2019

Generation of Memory of Infection During the CRISPR-Cas9 Immune Response

Robert Heler

Generation of Memory of Infection During the CRISPR-Cas9 Immune Response

A Thesis Presented to the Faculty of

The Rockefeller University

in Partial Fulfillment of the Requirements for

the degree of Doctor of Philosophy

by

Robert Heler

June 2019

Generation of Memory of Infection During the CRISPR-Cas9 Immune Response

Robert Heler, Ph.D.
The Rockefeller University 2019

Clustered regularly interspaced short palindromic repeat (CRISPR) loci and their associated (Cas) proteins provide adaptive immunity against viral attack in prokaryotes. Upon infection, short phage sequences known as spacers integrate between CRISPR repeats and are transcribed into small RNA molecules that guide the Cas9 nuclease to the viral targets (protospacers). *Streptococcus pyogenes* Cas9 cleavage of the viral genome requires the presence of a 5′-NGG-3′ protospacer adjacent motif (PAM) sequence immediately downstream of the viral target. Before my graduate work, it was not known whether and how viral sequences flanked by the correct PAM are chosen as new spacers. My work revealed that Cas9 selects functional spacers by recognizing their PAM during spacer acquisition. The replacement of *cas9* with alleles that lack the PAM recognition motif or recognize an NGGNG PAM eliminates or changes PAM specificity during spacer acquisition, respectively. Cas9 associates with other proteins of the acquisition machinery (Cas1, Cas2 and Csn2), presumably to provide PAM-specificity to this process. This was a newly identified function of Cas9 in the genesis of prokaryotic immunological memory.

To further explore the link between Cas9 and spacer acquisition, I performed random mutagenesis of the RNA-guided Cas9 nuclease to look for variants that provide enhanced immunity against viral infection. I identified a mutation, I473F, which increases the rate of spacer acquisition by more than two orders of magnitude. This patented variant of Cas9 highlights the enzyme's role during CRISPR immunization, provides a useful tool to study this otherwise rare process, and holds promise to be developed into a biotechnological application.

Researching Cas9 and spacer acquisition involved many rounds of high-throughput sequencing of millions of spacers acquired by bacteria during phage infection. These experiments revealed that the abundance of each spacer in the surviving population was highly uneven. Since the molecular mechanisms underlying this bias were not known, I decided to look into the factors that affect the distribution of individual spacer sequences during phage infection of cells harboring the CRISPR system from *Streptococcus pyogenes*. My work has shown that spacer patterns are established early during infection and correlate with spacer acquisition rates, but not with spacer targeting efficiency. The data suggests that the rate of spacer acquisition depends on unique sequence elements within the spacers and therefore determines the abundance of different spacers within the adapted population. These results elucidate a fundamental mechanism behind the generation of immunological diversity during the type II CRISPR-Cas response.

Acknowledgements

First, I would like to express my sincere gratitude to my advisor, Dr. Luciano Marraffini, for his continuous support of my Ph.D. work, his patience, trust and immense knowledge. I could not have imagined having a better advisor and mentor.

Second, I would like to thank the rest of my thesis committee, Dr. Daniel Mucida and Dr. Howard Hang, for their insightful comments and guidance which helped improve my research. In addition, I'd like to thank Dr. Harris Wang (Columbia University) for kindly agreeing to serve as the external committee evaluator of my thesis.

Third, my sincere appreciation goes to my collaborators outside of Rockefeller University, Dr. David Bikard (Institut Pasteur), Dr. Jennifer Doudna (UC Berkeley), Addison Wright (UC Berkley) and Dr. Marija Vucelja (University of Virginia), who provided expertise and helped advance my research projects.

Forth, I thank my fellow labmates for their invaluable research contributions to my thesis, in particular Poulami Samai, Gregory Goldberg and Joshua Modell. I'm also grateful for all the fun I've had over the past six years with the entire lab and for the friends I've made, especially Pascal Maguin, Nora Pyenson and Jakob Ros-.

Last but not least, I thank my husband Kyle Straub and our puppies, Dolly and Bee, for moral support and unconditional love.

Table of Contents

Chapter I. Introduction 1

1.1. Overview of CRISPR Immunity

1.2. Spacer Acquisition

1.2.1. The Protospacer Adjacent Motif

1.2.2. Naïve and Primed Acquisition

1.2.3. Spacer Integration

1.3. Unanswered Questions

Chapter II. Role of Cas9 in Spacer Acquisition 20

2.1. The protospacer adjacent motif is highly conserved

2.2. Cas9 is required for spacer acquisition

2.3. Cas9 specifies the PAM sequence of newly acquired spacers

2.4. Cas9 associates with other Cas proteins involved in spacer acquisition

2.5. The PAM binding motif of Cas9 is required for PAM selection

2.6. Discussion

Chapter III. Generation of Cas9 variants that increase spacer acquisition 45

3.1 Changing the PAM specificity of Cas9

3.2. Hyper Cas9 has increased NAG PAM specificity

3.3. Hyper Cas9 enhances CRISPR adaptive immunity by 100-fold

3.3.1. Hyper Cas9 provides wild-type cleavage efficiency

3.3.2. Hyper Cas9 promotes higher rates of spacer acquisition

3.4. Discussion

Chapter IV. Spacer Acquisition and Immunological Diversity 71

4.1. An uneven spacer distribution across the phage genome

4.2. Spacer distribution is biased and reproducible

4.2.1. Effects of DNA cleavage efficiency on spacer distribution

4.2.2. Effects of acquisition rates on spacer distribution

4.3. Analysis of spacer sequences that determine the rate of acquisition

4.4 Discussion

Chapter V. Perspectives 102

Chapter VI. Materials and Methods 105

References 138

Chapter I.

Introduction

Clustered, regularly interspaced, short palindromic repeats (CRISPR) loci and their associated genes (cas) confer bacteria and archaea with adaptive immunity against phages and other invading genetic elements. A fundamental requirement of any immune system is the ability to build a memory of past infections in order to deal more efficiently with recurrent infections. The adaptive feature of CRISPR-Cas immune systems relies on their ability to memorize DNA sequences of invading molecules and integrate them in between the repetitive sequences of the CRISPR array in the form of 'spacers'. The transcription of a spacer generates a small antisense RNA that is used by RNA-guided Cas nucleases to cleave the invading nucleic acid in order to protect the cell from infection. The acquisition of new spacers allows the CRISPR-Cas immune system to rapidly adapt against new threats and is therefore termed 'adaptation'. Recent studies have begun to elucidate the genetic requirements for adaptation and have demonstrated that rather than being a stochastic process, the selection of new spacers is influenced by several factors. This chapter reviews our current knowledge of the CRISPR adaptation mechanism.

Chapter 1.1. Overview of CRISPR Immunity

Bacteria and archaea have evolved to thrive in hostile environments under the constant threat of viral (phage) attack. As a result, these organisms have devised numerous strategies to prevent phage infection, including abortive infection, surface exclusion and restriction modification systems[1,2]. While highly effective, these innate defense strategies provide non-specific immunity. In contrast, the CRISPR-Cas immune system provides an adaptive defense mechanism against phages and other mobile genetic elements[3–5].

Since their discovery in *Escherichia coli* in 1987[6], CRISPR systems have proven to be widespread among bacteria and archaea[7–9]. Generally, a CRISPR locus contains the CRISPR-associated (*cas*) genes and the CRISPR array. The *cas* genes encode a diverse family of Cas proteins carrying predicted functional domains of proteins that participate in nucleic acids transactions, such as DNA-binding proteins, nucleases, polymerases and helicases[4,10]. The CRISPR array consists of identical nonadjacent sequences (repeats) interspaced by similarly sized variable sequences (spacers). An AT-rich leader sequence located upstream of the first repeat promotes the transcription of the CRISPR array[11,12] and is essential for spacer acquisition. Repeats are usually conserved within the same locus, and in most cases contain partially palindromic sequences[13]. Spacers are highly diverse even among closely related strains and were therefore initially exploited for strain typing purposes[14]. In 2005, independent bioinformatics studies revealed homology between spacer sequences and mobile genetic elements[3,15]. This observation led to the hypothesis that CRISPR may provide protection against invading phages and plasmids[4]. Soon, spacers were confirmed to provide sequence-specific interference against all prokaryotic routes of horizontal gene transfer, including bacteriophage infection[5,16,17], plasmid conjugation[18] and transformation[19,20].

CRISPR-Cas systems provide immunity against phages through a three-step defense pathway (Figure 1-1). First, a fragment of the invading nucleic acid (protospacer) is incorporated into the CRISPR array along with a synthesis of an additional repeat unit. This process is known as adaptation and is responsible for the unique adaptive features of CRISPR[5]. Second, during the crRNA biogenesis phase, the CRISPR locus is transcribed and then processed into mature guide RNAs (crRNAs)[21]. Third, crRNAs recruit effector complexes and guide them to their target by base pairing with the invading nucleic acids[22]. This last step of CRISPR immunity is known as interference and ends with the cleavage of the exogenous genetic element[23]. Despite this general mode of action, CRISPR systems have been classified into six types (I-VI), each of them with several subtypes, depending on the gene composition and architecture of the respective *cas* operons[24,25]. While studies of the mechanisms of crRNA biogenesis and interference are well advanced, CRISPR adaptation, also known as immunization or spacer acquisition, perhaps the most puzzling and fascinating aspect of these systems, remains poorly understood.

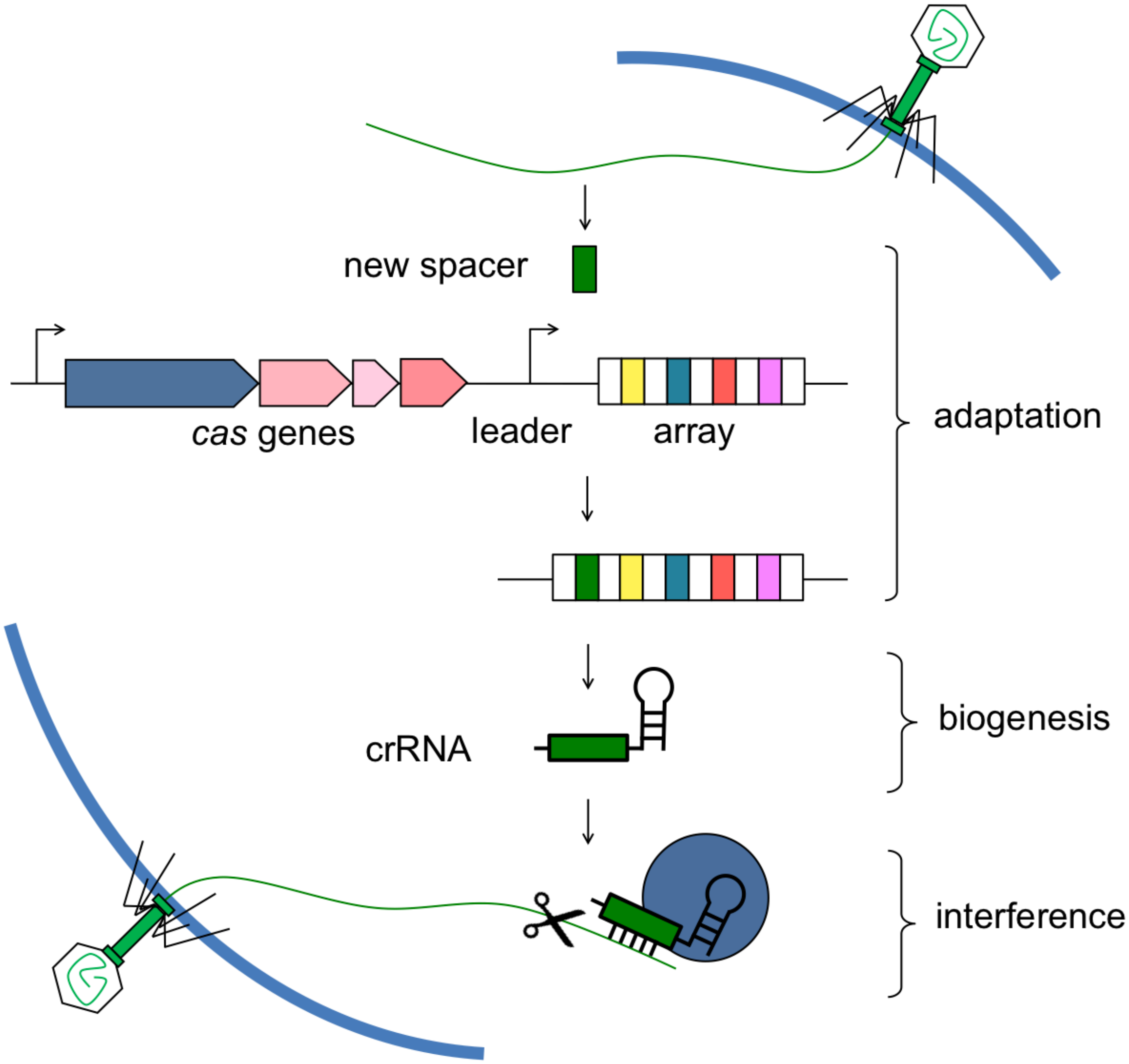


Figure 1-1. The three stages of CRISPR immunity. CRISPR loci consist of clusters of repeats (white rectangles) and spacers (colored rectangles) that are in proximity of the upstream leader sequence and CRISPR-associated (*cas*) genes. During adaptation, new spacers derived from the genome of the invading virus are incorporated into the CRISPR array along with a new repeat unit. During crRNA biogenesis, the array is transcribed, and the precursor transcript is processed by Cas endoribonucleases in order to generate small crRNAs. During interference, the crRNA guides a complex of Cas proteins to the matching target to initiate nucleolytic cleavage (scissors) of the invading nucleic acid.

Chapter 1.2. Spacer Acquisition

Spacer acquisition was first demonstrated under laboratory conditions in 2007 for the type II-A system of *Streptococcus thermophilus*[5]. In these studies, investigators examined the CRISPR array of phage-immunized bacteria and found the addition of new repeat-spacer units, with all new spacers perfectly matching regions of the genome of the challenging phage. Mutants that acquired spacers targeting sequences shared between two phages were resistant to both viruses. These results established CRISPR-Cas systems as an adaptive, sequence-specific immune system against phages and were corroborated in other bacteria and archaea containing different CRISPR-Cas Types: *Escherichia coli* type I-E[26–28], *Pseudomonas aeruginosa* type I-F[29], *Streptococcus agalactiae* type II-A[30], *Haloarcula hispanica* type I-B[31], and *Sulfolobus solfataricus* type I-A and III-B[32].

Chapter 1.2.1. The Protospacer Adjacent Motif

When investigators aligned newly acquired spacers from the *S. thermophilus* CRISPR-Cas system in search for common motifs, they found something unexpected. Instead of a common sequence *within* the spacers they found a conserved sequence *outside* the target (also known as protospacer), which was termed Protospacer Adjacent Motif (PAM)[17]. Therefore, it became clear early on that not all phage sequences are equal for the CRISPR-Cas system, suggesting that the adaptation machinery only acquires spacers that have adjacent PAMs. The PAM is not only important for the acquisition of spacer sequences, it is also required for the interference phase of CRISPR immunity since PAM mutations in types I and II prevent Cas nuclease cleavage[17,33–36]. This interference requirement is readily exploited by phages, which can avoid CRISPR immunity by mutating the PAM sequence[17]. The PAM is fundamental to avoid auto-immunity. If CRISPR immunity relied only on base-pair interactions between the crRNA and the target DNA, then the spacer sequence on the CRISPR array would be a target for the crRNA as well. Since the flanking sequences of a spacer in the CRISPR array are the CRISPR repeat, which lack a proper PAM, auto-immunity is prevented and only protospacers that are flanked with the correct PAM can be cleaved. Type III CRISPR-Cas systems, however, seem to be an exception, as no PAM is evident from the alignment of protospacer sequences acquired by these systems, nor it is required for target cleavage[37,38]. As a consequence, Type III systems developed a different mechanism to prevent auto-immunity[38].

While the recognition of a PAM by the acquisition machinery is essential for a protospacer to be selected, it appears that other mechanisms further influence the protospacer choice. Studies of spacer acquisition by the *E. coli* type I-E[28,39] and the *S. thermophilus* type II-A[40] CRISPR-Cas systems reported unequal distributions of protospacers across various targets. The expansion of the CRISPR arrays from *S. thermophilus* was monitored using DNA deep-sequencing upon infection with a lytic phage[40] and roughly half a million phage-derived spacer sequences were analyzed. Surprisingly, the top 10% most overrepresented spacers accounted for 99% of the identified sequences. In contrast, some candidate protospacers that could have been theoretically acquired from the target based on PAM compatibility were never sampled. Due to partial sequence similarities between some endogenous spacers and the target, Paez-Espino *et al.*[40] propose priming (Figure 1-2a) as a possible explanation for the strong overrepresentation of certain protospacers. In a similar study in type I-E, all potential protospacer sequences adjacent to a PAM were used as spacer donors, but the frequencies were indeed highly unequal[39]. While no correlation was observed between the frequency of protospacer incorporation and its nucleotide sequence, melting temperature, GC content, ssDNA secondary structure, or transcription pattern, other investigators detected additional sequence motifs besides the PAM that influence the acquisition efficiency of a protospacer[41]. By exchanging nucleotide blocks of various lengths upstream or downstream of one high-acquisition and one low-acquisition protospacer, investigators were able to reverse their rates of acquisition, suggesting that DNA motifs located at both ends of the highly acquired protospacer were responsible for its frequent incorporation. More specifically, in addition to the PAM, a dinucleotide AA motif termed Acquisition Affecting Motif (AAM) located at the 3’ end of the protospacer can boost the rate of incorporation of a given protospacer. Sequences of more than two million spacers confirmed the overrepresentation of the AA motif in highly sampled protospacers. Nonetheless, the AAM was not present in all highly sampled protospacers, indicating that other unidentified DNA motifs might influence the sampling frequencies of protospacers. In a different study, the AAM was not confirmed among plasmid-derived spacers.

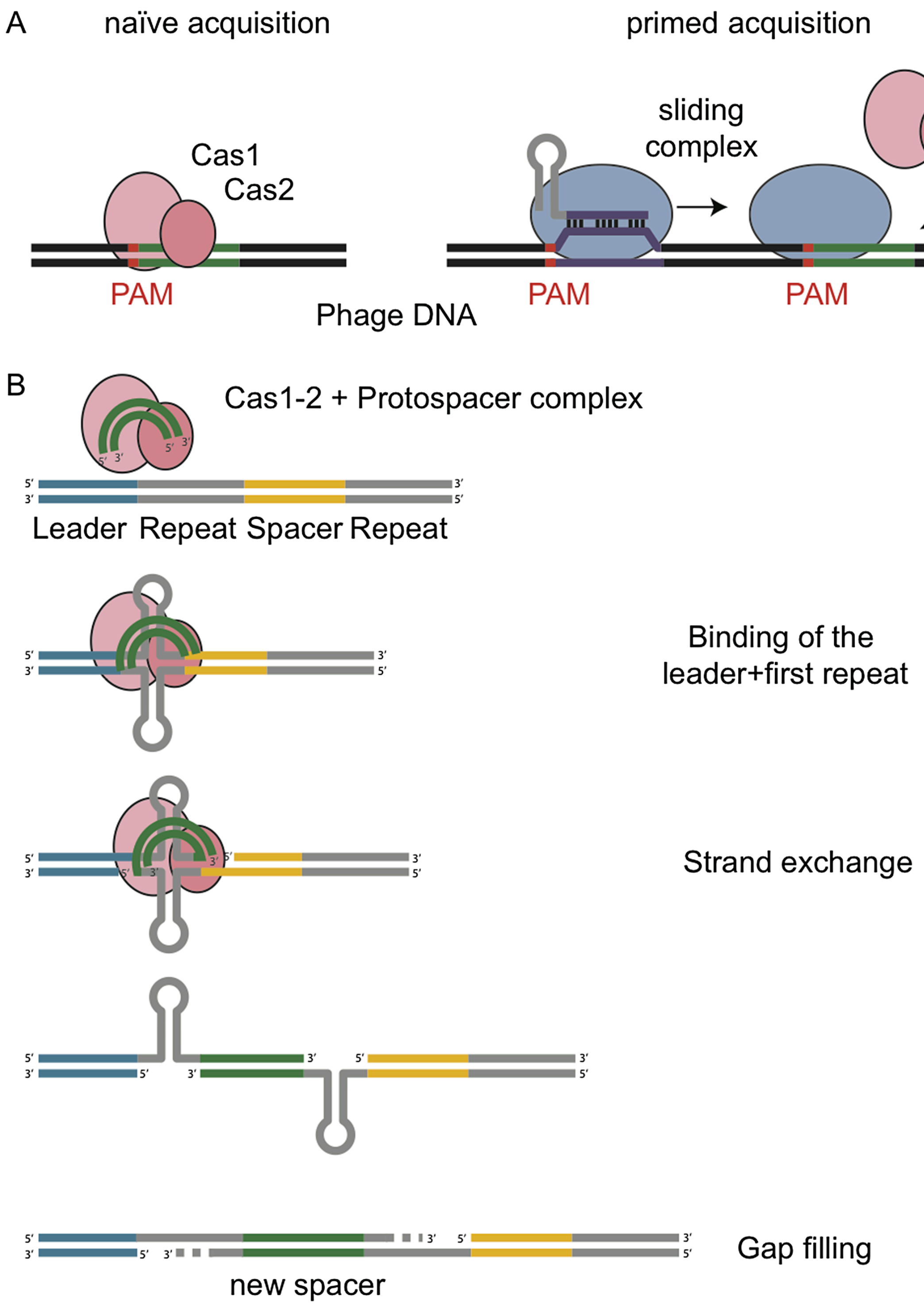
Chapter 1.2.2. Naïve and Primed Acquisition

Cas1 and Cas2 are the only Cas proteins universally conserved across all types and subtypes of CRISPR-Cas systems[10,24]. Initial studies on the role of Cas proteins revealed that mutations or deletions of Cas1 and Cas2 did not impact interference and crRNA maturation in type I[42], type II[43,44], and type III[45,46]. These observations led to the hypothesis that the two universal Cas proteins might be involved in adaptation. Indeed, in *E. coli*, overexpression of both *cas1* and *cas2* alone in the absence of the other *cas* genes was sufficient to acquire new plasmid-derived or host-derived spacers[26,28]. Yosef et al. (2012) also found that Cas1/2 mediate the preferential acquisition of spacers with a correct PAM, demonstrating that there is a mechanism to select spacer sequences flanking this motif, as opposed to the random acquisition of DNA sequences followed by the selection of those with correct PAMs. The biochemical properties of Cas1 support its role in spacer acquisition. *Pseudomonas aeruginosa* Cas1 has been shown to bind dsDNA in a sequence-independent manner with high affinity, and to work as a metal-dependent endonuclease which cleaves dsDNA into short fragments, that might serve as precursors for new spacers[47,48]. Nonetheless, it is still unclear whether new spacers are cut or copied from the invading molecule. Because Cas1 also has the ability to resolve Holliday junctions and thus promote DNA integration and recombination events, it could promote the integration of spacer sequences into the repeat-spacer array. While many Cas2 crystal structures have been solved and studied biochemically[49–51], a general consensus regarding its activity has not been reached.

The Cas1/2-mediated acquisition can add repeat-spacer units to a minimal CRISPR locus consisting of only one repeat sequence[28]. This indicates that this mechanism of acquisition does not require the presence of any other spacers, i.e. a previous exposure to the same or related phages and therefore is referred as "naïve" acquisition. This is in contrast to "primed" acquisition, where the presence of spacers with a full or partial match to the target DNA increases the frequency of acquiring another spacer (Figure 1-2a).

Primed acquisition has been studied in *E. coli*, which CRISPR-Cas system harbors the genes encoding the Cascade (CRISPR associated complex for antiviral defense) complex that contains the crRNA guide and is responsible for target recognition[42] and the Cas3 nuclease responsible for target cleavage[22], in addition to Cas1 and Cas2. One study showed that this CRISPR system can acquire spacers from a plasmid present in the cell, resulting in plasmid curing[27]. While the acquisition of a single spacer is enough to cure the plasmid, multiple

Figure 1-2. A model for the acquisition of new spacers. (a) Naïve vs. primed spacer acquisition. Upon lytic infection with a phage previously not encountered, incorporation of a new spacer into the CRISPR array ensures cell survival. Only protospacers adjacent to PAMs are sampled. Naïve adaptation (left) requires the concerted action of Cas1 and Cas2 alone. Primed acquisition (right) presupposes the existence of a non-targeting crRNA with partial homology to a region of the infecting phage (purple). Following the low-affinity target recognition by the interference machinery, the complex slides along the target DNA and, aided by Cas1 and Cas2, and recruits spacers from the same strand at a high rate. **(b)** A model for spacer integration onto the CRISPR array. The protospacer (green) is acquired from the viral genome and inserted into the CRISPR array at the leader-proximal end. Upon integration, the first pre-existent repeat serves as a template for the new repeat. In this model I speculate that the palindromic sequence of the repeat allows it to fold into DNA hairpins (grey).


A
naïve acquisition
primed acquisition
Cas1
Cas2
sliding
complex
PAM
Phage DNA
PAM
PAM
B
Cas1-2 + Protospacer complex
Leader
Repeat
Spacer
Repeat
Binding of the
leader+first repeat
Strand exchange
Gap filling
new spacer

spacers were frequently incorporated into the CRISPR array. Interestingly these were always acquired from the same strand of DNA. This led to the hypothesis that the acquisition of one spacer can trigger the acquisition of additional spacers from the same strand of the target DNA[27]. These authors hypothesize that the Cascade complex is directed to bind the foreign nucleic acid by a spacer already present in the array. If the match is good enough to trigger interference, Cas3 will degrade the dsDNA and these cleavage products can be used by Cas1 and Cas2 as precursors for new spacers recruited from the same strand.

Another study reported primed adaptation during infection of *E. coli* with the M13 phage[26]. The acquisition of new spacers was much more frequent when a spacer already targeting the M13 phage was present in the array. The authors showed how the orientation of the priming target determines the orientation of new protospacers. Interestingly, priming events occurred when mismatches that abolish interference were present between the crRNA and its target. This suggests that degradation of the target DNA by CRISPR interference is not necessary to prime adaptation. It is hypothesized that the Cascade complex can bind an imperfect target and trigger the acquisition of new spacers from the same molecule. Mutagenesis of the *cas* genes showed that in addition to Cas1 and Cas2, primed adaptation requires the Cascade complex and Cas3. Subsequent high-throughput analysis of spacer acquisition in *E. coli* confirmed that spacers are preferentially acquired from the primed strand with a 10-fold bias[39]. Assuming that spacers acquired from the non-primed strand are due to naïve adaptation, the authors conclude that primed adaptation occurs at much higher rates than naïve adaptation. Recently, a comprehensive study performed in type I-E showed that priming is nucleotide-dependent, as well as sensitive to the number of mutations and their locations with the target. Accordingly, high-throughput plasmid-loss assays revealed that priming tolerates up to 13 mutations within the PAM and protospacer. While the nucleotide-dependence of priming appears to be a more complex mechanism that needs to be further characterized, it appears that G-rich spacers are more likely to prime a better adaptation response.

These observations led to the "sliding" hypothesis for primed acquisition: after a low-affinity target recognition by the Cascade-crRNA complex, the complex slides along the target DNA randomly stopping at PAM sequences to recruit more spacers from the same strand (Figure 1-2a). This hypothesis was tested by several studies with results that both corroborated or challenged it. In a recent study of the *E. coli* type I-E system, Savitskaya *et al.*[39] argue that sliding from the priming position should lead to a preferential acquisition of nearby spacers, producing a gradient in spacer acquisition frequency relative to the priming location, which they did not observe. Moreover, insertion of poly-PAM blocks on the target molecule next to the priming site failed to halt the putative sliding acquisition machinery. Notably, these experiments were carried out on rather short (~ 3kb) circular plasmids that might have obstructed acquisition gradients caused by the priming site or poly-PAM brakes. In contrast, a recent study of *Haloarcula hispanica* type I-B system showed that protospacers in close-proximity of the priming protospacers were sampled more often than protospacers located farther away[31], thus supporting the sliding hypothesis. Regions located both upstream and downstream of the priming protospacer were highly sampled, but from opposite strands. Therefore, the authors proposed that sliding also involves stochastic Cas3 flipping from one strand of the DNA to the other. These contradicting results could be explained either by the presence of multiple priming sites or differences in the sliding dynamics, since a fast-sliding Cascade would prevent the generation of a positional gradient of acquired spacers.

The priming mechanism has likely evolved as a way to counteract phage mutants that escape CRISPR immunity by single point mutations in the target sequence. The spacers matching mutated targets cannot direct cleavage but can still be used to trigger the acquisition of new spacers and adapt against an evolving threat. Furthermore, priming favors the acquisition of multiple spacers targeting the same DNA molecule, which reduces the probability of escape and strengthens resistance. Notwithstanding the benefits of primed spacer acquisition, naïve adaptation remains crucial to detect unknown foreign molecules and is probably a universal feature of CRISPR systems.

Chapter 1.2.3. Spacer Integration

Once a target has been selected on the invading genome, it has to be incorporated into the CRISPR array. During this process, not only the spacer is incorporated, but also a new repeat is added to the array. Spacer insertion is polar, since the vast majority of new spacers are incorporated at the 5' end of the array, upstream of the first repeat[5,16,52–56]. Little is known about this process; however, the first repeat and the region immediately upstream of it, known as the leader sequence, seem to play a role.

The 200-500 bp region located upstream of the first repeat contains an A/T rich leader sequence that usually harbors the promoter of the CRISPR array, but it is also involved in adaptation. Deleting or scrambling the 60 nucleotides immediately adjacent to the array of the type I-E CRISPR-Cas system of *E. coli* prevents spacer acquisition[28]. This indicates that the leader contains specific sequence motifs essential for adaptation. Interestingly, deleting the Pribnow box required for the transcription of the array does not prevent spacer acquisition[28], suggesting that transcription is not essential for the adaptation process. It is believed that the leader sequence is recognized by the acquisition machinery. Evidence supporting this hypothesis comes from a study showing that Cas1 and Cas2 from the *E. coli* K12 strain can direct spacer acquisition in the CRISPR array of the O157:H7 strain, which carries a different leader sequence[57]. However, this artificial leader-Cas combination led to frequent abnormal acquisition events where the spacers were integrated in the wrong orientation. This suggests that the interaction between Cas proteins and the leader sequence determines the orientation of newly acquired spacers. Furthermore, in some instances, the insertion site was shifted by 2 bases, suggesting that the acquisition complex is anchored at the leader-repeat boundary where a first cut is made, and uses a ruler mechanism to cut the other strand on the other side of the repeat. The nucleotide content of the spacer is also thought to impact the orientation of newly acquired spacers. In type I-E, an underrepresentation of G and overrepresentation of C at the end of highly acquired spacers may serve as signals for insertion in the correct orientation.

The presence of a single repeat has been shown to be necessary and sufficient for both naïve and primed adaptation in the type I-E CRISPR system[26,58], and the presence of additional repeats does not increase the rate of acquisition of new spacers[41]. Interestingly, spacers incorporated into a minimal CRISPR array (one repeat, no preexisting spacers) have the correct length[58], suggesting that the protein machinery, rather than preexistent repeat-spacer units, dictates the size of additional spacers. In type I-E systems, the new repeat (29 nt long) is copied from the first repeat in the array since point mutations introduced in the first repeat are replicated in newly incorporated spacer-repeats units[26,28]. Interestingly, mutations of the last nucleotide of the repeat were not passed on to new repeats, indicating that only bases 1 through 28 of the repeat serve as a template for new repeats. In contrast, the 29th base originates from the protospacer and represents the last nucleotide of the PAM[26,59]. While the last nucleotide of the 5'-AWG-3' PAM is highly conserved in *E. coli*, this is not the case in many other systems where this mode of repeat duplication remains to be determined. Based on these results and on the known mechanisms of insertion of transposable elements[60] and retroviruses[61] a model for spacer acquisition has emerged (Figure 1-2b). The first repeat sequence of the CRISPR locus is subjected to ssDNA nicking at the 3’ end of each repeat strand. This cleavage could be facilitated by the stem-loop structure that can form on most repeats due to their partially palindromic sequences. Proximity to the leader would provide recognition of the first repeat and/or help recruit the spacer acquisition machinery. The free 3’-ends of repeats are ligated to the 5’ end of viral fragments, leading to the insertion of a new spacer and the generation of a staggered intermediate. The gaps are filled by DNA polymerase I, thus adding a new repeat to the array. Current research across different laboratories is testing this model.

Chapter 1.3. Unanswered Questions

Although recent studies have established molecular requirements as well as a general mechanism for the acquisition phase, many details of CRISPR adaptation are still poorly understood. An extra layer of complexity is added by the many different types of CRISPR-Cas systems, some of which could have different mechanisms of spacer acquisition. Indeed, variations in the way spacers are acquired from the target likely exist between CRISPR types and subtypes. In the type I-A of the crenarchaeon *Thermoproteus tenax*, Cas1 and Cas2 are fused as a single protein that forms the CRISPR-associated complex for the integration of spacers (Cascis) together with Csa1 and Cas4[62]. In type II, Csn2 was reported to be required for adaptation[5]. Structural studies have revealed that this protein forms a ring-like structure around DNA, suggesting that it might recruit other proteins to the protospacer and could form a sliding clamp that facilitates primed acquisition[63]. In type III, interference is transcription dependent and requires crRNAs that are antisense to their cognate RNA targets[64]. Since transcription of CRISPR loci is unidirectional, type III spacers need to be incorporated into the correct orientation in order to produce functional crRNAs. The underlying mechanism of this requirement is not understood.

I believe that future research will focus on understanding how the leader sequence is recognized, how the first repeat is cleaved, and the new spacer ligated in a way that allows the generation of an additional repeat, and whether the length of the array is regulated. In addition, it is still unknown why spacers are acquired preferentially from certain molecules or certain positions on a given molecule, and whether any mechanisms truly exist to prevent, or at least limit, the of self-targeting spacers. An interesting question also arises from the observation that the PAM motif is recognized both during acquisition and interference, and that the motif might be recognized by different protein complexes in each of these stages of the CRISPR immunity pathway. As the exact sequence requirements might be different for these two functions, it has been proposed to use the term Spacer Acquisition Motif (SAM) when referring to the sequence recognized by the acquisition machinery and the term Target Interference Motif (TIM) when referring to the sequence recognized by the interference machinery[65]. A mutation that affects SAM recognition might lead to the acquisition of spacers that will not be effective during the interference stage. Conversely, a mutation that affects the TIM recognition might render preexisting as well as newly acquired spacers useless. In the presence of this apparent evolutionary bottleneck, PAM sequences might be expected to be highly conserved, yet an extensive diversity has been described. How these two facts can be reconciled remains to be investigated.

Chapter II.

The Role of Cas9 in Spacer Acquisition

This chapter investigates the molecular requirements of spacer acquisition in the type II-A CRISPR system of *Streptococcus pyogenes*. Cas9 cleavage of the viral genome requires the presence of a 5′-NGG-3′ protospacer adjacent motif (PAM) sequence immediately downstream of the viral target. It is not known whether and how viral sequences flanked by the correct PAM are chosen as new spacers. Here I show that Cas9 selects functional spacers by recognizing their PAM during spacer acquisition. The replacement of *cas9*with alleles that lack the PAM recognition motif or recognize an NGGNG PAM eliminated or changed PAM specificity during spacer acquisition, respectively. Cas9 associates with other proteins of the acquisition machinery (Cas1, Cas2 and Csn2), presumably to provide PAM-specificity to this process. This chapter introduces a new function for Cas9 in the genesis of prokaryotic immunological memory.

2.1. The Protospacer Adjacent Motif is Highly Conserved

Based on their *cas* gene content, CRISPR-Cas systems can be classified into six distinct types, I-VI[24,25]. Each CRISPR-Cas type possesses different mechanisms of crRNA biogenesis, target destruction and prevention of autoimmunity. In the type II CRISPR-Cas system present in Streptococcus pyogenes the Cas9 nuclease inactivates infective phages using crRNAs as guides to introduce double-strand DNA breaks into the viral genome[23]. Cas9 cleavage requires the presence of a protospacer adjacent motif (PAM) sequence immediately downstream of the protospacer[34,67]. This requirement avoids the cleavage of the spacer sequence within the CRISPR array, i.e. autoimmunity, since the adjacent repeat lacks a PAM sequence. The importance of the PAM sequence for target recognition and cleavage[34,67–69] suggests the presence of a mechanism to ensure that newly acquired spacer sequences match protospacers flanked by a proper PAM sequence. For the type I-E CRISPR-Cas system of Escherichia coli, over-expression of cas1 and cas2 is sufficient for the acquisition of new spacers in the absence of phage infection. Reports indicate that spacers acquired in this fashion match preferentially (25–70%, depending on the study) to protospacers with the correct PAM (AWG, W=A/T)[26,28,57,70], suggesting that Cas1 and Cas2 are sufficient for spacer acquisition and have some intrinsic ability to recognize protospacers with the right PAM. In the type II system of S. pyogenes the PAM sequence is NGG (and also NAG at a much lower frequency)[4,33,34], where N is any nucleotide, and it is recognized and bound by a domain within the

Cas9 tracrRNA:crRNA-guided nuclease during target cleavage[67,71]. How spacers are acquired in this system, particularly how spacers with correct PAM sequences are selected during this process, is not known.

2.2. Cas9 is required for spacer acquisition

To investigate the mechanisms of recognition of PAM-adjacent protospacers during spacer acquisition, Icloned the type II-A CRISPR-Cas locus of S. pyogenes (Figure 2-1a) into the staphylococcal vector pC194 and introduced the resulting plasmid[64] into Staphylococcus aureus RN4220[72], a strain lacking CRISPR-Cas loci. Ichose this experimental system because it facilitates the genetic manipulation of the S. pyogenes CRISPR-Cas system. Ifirst tested the ability of the cells to mount adaptive CRISPR immunity by infecting them with the staphylococcal phage ϕNM4γ4, a lytic variant of ϕNM4[73] (see Methods for a description of ϕNM4γ4 isolation).

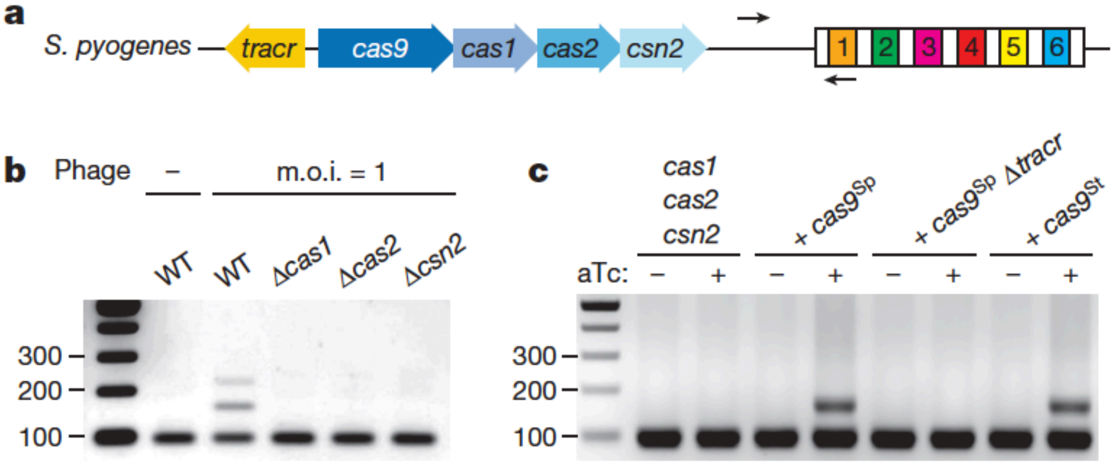
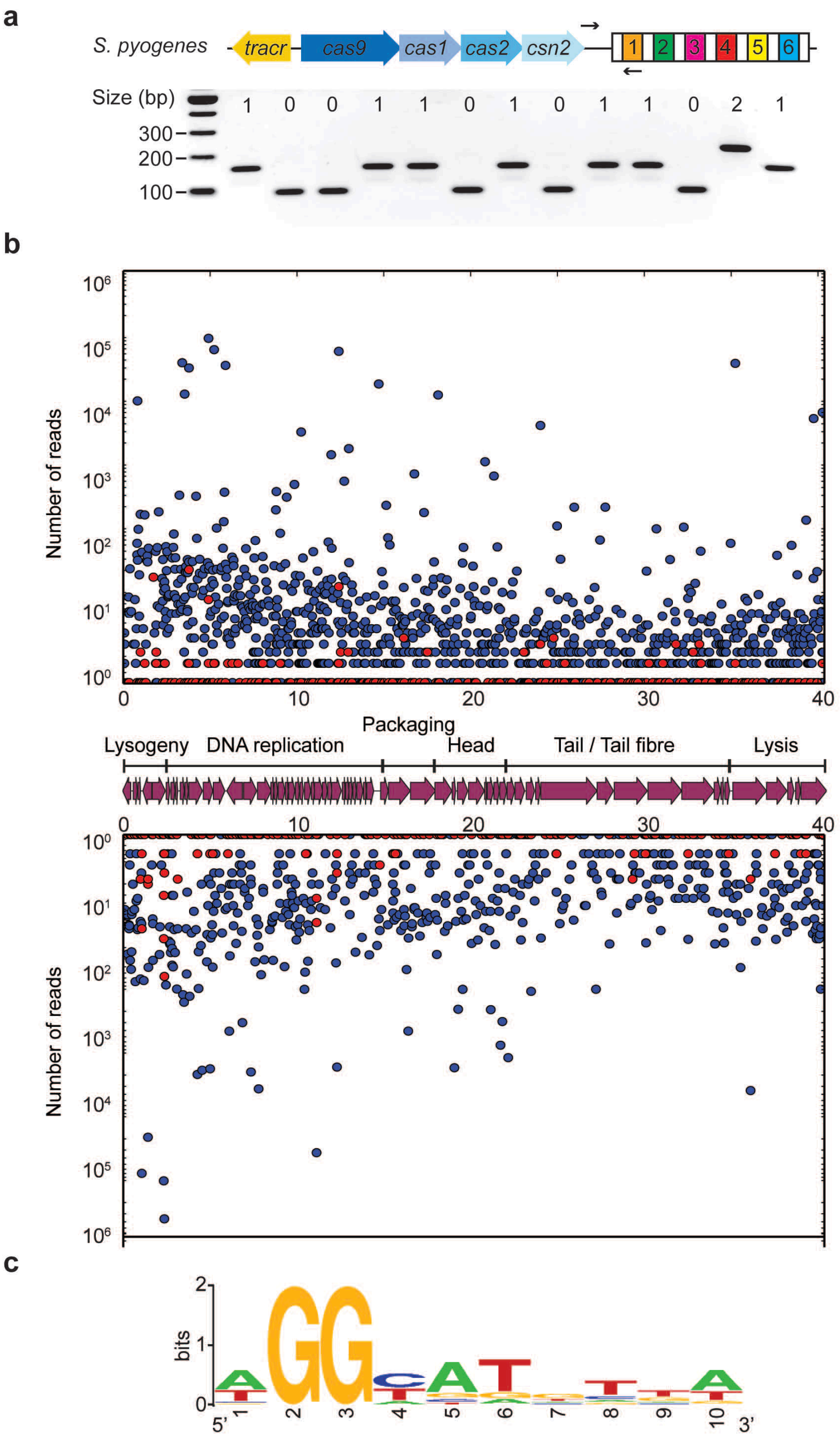


Figure 2-1. Cas9 is required for spacer acquisition. (a) Organization of the *S. pyogenes* type II-A CRISPR-Cas locus. Arrows indicate the annealing position of the primers used to check for the expansion of the CRISPR array. **(b)** PCR-based analysis of cultures to check for the acquisition of new spacer sequences. In the presence or the absence of phage ϕNM4γ4 infection. Wild-type (WT) as well as different *cas* mutants were analyzed. MOI; multiplicity of infection. **(c)** Cultures over-expressing Cas1, Cas2 and Csn2 under the control of a tetracycline-inducible promoter were analyzed using PCR for spacer acquisition in the absence of phage infection. The strain was complemented with plasmids carrying either Sp or St Cas9 (see Fig. S3). aTc; anhydrotetracycline.

Plate-based assays performed by mixing bacteria and phage in top agar allowed the selection of phage-resistant colonies that were checked by PCR to look for the expansion of the CRISPR array (Figure 2-2a). On average 50 % of the colonies acquired one or more spacers (8/13, 5/11 and 7/16 in three independent experiments), whereas the rest of the resistant colonies survived phage infection by a non-CRISPR mechanism, most likely including phage receptor mutations (Figure 2-3a). To maximize the capture of new spacer sequences, Iperformed the same assay in liquid and recovered surviving bacteria at the end of the phage challenge. These were analyzed by PCR of the CRISPR array and the amplification products of expanded loc I was subjected to Illumina MiSeq sequencing to determine the extent of spacer acquisition. Analysis of 2.96 million reads detected protospacers adjacent to 2083 out of 2687 NGG sequences present in the viral genome, although with variation in the frequency of acquisition of each sequence (Figure 2-2b). The data revealed a prominent selection of spacers matching protospacers with downstream NGG PAM sequences (99.97 %, Figure 2-2c). The acquisition of new spacers by cells in liquid culture proved to be simple and highly efficient, providing the possibility to look at millions of new spacers in a single step. It was therefore implemented in the rest of our studies.

To determine the genetic requirements for spacer acquisition Imade individual deletions of cas1, cas2 or csn2 and challenged the mutant strains with phage ϕNM4γ4. Spacer acquisition was decreased to levels below our limit of detection in each of these mutants (Figure 2-1b), corroborating previous experiments[26,43]. Therefore, while Cas1, Cas2 and Csn2 are dispensable for anti-phage immunity in the presence of a pre-existing spacer (Figure 2-3b and c), they are required for spacer acquisition.

Figure 2-2. The S. pyogenes type II CRISPR–Cas system displays a strong bias for the acquisition of spacers matching viral protospacers with NGG PAMs. (a) Analysis of bacteriophage-insensitive mutant colonies using PCR and agarose gel electrophoresis, representative of five technical replicates. Bacteria and phage were mixed in top agar and incubated overnight. DNA was isolated from individual colonies resistant to phage infection and used as template for a PCR reaction with primers (arrows) H182 and H183, which amplify the end of the S. pyogenes CRISPR array. The size of the PCR band indicates the number of new spacers (shown at the top of the gel). Cells without additional spacers resist infection by a CRISPR-independent mechanism, presumably envelope resistance. **(b)** Analysis of acquired spacers during phage infection of a population of bacteria carrying the S. pyogenes type II CRISPR–Cas system. Liquid cultures of bacteria were infected with phage, surviving cells were collected at the end of the infection, DNA extracted and used as template for a PCR reaction as described above. Amplification products were separated by agarose gel electrophoresis and the DNA of the bands corresponding to products with additional spacers was extracted and sent for MiSeq next generation sequencing. Reads corresponding to newly acquired spacers were plotted according to their position in the phage genome (x axis) and their abundance (y axis). Each dot represents a unique spacer sequence; blue and red dots indicate a corresponding protospacer with an NGG or non-NGG PAM. Top and bottom plots indicate protospacers in the top and bottom strands of the viral DNA. The map as well as the different functions of the phage genes are indicated in between the plots. **(c)** Weblogo showing the conservation of the 59 flanking sequences of 10,000 protospacers randomly selected from the experiment shown in b. Absolute conservation of the NGG PAM was observed.


a
S. pyogenes
tracr
cas9
cas1
cas2
csn2
1
2
3
4
5
6
Size (bp)
300
200
100
1 0 0 1 1 0 1 0 1 1 0 2 1
b
Number of reads
10^6
10^5
10^4
10^3
10^2
10^1
10^0
0
10
20
30
40
Packaging
Lysogeny
DNA replication
Head
Tail / Tail fibre
Lysis
c
bits
2
1
0
5'
1 2 3 4 5 6 7 8 9 10
3'

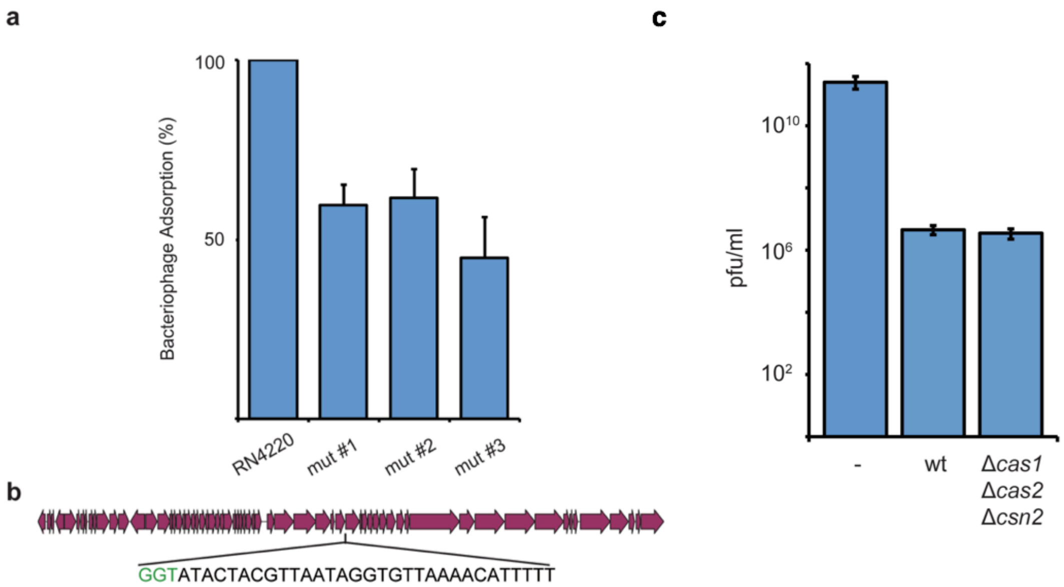
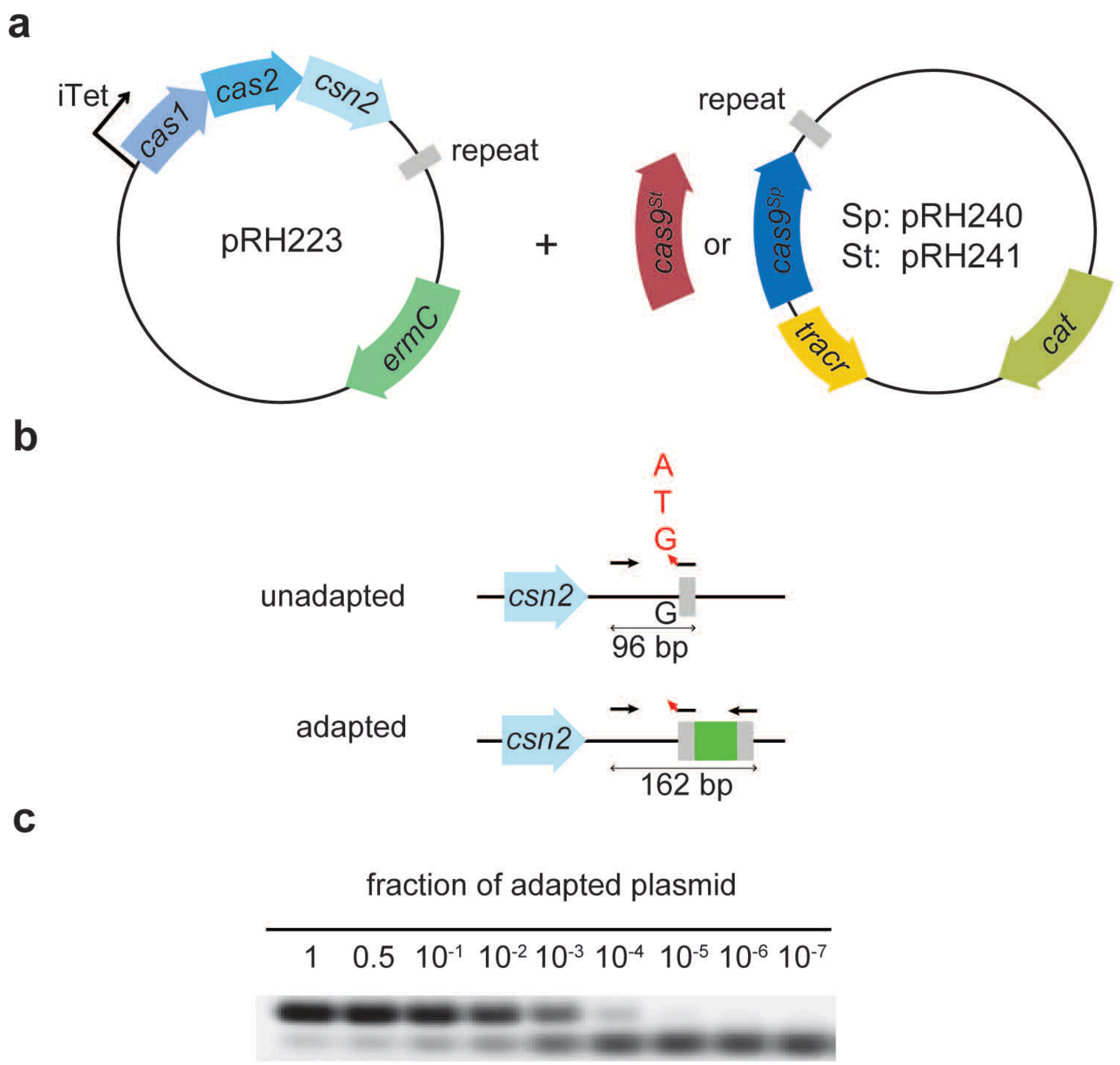


Figure 2-3. Cas1, cas2 and csn2 are not required for the execution of immunity. **(a)** Analysis of bacteriophage-resistant mutants that do not acquire a new spacer. Three colonies that survived phage infection in our in-plate adaptation assay (Figure 2-2) were subjected to phage adsorption assay. Briefly, surviving colonies as well as the wild-type S. aureus RN4220 control were grown in liquid and mixed with bacteriophage. After a brief incubation, cells were pelleted by centrifugation and the phages present in the supernatant (unable to bind and infect cells) were counted on a lawn of sensitive cells. The number of plaque-forming units (p.f.u.) of a control experiment in the absence of host cells were used to determine the 100% free phage, or 0% adsorption value. No plaques were observed in the control experiment using wild-type cells and this value was used to set the 100% adsorption limit. The three CRISPR-independent, bacteriophage-resistant mutants displayed a marked defect in phage adsorption (about 50%), indicating that most likely they carry envelope resistance mutations. **(b)** cas1, cas2 and csn2 are not required for the execution of immunity using previously acquired spacers. Position within the phage NM4 genome of the type II CRISPR–Cas target used in this experiment. The protospacer sequence is in the bottom strand (shown in 3’–5’ direction) and flanked by a TGG PAM (in green). **(c)** Comparison of immunity provided by a type II CRISPR–Cas system programmed to target the sequence shown in panel **(a)** in the presence (wild-type, wt) or absence (dcas1,dcas2, dcsn2) of cas1, cas2 and csn2. Immunity is measured as the p.f.u. of a phage lysate spotted on top agar lawns of S. aureus RN4220 cells containing no CRISPR system, a wild-type S. pyogenes CRISPR–Cas type II system (wt, pRH233), or the same CRISPR–Cas systems with a deletion of cas1, cas2 and csn2 genes (dcas1, dcas2, dcsn2, pRH079).

To determine whether these genes are also sufficient for this process, Iover-expressed cas1, cas2 and csn2 in the absence of cas9 using a tetracycline-inducible promoter in plasmid pRH223 and looked for the integration of new spacers in the absence of phage infection using a highly sensitive PCR assay (Figure 2-4). Iwere unable to detect new spacers even in the presence of the inducer (Fig. 2-1c). However, the addition of a second plasmid expressing tracrRNA and Cas9 from their native promoters (Figure 2-4a) enabled spacer acquisition only in the presence of the inducer, with all the new spacers matching chromosomal or plasmid sequences (Figure 2-1c). Although most likely the acquisition of such spacers causes cell death or plasmid curing, respectively, the acquisition event can still be detected in liquid culture using our highly sensitive PCR assay (Figure 2-4b and c). The tracRNA (Figure 2-1a) is a small RNA bound by Cas9 that is required for crRNA processing and Cas9 nuclease activity. Iwondered if Cas9 involvement in spacer acquisition also required the presence of the tracrRNA. Deletion of the tracrRNA prevented spacer acquisition in the absence of phage infection (Figure 2-1c), suggesting that apo-Cas9 is not sufficient to promote spacer acquisition and that association with its cofactor is also required. Altogether these data indicate that Cas1, Cas2 and Csn2 are necessary but not sufficient for the incorporation of new spacers and that tracrRNA/Cas9 is also required. This is in contrast to the type I-E CRISPR-Cas system of E. coli, where over-expression of Cas1 and Cas2 alone is sufficient for spacer acquisition. It is important to note that the CRISPR array used in this assay consists of a single repeat, without pre-existing spacers (Figure 2-4). Therefore, the Cas9 requirement is not a consequence of the phenomenon known as “primed” spacer acquisition. This refers to an increase in the frequency of spacer acquisition observed in type I CRISPR-Cas systems that relies on the presence of a pre-existing spacer with a partial match to the phage genome as well as the full targeting complex (Cascade)[26,31,74].

Figure 2-4. Generation of an experimental system for the overexpression of cas1, cas2 and csn2 and the detection of spacer acquisition in the absence of phage infection. (a) Plasmids used in the spacer acquisition experiments presented in Figure 2-1c and 2-6c and d. pRH223 contains cas1, cas2 and csn2 from S. pyogenes under a tetracycline-inducible promoter. Cells containing this plasmid only acquired spacers when a second plasmid expressing cas9 was introduced, pRH240 or pRH241, containing the tracrRNA gene, the leader and first repeat from the S. pyogenes type II CRISPR–Cas system as well as cas9 from S. pyogenes (cas9Sp) or S. thermophilus (cas9St), respectively. The leader is a short, AT-rich sequence immediately upstream of the first repeat that contains the promoter for the transcription of the CRISPR array. **(b)** Highly sensitive PCR assay to enrich for amplification products of adapted CRISPR loci. Arrows indicate primer annealing position and direction. The forward primer (JW8) anneals on the leader. For the reverse primer, a cocktail of JW3, JW4 and JW5 was used. The three reverse primers anneal on the repeat and differ only in their 3'-end nucleotide that never matches the last nucleotide of the leader (red arrowhead). Because this nucleotide is critical for the annealing of the primers, loci that acquire spacers ending in A, C or T are preferentially amplified over unadapted loci. **(c)** To quantify the sensitivity of this technique, Imixed pGG32 (one repeat, unadapted) with pRH087 (repeat-spacer-repeat, adapted) in known ratios. The amplification of adapted plasmid was detected even when it represented 0.01% of the total plasmid template, representative of three technical replicates. This highly sensitive PCR assay is not required to detect acquisition during phage infection, as in this case adapted cells survive and are enriched within the population, making their detection much easier.


a
iTet
cas1
cas2
csn2
repeat
pRH223
ermC
+
cas9St
or
repeat
cas9Sp
Sp: pRH240
St: pRH241
tracr
cat
b
A
T
G
unadapted
csn2
G
96 bp
adapted
csn2
162 bp
c
fraction of adapted plasmid
1
0.5
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}

2.3. Cas9 specifies the PAM sequence of newly acquired spacers

Given this newfound requirement in the CRISPR adaptation process and the well-established PAM recognition function of Cas9 during the surveillance and destruction of viral target sequences, I hypothesized that this nuclease could participate in the selection of PAM sequences during spacer acquisition. To test this I exchanged the cas9 genes of S. pyogenes (Sp) and S. thermophilus (St) CRISPR-Cas systems to create two chimeric CRISPR loci: tracrRNA^{Sp}-cas9St-cas1^{Sp}-cas2^{Sp}-csn2^{Sp} and tracrRNASt-cas9^{Sp}-cas1St-cas2St-csn2St (Figure 2-6a). I chose the type II-A CRISPR-Cas system of S. thermophilus (also known as CRISPR3) because it is an ortholog of the S. pyogenes system[75]. While the PAM sequence for the Sp CRISPR-Cas system is NGG, the PAM sequence for the St system is NGGNG[16] (Figure 2-6b). I infected each naïve strain with phage ϕNM4γ4, sequenced the newly acquired spacers, and obtained the PAM of the matching protospacers using WebLogo[76]. I found that each chimeric system acquired spacers with PAMs that correlated with the cas9, but not the tracrRNA, cas1, cas2 or csn2, allele present (Figure 2-6b). To rule out the possibility that non-functional spacers are negatively selected during phage infection, i.e. they are acquired randomly and only those cells containing spacers with a correct PAM for Cas9 cleavage provide immunity and allow cell survival, I sequenced the PAMs of spacers acquired in the absence of phage infection (Fig. 2-1c and 2-6c).

Either Cas9^{Sp} or Cas9St were produced in cells overexpressing Cas1^{Sp}, Cas2^{Sp} and Csn2^{Sp}. In this experiment, as explained above, spacers matching chromosomal or plasmid sequences will be acquired. The PCR products containing new spacers were cloned into a commercial vector from which they were sequenced. Expression of Cas9^{Sp} led to the incorporation of spacers matching protospacers with an NGG PAM sequence, whereas the expression of Cas9St in the same cells shifted the composition of the PAM to NGGNG (Fig. 2-6d). These results demonstrate that Cas9 specifies PAM sequences to ensure the acquisition of functional spacers during CRISPR adaptation.

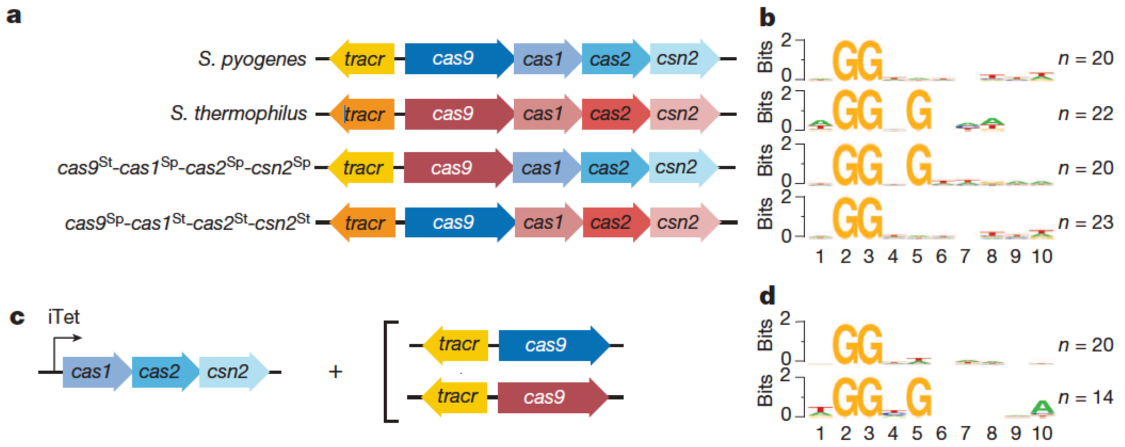


Figure 2-6. Cas9 determines the PAM sequence of acquired spacers. (a) (c) Genetic composition of the CRISPR–Cas loci tested for spacer during phage infection (a), or in the absence of infection (c), with the experimental set up shown in Figure 2-4. **(b) (d)** Sequence logos obtained after the alignment of the 3' flanking sequences of the protospacers matched by the newly acquired spacers in panels **(a)** and **(c)** respectively. Numbers indicate the positions of the flanking nucleotides downstream from the spacer. Number of sequences used in each alignment indicated as n.

2.4. Cas9 associates with other Cas proteins involved in spacer acquisition

In type I CRISPR-Cas systems, Cas1 and Cas2 form a complex[70] and the dsDNA nuclease activity of Cas1 has been implicated in the initial cleavage of the invading viral DNA to generate a new spacer[48]. The genetic analyses presented above suggest that in the type II S. pyogenes CRISPR-Cas system, the PAM-binding function of Cas9 observed in vitro[67] could specify a PAM-adjacent site of cleavage for Cas1, or other members of the spacer acquisition machinery. This would guarantee that newly acquired spacers have the correct PAM needed for Cas9 activity later in this immune pathway. This hypothesis predicts an interaction between Cas9 and Cas1, Cas2 and/or Csn2. To test this, I expressed the type II Cas operon in E. coli, using a histidyl tagged version of Cas9, and looked for other proteins that co-purify. I observed an abundant co-purifying protein with an apparent molecular weight close to 33 kDa, the expected size of Cas1 (Figure 2-5).

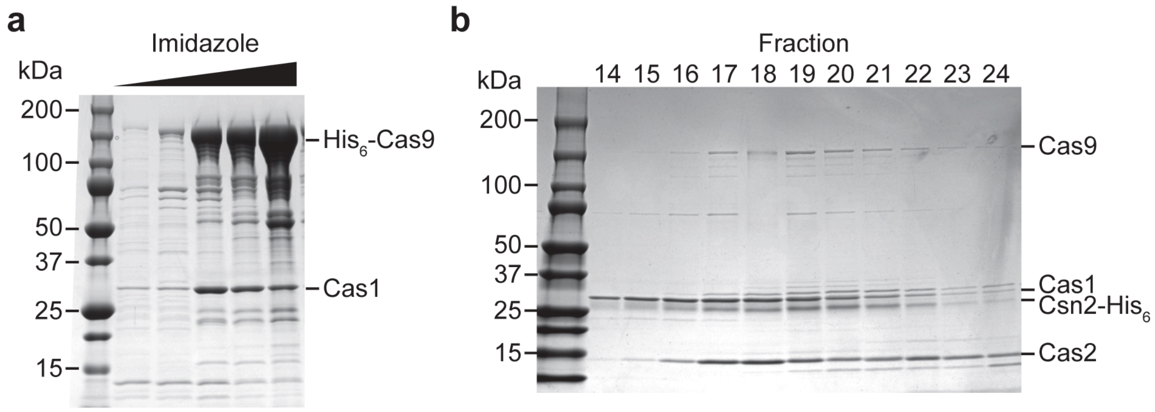
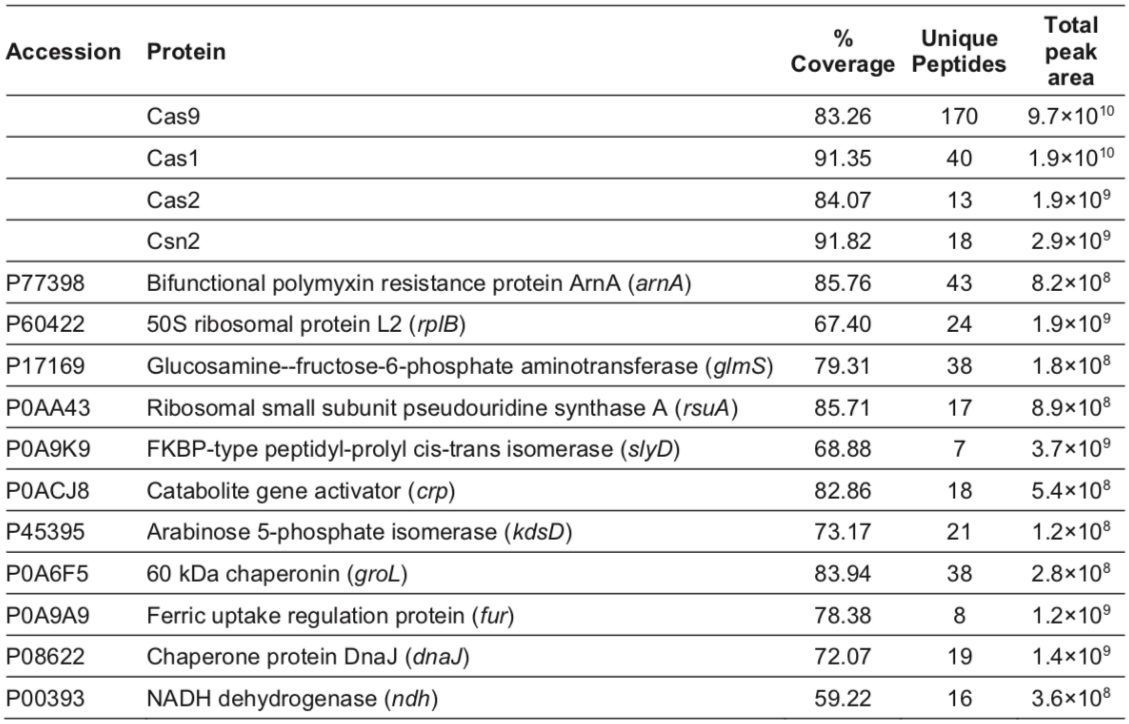


Figure 2-5. Purification of a Cas9–Cas1–Cas2–Csn2 complexes. (a) *The cas9–cas1–cas2–csn2* operon of S. pyogenes SF370 was cloned into the pET16b vector (generating pKW07) to add an N-terminal histidyl tag to Cas9 and express all proteins in E. coli. Purification was performed using Ni-NTA affinity chromatography. SDS–PAGE followed by Coomassie staining of the purified proteins revealed a co-purifying protein that was identified as Cas1 by mass spectrometry, in a result representative of five technical replicates. **(b)** The cas9–cas1– cas2–csn2 operon of S. pyogenes SF370 was cloned into the pET23a vector (generating pKW06) to add a C-terminal histidyl tag to Csn2 and express all proteins in E. coli. Purification was performed using Ni-NTA affinity chromatography followed by ion exchange chromatography. The elution fractions that constituted the peak containing the complex (Figure 2-7a) were separated by SDS–PAGE and visualized

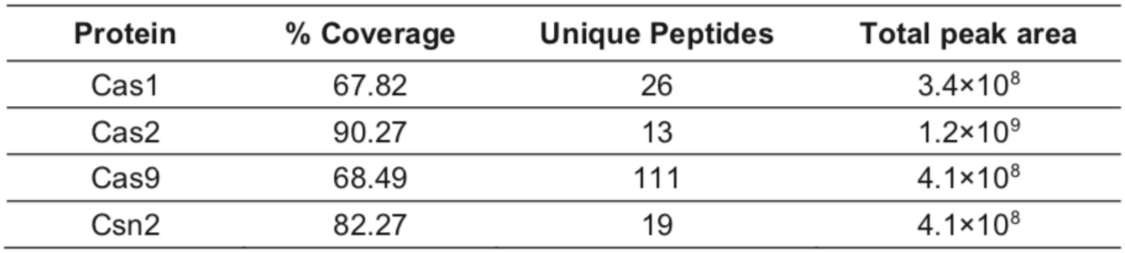
Mass spectrometry confirmed the identity of both of these proteins as well as the presence of Cas2 and Csn2 co-purifying with Cas9 (Table 1). This result suggested the formation of a Cas9-Cas1-Cas2-Csn2 complex and therefore I explored other purification strategies to unequivocally determine its existence. I was able to isolate a Cas9-Cas1-Cas2-Csn2 complex when the histidyl tag was added to Csn2 (Figure 2-7a and b). The identity of the purified proteins was confirmed by mass spectrometry (Table 2). This demonstrates a biochemical link between the Cas9 nuclease and the other Cas proteins that function exclusively to acquire new spacers, supporting the role of Cas9 as a PAM specificity factor in the adaptation phase of CRISPR immunity.

Table 1. Mass spectrometry analysis of proteins purified through Ni-NTA



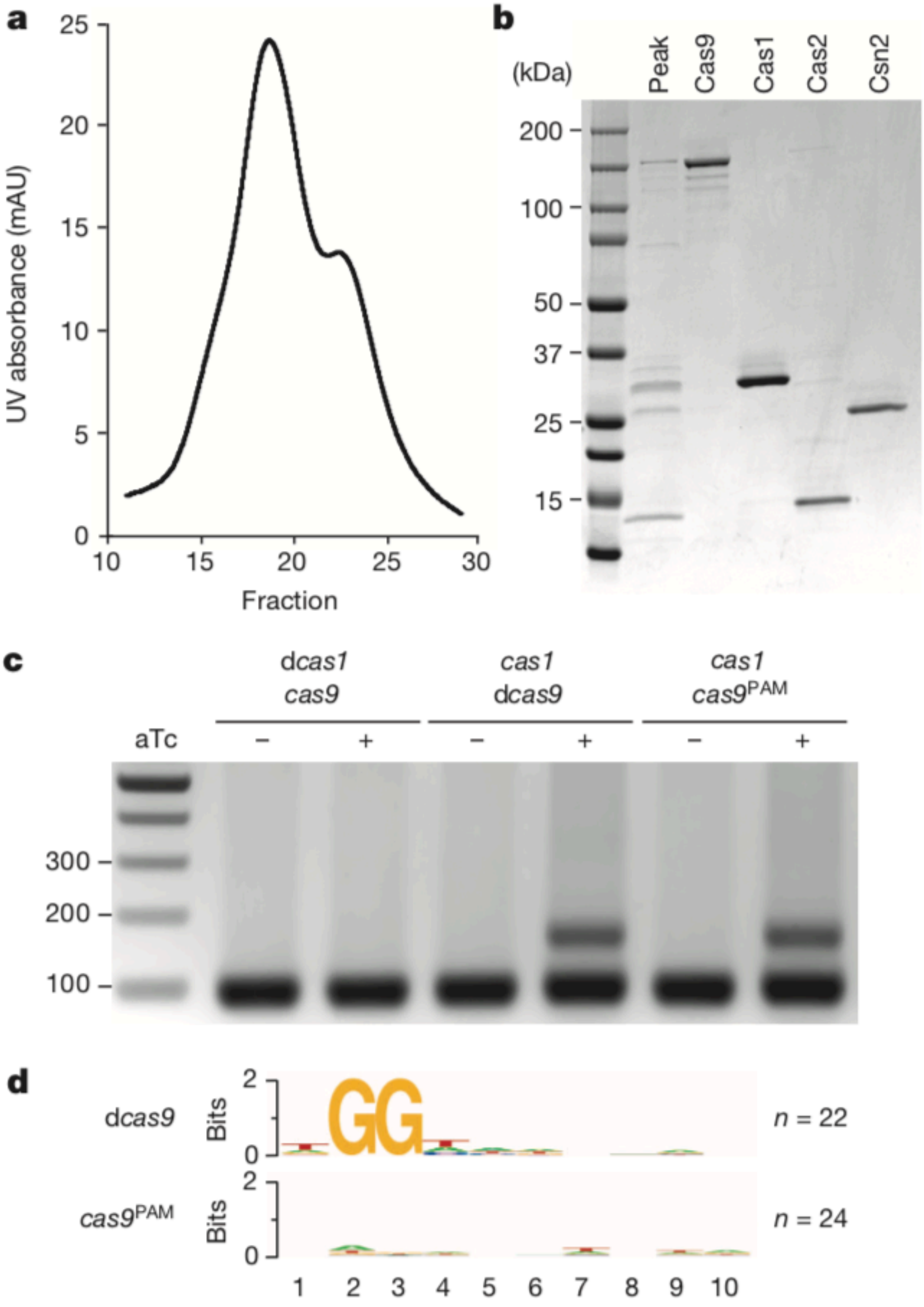
Accession	Protein	% Coverage	Unique Peptides	Total peak area
	Cas9	83.26	170	9.7×10^{10}
	Cas1	91.35	40	1.9×10^{10}
	Cas2	84.07	13	1.9×10^{9}
	Csn2	91.82	18	2.9×10^{9}
P77398	Bifunctional polymyxin resistance protein ArnA (*arnA*)	85.76	43	8.2×10^{8}
P60422	50S ribosomal protein L2 (*rplB*)	67.40	24	1.9×10^{9}
P17169	Glucosamine--fructose-6-phosphate aminotransferase (*glmS*)	79.31	38	1.8×10^{8}
P0AA43	Ribosomal small subunit pseudouridine synthase A (*rsuA*)	85.71	17	8.9×10^{8}
P0A9K9	FKBP-type peptidyl-prolyl cis-trans isomerase (*slyD*)	68.88	7	3.7×10^{9}
P0ACJ8	Catabolite gene activator (*crp*)	82.86	18	5.4×10^{8}
P45395	Arabinose 5-phosphate isomerase (*kdsD*)	73.17	21	1.2×10^{8}
P0A6F5	60 kDa chaperonin (*groL*)	83.94	38	2.8×10^{8}
P0A9A9	Ferric uptake regulation protein (*fur*)	78.38	8	1.2×10^{9}
P08622	Chaperone protein DnaJ (*dnaJ*)	72.07	19	1.4×10^{9}
P00393	NADH dehydrogenase (*ndh*)	59.22	16	3.6×10^{8}

Table 2. Mass spectrometry analysis of protein bands from the purified Cas9–Cas1–Cas2–Csn2 complex



Protein	% Coverage	Unique Peptides	Total peak area
Cas1	67.82	26	3.4×10^{8}
Cas2	90.27	13	1.2×10^{9}
Cas9	68.49	111	4.1×10^{8}
Csn2	82.27	19	4.1×10^{8}

Figure 2-7. S. pyogenes Cas9 PAM recognition domain is required for the acquisition of spacers with an NGG PAM sequence. (a) Separation of the Cas9–Cas1–Cas2–Csn2 complex by ion exchange chromatography. **(b)** SDS– PAGE of fraction 19 (peak) from the complex elution shown in panel (a), representative of five technical replicates. The four proteins of the complex were individually purified and run alongside the purified fraction to identify each protein in the complex. **(c)** Spacer acquisition was tested as in Figure 2-1c in the presence or absence of different Cas1 or Cas9 activities. Image is representative of eight technical replicates. dCas1, nuclease-dead Cas1 (E220A mutation); dCas9, nuclease-dead Cas9 (D10A, H840A mutations); Cas9PAM lacks the PAM recognition function (R1333Q, R1335Q mutations). **(d)** Sequence logos obtained after the alignment of the 39 flanking sequences of the protospacers matched by the newly acquired spacers in panel (c). Numbers indicate the positions of the flanking nucleotides downstream from the spacer. Number of sequences used in each alignment indicated as n.


a
25
20
15
10
5
0
UV absorbance (mAU)
10
15
20
25
30
Fraction
b
(kDa)
Peak
Cas9
Cas1
Cas2
Csn2
200
100
50
37
25
15
c
dcas1
cas9
cas1
dcas9
cas1
cas9PAM
aTc
−
+
−
+
−
+
300
200
100
d
dcas9
Bits
2
0
GG
n = 22
cas9PAM
Bits
2
0
n = 24
1
2
3
4
5
6
7
8
9
10

2.5. The PAM binding motif of Cas9 is required for PAM selection

Within this complex the PAM-binding domain of Cas9 would specify a functional spacer (one adjacent to a correct PAM) and the nuclease activity of Cas1 and/or Cas9 would cleave the invading DNA to extract the spacer sequence. To test this I performed adaptation studies in the absence of phage selection as described in Figure 2-4 but using different combinations of wild-type Cas1, Cas1^{E220A} (catalytically dead or dCas1[48]), wild-type Cas9, Cas9^{D10A,H840A} (catalytically dead or dCas9[34]) and Cas9^{R1333Q,R1335Q} (Cas9PAM, containing mutations in the PAM-binding motif that substantially reduces binding to target DNA sequences with NGG PAMs in vitro[71]). I observed that the nuclease activity of Cas1 is necessary for spacer acquisition (Figure 2-7c). In contrast, the nuclease activity and PAM-binding function of Cas9 are dispensable for this process. Next I determined the PAM of the acquired spacers in the presence of mutated Cas9 (Figure 2-7). I found that whereas spacers acquired in the presence of dCas9 displayed correct PAMs, those acquired in the presence of Cas9PAM matched DNA regions without a conserved flanking sequence, i.e. without a PAM sequence. The same result was obtained with St dCas9 (Figure 2-8). Altogether these results indicate that Cas1 and Cas9 are part of a complex dedicated to spacer acquisition which requires Cas1 nuclease activity and Cas9 PAM-binding properties for the selection of new spacer sequences.

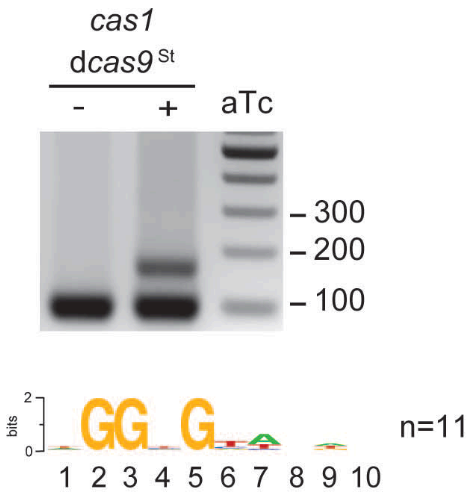


Figure 2-8. dCas9St can also support spacer acquisition. A plasmid derived from pRH241 containing mutations in the active site of S. thermophilus Cas9 (D10A, H847A; dCas9St) was used to characterize spacer acquisition in the absence of phage infection. Upon overexpression of Cas1, Cas2 and Csn2 using anydrotetracycline (aTc), I was able to detect spacer acquisition. Sequencing of spacers and alignment of the protospacer flanking sequences demonstrated the selection of an NGGNG PAM. The image is representative of three technical replicates.

2.6. Discussion

The selection of new spacers with a correct PAM is fundamental for the survival of the infected host during CRISPR-Cas immunity. In the simplest scenario there is no active selection of PAM-flanked protospacers; any spacer sequence can be acquired but only those with the correct PAM allow Cas9 cleavage of the invader and survival. Bacteria that acquire spacers with ineffective flanking sequences are killed by the virus and as a consequence PAM-flanking spacers are enriched in the population. Here I show that even in the absence of phage selection, the type II CRISPR-Cas system acquires new spacers with correct PAMs, a result that rules out the possibility of random spacer selection with subsequent selection for functional spacers. How are PAM-flanked protospacers selected during type II CRISPR-Cas immunity? One possibility is that the proteins exclusively dedicated to spacer acquisition perform the PAM-selection function. The inability of cells over-expressing only cas1, cas2 and csn2 to expand the CRISPR array strongly suggest that none of the proteins encoded by these genes can recognize and select correct PAMs. Another possibility is that the known PAM recognition function of Cas9[71,77], essential for destroying the invading virus, could also be used during spacer acquisition to recognize PAM-flanking viral sequences. Experiments showing that the cas9 allele, but not the cas1-cas2-csn2 alleles, determine the PAM sequence of the newly acquired spacers, demonstrated that this scenario is likely correct. How does Cas9 select new spacers with the correct PAMs? Our experiments demonstrate that Cas9 forms a stable complex with Cas1, Cas2 and Csn2 that presumably participates in the selection of new spacers.

The nuclease activity of Cas1, but not of Cas9, is required for spacer acquisition. The tracrRNA is also required, suggesting that the apo-Cas9 structure[77], very different from holo-Cas9[71], does not have the correct conformation to participate in spacer acquisition. The key residues involved in Cas9 PAM recognition are not required for spacer acquisition, but they are necessary for the incorporation of new spacers with the correct PAM sequence. This suggests that the reported non-specific DNA binding property of Cas9[34,67] is sufficient for spacer acquisition, but not for the selection of functional spacers. There are currently two models for the incorporation of new spacers into the CRISPR array, one where the future spacer sequence is cut from the invading viral DNA, the “cut and paste” model, and another where this sequence is copied from the viral genome, the “copy and paste” model[78]. In the context of the first model, our data suggests that, at a low frequency that may reflect the dynamics of spacer acquisition, Cas1 cleaves the invading genome to extract a new spacer sequence. However, on its own, Cas1 nuclease activity is non-specific[48]. Therefore, I propose that through the formation of the Cas9-Cas1-Cas2-Csn2 complex, Cas9 binding to PAM-adjacent sequences provides specificity to Cas1 endonuclease activity. In the “copy and paste” model, Cas1 nuclease activity is most likely necessary for downstream events, such as the cleavage of the repeat sequence that precedes spacer insertion, and Cas9 is required to “mark” sequences adjacent to GG motifs to be copied into the CRISPR array. In any case, following yet unknown processing and integration events, the selected DNA becomes a new functional spacer, i.e. its matching protospacer will have the correct PAM to license Cas9 cleavage (Figure 2-9). The molecular steps that take place after protospacer selection to incorporate it as a new spacer in the CRISPR array are still unknown. All genes of the type II-A CRISPR-Cas locus (tracrRNA, cas9, cas1, cas2 and csn2) are required for spacer acquisition, therefore most likely all the members of the Cas9-Cas1-Cas2-Csn2 complex participate in the process. Future work will address this and other aspects of the mechanisms of spacer integration in different CRISPR-Cas systems.

The present chapter reveals a new function for Cas9 in CRISPR immunity. This nuclease is fundamental for both the execution of immunity, participating in the surveillance and destruction of infectious target viruses, and the generation of immunological memory, selecting the viral sequences that allow adaptation and resistance to viral predators.

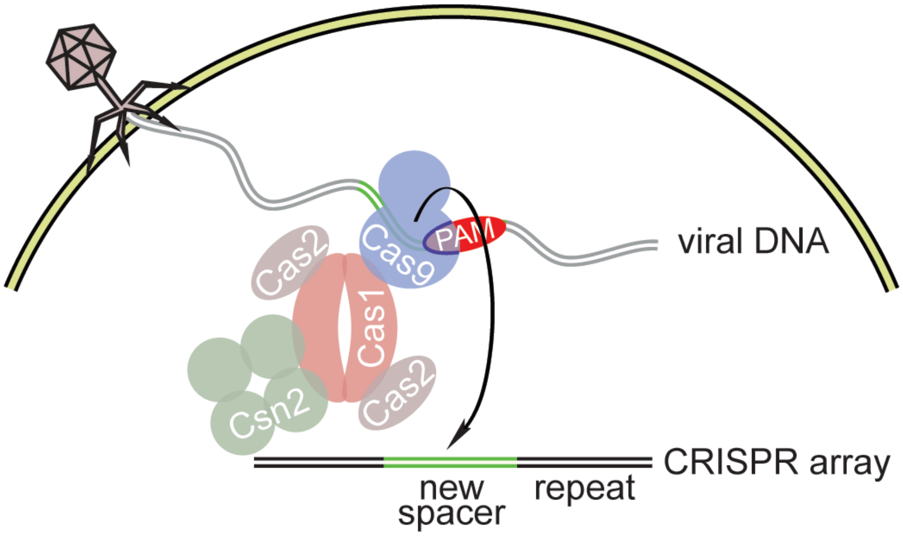


Figure 2-9. A model for the selection of PAM-flanking spacers by Cas9. After injection of the phage DNA, an adaptation complex formed by Cas9, Cas1, Cas2 and Csn2 uses the Cas9 PAM binding domain to specify functional protospacers, that is, that are followed by the correct PAM. It is not known how the protospacer sequence is extracted from the viral DNA to become a spacer. In the ‘cut and paste’ model, a nuclease, possibly Cas1, cuts the viral DNA to generate the spacer. In the ‘copy and paste’ model the protospacer sequence is copied first. Once loaded with the selected protospacer sequence, this complex promotes the integration of this sequence into the CRISPR array, thus becoming a new spacer. Previous studies demonstrated that Cas1 dimerizes and interacts with Cas2[70], Csn2 has been determined to forma tetramer[63].

Chapter III.

Generation of Cas9 Variants that Increase Spacer Acquisition

Having learned that Cas9 participates in spacer acquisition by specifying functional viral targets, I wanted to further explore this topic by engineering Cas9 mutants that provide enhanced CRISPR immunity. Here I performed random mutagenesis of the RNA-guided Cas9 nuclease to look for variants that provide enhanced immunity against viral infection. I identified a mutation, I473F, which increases the rate of spacer acquisition by more than two orders of magnitude. The results presented in this chapter highlight the role of Cas9 during CRISPR immunization and provide a useful tool to study this otherwise rare process and develop it as a biotechnological application.

3.1 Changing the PAM specificity of Cas9

Based on their *cas* genetic repertoire, CRISPR-Cas systems have been classified into six types, I through VI[24,83]. Cas9 is the crRNA-guided nuclease of the type II-A CRISPR-Cas system of *Streptococcus pyogenes*[34]. In addition to protospacer recognition by the crRNA, Cas9 target cleavage requires a 5'-NGG-

3’ protospacer adjacent motif (PAM) immediately downstream of the target[17,33,34,71]. Cas9 is also required for the immunization step of the CRISPR response[84,85], using its PAM binding domain to specify functional spacer sequences that are flanked by the required NGG motif[84]. In support of its role in spacer acquisition, Cas9 can associate in vivo with the other proteins encoded by the type II-A CRISPR-Cas system: Cas1, Cas2 and Csn2[84].

To further study the role of Cas9 in spacer acquisition, I decided to change its PAM specificity. Earlier work from our lab tested in vivo cleavage of targets having the same protospacer sequence but different PAMs displaying all possible trinucleotide combinations. I found that, in addition to the complete cleavage of targets with NGG PAMs, wild-type Cas9 displays approximately 50% of in vivo cleavage of targets with NAG PAMs. In an effort to understand how Cas9 affects the acquisition of spacers flanked by NGG motifs, I decided to evolve this weak but detectable affinity of the nuclease for NAG PAMs. After structural analysis determined the PAM interacting domain of Cas9[71,77], different groups have specifically mutated this domain to obtain a versatile set of nucleases for genome editing purposes and have obtained an NAG-recognizing Cas9[86]. I took a different approach and searched for mutations in any region of the nuclease that would increase its specificity for NAG-flanked targets. I found one such mutation, I473F, which provided partial immunity when directed (programmed) to recognize an NAG viral protospacer. Importantly, this mutation also expanded the levels of the CRISPR-Cas adaptive immune response, increasing the number of CRISPR-mediated, bacteriophage-resistant colonies by more than two orders of magnitude. I performed experiments to understand the molecular basis of the enhanced CRISPR-Cas immunity and determined that the I473F mutation mediates a significant increase in spacer acquisition. Our results highlight the role of Cas9 during CRISPR immunization and provide a useful tool to study this otherwise rare process.

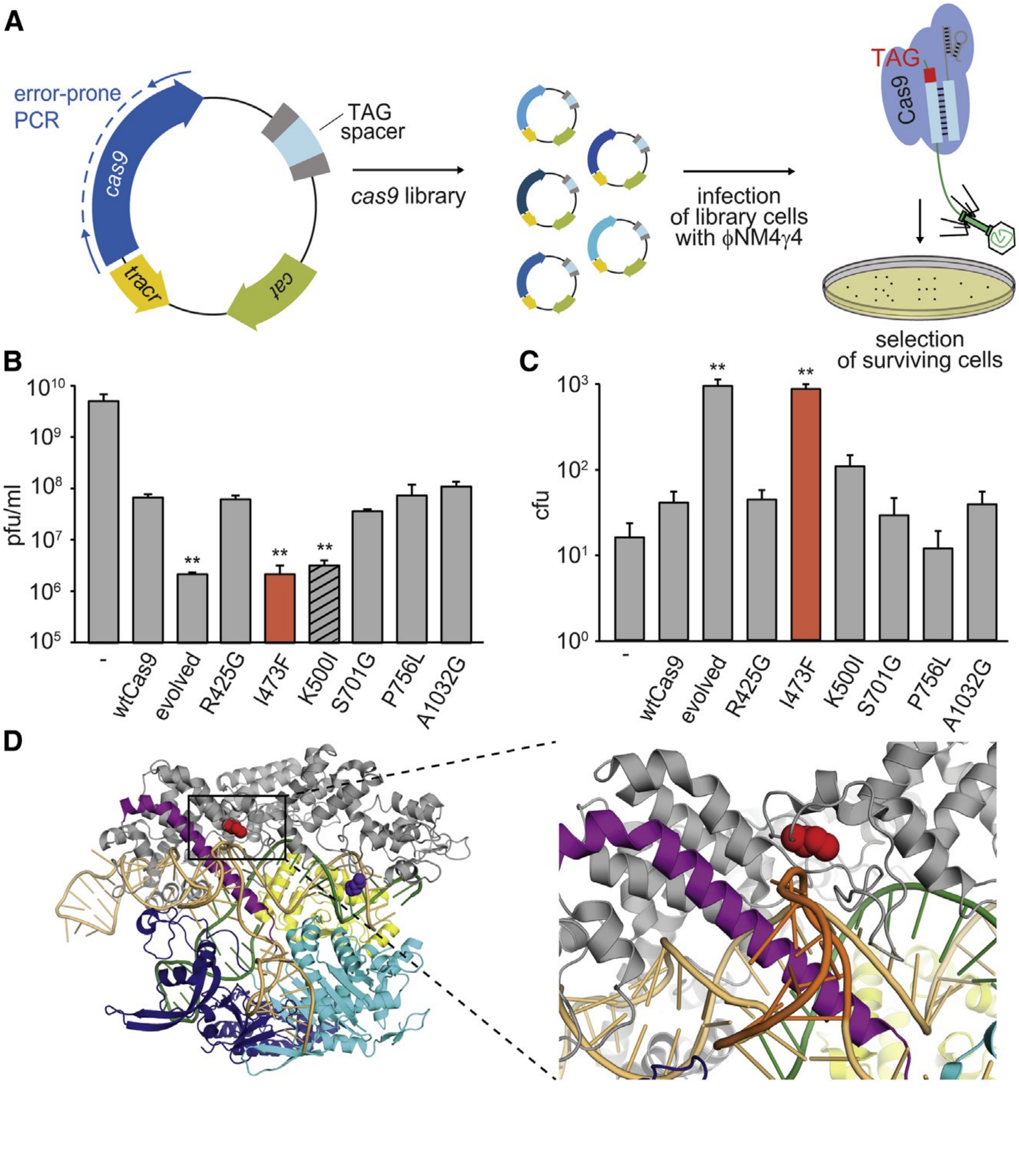
3.2. Evolved Cas9 has increased NAG PAM specificity

S. pyogenes Cas9 has an innate ability to cleave NAG-adjacent targets, but with much lower efficiency than it cleaves canonical (NGG) targets[33]. To enhance the ability of the nuclease to target protospacer sequences flanked by NAG PAMs, I constructed a library of plasmids carrying mutagenized *cas9* variants by subjecting the entire gene to error-prone PCR (Figure 3 -1a). The library plasmids also harbor the trans-activating crRNA (tracrRNA) gene[44] and a single-spacer CRISPR array targeting a protospacer sequence (AAAAACAAAAATGTTTTAACACCTATTAACG) followed by a TAG PAM on the genome of the lytic staphylococcal bacteriophage ϕNM4γ4[64]. The library was transformed into *Staphylococcus aureus* RN4220 cells that were subjected to phage infection on soft-agar plates to select for phage-resistant bacterial colonies. These colonies originated either from cells that acquired surface mutations preventing phage adsorption or cells harboring mutant *cas9* alleles with improved NAG cleaving efficiency that can sustain anti-viral immunity. To enrich for bacteria harboring desired Cas9 mutants, I isolated and re-transformed the plasmids of surviving colonies to perform a second round of phage infection.

Several colonies were obtained, and I proceeded with a further analysis of one of the “evolved” mutants that gained phage resistance. Sequencing of the plasmid revealed the presence of six single-nucleotide substitutions in the *cas9* gene, producing the following missense mutations: R425G, I473F, K500I, S701G, P756L and A1032G.

To evaluate the importance of each of these mutations in the gain-of-function phenotype I introduced them individually into the *cas9* gene and tested the ability of the resulting plasmid to prevent ϕNM4γ4 propagation by measuring the number of plaque forming units (pfu) that result after infection of the host cells (Figure 3-1b). Whereas cells harboring a control vector do not provide any immunity and allow high levels of phage propagation (up to ~10^{10} pfu/ml), cells containing wild-type Cas9 provide partial immunity and reduce phage propagation by about two orders of magnitude. Cas9 harboring the R425G, S701G, P756L and A1032G mutations allow wild-type levels of phage propagation and therefore do not contribute to the gain-of-function-phenotype of the evolved *cas9* allele I isolated. In contrast, cells containing Cas9 with the I473F or K500I mutations decrease phage propagation by about four orders of magnitude. This is close to the levels of immunity provided by wild-type Cas9 when programmed against NGG-flanked targets (a reduction of ~ 5 orders of magnitude, see Figure 3-3b). Similar results were obtained when other NAG PAMs were used in this assay (AAG, CAG, GAG, Fig. S1). Therefore, the I473F and K500I mutations enhance the ability of Cas9 to recognize targets with NAG flanking PAMs.

Figure 3-1. Directed Evolution of cas9 Generates Mutants with Increased Specificity for NAG Targets (a) Schematic diagram of the directed evolution assay. S. pyogenes cas9 was mutagenized by error-prone PCR and library amplicons were cloned into a plasmid carrying a spacer matching a TAG- adjacent target sequence on the fNM4g4 phage. Library cells were infected with lytic phage to screen for mutants displaying improved NAG cleaving efficiency. **(b)** Phage propagation was measured as the number of plaque-forming units (PFUs) per milliliter of stock on cells targeting the NAG-adjacent proto- spacer and harboring plasmids with different mutations on cas9: one of the "evolved" alleles or each of the six mutations present in this allele. Mutations with PFU values significantly different than wild- type are highlighted (**$p < 0.05$ compared to WTCas9). Data are represented as mean ± SD of three representative biological replicates. **(c)** Colony-forming units (CFUs) obtained after phage infection of naive cells (not programmed to target any viral sequence) harboring plasmids with different mutations in cas9. Mutations with CFU values significantly different than wild-type are highlighted. Data are represented as mean ± SD of three representative biological replicates. **(d)** Location of residues I473 and K500 on the Cas9:single-guide RNA ribonucleoprotein (PDB: 4UN3). Red, I473; purple, K500; orange, sgRNA; green, target DNA (the GG PAM highlighted in red); gray, a-helical (REC) lobe; yellow, HNH domain; light blue, RuvC domain; blue, PAM-interacting CTD.



A
error-prone PCR
cas9
tracr
cat
TAG spacer
cas9 library
infection of library cells with ϕNM4γ4
TAG
Cas9
selection of surviving cells
B
pfu/ml
10^{10}
10^{9}
10^{8}
10^{7}
10^{6}
10^{5}
**
**
**
-
wtCas9
evolved
R425G
I473F
K500I
S701G
P756L
A1032G
C
cfu
10^{3}
10^{2}
10^{1}
10^{0}
**
**
-
wtCas9
evolved
R425G
I473F
K500I
S701G
P756L
A1032G
D

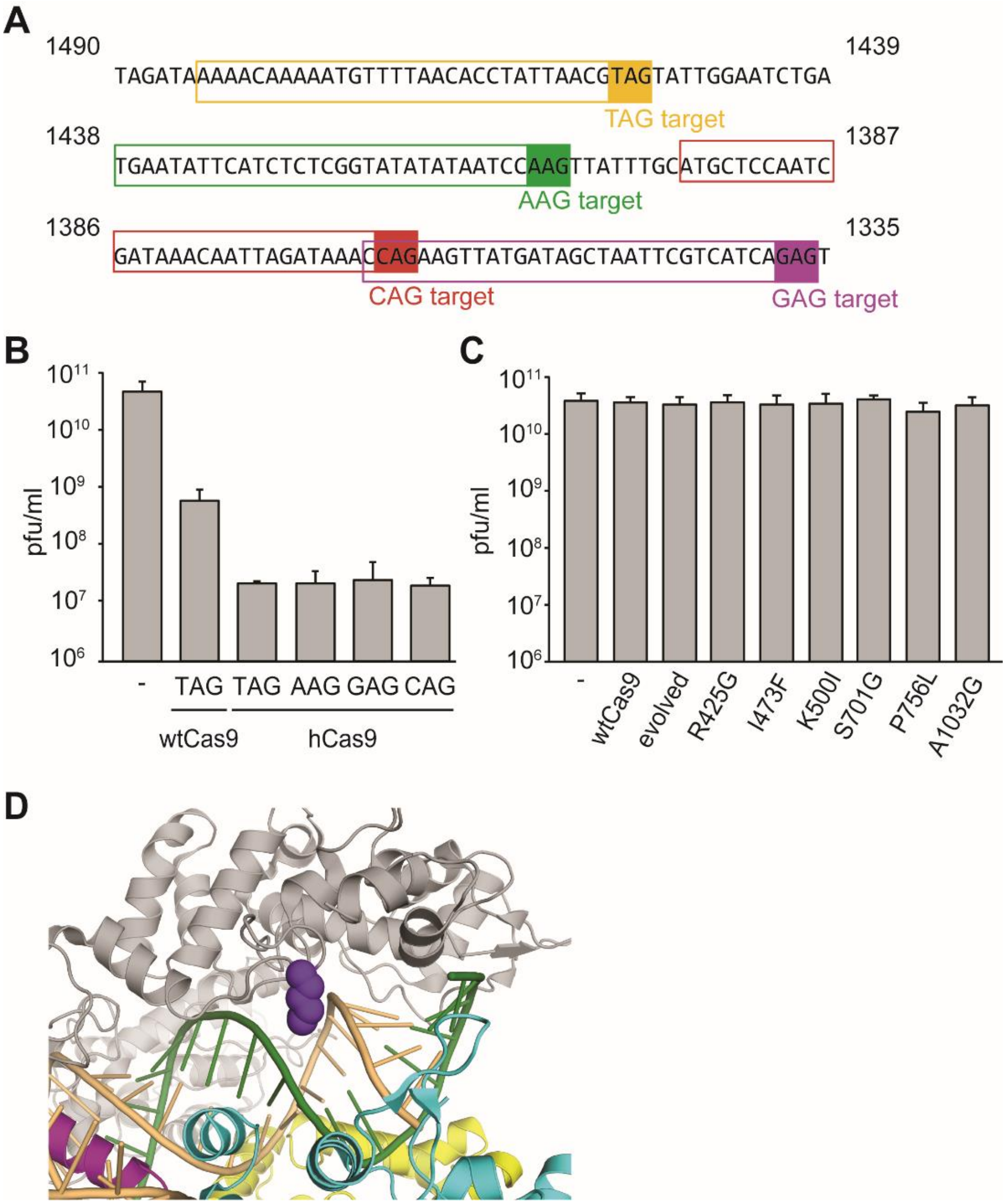


Figure 3-2. Protection of host cells by hCas9 programmed against different NAG-flanked targets. (**a**) The ability of hCas9 to target protospacers with different PAM was tested by measuring phage propagation in cells harboring CRISPR-Cas systems containing either wtCas9 or hCas9 and programmed to target the sequences shown, which are followed by TAG, AAG,

GAG or CAG PAMs. (**b**) Phage propagation was measured as the number of plaque forming units (pfu) per ml of stock, on cells targeting the TAG, AAG, GAG, and CAG-adjacent protospacers and hCas9. Data are represented as mean ± SD of three representative biological replicates. (**c**) Measurement of pfu formation on staphylococci carrying plasmids with different *cas9* mutations after infection with ϕ85, a phage lacking the target recognized in ϕNM4γ4. Data are represented as mean ± SD of three representative biological replicates. (**d**) Location of residue K500 on the Cas9:single-guide RNA ribonucleoprotein (PDB 4UN3). Purple, K500; orange, sgRNA; green, target DNA (the GG PAM highlighted in red); grey, alpha-helical (REC) lobe; yellow, HNH domain; light blue, RuvC domain; blue, PAM-interacting CTD.

Given the requirement of Cas9 for the immunization phase of the CRISPR-Cas immune response, i.e. the generation of phage-resistant bacteria through the acquisition of viral sequences as spacers[84,85], I wondered whether the evolved Cas9 as well as the individual mutants affected this process. To test this, I introduced the different alleles of *cas9* into a plasmid also harboring the *tracrRNA* coding sequence, the *S. pyogenes* SF370 CRISPR array (containing six spacers, none of them matching the genome of ϕNM4γ4) and the type II-A genes involved exclusively in the acquisition of new spacers, *cas1*, *cas2* and *csn2*[84,85]. *S. aureus* cells containing the different plasmids were infected with ϕNM4γ4 and the number of survivors were enumerated as colony forming units (cfu) (Figure 3-1c).

Cells harboring a vector control provide the threshold for the number of non-CRISPR phage resistant mutants. Only a small fraction of cells containing wild-type Cas9 are able to acquire new spacers, about 2-fold over the threshold control. In contrast, the evolved *cas9* allele containing all six mutations increased the number of CRISPR-surviving cells by about 60-fold. Analysis of single mutants revealed that this highly significant increase was provided almost exclusively by the I473F mutation (Figure 3-1c). Due to the sharp enhancement of the CRISPR-Cas immune response conferred by the I473F mutation I decided to name the Cas9^{I437F} mutant “hyper-Cas9”, or hCas9. I473 is located close to the surface of Cas9, outside of the PAM-interacting domain, and it is part of a projection from the Helical III domain that interacts with the nexus of the guide RNA[87] (Figure 3-1d). This position does not suggest an evident effect of the I473F mutation on Cas9 activity and therefore I decided to investigate the basis for its phenotype by performing a detailed comparison with the CRISPR-Cas immune response mediated by wild-type Cas9.

3.3. Mutant Cas9 enhances CRISPR adaptive immunity by 100-fold

To perform a more accurate comparison between wild-type (wtCas9) and hCas9, I counted the number of CRISPR-mediated, phage resistant cells that arise after phage infection. Figure 3-3a shows representative plates of infected cells containing plasmids with the wtCas9 or hCas9 *S. pyogenes* CRISPR-Cas locus, showing a striking difference in the number of surviving colonies. As mentioned before, most of these colonies arise from single cells that were able to acquire a new spacer matching the ϕNM4γ4 genome. However, a fraction of the surviving cells repels phage attack by non-CRISPR related mechanisms, such as envelope resistance[84]. To make a more accurate quantification of the CRISPR-Cas response, I analyzed individual colonies by PCR of the CRISPR array[28,84] to detect those in which new spacers were acquired, i.e. “adapted” cells (Figure 3-3b). Not only did many more resistant colonies originated from cells harboring hCas9 (an average of 31 cfu for wtCas9 vs 4,312 cfu for hCas9, Figure 3-3c), but also most of them showed CRISPR-mediated phage resistance (23% for wtCas9 vs 90% for hCas9, Figure 3-3c).

We wondered whether this was a consequence of the specific substitution of I473 by phenylalanine. To test this, I introduced an I473A mutation into Cas9 and compared this mutant with wtCas9 and hCas9 in this assay (Figure 3-4). I found that cells harboring the I473A mutant produced a number of CRISPR-mediated immune cfu comparable to cells carrying wtCas9, but 10 times lower than the cfu obtained from infection of cells expressing hCas9. Therefore, I conclude that the I473F mutation increases the CRISPR-adaptive immune response through a specific effect of the phenylalanine residue in position 473 and by more than two orders of magnitude: on average, approximately 7 cfu (31×0.23) per experiment for infected wtCas9-containing cells, and approximately 3,863 cfu (4,312×0.90) for infected hCas9-expressing bacteria. I sequenced PCR

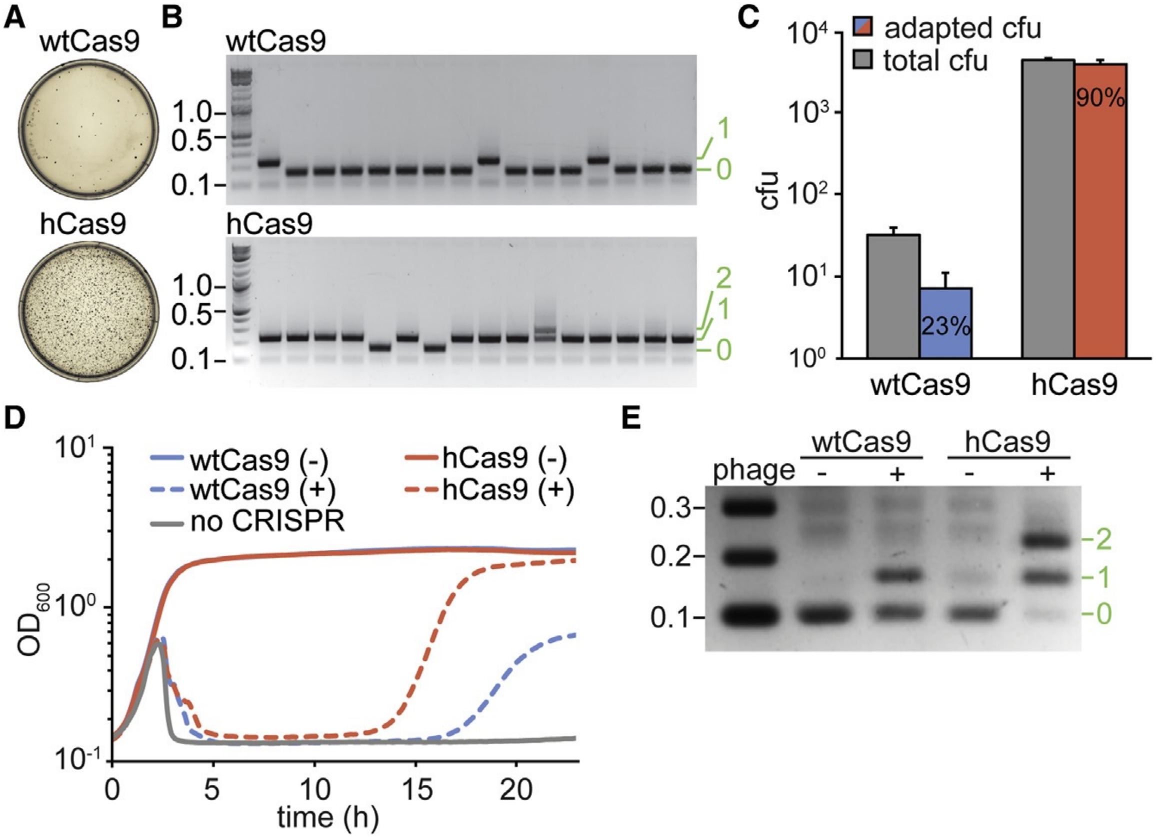


Figure 3-3. Cas9^{I473F} or hyper-Cas9 Mounts an Enhanced CRISPR Adaptive Immune Response. (a) Representative plates obtained after lytic infection of cells harboring the full CRISPR system of *S. pyogenes* with WTCas9 or hyperCas9 (hCas9) showing the number of surviving colonies. **(b)** Agarose gel electrophoresis of PCR products of the amplification of the CRISPR of arrays of surviving cells to detect newly acquired spacers (asterisks). Molecular markers (in kilobases) are indicated in black and the number of new spacers added in green. **(c)** Quantification of total surviving colonies (gray bars) and surviving colonies with newly incorporated spacers, as detected by PCR (blue and red bars). Data are represented as mean ± SD of three representative biological replicates. **(d)** Growth curves of cultures of cells harboring the full CRISPR system of *S. pyogenes* with WTCas9 or hCas9 with (+) or without (-) phage infection. **(e)** PCR-based analysis of the liquid cultures shown in (c) (at 24 hr post-infection) to check for the acquisition of new spacer sequences in the presence (+) or the absence (-) of phage infection by cells expressing WTCas9 or hCas9. Molecular markers (in kilobases) are indicated in black and the number of new spacers added in green. Image is representative of three technical replicates.

products to determine the PAM of the spacers acquired by 40 colonies expressing wtCas9 or hCas9. Interestingly, all 40 spacers acquired by cells expressing hCas9 matched targets with an NGG PAM, suggesting that this nuclease can still target sequences followed by the canonical PAM in addition to targets with NAG PAMs.

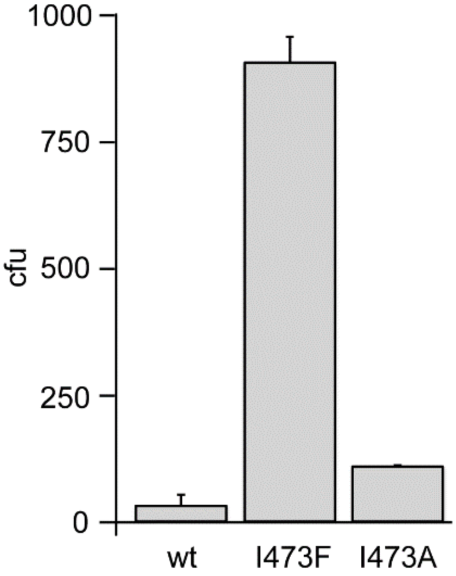


Figure 3-4. CRISPR-Cas immune response of cells expressing Cas9^{I473A}. Cultures harboring plasmids with *tracrRNA*, *cas1*, *cas2* and *csn2* genes, and either wild-type, I473F or I473A cas9 alleles, were infected with ΦNM4γ4 phage on top agar media and poured on plates. After 24 hours of incubation at 37 °C the CRISPR-surviving colonies were counted. Data are represented as mean ± SD of three representative biological replicates.

Similar results were observed when cells in culture carrying naïve wtCas9 or hCas9 CRISPR-Cas systems were infected with phage. Upon addition of ϕNM4γ4, the cultures lyse, as the vast majority of cells do not undergo spacer acquisition (Figure 3-3). Nonetheless, hCas9 cultures were able to regrow much earlier (~14 hours post-infection) than wtCas9 cultures (~17 hours post-infection). PCR analysis of the population of surviving cells (i.e., using DNA extracted from the whole culture, not individual resistant bacteria) at 24 hours post-infection corroborated the earlier observation that hCas9 cells mount a more robust CRISPR immune response (Figure 3-3e). Whereas the PCR products using DNA from immune cells carrying wtCas9 showed the presence of both adapted and non-adapted CRISPR arrays in the surviving population, the PCR results from cultures carrying hCas9 showed very little non-adapted CRISPR arrays, with the great majority of the cells acquiring one or two new spacers. Altogether these data show that the I473F mutation in Cas9 allows for a more robust CRISPR-Cas immune response due to a specific effect of the phenylalanine residue.

3.3.1. Hyper Cas9 provides wild-type cleavage efficiency

The CRISPR-Cas response can be divided into two distinct stages[7]. First there is spacer acquisition, where sequences from the invading virus are incorporated into the CRISPR array. This is followed by the second stage where the acquired spacers provide the crRNA guides to the Cas nucleases for the destruction of the viral DNA. Therefore, the enhanced immunity phenotype of hCas9 documented in Figure 3-3 could be in principle due to an increase in the frequency of spacer acquisition, a more robust cleavage by hCas9 of its targets, or both.

First, I considered the possibility that hCas9 could provide better cleavage of the infecting viral DNA. In this scenario both wtCas9 and hCas9 populations can acquire a similar number of new spacers but a more robust cleavage of the target DNA by hCas9 would lead to a faster recovery of the bacteria that acquired the spacers. This will result in the clonal expansion and the consequent increase in the number of surviving bacteria. To test this hypothesis, I infected cells carrying plasmids with either wtCas9 or hCas9 programmed to target the ϕNM4γ4 virus and the *tracrRNA* gene, but without the spacer acquisition machinery (*cas1*, *cas2* and *csn2*). This genetic background supports CRISPR-Cas anti-viral defense but does not allow the acquisition of new spacer sequences[84]. Because our data suggested that hCas9 can still target protospacers followed by NGG PAMs, I tested the immunity of cells programmed to attack targets with either an NAG or an NGG PAM located in the same region of the ϕNM4γ4 genome (Figure 3-5). As a control, cells harboring a vector control were also infected.

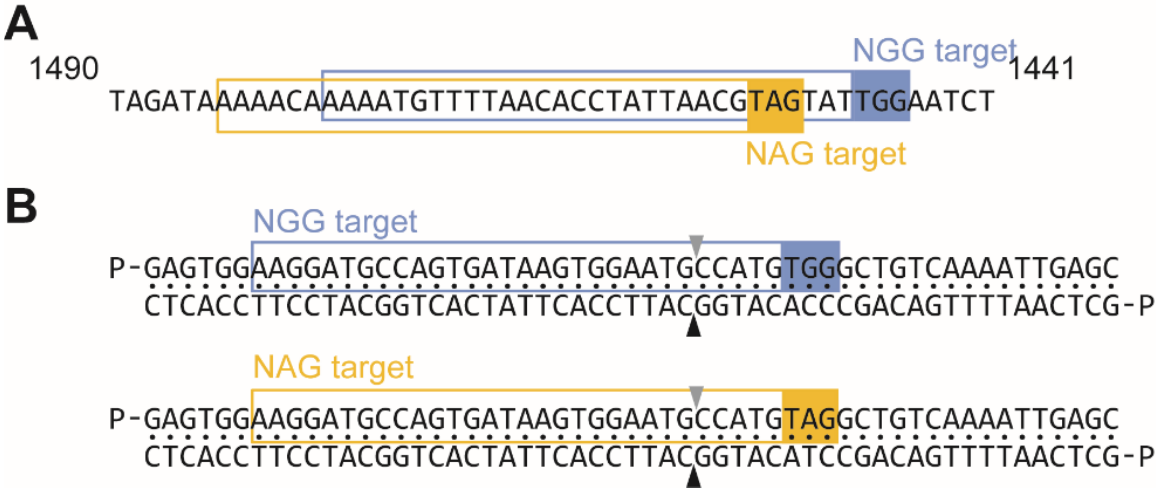

A
1490
NGG target
1441
TAGATAAAAACAAAAATGTTTTAACACCTATTAACGTAGTATTGGAATCT
NAG target
B
NGG target
P-GAGTGGAAGGATGCCAGTGATAAGTGGAATGCCATGTGGGCTGTCAAAATTGAGC
CTCACCTTCCTACGGTCACTATTCACCTTACGGTACACCCGACAGTTTTAACTCG-P
NAG target
P-GAGTGGAAGGATGCCAGTGATAAGTGGAATGCCATGTAGGCTGTCAAAATTGAGC
CTCACCTTCCTACGGTCACTATTCACCTTACGGTACATCCGACAGTTTTAACTCG-P

Figure 3-5. In vivo and in vitro targets. (**a**) Region of the ΦNM4γ4 phage genome (nucleotides 1441 to 1490) containing the TAG- and TGG-flanked protospacers, yellow and blue respectively, used in Figures 3-6a and 3-6b.(**b**) Sequences of the dsDNA target oligonucleotides used in Figure 3-6c. The protospacer sequence is the same, but it is flanked by either a TAG (yellow) or TGG (blue) PAM sequence. Radiolabel is at the 5' end (P). Grey and black arrowheads mark the cleavage sites of the RuvC and HNH domains, respectively.

Bacteria containing different plasmids were infected with phage during exponential growth and the optical density of the culture was followed over time to measure the immunity provided by Cas9 cleavage of the viral genome (Figure

3-6a). As expected, control cells were rapidly lysed by the addition of phage. In contrast, cells expressing wtCas9 programmed against an NGG target cleared the infection efficiently and continued the exponential growth. In contrast, the poor targeting of NAG-flanked protospacers by wtCas9 led to substantial lysis, although not as dramatic as the non-Cas9 control, suggesting a low level of target cleavage. The population of hCas9-containing cells targeting an NGG-flanked viral protospacer was protected to levels indistinguishable from the immunity provided by bacteria expressing wtCas9 programmed against the same target. Targeting of the protospacer followed by an NAG PAM was more efficient in cells having hCas9 than in the wtCas9 population. However, hCas9 did not provide full immunity as was the case for NGG-containing targets. Similar results were obtained when phage propagation was measured instead of cell survival (Figure 3-6b). To do this, I compared the number of plaques (measured as plaque forming units or pfu) obtained when a ϕNM4γ4 lysate was applied to plates seeded with the five different cultures used in Figure 3-6a. Cells harboring an empty vector allowed extensive phage propagation, up to more than 10^{10} pfu per ml of phage stock. In contrast, bacteria harboring wtCas9 or hCas9 programmed to target the NGG-flanked protospacer in the ϕNM4γ4 genome reduced phage proliferation by more than four orders of magnitude (~ 10^{6} pfu/ml). When the target contained an NAG PAM, wtCas9 only reduced viral multiplication by an order of magnitude compared to the non-CRISPR control, whereas hCas9 reduced it by about three orders of magnitude, but nevertheless failing to fully restrict the virus.

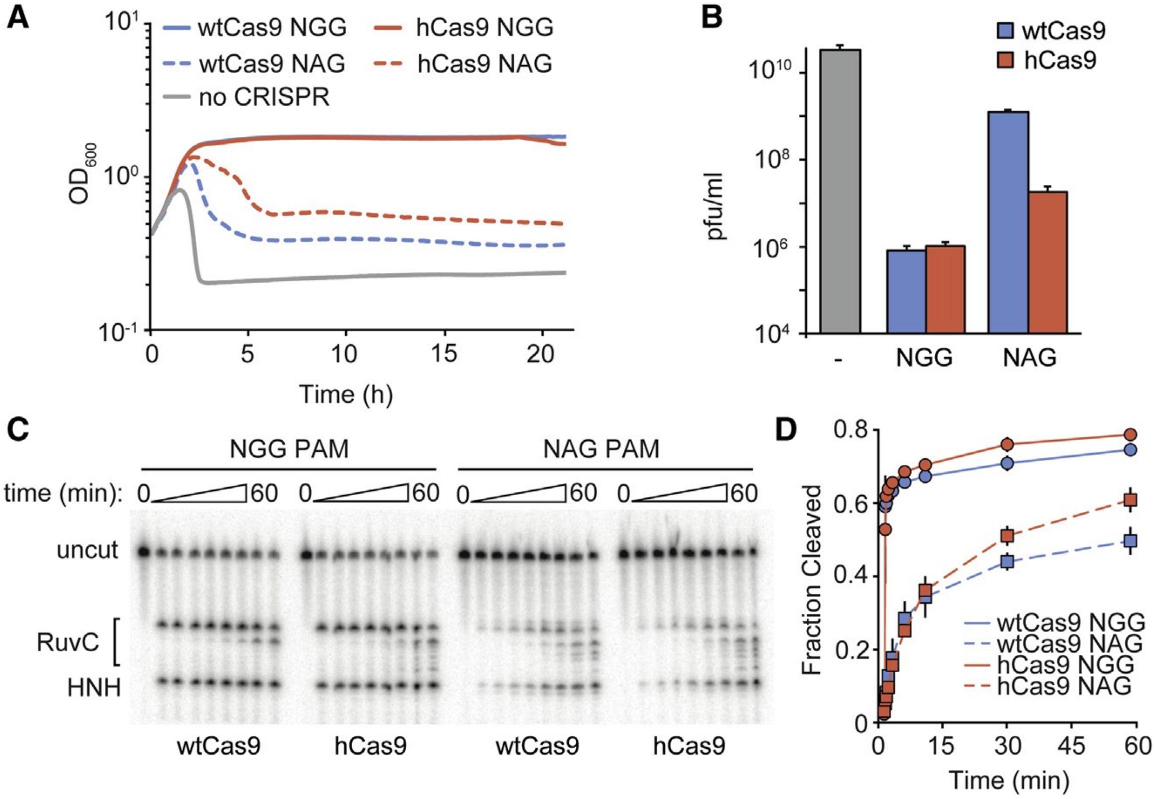


Figure 3-6. hCas9 has increased interference efficiency against NAG-adjacent, but Not NGG-adjacent, targets. (a) Growth curves of cultures infected with fNM4g4 harboring the WTCas9 or hCas9 (but not Cas1, Cas2, and Csn2) programmed to target either NAG- or NGG-flanked viral sequences. **(b)** Phage propagation, measured in PFU/mL, of the bacteria presented in (a). Data are represented as mean ± SD of three representative biological replicates. **(c)** Cleavage of radiolabeled dsDNA targets, flanked by either NGG or NAG PAMs, by WTCas9 or hCas9. **(d)** Quantification of the cleavage results shown in (c). Data are represented as mean ± SD of three representative biological replicates.

Both in vivo experiments measuring bacterial survival (Figure 3-6a) and phage propagation (Figure 3-6b) suggest that hCas9 has not improved efficiency of cleavage of NGG-flanked targets and displays only a small increase in the cleavage of NAG-flanked sequences. To unequivocally demonstrate this, I performed in vitro cleavage assays with purified wtCas9 and hCas9 (Figure 3-6c). In this case, I was able to compare cleavage of radiolabeled oligonucleotides containing the same protospacer sequence followed by either a TGG or TAG PAM (Figure 3-5b). Consistent with in vivo data, experiments showed similar cutting rates of the NGG target for wtCas9 and hCas9. Quantification of the cleavage products showed that hCas9 cleaved more of the NAG target than wtCas9 over longer timescales (Figure 3-6d). Altogether, the data presented in Figure 3 indicate that while there is a modest increase in the NAG-targeting properties of hCas9, this cannot explain the rise in the number of CRISPR-resistant colonies mediated by the I473F mutation (Figure 3-3c).

3.3.2. Hyper Cas9 promotes higher rates of spacer acquisition

A second hypothesis that could explain the increase in CRISPR-Cas immunity conferred by hCas9 is, as explained above, a possible increase in the frequency of spacer acquisition by the cells expressing this mutant. To test this, I performed a comparison of the spacer repertoires acquired by cells harboring wtCas9 or hCas9. I made two plasmid libraries, carrying the spacer acquisition genes *cas1*, *cas2* and *csn2* and wt*cas9* or h*cas9*, the *tracrRNA* gene and the *S.*

pyogenes array of repeats and spacers preceded by a "barcode" sequence of 10 nucleotides 50 bp immediately upstream of the CRISPR array (Figure 3-7a). Cells harboring each library were infected with phage ϕNM4γ4 and DNA from the adapted cells was used to amplify the CRISPR array via PCR and collect sequence information of all the new acquired spacers using next generation sequencing. The primers used also amplify the barcode sequence (Figure 3-7a) and therefore each new spacer sequence can be associated with a unique barcode, allowing us to count how many times a given spacer was independently acquired in each bacterial population. Over three million reads belonging to either library were analyzed.

The frequency of reads corresponding to each acquired spacer sequence was plotted according to its position in the ϕNM4γ4 genome (Figure 3-8a). Analysis of the PAMs of the acquired spacers showed that over 99.5% of the spacer reads contained the NGG sequence in both libraries (Figure 3-8b), corroborating our in vivo data showing that hCas9 retained NGG PAM specificity. In addition, I looked at the repertoire of unique different spacers independently of the number of reads per sequence (Figure 3-8c). Consistent with our previous finding that the PAM specificity of Cas9 is responsible for the PAM sequence of the new protospacers, the hCas9 library showed a 5-fold increase in the acquisition of spacers matching NAG-flanked targets. I also observed an increase in the total number of different spacer sequences, from 1980 for wtCas9 cells to 2500 for the hCas9 sample.

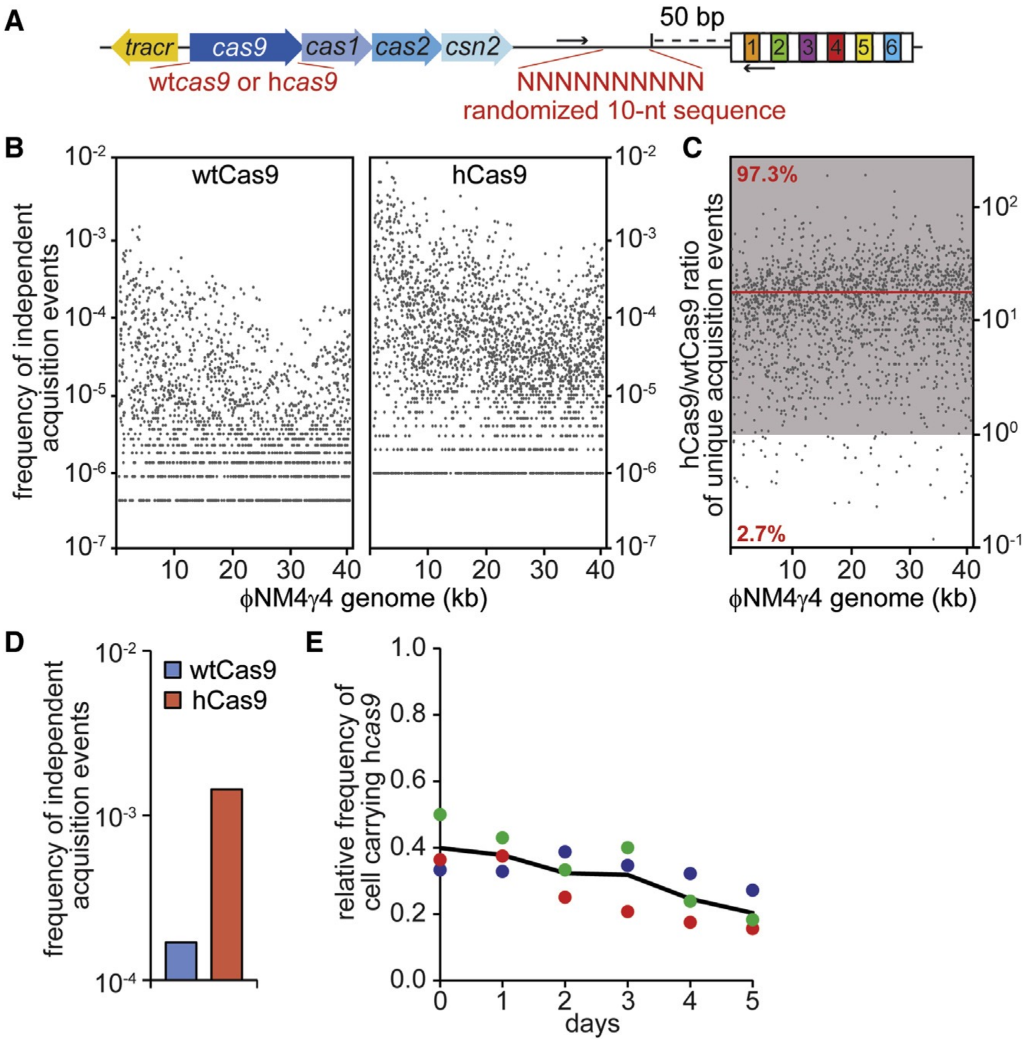
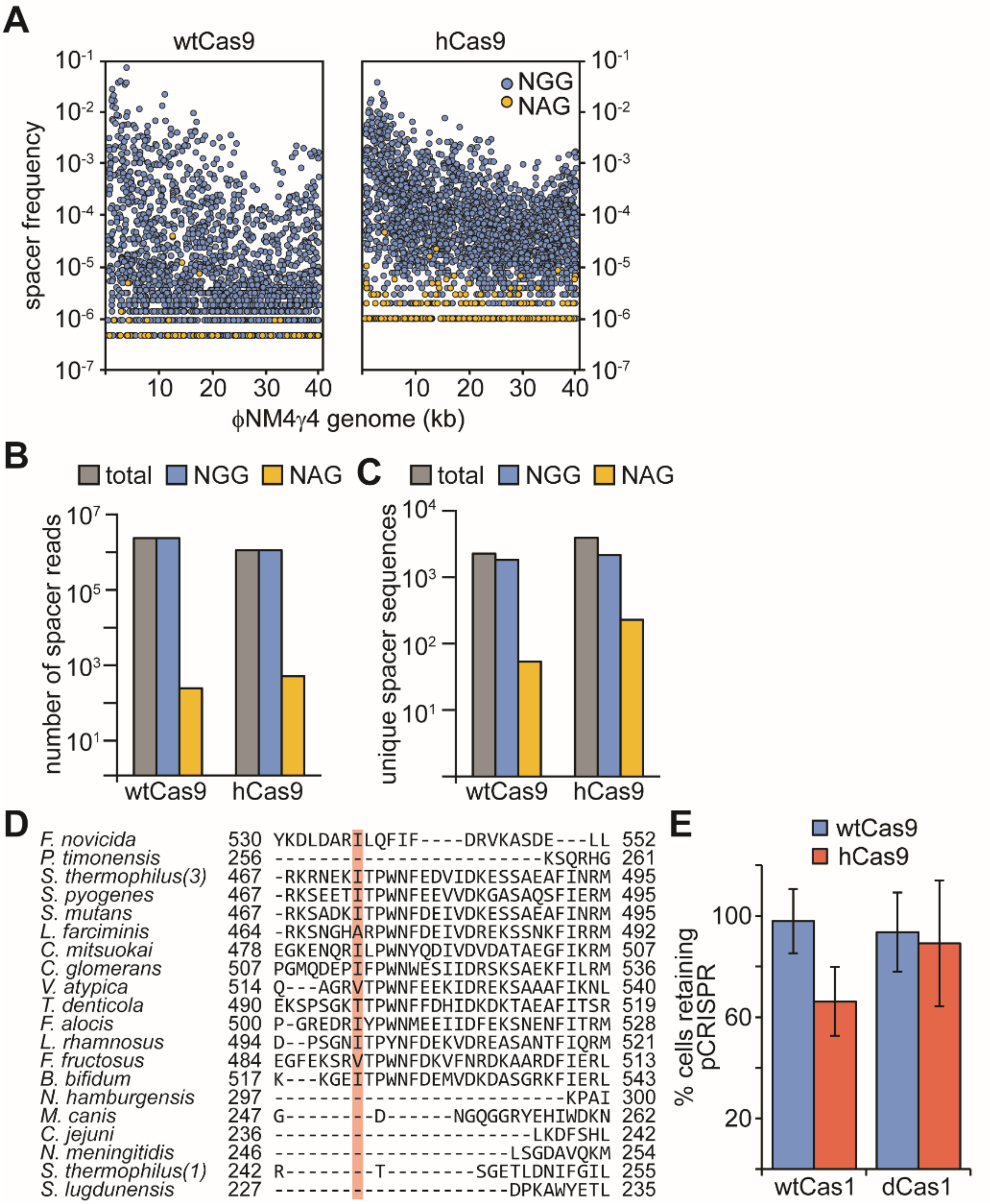


Figure 3-7. hCas9 Promotes Higher Rates of Spacer Acquisition. (a) Schematic diagram of the *S. pyogenes* CRISPR locus showing the barcode and primers (arrows) used to measure the number of independent spacer acquisition events. **(b)** Cultures expressing WTCas9 or hCas9 were infected with fNM4g4 phage, and surviving cells were collected after 24 hr, had DNA extracted, and were used as template for PCR of the CRISPR arrays. Amplification products were separated by agarose gel electrophoresis (not shown), and the DNA of the expanded CRISPR array was subject to MiSeq next-generation sequencing. The number of barcodes for each spacer sequence across the phage genome, normalized by the total number of spacer reads obtained, was plotted. **(c)** The hCas9/WTCas9 frequency of independent acquisition events ratio for 1,938 common spacer sequences was plotted across the phage genome. The zone where the ratio is greater than one is shown in gray. The red line shows the average ratio. **(d)** Same as (b) but without phage infection; i.e., a measure of acquisition of spacers derived from the host chromosome and resident plasmids. **(e)** Pairwise competition between staphylococci expressing WTCas9 or hCas9. The change in the relative frequency of cells carrying the h*cas9* allele (y axis) is plotted against the number of culture transfers (1 transfer/day, x axis).

To calculate the frequency of acquisition of every spacer I divided the number of different barcodes for a given spacer sequence by the total number of reads. This value was plotted according to its position in the ϕNM4γ4 genome (Figure 3-7b). The data show an increase in the frequency of acquisition in hCas9 cells, with a 6-fold increase in the average frequency. For all spacer sequences shared between the two libraries (1938 sequences were shared between the 1980 and 2500 different sequences for wtCas9 and hCas9, respectively), I calculated the ratio of unique adaptation events (i.e. number of different barcodes) for hCas9 reads compared to wtCas9 (Figure 3-7c). I found that more than 97% of the spacers were acquired more frequently in the hCas9 library (ratio > 1), with an average ratio of ~18. All together, these findings show that hCas9 provides the host bacterium with more efficient spacer acquisition and suggest that this is a major contribution to the enhanced CRISPR-Cas immunity granted by hCas9.

Figure 3-8. Analysis of next-generation sequencing results. (a) Data presented in Figure 3-7b was plotted as the number of reads for each spacer sequence across the phage genome, normalized by the total number of spacer reads obtained. Spacers matching protospacers with NGG PAMs are shown in blue, with NAG PAMs in yellow. (**b**) Quantification of the data shown in panel a. (**c**) Quantification of the data shown in Figure 3-7b. (**d**) Alignment of Cas9 protein sequences belonging to type II CRISPR-Cas systems. Highlighted in orange is the I473 residue. An equivalent residue is not found in some type II-B and II-C systems. (**e**) Fraction (%) of staphylococci retaining the plasmid harboring wt*cas9* and h*cas9* after 10 days of culture; with one transfer (1:100 dilution into fresh media) per day. Cells were plated in solid media with and without chloramphenicol, an antibiotic that selects for cells harboring the pCRISPR plasmid. The fraction of staphylococci carrying this plasmid was obtained dividing the chloramphenicol-resistant cfu by the total cfu count. Data are represented as mean ± SD of three representative biological replicates.


A
wtCas9
hCas9
spacer frequency
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10
20
30
40
● NGG
● NAG
ϕNM4γ4 genome (kb)
B
▨ total ▨ NGG ▨ NAG
number of spacer reads
10^7
10^5
10^3
10^1
wtCas9
hCas9
C
▨ total ▨ NGG ▨ NAG
unique spacer sequences
10^4
10^3
10^2
10^1
wtCas9
hCas9
D
F. novicida 530 YKDLDARILQFIF----DRVKASDE---LL 552
P. timonensis 256 ------------------------KSQRHG 261
S. thermophilus(3) 467 -RKRNEKITPWNFEDVIDKESSAEAFINRM 495
S. pyogenes 467 -RKSEETITPWNFEEVVDKGASAQSFIERM 495
S. mutans 467 -RKSADKITPWNFDEIVDKESSAEAFINRM 495
L. farciminis 464 -RKSNGHARPWNFDEIVDREKSSNKFIRRM 492
C. mitsuokai 478 EGKENQRILPWNYQDIVDVDATAEGFIKRM 507
C. glomerans 507 PGMQDEPIFPWNWESIIDRSKSAEKFILRM 536
V. atypica 514 Q---AGRVTPWNFEEKIDREKSAAAFIKNL 540
T. denticola 490 EKSPSGKTTPWNFFDHIDKDKTAEAFITSR 519
F. alocis 500 P-GREDRIYPWNMEEIIDFEKSNENFITRM 528
L. rhamnosus 494 D--PSGNITPYNFDEKVDREASANTFIQRM 521
F. fructosus 484 EGFEKSRVTPWNFDKVFNRDKAARDFIERL 513
B. bifidum 517 K---KGEITPWNFDEMVDKDASGRKFIERL 543
N. hamburgensis 297 --------------------------KPAI 300
M. canis 247 G--------D------NGQGGRYEHIWDKN 262
C. jejuni 236 -----------------------LKDFSHL 242
N. meningitidis 246 ---------------------LSGDAVQKM 254
S. thermophilus(1) 242 R--------T--------SGETLDNIFGIL 255
S. lugdunensis 227 ---------------------DPKAWYETL 235
E
▨ wtCas9
▨ hCas9
% cells retaining pCRISPR
100
60
20
wtCas1
dCas1

3.4. Discussion

Here I performed random mutagenesis on the *cas9* gene to identify mutants with an expanded CRISPR-Cas response. Specifically, I looked for mutants that would allow Cas9 to recognize not only NGG- but also NAG-containing targets. I isolated a mutant, harboring an I473F substitution, that displayed a modest increase in NAG-target recognition. More importantly, the mutation increased the CRISPR-Cas immune response of the bacterial host by more than two orders of magnitude, as measured by the number of CRISPR-mediated bacteriophage resistant colonies obtained after phage infection. Due to this hyper-activity in CRISPR immunity I named the mutant version of Cas9 hyper-Cas9, or hCas9. Deeper analysis of hCas9 revealed that it can perform crRNA-guided cleavage of targets containing an NAG PAM better than wtCas9. However, this improvement is minor and does not seem to contribute significantly to the rise of a high number of CRISPR-mediated resistant cells. On the other hand, upon phage infection bacteria expressing hCas9 are able to acquire many more viral spacers than those expressing wtCas9. I hypothesize that this high rate of spacer incorporation is the basis for the observed increase in the CRISPR-mediated phage resistant colonies.

At the molecular level, the mechanism by which the I473F mutation enables this increase in spacer acquisition is not clear. I previously reported the existence of a complex between the four Cas proteins encoded by the type II-A CRISPR locus, namely Cas9, Cas1, Cas2 and Csn2[84]. I hypothesized that these complex functions in spacer acquisition, with Cas9 selecting sequences flanked by NGG PAMs[84] and Cas1 and Cas2[70,79] being involved in the integration of these sequences into the CRISPR array. The precise role of Csn2 in spacer acquisition remains to be elucidated. I thought that the I473F mutation could affect the formation of the complex, since the mutated residue is located on Cas9 surface and could participate in its interaction with another Cas protein. The substitution could enhance protein-protein interactions and either increase the abundance or the stability of the complex, thus increasing the rate of spacer acquisition. To test this, I incubated the four proteins along with a single-guide RNA[34] and subjected them to gel filtration to detect the formation of the complex. However, I did not observe significant amounts of stable complexes neither in the presence of wtCas9 nor hCas9. In wtCas9, the isoleucine residue is in direct contact with bases of the tracrRNA (Figure 3-1d) that are equivalent to the nexus in the single-guide RNA[88]. Specifically, nucleotide U59 of the tracrRNA inserts into a hydrophobic pocket lined by I473 and its adjacent residues[87]. It is possible that the bulkier phenylalanine residue could interfere with the tracrRNA:Cas9 association, affecting the involvement of Cas9 in the immunization step of the CRISPR-Cas response. This hypothesis is supported by the wild-type phenotype of the I473A mutation (Figure 3-4), since the smaller alanine residue most likely will not interfere with the tracrRNA interaction. Another mutation in a residue close to I473, K500I, also seems to affect Cas9 target specificity, but not the rate of spacer acquisition. Future work will explore the importance of this region in Cas9 activity during the different phases of CRISPR-Cas immunity.

In a recent study[89], the *E. coli* type I-E CRISPR-Cas adaptation machinery has been repurposed as a recording device to store information (such as environmental signals) in the form of spacers in the CRISPR array. Because the adaptation frequency is relatively low, decoding requires deep sequencing of a population of cells. This limits the number of stimuli that can be recorded. Using hyperactive adaptation machinery such as hCas9 can boost the adaptation frequency and thus the recording capacity of such synthetic devices. Moreover, combined with introduction of sheared genomic DNA, the hyperactive CRISPR adaptation machinery can be used to generate diverse and unbiased gRNA libraries *in vivo*. I speculate that hCas9 is able to sample much larger genomes then the type I-E Cas1-Cas2 complex

Chapter IV.

Spacer Acquisition and Immunological Diversity

Previous studies have shown that the abundance of each spacer in the surviving population is highly uneven. However, the molecular mechanisms underlying this bias are poorly understood. Here, I studied the factors that affect the distribution of individual spacer sequences after phage infection of cells harboring the type II-A CRISPR system from *Streptococcus pyogenes*. I show that spacer patterns are established early during infection and correlate with spacer acquisition rates, but not with spacer targeting efficiency. I also show that the rate of spacer acquisition depends on unique sequence elements within the spacer, which in turn determines the abundance of different spacers within the adapted population. Our results elucidate a fundamental mechanism behind the generation of immunological diversity during the type II CRISPR-Cas response.

4.1. An uneven spacer distribution across the phage genome

In the type II-A CRISPR system from *Streptococcus pyogenes*, cleavage is performed by the crRNA-guided nuclease Cas9[34], whose catalytic activity depends on the recognition of a 5'-NGG-3' protospacer adjacent motif (PAM)[33,34] Cas9 contains a PAM-interacting domain to recognize this motif[33,71] that is not only required for target cleavage but also for the acquisition of spacers with the appropriate PAM[84].

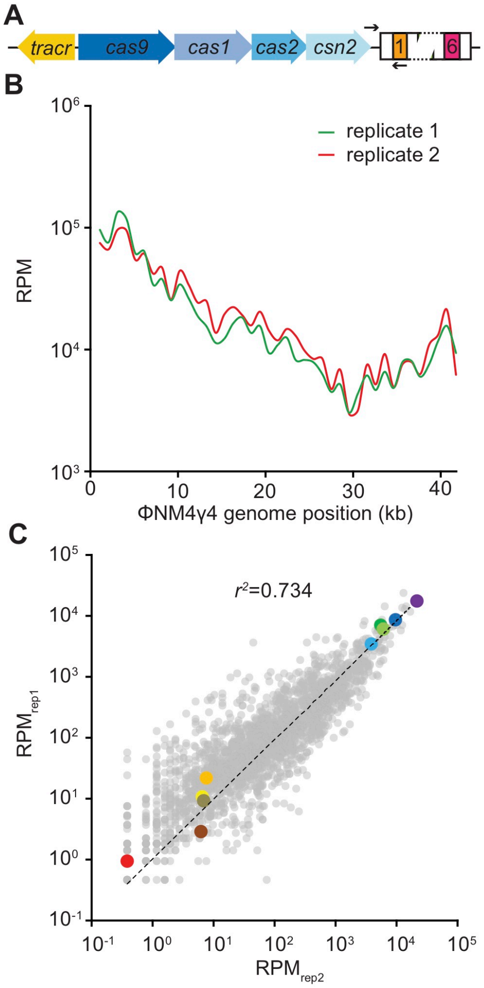
Besides the presence of a functional PAM, the rules that govern spacer acquisition in type II CRISPR-Cas systems are not completely understood. Multiple studies have shown an uneven pattern of spacer acquisition, where different spacer sequences have markedly different abundances within the population of cells that survive phage infection[40,84,90]. This observation led to the hypothesis that some spacers become overrepresented because they are more effective at directing targeting and/or cleavage by Cas9 and therefore have a selective advantage[40]. However, even when spacer acquisition was measured within 30 minutes of infection, i.e. before the viral lytic cycle is completed and the spacers cannot be selected for their abilities to guide DNA destruction, the pattern of spacer acquisition is constricted to the viral region that is first injected but with highly variable frequencies of acquisition for different spacers sequences within this genomic location[91]. These data suggest that the abundance of a spacer in the population of surviving bacteria can be independent of its targeting properties and determined solely by its acquisition rate.

Here I used the type II-A CRISPR system from *Streptococcus pyogenes* expressed in *Staphylococcus aureus* RN4220 cells[84] to investigate the mechanisms behind the pattern of spacer acquisition when cells are infected with the staphylococcal phage ϕNM4γ4[64,84]. First, I determined that this pattern is remarkably reproducible, with a set of spacer sequences consistently acquired at high frequencies. By measuring spacer abundance early and late during infection, I show that the frequency of individual spacers is mainly determined at the onset of infection and that there is little selection of spacer sequences thereafter. This led to the hypothesis that spacer abundance depends on the rate of acquisition rather than enhanced Cas9 cleavage activity. I tested this on selected spacer sequences at each end of the distribution spectrum by performing targeting assays and quantifying CRISPR acquisition of spacer-length oligonucleotides. These experiments demonstrated that high and low abundance spacers have similar targeting abilities but differ dramatically in their efficiency of acquisition. I established that the intrinsic spacer sequence dictates its acquisition rate, with the sequences proximal to the PAM being most critical. Our studies reveal that, for type II-A systems, spacer acquisition rates are fundamental to determine the distribution and diversity of the CRISPR-Cas immune response.

4.2. Spacer distribution is biased and reproducible

To analyze spacer distribution in the type II-A CRISPR system of *S. pyogenes* (Figure 4-1a) I performed infection assays with lytic phage ϕNM4γ4, as described previously[92]. DNA from surviving cells was used to amplify the CRISPR array by PCR and perform next generation sequencing of newly acquired spacers. I performed the infection in duplicate and obtained two libraries of 2.52 and 2.28 million phage-mapping reads, respectively. Of all the possible 2,318 NGG-adjacent protospacers on the genome of ϕNM4γ4, 2,096 were sampled in both libraries. The frequency of each spacer was normalized as reads per million (RPM) and plotted across the phage genome (1 kb bins, Figure 4-1b). I observed a similar pattern of spacer distribution for each duplicate experiment. To determine if the correlation is present not only in the groups of spacers within each 1 kb bin, but also at the level of the individual spacer sequences, I compared the RPM value for each of the 2,096 spacers (Figure 4-1c).

Figure 4-1. Acquired spacer sequences display a consistent distribution pattern. (a) Schematic diagram of the type II-A CRISPR system from Streptococcus pyogenes. Arrows indicate the position of the PCR primers used to check for spacer integration. **(b)** Average abundance (in reads per million per 1-kb bins, RPM) of ϕNM4γ4 viral sequences incorporated as spacers into the CRISPR array, mapped against location on the phage genome, in duplicate (red and green traces). **(c)** Individual spacers common to the two data sets in (b) were plotted with RPM values for replicate 1 on the x axis and replicate 2 on the y axis. The dotted line represents the linear regression fit. Ten spacers were color-coded based on their abundance (warm colors for low abundance and cold colors for high abundance).


A
tracr
cas9
cas1
cas2
csn2
1
6
B
10^6
10^5
10^4
10^3
RPM
replicate 1
replicate 2
0
10
20
30
40
ΦNM4γ4 genome position (kb)
C
10^5
10^4
10^3
10^2
10^1
10^0
10^-1
RPM_rep1
r²=0.734
10^-1
10^0
10^1
10^2
10^3
10^4
10^5
RPM_rep2

We found a remarkable correlation of the spacer frequencies in both replicas, particularly of the most abundant spacer sequences. I arbitrarily picked five spacer sequences with high or low RPM and marked them with different colors to follow their abundance over different experiments. This is an effort to illustrate the relative consistency in the distribution of individual spacer sequences, for example after mapping the spacers across the phage genome in our replicates (Figure 4-2a-c). To test if this correlation extends to experiments using other phages and type IIA CRISPR-Cas systems, I performed duplicate infection experiments of cells containing the *S. pyogenes* type IIA system with the phage ϕ85[93] (Figure 4-2d), or cells harboring the type IIA (also known as CRISPR3[17]) from *Streptococcus thermophilus* with ϕNM4γ4 (Figure 4-2e). Although I obtained fewer spacer reads in both cases (the efficiency of spacer acquisition is reduced under these conditions[84]), a very strong correlation for spacer abundance in the replicas was found. Altogether, these results indicate that the abundance of individual spacer sequences within the population of surviving cells is relatively constant during the type IIA CRISPR-Cas immune response.

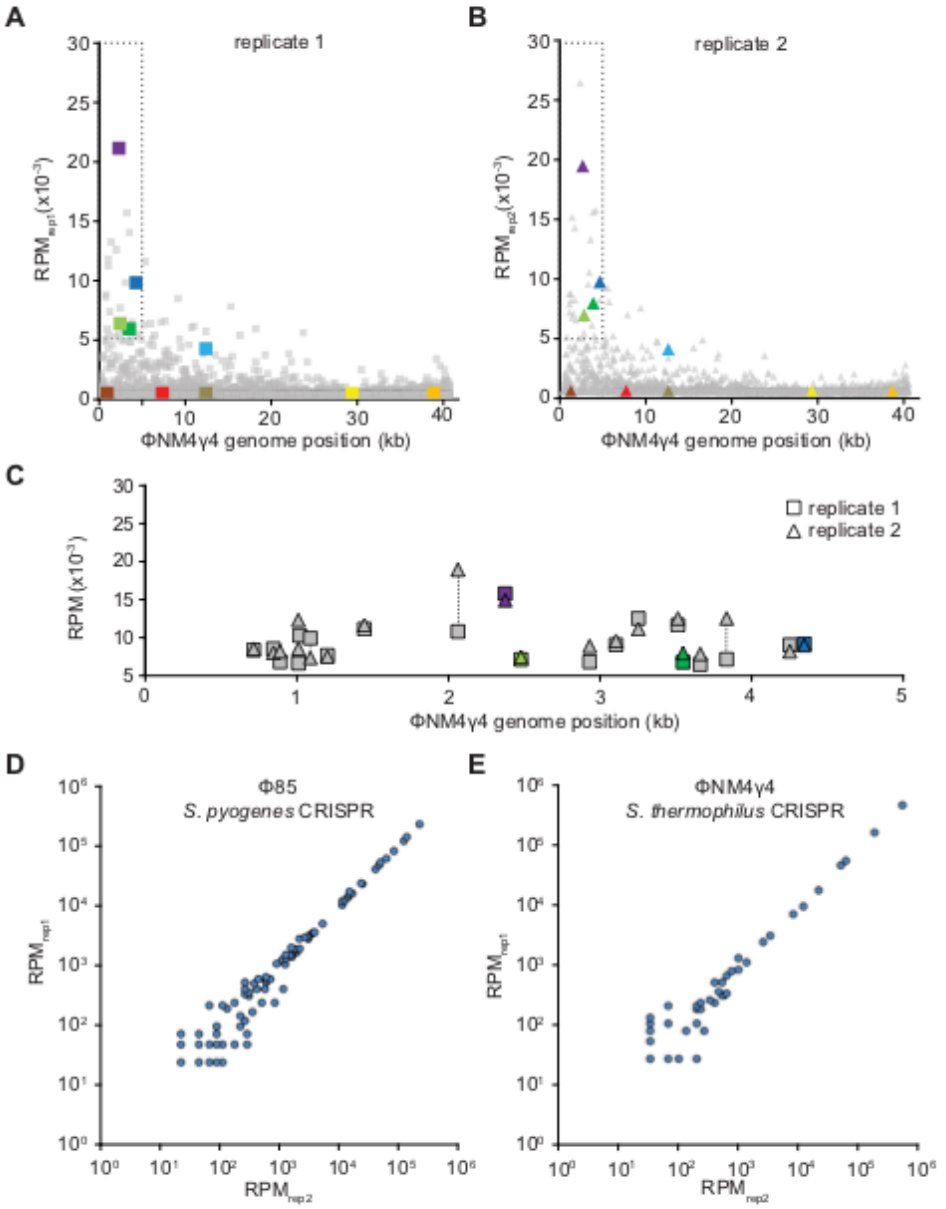
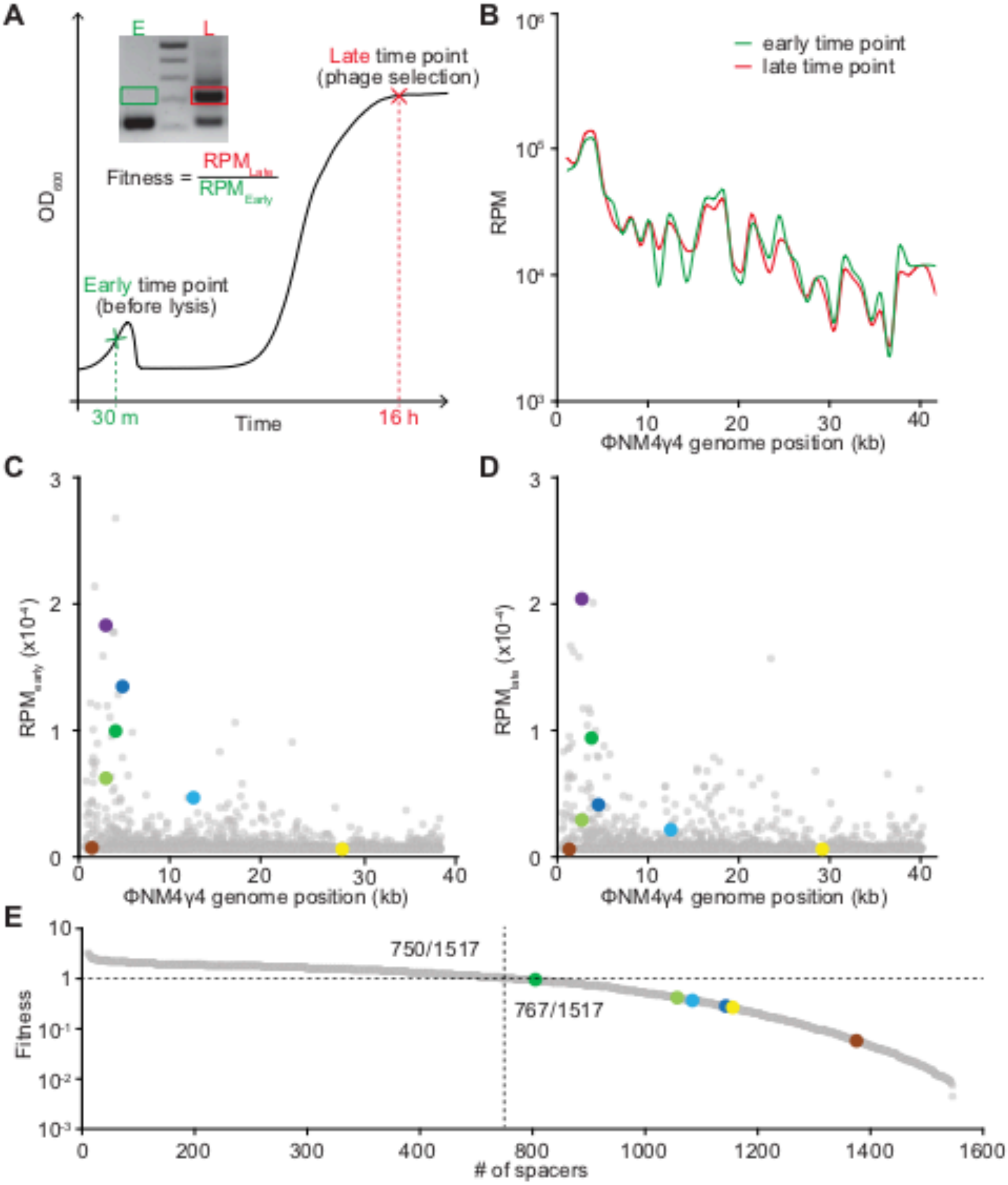


Figure 4-2. Biased sampling of phage DNA protospacers is a feature of other bacteriophages and type II CRISPR systems. (a), (b) Abundance (in reads per million, RPM) of ϕNM4γ4 viral sequences incorporated as spacers into the CRISPR array, mapped against location on the phage genome, in duplicate (raw data for Figure 4-1b and 4-1c). **(c)** Overlap of data in (a) and (b), zoomed on the first 5kb of the viral genome. Only spacers with RPM > 5,000 are shown **(d)**. RPM values of spacers sampled in two replicates during infection with lytic phage ϕ85 of cells harboring the Streptococcus pyogenes CRISPR system. **(e)** RPM values of spacers sampled in two replicates during infection with ϕNM4γ4 of cells harboring the Streptococcus thermophilus type II-A CRISPR3 system.

4.2.1. Effects of DNA cleavage efficiency on spacer distribution

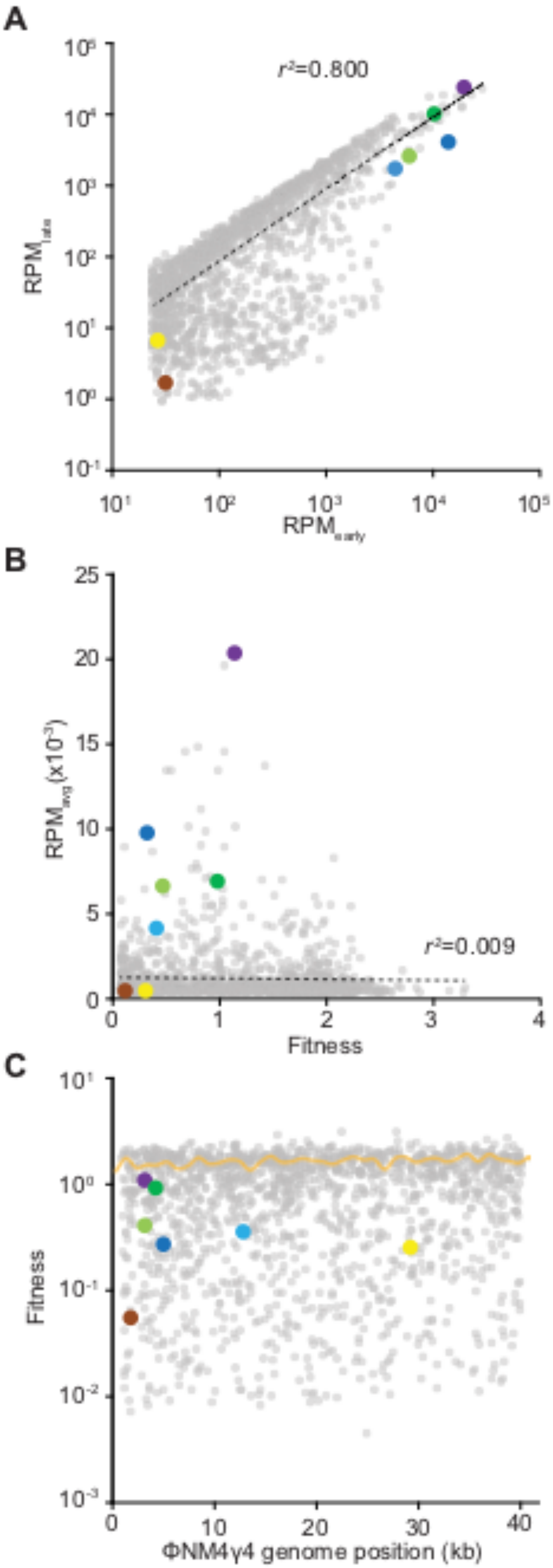
In principle, the different but reproducible abundance of spacers could be explained by two mutually non-exclusive forces that depend on their individual sequences: their inherent frequency of acquisition and/or their efficiency of viral targeting. To explore these possibilities, I compared the spacer distribution 30 minutes after infection, when the great majority of cells have not lysed yet (the ϕNM4γ4 viral cycle takes ~ 40 minutes), with the distribution obtained after 16 hours of infection, a time during which the acquired spacers can be selected against or for their targeting properties (Figure 4-3a).

Figure 4-3. Spacer distribution due to initial acquisition is not perturbed over the course of a live phage infection. (a) Diagram of assay used to measure the effects of interference efficiency on spacer abundance. **(b)** Average abundance (in reads per million per 1-kb bins, RPM) of ϕNM4γ4 viral sequences incorporated as spacers into the CRISPR array, mapped against location on the phage genome, in the early and late time point libraries (red and green traces). **(c), (d)** Abundance (in reads per million, RPM) of ϕNM4γ4 viral sequences incorporated as spacers into the CRISPR array, mapped against location on the phage genome in the early and late time point libraries (raw data for Figure 4-4). **(e)** Spacers ranked by decreasing fitness (ratio of abundance in late time point divided by abundance in early time point)


A
E
L
Late time point
(phage selection)
Fitness = RPM_Late / RPM_Early
OD600
Early time point
(before lysis)
30 m
Time
16 h
B
early time point
late time point
RPM
10^6
10^5
10^4
10^3
0
10
20
30
40
ΦNM4γ4 genome position (kb)
C
RPM_early (x10^-4)
3
2
1
0
0
10
20
30
40
ΦNM4γ4 genome position (kb)
D
RPM_late (x10^-4)
3
2
1
0
0
10
20
30
40
ΦNM4γ4 genome position (kb)
E
Fitness
10
1
10^-1
10^-2
10^-3
750/1517
767/1517
0
200
300
400
800
1000
1200
1400
1600
# of spacers

We analyzed over 0.72 million spacers for the early time point and 12.3 million spacers for the late time point, with 1,517 sequences shared between the two libraries. I observed a strong correlation for the values obtained at both time points for the frequency of each individual spacer (Figure 4-4a) and for their overall distribution across the phage genome (Figure 4-3b-d). This result suggests that spacer abundance is determined early after infection, and selection throughout the recovery of CRISPR-adapted cells has a minimal impact on shaping the spacer distribution. To explore this more directly, I calculated the fold-increase in abundance from the early time point to the late time point for each spacer. This value reflects the fitness of each sequence after its acquisition; i.e., the positive or negative selection suffered by a spacer due to its targeting abilities. I found that the fitness range of the entire spacer repertoire was narrow and did not correlate with spacer abundance (Figure 4-4b). For example, our set of highly abundant spacers had average finesses close to 1, even though they were order of magnitudes more frequent than other spacers with similar finesses (Figure 4-4b). Interestingly, I did not detect a strong positive selection for any spacer sequence (the maximum fitness value was 3.3), but there were 14 that displayed more than a 100-fold negative selection (Figure 4-4c, Figure 4-3e). On average, the acquired spacers have a fitness value close to 1 (Figure 4-4c), with approximately half of them displaying fitness higher than 1 and half lower than 1 (Figure 4-3e). These findings indicate that the relative abundance of spacer sequences is determined at their time of acquisition, early during the CRISPR-Cas immune response, and remains relatively constant during the targeting phase of CRISPR immunity.

Figure 4-4. The spacer distribution pattern is established early during infection. **(a)** Individual spacers common to the early and late time point samples plotted as RPM values against each other. **(b)** Spacer abundance in the live-phage sample (Figure 4-1) as a function of interference efficiency (fitness = abundance in late / early time point). **(c)** The fold increase in abundance in the late vs. early sample (fitness, y-axis) mapped across the phage genome. The yellow curve represents average fitness in 1-kb bins.


A
10^5
10^4
10^3
10^2
10^1
10^0
10^{-1}
RPM_{late}
r^2=0.800
10^1
10^2
10^3
10^4
10^5
RPM_{early}
B
25
20
15
10
5
0
RPM_{avg} ($x10^{-3}$)
r^2=0.009
0
1
2
3
4
Fitness
C
10^1
10^0
10^{-1}
10^{-2}
10^{-3}
Fitness
0
10
20
30
40
ΦNM4γ4 genome position (kb)

4.2.2. Effects of acquisition rates on spacer distribution

To test whether targeting efficiency affects the relative abundance of individual spacer sequences, I performed a barcoded, phage-free spacer acquisition experiment. For this I used a plasmid-based, modified type IIA CRISPR-Cas system (Figure 4-5a) in which a single-repeat CRISPR array was preceded by a random 10 nt sequence located 50 bp immediately upstream of the repeat, a barcoding strategy I previously used to count independent acquisition events[92]. In addition, expression of the *cas1*, *cas2* and *csn2* genes, essential for spacer acquisition, is controlled by an anhydrotetracycline-inducible promoter, allowing to turn on and off spacer integration[84,91]. Instead of using a live lytic virus, cells harboring this engineered CRISPR-Cas locus were transformed via electroporation with ϕNM4γ4 phage DNA, sheared into ~150 bp fragments by sonication, in the presence of anhydrotetracycline. After two hours the inducer was washed off, DNA was extracted from cells and the CRISPR loci along with barcoded leaders were amplified by PCR (Figure 4-5a) and subjected to next-generation sequencing. I analyzed 2.00 million spacer reads each with its respective barcode that sampled almost all (2,274) of the existing protospacers on the ϕNM4γ4 genome (Figure 4-6a-b). To test the barcoded system, I plotted the relative abundance versus the number of different barcodes for each individual sequence (Figure 4-6c). Assuming that different barcode sequences in front of the same spacer are the result of independent events of integration, this value reflects how many times a given spacer was acquired during

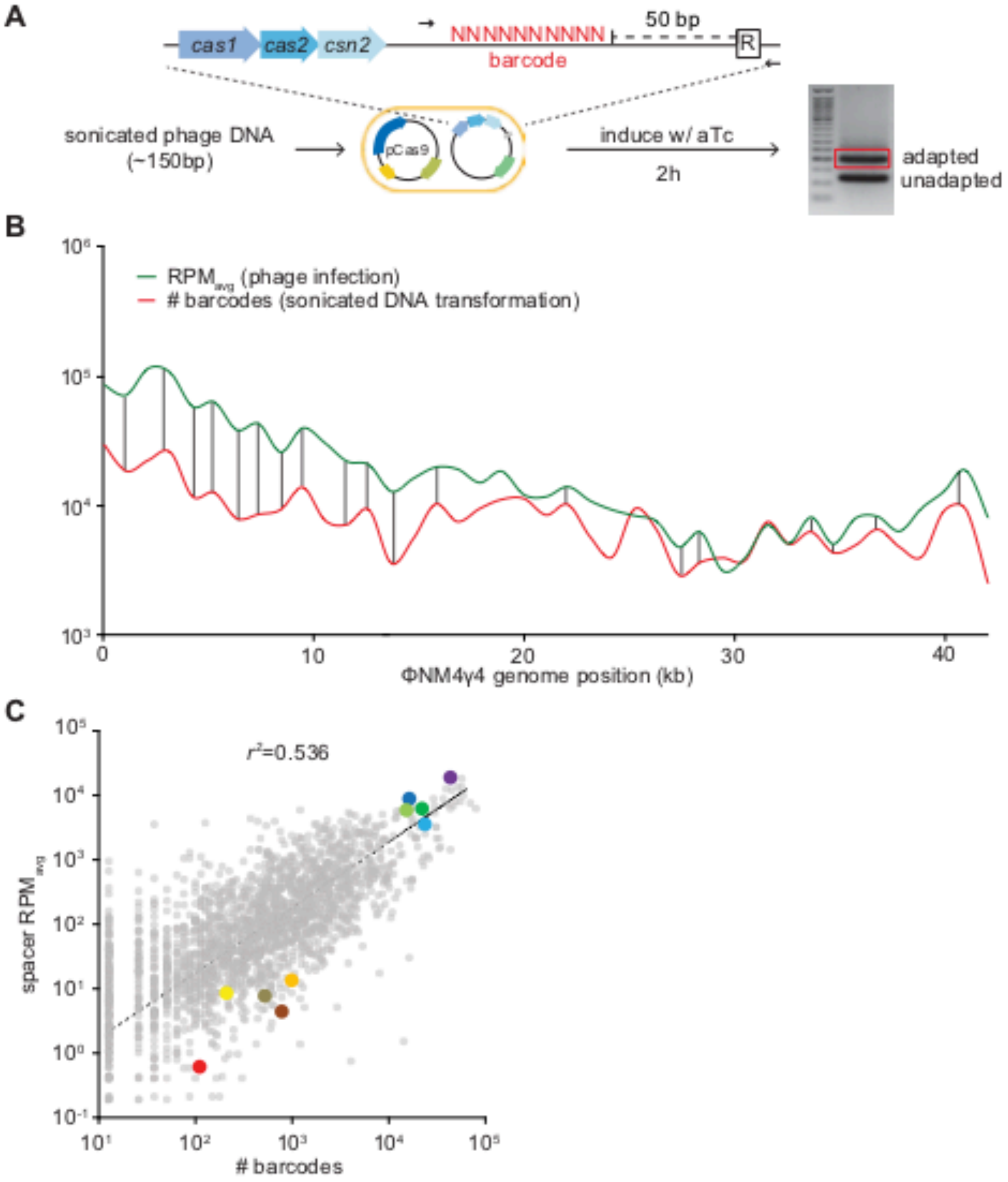


Figure 4-5. Spacer abundance is determined by the rate of acquisition. (a) Schematic diagram of the modified S. pyogenes CRISPR locus showing the location of the leader barcodes and primers (arrows) used to quantify the number of independent spacer acquisition events from sheared phage DNA. **(b)** Overlap of spacer distribution during a live phage infection (Figure 4-1) and number of barcodes as a measure of acquisition frequency, both plotted in 1-kb bins. **(c)** Comparison between abundance of individual spacers during a live phage infection and independent acquisition events from sheared phage DNA.

infection. I detected a strong correlation between the abundance of a spacer and its number of barcodes, a result that validates the use of barcode count as an absolute measure of the acquisition of a given spacer sequence present in the ϕNM4γ4 genome.

We then compared the number of barcodes with the number of reads obtained for each spacer sequence in the experiment using live phage presented Figure 1. In this way I can determine how much of the spacer distribution obtained after viral infection (measured as the average RPM of the replica experiments of Figure 4-1) can be explained by the intrinsic rate of acquisition of each viral spacer sequence (measured by the number of barcodes obtained in Figure 4-6). First I compared the distribution patterns across the ϕNM4γ4 genome (Fig. 4-5b). I found very similar distribution patterns, with a conservation of most peaks and valleys in both curves (note that the RPM and number of barcode values are intrinsically different and therefore the curves do not overlap). Next, I plotted both values against each other and found a good correlation, in which our ten selected spacers maintained their low or high abundance, and with an r^2 value of 0.536 (Fig. 4-5c). This indicates that the distribution of at least half of the spacers acquired in response to viral infection can be explained by their intrinsic rate of acquisition; i.e. independent of the targeting abilities of the spacer sequence.

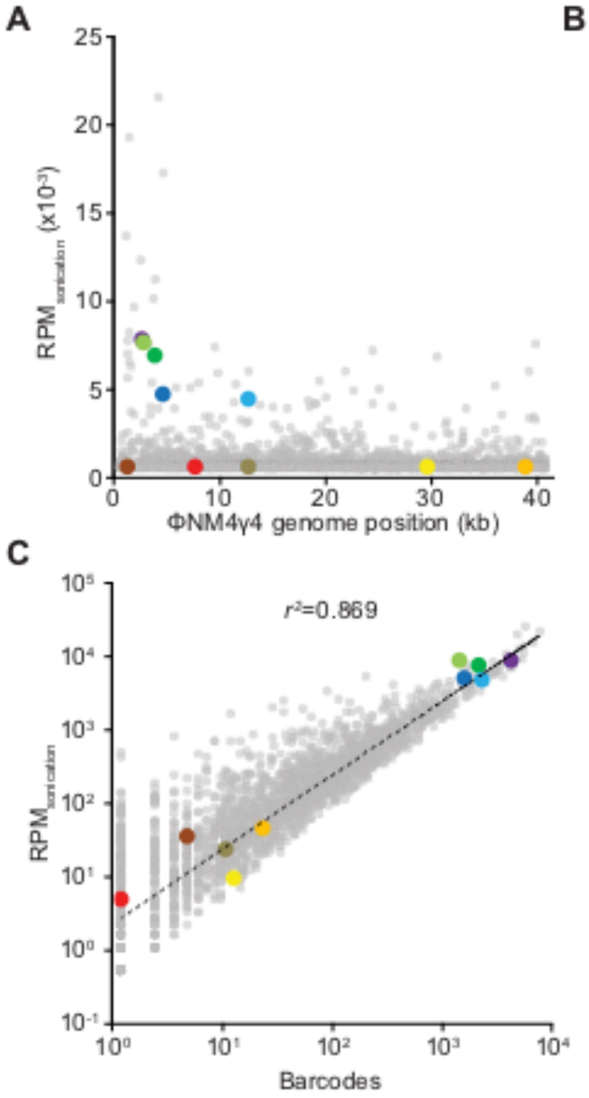


Figure 4-6. Oversampling of phage protospacers is due higher rates of acquisition. (a) (b) Abundance (in reads per million, RPM) of spacers incorporated into the CRISPR array, mapped against location on the phage genome, following electroporation of sheared phage DNA (raw data for Figure 4-5). **(c)** Abundance of individual spacers following electroporation of sonicated phage DNA plotted against barcodes as a measure of the number of times each spacer was acquired.

4.3. Analysis of spacer sequences that determine the rate of acquisition

Our data suggests that the efficiency of the spacer acquisition process, i.e. the selection and integration of a PAM-flanked phage sequence that happens early during infection, is the most important factor to determine the abundance of a spacer sequence during the CRISPR-Cas immune response. If this hypothesis is correct, when comparing a high- and low-abundance spacer it would be expected that (i) their ability to direct Cas9-mediated DNA cleavage should be similar, and (ii) their rate of acquisition should be dramatically different. To test these predictions, I selected two spacer sequences that were consistently over- and under-represented (the "dark green" and "red" spacers, respectively, in Figure 4-1c) in all our assays (Fig. 4-7a).

We tested the first prediction by comparing the efficiency of in vitro DNA cleavage by Cas9 using each of these spacers as guides and I found similar cleavage kinetics (Fig. 4-7b and Figure 4-8a-c). Second, I measured the cleaving capacity of each of these spacers in vivo, through the quantification of the reduction in phage propagation that they mediate (Fig. 4-7c). I did not detect significant differences between the spacers, a result that demonstrates that not only in vitro, but also in vivo, these sequences provide similar levels of defense. Because the sequences I decided to follow in these assays reside in different regions of the phage genome, which could be a variable affecting their targeting ability[91], I repeated these assays with two other sequences that overlap with each other but have markedly dissimilar abundances ("light blue" and "tan" in Figure 4-1c, Figure 4-9a). Again, I found no differences in DNA cleavage in vitro (Figures 4-8a,d,e and 4-9b) or in vivo (Fig. 4-9c). All these results corroborate the prediction that DNA targeting does not dictate spacer abundance.

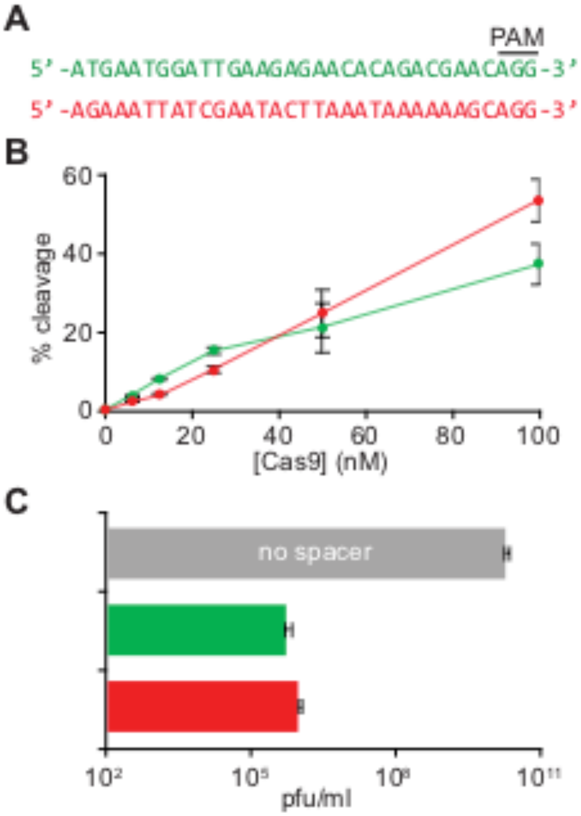


Figure 4-7. High and low abundance spacers have similar interference efficiencies. (a) Sequences of a high abundance (Figure 4-1, dark green) and low abundance (Figure 4-1, red) spacer, following infection of CRISPR cells with live phage. **(b)** Quantification of in vitro cleavage of a 2-kb phage target by various concentrations of Cas9 loaded with sgRNAs corresponding to the two spacers in (a). **(c)** Phage propagation measured as the number of plaque forming units (pfu) per ml of stock, on cells without CRISPR or cells loaded with Cas9 and a spacer targeting either one of the phage protospacers in (a).

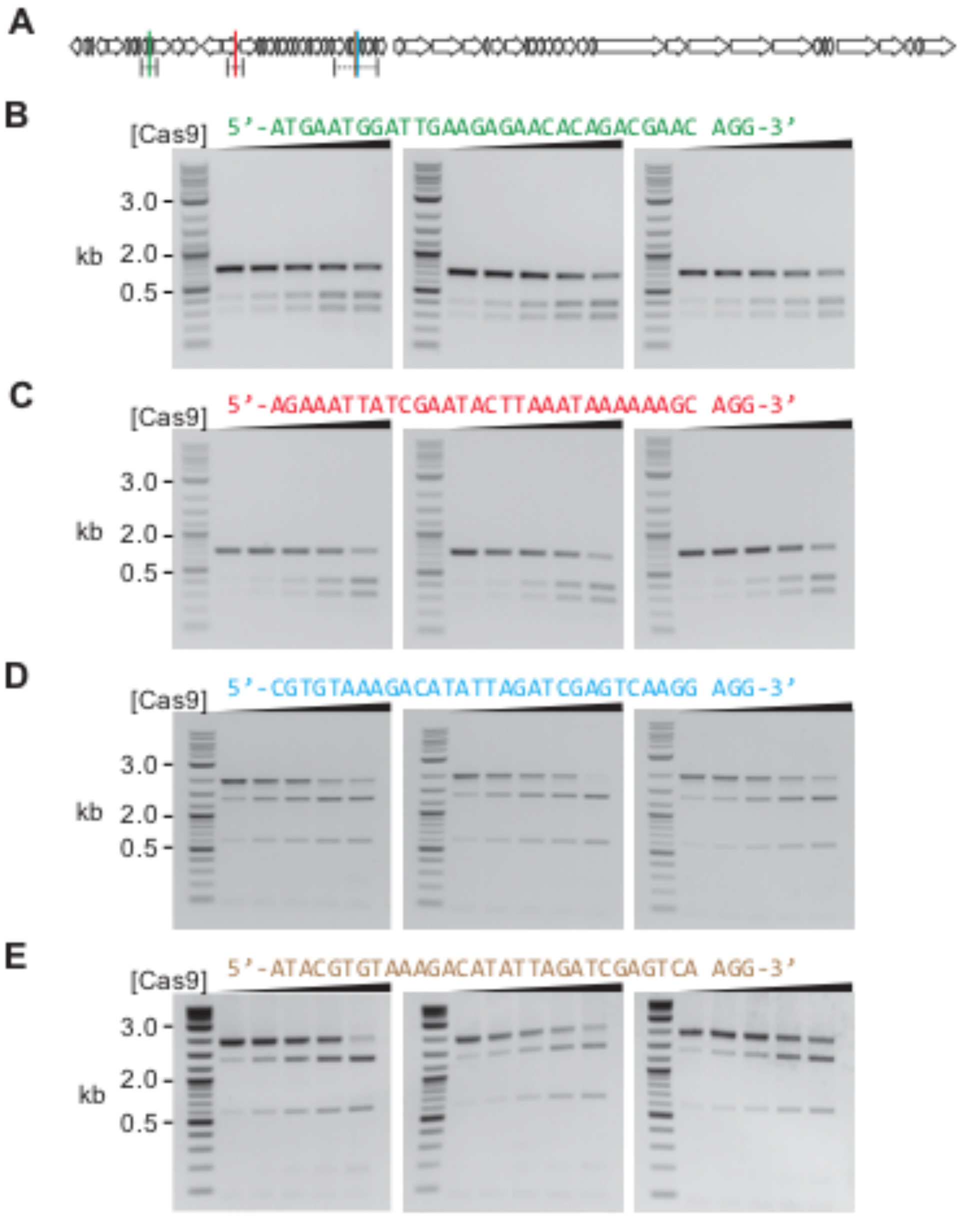


Figure 4-8. High and low abundance spacers have similar interference efficiencies. (a) Location on the phage genome of the spacers colored dark green, red, light blue and tan in Figure 4-1. **(b)** Agarose gels in triplicate of in vitro cleavage products of 2-kb phage targets by various concentrations of Cas9 loaded with sgRNAs corresponding to the four spacers in (a). The tested Cas9 concentrations were 6.26, 12.5, 25, 50 and 100nM.

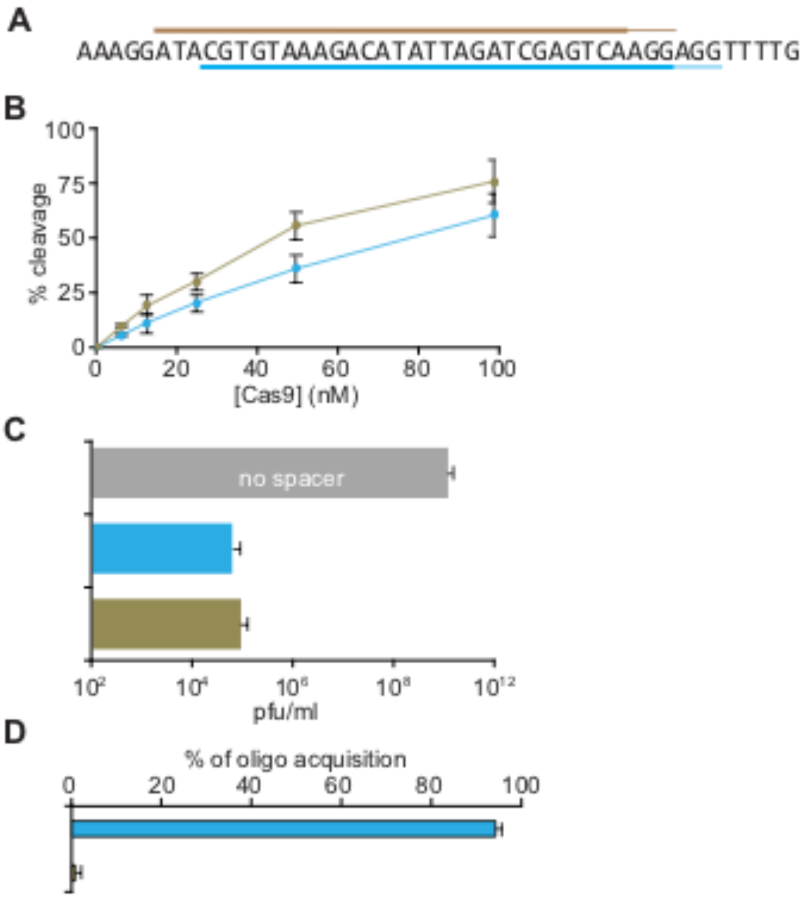
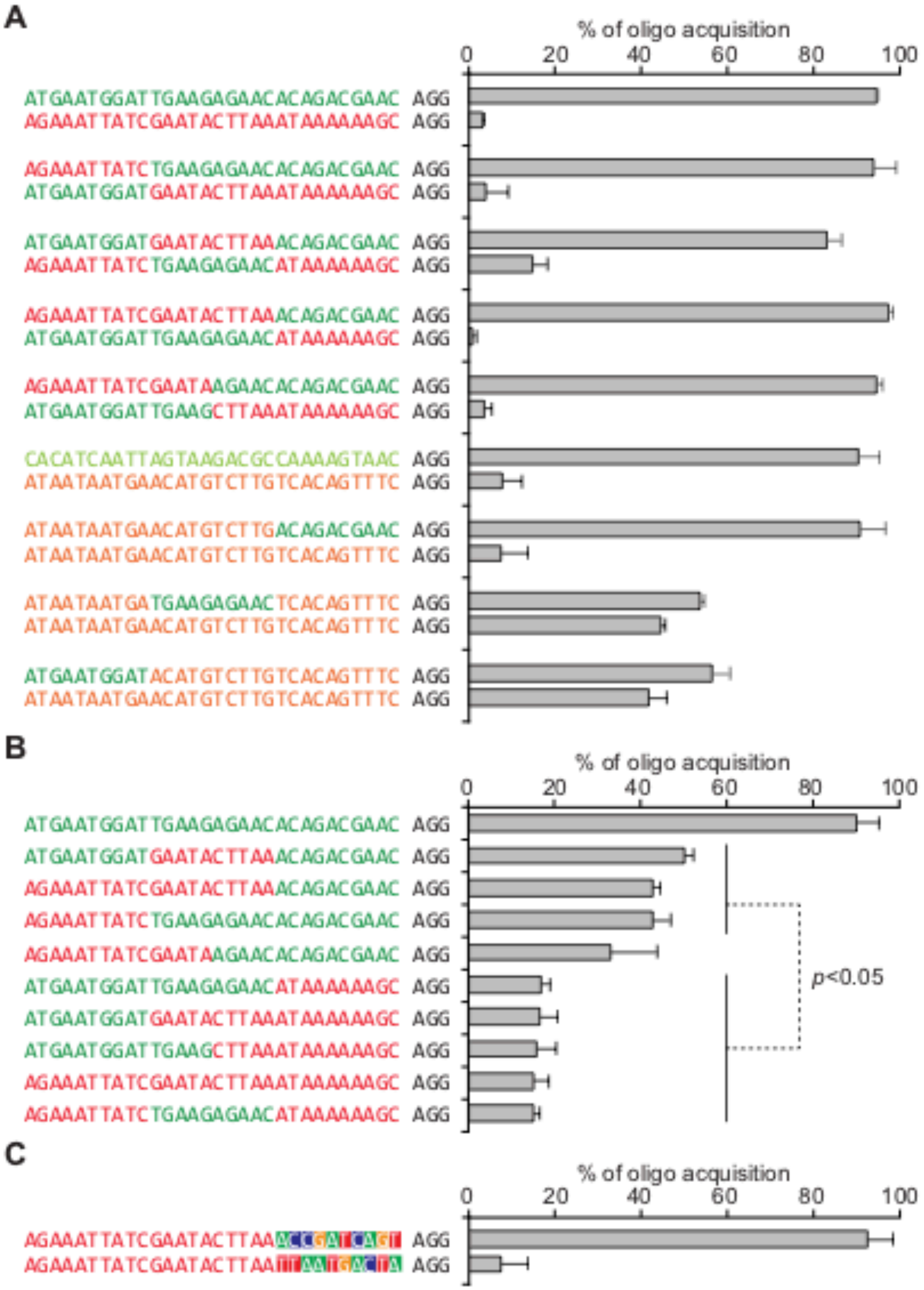


Figure 4-9. Partially overlapping spacers can provide similar interference efficiencies, in spite of highly dissimilar abundances. (a) Sequence of two high abundance (Figure 4-1, light blue) and low abundance (Figure 4-1, tan) overlapping spacers. **(b)** Quantification of in vitro cleavage of a 2-kb phage target by various concentrations of Cas9 loaded with sgRNAs corresponding to the two spacers in (a). **(c)** Phage propagation measured as the number of plaque forming units (pfu) per ml of stock, on cells without CRISPR or cells loaded with Cas9 and a spacer targeting either one of the phage protospacers in (a). **(d)** Relative acquisition rates (%) following electroporation of a single dsDNA oligonucleotide containing both protospacers.

To test the second prediction, that super-spacers have an intrinsic higher rate of acquisition, I designed an assay to compare the relative frequency of CRISPR incorporation of different sequences. I co-transformed staphylococci carrying the engineered type IIA CRISPR-Cas locus used for the acquisition of spacers from sheared phage DNA with annealed, dsDNA oligonucleotides at equimolar concentrations. Transformation was followed by next-generation sequencing of the amplified CRISPR array to quantify the relative frequency of acquisition for each transformed oligonucleotide. First, I compared the acquisition of the selected over- and under-represented sequence ("dark green" and "red" sequences, respectively, in Figure 4-1c), using oligonucleotides containing only the 30-nt spacer sequence followed by the 3-nt PAM (Figure 4-7a). I observed a striking difference in the number of reads, with ~ 96 % of the reads from oligo-derived spacers matching the highly abundant sequence (Fig. 4-10a).

Figure 4-10. Spacer sequences affect their rate of acquisition. Relative acquisition rates (%) of spacers following electroporation of various pairs **(a)(c)** or a set of 10 **(c)** dsDNA oligonucleotides mixed equimolar ratios.

To corroborate this finding, I performed spacer-specific PCR after transformation using each of the spacer sequences as reverse primers to amplify the CRISPR array. Consistent with our next generation sequencing data, I was able to detect a strong PCR product only when using the highly acquired spacer as reverse primer (Fig. 4-11a). I repeated these assays using extended oligonucleotides harboring additional (15-nt) phage sequences flanking the spacers and obtained similar results (Figure 4-11a-c). In addition, I compared the frequency of acquisition of another high- and low-abundance spacer pair (the “light green” and “orange” spacers in Figure 1c, respectively), and observed the same differential integration into the CRISPR array (Fig. 4-10a). Finally, I measured acquisition of the two overlapping spacers with dramatically different abundances (Figure 4-9a) and found that even when a single dsDNA oligonucleotide containing both sequences is used, mostly the abundant spacer sequence is acquired (Figure 4-9d). Altogether, these experiments demonstrate that for a given spacer sequence, its efficiency of acquisition but not its targeting capabilities, correlate with its abundance in the population of CRISPR-resistant cells.


A
% of oligo acquisition
0
20
40
60
80
100
ATGAATGGATTGAAGAGAACACAGACGAAC AGG
AGAAATTATCGAATACTTAAATAAAAAAGC AGG
AGAAATTATCTGAAGAGAACACAGACGAAC AGG
ATGAATGGATGAATACTTAAATAAAAAAGC AGG
ATGAATGGATGAATACTTAAACAGACGAAC AGG
AGAAATTATCTGAAGAGAACATAAAAAAGC AGG
AGAAATTATCGAATACTTAAACAGACGAAC AGG
ATGAATGGATTGAAGAGAACATAAAAAAGC AGG
AGAAATTATCGAATAAGAACACAGACGAAC AGG
ATGAATGGATTGAAGCTTAAATAAAAAAGC AGG
CACATCAATTAGTAAGACGCCAAAAGTAAC AGG
ATAATAATGAACATGTCTTGTCACAGTTTC AGG
ATAATAATGAACATGTCTTGACAGACGAAC AGG
ATAATAATGAACATGTCTTGTCACAGTTTC AGG
ATAATAATGATGAAGAGAACTCACAGTTTC AGG
ATAATAATGAACATGTCTTGTCACAGTTTC AGG
ATGAATGGATACATGTCTTGTCACAGTTTC AGG
ATAATAATGAACATGTCTTGTCACAGTTTC AGG
B
% of oligo acquisition
0
20
40
60
80
100
ATGAATGGATTGAAGAGAACACAGACGAAC AGG
ATGAATGGATGAATACTTAAACAGACGAAC AGG
AGAAATTATCGAATACTTAAACAGACGAAC AGG
AGAAATTATCTGAAGAGAACACAGACGAAC AGG
AGAAATTATCGAATAAGAACACAGACGAAC AGG
ATGAATGGATTGAAGAGAACATAAAAAAGC AGG
ATGAATGGATGAATACTTAAATAAAAAAGC AGG
ATGAATGGATTGAAGCTTAAATAAAAAAGC AGG
AGAAATTATCGAATACTTAAATAAAAAAGC AGG
AGAAATTATCTGAAGAGAACATAAAAAAGC AGG
p<0.05
C
% of oligo acquisition
0
20
40
60
80
100
AGAAATTATCGAATACTTAAACCGATCAGT AGG
AGAAATTATCGAATACTTAATTAATGACTA AGG

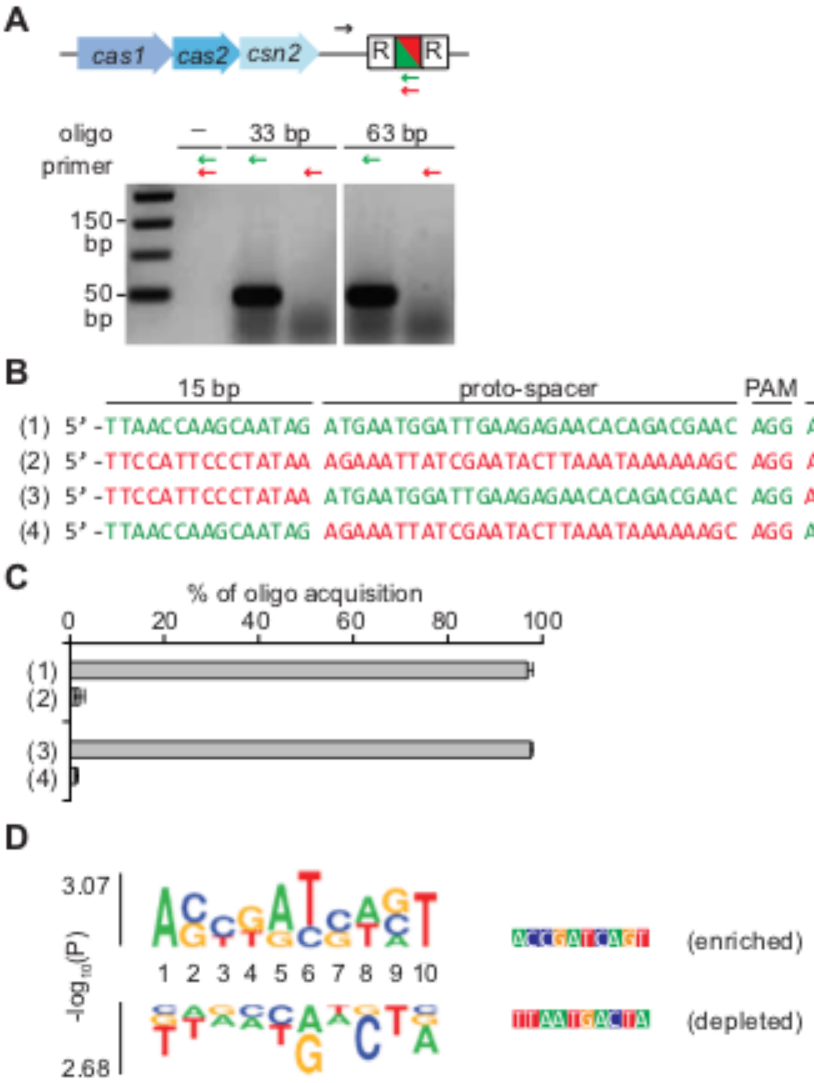


Figure 4-11. Sequence determinants within the protospacer sequence affect the rate of acquisition. (a) Qualitative PCR to assess the integration of a low abundance (Figure 4-1, red) and high abundance (Figure 4-1, dark green) spacer after electroporation of corresponding dsDNA oligonucleotides mixed in a 1:1 molar ratio. The oligonucleotides tested were 33bp long (protospacer + PAM only) or 63bp long (15bp upstream + protospacer + PAM + 15bp downstream). Reverse primers anneal on the integrated spacers. **(b)** Sequence of oligonucleotides containing the protospacers in (a) with 15bp upstream and downstream of the protospacer swapped or not. **(c)** Relative acquisition rates (%) of spacers in (b) following electroporation of pairs of dsDNA oligonucleotides mixed equimolar ratios. **(d)** Unweighted probability Logo of the top 1% protospacers generated using kpLogo (showing only 10bp upstream of the PAM). Nucleotides shown on top were enriched, while the ones shown on the bottom were depleted in the spacers used to create the logo. Enriched or depleted consensus sequences are shown to the right.

The above results suggest that there must be some element in the sequence of super-spacers that increases their rate of acquisition. To test this, I divided the sequence of the spacers into PAM-distal, middle and PAM-proximal 10-nt regions (Figure 4-10a) and swapped these regions in the super-spacer and low abundance spacer sequences. Electroporation with different pairs of swapped oligos, followed by next generation sequencing of expanded CRISPR arrays revealed that the presence of the 10-nt PAM-proximal region of the super-spacer was necessary and sufficient to ensure high levels of acquisition of a dsDNA oligo (Figure 4-10a). Moreover, the addition of the 10-nt PAM-proximal region of the “dark green” highly acquired spacer, but not the middle or PAM-distal sequences, was also sufficient to increase the frequency of acquisition of the “orange” low-abundance spacer (Figure 4-10a). To corroborate these findings, I co-transformed 10 different dsDNA oligonucleotides containing different combinations of 10-nt regions of the “dark green” and “red” spacer sequences (Figure 4-10b). Again, I found that dsDNA oligos containing the 10-nt PAM-proximal sequence of the highly acquired spacer were integrated into the CRISPR array at significantly higher frequencies than those having the same region from the low-abundance spacer. Finally, due to the impossibility of testing every acquired spacer via oligo transformation, I evaluated the importance of this sequence within the entire set of acquired spacers. To do this, I used kpLogo[94] to look for a conserved motif in the PAM-proximal 10-nt sequence of either the most abundant spacers (in the top 1 %, of average spacer reads in Fig. 1c). This analysis yielded two motifs, corresponding to the enriched and depleted consensus within this sequence (Figure 4-11d). I appended these motifs to the low abundance ("red") spacer to check for their influence in spacer acquisition. I found that the PAM-proximal motif derived from highly abundant spacers dramatically increased spacer acquisition (Figure 4-10c). The overall results of these experiments demonstrate that specific DNA sequences located immediately upstream of the PAM have important effects on the frequency of acquisition of the 30-nt spacer determined by that PAM.

4.4. Discussion

Early studies of the type II CRISPR-Cas response to phage infection have shown that the population of surviving bacteria has a diverse content of new spacer sequences, some much more abundant than others[40,90,91]. In principle, the abundance of a spacer should be determined by two factors: its frequency of integration into the CRISPR array and its targeting capabilities[95]. Here I found that the abundance of most spacers is determined early during infection, when positive or negative selection for good or bad targeting, respectively, is still not a factor at play. In addition, there is a strong correlation between the abundance of most spacers acquired during infection with live phage and their abundance after transformation with sheared phage DNA, again, when targeting is not required for survival. Finally, I showed that the frequency of most spacers in the surviving population correlates directly with their frequency of acquisition.

The data presented here show that the spacer abundance that emerges after the type II CRISPR-Cas immune response is basically determined shortly after infection, depending mostly on the acquisition rate of each acquired sequence and not on its properties as a guide for Cas9 DNA cleavage. This is also the prediction of theoretical analysis[95]. Modeling of the CRISPR-Cas immune response determined that high spacer acquisition probabilities will lead to greater diversity in the spacer distribution, while strong selection of spacers providing better phage clearance will tend to homogenize the population of spacers in favor of the most effective one ("winner takes all" situation). Previous studies in our lab that evaluated the effect of the concentration of CRISPR-adapted cells on immunity[91,96] could provide an explanation for such model for the impact of spacer acquisition on their distribution. The results showed that at very low concentrations of immune cells there is a marked effect on the recovery of these immune cells after infection. In this situation, equivalent to low acquisition rates, leads to the positive selection of spacers that are better at targeting due to their position in the CRISPR array[96] or because they guide Cas9 to the phage genome immediately after its injection[91]. On the other hand, when CRISPR-immune cells have high concentrations, the targeting efficiency of the spacers does not impact the host's growth after phage addition. Although both of these studies investigated single-spacer cultures, I believe that a similar scenario can happen during the infection of naïve cultures that acquired multiple (thousands) of new spacers. The high rate of acquisition of certain sequences would effectively create a high concentration of immune cells that will provide most of the immunity to the population, and no further selection of these sequences due to their targeting efficiency will take place.

Our findings showed that spacer abundance is mostly determined at the acquisition stage of type II CRISPR-Cas immunity. The uneven distribution of different spacer sequences could be in principle explained by the existence of phage genomic regions that are better substrates for spacer acquisition. Indeed, this is the case for the regions that first enter the host cell[91] and is a possible explanation for the clustering of highly abundant spacers from the 5' end of the ϕNM4γ4 genome (Figure 4-2a-b). However, even within this region (and also close to the *cos* site in ϕ12γ3[91]) there is a wide spectrum of spacer abundance. Here I showed that one explanation for these different abundances is the intrinsic frequency of acquisition of a given spacer sequence. Mutagenesis analysis revealed that the 10-nt sequence at the PAM end of a spacer is determinant for its frequency of acquisition and that there are conserved nucleotides within this region critical for the acquisition process. The molecular mechanisms behind this preferential acquisition are intriguing. The current model of spacer acquisition by type II CRISPR-Cas systems involves three steps: phage DNA is degraded by AddAB to create the spacer substrates[91,97], these are selected and processed by a Cas9-Cas1-Cas2-Csn2 complex[84] and finally the processed spacer sequence is integrated by the Cas1-Cas2 integrase into the CRISPR array[98]. Future experiments will investigate the impact of specific spacer sequences in the efficiency of these steps. In summary, our study begins to uncover the rules that govern the generation of immunological diversity during the type II CRISPR-Cas response to phage infection, revealing that spacer acquisition, a unique feature of these systems, is a key determinant to the structure of the surviving population.

Chapter V.

Perspectives

CRISPR-Cas is a DNA-encoded, sequence-specific immune system that protects bacteria and archaea against phages and other genetic elements. The adaptive feature of CRISPR-Cas immune systems relies on their ability to memorize DNA sequences from invading molecules (acquisition) and allows them to rapidly adapt against new threats. While recent research has drastically improved our understanding of adaptation, future studies will continue to address outstanding questions about the molecular mechanisms and technological applications of CRISPR, in general, and spacer acquisition, in particular.

Molecularly, the functions of Cas1 and Cas2, two signature proteins present in all CIRSPR systems, have been thoroughly investigated. In the type I CRISPR system of E. coli, the Cas1-Cas2 complex is bound to a partially duplexed dsDNA (pre-spacer)[79]. The complex recognizes specific sequences upstream the CRISPR array to ensure leader-polarized spacer integration. This process is facilitated by host factors such as IHF (integration host factor)[99,100]. Similar findings were reported in type II-A CRISPR system of Streptococcus pyogenes[96,101,102]. Integration of the spacer into the arrays is mediated by the integrase activity of Cas1[79] and the presence of a correct PAM in the pre-spacer facilitates integration in the right orientation[103].

By contrast, little is known about the other Cas proteins involved in spacer acquisition. In type I-C of *Bacillus halodurans*, Cas4 has been shown to enhance the formation of functional memory by assisting PAM-compatible spacer selection[104]. In type II, Csn2 and Cas9 form a complex with Cas1 and Cas2[84] and are involved in adaptation[5]. Structural studies have revealed that Csn2 forms a ring-like structure around DNA, suggesting that it might recruit other proteins involved in spacer selection or integration[63]. In addition, the tracrRNA is also required[84], suggesting that the apo-Cas9 structure[77], very different from holo-Cas9[71], does not have the correct conformation to participate in spacer acquisition. In addition to destroying the invading virus, Cas9 specifies PAM-flanking viral sequences during adaptation to ensure only functional spacers are acquired. This is in contrast to type I, where the Cas1-Cas2 complex is necessary and sufficient to direct incorporation of new spacers with correct PAMs. Therefore, the motif is sensed by only one protein in type II (Cas9)[89], but by different protein complexes in type I: Cas1-Cas2 during acquisition and the Cascade complex during interference[105]. Recognition of the PAM by the Cascade leads to accelerated integration of new spacers with correct PAMs during primed acquisition. By contrast, priming is yet to be shown to be a feature of type II CRISPR systems. Future work will address these and other mechanistical and molecular aspects of spacer acquisition in different CRISPR-Cas systems.

In recent studies[89,106], the *E. coli* type I-E CRISPR-Cas adaptation machinery has been repurposed as a recording device to store information (such as environmental signals) in the form of spacers in the CRISPR array. Because the adaptation frequency is relatively low, decoding requires deep sequencing of a population of cells. This limits the number of stimuli that can be recorded. Using hyperactive adaptation machinery such as hCas9 can boost the adaptation frequency and thus the recording capacity of such synthetic devices. Moreover, combined with introduction of sheared genomic DNA, the hyperactive CRISPR adaptation machinery is able to sample much larger genomes than the type I-E Cas1-Cas2 complex and can be used to generate diverse and unbiased gRNA libraries *in vivo*. Further optimization of CRISPR-based molecular recording technologies, such as TRACE[106], will enable high throughput parallel temporal recordings of biological states, such as fluctuations in gene expression or metabolite concentration.

Besides these direct applications of spacer acquisition, CRISPR has emerged as a powerful DNA-editing technology used across all fields of biomedical research. Furthermore, CRISPR is expected to have tremendous contributions to agriculture and treatment of human diseases. In the U.S., the first clinical trials of CRISPR to treat genetic disorders like beta-thalassemia and sickle cell disease are expected to begin in 2018.

Chapter VI.

Material and Methods

Bacterial strains and growth conditions

Cultivation of *Staphylococcus aureus* RN4220[72] was carried out in heart infusion broth (BHI) at 37°C. Whenever applicable, media were supplemented with chloramphenicol at 10 μg/mL, erythromycin at 10 μg/mL or spectinomycin at 250 μg/mL to ensure maintenance of pC194, pE194 and pLZ12 derived plasmids, respectively, or 5 mM CaCl2 for phage adsorption.

Directed evolution of *cas9*

The *cas9* gene was mutagenized at a low rate of 0-4.5 mutations/kb by error prone PCR using GeneMorph II Random Mutagenesis Kit. The mutant *cas9* amplicons were cloned into a backbone plasmid containing a spacer matching a TAG-adjacent target on ϕNM4γ4. The library was subjected to soft-agar lytic phage infection and surviving colonies were re-streaked on fresh plates. The TAG-cleaving efficiency of surviving colonies was individually assessed by phage propagation assays.

Spacer acquisition assay

Spacer acquisition assays of cells harboring the full CRISPR system of *Streptococcus pyogenes* were performed as described previously, both in liquid and on plate[84]. For plate acquisition assays, overnight cultures were launched from single colonies and diluted to equal optical densities. CRISPR arrays were amplified by PCR with primer pairs L400-H050 or L400-H052 (Supplementary table S3).

Bacterial growth curves

Overnight cultures were launched from single colonies and diluted 1:100 in BHI. After 1 hour of growth, optical density at 600 nm (OD_{600}) was measured for each culture, and samples were brought to equal cell densities and loaded into 96-well plates along with ϕNM4γ4 at MOI =1. Measurements were taken every 10 minutes for 24 hours.

Cas9 target cleavage assay

Cas9 was expressed and purified as previously described (Jinek et al., 2012). The I473F Cas9 expression vector was cloned by around-the-horn mutagenic PCR. crRNA and tracrRNA were transcribed using T7 RNA polymerase from single-stranded DNA templates and hybridized as previously described[34,67]. L2 oligonucleotides (Supplementary table S3) were hybridized to generate the two different target DNA duplexes and native PAGE-purified before 5' radiolabeling using [γ-^{32}P]-ATP (Perkin-Elmer) and T4 polynucleotide kinase (New England Biosciences). Cleavage assays were carried out essentially as previously described (Sternberg et al., 2014). In brief, Cas9 and crRNA:tracrRNA were allowed to form an RNP complex before addition of target DNA. Final concentration of RNP was 100 nM and target was 1 nM. Reactions were incubated at room temperature, and aliquots were taken at 0.25, 0.5, 1, 2, 5, 10, 30, and 60 minutes and quenched by addition of an equal volume of 95% formamide and 50 mM EDTA. Samples were run on 10% urea-PAGE, visualized by phosphorimaging, and quantified using ImageQuant (GE Healthcare). cleavage by Cas9 of various targets was assessed using the Guide-It Complete sgRNA Sreening System from Clontech (Cat. No. 632636) with minor modifications. Cas9 and the sgRNAs were pre-incubated for 5 min at 37C in equimolar ratio and then diluted into the cleavage reaction to final concentrations of 100, 50, 25, 12.5 and 6.25nM. All reactions contained 10nM of a phage-derived PCR template with the target site. All reactions were stopped after 5 minutes by heat inactivation at 80C for 5 minutes and stored at -80C until ready to be run on an agarose gel.

Phage Interference Assay

Overnight cultures were launched from single colonies. Serial dilutions of a stock of phage ϕNM4γ4[64] were spotted on fresh soft heart infusion agar (HIA) lawns of targeting cells containing chloramphenicol 10 µg/ml and 5 mM CaCl2. Plates were incubated at 37 °C overnight and interference efficiency was measured in plaque forming units (pfu).

Acquisition from live phage

Acquisition from live phage in cells harboring the CRISPR system of *Streptococcus pyogenes* (plasmid pWJ40) or CRISPR3 of *Streptococcus thermophilus* (pRH200) was performed as described previously[84]. In Figure 4-3 and 4-4, plasmid pWJ40* containing randomized leader barcodes was used instead of pWJ40[92]. The unweighted probability Logo of the top 1% protospacers was generated using kpLogo[94].

Acquisition from shredded phage DNA

Phage DNA was shredded by sonication to fragments of ~150bp as described[91]. Following dialysis, 100µg of phage DNA was electroporated into competent *S. aureus* cells carrying plasmids pRH317 and pRH318*. Cells were recovered for 2h in BHI supplemented with anhydrotetracycline at 1µg/µl.

Acquisition from dsDNA oligos

dsDNA substrates were obtained by annealing ssDNA oligos in Duplex Buffer from IDT. Following dialysis, 100nm of each competing dsDNA substrate were mixed and electroporated in competent *S. aureus* cells carrying plasmids pRH223 and pRH240[84]. Cells were recovered for 2h in BHI supplemented with anhydrotetracycline at 1µg/µl. Need to write the electroporated oligos for all samples either here or in a table.

High-throughput sequencing

Plasmid DNA was extracted from adapted cultures. 200 ng of plasmid DNA was used as template for Phusion PCR to amplify the CRISPR locus with primer pairs H370-H371 (Figure 4-1, 4-2d), H180-B153 (Figure 4-2e), H370-H366 (Figure 4-5, early timepoint), H372-H366 (Figure 4-5, late timepoint) and H186-H366 (oligo electroporation). Following gel extraction and purification of the adapted bands, samples were subject to Illumina MiSeq (Figures 4-1, 4-2, 4-5, 4-6, 4-7) or NextSeq (Figures 4-3, 4-4) sequencing. Data analysis was performed in Python: first, all newly acquired spacer sequences were extracted from raw MiSeq FASTA data files. Next, the frequency, number of different barcodes, the phage target location, and the flanking PAM were determined for each unique spacer sequence. Analysis was finished in Excel.

On-plate spacer acquisition assay

To detect individual adapted colonies on a plate, cells from overnight cultures were mixed with phage at a m.o.i. value of 1 in top agar containing appropriate antibiotic and 5 mM CaCl2. The mixture was poured on BHI plates with antibiotic and incubated at 37 °C overnight. Subsequently, colonies that survived phage infection were re-streaked on fresh BHI plates in order to remove contaminating virus and dead cells. Plates were incubated at 37 °C overnight. To check for spacer acquisition, individual colonies were resuspended in lysis buffer (250 mM KCl, 5 mM MgCl2 50 mM Tris-HCl at pH 9.0, 0.5% Triton X-100), treated with 50 ng µl−1 lysostaphin and incubated at 37 °C for 5 min, then 98 °C for 5 min. Following centrifugation (16,000g), a sample of the supernatant was used as template for TopTaq PCR amplification with primers L400 and H050. The PCR reactions were analyzed on 2% agarose gels (Fig. 2-1a).

Spacer Acquisition Enrichment PCR

Overnight cultures launched from single colonies were diluted 1:1,000 into a fresh 10-ml culture of BHI containing appropriate antibiotic and 5 mM CaCl2. When the cultures reached D600 nm of 0.4, depending on the experiment, they were either infected with phage MOI value of 1 (Fig. 2-1b) or induced with 1 µg ml−1 anhydrotetracycline (Fig. 2-1c). After 16 h, plasmids carrying the CRISPR systems were extracted using a slightly modified QIAprep Spin Miniprep Kit protocol: the pelleted bacterial cells were resuspended in 250 µl buffer P1 containing 50 ng µl−1 lysostaphin and incubated at 37 °C for 1 h, followed by the standard QIAprep protocol. 100 ng of plasmid DNA was used to amplify the CRISPR locus using Phusion DNA Polymerase (New England Biolabs) with the following primer mix: 3 parts JW8 and 1 part each of JW3, JW4 and JW5 (Extended Data Table 4). The following cycling conditions were used: (1) 98 °C for 30 s; (2) (for 30 times) 98 °C for 10 s, 64 °C for 20 s, 72 °C for 10 s; (3) 72 °C for 5 min. The PCR reactions were analyzed on 2% agarose gels. To sequence individual spacers, the adapted bands were extracted, gel-purified and cloned via Zero Blunt TOPO PCR Cloning Kit (Invitrogen). CRISPR loci of individual clones were checked for expansion of the arrays by PCR using the primers listed above and sent for sequencing.

Phage adsorption assay

The phage adsorption assay was performed as described previously30 with minor modifications. Cells were grown in BHI and 10 mM CaCl2 to a D600 nm (OD600) of 0.4. The phage solution was prepared at 106 plaque-forming units (p.f.u.) per ml and 100 *μ*l of this was added to 900 *μ*l of cells. The mixture was incubated for 10 min at 37 °C to allow adsorption of the phage to the cellular membrane. The mixture was centrifuged for 1 min at 16,000g and the number of phage particles left in the supernatant was determined by phage titer assay.

Plasmid construction

Construction of pWJ40 was described elsewhere17. For the construction of pC194-derived and pE194-derived plasmids, cloning was performed using chemically competent S. aureus cells, as described previously17. The Δcas1 (pRH059), Δcas2 (pRH061) and Δcsn2 (pRH063) mutants were constructed by one-piece Gibson assembly31 from pWJ40 using the pairs of primers H016–H017, H018–H019, H020–H021, respectively (Extended Data Table 4). Plasmid pRH087 containing the wild type cas genes of S. pyogenes was obtained by inserting the first spacer of S. pyogenes (annealed primers H049 and H050 containing compatible BsaI overhangs) in pDB184 using BsaI cloning32. BsaI cloning was also used to construct pRH079 and pRH233 by inserting a ϕNM4γ4 targeting spacer (annealed primers H029 and H030) into pDB114 and pDB184, respectively. Plasmid pRH200 harbours the wild-type CRISPR3 system from S. thermophilus LMD-9 amplified with H168 and H169 from genomic DNA. The fragment was inserted on pE194 via Gibson assembly using H166 and H167. pRH213 was constructed by replacing Cas9Sp on pRH087 with Cas9St from pRH200 using the primer pairs H232–H233 and H231–H234, respectively. pRH214 was constructed by replacing Cas9St on pRH200 with Cas9Sp from pRH087 using the primer pairs H227–H230 and H228–H229, respectively. pGG32 was created by reducing the CRISPR locus of pWJ40 to a single repeat. This was accomplished by 'round the horn' PCR33 using primers oGG82 and oGG83, followed by blunt ligation. pRH228 was constructed by replacing Cas9Sp on pGG32 with Cas9St from pRH200 using the primer pairs H232–H233 and H231–H234, respectively. pRH223 was constructed as a three-piece Gibson assembly combining TetR+ptet from pKL55-iTet (primers B534 and B616), pE194 (primers B532 and B617) and the cas1, cas2, csn2 genes and the array from pGG32 (primers H176–H177). pRH231 was constructed from pGG32 by one-piece Gibson assembly with primers H289–H290. pRH234 contains Cas1 E220A and was constructed via one-piece Gibson assembly from pRH223, respectively, using the primer pair H312–H313. pRH227 was constructed from pGG32 via two sequential single-piece Gibson assemblies: first, D10A was introduced with B337–B338 and second, H840A was introduced with B339–

B340. pRH229 was constructed via one-piece Gibson assembly from pGG32 using the primer pair H276–H277. Plasmids pRH240, pRH241, pRH242, pRH243 and pRH244 were constructed by one-piece Gibson assembly with primers H237–H238 from pGG32, pRH228, pRH227, pRH229 and pRH231, respectively. pRH245 was constructed from pRH241 via two sequential single-piece Gibson assemblies: first, D10A was introduced with H336–H337 and second, H847A was introduced with H338–H339. Plasmid pRH317 was constructed by deleting the CRISPR leader and array from pRH223[84] via a one-piece Gibson assembly reaction with primer pair JM126-JM127. Plasmid pRH318 was constructed by a two-piece Gibson assembly reaction from pRH240[84] and pLZ12 with primer pairs H558-H559 and H555-H557, respectively. Plasmid pRH318* (containing randomized leader barcodes) was constructed by a two-piece Gibson assembly with primers pairs H378-H294 and H379-H293. Plasmid pRH248, pRH249, pRH328 and pRH328 were constructed BsaI cloning as described in Heler Nature with annealed oligo pairs H433-H434, H435-H436, H641-H642, and H643-H645, respectively.

Isolation and sequencing of ϕNM4γ4

For the initial isolation of ϕNM4, supernatants from overnight cultures of S. aureus Newman were filtered and used to infect soft agar lawns of TB4:: ϕNM1,2 double lysogens. A single plaque was picked and then plaque-purified in two additional rounds of infection using TB4 soft agar lawns, and subsequently used to lysogenize TB4. For the resultant lysogen, specific primers were used to verify the presence of ϕNM4 and the absence of ϕNM1,2 by colony PCR. High titer lysates of ϕNM4 (~1011 p.f.u. per ml) were then prepared from this lineage and used for infection of TB4/pGG9 soft agar lawns harboring spacer 2B17. An escaper plaque was picked and then plaque-purified in two additional rounds of infection using TB4/pGG9 soft agar lawns. The resultant ϕNM4γ4 phage exhibited a clear plaque phenotype and was used to prepare a high titre lysate from which DNA was purified, deep sequenced, and assembled as described previously. The full sequence of the ϕNM4γ4 has been deposited in GenBank under accession number KP209285 and includes a 2,784 bp deletion encompassing the C-terminal 80% of the ϕNM4 cI-like repressor gene.

Protein purification of Cas9

pMJ806 (wild-type Cas9) plasmid was obtained from Addgene. The proteins were purified as described before6 with minor modifications as follows. The proteins were expressed in E. coli BL21 Rosetta 2(DE3) codon plus cells (EMD Millipore). Cultures (2 litres) were grown at 37 °C in Terrific Broth medium containing 50 µg ml−1 kanamycin and 34 µg ml−1 chloramphenicol until the D600nm reached 0.6. The cultures were supplemented with 0.2 mM isopropyl-1-thio-β-D-galactopyranoside and incubation was continued for 16 h at 16 °C with constant shaking. The cells were collected by centrifugation and the pellets stored at −80 °C. All subsequent steps were performed at 4 °C. Thawed bacteria were resuspended in 30 ml of buffer A (50 mM Tris–HCl pH 7.5, 500 mM NaCl, 200 mM Li2SO4, 10% sucrose, 15 mM imidazole) supplemented with complete EDTA free protease inhibitor tablet (Roche). Triton X-100 and lysozyme were added to final concentrations of 0.1% and 0.1 mg ml–1, respectively. After 30 min, the lysate was sonicated to reduce viscosity. Insoluble material was removed by centrifugation for 1 h at 16,200g in a Beckman JA-3050 rotor. The soluble extract was bound in batch to mixed for 1 h with 5 ml of Ni2+-Nitrilotriacetic acid-agarose resin (Qiagen) that had been pre-equilibrated with buffer A. The resin was recovered by centrifugation, and then washed extensively with buffer A. The bound protein was eluted step-wise with aliquots of IMAC buffer (50 mM Tris-HCl pH 7.5, 250 mM NaCl, 10% glycerol) containing increasing concentrations of imidazole. The 200 mM imidazole elutes containing the His6-MBP tagged Cas9 polypeptide was pooled together. The His6-MBP affinity tag was removed by cleavage with TEV protease during overnight dialysis against 20 mM Tris-HCl pH 7.5, 150 mM KCl, 1 mM TCEP and 10% glycerol. The tagless Cas9 protein was separated from the fusion tag by using a 5 ml SP Sepharose HiTrap column (GE Life Sciences). The protein was further purified by size exclusion chromatography using a Superdex 200 10/300 GL in 20 mM Tris HCl pH 7.5, 150 mM KCl, 1 mM TCEP, and 5% glycerol. The elution peak from the size exclusion was aliquoted, frozen and kept at –80 °C.

Protein purification of Cas1

Plasmid pKW01 (wild-type Cas1) was constructed by through amplification of pWJ40 as a template for polymerase chain reactions (PCRs) to clone Cas1 into pET28b-His10Smt3 using the primers PS192 and PS193 (Extended Data Table 4). Full sequencing of cloned DNA fragment confirmed perfect matches to the original sequence. The pKW01 plasmid was transformed into E. coli BL21 (DE3) Rosetta 2 cells (EMD Millipore). Cultures were grown and protein was purified by Ni-affinity chromatography step, as mentioned before in Cas9 purification. The 200 mM imidazole elutes containing the His10-Smt3 tagged Cas1 polypeptide was pooled together. The His10-Smt3 affinity tag was removed by cleavage with SUMO protease during overnight dialysis against 50 mM Tris-HCl pH 7.5, 250 mM NaCl, 20 mM imidazole and 10% glycerol. The tagless Cas1 protein was separated from the fusion tag by using a second Ni-NTA affinity step. The protein was further purified by size exclusion chromatography using a Superdex 200 10/300 GL in 20 mM Tris HCl pH 7.5, 500 mM KCl, 1 mM TCEP, and 5% glycerol. The elution peak from the size exclusion was aliquoted, frozen and kept at –80 °C.

Protein purification of Cas2

The sequence encoding Cas2 was PCR amplified with primers PS334 and PS335 from pWJ40 and inserted into a pET-His6 MBP TEV cloning vector (Addgene Plasmid number 29656) using ligation independent cloning (LIC). Sequencing of the resultant plasmid (pPS059) confirmed the matches to the wild-type sequence. The protein was expressed and purified following the same procedure as that for Cas9.

Protein purification of Csn2

Plasmid pPS060 was constructed by through amplification of pWJ40 as a template for polymerase chain reactions (PCRs) to clone Csn2 into pET28b-His10Smt3 using the primers PS336 and PS337. Full sequencing of cloned DNA fragment confirmed perfect matches to the original sequence. Csn2 was expressed and purified following the same method as that of Cas1. Previously Csn2 was shown to form a tetramer34. Protein concentrations for all the purifications were determined by using the Bradford dye reagent with BSA as the standard.

Protein purification of Cas9–Cas1–Cas2–Csn2 complex

pKW07 (His10-Cas9–Cas1–Cas2–Csn2) was constructed by amplification of pWJ40 with primers PS199/PS202 and pET16b (Novagen) with primers PS200/PS203, followed by Gibson assembly of the fragments. Full sequencing of cloned DNA fragment was done to confirm perfect matches to the original sequence. The proteins were expressed in E. coli BL21 Rosetta 2(DE3) codon plus cells (EMD Millipore). Cultures were grown and protein was purified by Ni-affinity chromatography step, as mentioned before in Cas9 purification with minor modifications. The 200 mM imidazole eluates were dialysed overnight against 20 mM Tris-HCl pH 7.5, 150 mM KCl, 1 mM TCEP and 10% glycerol and subjected to mass spectrometry for the identification of the co-purifying proteins. pKW06 (Cas9–Cas1–Cas2–Csn2–His6) was constructed by amplification of pWJ40 with primers PS204/PS205 and pET23a (Novagen) with primers PS206/PS207 (Extended Data Table 4), followed by Gibson assembly of the fragments. Full sequencing of cloned DNA fragment was done to confirm perfect matches to the original sequence. The proteins were expressed in E. coli BL21 Rosetta 2(DE3) codon plus cells (EMD Millipore). Cultures were grown and protein was purified by Ni-affinity chromatography step, as mentioned before in Cas9 purification with minor modifications. The 200 mM imidazole eluates were dialysed overnight against 20 mM Tris-HCl pH 7.5, 150 mM KCl, 1 mM TCEP and 10% glycerol. The proteins were further purified using a 5 ml SP Sepharose HiTrap column (GE Life Sciences), eluting with a linear gradient of 150 mM–1 M KCl.

Oligonucleotides Used

Name	Sequence
B337	GACGCTATTTGTGCCGATAGCTAAGCCTATTGAGTATTTC
B338	GAAATACTCAATAGGCTTAGCTATCGGCACAAATAGCGTC
B339	GGAAACTTTGTGGAACAATGGCATCGACATCATAATCACT
B340	AGTGATTATGATGTCGATGCCATTGTTCCACAAAGTTTCC
B532	CTTTTTCCGTGATGGTAACTGTTCATATTTATCAGAGCTCGTG
B534	GAGCTCTGATAAATATGAACAGTTACCATCACGGAAAAAGGTTATG
B616	TTATTTTAATTATGCTCTATCAA
B617	GAGTGATCGTTAAATTTATACTGC
H001	GGGCACTTTTTCACTCATTTTAGCTTCCTTAGCTCCTGAAAATC
H002	GGTGCCAGCCAATGATTTTTTTAAGGCAGTTATTGG
H003	GCTAAGGAAGCTAAAATGAGTGAAAAAGTGCCCGCC
H004	ACTGCCTTAAAAAAATCATTGGCTGGCACCAAGCAG

H005	GCTAAGGAAGCTAAAATGATTGAACAAGATGGATTGCAC
H006	ACTGCCTTAAAAAAATCAGAAGAACTCGTCAAGAAGGCG
H007	GACGAGTTCTTCTGATTTTTTTAAGGCAGTTATTGGTGC
H008	ATCTTGTTCAATCATTTTAGCTTCCTTAGCTCCTG
H009	TCCATCTTGTTCAATCATTTTAGCTTCCTTAGCTCCTGAAAATC
H010	GAGAAAGAGGGTTAATGGAAGCCGGCGGCACCTCGCTAAC
H011	GTGCCGCCGGCTTCCATTAACCCTCTTTCTCAAGTTATCA
H012	GCTATATGCGTTGAACCGGAATTGCCAGCTGGGGCGCCCT
H013	GGTGCCGCCGGCTTCCATTCAGAAGAACTCGTCAAGAAGGCG
H014	ACGAGTTCTTCTGAATGGAAGCCGGCGGCACCTCGCTAAC
H015	CCAGCTGGCAATTCCGGTTCAACGCATATAGCGCTAGCAG
H016	AGGAGGTGACTGATGGGAGTTCCTGAATTTAGGATATGAG
H017	TAAATTCAGGAACTCCCATCAGTCACCTCCTAGCTGACTC
H018	TTAGGATATGAGTGAGGCTTTTGATGAATCTTAATTTTTC
H019	TTCATCAAAAGCCTCACTCATATCCTAAATTCAGGAACTC
H020	TTTGATGAATCTTAATAAAAATATGGTATAATACTCTTAA
H021	TTATACCATATTTTTATTAAGATTCATCAAAAGCCTCCCC
H022	AAACACGAATATACAGGAAGAATACACGATGTTGG
H023	AAAACCAACATCGTGTATTCTTCCTGTATATTCGT
H024	AAACAAAAACAAAAATGTTTTAACACCTATTAACGG
H025	AAAACCGTTAATAGGTGTTAAAACATTTTTGTTTTT
H026	TGACGAGTTCTTCTGATTTTTTTAAGGCAGTTATTGGTGCCC
H027	TGACGAGTTCTTCTGATTTTTTTAAGGCAGTTATTGGTGCCCTTA
H029	AAACAAAAATGTTTTAACACCTATTAACGTAGTATG
H030	AAAACATACTACGTTAATAGGTGTTAAAACATTTTT
H031	GAACTTTGAAATCGGCTCAGGAAAAGGCCATTTTACCCTT
H032	TTTAAAGGGTAAAATGGCCTTTTCCTGAGCCGATTTCAAA
H033	GAACTTTGAGATCGGTTCTGGTAAGGGCCACTTCACTCTC
H034	TTTAGAGAGTGAAGTGGCCCTTACCAGAACCGATCTCAAA
H035	GAACATATCACACAAAGATAAACAAAAGTATAATTATTTC
H036	TTTAGAAATAATTATACTTTTGTTTATCTTTGTGTGATAT
H037	GAACATATCGCACAAGGACAAGCAGAAGTACAACTACTTT
H038	TTTAAAAGTAGTTGTACTTCTGCTTGTCCTTGTGCGATAT
H039	AAACCCCAGTCGACACCAGCAAAGTATTCTTTGATG
H040	AAAACATCAAAGAATACTTTGCTGGTGTCGACTGGG
H041	AAACCCATTGCACCTCAAGTATCGATGACTGATTCG
H042	AAAACGAATCAGTCATCGATACTTGAGGTGCAATGG
H043	AAACAAAAACGTTTTGACGCCCATCAACGTCGTGTG
H044	AAAACACACGACGTTGATGGGCGTCAAAACGTTTTT
H045	AAACAAGAACGTTTTGACCCCGATCAATGTCGTATG

H046	AAAACATACGACATTGATCGGGGTCAAAACGTTCTT
H047	AAACAAGAACGTGTTGACCCCGATCAATGTCGTCTG
H048	AAAACAGACGACATTGATCGGGGTCAACACGTTCTT
H049	AAACTGCGCTGGTTGATTTCTTCTTGCGCTTTTTG
H050	AAAACAAAAAGCGCAAGAAGAAATCAACCAGCGCA
H051	AAACTTATATGAACATAACTCAATTTGTAAAAAAG
H052	AAAACTTTTTTACAAATTGAGTTATGTTCATATAA
H053	AAACAGGAATATCCGCAATAATTAATTGCGCTCTG
H054	AAAACAGAGCGCAATTAATTATTGCGGATATTCCT
H055	AAACAGTGCCGAGGAAAAATTAGGTGCGCTTGGCG
H056	AAAACGCCAAGCGCACCTAATTTTTCCTCGGCACT
H057	AAACTAAATTTGTTTAGCAGGTAAACCGTGCTTTG
H058	AAAACAAAGCACGGTTTACCTGCTAAACAAATTTA
H059	AAACTTCAGCACACTGAGACTTGTTGAGTTCCATG
H060	AAAACATGGAACTCAACAAGTCTCAGTGTGCTGAA
H061	TTTTAGGAGGCAAAAATGGATAAGAAATACTCAATAGGCT
H062	CATCTAAAATATACTTCAGTCACCTCCTAGCTGACTCAAA
H063	CTAGGAGGTGACTGAAGTATATTTTAGATGAAGATTATTT
H064	GTATTTCTTATCCATTTTTGCCTCCTAAAATAAAAAGTTT
H065	GAT ATA ATG GGA GAT AAG ACG GTT C
H066	GGG ACC TCT TTA GCT CCT TG
H067	AAACAAATGTTTTAACACCTATTAACGTAGTATTGG
H068	AAAACCAATACTACGTTAATAGGTGTTAAAACATTT
H069	AAACAGATAAAAACAAAAATGTTTTAACACCTATTG
H070	AAAACAATAGGTGTTAAAACATTTTTGTTTTTATCT
H073	AAACAACAAAAATGTTTTAACACCTATTAACGTAGG
H074	AAAACCTACGTTAATAGGTGTTAAAACATTTTTGTT
H075	AAACTATTAACGTAGTATTGGAATCTGATGAATATG
H076	AAAACATATTCATCAGATTCCAATACTACGTTAATA
H077	AAACTATTTTTAGATAAAAACAAAAATGTTTTAACG
H078	AAAACGTTAAAACATTTTTGTTTTTATCTAAAAATA
H079	AAACGATAAAAACAAAAATGTTTTAACACCTATTAG
H080	AAAACTAATAGGTGTTAAAACATTTTTGTTTTTATC
H081	AAACTGTTTTAACACCTATTAACGTAGTATTGGAAG
H082	AAAACTTCCAATACTACGTTAATAGGTGTTAAAACA
H083	AAACAAAATGTTTTAACACCTATTAACGTAGTATTG
H084	AAAACAATACTACGTTAATAGGTGTTAAAACATTTT
H085	AAACTCATCTCTCGGTATATATAATCCAAGTTATTG
H086	AAAACAATAACTTGGATTATATATACCGAGAGATGA
H089	AAACAAAACAAAAATGTTTTAACACCTATTAACGTG

H090	AAAACACGTTAATAGGTGTTAAAACATTTTTGTTTT
H091	AAACAAACAAAAATGTTTTAACACCTATTAACGTAG
H092	AAAACTACGTTAATAGGTGTTAAAACATTTTTGTTT
H093	AAACATGTTTTAACACCTATTAACGTAGTATTGGAG
H094	AAAACTCCAATACTACGTTAATAGGTGTTAAAACAT
H095	AAACCAAAAATGTTTTAACACCTATTAACGTAGTAG
H096	AAAACTACTACGTTAATAGGTGTTAAAACATTTTTG
H097	AAACACAAAAATGTTTTAACACCTATTAACGTAGTG
H098	AAAACACTACGTTAATAGGTGTTAAAACATTTTTGT
H099	TCTATTTATTATTAATTATTGGGTAATATTTTTTGAAGAG
H100	AAATATTACCCAATAATTAATAATAAATAGATTATAACAC
H101	GCTATTTTGAGAGGACAAGAAGACTTTTATCC
H102	GGATAAAAGTCTTCTTGTCCTCTCAAAATAGC
H103	GGAAGTCTGAAGAAACATTTACCCCATGG
H104	CCATGGGGTAAATGTTTCTTCAGACTTCC
H105	GACAAACTTTGATATAAATCTTCCAAATGAAAAAGTACTACC
H106	GGTAGTACTTTTTCATTTGGAAGATTTATATCAAAGTTTGTC
H107	CCATGATGATGGTTTGACATTTAAAGAAGAC
H108	GTCTTCTTTAAATGTCAAACCATCATCATGG
H109	GGGCGGCATAAGCTAGAAAATATCG
H110	CGATATTTTCTAGCTTATGCCGCCC
H111	GCAAGAAATAGGCAAAGGAACCGC
H112	GCGGTTCCTTTGCCTATTTCTTGC
H113	AAACTTTAGCGATATTAATTATGCTCGTAAGAATG
H114	AAAACATTCTTACGAGCATAATTAATATCGCTAAA
H115	AAACTTTATTTTGCGTTAGAATTGACACCTCAAGAG
H116	AAAACTCTTGAGGTGTCAATTCTAACGCAAAATAAA
H117	AAACCTTTAAATGTTTTAAAAGAATAGCATCATTG
H118	AAAACAATGATGCTATTCTTTTAAAACATTTAAAG
H119	AAACACAGGAATTGAGACACCTCAATATATACTTGCG
H120	AAAACGCAAGTATATATTGAGGTGTCTCAATTCCTGT
H121	AAACAAAATGCAAGAATTAAACTACCCACCATATG
H122	AAAACATATGGTGGGTAGTTTAATTCTTGCATTTT
H123	AAACCTAAGATAGCTAAAGCAATACGTGATGATGTG
H124	AAAACACATCATCACGTATTGCTTTAGCTATCTTAG
H125	AAACATTTATATCCGATCTTATACGAAGTAAAGAG
H126	AAAACTCTTTACTTCGTATAAGATCGGATATAAAT
H127	TTTATCCATAAATTCGTTAAAGTCTTTACG
H128	TTATTTTGAGGATTTATAATGATGCTAGAG
H129	ATGAGTTATAGATATATGAGAATGATACTTATGTTTGATATGC

H130	ATTTGAGTCAGCTAGGAGGTGACTGATGATAGAGCTATCTAAATACAATATTTTAGTG
H131	GTATCATTCTCATATATCTATAACTCATGATTTATAAAATGAATATTGCTTAATATTTGG
H132	AAACAGGAATTGAGACACCTCAATATATACTTGCG
H133	AAAACGCAAGTATATATTGAGGTGTCTCAATTCCT
H134	AAACGTGCGAAAGATAGCAGACGAAGAAGGAATTG
H135	AAAACAATTCCTTCTTCGTCTGCTATCTTTCGCAC
H136	TTTGAGTCAGCTAGGAGGTGACTGATGAAGAGTAAAAAGCATCCTCAAATC
H137	TGTATTACTGCATTTATTAAGAGTACTCTAGCATCATTATAAATCCTCAAAATAATTAAG
H138	GAGTACTCTTAATAAATGCAGTAATACAGGGG
H139	TCAGTCACCTCCTAGCTGACTC
H140	TAACAACTACTATAACCTCTAGGCTTATGCCACTCTTATCCATCAATC
H141	AGCATCATTATAAATCCTCAAAATAACTCGTAGACTATTTTTGTCTAAAAAATTTTG
H142	TTATTTTGAGGATTTATAATGATGCTAGAGG
H143	AAGCCTAGAGGTTATAGTAGTTGTTAAAT
H144	GAACACTTTTGCGCTGGTTGATTTCTTCTTGCGCTTTTT
H145	TTTAAAAAAGCGCAAGAAGAAATCAACCAGCGCAAAAGT
H146	GAGCAAGTTAACATTAAATTAGATAAAACT
H147	AGTTTTATCTAATTTAATGTTAACTTGCTC
H148	AATATTTGGCGTAGTATGAAAGATTTAATT
H149	AATTAAATCTTTCATACTACGCCAAATATT
H150	GAACATAGGTAGCCTTTATACGGTCCATAAACATGGGGAT
H151	TTTAATCCCCATGTTTATGGACCGTATAAAGGCTACCTAT
H152	GGAAGAAGACAAGAACCATGAACGTCATC
H153	GATGACGTTCATGGTTCTTGTCTTCTTCC
H154	GATGAAGTTGCTTATCGTGAGAAATATCC
H155	GGATATTTCTCACGATAAGCAACTTCATC
H156	CTTAGCGCATATGTTTAAGTTTCGTG
H157	CACGAAACTTAAACATATGCGCTAAG
H158	CACAAGTGTTTGGACAAGGCGATAG
H159	CTATCGCCTTGTCCAAACACTTGTG
H160	GTTGTCGATAATGGTGCTTCAGCTC
H161	GAGCTGAAGCACCATTATCGACAAC
H162	GTGATGAAACAGTTTAAACGTCGCC
H163	GGCGACGTTTAAACTGTTTCATCAC
H164	GCCAAGTTAATCACTAAACGTAAGTTTG
H165	CAAACTTACGTTTAGTGATTAACTTGGC

H166	GAAATGTGAGAAGGGACCTCTGATAAATATGAACATGATGAGTGATCG
H167	GGACTCTTTTATCTCTACTCGTGCTATAATTATACTAATTTTATAAGGAGG
H168	AGTATAATTATAGCACGAGTAGAGATAAAAGAGTCCTTTGGATGATTCC
H169	TGTTCATATTTATCAGAGGTCCCTTCTCACATTTCAATACTAGACTC
H170	GCTAGTATTTTGTCAACGAATAATAAGAGG
H171	TTTTCTGTGATGATAAACGATTGCC
H172	GCGTTAAATCAGTTAGGTGAGG
H173	ATTAATTACTGATATTATAATGGCAGAGTG
H174	TATCGGCACAAATAGCGATGCCACTCTTATCCATCAATCC
H175	GGATAAGAGTGGCATCGCTATTTGTGCCGATATCTAAGCC
H176	TTGATAGAGCATAATTAAAATAAGATGCCACTCTTATCCATCAATCC
H177	GCAGTATAAATTTAACGATCACTCTAAAACCTCTCCAACTACCTCCC
H178	CCAATTTTCGTTTGATGTCTAAAAAATTTCGTAATCGCAC
H179	GAAATTTTTTAGACATCAAACGAAAATTGGATAAAGTGGG
H180	TCTGGTAGAAAAGATATCCTACGAG
H181	GAGCTTCCGAGACTGGTCTC
H182	NNNNNCAGCAAAATTTTTTAGACAAAAATAGTC
H183	NNNNNCAGAAGAAGAAATCAACCAGCGC
H184	NNNNNTCACAAAATTTTTTAGACAAAAATAGTC
H185	NNNNNTCAAAGAAGAAATCAACCAGCGC
H186	NNNNNGTCCAAAATTTTTTAGACAAAAATAGTC
H187	NNNNNGTCAAGAAGAAATCAACCAGCGC
H188	NNNNNAGTCAAAATTTTTTAGACAAAAATAGTC
H189	NNNNNAGTTAACCCTCTTTCTCAAGTTATC
H190	CCCCAGCGAATTTTGAAGAAGTTGTCGATAAAGGTGC
H191	CGACAACTTCTTCAAAATTCGCTGGGGTAATTGTTTCTTCAG
H192	ATTGCTCGTAAAAAAGACGCGGATCCAAAAAAATATGG
H193	CCACCATATTTTTTTGGATCCGCGTCTTTTTTACGAGC
H194	GAAGTCTGAAGAAACAATTACCGCAGCGGCTTTTGAAGAAGTTGTCG
H195	TCGACAACTTCTTCAAAAGCCGCTGCGGTAATTGTTTCTTCAGACTTCC
H196	AAGCTTATTGCTCGTAAAAAAGCCGCGGCTCCAAAAAAATATGGTGG
H197	CCACCATATTTTTTTGGAGCCGCGGCTTTTTTACGAGCAATAAGC
H198	TGAAAAAATCTTGACTTTTCGAATTCC
H199	AATACTCATAAAGCAAACTATGTTTTGG
H200	GGAATTCGAAAAGTCAAGATTTTTTCAATCTTCTCACG
H201	CCAAAACATAGTTTGCTTTATGAGTATTTTACGG
H202	CAATATTGTCAAGAAAACAGAAGTACAGAC
H203	TTAGCAACCACTAGGACTGAATAAGC
H204	GCCTGTCTGTACTTCTGTTTTCTTGAC

H205	GAAGTCTGAAGAAACATTTACCGCAGCGGCTTTTGAAGAAGTTGTCG
H206	ATCGACAACTTCTTCAAAAGCCGCTGCGGTAAATGTTTCTTCAGACTTCC
H207	GGAAGTCTGAAGAAACAGCTACCCCATGG
H208	CCATGGGGTAGCTGTTTCTTCAGACTTCC
H209	AATCACGGATTGGATAGAGGAAAACC
H210	TTAACCCTCTCCTAGTTTGGCAAGG
H211	CTTGCCAAACTAGGAGAGGGTTAATCCATCACTGGTCTTTATGAAACACG
H212	GTTTTCCTCTATCCAATCCGTGATTTGTTTTCTTGACAATATTGACTTGG
H213	AAGTACAGACAGGCGGATTCTCC
H214	CAGTCACCTCCTAGCTGACTC
H215	TTGATTTGAGTCAGCTAGGAGGTGACTGACAATCTGTTACAGGCCTC
H216	CTTGGAGAATCCGCCTGTCTGTACTTCCTGTTCCTCAACTTTTTTCACAAC
H221	GATAAAGGTGCTTCAGCTCAATC
H222	AGACTTCCGAGTCATCCATGC
H223	GGTTTTGATAGTCCAACGGTAGC
H224	AAGCTTGTCCGAATTTCTTTTTGG
H225	GCCGCGGCTCCAAAAAAATATGGTGGTTTTGATAGTCC
H226	GCAGCGGCTTTTGAAGAAGTTGTCGATAAAGGTGC
H227	TAATGGCAGGTTGGAGAACAGTAGTC
H228	ACTACTGTTCTCCAACCTGCCATTAGTCACCTCCTAGCTGACTC
H229	AGATTTTTCAAATAAGGAGAAATGTTTGAAATCATCAAACTCATTATGGATTTAATTTAAACTTTTTATTTTAGG
H230	ACATTTCTCCTTATTTGAAAAATCTAAATTTATAGAAATTATTATACGC
H231	AACTTTTTATTTTAGGAGGCAAAAAGCGTATAATAATTTCTATAAATTTAGATTTTTCAAATAAGG
H232	TTTTGCCTCCTAAAATAAAAAGTTTAAATTAAATCCATAATGAG
H233	TGATGGCTGGTTGGCGTAC
H234	CAACAGTACGCCAACCAGCCATCAACCCTCTCCTAGTTTGGC
H235	GATATCGGCACAAATAGCTTAGATGCCACTCTTATCCATCAATCC
H236	AAGAGTGGCATCTAAGCTATTTGTGCCGATATCTAAGCC
H237	GGCGTACTGATGAAGATTATTTCTTAATAACTAAAAATATGG
H238	TTTAGTTATTAAGAAATAATCTTCATCAGTACGCCAACCAGCC
H239	TCAATTGGACTTGATATTATAGACCTTGCCAAACTAGGAG
H240	TTTGGCAAGGTCTATAATATCAAGTCCAATTGAGTATGGC
H241	AACAGTAGTCATTTTAGACAAGGATTATATTTTGATGCCC
H242	ATAATCCTTGTCTAAAATGACTACTGTTCTCCAACCTGCC
H243	NNNNNTCGCAAAATTTTTTAGACAAAAATAGTC
H244	NNNNNTCGAAGAAGAAATCAACCAGCGC

H245	NNNNNAGCCAAAATTTTTTAGACAAAAATAGTC
H246	NNNNNAGCAAGAAGAAATCAACCAGCGC
H247	NNNNNCATCAAAATTTTTTAGACAAAAATAGTC
H248	NNNNNCATAAGAAGAAATCAACCAGCGC
H249	NNNNNGACCAAAATTTTTTAGACAAAAATAGTC
H250	NNNNNGACAAGAAGAAATCAACCAGCGC
H251	NNNNNACTCAAAATTTTTTAGACAAAAATAGTC
H252	NNNNNACTAAGAAGAAATCAACCAGCGC
H253	NNNNNCTGCAAAATTTTTTAGACAAAAATAGTC
H254	NNNNNCTGAAGAAGAAATCAACCAGCGC
H255	NNNNNTGACAAAATTTTTTAGACAAAAATAGTC
H256	NNNNNTGATAACCCTCTTTCTCAAGTTATC
H257	AAGCTTTATATAACTCTCTTGCA
H258	ATGGTAAAGCTTTTGTAAAAACCTG
H259	GCTCTAGAAGCTTCAAAGTTTTAC
H260	GCGAAAAGATAAACGAAAGCTTG
H261	CTTAGAAGCTTGTACTAAGCCG
H262	CTTCGACAGTAGCTTTAGTTGC
H263	CAGAAAAACAATAACAGAAGCTTGGAA
H264	CTTGTTGTTTAGTAAAAGCTTGAG
H265	NNNNNGGGCAAAATTTTTTAGACAAAAATAGTC
H266	NNNNNGGGAAGAAGAAATCAACCAGCGC
H267	GGAATTATTTTGAAGCTGAAGTCATG
H268	AAACAACCAAAAAAGGGAAGGGCTCGGTTGTACAG
H269	AAAACTGTACAACCGAGCCCTTCCCTTTTTTGGTT
H270	AAACAACCAAAAAAGGGAAGGGCTCGGTTGTATCG
H271	AAAACGATACAACCGAGCCCTTCCCTTTTTTGGTT
H272	CAATTGATCAAAAACGATATACGTCTAC
H273	ATATCGTTTTTGATCAATTGTTGTATC
H274	ATCGTAAACAATATACGTCTACAAAAGAAG
H275	TAGACGTATATTGTTTACGATCAATTGTTG
H276	TTGATCAAAAACAATATACGTCTACAAAAGAAG
H277	TAGACGTATATTGTTTTTGATCAATTGTTGTATCAA
H278	ORDERED. CHECK SEQUENCE
H279	GATGCCACTCTTATCCATCAATCC
H280	TTTTTATTTTAATTATGCTCTATCAATGATAGAGTGTC
H281	GATAGAGCATAATTAAAATAAAAAGCATATTAAACTAATTTCGG
H282	TGGATTGATGGATAAGAGTGGCATCTAAAACTTCTTTTGTAGACG
H283	TGCGACTACAAAATTTTTTAGACAAAAATAGTCTACGAGG
H284	TTTGTCTAAAAAATTTTGTAGTCGCACTATTTGTCTCAGC

H285	TGCGAGTACAAAATTTTTTAGACAAAAATAGTCTACGAGG
H286	TTTGTCTAAAAAATTTTGTACTCGCACTATTTGTCTCAGC
H287	GTACAATTCTTGTGTTGCTTATTTTTGTCAATAGCGGAGC
H288	CAAAAATAAGCAACACAAGAATTGTACCGCCTCTTAATGG
H289	AGCGCTTGGGAGAAATTCAAAGAAATTTATCAGCC
H290	TTTCTTTGAATTTCTCCCAAGCGCTTTCAAAACGC
H291	TCCAAGTTATTTGCATGCTCC
H292	AAAGGTGGTGAAAAGAAATGCC
H293	GCAAAAATGGATAAGAAATACTCAATAGGC
H294	TATTGAGTATTTCTTATCCATTTTTGCCTCC
H295	AACACGCATTGATTTGAGTCAGC
H296	TCCTAGCTGACTCAAATCAATGCG
H297	AAACCGAATAACTCACGTTCCATTGAATACTGTGTG
H298	AAAACACACAGTATTCAATGGAACGTGAGTTATTCG
H299	AAACACGTTCCATTGAATACTGTGTAGGCATGTTAG
H300	AAAACTAACATGCCTACACAGTATTCAATGGAACGT
H302	TTGGAGCTCCCGCTGC
H303	GGATTGATGGATAAGAGTGGC
H304	ATATTTAAAAGCAGCGGGAGC
H305	GTTTTAGATGCCACTCTTATCC
H306	CGTTTTGCATGGATGACTCG
H307	TGATTGAGCTGAAGCACC
H308	TGCGATCACAAAATTTTTTAGACAAAAATAGTCTACGAGG
H309	TTTGTCTAAAAAATTTTGTGATCGCACTATTTGTCTCAGC
H310	TGCGATGACAAAATTTTTTAGACAAAAATAGTCTACGAGG
H311	TTTGTCTAAAAAATTTTGTCATCGCACTATTTGTCTCAGC
H312	GATATTATGGCACCATTTAGGCCTTTAGTGG
H313	AAAGGCCTAAATGGTGCCATAATATCGCTAGC
H314	TATATCATGGTTTGGCAAATTTTGATCCGAG
H315	ATCAAAATTTGCCAAACCATGATATAAATCC
H316	TTTTGCTAGCGCTATTATGGAACCATTTAGGCC
H317	TGGTTCCATAATAGCGCTAGCAAAATTGAACTG
H318	CGGACACCGCTGAGGCACGAAAAGCCTATCG
H319	GCTTTTCGTGCCTCAGCGGTGTCCGTCGGC
H320	GAGGAAGCAAAAGCCTATCGAAAATTTCGG
H321	ATTTTCGATAGGCTTTTGCTTCCTCAGCGGTG
H322	GAGGAACGAAAAGCCTATGCAAAATTTCGG
H323	ATTTTGCATAGGCTTTTCGTTCCTCAGCGG
H324	ATCCTGGCATTGATTAAGTCCTTAGGAG
H325	AAGGACTTAATCAATGCCAGGATTGTG

H326	TTAGGAGTAAAAGTAGCAACGCAAAGTG
H327	TTGCGTTGCTACTTTTACTCCTAAGGAC
H328	TTAGGAGTAGAAGTAGAAACGCAAAGTG
H329	TTGCGTTTCTACTTCTACTCCTAAGGAC
H330	AATTTAATGAGGAACCCGAAGTGAAATCG
H331	TTTCACTTCGGGTTCCTCATTAAATTGAG
H332	AAACTTTTAAGCTATTCATTTTAAAAGGTCATATG
H333	AAAACATATGACCTTTTAAAATGAATAGCTTAAAA
H334	AAACATTTTAAGCTATTCATTTTAAAAGGTCATAG
H335	AAAACTATGACCTTTTAAAATGAATAGCTTAAAAT
H336	CATACTCAATTGGACTTGCTATTGGAACGAATAGTGTTGG
H337	CGTTCCAATAGCAAGTCCAATTGAGTATGGCTTAGTC
H338	GTAATTATGATATTGATGCTATTATTCCTCAAGC
H339	GAGGAATAATAGCATCAATATCATAATTACTTAATC
H340	AAACTGCCTATTTTTTTATGTTATAGCTAGCCTTG
H341	AAAACAAGGCTAGCTATAACATAAAAAAATAGGCA
H342	AAACAATTCCTTGAATCGAAAGGAGGTTAGCCTTG
H343	AAAACAAGGCTAACCTCCTTTCGATTCAAGGAATT
H344	AAACCGTGTAAAGACATATTAGATCGAGTCAAGGG
H345	AAAACCCTTGACTCGATCTAATATGTCTTTACACG
H346	AAACATACGTGTAAAGACATATTAGATCGAGTCAG
H347	AAAACTGACTCGATCTAATATGTCTTTACACGTAT
H348	AAACAAGACATATTAGATCGAGTCAAGGAGGTTTG
H349	AAAACAAACCTCCTTGACTCGATCTAATATGTCTT
H350	AAACAGACATATTAGATCGAGTCAAGGAGGTTTTG
H351	AAAACAAAACCTCCTTGACTCGATCTAATATGTCT
H352	AAACGACATATTAGATCGAGTCAAGGAGGTTTTGG
H353	AAAACCAAAACCTCCTTGACTCGATCTAATATGTC
H354	CCCCAACAAGAGAATTGGC
H355	CACCACCATAAACAACACAAGG
H356	CAAATGGGACATCTCGATGG
H357	GTCAGGACCTTTAGGTCATAGC
H358	NNNNNCAGTAGCTGAGACAAATAGTGCG
H359	NNNNNCAGCTCAACAAGTCTCAGTGTGC
H360	NNNNNGCTTAGCTGAGACAAATAGTGCG
H361	NNNNNGCTCTCAACAAGTCTCAGTGTGC
H362	NNNNNCAGAAAACAGCATAGCTCTAAAACG
H363	NNNNNCAGAAAACAGCATAGCTCTAAAACA
H364	NNNNNCAGAAAACAGCATAGCTCTAAAACT
H365	NNNNNCAGGGCTTTTCAAGACTGAAGTCTAG

H366	NNNNNGCTAAAACAGCATAGCTCTAAAACG
H367	NNNNNGCTAAAACAGCATAGCTCTAAAACA
H368	NNNNNGCTAAAACAGCATAGCTCTAAAACT
H369	NNNNNGCTGGCTTTTCAAGACTGAAGTCTAG
H370	NNNNNGACAGGGGCTTTTCAAGACTG
H371	NNNNNGACGAAGAAATCAACCAGCGC
H372	NNNNNACTAGGGGCTTTTCAAGACTG
H373	NNNNNACTGAAGAAATCAACCAGCGC
H374	NNNNNCTGAGGGGCTTTTCAAGACTG
H375	NNNNNCTGGAAGAAATCAACCAGCGC
H376	NNNNNTGAAGGGGCTTTTCAAGACTG
H377	NNNNNTGAGAAGAAATCAACCAGCGC
H378	CAGGGGCTTTTCAAGACTGNNNNNNNNNNGAGACAAATAGTGCG
H379	CAGTCTTGAAAAGCCCCTG
H380	TTCAAGGTAAGTTTTGTCGTATCGTTCAATTTTATTCCGATCAGGCAATAGTTGAACTTT
H381	TGAAAAAGTTCAACTATTGCCTGATCGGAATAAAATTGAACGATACGACAAAACTTACCT
H382	TAGAATATGAGTTATAGATATATGAGAATGATACTTATGTTTGATATGCC
H383	CATTCTCATATATCTATAACTCATATTCTAAATTCAGGAATTTCCTCACC
H384	AAATGTAGAATGATAAAATAGAGATAAAAGAGTCCTTTGG
H385	GACTCTTTTATCTCTATTTTATCATTCTACATTTAGGCGC
H386	TTCTATAAATTTAGATTTTAGTATTGGGTAATATTTTTTGAAGAG
H387	TATTACCCAATACTAAAATCTAAATTTATAGAAATTATTATACGC
H388	TTCAAGGTAAGTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC
H389	TGAAGTTTTGGGACCATTCAAAACAGCATAGCTCTAAAACTTACCT
H390	TTCATCCATTGTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC
H391	TGAAGTTTTGGGACCATTCAAAACAGCATAGCTCTAAAACAATGGA
H392	TTCAGTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAACAATGGA
H393	TGAATCCATTGTTTTGGGACCATTCAAAACAGCATAGCTCTAAAAC
H394	TTCAGTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAACTTACCT
H395	TGAAAGGTAAGTTTTGGGACCATTCAAAACAGCATAGCTCTAAAAC
H396	TTCAGAATAACTCACGTTCCATTGAATACTGTGTAGG
H397	TGAACCTACACAGTATTCAATGGAACGTGAGTTATTC
H398	TTCAGAATAACTCACGTTCCATTGAATACTGTGTAGGCATGTTATAATCCACACCCTTGC
H399	TGAAGCAAGGGTGTGGATTATAACATGCCTACACAGTATTCAATGGAACGTGAGTTATTC
H400	TTC AGT TTT GGG ACC ATT CAA AAC AGC ATA GCT CTA AAA CTT ACC T
H401	TGA AAG GTA AGT TTT AGA GCT ATG CTG TTT TGA ATG GTC CCA AAA C

H402	NNNNNCAGAGGGGCTTTTCAAGACTG
H403	NNNNNCAGGAAGAAATCAACCAGCGC
H404	NNNNNTCAAGGGGCTTTTCAAGACTG
H405	NNNNNTCAGAAGAAATCAACCAGCGC
H406	NNNNNGTCAGGGGCTTTTCAAGACTG
H407	NNNNNGTCGAAGAAATCAACCAGCGC
H408	NNNNNAGTAGGGGCTTTTCAAGACTG
H409	NNNNNAGTGAAGAAATCAACCAGCGC
H410	TTCCCAAAATAAAGGAAGAGTTATTTACTTTGTTTTCAG
H411	AAGTAAATAACTCTTCCTTTATTTTGGGAAAAGGCTG
H412	CCAAAATAAAGAGAAAGTTATTTACTTTGTTTTCAGATAC
H413	ACAAAGTAAATAACTTTCTCTTTATTTTGGGAAAAGGCTG
H414	TTTGTTTTCACGTACATTTTCATATAATGGTAAAGAGATG
H415	TACCATTATATGAAAATGTACGTGAAAACAAAGTAAATAACTCTCTC
H416	AAAAACAGTTTGCAGAAATGATTTATTTACATGGTGAAAG
H417	TGTAAATAAATCATTTCTGCAAACTGTTTTTCCGTGACCG
H418	TATTTACATGGTAAAAGAAATAATTGTATTGCAAACTCCG
H419	TACAATTATTTCTTTTACCATGTAAATAAATCATTCGTGC
H420	TGTATTGCAAACTCCGATAAAAGACTTGTATTTCTTGGGG
H421	TACAAGTCTTTTATCGGAGTTTGCAATACAATTATTTCTTTCACC
H422	GTATTTCTTGGGAAGGCTTTTGATGAATCTTAATTTTTCC
H423	AGATTCATCAAAAGCCTTCCCAAGAAATACAAGTCTTTCATCGG
H424	ACAATTGATCGTAAACAATATAGGTCTACAAAAGAAGTTTTAGATGCCAC
H425	TTGTAGACCTATATTGTTTACGATCAATTGTTGTATCAAAATATTTAAAAGCAGC
H426	AAATATGGTGGTTTTGAAAGTCCAACGGTAGCTTATTCAGTCC
H427	TACCGTTGGACTTTCAAAACCACCATATTTTTTTGGATCCC
H428	AAACCGAAATCATACGCAAAAATATTCATGTTAAG
H429	TTAACATGAATATTTTTGCGTATGATTTCGCAAAA
H430	AAACGAAATCATACGCAAAAATATTCATGTTAACG
H431	GTTAACATGAATATTTTTGCGTATGATTTCCAAAA
H432	CCAAAATAAAGAGAAAGTTATTTACTTTGTTTTCACGTAC
H433	AAACAAACAGTGACAGAAACTATTGAGTACGAGGG
H434	AAAACCCTCGTACTCAATAGTTTCTGTCACTGTTT
H435	AAACAGAAAACAGTGACAGAAACTATTGAGTACGG
H436	AAAACCGTACTCAATAGTTTCTGTCACTGTTTTCT
H437	TTGTTTCCCAAAACACCTATACCTG
H438	ATTTTCAGGTATAGGTGTTTTGGG
H439	GGAAGTCTGAAGAAACAGCTACCCCATGGAATTTTGAAG

H440	AAACTATTCAAATTGTTACTTCATAATCTTTTGTG
H441	AAAACACAAAAGATTATGAAGTAACAATTTGAATA
H442	AAACAAGCTAGGAAGATTTAACTTAATACCTAATG
H443	AAAACATTAGGTATTAAGTTAAATCTTCCTAGCTT
H444	AAACGAGCTAGGGAGTTTAACGGTATGGAAGAAGG
H445	AAAACCTTCTTCCATACCGTTAAACTCCCTAGCTC
H446	AAACACATGTAACAGAGAAAATGGACAGAGAGTTG
H447	AAAACAACTCTCTGTCCATTTTCTCTGTTACATGT
H448	AAACATGTCACTTATAACCAAATGTTCAAGAAATG
H449	AAAACATTTCTTGAACATTTGGTTATAAGTGACAT
H450	AAACCAAATGTTCAAGAAATGGAGTGAAGCATAAG
H451	AAAACTTATGCTTCACTCCATTTCTTGAACATTTG
H452	AAACTTCACTTAATTGACGCATATGATTTAACAAG
H453	AAAACTTGTTAAATCATATGCGTCAATTAAGTGAA
H454	AAACTAGCATACCTTGTGTTACGCGGTATGGGTAG
H455	AAAACTACCCATACCGCGTAACACAAGGTATGCTA
H456	AAACAAGAAGATGCTTATATAGAAAAATTCCTTAG
H457	AAAACTAAGGAATTTTTCTATATAAGCATCTTCTT
H458	AAACAGATACGTATGCACATTACACAAGGTATTAG
H459	AAAACTAATACCTTGTGTAATGTGCATACGTATCT
H460	AAACCTGGTGGTAGTCGTGCTACAAAGATTCCGTG
H461	AAAACACGGAATCTTTGTAGCACGACTACCACCAG
H462	AAACGTACAATCACTAATTTTGTTAGCAGTATTTG
H463	AAAACAAATACTGCTAACAAAATTAGTGATTGTAC
H464	AAACTCCGCCAATAAACTTATGTGTGTATGCCTTG
H465	AAAACTCCGCCAATAAACTTATGTGTGTATGCCTT
H466	AAACGACATCATCGCAACATGTTTAGCTACATCAG
H467	AAAACTGATGTAGCTAAACATGTTGCGATGATGTC
H468	AAACGCTTTTATGTTATAATTGCTTTTATATAGTG
H469	AAAACACTATATAAAAGCAATTATAACATAAAAGC
H470	AAACTTTTTAACTTCAGGTCGTTGATAATACTCTG
H471	AAAACAGAGTATTATCAACGACCTGAAGTTAAAAA
H472	AAACATGACTTTAGCATTCCCGTATAACAGTTTAG
H473	AAAACTAAACTGTTATACGGGAATGCTAAAGTCAT
H474	AAACCTTTTATATAGTAGGAGTGAACTATATAGCG
H475	AAAACGCTATATAGTTCACTCCTACTATATAAAAG
H476	AAACCCGTTATGGCCTAGAATCATATTGCTAAAAG
H477	AAAACTTTTAGCAATATGATTCTAGGCCATAACGG
H478	AAACTTATTTTGCGTTAGAATTGACACCTCAAGAG
H479	AAAACTCTTGAGGTGTCAATTCTAACGCAAAATAA

H480	AAACTTATCGTGAGTGGGAGAAATATAAGCGAAAG
H481	AAAACTTTCGCTTATATTTCTCCCACTCACGATAA
H482	AAACGACAAATGCTATTCAACATTCAGTTAAAGAG
H483	AAAACTCTTTAACTGAATGTTGAATAGCATTTGTC
H484	AAACGCTAAAACAAAAGATTTTATTAAAGCAAGAG
H485	AAAACTCTTGCTTTAATAAAATCTTTTGTTTTAGC
H486	AAACTGATACATTAACATTTAGTAAATCATTACGG
H487	AAAACCGTAATGATTTACTAAATGTTAATGTATCA
H488	AAACTTGTTTATCGATTGGAGCATGCAAATAACTG
H489	AAAACAGTTATTTGCATGCTCCAATCGATAAACAA
H490	AAACAGTTGGATTTAGATGCAAACCCCGCTAAAAG
H491	AAAACTTTTAGCGGGGTTTGCATCTAAATCCAACT
H492	AAACATTACTTAACACACTGCTAACAGCTGCAATG
H493	AAAACATTGCAGCTGTTAGCAGTGTGTTAAGTAAT
H494	AAACGGATATTGTCGTTTTCCCGTCAAAGTATGGG
H495	AAAACCCATACTTTGACGGGAAAACGACAATATCC
H496	AGAGGTTGAACTACGTAAGAGG
H497	TTGTGGTGGATACTGTGCC
H498	TACGATAATACTTATTATTATGTATTTCGAGG
H499	TTTACTGACTTTGCAAAACGC
H500	GAGAAATATCAAAATGATGATGTG
H501	GACTGTTTCTCTCATTGTTGCG
H502	TTTTTTGTTATGATGTGTTACACATGC
H503	AAGGAAGATGTCTCCTGTGG
H504	GTGATGAAGAAGAAATATTTAAGATGG
H505	TGTTAGATGAAGGTATGAGC
H506	GATCTTGCAATGTCTTATGACC
H507	ATGTTTTAACCATATCTAAATCAGC
H508	AGGAAGACACTAATGAATAACCG
H509	GGTATGGATTTCAGTGTTATGATTACG
H510	CATGAATCGCACCGGC
H511	TACTGCAATGGCTCCTATAGC
H512	AAACAATTAGGTTTTATTACTAATAAAAATGATAG
H513	AAAACTATCATTTTTATTAGTAATAAAACCTAATT
H514	AAACGTTGGACTAATGGCGTTGCGCAACCTGGTTG
H515	AAAACAACCAGGTTGCGCAACGCCATTAGTCCAAC
H516	AAACCAAGCAGAAAAATGGTTTGACAATTCATTAG
H517	AAAACTAATGAATTGTCAAACCATTTTTCTGCTTG
H518	AAACACGGTTATTCAACTAATTCAAGAATTACAGG
H519	AAAACCTGTAATTCTTGAATTAGTTGAATAACCGT

H520	AAGCAACAGGACAAGCACC
H521	GCGGCCTCTAATACGACTCACTATAGGGCATATTAGATCGAGTCAAGGGTTTTAGAGCTAGAAATAGCA
H522	GCGGCCTCTAATACGACTCACTATAGGGAGACATATTAGATCGAGTCAGTTTTAGAGCTAGAAATAGCA
H523	GCGGCCTCTAATACGACTCACTATAGGGAGATCGAGTCAAGGAGGTTTGTTTTAGAGCTAGAAATAGCA
H524	GCGGCCTCTAATACGACTCACTATAGGGGATCGAGTCAAGGAGGTTTTGTTTTAGAGCTAGAAATAGCA
H525	GCGGCCTCTAATACGACTCACTATAGGGATCGAGTCAAGGAGGTTTTGGTTTTAGAGCTAGAAATAGCA
H526	AGCAGTAGGGATTATGACGG
H527	ATCAAATCAGACTGATCGCTC
H528	AAACACGCAGATTGTTTGAGTGGTTACGTCAAAAG
H529	AAAACTTTTGACGTAACCACTCAAACAATCTGCGT
H530	AAACTTTAGCGATATTAATTATGCTCGTAAGAATG
H531	AAAACATTCTTACGAGCATAATTAATATCGCTAAA
H532	AAACCTCTGATGACGAATTAGCTATCATAACTTCG
H533	AAAACGAAGTTATGATAGCTAATTCGTCATCAGAG
H534	AAACCATTTTAGATTTCAAAAGTTTAGTATCTATG
H535	AAAACATAGATACTAAACTTTTGAAATCTAAAATG
H536	AAACGTATCTCTATTGACACCAATTTCTTCAGAAG
H537	AAAACTTCTGAAGAAATTGGTGTCAATAGAGATAC
H538	AAACATAGGGATTTTACAAGTGTACTTACAAGTAG
H539	AAAACTACTTGTAAGTACACTTGTAAAATCCCTAT
H540	AAACGAAATTAACTTGAAGCATTTCAAAGAAAATG
H541	AAAACATTTTCTTTGAAATGCTTCAAGTTAATTTC
H542	AAACAGTAGCTACTGCATCTGCAAATACAATTTTG
H543	AAAACAAAATTGTATTTGCAGATGCAGTAGCTACT
H544	AAACGAATAACTCACGTTCCATTGAATACTGTGTG
H545	AAAACACACAGTATTCAATGGAACGTGAGTTATTC
H546	AAACTGAATATTCATCTCTCGGTATATATAATCCG
H547	AAAACGGATTATATATACCGAGAGATGAATATTCA
H548	AAACCCAGAAGTTATGATAGCTAATTCGTCATCAG
H549	AAAACTGATGACGAATTAGCTATCATAACTTCTGG
H550	AAACATGCTCCAATCGATAAACAATTAGATAAACG
H551	AAAACGTTTATCTAATTGTTTATCGATTGGAGCAT
H552	NNNNNGTCTCTGGTAGAAAAGATATCCTACGAG
H553	NNNNNTCATCTGGTAGAAAAGATATCCTACGAG
H554	NNNNNGTCCTCGTACAGTGAACCTTTTTCACC
H555	NNNNNTCACTCGTACAGTGAACCTTTTTCACC

H556	AGTTCAACAAACGGGTCATAACCTGAAGGAAGATCTGG
H557	CAGAATCCACGAGATCTGTGCCAGTTCGTAATGTCTGG
H558	TACGAACTGGCACAGATCTCGTGGATTCTGTGATTTGG
H559	CCTTCAGGTTATGACCCGTTTGTTGAACTAATGGGTGC
H560	AAACGATTGAACAATAGATTGTCTAAAGTTGAGAG
H561	AAAACTCTCAACTTTAGACAATCTATTGTTCAATC
H562	AAACTGTGGGAAAGTGGAAGAACTGAACCTAGAAG
H563	AAAACTTCTAGGTTCAGTTCTTCCACTTTCCCACA
H564	AAACTTTGTTCAATGTTTCTAAAGGTTATCTCTTG
H565	AAAACAAGAGATAACCTTTAGAAACATTGAACAAA
H566	AAACACCTAGCGAATGTATAGCACTAAAAATAAAG
H567	AAAACTTTATTTTTAGTGCTATACATTCGCTAGGT
H568	AAACACAATCTATTGTTCAATCTGATTTCTTTTAG
H569	AAAACTAAAAGAAATCAGATTGAACAATAGATTGT
H570	AAACACTTGAAATTTTTTCGACCATACCCATTCTG
H571	AAAACAGAATGGGTATGGTCGAAAAAATTTCAAGT
H572	AAACATATGGAACCTCGATTTCGCTATCAAATTCG
H573	AAAACGAATTTGATAGCGAAATCGAGGTTCCATAT
H574	AAACTCCGTTTATTTTTAGTGCTATACATTCGCTG
H575	AAAACAGCGAATGTATAGCACTAAAAATAAACGGA
H576	AAACATAAAAGTGTAAAAACATTATATATAAGGAG
H577	AAAACTCCTTATATATAATGTTTTTACACTTTTAT
H578	AATAATTCTGTTGATTTCGTGCCACTGTGCGGG
H579	CCCGCACAGTGGCACGAAATCAACAGAATTATT
H580	ACTAAATTGTCCGTCAATAATTCTGTTGATTTCGTGCCACTGTGCGGGTGTGAATTGCTTTCT
H581	AGAAAGCAATTCACACCCGCACAGTGGCACGAAATCAACAGAATTATTGACGGACAATTTAGT
H582	TGCACAAGCAGAAATGGAAGCTAAGAAAATTGG
H583	CCAATTTTCTTAGCTTCCATTTCTGCTTGTGCA
H584	TTCTAAGCCTGAATATGCACAAGCAGAAATGGAAGCTAAGAAAATTGGTGTAATTATTCCGTT
H585	AACGGAATAATTACACCAATTTTCTTAGCTTCCATTTCTGCTTGTGCATATTCAGGCTTAGAA
H586	AACAACATCACCTATTTTAGGGTTAGCTTCTGG
H587	CCAGAAGCTAACCCTAAAATAGGTGATGTTGTT
H588	AGAATCCACCACTCTAACAACATCACCTATTTTAGGGTTAGCTTCTGGGAAATGTTCACGTAA
H589	TTACGTGAACATTTCCCAGAAGCTAACCCTAAAATAGGTGATGTTGTTAGAGTGGTGGATTCT
H590	NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAGG
H591	AATAATTCTGTTGATTTCGTGCCACTGTGC

H592	GCACAGTGGCACGAAATCAACAGAATTATT
H593	GCACAGTGGCACGAAATCAACAG
H594	GTTTGGAGTATGTAGAAGTACAGTATACAACTGG
H595	CCAGTTGTATACTGTACTTCTACATACTCCAAAC
H596	GTTTGGAGTATGTAGAAGTACAGTATACAACTAT
H597	ATAGTTGTATACTGTACTTCTACATACTCCAAAC
H598	GTTTGGAGTATGTAGAAGTACAGTATACAAC
H599	GTTGTATACTGTACTTCTACATACTCCAAAC
H600	GTAGAAGTACAGTATACAACTGG
H601	CCAGTTGTATACTGTACTTCTAC
H602	GTAGAAGTACAGTATACAACTAT
H603	ATAGTTGTATACTGTACTTCTAC
H604	GTAGAAGTACAGTATACAAC
H605	GTTGTATACTGTACTTCTAC
H606	NNNNNGTCGGCTTTTCAAGACTGAAGTCTAG
H607	NNNNNCAGGGCTTTTCAAGACTGAAGTCTAG
H608	NNNNNAGTGGCTTTTCAAGACTGAAGTCTAG
H609	NNNNNTCAGGCTTTTCAAGACTGAAGTCTAG
H610	NNNNNGCTGGCTTTTCAAGACTGAAGTCTAG
H611	NNNNNCGAGGCTTTTCAAGACTGAAGTCTAG
H612	ATGAATGGATTGAAGAGAACACAGACGAACAGG
H613	CCTGTTCGTCTGTGTTCTCTTCAATCCATTCAT
H614	TTAACCAAGCAATAGATGAATGGATTGAAGAGAACACAGACGAACAGGACAGACTAATTAACT
H615	AGTTAATTAGTCTGTCCTGTTCGTCTGTGTTCTCTTCAATCCATTCATCTATTGCTTGGTTAA
H616	AGAAATTATCGAATACTTAAATAAAAAAGCAGG
H617	CCTGCTTTTTTATTTAAGTATTCGATAATTTCT
H618	TTCCATTCCCTATAAAGAAATTATCGAATACTTAAATAAAAAAGCAGGAAAGCATTTTAAACA
H619	TGTTTAAAATGCTTTCCTGCTTTTTTATTTAAGTATTCGATAATTTCTTTATAGGGAATGGAA
H620	ATGAATGGATTGAAGCTTAAATAAAAAAGCAGG
H621	CCTGCTTTTTTATTTAAGCTTCAATCCATTCAT
H622	AGAAATTATCGAATAAGAACACAGACGAACAGG
H623	CCTGTTCGTCTGTGTTCTTATTCGATAATTTCT
H624	ATGAATGGATTGAAGAGAACATAAAAAAGCAGG
H625	CCTGCTTTTTTATGTTCTCTTCAATCCATTCAT
H626	ATGAATGGATGAATACTTAAACAGACGAACAGG
H627	CCTGTTCGTCTGTTTAAGTATTCATCCATTCAT
H628	ATGAATGGATGAATACTTAAATAAAAAAGCAGG

H629	CCTGCTTTTTTATTTAAGTATTCATCCATTCAT
H630	AGAAATTATCTGAAGAGAACACAGACGAACAGG
H631	CCTGTTCGTCTGTGTTCTCTTCAGATAATTTCT
H632	AGAAATTATCTGAAGAGAACATAAAAAAGCAGG
H633	CCTGCTTTTTTATGTTCTCTTCAGATAATTTCT
H634	AGAAATTATCGAATACTTAAACAGACGAACAGG
H635	CCTGTTCGTCTGTTTAAGTATTCGATAATTTCT
H636	TTCCATTCCCTATAAATGAATGGATTGAAGAGAACACAGACGAACAGGAAAGCATTTTAAACA
H637	TGTTTAAAATGCTTTCCTGTTCGTCTGTGTTCTCTTCAATCCATTCATTTATAGGGAATGGAA
H638	TTAACCAAGCAATAGAGAAATTATCGAATACTTAAATAAAAAAGCAGGACAGACTAATTAACT
H639	AGTTAATTAGTCTGTCCTGCTTTTTTATTTAAGTATTCGATAATTTCTCTATTGCTTGGTTAA
H641	AAACATGAATGGATTGAAGAGAACACAGACGAACG
H642	AAAACGTTCGTCTGTGTTCTCTTCAATCCATTCAT
H643	AAACAGAAATTATCGAATACTTAAATAAAAAAGCG
H644	AAAACGCTTTTTTATTTAAGTATTCGATAATTTCT
H645	ATGAATGGATTGAAGAGAACACAGACGAACATT
H646	AATGTTCGTCTGTGTTCTCTTCAATCCATTCAT
H647	TTAACCAAGCAATAGATGAATGGATTGAAGAGAACACAGACGAACATTACAGACTAATTAACT
H648	AGTTAATTAGTCTGTAATGTTCGTCTGTGTTCTCTTCAATCCATTCATCTATTGCTTGGTTAA
H649	AGAAATTATCGAATACTTAAATAAAAAAGCATT
H650	AATGCTTTTTTATTTAAGTATTCGATAATTTCT
H651	TTCCATTCCCTATAAAGAAATTATCGAATACTTAAATAAAAAAGCATTAAAGCATTTTAAACA
H652	TGTTTAAAATGCTTTAATGCTTTTTTATTTAAGTATTCGATAATTTCTTTATAGGGAATGGAA
H653	TTCCATTCCCTATAAAGAAATTATCGAATACTTAAATAAAAAAGCAGGGGAGGGGTTTAAACA
H654	TGTTTAAACCCCTCCCCTGCTTTTTTATTTAAGTATTCGATAATTTCTTTATAGGGAATGGAA
H655	CACATCAATTAGTAAGACGCCAAAAGTAACAGG
H656	CCTGTTACTTTTGGCGTCTTACTAATTGATGTG
H657	ATAATAATGAACATGTCTTGTCACAGTTTCAGG
H658	CCTGAAACTGTGACAAGACATGTTCATTATTAT
H659	ATGAATGGATTGAAGAGAACCAAAAGTAACAGG
H660	CCTGTTACTTTTGGTTCTCTTCAATCCATTCAT
H661	AGAAATTATCGAATACTTAATCACAGTTTCAGG
H662	CCTGAAACTGTGATTAAGTATTCGATAATTTCT

H663	NNNNNCCCCAAAATTTTTTAGACAAAAATAGTC
H664	NNNNNAAACAAAATTTTTTAGACAAAAATAGTC
H665	NNNNNTTTCAAAATTTTTTAGACAAAAATAGTC
H666	NNNNNAGTCCAAAATTTTTTAGACAAAAATAGTC
H667	NNNNNCTGACAAAATTTTTTAGACAAAAATAGTC
H668	ATAATAATGAACATGTCTTGACAGACGAACAGG
H669	CCTGTTCGTCTGTCAAGACATGTTCATTATTAT
H670	ATAATAATGATGAAGAGAACTCACAGTTTCAGG
H671	CCTGAAACTGTGAGTTCTCTTCATCATTATTAT
H672	ATGAATGGATACATGTCTTGTCACAGTTTCAGG
H673	CCTGAAACTGTGACAAGACATGTATCCATTCAT
H674	ATGGATTGAAGAGAACACAGACGAAC
H675	GTCTGTGTTCTCTTCAATCCATTCAT
H676	ATTATCGAATACTTAAATAAAAAAGC
H677	TTTTATTTAAGTATTCGATAATTTCT
H678	ATGGATTGAAGAGAACATAAAAAAGC
H679	TTTTATGTTCTCTTCAATCCATTCAT
H680	ATGGATGAATACTTAAACAGACGAAC
H681	GTCTGTTTAAGTATTCATCCATTCAT
H682	ATTATCTGAAGAGAACACAGACGAAC
H683	GTCTGTGTTCTCTTCAGATAATTTCT
H684	ATGGATGAATACTTAAATAAAAAAGC
H685	TTTTATTTAAGTATTCATCCATTCAT
H686	ATTATCTGAAGAGAACATAAAAAAGC
H687	TTTTATGTTCTCTTCAGATAATTTCT
H688	ATTATCGAATACTTAAACAGACGAAC
H689	GTCTGTTTAAGTATTCGATAATTTCT
H690	AAAGGATACGTGTAAAGACATATTAGATCGAGTCAAGGAGGTTTTG
H691	CAAAACCTCCTTGACTCGATCTAATATGTCTTTACACGTATCCTTT
H692	AAAGGATACGTGTAAAGACATATTAGATCGAGTCAAGGAGGTTTTGGGGAAGTG
H693	CACTTCCCCAAAACCTCCTTGACTCGATCTAATATGTCTTTACACGTATCCTTT
H694	CCTCTAATACGACTCACTATAGGTGAAGAGAACACAGACGAACGTTTAAGAGCTATGC
H695	CCTCTAATACGACTCACTATAGGAATACTTAAATAAAAAAGCGTTTAAGAGCTATGC
H696	AACTGCTACTTGTTGGAGC
H697	TTATCTCTTGTAGCAAACGTGG
H698	TGGCGTTCAAGAACTTATGG
H699	TACTCGTAACCATTCGGGTG

H700	TATAAAGAAATTATCGAATACTTAAACCGATCAGTAGGAAAGC
H701	GCTTTCCTACTGATCGGTTTAAGTATTCGATAATTTCTTTATA
H702	TATAAAGAAATTATCGAATACTTAATTAATGACTAAGGAAAGC
H703	GCTTTCCTTAGTCATTAATTAAGTATTCGATAATTTCTTTATA
JW3	AAAACAGCATAGCTCTAAAACG
JW4	AAAACAGCATAGCTCTAAAACA
JW5	AAAACAGCATAGCTCTAAAACT
JW8	GGCTTTTCAAGACTGAAGTCTAG
L400	CGAAATTTTTTAGACAAAAATAGTC
oGG82	AACATTGCCGATGATAACTTGAG
oGG83	GTTTTGGGACCATTCAAAACAGCATAGCTCTAAAACCTCGTAG
PS192	CGCGGATCCATGGCTGGTTGGCGTACTGTTGTGG
PS193	CGCCTCGAGTCATATCCTAAATTCAGGAACTCC
PS199	CGAGCATATGACGACCTTCGATATGATCGGCAATGTTGAATGGAGACCATTC
PS200	GAATGGTCTCCATTCAACATTGCCGATCATATCGAAGGTCGTCATATGCTCG
PS202	CATCATCATCATCATCACAGCAGCGGCATGGATAAGAAATACTCAATAGG
PS203	CCTATTGAGTATTTCTTATCCATGCCGCTGCTGTGATGATGATGATGATG
PS204	CGACAAGCTTGCGGCCGCACTCGAGCTTTTTATTTTAGGAGGCAAAAATG
PS205	GGATCTCAGTGGTGGTGGTGGTGGTGTACCATATTTTTAGTTATTAAGAAATAATC
PS206	GATTATTTCTTAATAACTAAAAATATGGTACACCACCACCACCACCACTGAGATCC
PS207	CATTTTTGCCTCCTAAAATAAAAAGCTCGAGTGCGGCCGCAAGCTTGTCG
PS284	GCTAGCGATATTATGGCACCATTTAGGCCTTTAG
PS285	CTAAAGGCCTAAATGGTGCCATAATATCGCTAGC
PS334	TACTTCCAATCCAATGCAATGAGCTATCGCTATATG
PS335	TTATCCACTTCCAATGTTATTATTAGCTTTCATCAAAGGC
PS336	CGCGGATCCATGAACCTGAACTTTAGCCTGCTGG
PS337	CGCCTCGAGTTACACCATATTTTTGGTAATCAG
PS354	GTTCCTGAATTTAGGATATGAAACATTGCCGATCATATCGAAGG
PS355	CCTTCGATATGATCGGCAATGTTTCATATCCTAAATTCAGGAAC

References

1. Thomas, C. M. & Nielsen, K. M. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature Reviews Microbiology* **3,** 711–721 (2005).
2. Bikard, D. & Marraffini, L. A. Innate and adaptive immunity in bacteria: Mechanisms of programmed genetic variation to fight bacteriophages. *Current Opinion in Immunology* **24,** 15–20 (2012).
3. Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol.* **60,** 174–182 (2005).
4. Makarova, K. K. S. *et al.* 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).
5. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science (80-.).* **315,** 1709–1712 (2007).
6. Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. & Nakata, A. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. *J. Bacteriol.* **169,** 5429–5433 (1987).
7. Marraffini, L. A. CRISPR-Cas Immunity against Phages: Its Effects on the Evolution and Survival of Bacterial Pathogens. *PLoS Pathog.* **9,** 1–4 (2013).
8. Jansen, R., Embden, J. D. A. van, Gaastra, W. & Schouls, L. M. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.* **43,** 1565–1575 (2002).
9. Sorek, R., Kunin, V. & Hugenholtz, P. CRISPR - A widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat. Rev. Microbiol.* **6,** 181–186 (2008).
10. Haft, D. H., Selengut, J., Mongodin, E. F. & Nelson, K. E. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/cas subtypes exist in prokaryotic genomes. *PLoS Comput. Biol.* **1,** 0474–0483 (2005).
11. Agari, Y. *et al.* Transcription Profile of Thermus thermophilus CRISPR Systems after Phage Infection. *J. Mol. Biol.* **395,** 270–281 (2010).
12. Pougach, K. *et al.* Transcription, processing and function of CRISPR cassettes in Escherichia coli. *Mol. Microbiol.* **77,** 1367–1379 (2010).
13. Kunin, V., Sorek, R. & Hugenholtz, P. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol.* **8,** (2007).
14. Kamerbeek, J. *et al.* Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. *J. Clin.*

Microbiol. **35,** 907–914 (1997).

15. Bolotin, A., Quinquis, B., Sorokin, A. & Dusko Ehrlich, S. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151,** 2551–2561 (2005).
16. Horvath, P. *et al.* Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. *J. Bacteriol.* **190,** 1401–1412 (2008).
17. Deveau, H. *et al.* Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *J. Bacteriol.* **190,** 1390–1400 (2008).
18. Marraffini, L. A. & Sontheimer, E. J. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science (80-.).* **322,** 1843–1845 (2008).
19. Bikard, D., Hatoum-Aslan, A., Mucida, D. & Marraffini, L. A. CRISPR interference can prevent natural transformation and virulence acquisition during in vivo bacterial infection. *Cell Host Microbe* **12,** 177–186 (2012).
20. Zhang, Y. *et al.* Processing-Independent CRISPR RNAs Limit Natural Transformation in Neisseria meningitidis. *Mol. Cell* **50,** 488–503 (2013).
21. Haurwitz, R. E., Jinek, M., Wiedenheft, B., Zhou, K. & Doudna, J. A. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science (80-.).* **329,** 1355–1358 (2010).
22. Westra, E. R. *et al.* CRISPR Immunity Relies on the Consecutive Binding and Degradation of Negatively Supercoiled Invader DNA by Cascade and Cas3. *Mol. Cell* **46,** 595–605 (2012).
23. Garneau, J. E. *et al.* The CRISPR/cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468,** 67–71 (2010).
24. Makarova, K. S. *et al.* Evolution and classification of the CRISPR–Cas systems. *Nat. Rev. Microbiol.* **9,** 467–477 (2011).
25. Koonin, E. V., Makarova, K. S. & Zhang, F. Diversity, classification and evolution of CRISPR-Cas systems. *Current Opinion in Microbiology* **37,** 67–78 (2017).
26. Datsenko, K. A. *et al.* Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat. Commun.* **3,** (2012).
27. Swarts, D. C., Mosterd, C., van Passel, M. W. J. & Brouns, S. J. J. CRISPR interference directs strand specific spacer acquisition. *PLoS One* **7,** (2012).
28. Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res.* **40,** 5569–76 (2012).
29. Cady, K. C., Bondy-Denomy, J., Heussler, G. E., Davidson, A. R. & O'Toole, G. A. The CRISPR/Cas adaptive immune system of Pseudomonas aeruginosa mediates resistance to naturally occurring and engineered phages. *Journal of Bacteriology* **194,** 5728–5738 (2012).
30. Lopez-Sanchez, M. J. *et al.* The highly dynamic CRISPR1 system of Streptococcus agalactiae controls the diversity of its mobilome. *Mol. Microbiol.* **85,** 1057–1071 (2012).
31. Li, M., Wang, R., Zhao, D. & Xiang, H. Adaptation of the Haloarcula

hispanica CRISPR-Cas system to a purified virus strictly requires a priming process. *Nucleic Acids Res.* **42,** 2483–2492 (2014).

32. Erdmann, S. & Garrett, R. A. Selective and hyperactive uptake of foreign DNA by adaptive immune systems of an archaeon via two distinct mechanisms. *Mol. Microbiol.* **85,** 1044–1056 (2012).
33. Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L. A. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat. Biotechnol.* **31,** 233–239 (2013).
34. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (80-.).* **337,** 816–821 (2012).
35. Semenova, E. *et al.* Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc. Natl. Acad. Sci.* **108,** 10098–10103 (2011).
36. Westra, E. R. *et al.* Type I-E CRISPR-Cas Systems Discriminate Target from Non-Target DNA through Base Pairing-Independent PAM Recognition. *PLoS Genet.* **9,** (2013).
37. Hale, C. R. *et al.* RNA-Guided RNA Cleavage by a CRISPR RNA-Cas Protein Complex. *Cell* **139,** 945–956 (2009).
38. Marraffini, L. A. & Sontheimer, E. J. Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature* **463,** 568–571 (2010).
39. Savitskaya, E., Semenova, E., Dedkov, V., Metlitskaya, A. & Severinov, K. High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in E. coli. *RNA Biol.* **10,** 716–725 (2013).
40. Paez-Espino, D. *et al.* Strong bias in the bacterial CRISPR elements that confer immunity to phage. *Nat. Commun.* **4,** (2013).
41. Yosef, I. *et al.* DNA motifs determining the efficiency of adaptation into the Escherichia coli CRISPR array. *Proc. Natl. Acad. Sci.* **110,** 14396–14401 (2013).
42. Brouns, S. J. J. *et al.* Small Crispr Rnas Guide Antiviral Defense in Prokaryotes. *Cancer Epidemiol. Biomarkers Prev.* **2,** 531–535 (2008).
43. Sapranauskas, R. *et al.* The Streptococcus thermophilus CRISPR / Cas system provides immunity in Escherichia coli. *Nucleic Acids Res.* **39,** 1–8 (2011).
44. Deltcheva, E., Chylinski, K., Sharma, C. M. & Gonzales, K. CRISPR RNA maturation by trans -encoded small RNA and host factor RNase III. *Nature* **471,** 602–607 (2011).
45. Hatoum-Aslan, A., Maniv, I. & Marraffini, L. A. Mature clustered, regularly interspaced, short palindromic repeats RNA (crRNA) length is measured by a ruler mechanism anchored at the precursor processing site. *Proc. Natl. Acad. Sci.* **108,** 21218–21222 (2011).
46. Hatoum-Aslan, A., Maniv, I., Samai, P. & Marraffini, L. A. Genetic characterization of antiplasmid immunity through a type III-A CRISPR-cas system. *J. Bacteriol.* **196,** 310–317 (2014).
47. Han, D., Lehmann, K. & Krauss, G. SSO1450 - A CAS1 protein from

Sulfolobus solfataricus P2 with high affinity for RNA and DNA. *FEBS Lett.* **583,** 1928–1932 (2009).

48. Wiedenheft, B. *et al.* Structural Basis for DNase Activity of a Conserved Protein Implicated in CRISPR-Mediated Genome Defense. *Structure* **17,** 904–912 (2009).
49. Beloglazova, N. *et al.* A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. *J. Biol. Chem.* **283,** 20361–20371 (2008).
50. Samai, P., Smith, P. & Shuman, S. Structure of a CRISPR-associated protein Cas2 from Desulfovibrio vulgaris. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **66,** 1552–1556 (2010).
51. Nam, K. H. *et al.* Double-stranded endonuclease activity in Bacillus halodurans clustered regularly interspaced short palindromic repeats (CRISPR)-associated Cas2 protein. *J. Biol. Chem.* **287,** 35943–35952 (2012).
52. Andersson, A. F. & Banfield, J. F. Virus population dynamics and acquired virus resistance in natural microbial communities. *Science* **320,** 1047–50 (2008).
53. Tyson, G. W. & Banfield, J. F. Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ. Microbiol.* **10,** 200–207 (2008).
54. Horvath, P. *et al.* Comparative analysis of CRISPR loci in lactic acid bacteria genomes. *Int. J. Food Microbiol.* **131,** 62–70 (2009).
55. Pride, D. T. *et al.* Analysis of streptococcal CRISPRs from human saliva reveals substantial sequence diversity within and between subjects over time. *Genome Res.* **21,** 126–136 (2011).
56. Pride, D. T., Salzman, J. & Relman, D. A. Comparisons of clustered regularly interspaced short palindromic repeats and viromes in human saliva reveal bacterial adaptations to salivary viruses. *Environ. Microbiol.* **14,** 2564–2576 (2012).
57. Díez-Villaseñor, C., Guzmán, N. M., Almendros, C., García-Martínez, J. & Mojica, F. J. M. CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of Escherichia coli. *RNA Biol.* **10,** 792–802 (2013).
58. Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic Acids Res.* **40,** 5569–5576 (2012).
59. Goren, M. G., Yosef, I., Edgar, R. & Qimron, U. The bacterial CRISPR/Cas system as analog of the mammalian adaptive immune system. *RNA Biology* **9,** 549–554 (2012).
60. Goryshin, I. Y., Miller, J. A., Kil, Y. V., Lanzov, V. A. & Reznikoff, W. S. Tn5/IS50 target recognition. *Proc. Natl. Acad. Sci.* **95,** 10716–10721 (1998).
61. Merkel, G., Andrake, M. D., Ramcharan, J. & Skalka, A. M.

Oligonucleotide-based assays for integrase activity. *Methods* **47,** 243–248 (2009).

62. Plagens, A., Tjaden, B., Hagemann, A., Randau, L. & Hensel, R. Characterization of the CRISPR/Cas subtype I-A system of the hyperthermophilic crenarchaeon Thermoproteus tenax. *J. Bacteriol.* **194,** 2491–2500 (2012).
63. Arslan, Z. *et al.* Double-strand DNA end-binding and sliding of the toroidal CRISPR-associated protein Csn2. *Nucleic Acids Res.* **41,** 6347–6359 (2013).
64. Goldberg, G. W., Jiang, W., Bikard, D. & Marraffini, L. A. Conditional tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting. *Nature* **514,** 633–637 (2014).
65. Almendros, C., Guzmán, N. M., Díez-Villaseñor, C., García-Martínez, J. & Mojica, F. J. M. Target Motifs Affecting Natural Immunity by a Constitutive CRISPR-Cas System in Escherichia coli. *PLoS One* **7,** (2012).
66. Barrangou, R. & Marraffini, L. A. CRISPR-cas systems: Prokaryotes upgrade to adaptive immunity. *Molecular Cell* **54,** 234–244 (2014).
67. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507,** 62–67 (2014).
68. Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl. Acad. Sci.* **109,** E2579–E2586 (2012).
69. Szczelkun, M. D. *et al.* Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. *Proc. Natl. Acad. Sci.* **111,** 9798–9803 (2014).
70. Nuñez, J. K. *et al.* Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nat. Struct. Mol. Biol.* **21,** 528–534 (2014).
71. Anders, C., Niewoehner, O., Duerst, A. & Jinek, M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **513,** 569–573 (2014).
72. Kreiswirth, B. N. *et al.* The toxic shock syndrome exotoxin structural gene is not detectably transmitted by a prophage. *Nature* **305,** 709–712 (1983).
73. Bae, T., Baba, T., Hiramatsu, K. & Schneewind, O. Prophages of Staphylococcus aureus Newman and their contribution to virulence. *Mol. Microbiol.* **62,** 1035–1047 (2006).
74. Richter, C. *et al.* Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. *Nucleic Acids Res.* **42,** 8516–8526 (2014).
75. Fonfara, I. *et al.* Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. *Nucleic Acids Res.* **42,** 2577–2590 (2014).
76. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: A

sequence logo generator. *Genome Res.* **14,** 1188–1190 (2004).

77. Jinek, M. *et al.* Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science (80-.).* **343,** (2014).
78. Arslan, Z., Hermanns, V., Wurm, R., Wagner, R. & Pul, Ü. Detection and characterization of spacer integration intermediates in type I-E CRISPR-Cas system. *Nucleic Acids Res.* **42,** 7884–7893 (2014).
79. Nuñez, J. K., Lee, A. S. Y., Engelman, A. & Doudna, J. A. Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature* **519,** 193–198 (2015).
80. Carte, J., Wang, R., Li, H., Terns, R. M. & Terns, M. P. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes Dev.* **22,** 3489–3496 (2008).
81. Jore, M. M. *et al.* Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nat. Struct. Mol. Biol.* **18,** 529–536 (2011).
82. Samai, P. *et al.* Co-transcriptional DNA and RNA cleavage during type III CRISPR-cas immunity. *Cell* **161,** 1164–1174 (2015).
83. Shmakov, S. *et al.* Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. *Mol. Cell* **60,** 385–397 (2015).
84. Heler, R. *et al.* Cas9 specifies functional viral targets during CRISPR-Cas adaptation. *Nature* **519,** 199–202 (2015).
85. Wei, Y., Terns, R. M. & Terns, M. P. Cas9 function and host genome sampling in type II-A CRISPR–cas adaptation. *Genes Dev.* **29,** 356–361 (2015).
86. Kleinstiver, B. P. *et al.* Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* **523,** 481–485 (2015).
87. Jiang, F. *et al.* Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science (80-.).* **351,** 867–871 (2016).
88. Briner, A. E. *et al.* Guide RNA functional modules direct Cas9 activity and orthogonality. *Mol. Cell* **56,** 333–339 (2014).
89. Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. Molecular recordings by directed CRISPR spacer acquisition. *Science (80-.).* **353,** (2016).
90. Paez-Espino, D. *et al.* CRISPR immunity drives rapid phage genome evolution in streptococcus thermophilus. *MBio* **6,** 1–9 (2015).
91. Modell, J. W., Jiang, W. & Marraffini, L. A. CRISPR-Cas systems exploit viral DNA injection to establish and maintain adaptive immunity. *Nature* **544,** 101–104 (2017).
92. Heler, R. *et al.* Mutations in Cas9 Enhance the Rate of Acquisition of Viral Spacer Sequences during the CRISPR-Cas Immune Response. *Mol. Cell* **65,** 168–175 (2017).
93. Kwan, T., Liu, J., DuBow, M., Gros, P. & Pelletier, J. The complete genomes and proteomes of 27 Staphylococcus aureus bacteriophages. *Proc. Natl. Acad. Sci.* **102,** 5174–5179 (2005).
94. Wu, X. & Bartel, D. P. KpLogo: Positional k -mer analysis reveals hidden

specificity in biological sequences. *Nucleic Acids Res.* **45,** W534–W538 (2017).

95. Bradde, S., Vucelja, M., Teşileanu, T. & Balasubramanian, V. Dynamics of adaptive immunity against phage in bacterial populations. *PLoS Comput. Biol.* **13,** (2017).

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

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

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

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

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

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

102. Ivančić-Bace, I., Cass, S. D., Wearne, S. J. & Bolt, E. L. Different genome stability proteins underpin primed and Naïve adaptation in E. Coli CRISPR-Cas immunity. *Nucleic Acids Res.* **43,** 10821–10830 (2015).

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

104. Lee, H., Zhou, Y., Taylor, D. W. & Sashital, D. G. Cas4-Dependent Prespacer Processing Ensures High-Fidelity Programming of CRISPR Arrays. *Mol. Cell* **70,** 48–59.e5 (2018).

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

106. Sheth, R. U., Yim, S. S., Wu, F. L. & Wang, H. H. Multiplex recording of cellular events over time on CRISPR biological tape. *Science (80-.).* **358,** 1457–1461 (2017).