularen Ebene etabliert wird. Cas Proteine involviert in Adaptation und Interferenz haben sich in wechselseitiger Anpassung entwickelt, um die prokaryotische Zelle vor viraler Vermehrung zu schützen. Diese Schutzfunktion resultiert entweder aus der Erschaffung von antiviralem Gedächtnis *de novo* oder aus der Unterstützung und Verstärkung bereits bestehender Immunität. Unsere experimentellen Einsichten erhellen die Rolle des Cas4 Proteins als ein der Adaptation gewidmetem Co-Faktors, der die Selektion, die Prozessierung und Integration von Spacers unterstützt und somit die effiziente Zerstörung von Eindringlingen ermöglicht. Daneben ist der Sequenzkontext, in der sich der CRISPR befindet, maßgeblich für die Lokalisierung des korrekten Integrationsorts. Diese Eigenschaften sind ein Resultat des konstanten evolutionären Drucks, dem die Organismen, die diese Immunsysteme in sich tragen, ausgesetzt sind. Jedoch auch die Eindringlinge, die Ziel dieser Immunität sind, passen sich dementsprechend an. Durch das Mutieren von Zielsequenzen in ihrem eigenen Erbgut schaffen es diese mobilen genetischen Elemente, ihrer Unschädlichmachung zu entkommen. Unsere Einsichten, wie die Cas3 Nuklease-Helikase zur regelmäßigen Auffrischung des Immungedächtnisses beiträgt, unterstreichen nochmals die genialen Strategien, auf denen das Überleben von Bakterien und Archaeen in diesem von Viren beherrschten Mikrokosmos beruht. Ein Mikrokosmos, in dem der Jäger schnell zum Gejagten verkommt und in dem ein kleiner Unterschied in der DNA Sequenz den Unterschied ausmacht, ob das Rennen gegen die allgegenwärtige Entropie gewonnen oder verloren wird.

ACKNOWLEDGEMENTS

Having spent many hours on compiling this thesis and having the feeling that I am now approaching its completion, the time has come to think back and dig through some memories of the last 5 years. Five years with countless ups and downs, happy moments and seemingly desperate moments. Moments filled with excitement that shortly after could dissolve in perceived devastation. And the realization that this is - Science.

Contrary to the perception of certain members of society, the lone-wolf ingenious scientist that changes the world forever is more an exception than the rule. Science in the end is mostly a team effort. And that's why I want to thank first and foremost my team of the Broun's lab and of course it's head: **Stan**. Stan, I joined your group at a stage of major change. A stage that came with steep learning curves for everyone involved. But I cannot say that I wasn't prepared for it. I remember well sitting at my parents place and drafting my motivation letter, which turned out to be a rather long letter! Few weeks later I decided to do the trip to Wageningen and to step on your feet personally in order to get this great opportunity! Already in this spontaneous interview you informed me about the possibility of relocation to Delft University of Technology and I was getting more and more excited. Not only because of the reputation of this university, but also due to my very poor geographic knowledge that made me believe moving to Delft would bring me closer to my actual home. But even after I realized that this couldn't be further from the truth, obtaining my PhD from TU Delft seemed to promise a great adventure and the potential to grow! So in the end we both came to the agreement that enabled me to become part of your group and I would like to express my gratitude for making this possible! I was not only allowed to work on exciting topics that enabled me to publish my first scientific papers, but I also gained experience and confidence that nobody can take away from me. Yes, like in every human interaction there was tension and friction, opposing opinions and annoyance, but seeing the draft of my thesis is something that would not have been possible without you and I am proud of what we were able to accomplish together! In that sense, thanks Stan!

Of course, a group with only its head would not be a group (logically) and I was happy how the Brouns lab welcomed me and integrated me

into the group! **Jochem** and **Patrick**, both of you were my very first office neighbors back in the old and charming Microbiology building at the Dreijenlaan. Who could have known back then how everything would develop and how our different fates would unfold. Both of you are extremely smart guys that made it through a lot and I wish you all the best with your future endeavors! Thanks for the beers we drank together, the experiences we had together, the scientific and personal conversations! **Becca**, you were also one of the first members I encountered and I do remember how you noticed my stiff introduction, and most notably the mandatory handshake! Knowing that years later this handshake would become a relic of the past, I should have continued with shaking hands every single day until COVID hit. Jokes aside, I am very happy that you were there! I feel grateful that we walked this stony path together, both, from the scientific and personal perspective. That's why I also write down this heartfelt wish for you and also Jochem, that one day you can live a life that comes as close as possible to your dreams! Both of you can be very proud of your resilience and optimism!

Franklin my friend! I think you're one of the few people for which I was both happy and sad that you managed to leave your affinity to nicotine behind. One of the reasons why I still haven't managed to take this step is the rare opportunity to have conversations "out of the box" (lol). I enjoyed every single smoke that we could spent outside together, talking about personal and professional life. You are a truly a unique person that you don't encounter very often. Your dedication to science is something I haven't seen in many people and I deeply admire your attitude and knowledgeability. But what I probably admire the most is your courage of saying things out loud that many people would keep for themselves. Of course that comes with a price and even we had difficult times together, undoubtedly. But I'm happy that in the end everything straightened out. I do know that a new start always comes with some hiccups and this is most likely also true for your new position. But I'm confident that after you have overcome those hiccups, especially the COVID one, you will excel at what you're doing! I definitely wish you all the best and that your life will not only be recognized and successful, but also filled with love and joy!

And even though the first months of my PhD feel like an eternity away, I still have vivid memories of my old student town Wageningen and its wonderful people! **Tim**, science brought us together twice, once

at the beginning of your own PhD trajectory and once at the beginning of my own. Thanks a lot for introducing me to the tricky *in vitro* world of CRISPR adaptation and the head start that came with it! I wish you and your family all the best! Also, **Wen** and **Prarthana**, even though we hadn't had the opportunity lately, I still have you in good memory! Wen, your infectious optimism and smile are truly unique and have taught me a great deal of how to look at things, even when they don't go the way I would like them to go. Prarthana, you'll also be remembered in a good way and I hope your future scientific journeys keep being exciting and joyful!

Yannis and **Johanna**, isn't it nice to know such great people as you are? When I moved to Delft both of you were among the reasons why I felt sad leaving this environment behind and I can still say that I wish that our paths would have not separated so prematurely. I hope one day in the future we'll have the chance to catch up again! **Indra**, **Hanne**, **Alex**, **Yifan** and **Nico**! You great people and great scientists! Some of you I still knew from my B.Sc. thesis project back in 2012 and I was very happy that we had this small, unfortunately too short, reunion in Microbiology. You were part of the already experienced generation and I admired the path you had taken. Even if not all of you will read this, I am happy to know you and I keep wishing you a successful and bright future! Another someone that I did not expect to see back again is you **Sanne**! I remember how we were together in the Immunotechnology course (with Jasper!) and sweating through those group assignments. That our paths would cross again during the PhD trajectory was a pleasant surprise! Another pleasant surprise, although not very surprising, was your massive impact Nature publication, respect! Hope it'll continue like that for you. Going back even further than Immunotech was the basic practical microbiology in which **Wim** and **Tom** appeared for the first time! First then, followed by my short B.Sc. thesis visit at Microbiology, up to my PhD. Both of you, Tom and Wim, were always a constant in the laboratory of Microbiology and you helped many people, including me, and that's why I also would like to acknowledge you here! **Rob Koehorst**, even though I only met you once or twice, you actually contributed something to this thesis that made all of it possible! I remember that you were very passionate about the *Synechocystis* sp. 6803 strain that you cultured years back in the Biophysics department. And I also do remember that you asked me to stock more cells in case I'd grow them. I'm sorry dear Rob, I

did not culture them a single time. But they served as amazing PCR templates and none of my research would have been possible without that little tube filled with dark green stuff! Thanks a lot Rob, I sincerely hope you're doing well and that one way or the other, you will see those lines. Also you **Daan S**. were part of the old crew when I was still a little B.Sc. student! Already back then you delivered impressive work on Argonaute and it didn't stop there. I'm happy that you had the experience in Jinek's lab and that you managed to settle with your family back in the city where everything began. Thanks a lot for your efforts to structurally tame the I-D adaptation complex and best of luck with your new (well not so new anymore) position at Biochemistry!

Next I'd like to jump back in time a bit further than necessary, but nonetheless, these people left lasting impressions that contributed to my development and hence this work. All the years of studying in Wageningen would have been a lot harder without my BioTech/Mol. Life Science companions. **Robert** and **Sabine**, how much I was looking forward to attending your great wedding, but COVID changed a lot of things for everyone. Nonetheless, I wish both of you all the best, that you stay together happy and healthy! Also **Herr Müller**, you made staying in Wageningen for the first months of my PhD possible so thanks again for housing me and overall, the time we spent! Also **Harm**, **Arne** and **Anne** thanks for the precious memories I have of you, I'm hoping our paths will cross again. Also thanks to **Paulus den Hollander** and **Jan van Lent** for sparking my passion for virology, I honestly really enjoyed my MSc. Thesis time with you, the courses you taught and the opportunity to pass on some of this enthusiasm when you made me your teaching assistant.

Before I mentally leave Wageningen I'd like to especially thank my good friends **Raymond** and **Eric**! It almost feels like a multi-generation thing with you guys. Raymond, starting out as my smoking buddy, then supervisor and lastly my friend (that I, to my great shame, not contact often enough. However, always in my heart, a very true motto. Follow your dreams, you can reach you goals, I'm living proof: **!!!BEEFCAKE!!!**). Same with you Eric … just in reverse: Smoking buddy, my internship student and above all a good friend. You guys are truly a unique bunch and I'm happy that I know you! Both of you would have been great paranymphs and I feel betrayed that the current restrictions don't allow you guys guarding me through this epic battle of finishing off the beast called doctor of philosophy. In any way,

I wish both of you the best of luck, stay the way you are!

And even though I say above that I was prepared for Delft, I kind of have to say that I wasn't. After spending a tad more than half a year in Wageningen, I almost felt a bit homesick and missed the people that I acknowledge here. Nevertheless, the human is a creature of habit and even difficult changes become normal at some stage. Although with a slight delay, the Brouns lab reassembled at TU Delft and soon after, was back and kicking. This time we had some additions though and **Anna** was one of them. Thanks a lot Anna for your support with taming the AKTA systems, purifying proteins and overall, always being open for a little chat! It was good having you around and that you took care of a lot of things that otherwise would have slipped. All the best for you and your family! Also you **Rita** were a great addition, in every sense. It happens rarely that you meet such genuine and kind hearted people and I'm glad that you were around for most of the time of my PhD. With that said I really wish you (and of course also Franklin) have a bright future, with everything included that you value and is important to you. Stay healthy, stay happy and most importantly, stay the great person you are!

Not so long after the Brouns lab restarted in Delft, **Cristóbal** joined us! Man what should I say? Thank you so much for all the help and support I got from your site! I remember how I was loading one of those dreadful denaturing PAGE when you first showed up, not knowing that at some point you would go through the same kind of struggle. Nevertheless, not that much later you came to me saying: Yo, I have a present for you! And indeed, some PCR magic made it possible to first detect adaptation in the type I-D system and from there things started to take off! I'm proud of our close collaboration and the things we accomplished together, going through good and bad times. This thesis would have been a different one without you and I greatly appreciate your help and I hope I could give you useful lessons in return. I'm happy about the years we spent together, the conferences we attended, the beers and conversations we shared! I not only had a great colleague with you but also found a great friend in you. Man, from deep down I wish you all the best! Of course the same to your wonderful wife **Patri**, thanks both of you for the great time. Please stay the way you are but most importantly always healthy and happy! Pretty much at the same time also you **Cristian** became part of the team. And even though we're both equally bad at staying in touch I do hope that soon we can

have a drink together, first for my graduation and later on to celebrate yours! I do know that science can be tough on you and despite the endless hours you spend in the lab, success is nothing that can be taken for granted. You're a very hard working guy and I seriously wish you can enjoy the fruit of your hard work, sooner or later! I wish you all the best for the completion of your thesis and of course also for the things that will follow. Man you're such a good guy, stay like that!

Next also many thanks to **Boris** and **Teunke** who helped keeping the lab up running and also otherwise were great to have around! Interesting that both of you were already present in my pre-thesis times and it was nice that you kept reappearing. So who knows what the future will bring? In any way I hope that you're happy with whatever may come and that you stay the good people I got to know you! Also thanks **Benjamin** for organizing this amazing trip to the ESA swimming pool, I've never seen such an impressive swimming pool (through a window). The last person I personally got to know joining the Brouns lab was you **Sam**! First and foremost a big fat thank you for helping out so much with the Dutch summary of this thesis, always pays off (for you and for me xD) knowing a writer! This saved me a ton of work so thanks a lot for that! I also wish you the best of luck with the coming years of your PhD and the work connected to it and I hope that you can keep up with your own expectations!

We also had many great students that brought some fresh views and change to the group! Thanks **Marre** for your initial contributions to the Cas4 story … I think back then nobody expected that it would drag on for another 4 years (how matching). Also you **Ana** left a lasting impression, I definitely wish you good luck with your own PhD and I'm convinced you'll do a great job! Same for you **Rodrigo**, good luck with your Cas4 review and overall I hope you enjoy your own academic adventure. **Monique** and **Marrit** you were also fun to have around and your work was a nice contribution to the Brouns lab! Also thanks a lot **Karlijn** for your interest and an open ear, it was good to have you around and I wish you your own PhD success story! The same for you **Ilma**, and also many thanks for helping me out during my uniQure application procedure, very much appreciated! I'm sure I forgot to mention each and every student that joined the group during the years that I was part of this lab, but each and everyone of you did your part to keep up an interesting and everchanging atmosphere!

And, although the Applied Science building did not exactly facilitate interaction and exchange between different groups of people, I nonetheless left this department with many good memories of all the different people. Many thanks to the **JooCs**! **Luuk**, **Sungchul**, **Laura** you were already around even before I made the move from Wageningen and it was good to at least remotely know people at the place I was moving to. Also **Mike**, **Ilja**, **Thijs**, **Stanley**, **Ivo** and **Mohammed** you made the Joo lab a fun bunch of people that were good to have around. Not to forget about you **Chirlmin**! Having conversations and discussions with you was always fruitful and I can fully understand the popularity you enjoy among your group members and the department! Even that you reserved time for me to discuss my future ambitions in which you gave me substantial food for thoughts is something I very much appreciate! Thanks for acting as my promoter as well! I really do wish you all the best for the future that lies ahead of you and of course only health and happiness for your wife and daughter.

Also thanks a lot to **Martin** and **Misha**, it was always a pleasant kind of change seeing how differently you approached the CRISPR field! Same holds true for you **Bertus**, also thanks a lot for the conversations we had! Of course not to forget about **Sacha**, **Anke** and **Jan**! You were really the people that made sure that we get all the support needed to focus on the science part and that's something definitely worth acknowledging! Thanks a lot for all your help and on top for the conversations that were always nice to introduce some change and loosen up the daily routine a little bit! It was great meeting you and I wish all of you a long and happy life. Also whenever something wasn't going smoothly with the AKTA, it was nice having you around **Cecilia** and **Eli** and I also greatly appreciated the chats we had! Good luck with finishing your thesis **Tanja** and **Stefan**, alles Gute euch beiden! I'm also very happy for having you guys around, **Richard** and **Louis K** – and I do have to admit that I miss the chats we had every now and then! Hope you guys will find happiness and a position that you find interesting and fulfilling! **David** and **Louis R**, being involved in the iGem teams of 2017 and 2018 was a great experience and that's also thanks to you! Ich wünsche euch beiden wirklich alles erdenklich Gute, auf dass ihr euer Glück findet (gilt natürlich auch für dich Richard, ich drücke dir nach wie vor die Daumen!). **Michel** and your lovely daughter Maddi, again my respect for what you have accomplished – very impressive how you managed to juggle all those things. Good luck

with your new position, hope you find joy!

Although you're a bit of an outlier in this listing, I still would like to express my gratitude that we met, **Dr Filonenko**! I've no idea what you're actually doing in ChemE but you do seem to be on top of it! Always very impressive seeing highly motivated scientists being fully submerged in their field and topic. But even more satisfying meeting the person behind it and believe me, such outspoken and charismatic folk you don't meet very often! In that sense, dude, stay the cool guy you are and I hope one way or the other we'll meet again!

Also thanks, but no thanks, to **Sodexo**. The punishment for forgetting your lunch at home.

And even this traces back before my PhD I'd also like to acknowledge **Peter** and his awesome group! Thanks a lot Peter for hosting me back in the days, it was truly an amazing experience and I learned so much from this stay! That of course includes **Hannah**, **Adrian**, **Bridget**, **Corinda**, **Max** and **Rita**. Those are some fantastic memories and I hope that one day I'll see all of you again! **Simon** and **Rob**, you joined after I left Peters group, but nonetheless, the occasions we had during conferences will be remembered!

Also dear **Brenda**, I have to thank you for many things, but mostly for being a great supervisor that, despite the Boss, made it a great stay at Twincore! It's nice to know that you found a place now where you are happy. And you also helped me a lot with my current position, who knows, without you this could have been a different outcome. I wish you good luck at the Dutch cancer institute and I hope your work will have real impact on the life of patients!

Before I start some excursion into German, I also would like to thank you **Pawel**! It was great having you around and you're still dearly missed here in Delft. Regardless if you were here or now in France, I always keep my fingers crossed that you're being happy and enjoying life!

German Intermezzo

Liebe **Mama**, lieber **Papa**, es klingt natürlich erstmal ziemlich schmalzig, aber es entspricht der Wahrheit: Ohne euch wäre all das hier nicht möglich gewesen. Ich könnte euch eine Liste von Dingen schreiben, für die ich euch dankbar bin, die wahrscheinlich die Länge

der gesamten Danksagungen hätte. Also versuche ich das ganze etwas zu komprimieren. Ihr seid für mich die besten Eltern die ich mir hätte wünschen können. Dank euch habe ich glückliche Kindheitserinnerungen. Ihr habt mir mein Studium ermöglicht, ohne welches diese Dissertation nicht zustande gekommen wäre. Ihr habt mir immer nach bestem Wissen und Gewissen zur Seite gestanden und ein nicht unerheblicher Teil eurer grauen Haare kamen wahrscheinlich durch mich zustande. Ich bin so dankbar euch zu haben und auch wenn uns schon seit einigen Jahren mehrere hundert Kilometer Distanz trennen, seid ihr in meinen Gedanken immer bei mir. Ihr habt mir so viel ermöglicht, mir so viel geholfen und sowohl Freude als auch Leid mit mir geteilt. Dafür werde ich euch für immer dankbar sein! Ich habe euch lieb, euer Dr. Sohn.

Geliebte **Oma**, was hätte ich mir gewünscht, dass du jetzt noch bei uns wärst und diesen Moment mit uns teilen könntest. Du wirst uns immer fehlen. Ich widme dir diese Zeilen in der Hoffnung, dass sie dich dennoch irgendwie erreichen. Ich war und bin stolz auf dich, für mich warst du immer die starke Frau die trotz der widrigsten Umstände niemals aufgegeben hat. Wie oft denke ich an dich und bin dankbar für das Vorbild das du mir gewesen bist. Selbstlos, hilfsbereit – das Wohl anderer immer über dein eigenes gestellt. Ich glaube diese Welt wäre eine Bessere, gäbe es mehr Menschen wie dich. Auch wenn der Schmerz und die Traurigkeit langsam weniger werden, ganz aufhören wird es nie, denn du bist durch nichts zu ersetzen. Ich habe so vieles von dir gelernt was mich bis an mein eigenes Ende begleiten wird. Ich hoffe und wünsche, dass es irgendwann einmal ein Wiedersehen gibt. Dein Enkel Sebastian.

Und auch dir möchte ich eine kurze Zeile widmen lieber **Ralph**. Du hast uns in einer sehr schweren Zeit zur Seite gestanden und unserer Familie sehr geholfen. Dafür danke ich dir!

Natürlich wäre ich nicht derselbe ohne den Rest meiner Familie! Ich glaube, dass der Familien- und Freundeskreis einen großen Einfluss auf den Werdegang eines Menschen hat, denn all dies beeinflusst den Charakter und das Wesen. Ich bin immer stolz auf meine Familie gewesen und schwelge gerne in Erinnerungen an vergangene Zeiten. Lieber **Günther** und **Anne**, lieber **Christoph** und **Olaf**, liebe **Kerstin** und **Klaus**, **Julia** und **Mark**. Liebe **Barbara**. Ja, auch du, **Armin**! Lieber **Henning** und **Petra**. Ihr alle habt eine Rolle in mei-

nem Leben gespielt und ich bin dankbar für das Zusammen erlebte und die Erinnerungen. Bleibt so wie ihr seid, aber vor allem glücklich, froh, gesund und munter!

Bevor es wieder zurück ins Englische geht, möchte ich auch dir noch einmal danken liebe **Katja**! Mensch, verrückt was? Mein erstes und wahrscheinlich letztes Buch, da muss ich dich natürlich auch würdigen! Wir haben uns in merkwürdigen Zeiten kennen gelernt und ich bin manchmal immer noch erstaunt, wie lange es wir trotz der Distanz miteinander ausgehalten haben. Ich hätte wirklich besser zu dir sein müssen und ich danke dir, dass du mich nicht völlig blöd findest. Ich freue mich jedes Mal von dir zu hören und ich bin sehr froh darüber, dass du dein Glück gefunden hast! Ich wünsche dir alles erdenklich Gute, bleib gesund!

Und auch du mein lieber **Vincent**, warst ja immer mein Bester. Schade, dass sich die Dinge so verlaufen haben, aber ich hoffe, wir werden uns nicht völlig aus den Augen verlieren! Ich wünsch dir und deiner **Merle** jedenfalls alles Gute, dass ihr beiden immer schön gesund und glücklich bleibt (und liebe Grüße an deinen Papa und deine Geschwister!). Dir **Julius** und **Alex**, so wie euren Familien, wünsche ich auch immer viel Erfolg, Glück und Gesundheit! **Sina, Keule**, **Bosse**, **Jochen**&**Kenny**, schön dass es euch gibt! Dir liebe **Anne**, gelten dieselben Wünsche wie den zuvor benannten! Finde es schon irre, dass man sich trotz der langen Zeit nicht aus den Augen verloren hat!

End of German intermezzo

At this point I want to acknowledge the **Foundation for Fundamental Research on Matter** which is part of the **Dutch Research council**. Not only were you paying my salary for the last years but you also organized some very nice courses for me! My sincere gratitude also for being very understanding what kind of impact a pandemic has on a research trajectory. Extending my contract was a very nice gesture and I'm happy that you did, thanks a lot for that!

Of course I also would like to thank my new working environment for integrating me into this new and exciting world outside of academia! **Angga**, you were my first contact at this company and since then you have been a great mentor and colleague! I'm happy that I'm allowed to work on such exciting topics that hopefully will translate into

a better future for our patients. Also you **Fiona**, **Tom**, **Bas**, **Giorgia**, **Karin**, **Rudy**, **Seyda**, **Astrid**, **Irena**, **Jiali**, **Lukas**, **Lisa**, **Vanessa**, **Sonay**, **Morgane**, **Bianca**, **Anna** and **Melvin** (and all the others that I forgot mentioning here) are a great bunch of people that I greatly enjoy working with! I'm grateful that I found a follow up position to my PhD that allows me to learn, grow, develop and after all become a more experienced scientist!

Last but not least I want to acknowledge the most influential person of the last 4 years, **Viktorija**. I do remember well what a good job you did at actively ignoring me in the beginning! At that point it was far beyond imagination that we would have this amazing little girl named Matilda. We have been through a lot together, positive and negative, but somehow we have managed to land on our feet again. There has been a lot of change in our life and we tried to make the best out of it. In any case, thank you so much for being part of this journey because without you, this thesis and this existence would have been a different one. This experience is something that cannot be taken away and it will always connect us, no matter what. I definitely wish you all the happiness that you deserve in your life! That of course includes good health, a job that you really enjoy and that you're always surrounded by people that love you and care for you!

My little **Matilda**. Right now this doesn't mean anything to you and that's the beauty of being a child. You don't have to take care of deadlines, don't have to worry about anything (except maybe where your stupid parents have hidden the sweets again). But one day you'll be old enough to read and understand those lines: When your dad started to work for this book, there wasn't even a thought of you. But now, while I'm writing these lines, you're already 2 ½ years old, and it's amazing to what little human you already have developed in this short time span. You can eat yourself, you can say what you like (and even more strongly what you dislike), you can even count to ten (one, two, three, short break, eight, ten!). And the most amazing thing is: This is only the very beginning. I'm confident you'll not stay this amazing little girl forever (amazing yes, little not). After some years of being a horrible teenager, you'll grow up to be an amazing adult. I'm immensely proud of you. Proud of the little girl that you are now, and proud of that person you will be in the future. I hope that you will find your own way that makes you a happy and fulfilled individual. Your world will most likely be very different from the one it is now and I wish for you that

it will be a better one. I hope when you read this you will have had the same kind of happy childhood that I like to remember myself. In any way, always remember that you're being loved and valued and I wish you all the luck, love and happiness of this world!

Curriculum Vitae

Sebastian Niklas Kieper

22 November 1988	Born in Hanover, Germany
2001 - 2008	Higher Education Entrance Qualification (Abitur) Käthe-Kollwitz-Gymnasium, Hanover, Germany
2009 - 2013	B.Sc. in Medical Biotechnology Wageningen University and Research Wageningen, The Netherlands
2013 - 2015	M.Sc. in Medical Biotechnology Wageningen University and Research Wageningen, The Netherlands
2016 - 2021	PhD in Molecular Biology/Biochemistry Title "*CRISPR's little helpers: CRISPR-Cas Proteins involved in PAM-selection*" Promotor: Dr. ir. S.J.J.Brouns Promotor: Dr. C. Joo Department of Bionanoscience Technical University Delft, The Netherlands
2020 - present	Junior Scientist, Vector Technology uniQure Biopharma B.V. Amsterdam, The Netherlands

List of Publications

1. **Kieper S.N.**, Almendros C., Haagsma A.C., Barendregt A., Heck A. J.R., Brouns S.J.J., "Cas4-Cas1 is a PAM-processing Factor mediating Half-Site integration during CRISPR adaptation" Accepted for Publication in The CRISPR Journal (2021).
2. Almendros C., **Kieper S.N.**, Brouns S.J.J., "CRISPR-Cas Systems Reduced to a Minimum", Molecular Cell, 73(4):641-642, (2019).
3. **Kieper S.N.**, Almendros C., Brouns S.J.J., "Conserved motifs in the CRISPR leader sequence control spacer acquisition levels in Type I-D CRISPR-Cas systems", FEMS Microbiology Letters 1;366(11):fnz129, (2019).
4. **Kieper S.N.**, Almendros C., Behler J., McKenzie R.E., Nobrega F.L., Haagsma A.C., Vink J.N.A., Hess W.R., Brouns S.J.J. "Cas4 Facilitates PAM-Compatible Spacer Selection during CRISPR Adaptation." Cell Reports, 22(13):3377-3384, (2018).
5. Jackson S.A., McKenzie R.E., Fagerlund R.D., **Kieper S.N.**, Fineran P.C., Brouns S.J.J., "CRISPR-Cas: Adapting to change", Science, 356(6333):eaal5056, (2017).
6. Fagerlund R.D., Wilkinson M.E., Klykov O., Barendregt A., Pearce F.G., **Kieper S.N.**, Maxwell H.W.R., Capolupo A., Heck A.J.R., Krause K.L., Bostina M., Scheltema R.A., Staals R.H.J., Fineran P.C., "Spacer capture and integration by a type I-F Cas1-Cas2-3 CRISPR adaptation complex", PNAS, 114(26):E5122-E5128, (2017).
7. Künne T., **Kieper S.N.**, Bannenberg J.W., Vogel A.I., Miellet W.R., Klein M., Depken M., Suarez-Diez M., Brouns S.J.J., " Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation", Molecular Cell, 63(5):852-64, (2016).
8. den Hollander P.W., **Kieper S.N.**, Borst J.W., van Lent J.W., "The role of plasmodesma-located proteins in tubule-guided virus transport is limited to the plasmodesmata", Archives of Virology, 161(9):2431-40, (2016).
9. Wilkinson M.E., Nakatani Y., Staals R.H., **Kieper S.N.**, Opel-Reading H.K., McKenzie R.E., Fineran P.C., Krause K.L., " Structural plasticity and in vivo activity of Cas1 from the type I-F CRISPR-Cas system", Biochemical Journal, 473(8):1063-72, (2016).