CRISPR-mediated antiviral defence in prokaryotes

Matthijs M. Jore

Thesis committee

Thesis supervisor
Prof. Dr. John van der Oost
Personal Chair at the Laboratory of Microbiology
Wageningen University

Thesis co-supervisor
Dr. Ir. Stan J.J. Brouns
Researcher at the Laboratory of Microbiology
Wageningen University

Members
Prof. Dr. Just M. Vlak
Wageningen University

Prof. Dr. Willem J.H. van Berkel
Wageningen University

Dr. Luciano A. Marraffini,
Rockefeller University, New York (NY), USA

Prof. Dr. Michael P. Terns
University of Georgia, Athens (GA), USA

This research was conducted under the auspices of the Graduate School VLAG

CRISPR-mediated antiviral defence in prokaryotes

Matthijs M. Jore

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. dr. M.J. Kropff,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Wednesday 20 October 2010
at 4 p.m. in the Aula.

Matthijs M. Jore
CRISPR-mediated antiviral defence in prokaryotes,
160 pages.

Thesis, Wageningen University, Wageningen, NL (2010)
With references, with summaries in Dutch and English

ISBN 978-90-8585-781-5

Table of contents

Preface and Outline

Like eukaryotes, bacteria are constantly being attacked by viruses. Because infections by these (bacterio)phages are generally disadvantageous for bacteria, they have developed several defense mechanisms. Well-known examples include mutations of receptors that the phages use for attachment (and entry) to the host, and the restriction-modification system that degrades DNA with a “non-self” methylation pattern. Phages in turn have developed methods to bypass these defense mechanisms, illustrating the fascinating ongoing evolutionary battle. Bacteriophages and their life cycles have been extensively studied in the early days of molecular biology research. However, the development of high throughput sequencing has initiated a new era for this research field. Since 1995, analyses of prokaryotic genome sequences (both bacteria and archaea), revealed the presence of typical repetitive regions, the CRISPR loci. These CRISPRs are composed of short direct repeats that are interspaced with unique spacer sequences of similar length. In 2005, several groups performed detailed analyses of the spacer sequences, resulting in the discovery of homology with phage and plasmid DNA. This has lead to the hypothesis that the CRISPRs might provide the bacteria with yet another defense mechanism against invading nucleic acids. A year later the group of Eugene Koonin released an article, which comprehensively described a putative model of the CRISPR-mediated defense system. This model was based on bioinformatics analysis of the CRISPRs and CRISPR-associated (*cas*) genes, and limited experimental data that was available at that time. At this point the research described in this thesis was initiated. We aimed at unraveling the molecular pathway of this fascinating heritable and adaptive immune system in our model organism *Escherichia coli*.

Chapter 1 gives an overview of the different stages of CRISPR-mediated defense. It compares the different CRISPR/Cas subtypes and discusses the experimental data that is available in literature. Basically three stages of CRISPR defense are distinguished. During the first stage, fragments from invading nucleic acids are integrated as spacers into a CRISPR locus. During the second stage the CRISPR is transcribed and the CRISPR RNA is cleaved into small CRISPR RNAs (crRNAs) which each contain one spacer and part(s) of the repeat sequence. During the third and final stage these crRNAs guide the antiviral Cas protein machinery to the target; the crRNAs can base pair with invading nucleic acids that have been previously filed into the CRISPR locus. CRISPR/Cas thus provides prokaryotes with an adaptive and heritable immune system against potentially hazardous nucleic acids. Analogies with RNA interference (RNAi) in eukaryotes, which can also serve as an antiviral defense, are discussed.

Chapter 2 describes the identification of a CRISPR-associated complex for antiviral defense (Cascade) that is composed of 5 different subunits (CasABCDE). It is demonstrated that the CRISPR RNA transcript (pre-crRNA) is cleaved by the Cascade complex into small CRISPR RNAs (crRNAs) during the second stage of CRISPR-mediated defense. The CasE subunit is identified as a metal-independent endoribonuclease that specifically recognizes repeat sequences of the *E. coli* K12 CRISPR. This chapter also describes the engineering of an *E. coli* strain to become resistant against the well studied bacteriophage lambda (λ).

The next stage after CRISPR expression and processing is that of target recognition, a process that has been studied in more detail in **Chapter 3**. It addresses the question how mature crRNAs can recognize and bind to their complementary target. An integrated analysis has been used to reveal insights in the structure and function of the Cascade complex. Mass spectrometry has been used to determine the subunit stoichiometry. In addition, the architecture of the crRNA has been studied in more detail. By using electron microscopy and small angle X-ray scattering, a first low resolution model of Cascade has been obtained.

In **Chapter 4** it is shown that the CRISPR/Cas system can also inhibit plasmid transformation. This enabled us to study the sequential features of the target that are required for successful interference. A mutant library of CRISPR-targeted plasmids was created and transformed. Plasmids that were able to evade the CRISPR/Cas system could cause colony formation. Sequence analyses of these escape mutants revealed mutations in the protospacer (the targeted sequence identical to the spacer in the crRNA) and the conserved Protospacer Adjacent Motif (PAM).

One phage lambda particle can infect, replicate and release 50-100 particles in about 30 minutes, not only destroying the host cell but also threatening neighboring cells. Phages infect and replicate at the cellular poles. In **Chapter 5** we checked the hypothesis that the Cas protein machinery might be located in this same location to rapidly counteract a potentially devastating infection. Indeed we observed with high resolution microscopy that Cas proteins may be located at the cellular poles.

Transformation of a phage lambda targeting CRISPR into *E. coli* K12 did not result in elevated immunity, due to transcriptional silencing of the *cas* genes and CRISPR. In **Chapter 6** two regulators of the CRISPR/Cas system in *E. coli* are identified and characterized. H-NS functions as a transcription repressor, while its antagonist LeuO acts as a transcription activator. This regulation has been analyzed both at transcription level and functional level by means of plaque assays.

Chapter 7 finally summarizes and discusses the research described in this thesis. Remaining questions about the mechanism of CRISPR/Cas are addressed. It also describes the wide variety of potential applications of the fundamental research carried out here.

Chapter 1

Introduction

Partly adapted from:

Matthijs M. Jore, Stan J. J. Brouns, John van der Oost
RNA in Defense: CRISPRs protect prokaryotes against mobile genetic elements
in RNA Worlds: From Life's Origins to Diversity in Gene Regulation, *in press*

The evolution of micro-organisms is significantly influenced both qualitatively and quantitatively by the continuous exchange of genomic material with mobile genetic elements: viruses and plasmids. Viruses are among the most abundant entities on earth (Bergh et al. 1989; Wommack and Colwell 2000) and they proliferate by a series of events: adsorption of the virion to the host's cell wall, injection of the viral genome (DNA, RNA) through the cell membrane(s), expression of viral genes, replication of the viral genome and assembly of viral protein capsids, and finally release of progeny virions (Sturino and Klaenhammer 2004). Plasmids are another main class of selfish mobile elements. After entry plasmid DNA resides in a host, either free in the cytoplasm or integrated in the host genome. Plasmids can be transferred from donor to recipient via conjugation, making use of dedicated transfer systems (Llosa et al. 2002).

Despite the occasional gain of function as a result of horizontal gene transfer, recombination with mobile elements can also cause severe damage; disruption of either structural or regulatory regions on the host genome leads to loss of function. Additionally phage infections can eventually lead to host cell lysis. To avoid these detrimental effects, sophisticated mechanisms have evolved to defend host organisms against nucleic acids from invading mobile elements. Several defense systems have been recognized in prokaryotes that are very different from the eukaryotic immune systems. A passive defense mechanism may act at the level of virion adsorption and/or injection of its genomic material. Spontaneous mutations in virus receptor proteins of the host can perturb virus attachment and genome injection, not affecting host fitness, e.g. the maltoporin of *E. coli* used by phage lambda (Hofnung et al. 1976). A well known active defense mechanism is the restriction-modification (R-M) system. Dedicated methyltransferases modify potential cleavage sites of the host DNA, preventing strand cleavage by restriction enzymes. Incoming invader DNA lacks these modifications, and is therefore a target for digestion by these endonucleases (Tock and Dryden 2005). An additional mechanism that appears functionally analogous to eukaryotic apoptosis, is the prokaryotic Abortive infection mechanism (Abi). This mechanism inhibits phage multiplication either by blocking the phage replication machinery, or by inhibiting host translation. This results in death of both host and virus, a sacrifice that will save the rest of the population (Chopin et al. 2005).

Recently another defense mechanism has been discovered that is based on Clusters of Regularly Interspaced Short Palindromic Repeats (CRISPRs) and CRISPR-associated genes (*cas* genes). A timeline of major breakthroughs in CRISPR research is given in Figure 1.1. The CRISPR/Cas system can integrate nucleic acid fragments from invading mobile elements into the CRISPR locus. The CRISPR is transcribed and cleaved

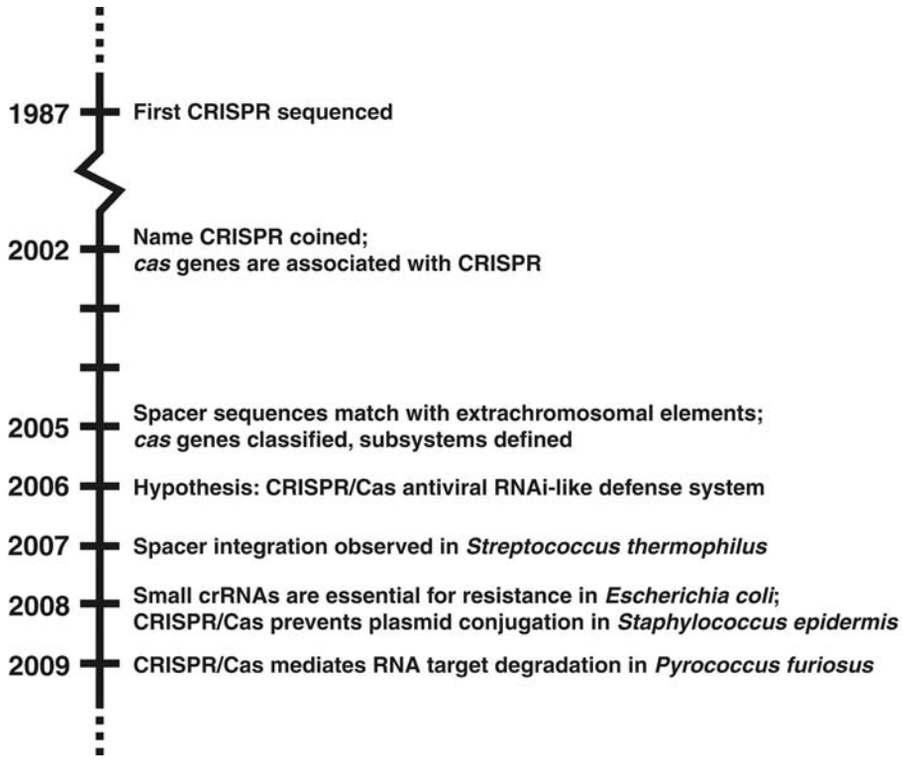


Figure 1.1. Timeline of major advances in CRISPR research. See text for references.

into short mature RNAs (crRNAs). These crRNAs specifically guide the Cas protein machinery to their complementary targets: either DNA or RNA from invading viruses or plasmids. Thus the CRISPR/Cas system can provide the host with acquired and heritable resistance (reviewed in (Sorek et al. 2008; van der Oost et al. 2009; Horvath and Barrangou 2010; Karginov and Hannon 2010; Marraffini and Sontheimer 2010a). In this chapter we will describe mechanistic features of the CRISPR/Cas system, and we will discuss the similarities and differences with RNA interference in eukaryotes.

CRISPR loci and *cas* genes

CRISPRs were first discovered in 1987 when a chromosomal fragment from *Escherichia coli* K12 was sequenced (Ishino et al. 1987). Since then many CRISPR sequences have been identified in prokaryotic genomes (for overview, see: http://crispr.u-psud.fr/crispr/CRISPRdatabase.php). CRISPRs have been detected in 48% of the sequenced bacterial genomes and in 95% of the sequenced archaeal genomes. CRISPRs are composed of a cluster of identical repetitive sequences that are separated by non-identical spacer sequences of similar length (see below). The CRISPR array is often preceded by an AT-rich leader sequence of up to 500 basepairs (Jansen et al. 2002). The number of CRISPR loci per genome ranges from 1 to 20, varying in length from a few to hundreds of repeat-spacer pairs; the present record holder is a CRISPR of *Chloroflexus* sp. with 374 repeats and spacers. Twelve major types of CRISPR have

been proposed, based on sequence similarity of the repeats (Kunin et al. 2007). The repeat size varies from 24-47 bp, whereas the spacer size ranges from 24-72 bp; the size of repeats and spacers are typically around 30 bp. Some repeats have palindromic sequences that encode CRISPR RNAs with potentially strong secondary structures, while other sequences appear to lack such structures (Fig. 1.2B). Each cluster correlates mainly with one Cas subtype, as discussed below. In 2005, three different research groups independently observed that at least a subset of the spacer sequences are identical to phage and plasmid DNA sequences (Bolotin et al. 2005; Mojica et al. 2005; Pourcel et al. 2005). The virus or plasmid fragment matching the spacer sequence is called the proto-spacer (Deveau et al. 2008). The observation that spacer sequences were derived from viral sequences has led to the hypothesis that the CRISPR/Cas system might be involved in prokaryotic resistance to alien nucleic acids (reviewed in (Makarova et al. 2006)). The composition of the CRISPR is hypervariable and is rapidly shaped by extra-chromosomal elements in the host's environment (Lillestøl et al. 2006; Andersson and Banfield 2008; Tyson and Banfield 2008; Banfield and Young 2009; Held and Whitaker 2009; Lillestol et al. 2009). Extrachromosomal elements in turn respond by extensive gene shuffling (Andersson and Banfield 2008) or mutations (Deveau et al. 2008; Heidelberg et al. 2009; Semenova et al. 2009; van der Ploeg 2009) to escape the CRISPR defense mechanism, illustrating that the ongoing battle between hosts and their predators.

A set of conserved *cas* genes can be found in close proximity of the CRISPR array. The encoded proteins were initially thought to be involved in DNA repair, because they had predicted nucleic acid related functions (Makarova et al. 2002). The link between *cas* genes and CRISPRs was made shortly thereafter (Jansen et al. 2002) and the four most conserved *cas* genes were identified. The *cas* gene products were further classified into ~45 distinct families (Haft et al. 2005); that number was later reduced to ~25 families (Makarova et al. 2006). The set of Cas proteins is composed of core proteins (Cas1-6), a diverse group of Repeat-Associated Mysterious Proteins (RAMPs) and more loosely associated Cas proteins, such as a polymerase. Based on the composition of the *cas* operons, eight Cas subtypes (Csa, Csd, Cse, Csh, Csm, Csn, Cst, Csy) by (Haft et al. 2005), or seven Cas systems (CASS1-7) have been proposed by (Makarova et al. 2006), that each contain a certain set of core proteins and a subtype specific module that in most cases contains at least one RAMP (Fig. 1.2A). An additional subtype (Cas module RAMP, Cmr) includes many RAMP proteins and a polymerase/nuclease; this system seems to share core Cas proteins with another subtype that resides on the same genome. As discussed below, the Cmr cluster at least to some extent resembles the Csm-subtype. The distribution of the closely related subtypes in phylogenetically distant

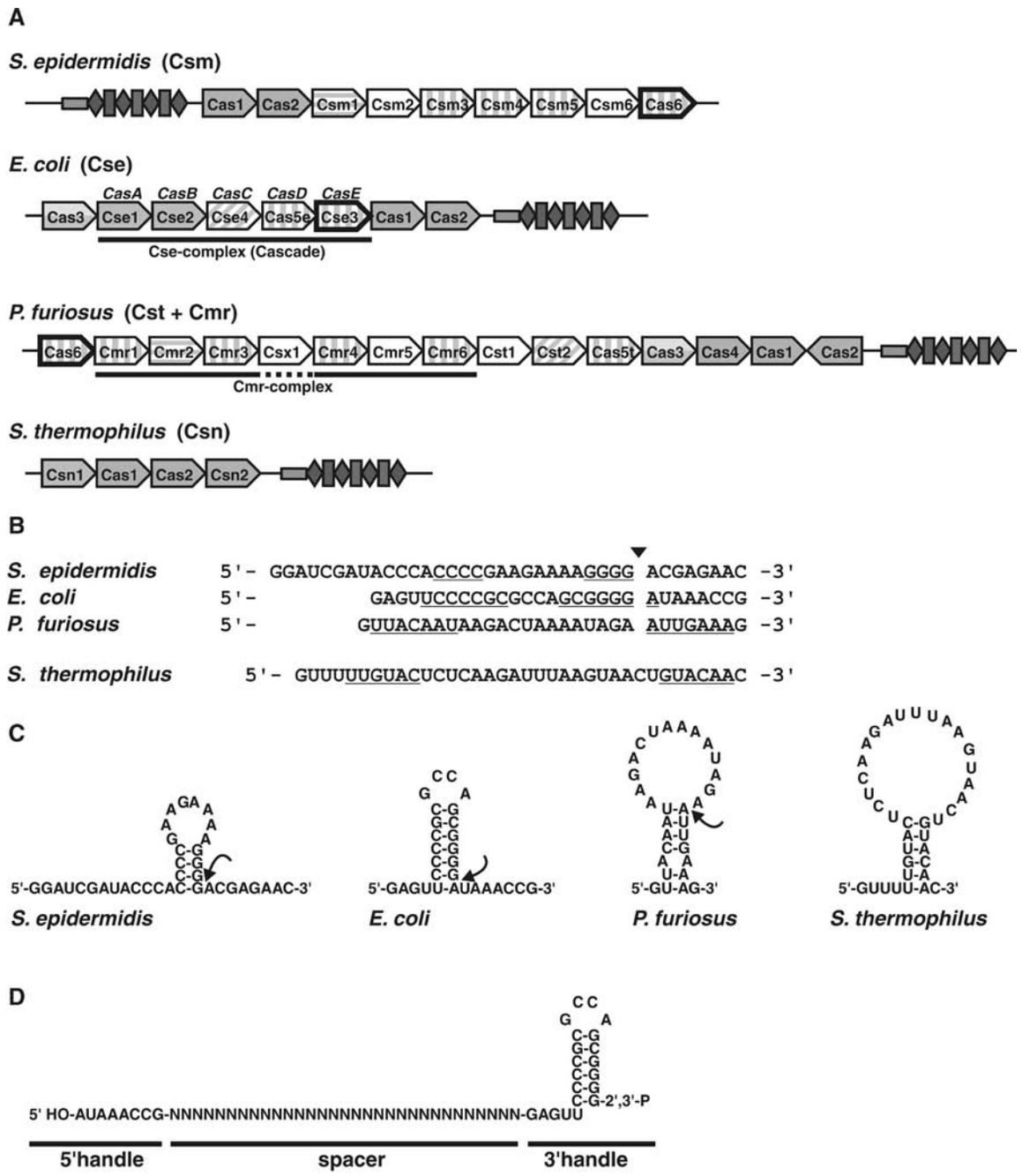


Figure 1.2. Overview of the four CRISPR/Cas subtypes that are described in this chapter. For full colour version see page 126. For an overview of all eight CRISPR/Cas subtypes, see (Haft et al. 2005; van der Oost et al. 2009). (A) *cas* gene neighborhoods in 4 experimentally studied organisms, each representing a different subtype indicated between brackets. CRISPRs consist of a leader (grey box), repeats (red diamonds) and spacers (blue boxes); only a fragment of the CRISPR is shown. Genes are indicated as arrows. Blue arrows indicate genes that are (possibly) involved in spacer acquisition. Yellow arrows indicate genes that are involved in CRISPR transcription and processing and target interference. The endonucleases that cleave pre-crRNA generating crRNA are highlighted as bold arrows. Hatching patterns indicate gene similarity: RAMP genes have vertical lines, polymerase genes have horizontal lines, CasC homologues have diagonal lines, and other genes that are not related to each other are filled. Genes that encode proteins from isolated complexes (Cse-complex from *E. coli* and Cmr-complex from *P. furiosus*) are underlined. (B) CRISPR RNA repeat sequences from each organism are given. The cleavage site is indicated by a triangle. Although the repeat sequences are different, all CRISPR RNA cleavage events generate an 8 nucleotide 5' handle. Please note that the cleavage site in *S. thermophilus* CRISPR RNA has not been determined. Palindromic sequences are underlined. (C) Predicted secondary structures of the different CRISPR RNA repeats. Cleavage sites are indicated with an arrow. As described previously by Kunin et al., the repeat of *P. furiosus* is not likely to form a stem loop (Kunin et al. 2007). (D) Mature crRNA from *E. coli* K12.

organisms, suggests that the CRISPR/Cas system has frequently been exchanged by horizontal gene transfer between distant micro-organisms (Makarova et al. 2006; Horvath et al. 2009). This hypothesis is supported by the observation that the CRISPR/Cas can be located on plasmids (e.g. the megaplasmids from *Thermus thermophilus*); other plasmids have been reported to contain a CRISPR locus without associated *cas* genes (e.g. the pNOB8 conjugative plasmid from *Sulfolobus solfataricus*) (Godde and Bickerton 2006).

The best conserved *cas* genes are *cas1* and *cas2* that are present in all subtypes (Haft et al. 2005). Therefore they are suitable markers for the presence of CRISPR/Cas. The putative nuclease/integrase Cas1 (Makarova et al. 2006) has been demonstrated to be a metal dependent nuclease that cleaves ssDNA and dsDNA, generating ~80 bp DNA fragments from dsDNA. The Cas1 structure reveals a novel fold with a two-domain architecture (Wiedenheft et al. 2009). The small Cas2 protein cleaves ssRNAs in U-rich regions. Crystal structures of Cas2 from several species have been solved, revealing a ferredoxin fold, which is not common for endoribonucleases (Beloglazova et al. 2008). Cas1 has been proposed to be involved in spacer integration (Makarova et al. 2006), a prediction that is in agreement with the observation that Cas1 and Cas2 in *E. coli* are not involved in the antiviral defense stage of the mechanism when a spacer is already present in the CRISPR array (Brouns et al. 2008; Hale et al. 2009). Fusion of *cas1* and *cas4* genes in several genomes, including that of *Geobacter sulfurreducens*, suggests that Cas4, a putative RecB-like nuclease (Makarova et al. 2006), might also be involved in spacer acquisition (van der Oost and Brouns 2009). Cas3 is a special case, typically being a single polypeptide composed of two domains: an HD domain that has metal-dependent nuclease activity on double-stranded oligonucleotides (Aravind and Koonin 1998; Han and Krauss 2009) and a DEAD/H box helicase domain (Makarova et al. 2006). Interestingly, in the Csa-subtype the domains are separated, and in the Csy-subtype Cas3 is fused to Cas2 (Makarova et al. 2006). Cas5 and Cas6, previously annotated as core Cas proteins as well, represent a group of distantly related Cas proteins referred to as RAMPs; they appear to have similar 3D structures, and share at least a C-terminal glycine-rich loop (Makarova et al. 2002). Two RAMP proteins (CasE and Cas6) have recently been demonstrated to be metal-independent endonucleases involved in the processing of CRISPR RNA (pre-crRNA), as described below (Brouns et al. 2008; Carte et al. 2008). Additionally, two types of multi-subunit Cas complexes have recently been characterized. In *E. coli* a complex is encoded by 5 clustered genes *cse1-4* and *cas5e* (Cas5e and Cse3 are RAMPs) and the gene products form a Cse-complex termed Cascade (CRISPR-associated complex for antiviral defense) (Brouns et al. 2008). A crRNA-binding Cmr-complex comprising Cmr1-6 has been isolated from

Table 1.1. Cas proteins involved in different stages. Experimentally determined activities and PDB ID codes are indicated.

Stage involved	Name	Activity	Remarks
Acquisition	Cas1	DNA endonuclease (Wiedenheft et al. 2009), RNA and DNA binding (Han et al. 2009)	Crystal structures (Wiedenheft et al. 2009) (3GOD, 2YZS)
	Cas2	Ribonuclease activity (Beloglazova et al. 2008)	Crystal structures (Beloglazova et al. 2008) (2I8E, 1ZPW, 2I0X and 2IVY)
	Cas4		RecB-like nuclease
Processing	CasE (part of Cascade), Cas6	pre-crRNA cleavage (Brouns et al. 2008; Carte et al. 2008)	Crystal structures of CasE (Ebihara et al. 2006) (1WJ9) and Cas6 (Carte et al. 2008) (3I4H)
Interference	Cse-complex (Cascade)	Strand displacement of complementary target DNA (Chapter 3)	Comprises CasA-E. Crystal structures solved of CasB (Agari et al. 2008) (2ZCA) and CasE (see above)
	Cas3	Nuclease activity of HD domain (Han and Krauss 2009)	Helicase, often fused to HD-domain (Makarova et al. 2006). Possibly also involved in acquisition, according to fusion to Cas2 in Csy-subtype (van der Oost et al. 2009)
	Cmr-complex	Cleavage of RNA complementary to crRNA (Hale et al. 2009)	Comprises Cmr1-6. Two structures of Cmr5 are available (Sakamoto et al. 2009) (2OEB and 2ZOP)

Pyrococcus furiosus (Hale et al. 2009). An overview of experimentally determined and putative activities and structures of core Cas proteins and Cas complexes is provided in Table 1.1.

Mode of action

The CRISPR/Cas mechanism can be divided into three distinct stages. The first stage concerns the integration of nucleic acid fragments of invading mobile genetic elements as new spacers into the CRISPR locus. In the second stage, the CRISPR is transcribed as a precursor (pre-crRNA), which is subsequently cleaved by a dedicated endoribonuclease, resulting in mature crRNAs that remain associated with a Cas protein complex. During the third and final stage the crRNA guides the Cas complex to known, invading nucleic acids to neutralize the invader, most likely by cleavage.

1. Integration of new spacers

The first experimental evidence that the CRISPR/Cas system is indeed an antiviral defense system was obtained from phage infection experiments of the lactic acid bacterium *Streptococcus thermophilus* that has a CRISPR/Cas locus of the Csn-subtype (Fig. 1.2A) (Barrangou et al. 2007; Deveau et al. 2008; Horvath et al. 2008). Screening for adaptation of the CRISPR locus in the surviving bacteria revealed that

a sub-population of survivors had acquired new phage-specific spacer sequences (Fig. 1.3). Subsequent deletion of these new spacer sequences resulted in loss of the acquired resistance, demonstrating the correlation of spacer presence and phage resistance (Barrangou et al. 2007). Comparative analysis of the spacer-targeted region of the viral genome revealed a sequence motif called CRISPR motif or proto-spacer adjacent motif (PAM) (i.e. NNAGAAW) downstream of the proto-spacer. The phages responded to the resistance of the host by mutations in the proto-spacer, but also by mutations in the motif (Deveau et al. 2008), illustrating the constant battle between phages and bacteria. Spacers that did not have a perfect PAM were integrated as well, but these were always accompanied by spacers with a perfect PAM (Deveau et al. 2008), the latter probably being essential for resistance. The data suggests that the PAM is crucial for target interference, but not for integration of new spacers. Genomic analysis of proto-spacers revealed the presence of slightly different PAMs in many extrachromosomal elements, and the conservation of PAM sequences correlates well with the CRISPR repeat types, and thus cas gene subtypes (Mojica et al. 2009).

The spacers in *S. thermophilus* were integrated at the leader proximal end of the CRISPR locus, suggesting that a CRISPR array is a chronological record of past phage infections. The role of the leader sequence is not exactly known, but it is possibly required for repeat duplication and/or spacer integration. Polarity of the CRISPR integration had been previously predicted by Lillestol and coworkers after comparing the CRISPRs of two *Sulfolobus solfataricus* strains with major variation at the leader-side of the CRISPR (Lillestøl et al. 2006). Next to acquired resistance by spacer integration in *S. thermophilus*, the above mentioned study also showed that *cas* genes are involved in CRISPR-mediated defense. Disruption of the *csn1* gene resulted in loss of viral resistance. An interrupted c*sn2* gene did not lead to loss of resistance, but rather to a disrupted ability to integrate new spacers. This does not only show that both

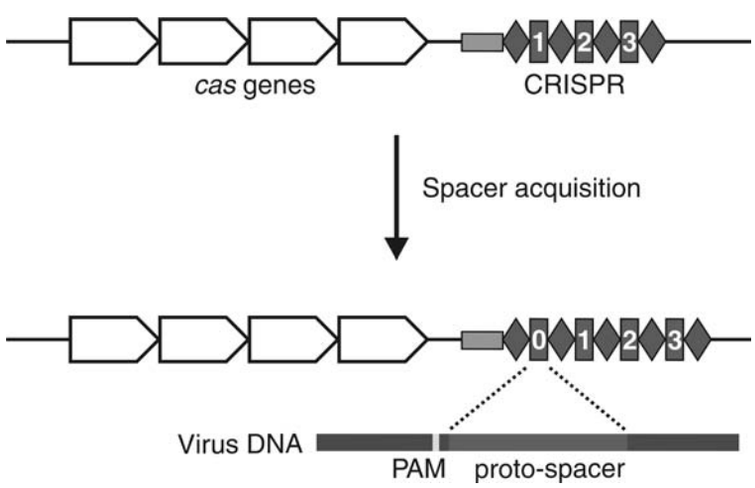


Figure 1.3. Integration of a new spacer. For full colour version see page 127. A new spacer is acquired at the leader proximal side of the CRISPR during virus infection, resulting in resistance. The CRISPR consist of a leader (grey box), repeats (red diamonds) and spacers (blue boxes). The newly acquired spacer is numbered 0 and matches the sequence of the virus (proto-spacer). The protospacer adjacent motif (PAM) is located upstream or downstream the protospacer.

cas genes are involved in resistance, but also that they play a role at different stages. *Csn2* is apparently needed for integration of new spacers, but it is restricted to the Csn-subtype. In *Wollinella succinogenes*, which carries the same subtype, *csn2* is absent and seems to be replaced by *cas4* (van der Oost and Brouns 2009). Thus, despite the unknown role of Csn2 in spacer integration, it is likely that it is replaced by Cas proteins with analogous functions in other Cas subtypes, Cas4 being a likely candidate. Csn1 may have a function analogous to Cascade/Cas3 (as discussed below, Cas3 is also involved in target interference)(van der Oost and Brouns 2009).

2. CRISPR transcription and processing

Before it was hypothesized that the CRISPR/Cas system protects the host against invading nucleic acids, studies on short RNAs in the archaea *Archaeoglobus fulgidus* and *S. solfataricus* had already indicated that CRISPRs are actively transcribed and the pre-crRNA processed (Tang et al. 2002; Tang et al. 2005). Although studies on transcription and processing in *S. solfataricus* showed that transcription can be bidirectional (Lillestøl et al. 2006; Lillestol et al. 2009), most other studies have reported unidirectional transcription from the leader proximal side(Brouns et al. 2008; Hale et al. 2008; Marraffini and Sontheimer 2008; Semenova et al. 2009). It is therefore anticipated that in most cases a single promoter at the leader side controls transcription of a CRISPR locus. It has recently been shown that in the case of *E. coli* K12 the promoter indeed resides in the leader region (Pul et al. 2010). CRISPR transcription and processing has been studied in more detail for this *E. coli* K12 system, which consists of 8 *cas* genes upstream of a CRISPR (Fig. 1.2A). The Cas proteins were overexpressed in *E. coli* BL21, a strain that lacks endogenous *cas* genes (Studier et al. 2009). Pull down analysis revealed the presence of Cascade, a protein complex that contains 5 different subunits, CasABCDE. Northern blot analysis of crRNAs in *E.coli* K12 revealed short RNAs, the size of which corresponded to approximately 1 spacer and 1 repeat. Omitting the *cas* genes one by one identified CasE, a RAMP protein, as a potential candidate for pre-crRNA processing. This finding was confirmed by *in vitro* activity assays showing that CasE is a metal-independent endoribonuclease that specifically cleaves a precursor (pre-crRNA) into mature crRNAs. After processing, these crRNAs remain tightly bound to the Cascade complex. Cloning and subsequent sequence analysis revealed more crRNA products derived from the leader proximal end of the CRISPR (Brouns et al. 2008). Independent analysis of crRNA from *P. furiosus* also uncovered more products from the leader proximal side of the CRISPR, and fewer from the distal end. This can be explained by premature termination of CRISPR transcription (Hale et al. 2008). Mass spectrometry analysis of the *E. coli* crRNA products revealed that the product contains 8 nucleotides of the repeat termed

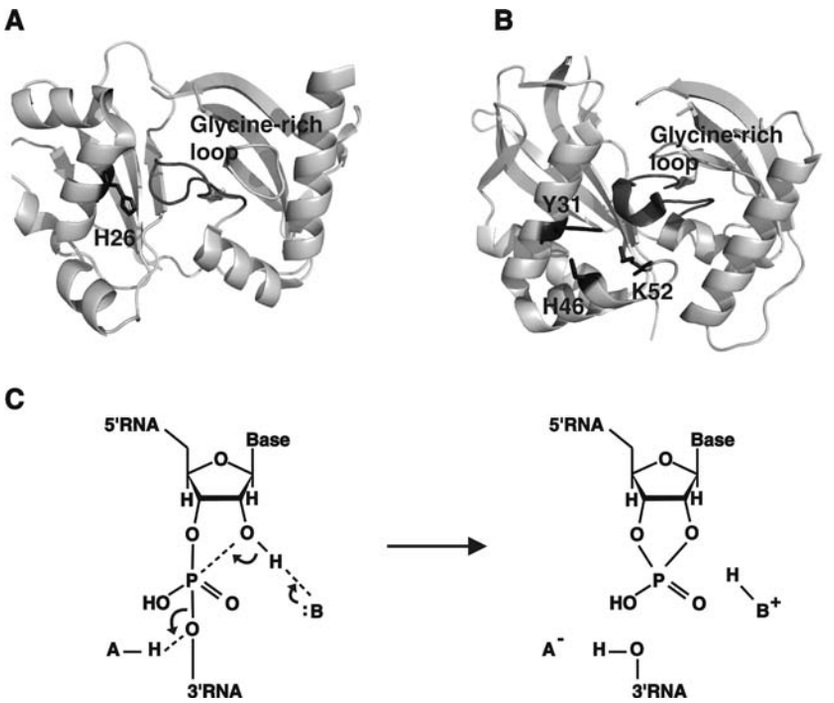


Figure 1.4. The catalytic sites of CasE and Cas6, and the proposed reaction mechanism of pre-crRNA cleavage. For full colour version see page 127. (A) Proposed catalytic site of CasE from *T. thermophilus* showing the conserved histidine residue (H26) and the glycine-rich C-terminal loop. The catalytic site of Cas6 from *P. furiosus* (B) contains a catalytic triad of tyrosine (Y31), histidine (H46) and lysine (K52) and a glycine-rich C-terminal loop. The loop and the overall duplicated ferredoxin fold are conserved among CasE and Cas6. Pre-crRNA cleavage might follow a general acid-base hydrolysis mechanism (D). A base (B) draws a proton from the 2'OH of the ribose ring. A subsequent nucleophilic attack on the phosphorus atom is simultaneously compensated by the acid (A) that donates a proton to the leaving 3'RNA. The tyrosine residue of Cas6 is proposed to be the base and the histidine the acid residue (Carte et al. 2008). In CasE the histidine and a water molecule might be the catalytic residues. Pictures in (A) and (B) are generated with pymol (www.pymol.org), potential catalytic residues are depicted in blue; the glycine-rich loop is depicted in red. Coordinates were obtained from the Protein Data Bank (www.pdb.org).

the 5' handle, the spacer, and a large part of the next repeat including the stem-loop termed the 3' handle (Brouns et al. 2008) (Fig. 1.2D). It has been proposed that these handles, the conserved parts of the crRNAs, are bound by subunits of Cascade (Brouns et al. 2008). Based on sequence conservation in CasE and the crystal structure of a CasE homolog from *Thermus thermophilus* (Ebihara et al. 2006) the histidine residue at position 20 has been predicted to be a residue involved in endonuclease activity. Indeed, the activity was lost when the histidine was substituted by an alanine. *In vivo* analysis showed that the mutation resulted in loss of resistance, thus showing that pre-crRNA cleavage is a mechanistic requirement.

Despite the fact that Cas6 from *P. furiosus* shares low sequence identity (except for the C-terminal glycine-rich loop, a common feature of RAMP proteins), its structure, a duplicated ferredoxin fold, is surprisingly similar to CasE from *T. thermophilus* (Fig. 1.4) (van der Oost et al. 2009). Like CasE, Cas6 displays metal independent

endoribonuclease activity (Carte et al. 2008). Although the folded secondary structure of the repeat RNA of *P. furiosus* is debatable (Kunin et al. 2007), its crRNA cleavage product also contains an 8 nucleotide 5' handle (psi-tag; Fig. 1.2B and C). Unlike the proposed single histidine site in CasE (Brouns et al. 2008), a potential catalytic triad with a histidine, tyrosine and a lysine residue is present in Cas6 (Fig. 1.4) (Carte et al. 2008). This predicted catalytic site is structurally similar to that of tRNA splicing enzymes. Analysis of the cleaved pre-crRNA products revealed that cleavage occurs at the 3' side of the phosphodiester bond, generating a 5' end hydroxyl group and a 2', 3' end cyclic phosphate group, analogous to tRNA splicing enzymes (Calvin and Li 2008; Carte et al. 2008). Like the tRNA splicing mechanism, recent analysis has revealed that crRNA from *E. coli* also possess a 2',3'-cyclic phosphate (Chapter 3), as was predicted (Calvin and Li 2008; Carte et al. 2008). Both CasE and Cas6 probably cleave the pre-crRNA following a general acid-base hydrolysis mechanism (Fig. 1.4C). Besides its endoribonucleolytic activity Cas6 binds the 3' handle of the crRNA. Unlike the situation in *E. coli*, in *P. furiosus* the endonucleolytic product is further trimmed to active mature crRNAs lacking a 3' handle (Hale et al. 2008; Hale et al. 2009), whereas mature crRNAs in *E. coli* do contain a 3' handle. Another difference is the fact that CasE remains part of the Cascade complex (Brouns et al. 2008) whereas Cas6 has not been identified as part of the pyrococcal Cmr-complex (see below) (Hale et al. 2009); apparently cleaved crRNAs in *P. furiosus* are transferred to the Cmr-complex. Besides the Cmr-complex, an additional Cst-complex might be present in *P. furiosus*, that is encoded by three *cst* genes downstream of the *cmr* module (Fig. 1.2A). Overall, it has become clear that despite some mechanistic similarity, also substantial differences exists between the different CRISPR/Cas subtypes.

3. Target interference

Although it was initially hypothesized that the CRISPR/Cas system would target alien RNA, analogous to RNAi (Makarova et al. 2006), several studies have indicated that the target of the CRISPR system rather is invading DNA. The first observation supporting this hypothesis was made in virus infection studies with *Streptococcus thermophilus*, revealing that spacer sequences corresponding to either the coding or non-coding strand were integrated (Barrangou et al. 2007). The CRISPR locus has been demonstrated to be transcribed only from the leader proximal side in *E. coli*, *S. epidermidis* and *P. furiosus* (Brouns et al. 2008; Hale et al. 2008; Marraffini and Sontheimer 2008); to date the only exception appears to be the case of *S. solfataricus* (Lillestol et al. 2006; Lillestol et al. 2009). A consequence of mono-directional CRISPR transcription would be that generated crRNAs have to be complementary to the mRNA of the virus, only spacers from one strand should be incorporated in case of an antisense RNA mechanism.

The observation that this is generally not the case suggests that DNA is the target. Additional evidence that DNA is the target was provided by a study on an engineered *E. coli* Cas system in which artificial spacers were unidirectionally transcribed. Generating crRNAs complementary to both the coding strand and the template strand were successful in inhibiting virus proliferation (Brouns et al. 2008). Furthermore, a study on plasmid conjugation in *Staphylococcus epidermidis* convincingly proved that DNA is being targeted (Marraffini and Sontheimer 2008). A natural spacer from the CRISPR of *S. epidermidis* has a perfect match with a gene of a conjugative plasmid. This spacer confers resistance and prevents conjugation of this plasmid. A self-splicing intron was inserted into the center of the proto-spacer sequence in the plasmid. After conjugation and subsequent transcription this intron is spliced, generating a mature mRNA that contains a fully complementary sequence to the crRNA. The plasmid was able to escape the CRISPR/Cas system showing that mRNA is not being targeted; hence, the target must be DNA that could not be recognized by the crRNA due to the intron DNA sequence interrupting the proto-spacer. This finding is confirmed by an additional experiment in which a fragment of the targeted gene is inserted in a plasmid, in both orientations. The fragment is not essential for propagation of the plasmid. Nevertheless, the inserted fragment, in both orientations, dramatically decreased the transformation efficiencies, again indicating that DNA is being targeted. Recently an exceptional CRISPR/Cas system has been described, i.e. the Cmr-subtype from *P. furiosus* that appears to target and degrade RNA (discussed below). For this system however, no natural targets have been found yet.

Most biochemical and mechanistic information on the pre-crRNA and target interference is derived from the above mentioned *E. coli* and *P. furiosus* model systems. Requirements for CRISPR-based resistance were determined in *E. coli* BL21. Artificial CRISPRs were designed to target four different genes of phage lambda. Overexpression of Cascade and CRISPR RNA was sufficient to yield mature crRNAs that were bound and protected by Cascade (Brouns et al. 2008). No resistance was observed when the CRISPRs targeting phage lambda were co-expressed with Cascade only. However, co-expression of this crRNA-loaded Cascade with the Cas3 protein did result in a dramatic increase of resistance towards phage lambda infection. A recent study revealed that all the Cascade subunits are essential for immunity (Chapter 3). The mechanism of target recognition was studied *in vitro* by use of Electrophoretic Mobility Shift Assays (EMSAs) (Chapter 3). When loaded with crRNA, Cascade or CasCDE (a minimal core that is still capable of processing, binding and protecting crRNA) (Chapter 3) was shown to bind complementary ssDNA and ssRNA. Interestingly, the complexes are also able to interact with complementary dsDNA, but not dsRNA. The observation that only the

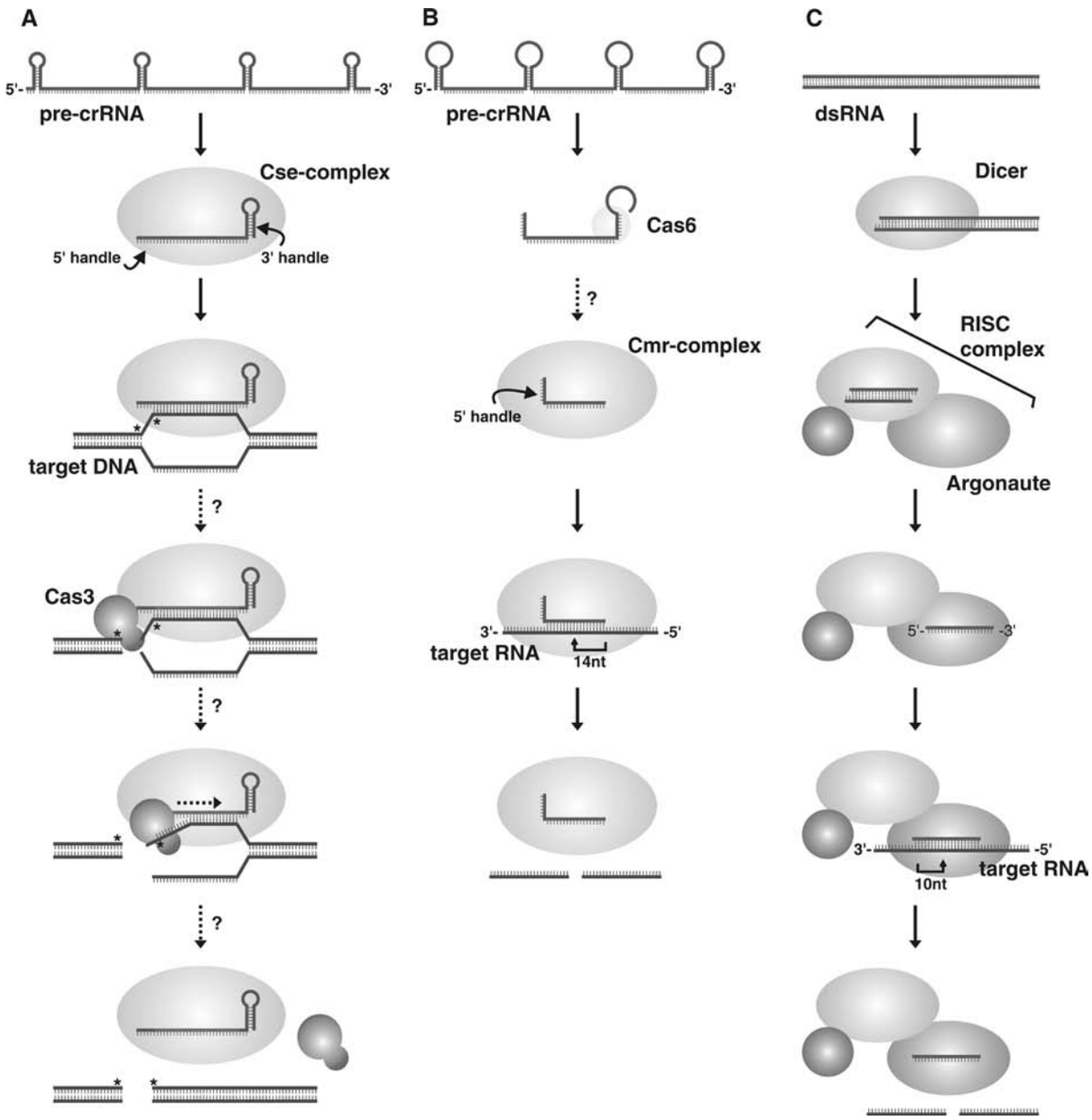


Figure 1.5. Antiviral DNA and RNA silencing pathways in prokaryotes and eukaryotes. For full colour version see page 128. (A) crRNA mediated DNA silencing pathway in *E. coli*. pre-crRNA is cleaved by the CasE subunit of Cascade (Cse-complex) and the mature crRNA remains bound to Cascade. When encountering viral dsDNA containing a sequence identical to the spacer sequence of the crRNA, it may basepair with the complementary DNA strand by a strand displacement event. The HD-domain of Cas3 is likely to be activated and cleave the viral DNA only when the 2 bases PAM on the viral DNA is present (marked with an asterisk). The helicase domain might subsequently separate the RNA:DNA duplex generating free Cascade that can be used in a next cleavage event. (B) crRNA mediated RNA silencing pathway in *P. furiosus*. Pre-crRNA is cleaved by Cas6 and then further trimmed to generate crRNAs of two different lengths. These crRNAs are bound by the Cmr-complex. This loaded Cmr-complex specifically binds viral RNA and cleaves the complementary strand 14 nucleotides away from the 3' end of the crRNA. This pathway shares functional analogies with siRNA mediated antiviral resistance in eukaryotes (C) siRNAs are generated from viral dsRNA by dicer. The first (random) cleavage event by dicer generates dsRNA with a 3' dinucleotide overhang. The second cleavage by dicer takes place 20-25 bases away from the overhang generating short dsRNAs. The dsRNA is transferred to the Argonaute protein of the RISC complex and the passenger strand is removed. The retained guide strand can basepair with a complementary viral mRNA molecule, followed by a cleavage of the scissile bond between the 10th and 11th base from the 3' end of the guide strand. The cleaved target RNA dissociates and the recycled RISC can be used in a second round of RNA binding and cleavage. Please note that dashed arrows indicate processes that are based on hypotheses.

complementary strand remains bound to Cascade suggests this dsDNA binding occurs by local strand displacement. The ability to specifically target dsDNA indicates that the Cas machinery can immediately attack the incoming dsDNA from phage Lambda. The *in vitro* DNA binding is enhanced in the presence of CasA. In conclusion, Cascade loaded with crRNA recognizes and binds the incoming double stranded phage DNA. The role of the two-domain protein Cas3 remains to be elucidated, but it is tempting to speculate that the HD nuclease domain cleaves targeted DNA (Fig. 1.5A). This Cse-subtype from *E. coli* K12 targets DNA, but how does it discriminate between self DNA (namely the CRISPR that contains a spacer complementary to the crRNA) and non-self DNA (the invader)? A recent study in *S. epidermidis* revealed that the flanking sequence of the proto-spacer is crucial for self versus non-self DNA recognition, and thus for interference (Marraffini and Sontheimer 2010b). It was shown that a targeted conjugative plasmid could bypass host resistance by the CRISPR/Cas system if three bases downstream of the proto-spacer were complementary to the CRISPR repeat sequence (Marraffini and Sontheimer 2010b). In other subtypes however, the PAM might help prevent auto-immunity by being present in the targeted invading DNA sequence, and absent from the CRISPR DNA sequence. The PAM determines if invading nucleic acids are being targeted in the case of *S. thermophilus*, as can be deduced from phages that have mutated their PAM sequence and thus can escape the CRISPR/Cas system (Deveau et al. 2008). The two mechanisms of self versus non-self DNA recognition appear to be fundamentally different. In *S. epidermidis* the potential of the downstream sequence of the proto-spacer to basepair with the CRISPR repeat determines whether the DNA is being targeted, while in *e.g. S. thermophilus* the PAM determines whether the DNA is being targeted. To bypass immunity in *S. epidermidis* the invader can only mutate its flanking nucleic acids to a sequence complementary to the repeat DNA, while in *S. thermophilus* the invader can mutate its PAM to any other sequence.

Interestingly, the Cas-system from *P. furiosus* has recently been reported to be capable of interfering with target RNA rather than DNA (Hale et al. 2009). With native Northern blot analysis a Cmr-type protein complex was identified that forms a stable interaction of crRNAs (psiRNA) (Hale et al. 2008; Hale et al. 2009). The crRNAs are 39 and 45 nucleotides in length, containing an identical 5' handle but different 3' ends. The RNP complex isolated from *Pyrococcus* comprised 6 distinct Cas proteins: Cmr1-6. Cas6 was not part of the complex (Hale et al. 2009), unlike its functional analog CasE that is a core subunit of the Cascade complex (Cse-complex) in *E. coli* (Brouns et al. 2008). The isolated RNP complex cleaved complementary RNA but not ssDNA. The 39 and 45 nucleotide long crRNAs resulted in two different cleavage sites in the target mRNA; both cleaved 14 nucleotides upstream from the 3'end of the crRNA,

suggesting a molecular ruler mechanism for cleavage (Fig. 1.5B). Reconstitution of the RNP complex from purified subunits revealed that all except Cmr5 are essential for target RNA cleavage (Hale et al. 2008; Hale et al. 2009). The Cmr2 protein that is part of the Cmr-complex contains a PALM (polymerase) domain, fused to a HD-nuclease domain. Whereas the biological function of the polymerase is not known, the HD-nuclease might be responsible for degradation of the target RNA. At least two proteins from the Cmr-complex are related to proteins that are encoded by the *csm* module in *S. epidermidis* (Fig. 1.2A); the polymerase Cmr2 is related to Csm1 and the RAMP protein Cmr4 is related to Csm3 (Haft et al. 2005). The Csm-type system from *S. epidermidis* has been demonstrated to target DNA *in vivo* which is in contrast to the observed *in vitro* RNA cleavage activity of the Cmr-complex. Future biochemical and *in vivo* analyses are required to resolve this apparent contradiction.

Analogy with RNAi in eukaryotes

The function and biogenesis of crRNAs and the mechanism of target interference display striking analogies with small regulatory RNAs in eukaryotes. The eukaryotic regulatory RNAs can be divided in three major groups; endogenous miRNAs that generally silence host gene expression, piRNAs that silence transposable elements in animal germ cells, and siRNAs that can be involved in viral RNA silencing (see elsewhere in this book for more details on eukaryotic small RNAs). In short, siRNAs are derived from dsRNA that is randomly cleaved by Dicer generating fragments with a 3' dinucleotide overhang. Dicer subsequently binds the overhang and cleaves ~20 bases away from the first cleavage site following a molecular ruler mechanism, generating short dsRNAs with 3' overhangs and 5' phosphates (Bernstein et al. 2001; Macrae et al. 2006; MacRae and Doudna 2007). These dsRNAs are transferred to the Argonaute protein in the RNA induced silencing complex (RISC) that also consists of Dicer and a dsRNA binding protein; the latter determines in which orientation the dsRNA molecule is loaded onto Argonaute (Tomari et al. 2004). Argonaute recognizes the passenger strand and degrades it, retaining the guide strand (Rand et al. 2005). When RISC encounters a perfectly complementary RNA it can interact with it by Watson-Crick base pairing. Argonaute subsequently cleaves the complementary target RNA strand, 10 nucleotides away from the 5' end of the guide, after which the target fragments are released and the RISC complex is recycled for a new target degradation event, eventually resulting in silencing the virus (Baulcombe 2004) (Fig. 1.5C).

Although sequence comparison of the prokaryotic CRISPR/Cas and the eukaryotic siRNA system indicates they are phylogenetically unrelated (Makarova et al. 2006), they share some functional and mechanistic analogies. Both crRNA and siRNA are derived

from large RNA precursors. Furthermore, both the Cmr-complex from *P. furiosus* and RISC complex specifically bind their target RNA by base pairing and degrade it by a molecular ruler mechanism. The major difference between the two systems is that crRNAs are transcribed from the host chromosome and are part of an adaptive immune system, while antiviral siRNAs are derived from the invaders and are part of an innate immune system. Moreover, the Cse- and Csm-type CRISPR/Cas systems target DNA rather than RNA. The RNAi pathways in which Dicer and Argonaute are the key players are restricted to eukaryotes. The only conserved component among prokaryotes and eukaryotes is the Argonaute protein. The ones from prokaryotes have been useful models in crystallization studies (Jinek and Doudna 2009; Wang et al. 2009). It has recently been hypothesized that prokaryotic Argonaute proteins are involved in yet another DNA-mediated antiviral defense mechanism (Makarova et al. 2009).

Chapter 2

Small CRISPR RNAs guide antiviral defense in prokaryotes

Matthijs M. Jore*, Stan J. J. Brouns*, Magnus Lundgren, Edze R. Westra, Rik J. H. Slijkhuis, Ambrosius P. L. Snijders, Mark J. Dickman, Kira S. Makarova, Eugene V. Koonin, John van der Oost

*Contributed equally

Science. 2008 Aug 15; **321**(5891): 960-4.

Abstract

Prokaryotes acquire virus resistance by integrating short fragments of viral nucleic acid into clusters of regularly interspaced short palindromic repeats (CRISPRs). Here we show how virus-derived sequences contained in CRISPRs are used by CRISPR-associated (Cas) proteins from the host to mediate an antiviral response that counteracts infection. After transcription of the CRISPR, a complex of Cas proteins termed Cascade cleaves a CRISPR RNA precursor in each repeat and retains the cleavage products containing the virus-derived sequence. Assisted by the helicase Cas3, these mature CRISPR RNAs then serve as small guide RNAs that enable Cascade to interfere with virus proliferation. Our results demonstrate that the formation of mature guide RNAs by the CRISPR RNA endonuclease subunit of Cascade is a mechanistic requirement for antiviral defense.

Small CRISPR RNAs guide antiviral defense in prokaryotes

The clusters of regularly interspaced short palindromic repeat (CRISPR)–based defense system protects many bacteria and archaea against invading conjugative plasmids, transposable elements, and viruses (Bolotin et al. 2005; Mojica et al. 2005; Pourcel et al. 2005; Godde and Bickerton 2006; Lillestøl et al. 2006; Barrangou et al. 2007; Sorek et al. 2008; Tyson and Banfield 2008). Resistance is acquired by incorporating short stretches of invading DNA sequences in genomic CRISPR loci (Barrangou et al. 2007; Deveau et al. 2008; Horvath et al. 2008). These integrated sequences are thought to function as a genetic memory that prevents the host from being infected by viruses containing this recognition sequence. A number of CRISPR-associated (*cas*) genes (Jansen et al. 2002; Haft et al. 2005; Makarova et al. 2006) has been reported to be essential for the phage-resistant phenotype (Barrangou et al. 2007). However, the molecular mechanism of this adaptive and inheritable defense system in prokaryotes has remained unknown.

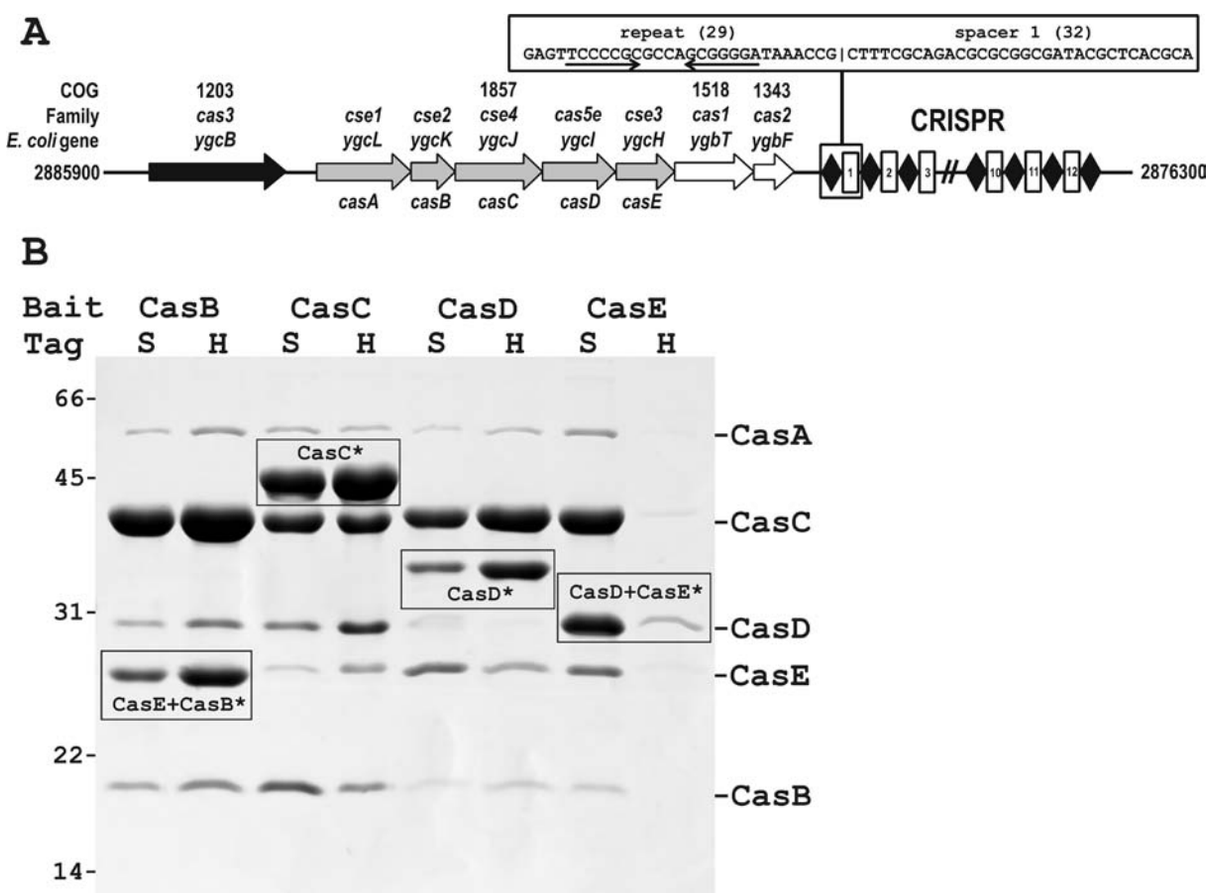


Figure 2.1. The composition of the Cascade complex. (A) Schematic diagram of the CRISPR/*cas* gene cluster of *E. coli* K12 W3110. Repeats and spacers are indicated by diamonds and rectangles, respectively. A palindrome in the repeat is marked by convergently pointing arrows. Protein family nomenclature is as described in (Jansen et al. 2002; Haft et al. 2005). (B) Coomassie blue–stained SDS-polyacrylamide gel of the affinity purified protein complex using either the N-terminal StrepII-tag (S) or C-terminal His-tag (H) of each of the subunits CasB, CasC, CasD, or CasE as bait. Asterisks indicate the 5.5 kD larger double-tagged subunits. Marker sizes in kilodaltons on the left; location of untagged subunits on the right.

The *Escherichia coli* K12 CRISPR/*cas* system comprises eight *cas* genes: *cas3* (predicted HD-nuclease fused to a DEAD-box helicase), five genes designated *casABCDE*, *cas1* (predicted integrase) (Makarova et al. 2006), and the endoribonuclease gene *cas2* (Beloglazova et al. 2008) (Fig. 2.1A and Table S2.1). In separate experiments, each Cas protein was tagged at both the N and C terminus and produced along with the complete set of untagged Cas proteins. Affinity purification of the tagged component

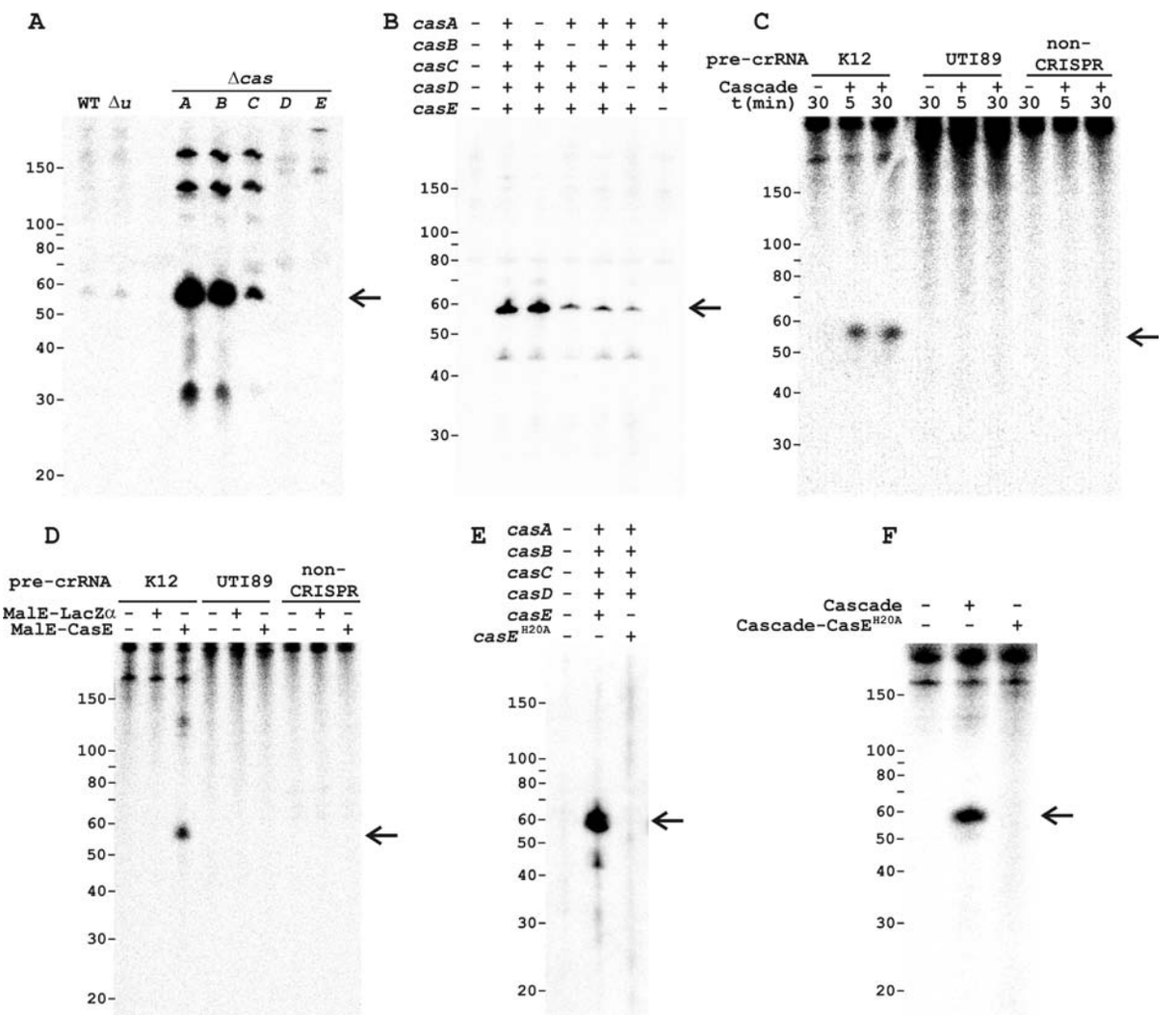


Figure 2.2. Cascade cleaves CRISPR RNA precursors into small RNAs of ~57 nucleotides (marked by arrows). (A) Northern analysis of total RNA of WT *E. coli* K12 (WT), a non-*cas* gene knockout (Δ*u*, *uidA*, β-glucuronidase), and Cascade gene knockouts using the singlestranded spacer sequence BG2349 (table S2.2) as a probe. (B) Northern blot as in (A) of total RNA from *E. coli* BL21 (DE3) expressing the *E. coli* K12 pre-crRNA and either the complete or incomplete Cascade complex. (C) Activity assays with purified Cascade using *in vitro* transcribed γ-^{32}P-uridine triphosphate-labeled pre-crRNA from *E. coli* K12 (repeat sequence: GAGUUCCCCGCCAGCGGG-GAUAAACCG), *E. coli* UTI89 (repeat sequence: GUUCACUGCCGUACAGGCAGCUUAGAAA), and non-crRNA as substrates. (D) Activity assays as shown in (C) for 15 min with purified MalE-LacZα and MalE-CasE fusion proteins. (E) Northern blot as shown in (B) with Cascade or Cascade-CasE^{H20A}. (F) Activity assays as shown in (C) for 30 min with purified Cascade or Cascade-CasE^{H20A}.

enabled the identification of a protein complex composed of five Cas proteins: CasA, CasB, CasC, CasD, and CasE (Fig. 2.1B). The complex, denoted Cascade (CRISPR-associated complex for antiviral defense), could be isolated from *E. coli* lysates using any of the tagged subunits of the complex as bait, except for CasA.

The function of Cascade was studied by analyzing the effect of in-frame *cas* gene knockouts (Baba and Mori 2008) on the formation of transcripts of the CRISPR region in *E. coli* K12 (Fig. 2.1A). Northern analysis of total RNA with single-stranded spacer sequences as a probe showed transcription of the CRISPR region in the direction downstream of the *cas2* gene (Figs. 2.1A and 2.2A) and no transcription in the opposite direction. Analysis of control strains (wild type and a non-cas gene knockout) revealed a small CRISPR-RNA (crRNA) product of ~57 nucleotides (Fig. 2.2A). The same product was present in much higher amounts in the *casA*, *casB*, and *casC* knockout strains but absent from strains lacking the overlapping genes *casD* and *casE* (Fig. 2.2A). The small crRNAs seem to be cleaved from a multiunit crRNA precursor (pre-crRNA) (Tang et al. 2002; Tang et al. 2005; Lillestøl et al. 2006), as is evident from the presence of two and three repeat-spacer units (~120 and ~180 nucleotides) that show up in the Δ*casA*, Δ*casB*, and Δ*casC* strains (Fig. 2.2A). The Δ*casE* strain contained a large pre-crRNA, suggesting that the disruption of this gene prevents pre-crRNA cleavage.

To study the accumulation and cleavage patterns of crRNAs in the *E. coli* K12 knockout strains in more detail and to rule out any effects of the gene disruptions on the expression of downstream or upstream *cas* genes, the five subunits of Cascade and the K12-type pre-crRNA were expressed in *E. coli* BL21(DE3), which lacks endogenous *cas* genes (Kim et al.). Northern analysis showed that crRNAs of ~57 nucleotides were only produced in strains containing the Cascade complex (Fig. 2.2B). By omitting the individual subunits one by one, it became apparent that the small crRNA was absent only in the strain that lacked *casE* (Fig. 2.2B), indicating that this is the only Cascade subunit essential for pre-crRNA cleavage.

Activity assays with purified Cascade showed that the complex is capable of cleaving the E. coli K12 pre-crRNA into fragments of ~57 nucleotides in vitro (Fig. 2.2C). However, no cleavage was observed with either pre-crRNA from E. coli UTI89, which contains repeats with a different sequence (Kunin et al. 2007), or a non-crRNA template (Fig. 2.2C). The RNA cleavage reaction proceeded in the absence of divalent metal ions and adenosine triphosphate and reached saturation level within 5 min. To investigate whether the CasE subunit is sufficient for pre-crRNA cleavage activity, it was overproduced as a fusion with the *E. coli* maltose binding protein (MalE). Like the complete Cascade,

the CasE fusion protein cleaved only the K12-type pre-crRNA (Fig. 2.2D), showing that CasE is an unusual endoribonuclease that does not require the other Cascade subunits. We cannot rule out the possibility that pre-crRNA cleavage is an autocatalytic, ribozyme-like reaction, in which CasE is an essential RNA chaperone.

CasE belongs to one of the numerous families of repeat-associated mysterious proteins, the largest and most diverse class of Cas proteins (Haft et al. 2005; Makarova et al. 2006). The crystal structure of a CasE homolog from *Thermus thermophilus* HB8 shows that the protein contains two domains with a ferredoxin-like fold, and displays overall structural similarity to a variety of RNA-binding proteins (Ebihara et al. 2006; Makarova et al. 2006). On the basis of structure and amino acid conservation analysis of this protein family (Fig. S2.1), the invariant residue His^{20} was mutated to Ala to analyze the effect on pre-crRNA cleavage. Northern blots indicated that crRNAs of ~57 nucleotides were no longer formed in the strain containing Cascade-$CasE^{H20A}$ (Fig. 2.2E). Moreover, although the mutated CasE was still incorporated into Cascade, the pre-crRNA cleaving ability of purified Cascade was abolished (Fig. 2.2F), providing further support for the essential role of CasE in pre-crRNA cleavage and suggesting that the conserved His residue is involved in catalysis.

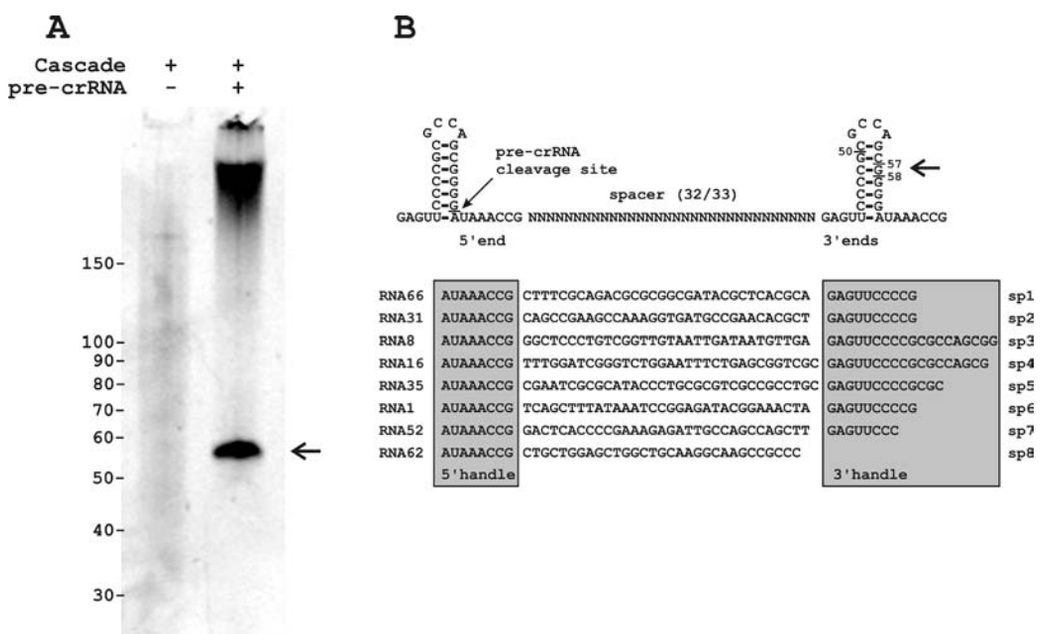


Figure 2.3. Cleaved crRNAs remain bound by Cascade. (A) Denaturing polyacrylamide gel showing the crRNA (marked by the arrow) isolated from purified Cascade in the absence and presence of co-expressed pre-crRNA. (B) Secondary structure of pre-crRNA repeats and example sequences of cloned crRNAs indicating the PCS and crRNA handles.

The crRNA cleavage sites were examined by simultaneous expression of K12-type pre-crRNA and Cascade. Under these conditions, the purification of Cascade yielded substantial amounts of copurified RNAs of ~57 nucleotides (Fig. 2.3A). Cloning and sequencing of this Cascade-bound RNA revealed that 85% of the clones [67 out of 79 clones (67/79)] were derived from crRNAs, of which 78% (52/67) started with the last eight bases of the repeat sequence (AUAAACCG) (Fig. 2.3B and Fig. S2.2). This well-defined 5′ end was followed by a complete spacer sequence and a less well-defined 3′ sequence ending in the next repeat region. A transcript of a single palindromic repeat can fold as a stable stemloop of seven base pairs, which may facilitate recognition by RNA-binding Cas proteins (Kunin et al. 2007; Sorek et al. 2008), such as CasE. The pre-crRNA cleavage site (PCS) appeared to be located immediately upstream of the 3′ terminal base of the stem-loop formed by the repeat (Fig. 2.3B). The clone library did not contain crRNAs of 61 nucleotides, which would be the result of a single endonuclease cleavage event in each repeat, given the size of a repeat (29 nucleotides) and most spacers (32 nucleotides). Instead, in agreement with experimental observations (Figs. 2.2 and 2.3A), the crRNAs were truncated at the 3′ end by at least two guanosine bases from the endonuclease cleavage site, removing several stem-forming bases.

To test whether crRNA-loaded Cascade gives rise to phage resistance, two artificial CRISPRs were designed against phage Lambda (λ). Each of these CRISPRs targeted four essential λ genes (Fig. S2.3). The coding CRISPR ($C_{1–4}$) produced crRNAs complementary to both the mRNA and the coding strand of these four genes, whereas the template CRISPR ($T_{1–4}$) targeted only the template strand of the same proto-spacer regions (Fig. S2.3). A nontargeting CRISPR containing wild-type (WT) spacers with no similarity to the phage genome served as a control. Plaque assays with *E. coli* showed that the introduction of either one of these anti-λ phage CRISPRs in a strain expressing only Cascade did not result in reduced sensitivity of the host to a virulent Lambda phage (λvir) (Fig. 2.4A). However, strains that expressed Cascade and Cas3 were much less sensitive to phage infection. The template CRISPR rendered the strain insensitive to the phage at the highest phage titer tested (>10^7-fold less sensitive than the control strain), whereas the coding CRISPR reduced the sensitivity 10^2- fold (Fig. 2.4A) and produced plaques with a diameter ~1/10 of the standard l plaque. The phage resistance phenotype was lost when Cascade was omitted (Fig. 2.4A), proving that both Cascade and Cas3 are required in this process. Moreover, strains containing Cas3 and Cascade-CasE^{H20A} displayed a sensitive phenotype, which shows that pre-crRNA cleavage is mechanistically required for phage resistance. The co-expression of Cas1 and Cas2 had no effect on the sensitivity profile of the strain (Fig. 2.4A), suggesting that these proteins are involved in other stages of the CRISPR/*cas*

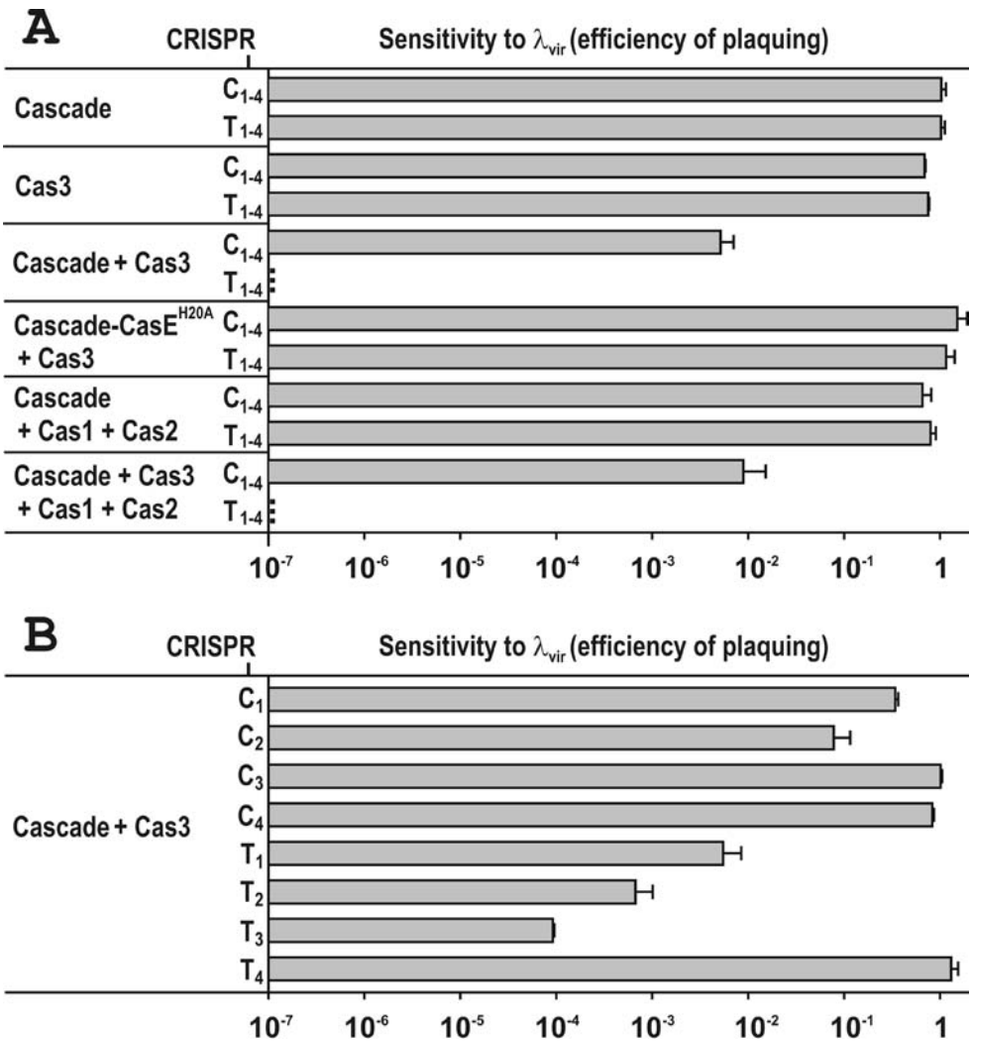


Figure 2.4. Engineered CRISPRs confer resistance to λ in the presence of Cascade and Cas3. (A) Effect of the presence of different sets of *cas* genes on the sensitivity of *E. coli* to phage λ_{vir}. Cells were equipped with one of two engineered CRISPRs containing four anti-λ spacers each (fig. S2.3). The $C_{1–4}$ CRISPR produces crRNA complementary to the coding strand and mRNA of λ_{vir}, and the $T_{1–4}$ CRISPR targets only the template strand. The sensitivity of each strain to phage λ_{vir} is represented as a histogram of the efficiency of plaquing, which is the plaque count ratio of the anti-λ CRISPR to that of the nontargeting control CRISPR. (B) Effect of single anti-λ spacers (fig. S2.3) on the sensitivity of *E. coli* to λ_{vir}. Error bars indicate 1 SD.

mechanism. Plaque assays with single anti-λ spacers (Fig. S2.3) showed that the total reduction of sensitivity observed with the four anti-λ spacers ($C_{1–4}$ and $T_{1–4}$) (Fig. 2.4A) results from a synergistic effect of the individual spacers (C_1 to T_4) (Fig. 2.4B).

Our results demonstrate that a complex of five Cas proteins is responsible for the maturation of pre-crRNA to small crRNAs that are critical for mediating an antiviral response. These mature crRNAs contain the antiviral spacer unit flanked by short RNA sequences derived from the repeat on either side termed the 5′ and 3′ handle, which may serve as conserved binding sites for Cascade subunits, as has been suggested previously (Kunin et al. 2007). The Cascade-bound crRNA serves as a guide to direct

the complex to viral nucleic acids to mediate an antiviral response. We hypothesize that crRNAs target virus DNA, because anti-λ CRISPRs of both polarities lead to a reduction of sensitivity to the phage. The model is supported by previous observations that virus-derived sequences are integrated into CRISPR loci, irrespective of their orientation in the virus genome (Bolotin et al. 2005; Mojica et al. 2005; Pourcel et al. 2005; Lillestøl et al. 2006; Makarova et al. 2006; Barrangou et al. 2007; Deveau et al. 2008; Horvath et al. 2008). We conclude that the transcription of CRISPR regions - and the cleavage of pre-crRNA to mature crRNAs by Cas proteins - is the molecular basis of the antiviral defense stage of the CRISPR/*cas* system, which enables prokaryotes to effectively prevent phage predation.

Materials and Methods

Strains

E. coli K12 gene knockouts from the KEIO collection were kindly provided by the National BioResource Project (NBRP, NIG, Japan) (Table S2.2).

Gene cloning, protein production and purification

The *cas* genes and CRISPRs were PCR amplified from *E. coli* K12 W3110 (BW25113) genomic DNA, and directionally cloned into a compatible expression vector set consisting of pET-52b (Amp^R, 100 µg/ml), pCDF-1b (Str^R, 50 µg/ml), pRSF-1b (Kan^R, 50 µg/ml) and pACYCduet-1 (Cam^R, 34 µg/ml) (Novagen), or pIH1119 (Amp^R) (New England Biolabs) as indicated in Table S2.2 and S2.3. Primers for the *cas* genes were designed based on their predicted translation start sites according to their latest annotation (Hayashi et al. 2006) (Table S2.2). Mutations were introduced using the QuikChange site-directed mutagenesis kit (Stratagene) (Table S2.2). Plasmids were transformed into *E. coli* BL21(DE3) (Novagen) which lacks endogenous *cas* genes, or *E. coli* DH5α in the case of pIH1119 and overproduced by inducing with 0.1 mM isopropyl-β-D-thiogalactopyranoside (IPTG) (Invitrogen) at an optical cell density at 600 nm of 0.6 for 16 h at 37 °C. Cells were harvested, resuspended in 20 mM Tris-HCl (pH 8.0) supplemented with 0.1 M NaCl, and disrupted using a French Pressure Cell. Tagged proteins were isolated using Strep-Tactin (IBA, Germany), HIS-Select (Sigma-Aldrich), or Amylose (New England Biolabs) affinity chromatography following manufacturer's instructions. The identity of the proteins was determined by mass spectrometry, as described (Snijders et al. 2006).

Northern blotting

Total RNA from 11 ml of exponentially grown *E. coli* cells was isolated using the

*mir*Vana miRNA isolation kit (Ambion). Strain *E. coli* BL21(DE3) containing various plasmid combinations (Table S2.3) was grown without IPTG induction, which resulted in low expression levels of the *cas* genes and CRISPR due to leakage of the expression system. Northern blots were performed by running 10 µg of RNA on a 9% polyacrylamide gels with 7 M urea in 0.5x TBE buffer (Sambrook et al. 1989). The RNA was then transferred to a Genescreen Plus membrane (PerkinElmer) by semi-dry blotting using a Trans-blot SD (Bio-Rad). After 1 min of UV-crosslinking and baking (80 °C, 15 min), the membrane was hybridized with QuikHyb (Stratagene) at 42 °C. Blots were probed for 12 hours with a ^{32}P-5'-labelled DNA oligonucleotide of spacer 4 in the *E. coli* K12 CRISPR (oligonucleotide BG2349, Table S2.2). The blots were subsequently washed with 2x SSC buffer (Sambrook et al. 1989) containing 0.1% SDS for 30 min, and 0.1x SSC buffer containing 0.1% SDS for 30 min. Blots were visualized using phosphorimaging with a Personal FX phosphorimager (Bio-Rad). RNA sizes were estimated by comparison with ^{32}P-labeled Decade RNA marker (Ambion).

Cleavage reactions

Internally radiolabelled transcripts were generated by *in vitro* transcription using the MAXIscript T7 kit (Ambion) and α-^{32}P-UTP (GE) (Table S2.2). Templates for *in vitro* transcription were generated by PCR using primers BG2559 and BG2374 for the *E. coli* K12 CRISPR, BG2462 and BG2463 for *E. coli* UTI89, and BG2452 and BG2461 for the Full-length RNA substrates were gel-isolated from denaturing 2% agarose gels as described (Locker, 1979). Cleavage reactions were set up at 37 °C in 20 mM Tris-HCl (pH 8.0) supplemented with 0.1 M NaCl and 1 mM EDTA. Protein samples (Table S2.3) were treated with 10 mM EDTA prior to the cleavage assay. Assays were started by adding 0.3 µg of Cascade, or 0.1 µg of MalE-CasE to a reaction containing 10 ng of gel-purified internally ^{32}P labeled transcripts in a total volume of 10 µl. Samples were treated with 1 U of proteinase K for 5 min at 37 °C (Fluka) and acid-phenol extracted as described (Elbashir et al. 2001). Reaction products were analyzed using 8% polyacrylamide gels containing 7 M urea.

RNA cloning

Protein-bound total RNA was isolated from Strep-Tactin-purified Cascade (Table S2.3) using the *mir*Vana miRNA isolation kit (Ambion). Approximately 6 µg of RNA was denatured for 10 min at 65 °C, and subsequently 3' polyadenylated for 75 min at 41 °C using Poly(A) polymerase (Ambion) as described (Botero et al. 2005). RNA products were separated from unincorporated nucleotides and enzymes using a NucAway gel filtration spin column (Ambion), and reverse transcribed with the Superscript III kit (Invitrogen) for 50 min at 50 °C with anchored primer BG2164 (5'-

GCCCGCCC**GGATCC**TTTTTTTTT-TTTTTTTTTTTTTTTTTTTTTVN-3') (BamHI site in bold face) (Botero et al. 2005). The RNA strand was degraded by RNase A (Fermentas) and RNase HI (Promega) for 15 min at 37 °C. Single-stranded cDNA was purified using the MinElute PCR purification kit (Qiagen). A 3' poly(dG) tail was added to the cDNA by 60 min incubation with Terminal Deoxynucleotidyl Transferase (Invitrogen) at 37 °C, followed by purification using MinElute reaction cleanup kit (Qiagen). The single stranded cDNA was used as a template in a PCR using Native Pfu polymerase (Stratagene) and primers BG2220 (5'-GCGCCCGC**GGATCC**CCCCCCCCDN-3') and BG2222 (5'-GCCCGCCC**GGATCC-**TT-3'). The PCR products were cloned into vector pUC29 and transformed into *E. coli* NEB5α (New England Biolabs).

Phage studies

Host sensitivity to phages was tested using a virulent variant of phage Lambda (λ_{vir}) (Jacob and Wollman, 1954) obtained from Centraalbureau voor Schimmelcultures (Utrecht, Netherlands) and *E. coli* BL21-AI (Invitrogen) as a host (Table S2.2 and S2.3). Strains were grown in 2YT-Lambda (2YTL) media (16 g/L tryptone, 10 g/L yeast extract, 5 g/L NaCl, 10 mM $MgSO_4$, 0.2 % maltose) until the optical density ($OD_{600\ nm}$) reached 0.3. Cas protein and pre-crRNA production was then induced for 30 min by adding a final concentration of 0.2 % L-arabinose (Sigma-Aldrich) and 0.1 mM IPTG. Cells were spun down and resuspended in 10 mM $MgSO_4$, before being used in plaque assays according to standard procedures (Sambrook et al. 1989). Plaque assays were performed in triplicate. Plates and top-agar contained 2YTL and above mentioned concentrations of inducers. The sensitivity of the host to phage infection was calculated as the efficiency of plaquing (Barrangou et al. 2007), which is the plaque count ratio of a strain containing an anti-λ_{vir} CRISPR to that of a strain containing a CRISPR with non-targeting spacers (N). Error-bars were calculated as one standard deviation. Anti-λ CRISPRs were designed by randomly picking proto-spacer sequences in four genes of the λ genome (Fig.S2.3). The artificial anti-λ CRISPR design did not take any *S. thermophilus* CRISPR motifs into account (Deveau et al. 2008, Horvath et al. 2008). The motifs are conserved nucleotide sequences located downstream of proto-spacers on the virus genome, which are important for the phage resistant phenotype in *S. thermophilus*. No CRISPR motif could be identified for the *E. coli* K12 CRISPR/*cas* system using the flanking regions of the four known proto-spacers in phage P1 and plasmid F (Mojica et al. 2005). The anti-λ CRISPRs were synthesized by Geneart AG (Regensburg, Germany), and subcloned into vector pACYCduet-1 vector (Novagen) ($C_{1\text{-}4}$, $T_{1\text{-}4}$, Table S2.2) using restriction sites NcoI and Acc65I. CRISPRs with single targeting spacers (C_1, C_2, C_3, C_4, T_1, T_2, T_3, T_4) were obtained by exchange of single non-targeting spacers of the N CRISPR with the corresponding single targeting spacers

of C_{1-4} and T_{1-4} CRISPRs using restriction enzyme pairs NcoI and EcoRI, EcoRI and BamHI, BamHI and NsiI, NsiI and Acc65I (Fig.S2.3). The CRISPR sequences are provided in Figure S2.4.

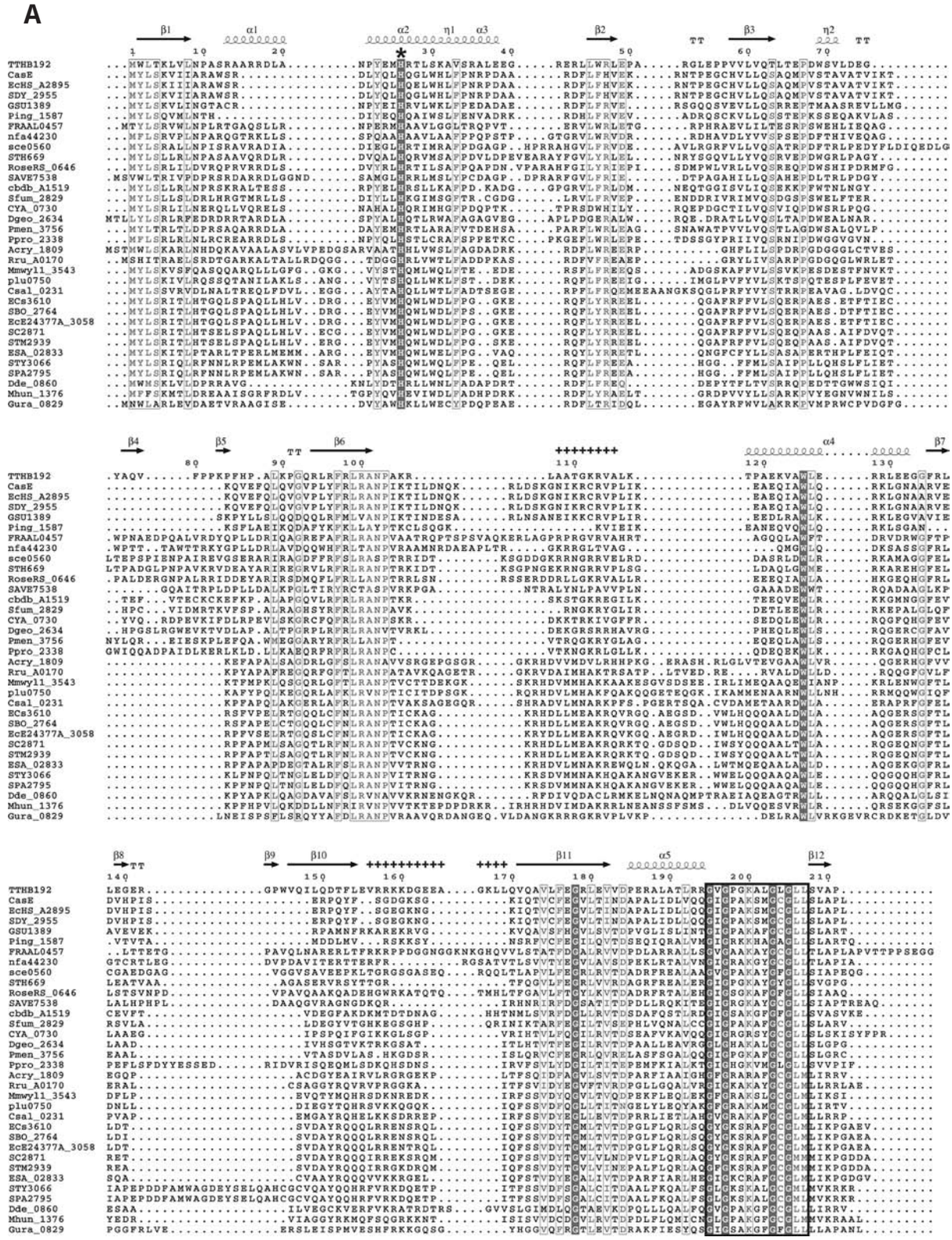
Acknowledgments

We thank T. Verweij, C. G. J. van Houte, and M. R. Beijer for experimental contributions and T. Goosen (Hogeschool van Arnhem en Nijmegen BioCentre), M. J. Young (Montana State University), T. Bisseling, and W. M. de Vos (Wageningen University) for helpful discussions. We are grateful for receiving strains from the KEIO collection distributed by National BioResource Project (National Institute of Genetics, Japan). We thank U. Dobrindt (University of Würzburg) for sending genomic material of *E. coli* UTI89. This work was financially supported by a Vici grant from the Dutch Organization for Scientific Research (Nederlandse Organisatie voor Wetenschappelijk Onderzoek) and a Marie Curie grant from the European Union. M.L. was supported by the Wenner-Gren Foundations.

Supporting Online Material, including Fig. S2.2, Fig. S2.4, Table S2.2 and Table S2.3, can be found at www.sciencemag.org/cgi/content/full/321/5891/[page]/DC1

Supplementary figures and table

A



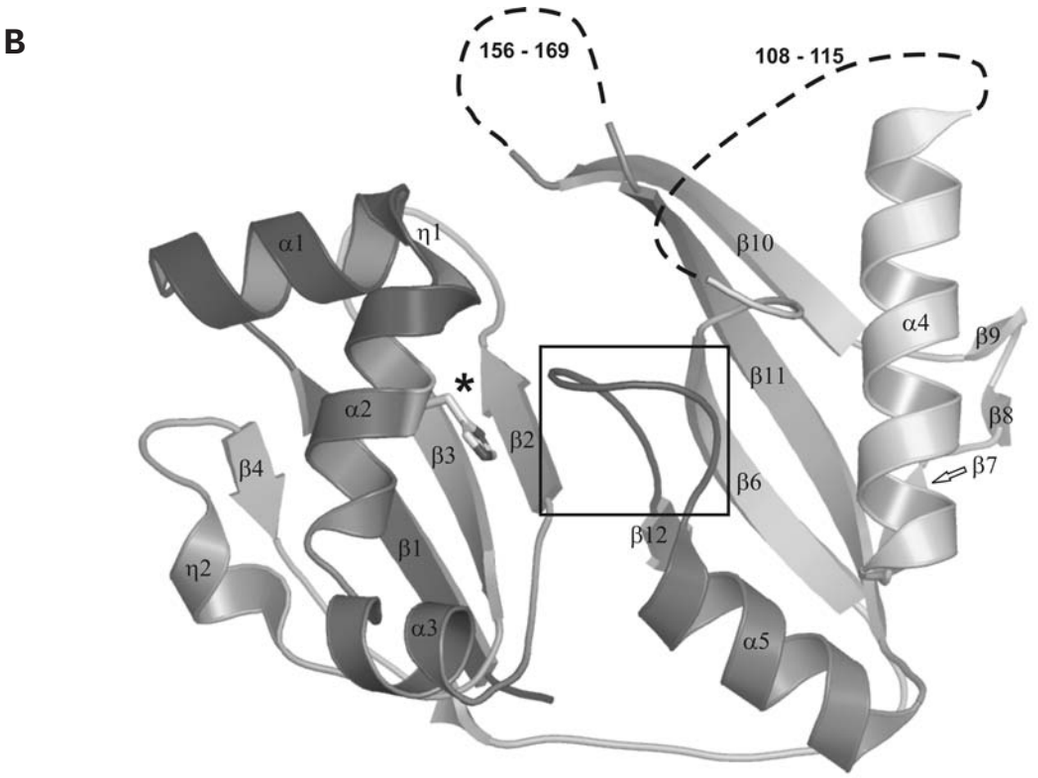


Figure S2.1 For full colour version see page 129.

A. Multiple sequence alignment of CasE homologs. Sequences were aligned with TCoffee (Poirot et al. 2003) and aligned to the structure of TTHB192 (PDB ID: 1WJ9) (Ebihara et al. 2006) using ESPript (Gouet et al. 2003). TTHB192, *Thermus thermophilus* HB8; CasE, *Escherichia coli* K12 W3110; EcHS_A2895, *E. coli* HS; SDY_2955, *Shigella dysenteriae* Sd197; GSU1389, *Geobacter sulfurreducens* PCA; Ping_1587, *Psychromonas ingrahamii* 37; FRAAL0457, *Frankia alni* ACN14a; nfa44230, *Nocardia farcinica* IFM 10152; sce0560, *Sorangium cellulosum* 'So ce 56'; STH669, *Symbiobacterium thermophilum* IAM 14863; RoseRS_0646, *Roseiflexus* sp. RS-1; SAVE7538, *Streptomyces avermitilis* MA-4680; cbdb_A1519, *Dehalococcoides* sp. CBDB1; Sfum_2829, *Syntrophobacter fumaroxidans* MPOB; CYA_0730, *Synechococcus* sp. JA-3-3Ab; Dgeo_2534, *Deinococcus geothermalis* DSM 11300; Pmen_3756, *Pseudomonas mendocina* ymp; Ppro_2338, *Pelobacter propionicus* DSM 2379; Acry_1809, *Acidiphilium cryptum* JF-5; Rru_A0170, *Rhodospirillum rubrum* ATCC 11170; Mmwyl1_3543, *Marinomonas* sp. MWYL1; plu0750, *Photorabdus luminescens* subsp. laumondii TT01; Csal_0231, *Chromohalobacter salexigens* DSM 3043; Ecs3610*, E. coli* 0157:H7 str. Sakai; SBO_2764, *Shigella boydii* Sb227; EcE24377A_3058, *E. coli* E24377A; SC2871, *Salmonella enterica* subsp. Enterica serovar Choleraesuis str. SC-B67; STM2939, *Salmonella typhimurium* LT2; ESA_02833, *Enterobacter sakazakii* ATCC BAA-894; STY3066, *S. enterica* subsp. Enterica serovar Typhi str. CT18; SPA2795, *S. enterica* subsp. Enterica serovar Paratyphi A str. ATCC9150; Dde_0860, *Desulfovivrio desulfuricans* G20; Mhun_1376, *Methanospirillum hungatei* JF-1; Gura_0829, *Geobacter uraniireducens* Rf4. Secondary structural elements and amino acid numbering follows the TTHB192 structure and sequence. Disordered loops are indicated with a plus (**+**), and the highly conserved residue His26 is marked with an asterisk (*****). The highly conserved C-terminal glycine-rich loop, which is the hallmark of this protein family, is highlighted with a box.

B. Ribbon diagram of the structure of TTHB192, a CasE homolog from *Thermus thermophilus* HB8 (PDB ID: 1WJ9) (Ebihara et al. 2006). Structural features are indicated as in **A**. Structurally disordered residues 108 to 115 and 156 to 169 are depicted by dashed lines. Note that the highly conserved glycine-rich loop between secondary structure elements α5 and β12 is spatially close to His26.

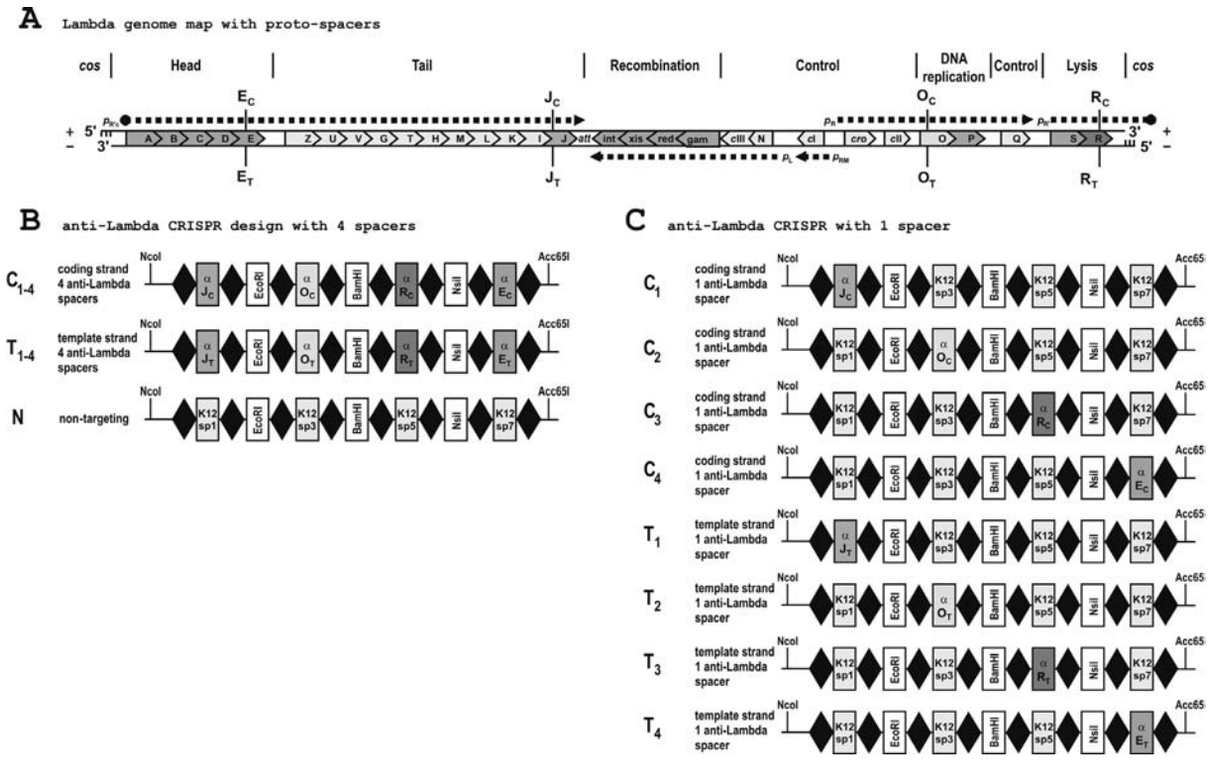


Figure S2.3

A Phage λ genome map indicating the main genes and transcripts (dotted arrows), and the positions of the proto-spacers on the coding or template strand.

B Design of two anti-λCRISPRs (repeats: diamonds, spacers: rectangles) producing crRNAs complementary to (indicated by α) either the coding strand (C_{1-4}) of the four genes J, O, R and E (*i.e.* mRNA and plus strand of the viral genome), or the template strand (T_{1-4}) of these four genes (*i.e.* minus strand). A third CRISPR (N) was designed as a non-targeting control containing the naturally occurring spacers 1, 3, 5, and 7 from *E. coli* K12, which have no homology to any known phage. The number of plaque forming units obtained in the presence of this CRISPR was used to calculate the efficiency of plaquing (Fig.2.4). Restriction sites were introduced in spacer 2, 4 and 6 (EcoRI, BamHI and NsiI, respectively).

C CRISPRs with single targeting spacers (C_1, C_2, C_3, C_4, T_1, T_2, T_3, T_4) were obtained by exchange of single non-targeting spacers of the N CRISPR with the corresponding single targeting spacers of C_{1-4} and T_{1-4} CRISPRs using restriction enzyme pairs NcoI and EcoRI, EcoRI and BamHI, BamHI and NsiI, NsiI and Acc65I.

Table S2.1. Overview of different Cas systems. This Table focuses on well conserved Cas proteins (Cas1-4), and subunits of the *E. coli* K12 Cascade complex (including Cas5); for more extensive comparative analyses, see Haft *et al.* (2005) and Makarova *et al.* (2006). *E. coli* strains discussed in this study are in bold.

Cas protein[a]	Family[b]		Cas system (CASS)[c]								
			none	CASS1	CASS2	CASS3	CASS4	CASS5	CASS6	CASS7	Function predicted (p)[a], or demonstrated experimentally (e)
		Species Strain	***Escherichia coli*** **BL21(DE3)**	*Bacillus halodurans* C-125	***Escherichia coli*** **K12**	***Escherichia coli*** **UTI89**	*Streptococcus thermophilus* CNRZ1066	*Sulfolobus solfataricus* P2	*Nitrosomonas europaea* ATCC 19718	*Archaeoglobus fulgidus* VC-16	
Cas1	COG1518		-	BH0341	b2755 (YgbG)	C0890	str0658	SSO1450	NE0111	AF1878	nuclease/integrase (p)
Cas2	COG1343		-	BH0342	b2754 (YgbF)	C0891[f]	str0659	SSO8090	NE0112	AF1876	RNA endonuclease (e) (*S17*)
Cas3	COG1203		-	BH0336	b2761 (YgbB)	C0891[f]	-	SSO1440	-	AF1874	DEAD-box helicase, often fused to HD nuclease (p)
Cas4	COG1468		-	BH0340	-	-	-[d]	SSO1451	-	AF1877	RecB-like nuclease (p), often has C-terminal Zn clusters
Cascade complex[e]											
CasA (Cse1)	YgcL		-	-	b2760 (YgcL)	-	-	-	-	-	Zn-finger containing protein (p)
CasB (cse2)	YgcK		-	-	b2759 (YgcK)	-	-	-	-	-	α-helical protein (p)
CasC (Cse4)	YgcJ (COG1857)		-	BH0339 (COG3649)	b2758 (YgcJ)	C0893 (y1725)	-	SSO1442	-	AF1871	α/β protein, nuclease (p)
CasD (Cas5/5e)	YgcI (COG1688)		-	BH0337	b2757 (YgcI)	C0893/894 (y1726)	-	SSO1441	-	AF1872	RAMP (p)
CasE (Cse3)	YgcH		-	-	b2756 (YgcH)	C0896 (y1727)	-	-	-	-	RAMP (p), crRNA endonuclease (e)[e]

[a] Cas protein nomenclature and functional prediction according to Haft *et al.* (2005), Jansen *et al.* and Makarova *et al.* (2006)
[b] Family nomenclature according to Makarova *et al.* (2006); family may contain several COGs
[c] CASS nomenclature according to Makarova *et al.* (2006)
[d] Cas4 is absent in CASS4 (*S. thermophilus*), but is present in CASS4a (Makarova *et al.*, 2006)
[e] this study
[f] Beloglazova *et al.* (2008)

Chapter 3

Structural basis for CRISPR RNA-guided DNA recognition by Cascade

Matthijs M. Jore*, Magnus Lundgren*, Esther van Duijn*, Jelle B. Bultema*, Edze R. Westra, Sakharam P. Waghmare, Blake Wiedenheft, Marieke R. Beijer, Arjan Barendregt, Kaihong Zhou, Ambrosius P.L. Snijders, Mark J. Dickman, Jennifer A. Doudna, Egbert J. Boekema, Albert J. R. Heck, John van der Oost, Stan J.J. Brouns

*Contributed equally

Submitted

Abstract

The CRISPR immune system in prokaryotes utilizes small guide RNAs to neutralize invading viruses and plasmids. In *Escherichia coli*, immunity is dependent on a ribonucleoprotein complex called Cascade. Here we present the composition and low-resolution structure of Cascade and show how it recognizes double-stranded DNA targets sequence specifically. Cascade is a 405 kDa complex comprising five functionally essential Cas proteins ($CasA_1B_2C_6D_1E_1$) and a 61 nucleotide crRNA with 5'-hydroxyl and 2', 3'-cyclic phosphate termini. The crRNA guides Cascade to the complementary sequence of double-stranded DNA by ATP-independent strand displacement, indicating that invader DNA surveillance takes place without continuous investment of resources. The structure of Cascade reveals an unusual seahorse-shape that undergoes conformational changes upon target DNA binding. Based on the composition of Cascade and structures of Cascade core subcomplexes a structural model is presented that provides insight into the molecular basis of crRNA-guided target DNA recognition.

Introduction

The constant pressure of invading viruses and conjugative plasmids has shaped the evolution of host defence systems in prokaryotes. The widely distributed CRISPR (clustered regularly interspaced short palindromic repeats) immune system represents the most recently discovered prokaryotic defence strategy (reviewed by (van der Oost et al. 2009; Horvath and Barrangou 2010; Karginov and Hannon 2010; Marraffini and Sontheimer 2010a)). The system consists of repeats that are interspaced by unique sequences called spacers, which are derived from viral and plasmid DNA (Bolotin et al. 2005; Mojica et al. 2005; Pourcel et al. 2005). CRISPR-based immunity is adaptive and inheritable because it can both memorize invaders by storing fragments of their DNA, and pass that information on to subsequent generations following a Lamarckian type of evolution (Koonin and Wolf 2009; van der Oost et al. 2009).

The CRISPR-associated (Cas) protein machinery is encoded by gene clusters that are located in close proximity of the CRISPR locus (Jansen et al. 2002) which has allowed for the extensive horizontal transfer of complete CRISPR/Cas systems (Godde and Bickerton 2006). Multiple types of *cas*-gene sets have been recognized (Haft et al. 2005; Makarova et al. 2006) that correlate with specific families of repeat sequences (Kunin et al. 2007).

The mechanism of CRISPR/Cas-induced immunity has been divided into three stages. In the first stage, CRISPR adaptation, the host encounters an invader and integrates a fragment of foreign DNA non-directionally into the CRISPR as a new spacer, resulting in resistance to foreign genetic elements carrying this sequence (Barrangou et al. 2007; Horvath et al. 2008; van der Ploeg 2009). Although the molecular determinants of foreign DNA recognition have not been elucidated, the crystal structure and metal-dependent DNase activity of Cas1 indicated that this enzyme maybe involved in generating small DNA fragments that are used as precursors for CRISPR adaptation (Wiedenheft et al. 2009). Newly acquired spacers from both the coding and template strand of the viral genome have been shown to confer immunity (Barrangou et al. 2007).

In the second stage, CRISPR expression, the CRISPR locus is transcribed and the repeat regions within the precursor CRISPR RNA (pre-crRNA) are cleaved (Tang et al. 2002; Hale et al. 2008; Lillestol et al. 2009) by a specific Cas endoribonuclease. Two pre-crRNA processing endonucleases have been described: CasE from *Escherichia coli* (Brouns et al. 2008) and Cas6 from *Pyrococcus furiosus* (Carte et al. 2008). In *E. coli*, CasE is an essential component of the Cascade complex (CRISPR-associated complex for antiviral defence), which consists of five Cas proteins from the Cse-type

(CasABCDE) (Fig. 3.1A). CasE-generated mature crRNAs remain bound to Cascade to guide host defence. This is distinct from the situation in *P. furiosus* in which Cas6-generated crRNAs (also called psiRNAs) end up in the Cmr-complex encoded by the RAMP module of Cas proteins, where they are further trimmed at the 3'-end from ~67 to 39 or 45 nucleotides. The guide RNA-loaded Cmr-complex cleaves single stranded target RNA sequence-specifically (Hale et al. 2009). In *E. coli*, the third stage, CRISPR interference, not only requires Cascade loaded with anti-invader crRNA, but also the participation of the predicted nuclease/helicase Cas3. Because crRNAs complementary to the either strand of the phage DNA provided resistance, it was proposed that Cascade is a crRNA-guided complex that targets DNA rather than mRNA (Brouns et al. 2008). A series of genetic experiments in *Staphylococcus epidermidis* involving a conjugative plasmid showed that the Csm-type CRISPR/Cas system targets DNA as well (Marraffini and Sontheimer 2008). Although DNA is the prime candidate for the target molecule in the *E. coli* and *S. epidermidis* model systems, direct molecular evidence of Cas proteins interacting with their target DNA has been lacking. In this study we show how Cascade recognizes target DNA and present a structural model of Cascade that provides insight into the mechanism of crRNA-guided recognition of DNA targets.

Results

Core sub-complexes of Cascade

The *E. coli* K12 CRISPR/Cas system (Cse-subtype) consists of a gene cluster encoding eight *cas* genes (*cas3*, *casABCDE*, *cas1* and *cas2*) and a downstream CRISPR locus (Fig. 3.1A). Five Cas proteins (CasABCDE) form the Cascade complex, which cleaves a long precursor transcript of the CRISPR region (pre-crRNA) into small crRNA molecules. These crRNAs remain bound to the complex to guide antiviral defence (Brouns et al. 2008). To investigate the role of the individual subunits, we first tested whether each subunit is required for antiviral defence. Viral plaque assays with Cas3 and Cascade lacking one type of protein subunit showed that all protein components of Cascade are indispensible for the virus resistant phenotype of *E. coli* (Fig. S3.1).

We then systematically overproduced and affinity purified different combinations of Cascade subunits and checked for the presence of mature crRNA. This analysis showed that CasA or CasAB could be omitted without affecting the apparent stoichiometry of the remaining subunits or the mature crRNA (Figs. 3.1B and 3.1C). We noticed that Cascade, unlike CasBCDE and CasCDE, always co-purified with large nucleic acid molecules (>300 nt) (Fig. 3.1C). Removal of the Cas proteins followed by nuclease

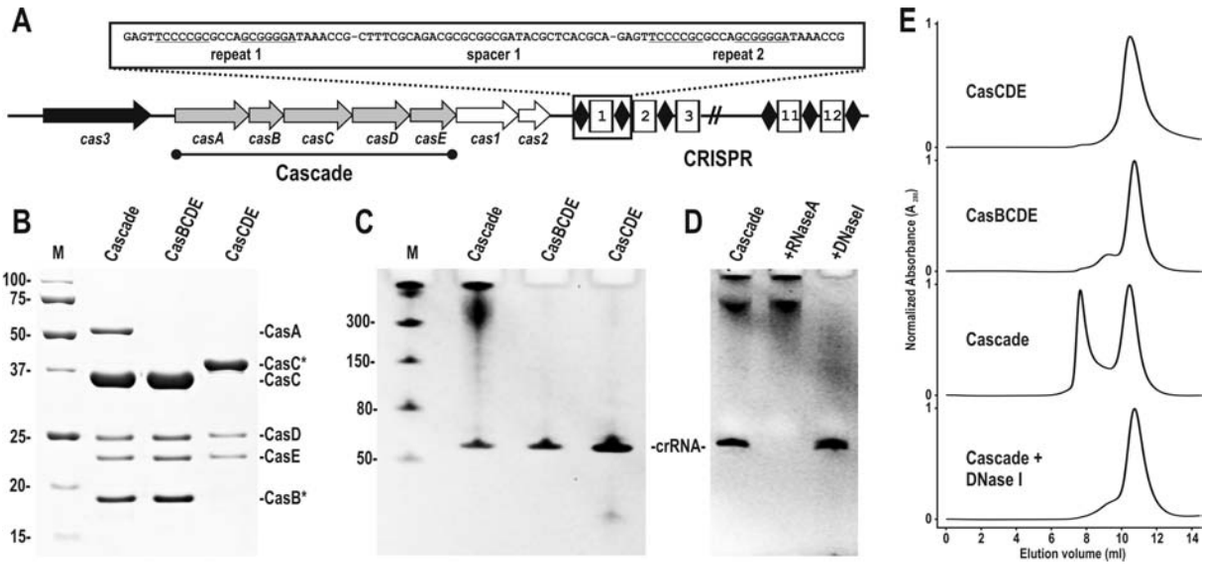


Figure 3.1. Core complexes of Cascade retain crRNA. A) Schematic diagram of the CRISPR/Cas locus in *E. coli* K12 containing *cas3* (*ygcB*), *casA* (*cse1*, *ygcL*), *casB* (*cse2*, *ygcK*), *casC* (*cse4*, *ygcJ*), *casD* (*cas5e*, *ygcI*), *casE* (*cse3*, *ygcH*), *cas1* (*ygbT*), *cas2* (*ygbF*) (Jansen et al. 2002; Haft et al. 2005). B) Coomassie blue-stained SDS-polyacrylamide gel showing StrepTactin purified Cascade, CasBCDE and CasCDE. Protein marker sizes in kDa. The asterisk marks the Strep-tagged subunit. C) Ethidium bromide-stained denaturing PAA-gel showing nucleic acids isolated from purified Cascade (sub-)complexes. RNA marker sizes in nucleotides. D) RNase A or DNase I treatment of Cascade bound nucleic acids from Cascade. E) Size exclusion elution profiles of CasCDE, CasBCDE and Cascade before and after DNase I treatment.

treatments showed that RNase A only hydrolyzed the crRNA, while DNase I removed the long nucleic acids, thereby identifying the co-purified nucleic acid as DNA (Fig. 3.1D). Size exclusion chromatography of the three types of complexes revealed that the vast majority of CasBCDE and CasCDE were present in a single form, whereas Cascade displayed a substantial void peak in addition to a discrete peak at ~11 ml (Fig. 3.1E). DNase I treatment prior to gel filtration eliminates the void peak without disruption of the discrete Cascade peak, again indicating the presence of Cascade-bound DNA (Fig. 3.1E).

Architecture of crRNA

The characteristics of the mature crRNA species were accurately determined by subjecting mature crRNAs isolated from Cascade to denaturing RNA chromatography (Dickman and Hornby 2006; Waghmare et al. 2009) and electrospray ionization mass spectrometry (ESI-MS). To simplify the analysis, a uniform crRNA preparation was obtained by co-expressing Cascade with a designed CRISPR containing eight repeats and seven identical spacers (denoted R44 CRISPR (Fig. S3.2)). This setup resulted in a Cascade preparation in which each molecule was loaded with the same

crRNA. Chromatography demonstrated the purity and homogeneity of this crRNA preparation (Fig. 3.2A). Furthermore, the observed retention time was consistent with an approximate length of 60 nt. The ESI-MS spectra indicated that the crRNA had a molecular weight of 19,660.80 Da (Fig. 3.2B), which corresponds well to an expected molecular weight of 19,660.82 Da for a 61 nt crRNA resulting from a single CasE endoribonuclease cleavage event in each repeat. The purified mature crRNA was also analyzed using ESI-MS/MS analysis following RNase T1 and RNase A digestion. A number of oligoribonucleotide digests were assigned to the mature crRNA sequence (Fig. S3.3) and were consistent with the previously determined CasE cleavage site, 5' of the terminal base of the hairpin (Brouns et al. 2008). The molecular weight analysis of the crRNA indicated a 5'-hydroxyl group and a 2',3'-cyclic phosphate terminus. The presence of a cyclic phosphate terminus was confirmed by acid treatment of the crRNA, which showed a mass shift of 18 Da, corresponding to the hydrolysis of the 2',3'-cyclic phosphate to a 2' or 3' phosphate (Fig. 3.2C). Mature crRNA is 61 nucleotides long and contains the 32 nucleotide spacer sequence, flanked by repeat-derived sequences on either end: 8 bases at the 5' terminus (5'-handle) and 21 bases forming a hairpin with a tetra-nucleotide loop at the 3' terminus (3'-handle) (Fig. 3.2D).

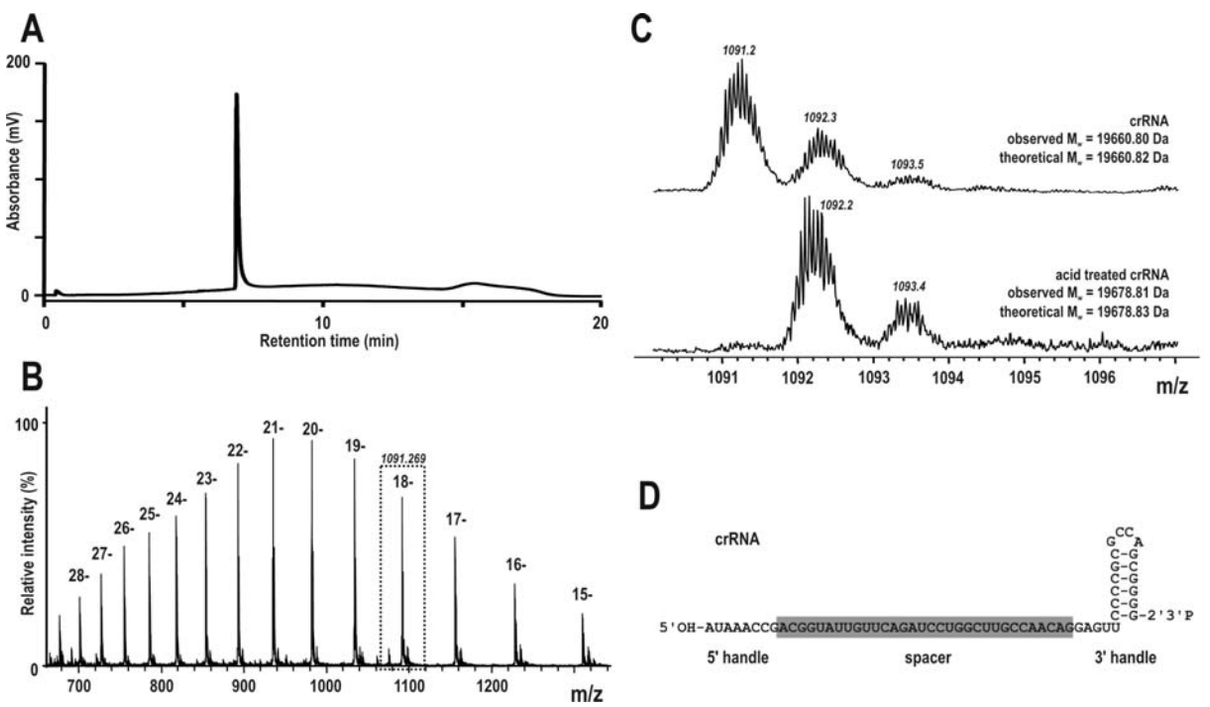


Figure 3.2. Architecture of crRNA. A) Ion-pair reversed-phase HPLC purification of mature R44 crRNA (Fig.S3.2) at 75 °C. B) Multiple-charged ESI-MS spectra of the purified mature crRNA. C) Enhanced view of the 18- charged species before (upper graph) and after acid treatment (lower graph) indicating the hydrolysis of the 2',3'-cyclic phosphate. D) Diagram of mature crRNA derived from the R44 CRISPR showing the 5' hydroxyl group and 2',3'-cyclic phosphate.

Target recognition by Cascade

The observation that DNA co-purified with Cascade (Figs. 3.1C and 3.1D) prompted us to analyze the DNA binding behaviour of Cascade in detail. Electrophoretic mobility shift assays (EMSAs) demonstrated that Cascade was able to bind single-stranded (ss) DNA containing the protospacer, a sequence complementary to the spacer sequence of the crRNA (Figs. 3.3A, 3.3C and S3.4). Double-stranded (ds) target DNA was also bound, without the need for additional co-factors such as divalent metal-ions or ATP (Figs. 3.3B and 3.3D). The dissociation constant (Kd) of Cascade for single- and double-stranded target DNA was 8 and 790 nM, respectively. In addition to target DNA, Cascade also displayed weak non-target DNA binding, i.e. DNA without a protospacer (Figs. 3.3C and 3.3D). Competitor DNA blocked Cascade from binding non-target DNA, and at very high competitor concentrations also from binding target DNA (Figs. 3.3A-D). The competitor had little effect on preformed Cascade-target DNA complexes, indicating a stable interaction between Cascade and complementary DNA substrates, while non-target interaction was transient (not shown). At high competitor concentrations the binding of target DNA by a proportion of Cascade lacking CasA was observed, as was evident from the faster migration rate of the CasBCDE-DNA complex (Figs. 3.3A, 3.3C and S3.4).

Cascade sub-complexes lacking CasA (CasBCDE and CasCDE) displayed only sequence-specific binding to ss- and dsDNA targets, and did not bind non-target DNA (Figs. 3.3E-H). Consistent with that observation, the target DNA binding behaviour of these subcomplexes was not affected by the addition of competitor DNA. The addition of purified CasA to CasBCDE preparations restored Cascade-like non-specific DNA binding ability (Fig. S3.5), but neither CasA alone (Fig. S3.5) nor the combination CasA and CasCDE (not shown) displayed non-specific DNA binding. Because the mobility shift caused by Cascade or CasBCDE binding could be distinguished (Figs. 3.3A, 3.3C and S3.4), competition assays were performed between the two types of complexes.

This showed that addition of target to a mix of equal amounts of Cascade and CasBCDE resulted in substantially more target DNA being bound by Cascade; a CasBCDE:Cascade ratio of 25:1 was required to distribute the target equally between the types of complexes (Fig. 3.4A). Furthermore, less dsDNA target was shifted by CasBCDE than Cascade at equivalent conditions (Figs. 3.3D and 3.3F), again indicating that CasA enhances target DNA localization.

It appears that Cascade recognizes dsDNA targets by base pairing of the crRNA spacer sequence with the complementary DNA strand. Analysis of long dsDNA targets

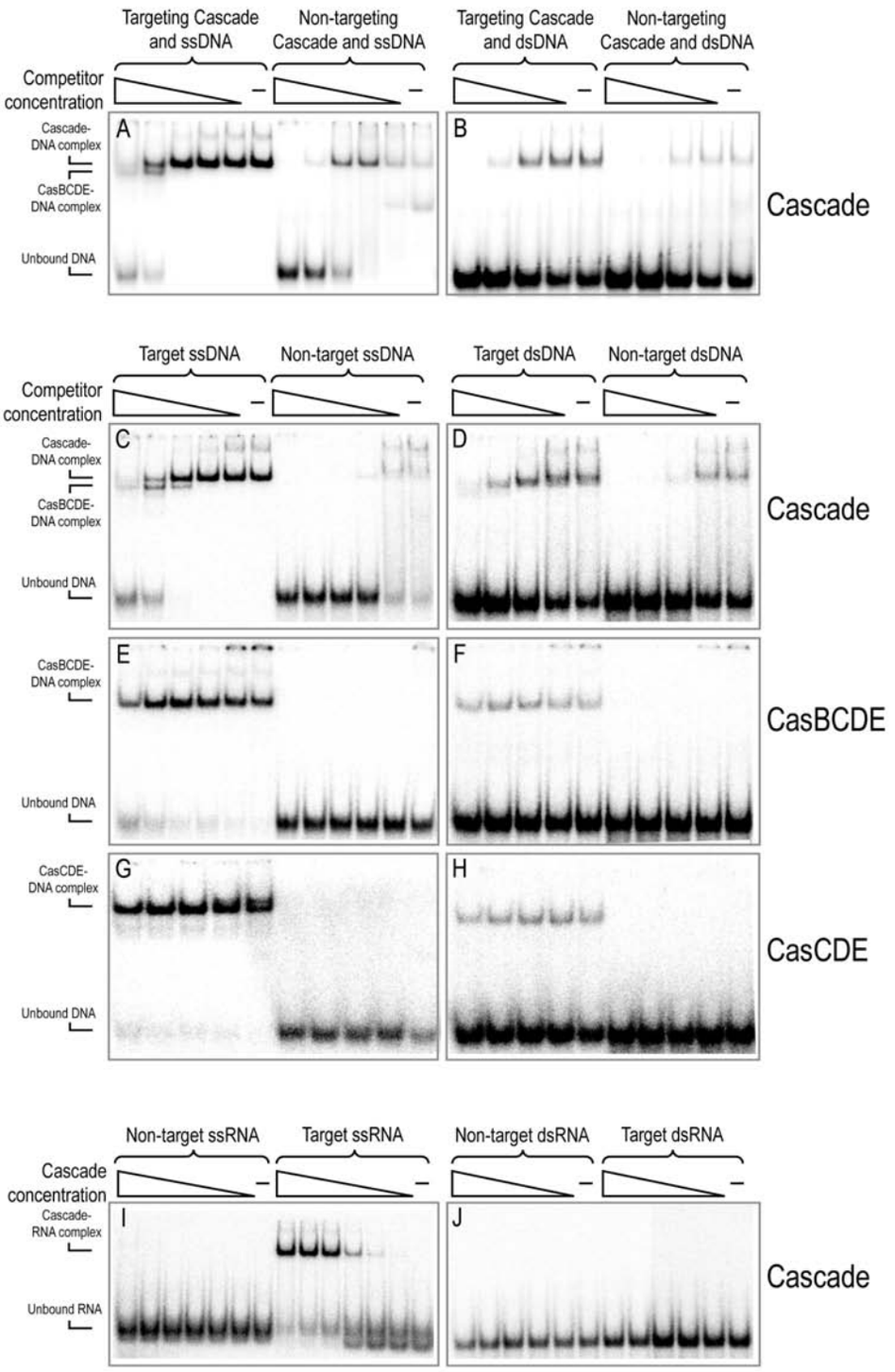


Figure 3.3. Target recognition by Cascade. A-B) Effect of the type of crRNA bound. Cascade was loaded with either targeting crRNA (derived from the R44 CRISPR, Fig. S3.2) or non-targeting crRNA (derived from the K12 CRISPR). The binding of these two types of Cascade complexes to one type of probe is shown. DNA probes are 86 nucleotide ssDNA or dsDNA sequences containing the R44 protospacer (32 nucleotides), flanked by 27 nucleotides on either end C-H) Effect of uniform crRNA-loaded complexes (R44 CRISPR) on the binding of DNA probes with and without protospacer (denoted target and non-target, respectively). The binding of Cascade, CasBCDE and CasCDE to an 86 nucleotide target or non-target ss/dsDNA is shown. Non-target DNA probes contain a scrambled R44 protospacer sequence. I-J) Effect of uniform crRNA-loaded Cascade (R44 CRISPR) on the binding of target and non-target ssRNA and dsRNA. A-H) Labeled probe concentration 1 nM. DNA competitor concentration is 2500, 500, 50, 5 and 0.5 ng/µl (the highest concentration was not used for CasCDE), protein concentration is 200-300 nM except in I-J where the Cascade concentration range is 200, 100, 50, 25 and 12.5 nM.

(protospacer with 27 bp flanks) showed that both strands shifted (Fig. 3.4B) due to base pairing of the flanking regions. Only the complementary strand shifted when short dsDNA targets (corresponding to the protospacer) were used, indicating that crRNA base pairing with the complementary strand caused displacement of the non-complementary strand (Fig. 3.4C). Cascade-mediated target DNA cleavage was not observed under any of the conditions tested. Binding to complementary ssRNA could be demonstrated (Fig. 3.3I), but this did not result in target RNA cleavage. In contrast

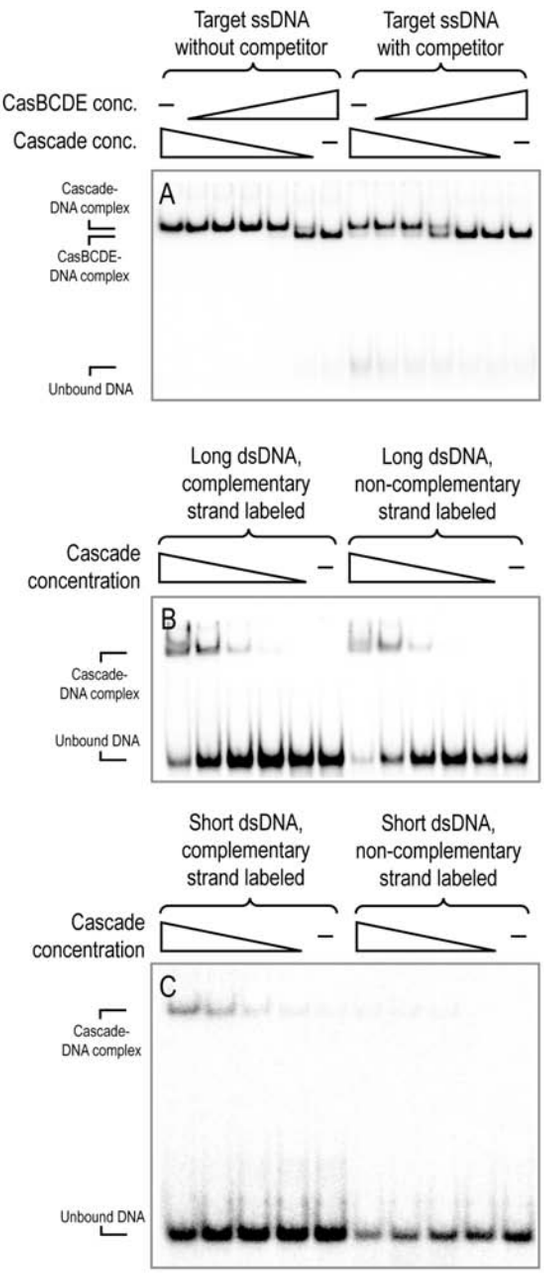


Figure 3.4. A) Competition assay between R44 crRNA-loaded Cascade and CasBCDE for R44 ssDNA target. Total protein concentration is 500 nM in each reaction, and the Cascade:CasBCDE ratio is 1:0, 100:1, 10:1, 1:1, 1:10, 1:100 and 0:1. When used, DNA competitor concentration is 1 mg/µl. B) Effect of labeling the complementary or non-complementary strand of a long dsDNA target containing the R44 protospacer with 27 bp flanks. C) as A) but with a short dsDNA target encompassing only the R44 protospacer. B-C) Cascade concentration range is 1500, 300, 60 and 12.5 nM.

to dsDNA, no binding to dsRNA was observed (Fig. 3.3J). In addition, very little non-specific interaction occurred between Cascade and ss- or dsRNA (Figs. 3.3I and 3.3J).

Subunit stoichiometry of Cascade

To understand the structural basis of the interaction between Cascade and target DNA, the composition of the Cascade protein assembly was determined using an array of mass spectrometric analyses. Denaturing and tandem mass spectrometry analyses resulted in accurate mass measurements for each component of Cascade (Table S3.1). The measured masses CasA, CasB and CasD were in agreement with the expected values, and the mass of CasC and CasE with the primary amino acid sequence lacking the N-terminal methionine. A complex composed of one copy of each Cascade component ($CasA_1B_1C_1D_1E_1$/$crRNA_1$) would have a mass of 184 kDa. However, analysis of the intact assembly by native mass spectrometry (Heck 2008) showed two major charge state distributions, corresponding to masses of 405,365 ± 135 Da and 349,399 ± 84 Da (Fig. 3.5A and Table S3.1). A third low intensity charge

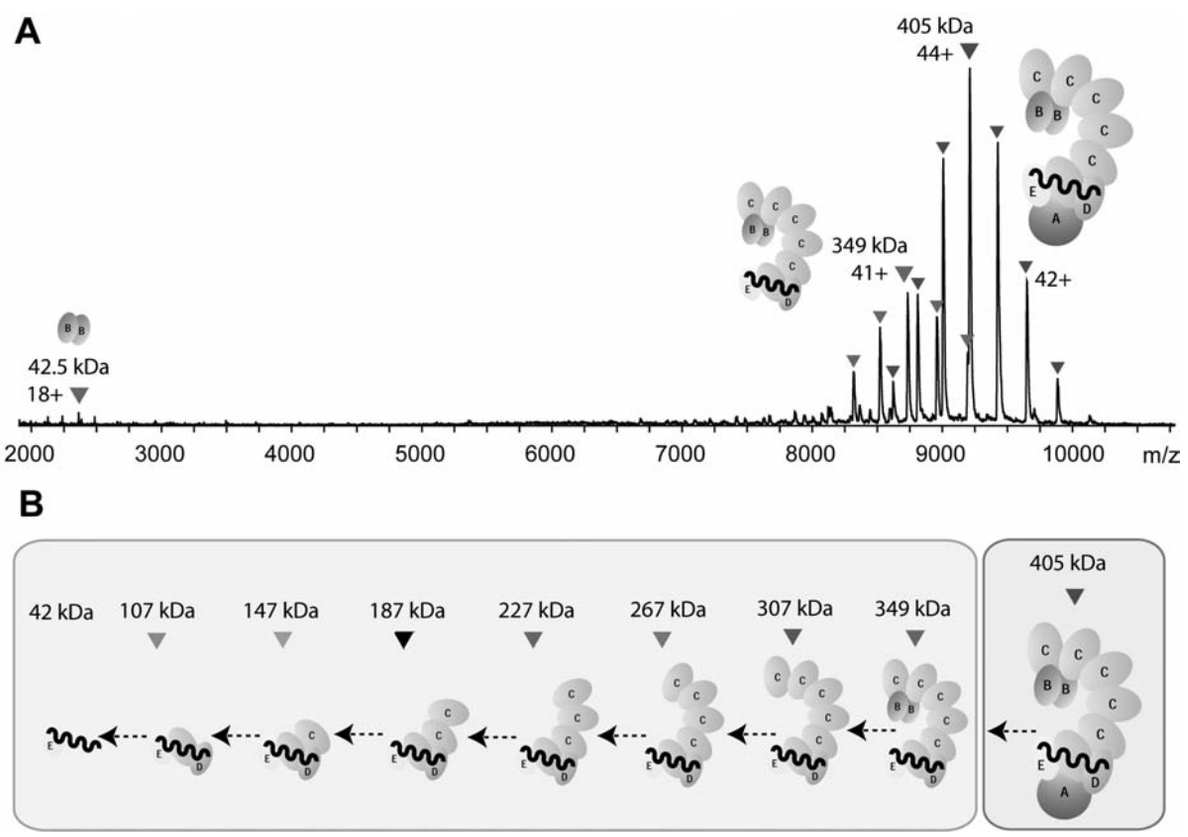


Figure 3.5. For full colour version see page 129. A) Native nano-ESI mass spectrum of Cascade. Two charge state distributions are present at high m/z values, corresponding to complexes of 405 kDa (purple) and 349 kDa (pink). The charge state distribution indicated in red indicates the CasB dimer. B) Cascade (sub-)complexes analyzed by native mass spectrometry. The sub-complexes were formed in solution after adding 5% 2-propanol to the buffer solution containing Cascade.

state distribution around 2,200 m/z was observed with a mass of 42,524 ± 8 Da, which is close to the theoretical molecular weight of a CasB dimer (42,521 Da). Proteolytic removal of the affinity tag on CasB unambiguously confirmed the presence of two CasB copies (Fig. S3.6A).

The two major complexes of 405 and 349 kDa likely resemble the intact Cascade and a Cascade sub-complex lacking CasA, in agreement with the mass difference of 55,966 Da. The presence of this sub-complex suggests that at least one CasA copy is located at the periphery of Cascade with a rather low affinity. Tandem mass spectrometry experiments on intact Cascade ions revealed that CasA was the first subunit to be expelled from the complex under collisional activation conditions, again indicative of a peripheral position of CasA within the complex (Fig. S3.6B). In addition to the elimination of CasA, also the loss of CasC was observed by tandem mass spectrometry. Similarly, selection and activation of the 349 kDa Cascade sub-complex showed the loss of one CasD and up to two CasC subunits. Unfortunately, Cascade could not be further disrupted by tandem mass spectrometry, hampering the full assignment of its stoichiometry. Therefore a number of alternative strategies were devised.

Since Cascade was loaded with a single type of crRNA (R44, Fig. S3.2), the number of bound crRNA molecules could be determined by adding a complementary ssDNA probe. The total molecular weight of Cascade increased by the mass of a single ssDNA-probe, indicating the presence of one accessible crRNA (Fig. 3.6C). For further characterization we used a combined approach of in solution and in gasphase dissociation of Cascade (Lorenzen et al. 2007; Zhou et al. 2008). By adding a low percentage of 2-propanol, Cascade was partially disrupted, resulting in a variety of Cascade sub-complexes (Fig. 3.5B, Table S3.2)). In addition to the intact Cascade (405 kDa), we also detected Cascade lacking CasA (349 kDa), and seven additional sub-complexes, the largest of which is a 307 kDa species. The difference between the 349 kDa and 307 kDa sub-complexes is 42,442 Da and likely reflects the loss of the CasB dimer, which was previously observed (Fig. 3.5A).

The apparent consecutive loss of five times a mass of around 40 kDa (from 307 kDa down to 107 kDa) is directly evident from this catalogue of sub-complexes (Fig. 3.5B). This implies that at least five CasC subunits are present in intact Cascade. This hypothesis was confirmed by subjecting each of these sub-complexes to tandem mass spectrometry (Table S3.2). In addition to CasD, CasE and one crRNA molecule, this analysis revealed the presence of even a sixth CasC subunit in the 107 kDa sub-complex (Fig. 3.5B). Combining all mass spectrometry data resulted in a Cascade

stoichiometry of $CasA_1B_2C_6D_1E_1/crRNA_1$. The theoretical mass of this complex is in excellent agreement with the experimental mass (405,095 Da versus 405,365 ± 135 Da), band intensities on protein gels (Fig. 3.1B), and elution profiles on a calibrated size exclusion column (Fig. 3.1E). Similar analyses of the purified sub-complexes revealed masses and compositions consistent with Cascade: 349 kDa for $CasB_2C_6D_1E_1/crRNA_1$ and 324 kDa for $CasC_6D_1E_1/crRNA_1$ (data not shown). The constituency of the complexes and the elimination of specific subunits in solution and under tandem mass spectrometry conditions reinforced the assignments of subunits in the structural model presented below.

Structure of Cascade

First insights into the structural organization of Cascade were obtained by single particle electron microscopy (EM) and small angle X-ray scattering (SAXS). The Cascade complex is an elongated particle with no discernible symmetry. The particle has approximate dimensions of 10 x 20 nm, and resembles a seahorse with a curled-up tail (Figs. 3.6A-C). The particle displays a striking indentation on one side which gives rise to the head and neck features of the seahorse. Three types of projection maps with minor variation were obtained for Cascade (Figs. 3.6A-C), indicative of a strong adsorption orientation preference on the carbon support film. In addition to DNA-free Cascade, we also examined the structure of Cascade with target ssDNA bound, resulting in six diverse groups of projections in multiple adsorption orientations (Figs. 3.6D-I). The comparison between DNA-free and DNA-bound Cascade projections in a similar orientation reveals differences in the head and back of the seahorse-shaped morphology of Cascade (Figs. 3.6J-L), which suggest that Cascade undergoes conformational changes upon target DNA binding.

Subunit localization

Given the availability of stable sub-complexes of Cascade and knowing their subunit compositions allowed for the investigation of subunit localization within the Cascade structure. Some of the projections show a regularly shaped and evenly spaced feature with sharp edges that spans the torso of the complex (Fig. 3.6I). This repeated feature traverses the spine of this structure and is consistent with the six copies of CasC that are present in the complex and comprise the backbone of Cascade. The position of the target DNA recognition enhancing subunit CasA was determined from difference maps between target DNA bound forms of Cascade and Cascade lacking CasA (CasBCDE) (Figs. 3.6M-O). The analysis revealed that CasA is located in the curled-up tail of the seahorse (Fig. 3.6O). The position of CasB was identified using difference maps between target DNA bound CasBCDE and CasCDE complexes (Figs. 3.6P-R). Apart

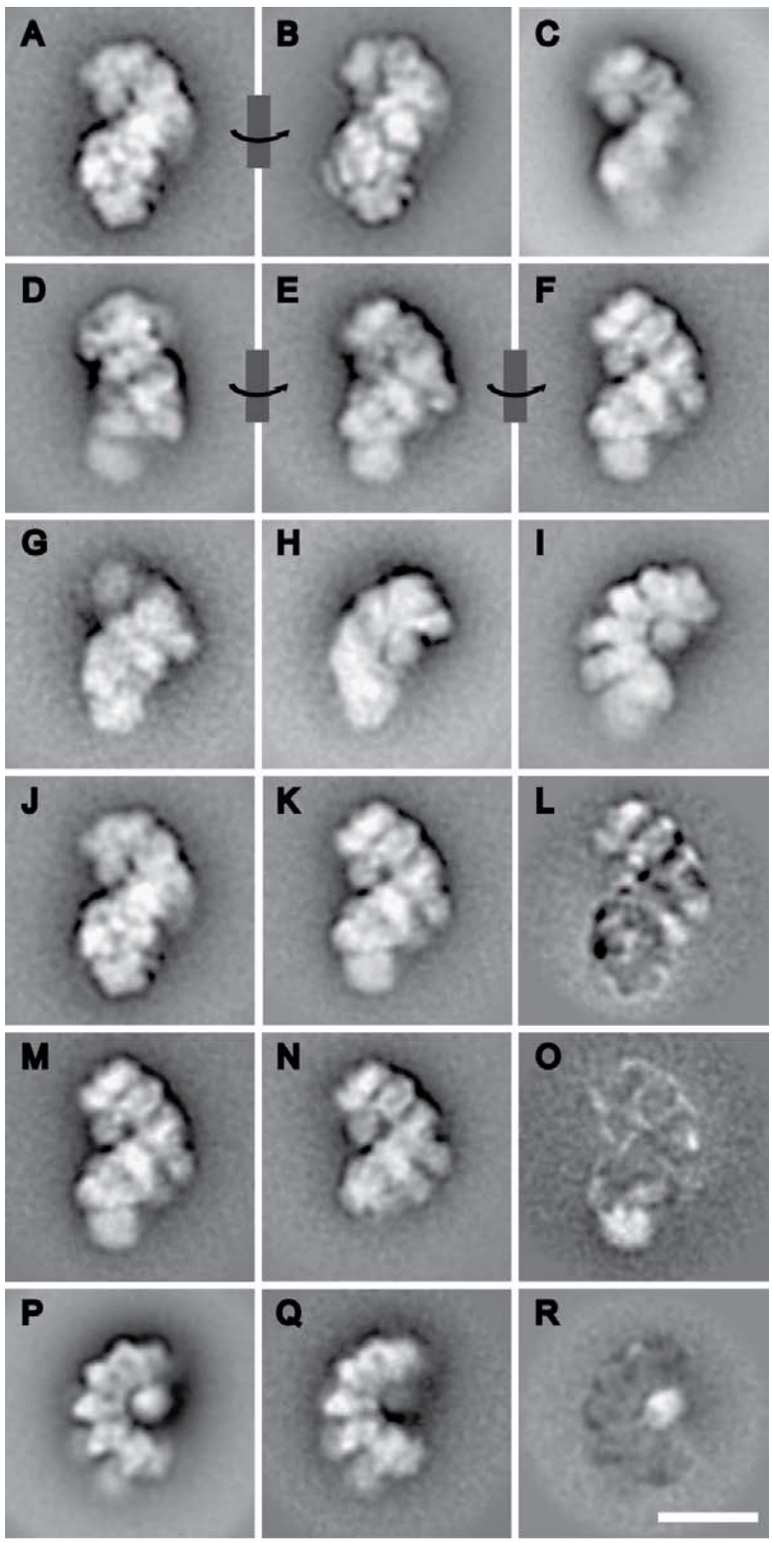


Figure 3.6. Electron Microscopic structure of Cascade. A-C) Cascade projections showing an elongated particle with 20 x 10 nm dimensions. D-I) Cascade projections bound to target ssDNA. J-L) Difference map indicating morphological changes to the Cascade particle that result from target DNA binding (L). The difference map was generated from Cascade (J) and Cascade with target ssDNA bound (K). M-O) Difference map showing the location of the CasA subunit (O). The difference map was generated from Cascade (M) and CasBCDE (N) projection maps with target ssDNA bound. P-R) Difference map showing the location of the CasB subunit (R). The difference map was generated from CasBCDE (P) and CasCDE (Q) projection maps with target ssDNA bound. On average, each image is the sum of 1,500 single particle projections. The scale bar equals 10 nm.

from the six better resolved CasC subunits in these projections of CasBCDE and CasCDE (Figs. 3.6P and 3.6Q), the difference map showed one region compatible with two CasB copies (Fig. 3.6R), consistent with the CasB homodimer observed with mass spectrometry (Fig. 3.5A). The CasB dimer is contributing the nose to the seahorse-shaped morphology of Cascade (Fig. 3.6R).

In addition to the position of some of the subunits in the Cascade structure, we also investigated the topological constraints of Cascade. Protein pulldown experiments between CasC and one of the remaining Cascade components showed that CasB, CasD and CasE form stable two-component complexes with CasC, whereas CasA does not (Fig. S3.7). The association of this subunit with the CasC backbone is mediated by CasE (Fig. S3.7).

Solution structure of Cascade

Ab initio small-angle x-ray scattering (SAXS)-based reconstruction of Cascade and Cascade bound to target DNA result in well-supported models (Fig. 3.7) that are of similar shape and size to what was observed by EM (Fig. 3.6). Scattering curves for Cascade (Fig. 3.7A) and Cascade bound to target DNA (Fig. 3.7B) were generated from data collected at two protein concentrations and include scattering vectors (q), ranging from 0.015 $Å^{-1}$ to 0.127 $Å^{-1}$ and 0.015 $Å^{-1}$ to 0.133 $Å^{-1}$ for the two complexes respectively. Guinier approximations of each curve estimates the radius of gyration (R(g)) for Cascade and for Cascade bound to target DNA at 5.6 nm. The pair-distribution function (P(r), Fig. 3.7A and 3.7B), which is the frequency of interatomic vector lengths within the scattering particle and provides real-space information about the molecule shape, was calculated from the entire scattering curve I(q). This analysis indicates that the most frequently sampled interatomic distance is 5.6 nm for each of the two complexes, which is consistent with the R(g) estimated by the Guinier analysis. The SAXS models of Cascade and Cascade bound to target DNA were generated from a total of ten independent reconstructions for each complex. These individual reconstructions were in good agreement with one another and used to generate the average models (Fig. 3.7C and 3.7D) with resulting particle volumes of 1024 and 1235 nm^3, respectively.

Cascade's association with target DNA induces a conformational change that results in a shape with fewer prominent features, including the loss of the indentation on one side which gives rise to the head and neck features of the seahorse (Fig. 3.7D). These structural changes in Cascade involve regions of the complex assigned to CasB and CasC subunits. The conformational change is more apparent when the DNA-free and

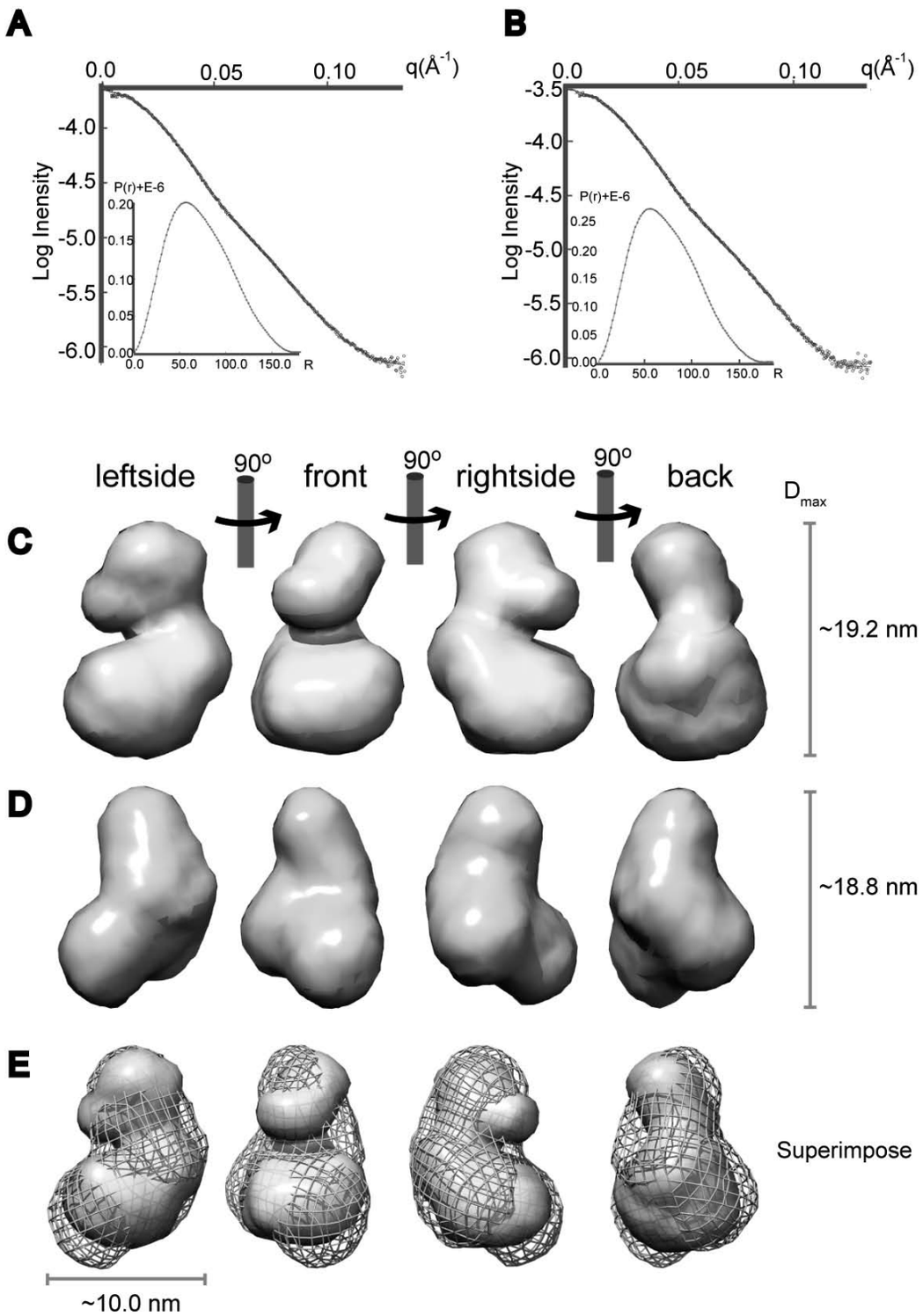


Figure 3.7. Solution scattering model of Cascade obtained with Small-Angle X-ray Scattering. For full colour version see page 130. Scattering data for Cascade were collected at 10 keV (1.24 Å λ) from two protein concentrations, and include scattering vectors (q), ranging from A) 0.015 $Å^{-1}$ to 0.127 $Å^{-1}$ for Cascade and B) 0.015 $Å^{-1}$ to 0.133 $Å^{-1}$ for Cascade bound to target DNA. The pair-distribution function (insert) indicates that the radius of gyration for both particles is ~5.6 nm. C) *Ab initio* reconstructions of Cascade reveal a seahorse shaped complex, consistent with EM imaging. D) DNA binding induces a conformational change in Cascade. E) Superposition of the solution structures of Cascade without (yellow) and with target DNA (mesh) suggest that regions of the complex assigned to the CasA and CasB are repositioned in the DNA bound state. Images have been rendered using Chimera (Goddard et al. 2005).

DNA-bound models are superimposed (Fig. 3.7E). In addition to the loss of the nose-shape feature consisting of CasB, which is so prominent in a DNA-free complex (Fig. 3.6A and 3.7A) is gone in the DNA-bound state, and the CasA tail appears to be more extended in the DNA-bound form. These changes in shape are not a consequence of DNA-induced changes to the subunit stoichiometry of the complex, as determined by mass spectrometry, but are indicative of a ligand-induced conformational change.

Discussion

Composition of Cascade

Cascade is a ribonucleoprotein complex in *E. coli* that plays a central role in CRISPR-based defence against mobile genetic elements, such as phages and conjugative plasmids. Despite its crucial role as the main effector complex, very little is known about how the host counterattack is accomplished. Cascade has a mass of 405 kDa and consists of five proteins (CasA, CasB, CasC, CasD and CasE) and one guide RNA (Brouns et al. 2008). The core of the protein complex consists of six CasC (Cse4) subunits, presumably to provide a structural backbone for the other subunits and the crRNA. Given the presence of CasC-homologues (COG1857) in the majority of Cas subtypes (Cse4, Csd2, Csh2, Cst2, Csa2, Csy2) (Makarova et al. 2006), these proteins could be components of Cascade-like protein complexes from other microbes. The next most frequently found subunit of the complex is CasD (Cas5e), which belongs to the Cas5-type (Cas5e, Cas5d, Cas5h, Cas5t, Cas5a, Csy3) RAMP protein class (Repeat-Associated Mysterious Proteins). This protein co-occurs with CasC-homologues in five subtypes (Makarova et al. 2006). CasD is present in a single copy in Cascade and together with six CasC subunits and one CasE (Cse3) subunit part of a minimal Cascade core that accommodates a single mature crRNA. This minimal core is expanded by a dimer of CasB (Cse2), a positively charged protein (pI 9.2). A crystal structure of the CasB from *Thermus thermophilus* HB8 shows a strictly α- helical protein with a conserved basic surface patch that suggests a role in nucleic acid interactions (Agari et al. 2008). Cascade is completed by one copy of CasA (Cse1), the largest and most loosely attached subunit of the five. Gene synteny analysis of the various Cas gene clusters shows that CasA and CasB only occur in the Cse-subtype, and that the CasAB pair is substituted by other Cas proteins (*e.g.* Csy1) containing the CasCD core (van der Oost et al. 2009).

The crRNA is remarkably stable when bound by Cascade or the CasBCDE and CasCDE sub-complexes, indicating that it is tightly associated with the CasCDE core of the protein complex and well shielded from cellular ribonuclease activities. Yet it

is sufficiently exposed to allow for base pairing with complementary nucleic acids. Interestingly, the length of the 5' handle is conserved among crRNAs from *E. coli*, *S. epidermidis* and *P. furiosus* (Brouns et al. 2008; Carte et al. 2008; Marraffini and Sontheimer 2008), suggesting a general mechanism of binding. A recent study in *S. epidermidis* (Csm-type) showed that differential complementarity of the 5'-handle of the crRNA with the downstream protospacer flank allows discrimination between self DNA (the CRISPR) and non-self DNA (the target) (Marraffini and Sontheimer 2010b). While non-complementarity of the 5'-handle results in a sequence that is targeted, base pairing of at least nucleotide 5, 6 and 7 of the 5'-handle provides self protection (Marraffini and Sontheimer 2010b). In other CRISPR/Cas systems the regions flanking the protospacer (i.e. opposite of the 5' or 3'-handle of the crRNA) often harbour short sequence motifs that are known as CRISPR motifs (Deveau et al. 2008; Horvath et al. 2008) or protospacer adjacent motifs (PAM) (Mojica et al. 2009). The relevance of these short nucleotide sequences was originally shown in *S. thermophilus* by sequencing phages that had overcome host immunity by mutating a single nucleotide of the motif (Deveau et al. 2008). In *E. coli* this region contains the PAM sequence 5'-AWG-protospacer-3', which gives rise to non-basepaired rC:dT, rC:dW and a basepaired rG:dC between the sixth, seventh and eighth base of the 5'-handle, respectively (Fig. 3.8A). This signature in the target DNA-Cascade complex might serve to verify that a target sequence has been located before downstream interference processes are commenced.

The crRNA is further characterized by the presence of a 5'-hydroxyl group, as was observed for Cas6-generated crRNA (Carte et al. 2008). Eukaryotic small interfering RNA (siRNA) and microRNA (miRNA) by contrast need to be 5'-phosphorylated in order to bind to Argonaute and serve as a guide for the RNA-induced silencing complex (RISC) (Ma et al. 2005). *E. coli* crRNAs are unmodified, unlike plant siRNAs and miRNAs as

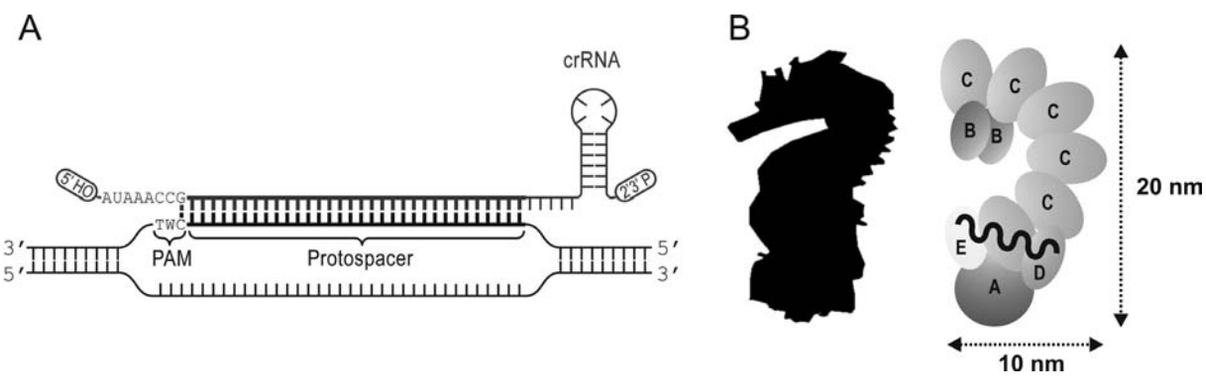


Figure 3.8. For full colour version see page 130. A) Schematic diagram of crRNA base paired to double stranded target DNA, indicating the local strand displacement, and the additional rG-dC basepair between the eighth base of the crRNA (rG) with the PAM (dC). B) Seahorse morphology and structural model of Cascade.

well as vertebrate piwi-interacting RNAs for example, which are methylated at the 2'-hydroxyl group of the 3'-terminal ribonucleotide to prevent uridylation and associated destabilization (Li et al. 2005; Houwing et al. 2007). The cyclic 2',3'-phosphate of crRNA results from a metal ion-independent endonuclease activity on the 3'-side of the phosphodiester bond of pre-crRNA, as suggested previously (Carte et al. 2008). These initial pre-crRNA endonuclease cleavage products are the mature form of the crRNA in *E. coli*, in contrast to crRNAs from *P. furiosus*, *S. solfataricus* and *S. epidermidis* which are trimmed at the 3'-end (Hale et al. 2009; Lillestol et al. 2009; Marraffini and Sontheimer 2010b). It is worth mentioning that the cyclic phosphate 3'-end is not a substrate for *E. coli* poly(A) RNA polymerase (Zaug et al. 1996), and this explains why only shorter, apparently partly degraded crRNAs were cloned and sequenced previously (Brouns et al. 2008).

Cascade targets DNA

Multiple lines of evidence suggest that crRNAs directly target invader DNA in the *E. coli* (Cse), *S. epidermidis* (Csm) and likely also *S. thermophilus* (Csn) model systems (Barrangou et al. 2007; Brouns et al. 2008; Marraffini and Sontheimer 2008). However, molecular evidence of Cas-effector complexes recognizing their target DNA has been lacking, limiting our understanding of how CRISPR/Cas systems operate. Binding studies showed that Cascade is guided by the crRNA to dsDNA sequences containing the protospacer without supplemented co-factors such as ATP. This surprising characteristic of Cascade is physiologically important, since most cellular and invader DNA is of double-stranded nature. The ATP-independence of this scanning process allows cells to continuously survey nucleic acids for crRNA matches without major energy investments. In addition to sequence-specific DNA recognition, Cascade also interacts non-specifically with DNA. Cascade sub-complexes lacking CasA, however, display only sequence-specific DNA recognition. Direct competition assays between Cascade and CasBCDE demonstrate the superiority of Cascade in locating the protospacer DNA, suggesting a role for CasA as an enhancer. Cascade thus appears to interact with DNA in two different modes, non-specifically via CasA and sequence-specifically via Watson-Crick base pairing of the crRNA (Fig. 3.8A). We hypothesize that the non-specific affinity of Cascade for DNA enables Cascade to associate fast enough with incoming foreign DNA to neutralize a rapidly proliferating phage or conjugative plasmid in a cell. Further analysis showed that although Cascade exhibits little non-specific affinity for RNA, it binds complementary ssRNA as well. Base pairing of the crRNA with mRNA occurs when the coding strand of an invading nucleic acid is targeted. This could titrate out unpaired Cascade-bound crRNAs, reducing the effectiveness of the immune response when the coding strand is targeted (Brouns et al. 2008). Cascade

does not recognize a protospacer within a dsRNA molecule, indicating that the ability to locate protospacers in a nucleic acid duplex is restricted to dsDNA. Recently, a distinct Cas protein complex of the Cmr-subtype from *P. furiosus* was shown to cleave ssRNA complementary to the bound guide RNA (Hale et al. 2009), akin to RISC in eukaryotes. Target RNA hydrolysis takes place 14 bases from the 3'-end of the guide RNA, suggesting that target RNA cleavage occurs by a ruler mechanism. The difference in target nucleic acid - RNA or DNA - between the Cmr- and Cse-subtypes may illustrate the remarkable diversity present among CRISPR/Cas systems.

Structural basis for target DNA recognition

Cascade displays an unusual seahorse-shaped architecture. The backbone of this structure is composed of six CasC subunits that are arranged in an arch, representing the torso of the Cascade particle. CasC only assembles into this stable hexameric arrangement in the presence of CasD, CasE and crRNA. Although the location of these latter three components in the structure could not be determined by direct methods, mass spectrometry showed the existence of a 107 kDa 2-propanol-induced sub-complex containing single copies of CasC, CasD, CasE and crRNA (Fig. 3.5B). This subassembly of Cascade implies that these four components are in close proximity of each other. Further disassembly of Cascade reveals a CasE-crRNA particle, suggesting that CasE holds on tightly to the crRNA after pre-crRNA cleavage. Despite the fact that the location of the crRNA remains unclear, we expect that the 61 nucleotide molecule is bound by a substantial part of Cascade. In an unpaired state, the crRNA likely experiences some conformational flexibility, whereas in a DNA base-paired state, the conformation of the spacer sequence is constrained by the formation of the RNA:DNA hybrid helix, which in contrast to the B-type dsDNA helix is close to A-type (Noy et al. 2005). This constraint might induce the conformational change observed in the EM and SAXS structures of Cascade, and may serve to recruit Cas3, which is required for neutralizing the targeted invader. The length of the 32 bp spacer sequence base paired to its target ssDNA is approximately 8.2 nm. Together with the 5' and 3' handles, the crRNA could be up to 10 nm in length.

These combined observations result in a structural model in which the position of CasA, CasB and CasC can be assigned with confidence, contributing the curled-up tail, nose and torso of the seahorse, respectively (Fig. 3.8B). The model predicts that CasD, CasE and at least part of the crRNA are located in the main body of Cascade at the tail-end of the CasC spine, in the proximity of CasA. The crRNA could also be bound along the spine of the CasC backbone and in this way define the number of CasC subunits by the length of the crRNA.

In conclusion, Cascade is a highly unusual ribonucleoprotein complex capable of specifically recognizing double stranded invader DNA archived in the CRISPR blacklist. The direct attack on invader DNA rather than its RNA transcripts may be a defence strategy that enables prokaryotes to neutralize the invasive source of selfish genetic elements effectively.

Methods

Protein production and purification

Cascade complexes were produced and purified as described previously (Brouns et al. 2008), using the expression plasmids listed in Table S3.3. Cascade was routinely purified with an N-terminal Strep-tag II fused to CasB (or CasC in CasCDE). Size exclusion chromatography (Superdex 200 HR 10/30 (GE)) was performed using 20 mM Tris-HCl (pH 8.0), 0.1 M NaCl, 1 mM dithiotreitol. Cascade preparations (~0.3 mg) were incubated with DNase I (Invitrogen) in the presence of 2.5 mM $MgCl_2$ for 15 min at 37 °C prior to size exclusion analysis. Copurified nucleic acids were isolated by extraction using an equal volume of phenol:chloroform:isoamylalcohol (25:24:1) pH 8.0 (Fluka), and incubated with either DNase I (Invitrogen) supplemented with 2.5 mM MgCl2 or RNase A (Fermentas) for 10 min at 37 ºC.

Lambda phage infection

Plaque assays were performed using bacteriophage Lambda and the efficiency of plaquing (EOP) was calculated as described previously (Brouns et al. 2008).

HPLC purification of crRNA

All samples were analyzed by ion-pair reversed-phased-HPLC on an Agilent 1100 HPLC with UV_{260nm} detector (Agilent) using a DNAsep column 50 mm × 4.6 mm I. D. (Transgenomic, San Jose, CA). The chromatographic analysis was performed using the following buffer conditions: A) 0.1 M triethylammonium acetate (TEAA) (pH 7.0) (Fluka); B) buffer A with 25% LC MS grade acetonitrile (v/v) (Fisher). The crRNA was obtained by injecting purified intact Cascade at 75 °C using a linear gradient starting at 15% buffer B and extending to 60% B in 12.5 min, followed by a linear extension to 100% B over 2 min at a flow rate of 1.0 ml/min. Hydrolysis of the cyclic phosphate terminus was performed by incubating the HPLC-purified crRNA in a final concentration of 0.1 M HCl at 4 ºC for 1 hour. The samples were concentrated to 5-10 µl on a vacuum concentrator (Eppendorf) prior to ESI-MS analysis.

ESI-MS analysis of crRNA

Electrospray Ionization Mass spectrometry was performed in negative mode using an UHRTOF mass spectrometer (maXis) or an HCT Ultra PTM Discovery instrument (both Bruker Daltonics), coupled to an online capillary liquid chromatography system (Ultimate 3000, Dionex, UK). RNA separations were performed using a monolithic (PS-DVB) capillary column (200 mm ´ 50 mm I.D., Dionex, UK). The chromatography was performed using the following buffer conditions: C) 0.4 M 1,1,1,3,3,3,-Hexafluoro-2-propanol (HFIP, Sigma- Aldrich) adjusted with triethylamine (TEA) to pH 7.0 and 0.1 mM TEAA, and D) buffer C with 50% methanol (v/v) (Fisher). RNA analysis was performed at 50 ºC with 20% buffer D, extending to 40% D in 5 min followed by a linear extension to 60% D over 8 min at a flow rate of 2 µl/min.

Protein and Native Mass spectrometry

Cascade was analyzed in 0.15 M ammonium acetate (pH 8.0) at a protein concentration of 5 µM. This protein preparation was obtained by five sequential concentration and dilution steps at 4 ºC using a centrifugal filter with a cut-off of 10 kDa (Millipore). Proteins were sprayed from borosilicate glass capillaries and analyzed on a LCT electrospray time-of-flight or modified quadrupole time-of-flight instruments (both Waters, UK) adjusted for optimal performance in high mass detection (Tahallah et al. 2001; van den Heuvel et al. 2006). Exact mass measurements of the individual Cas proteins were acquired under denaturing conditions (50% acetonitrile, 50% MQ, 0.1% formic acid). Sub-complexes in solution were generated by the addition of 2-propanol to the spray solution to a final concentration of 5% (v/v). Instrument settings were as follows; needle voltage ~1.2 kV, cone voltage ~175 V, source pressure 9 mbar. Xenon was used as the collision gas for tandem mass spectrometric analysis at a pressure of 1.5 10^{-2} mbar. The collision voltage varied between 10-200 V.

Electrophoretic mobility shift assays (EMSA)

EMSA was performed by incubating Cascade, CasBCDE or CasCDE with 1nM labelled nucleic acid in 50 mM Tris-Cl pH 7.5, 100 mM NaCl. Salmon sperm DNA (Invitrogen) was used as competitor. The EMSA reactions were incubated at 37 °C for 20-30 min prior to electrophoresis on 5% polyacrylamide gels. The gels were dried and analyzed using a PMI phosphor imager (Bio-Rad). DNA targets were gel-purified long oligonucleotides (Isogen Life Sciences or Biolegio), listed in Table S3.3. The oligonucleotides were end-labelled using γ ^{32}P-ATP (PerkinElmer) and T4 kinase (Fermentas). Double-stranded DNA targets were prepared by annealing complementary oligonucleotides and digesting remaining ssDNA with Exonuclease I (Fermentas). Labelled RNA targets were in vitro transcribed using T7 Maxiscript or T7 Mega Shortscript kits (Ambion) with

α ^{32}P-CTP (PerkinElmer) and removing template by DNase I (Fermentas) digestion. Double stranded RNA targets were prepared by annealing complementary RNAs and digesting surplus ssRNA with RNase T1 (Fermentas), followed by phenol extraction.

Electron microscopy

Purified protein samples were negatively stained with 2% uranyl acetate on glow-discharged carbon-coated copper grids. R44 target ssDNA (BG3028, Table S3.3) was added to Cascade sub-complexes loaded with R44 crRNA in a two-fold excess at least 5 min prior to sample preparation. Electron microscopy was performed on a Philips CM120 equipped with a LaB6 tip operating at 120 kV. Images were recorded with a Gatan 4000 SP 4K slow-scan CCD camera at 130,000× magnification at a pixel size (after binning the images) of 0.23 nm at the specimen level with GRACE software for semi-automated specimen selection and data acquisition (Oostergetel et al. 1998). Single particle projections were selected from micrographs mainly by reference-based automated particle selection procedure incorporated into GRIP (GRoningen Image Processing) software (van Heel et al. 2000). Approximately 400,000 single particles were selected and extracted from 17,000 electron micrographs. Single particle data sets were analyzed with the GRIP software using multi-reference alignments and no-reference alignments, multivariate statistical analysis, and hierarchical ascendant classification. The final two-dimensional projection maps were calculated from the best resolved classes by summing the best 5–20% of the projections based on the correlation coefficient determined in the alignment step.

Small Angle X-ray Scattering

SAXS data were collected at the Advanced Light Source (Lawrence Berkeley National Laboratory) on beamline 7.3.3. Solution scattering of Cascade was collected at room temperature (~22 °C) using at least two different concentrations (between 1 mg/ml and 18 mg/ml) in a 20 µl sample at 10 keV (1.24 Å λ). The sample-to-detector distance was set to 3056.69 m resulting in scattering vectors (q) ranging from 0.012 Å^{-1} to 0.127 Å^{-1} for Cascade, and 0.0147 Å^{-1} to 1.334 Å^{-1} for Cascade bound to target DNA. Scatter plots for the low and high concentrations were merged and background subtracted using PRIMUS (Konarev et al. 2003). One-dimensional scatter curves were transformed and distance distribution functions P(r) were calculated using GNOM (Svergun 1992). The pair-distribution function P(r), the frequency of interatomic vector lengths within the scattering particle, was calculated from the entire scattering curve I(q), by an indirect Fourier transform using GNOM. Ten independent models of each complex were generated using a simulated annealing method in DAMMIF (Franke and Svergun 2009). *Ab initio* reconstructions for each complex were aligned, filtered

and averaged based on occupancy using DAMAVER (Volkov and Svergun 2003). The SAXS bead models were converted to volumetric format using the pdb2vol convolution kernel in the Situs software package (Wriggers et al. 1999). Volumes were calculated using VOIDOO (Kleywegt and Jones 1994).

Acknowledgments

We thank Luc van Heereveld and Man H. Lai for experimental contributions, and Eric Schaible, Peter Zwart and Marcel Bokhove for technical support and for assistance with post processing of SAXS data. This work was financially supported by an NWO Vici grant to JvdO (865.05.001), Veni grants to SJJB (863.08.014) and EvD (700.58.402), NWO TOP grant to EJB, EPSRC and BBSRC grants to MD. ML was financially supported by the Wenner-Gren Foundations, ERW by Spinoza resources awarded to Willem M. de Vos. APLS is an RCUK Academic Fellow. BW is a Howard Hughes Medical Institute Fellow of the Life Sciences Research Foundation. We thank the Netherlands Proteomics Center for financial support.

Supplementary figures

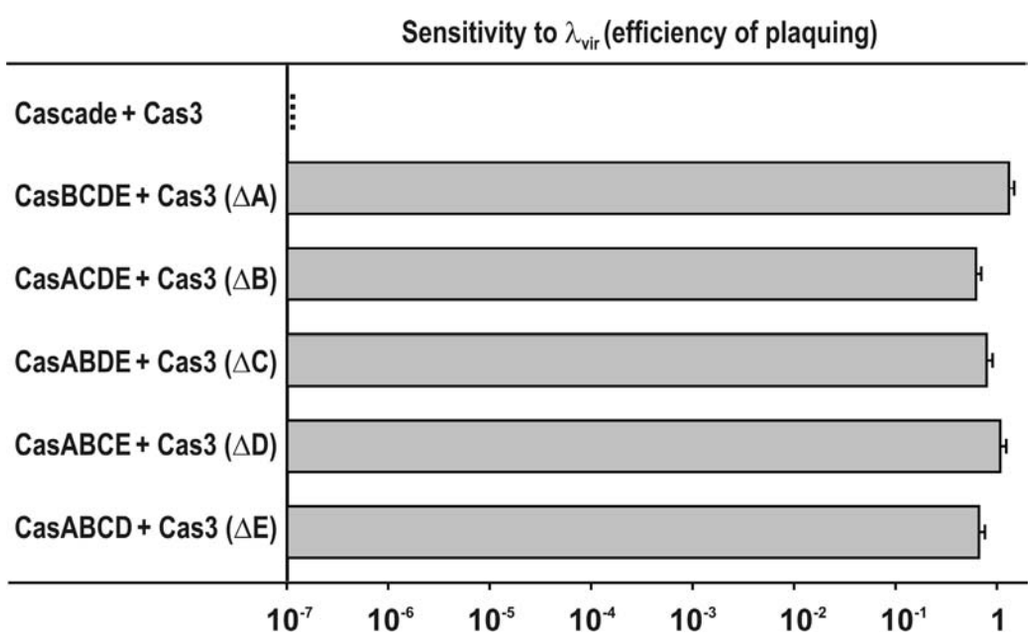


Figure S3.1. All Cascade subunits are required for immunity. The efficiency of plaquing (EOP) was determined in *E. coli* BL21-AI containing complete Cascade, phage λ-targetting $T_{1\text{-}4}$ CRISPR (see Fig. S3.2) and Cas3 (Brouns et al. 2008), and for strains expressing Cascade sub-complexes lacking one of the Cas proteins. Host strain *E. coli* BL21-AI does not contain *cas* genes (Barrick et al. 2009). Error bars indicate one standard deviation.

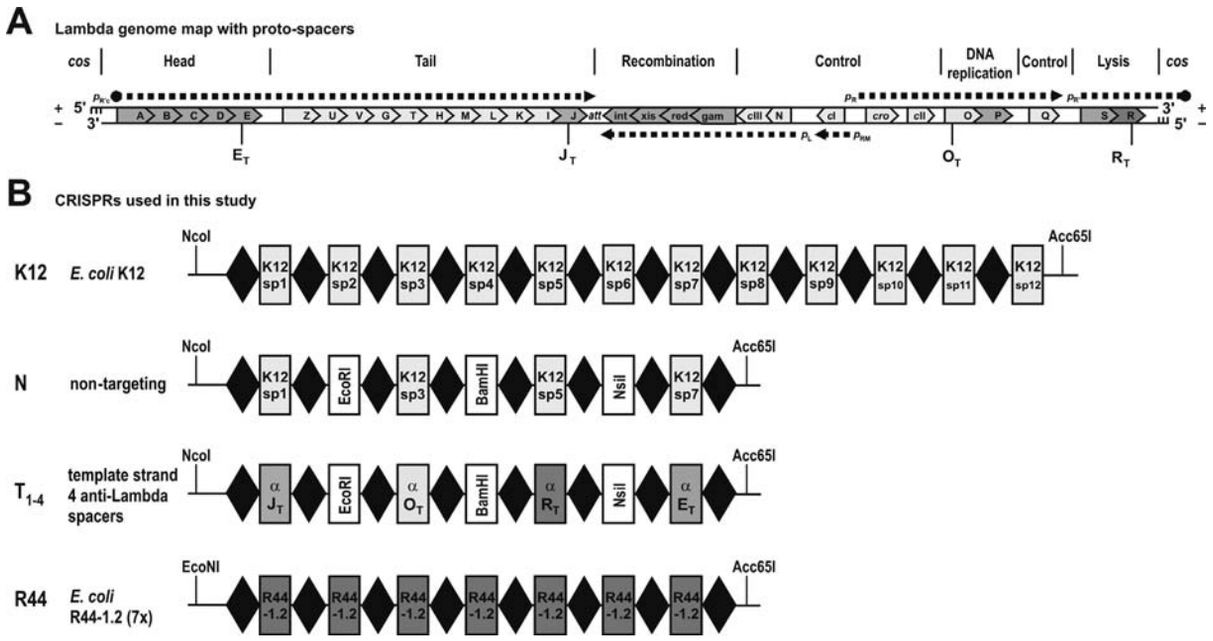


Figure S3.2. CRISPRs used in this study. A) Phage λ genome map indicating the main genes and transcripts (dotted arrows), and the positions of the protospacers on the coding or template strand. B) Schematic overview of CRISPRs used in this study (repeats: diamonds, spacers: rectangles). The N CRISPR was designed as a non-targeting control containing the naturally occurring spacers 1, 3, 5, and 7 from *E. coli* K12, which have no homology to any known phage (Brouns et al. 2008). The number of plaque forming units obtained in the presence of this CRISPR was used to calculate the efficiency of plaquing (Fig. S3.1). The uniform CRISPR R44 was designed based on a natural spacer at the second position of the CRISPR in *E. coli* R44 (ECOR44) (Ochman and Selander, 1984; Mojica et al. 2005). This spacer was derived from the *upfA* gene of phage P1. The R44 CRISPR was designed such that the spacer was repeated seven times. The sequences of the CRISPRs are provided below.

K12 CRISPR, pWUR396

GCGTACCATGGCATAAGGAAATGTTACATTAAGGTTGGTGGGTTGTTTTTATGGGAAAAAATGCTTTAAGAACAAATGTATACTTTTAGAGAGT
TCCCCGCGCCAGCGGGGATAAACCGCTTTCGCAGACGCGCGGCGATACGCTCACGCAGAGTTCCCCGCGCCAGCGGGGATAAACCGCAGCCGAA
GCCAAAGGTGATGCCGAACACGCTGAGTTCCCCGCGCCAGCGGGGATAAACCGGGCTCCCTGTCGGTTGTAATTGATAATGTTGAGAGTTCCCC
GCGCCAGCGGGGATAAACCGTTTGGATCGGGTCTGGAATTTCTGAGCGGTCGCGAGTTCCCCGCGCCAGCGGGGATAAACCGCGAATCGCGCAT
ACCCTGCGCGTCGCCGCCTGCGAGTTCCCCGCGCCAGCGGGGATAAACCGTCAGCTTTATAAATCCGGAGATACGGAAACTAGAGTTCCCCGCG
CCAGCGGGGATAAACCGGACTCACCCCGAAAGAGATTGCCAGCCAGCTTGAGTTCCCCGCGCCAGCGGGGATAAACCGCTGCTGGAGCTGGCTG
CAAGGCAAGCCGCCCAGAGTTCCCCGCGCCAGCGGGGATAAACCGGGGGGGCGCATGACCGTAAACATTATCCCCCGGGAGTTCCCCGCGCCAGC
GGGGATAAACCGGGAGTTCAGACATAGGTGGAATGATGGACTACGAGTTCCCCGCGTTAGCGGGGATAAACCGCCCGGTAGCCAGGTTTGCAAC
GCCTGAACCGAGAGTTCCCCGCGCCAGCAGGGATAAACCGGCAACGACGGTGAGATTTCACGCCTGACGCTGGGTACCGGACC

Non-targeting CRISPR (N), pWUR477

GGCGCGCCATGGAAACAAAGAATTAGCTGATCTTTAATAATAAGGAAATGTTACATTAAGGTTGGTGGGTTGTTTTTATGGGAAAAAATGCTTT
AAGAACAAATGTATACTTTTAGAGAGTTCCCCGCGCCAGCGGGGATAAACCGCTTTCGCAGACGCGCGGCGATACGCTCACGCAGAGTTCCCCG
CGCCAGCGGGGATAAACCGCAGCCGAAGCCAAAGAATTCGCCGAACACGCTGAGTTCCCCGCGCCAGCGGGGATAAACCGGGCTCCCTGTCGGT
TGTAATTGATAATGTTGAGAGTTCCCCGCGCCAGCGGGGATAAACCGTTTGGATCGGGTCTGGATCCTCTGAGCGGTCGGAGTTCCCCGCGCCA
GCGGGGATAAACCGCGAATCGCGCATACCCTGCGCGTCGCCGCCTGGAGTTCCCCGCGCCAGCGGGGATAAACCGTCAGCTTTATAAATATGCA
TATACGGAAACTAGAGTTCCCCGCGCCAGCGGGGATAAACCGGACTCACCCCGAAAGAGATTGCCAGCCAGCTTGAGTTCCCCGCGCCAGCGGG
GATAAACCGCAGCTCCCATTTTCAAACCCATCAAGACGCGGTACCTTAATTAA

Template CRISPR (T_{1-4}), pWUR478

GGCGCGCCATGGAAACAAAGAATTAGCTGATCTTTAATAATAAGGAAATGTTACATTAAGGTTGGTGGGTTGTTTTTATGGGAAAAAATGCTTT
AAGAACAAATGTATACTTTTAGAGAGTTCCCCGCGCCAGCGGGGATAAACCGCTGAGTGTGATCGATGCCATCAGCGAAGGGCCGAGTTCCCCG
CGCCAGCGGGGATAAACCGCAGCCGAAGCCAAAGAATTCGCCGAACACGCTGAGTTCCCCGCGCCAGCGGGGATAAACCGCAAGCAACAGGCAG
GCGTGACAGCCAGCAAACGAGTTCCCCGCGCCAGCGGGGATAAACCGTTTGGATCGGGTCTGGATCCTCTGAGCGGTCGGAGTTCCCCGCGCCA
GCGGGGATAAACCGTGGGATGCCTACCGCAAGCAGCTTGGCCTGAAGAGTTCCCCGCGCCAGCGGGGATAAACCGTCAGCTTTATAAATATGCA
TATACGGAAACTAGAGTTCCCCGCGCCAGCGGGGATAAACCGTGACAAGTCCACGTATGACCCGACCGACGATAGAGTTCCCCGCGCCAGCGGG
GATAAACCGCAGCTCCCATTTTCAAACCCATCAAGACGCGGTACCTTAATTAA

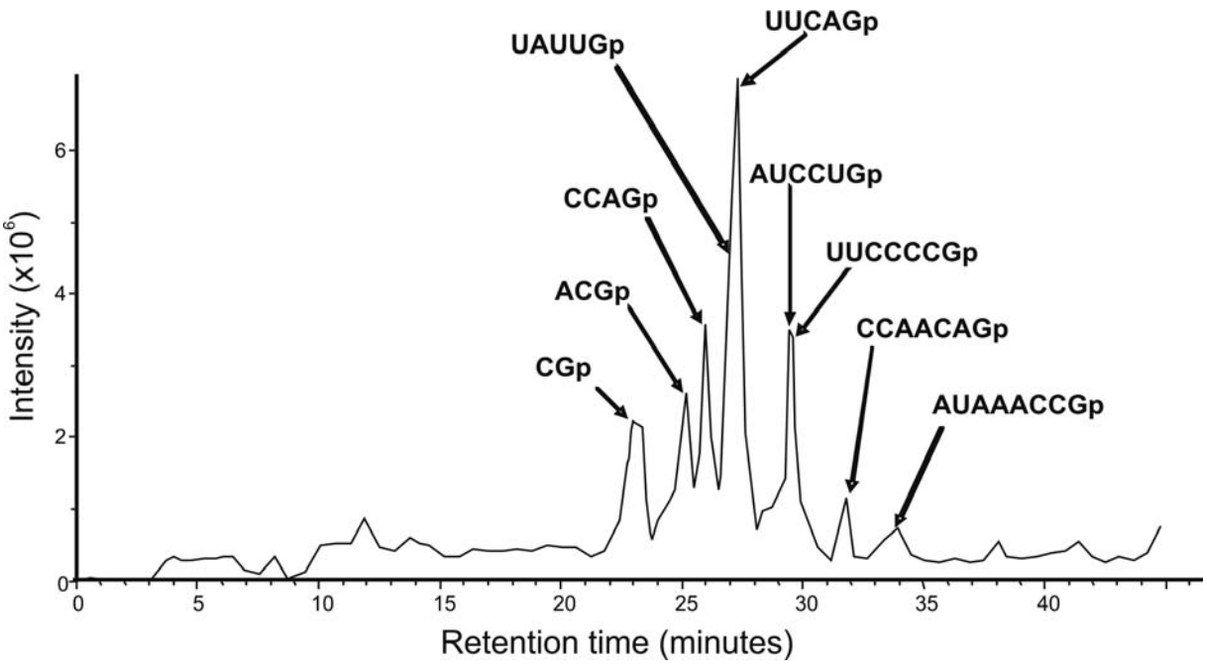
ECOR44-1.2 (7x), pWUR547

CCTGCATTAGGTAATACGACTCACTATAGGATAAACCGACGGTATTGTTCAGATCCTGGCTTGCCAACAGGAGTTCCCCGCGCCAGCGGGGATA
AACCGACGGTATTGTTCAGATCCTGGCTTGCCAACAGGAGTTCCCCGCGCCAGCGGGGATAAACCGACGGTATTGTTCAGATCCTGGCTTGCCA
ACAGGAGTTCCCCGCGCCAGCGGGGATAAACCGACGGTATTGTTCAGATCCTGGCTTGCCAACAGGAGTTCCCCGCGCCAGCGGGGATAAACCG
ACGGTATTGTTCAGATCCTGGCTTGCCAACAGGAGTTCCCCGCGCCAGCGGGGATAAACCGACGGTATTGTTCAGATCCTGGCTTGCCAACAGG
AGTTCCCCGCGCCAGCGGGGATAAACCGACGGTATTGTTCAGATCCTGGCTTGCCAACAGGAGTTCCCCGCGCCAGCGGGGATAAACCGGGTAC
C

A)

5'OH–AUAAACCGACGGUAUUGUUCAGAUCCUGGCUUGCCAACAGGAGUUC-CCCGCGCCAGCGGGG–2', 3'-P

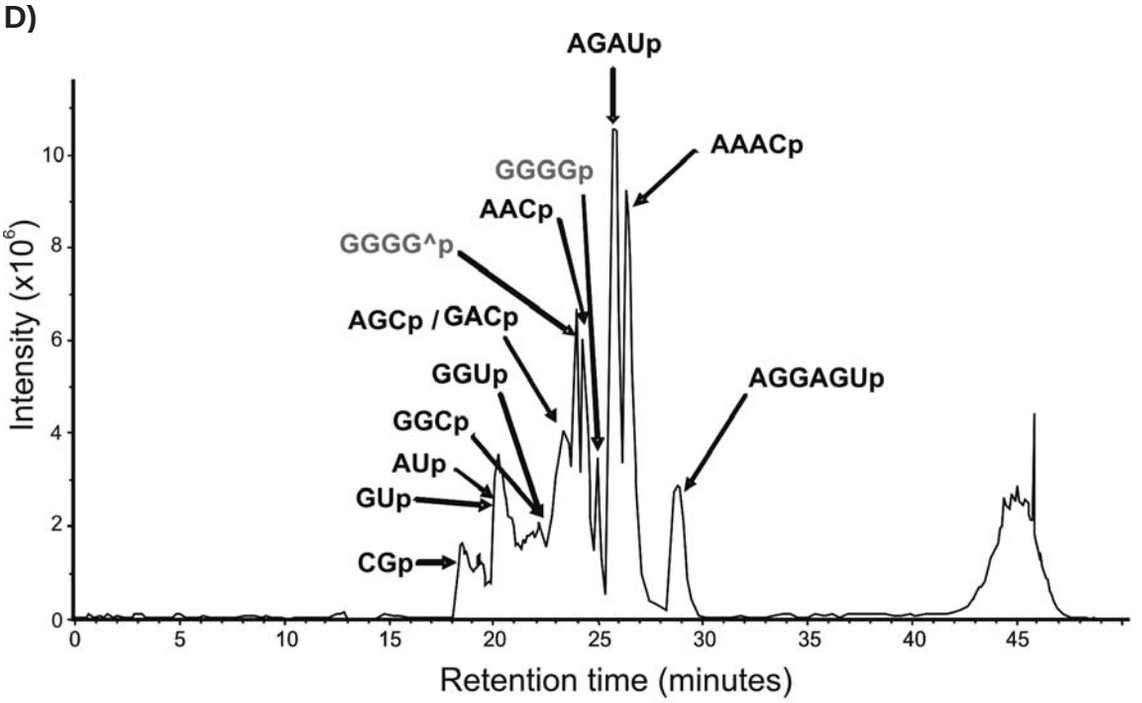
B)



C)

Sequence	Position	Theoretical mass (Da)	Experimental mass (Da)
AUAAACCGp	A1:G8	2596.597	2596.2
ACGp	A9:G11	997.617	997.1
UAUUGp	U13:G17	1610.94	1610.1
UUCAGp	U18:G22	1609.955	1609.3
AUCCUGp	A23:G28	1915.139	1914.9
CUUGp	C30:G33	1280.745	1280.2
CCAACAGp	C34:G40	2266.403	2266.1
AGp	A42:G43	692.433	692.1
UUCCCCGp	U44:G50	2196.298	2195.3
CCAGp	C53:G56	1302.801	1302.3

D)



E)

Sequence	Position	Theoretical mass (Da)	Experimental mass (Da)
AUp	A1:U2 / A14:U15	653.393	653.1
AAACp	A3:C6	1310.826	1310.4
GACp / AGCp	G8:C10 / A55:C57	997.617	997.1
GGUp	G11:U13	1014.602	1014.2
GUp	G17:U18	669.393	669.0
AGAUp	A21:U24	1327.811	1327.3
GGCp	G28:C30	1013.617	1013.4
AACp	A36:C38	981.617	981.2
AGGAGUp	A39:U44	2018.229	2017.7
GGGGp	G58:G61	1398.851	1398.4
GGGG^p	G58:G61	1380.836	1380.4

Figure S3.3. RNase digests of the crRNA. A) Sequence of mature crRNA with the R44 spacer sequence shaded, and the hairpin underlined. B) Base peak chromatogram of the RNase T1 digest of mature crRNA. The predominant oligoribonucleotide peaks assigned to the mature crRNA are highlighted. C) Summary of the identified oligoribonucleotides assigned to mature crRNA from the RNase T1 digest. D) Base peak chromatogram of the RNase A digest of mature crRNA. The predominant oligoribonucleotide peaks assigned to the mature crRNA are highlighted. E) Summary of the identified oligoribonucleotides assigned to mature crRNA from the RNase A digest. ^ indicates cyclic 2',3'-phosphate.

MS analysis of RNase digests. RNase T1 and RNase A digests of mature crRNA was performed using an ESI-MS/MS using a HCT Esquire Quadruple Ion Trap (Bruker Daltonics) coupled to a Dionex. The oligoribonucleotide mixture was separated on a PepMap C-18 RP capillary column (300 µm x 150 mm I.D., Dionex, UK) at 50 °C using a gradient condition starting at 20% buffer D (0.4 M 1,1,1,3,3,3,-Hexafluoro-2-propanol (Sigma-Aldrich) adjusted with triethylamine (TEA) to pH 7.0, 0.1 mM TEAA, and 50% methanol (v/v) (Fisher)) and extending to 35% D in 20 min at a flow rate of 2 µl /min. The mass spectrometer was set select a mass range of 250−1500 m/z and the capillary voltage was kept at -3650 V. Oligoribonucleotides with -2 to -4 charge states were selected for tandem mass spectrometry using collision induced dissociation. The theoretical masses of the crRNA and predicted digests were determined using the Mongo Oligo Mass Calculator.

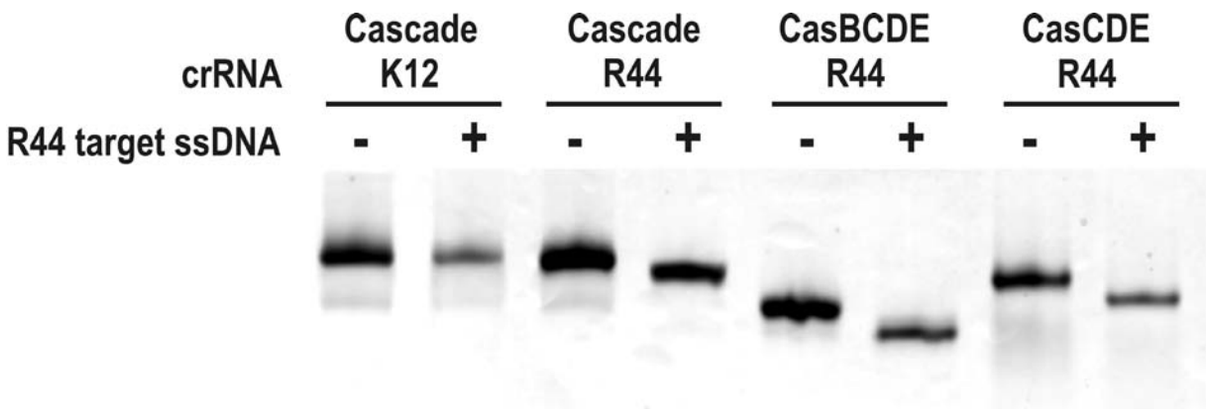


Figure S3.4. Native PAGE of Cascade-target DNA complexes. Coomassie-blue stained native PAGE analysis of Cascade and sub-complexes binding ssDNA oligonucleotides complementary to the R44 crRNA (BG3028, see Table S3.3). The gel shows the increased migration rates of Cascade complexes loaded with uniform R44 crRNA in the presence their complementary target ssDNA due to the additional charge negative charge of the ssDNA. By contrast, the migration rate of Cascade loaded with different crRNAs derived from the *E. coli* K12 CRISPR I array that does not contain the R44 spacer is not affected. In addition, differences in migration rates of the various complexes (Cascade, CasBCDE and CasCDE) are visible.

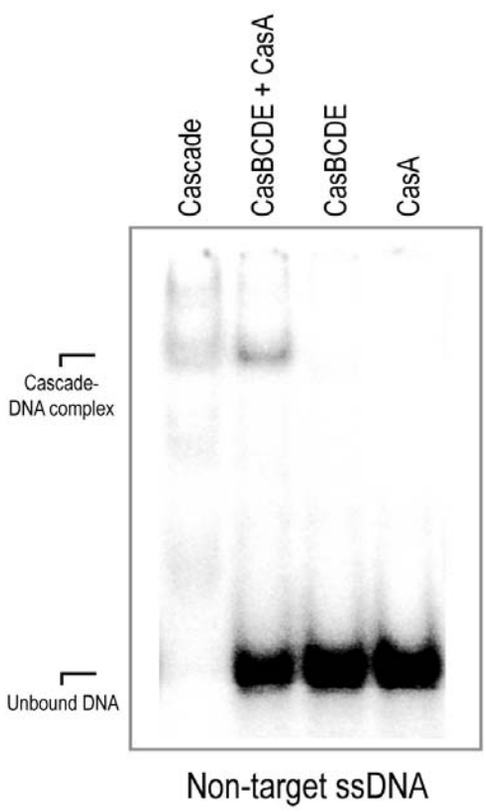


Figure S3.5. CasA complementation of non-specific DNA binding. Binding ability of uniform crRNA-loaded (R44) Cascade, CasBCDE + CasA, CasBCDE and CasA to non-target ssDNA is shown. While CasA does not exhibit non-specific DNA binding on by itself, it is able to restore the non-specific DNA binding of CasBCDE.

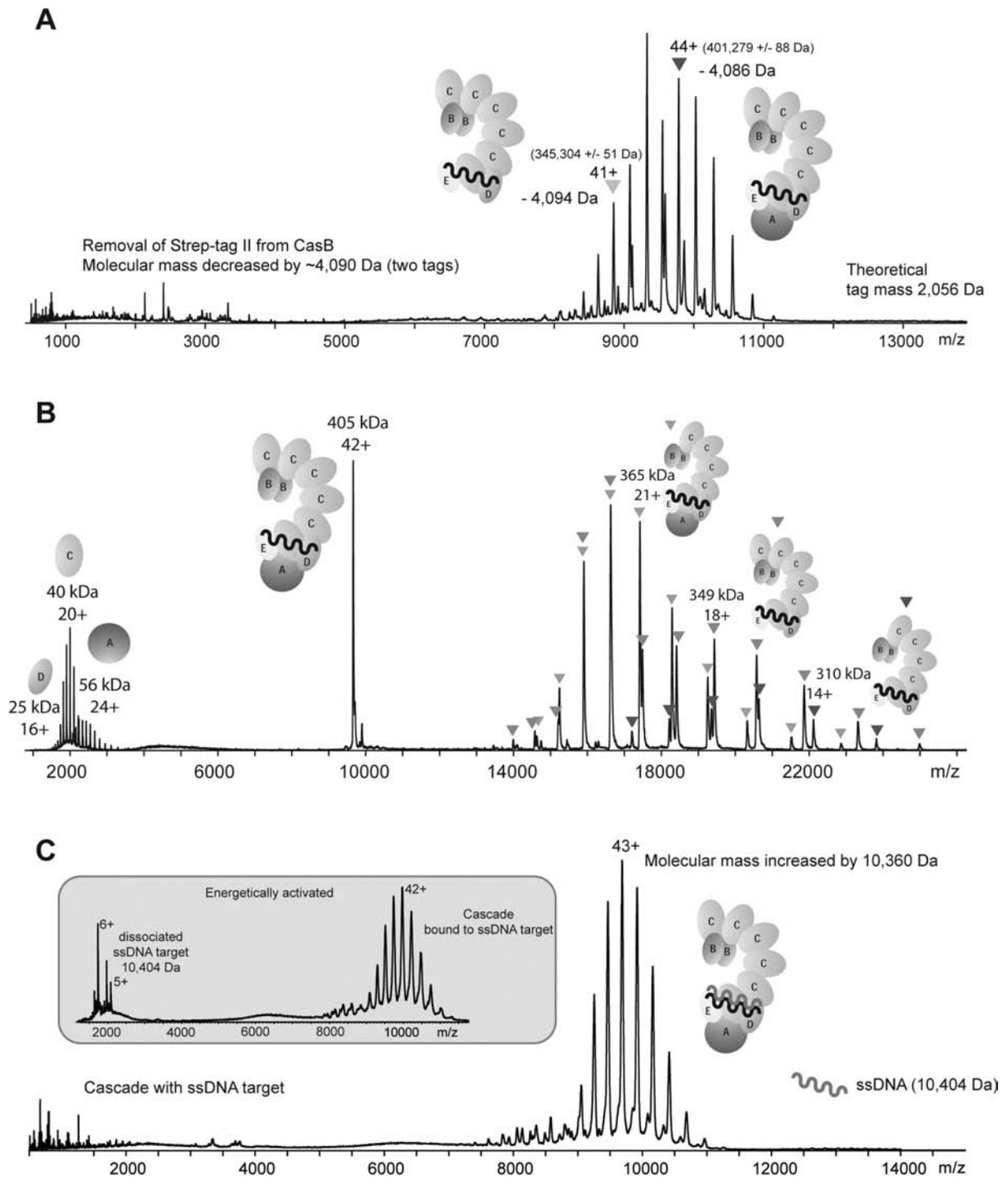


Figure S3.6. For full colour version see page 131. A) Native mass spectrum of Cascade after treatment with HRV3C protease. A dominant species with a mass of 401,279 Da (blue triangle) was observed, confirming the presence of two copies of CasB in the intact assembly. Indicated by the green triangles is the complex lacking CasA. B) Tandem mass spectrum of the 42+ ion of Cascade. Besides the dissociation of CasA (green) also CasC (orange) dissociated from the complex. The complex lacking CasA further expels a CasC subunit to form a 310 kDa Cascade subcomplex (blue). The low m/z region of the spectrum shows the dissociated CasA, CasC and CasD proteins. Overlapping peaks of two different complexes are indicated by two colours. C) Native mass spectrum of Cascade bound to the ssDNA-probe. The mass of the complex increased by 10,201 Da, indicating the presence of one crRNA per Cascade. The inset shows the same spectrum after energetically activating the Cascade-ssDNA probe complex. The charge state distribution for the ssDNA probe is centred around 2,000 m/z.

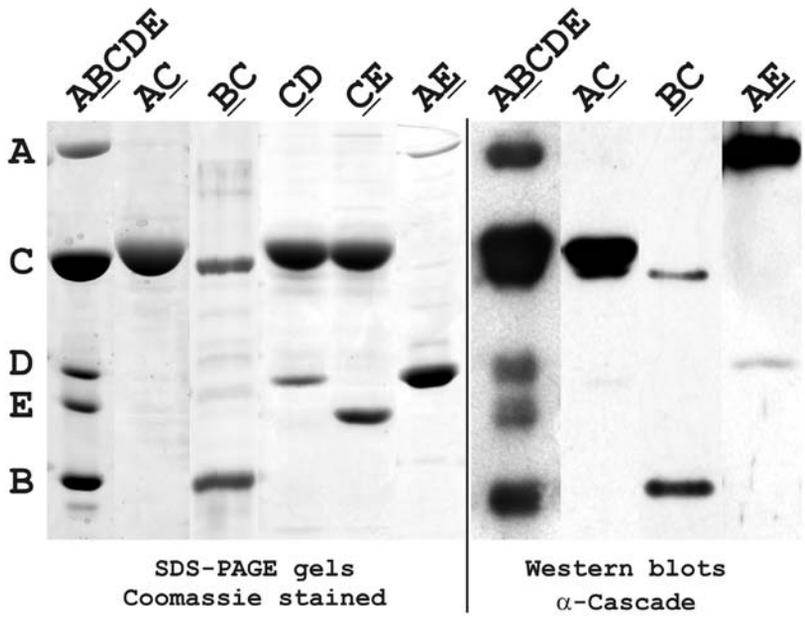


Figure S3.7. Interactions between individual Cascade subunits. Composite of Coomassie blue-stained SDS-polyacrylamide gels and Western blots of affinity purified tagged Cascade subunits (underlined) co-expressed with untagged Cascade subunits and pre-crRNA. All subunits directly interact with CasC except CasA, which interacts with CasE. Western blots were probed with anti-Cascade antibodies. Note that the N-terminal Strep-tagged Cascade subunits migrate slower through the gel than the untagged proteins.

Western blotting. The purified proteins transferred to nitrocellulose membrane (Perkin-Elmer), and incubated for 1 h in Phosphate Buffered Saline (PBS) blocking buffer pH 7.5, supplemented with 0.1% Tween 20 and 5% milk powder. The membranes were incubated with anti-Cascade serum raised in rabbits (Eurogentec, Belgium) (1:1000) in blocking buffer for 12 h, followed by three 15 min wash steps in PBS supplemented with 0.1% Tween 20. The membrane was then incubated with HRP-conjugated anti-rabbit antibodies (GE) (1:5000) in wash buffer for 1 h, followed by three 15 min wash steps. Photographic film detection (KODAK) of the signal was performed using ECL-plus substrate (GE).

Supplementary Tables

Table S3.1. Exact masses of individual Cas protein subunits of Cascade. The masses of the Cascade (sub-) complexes observed under native mass spectrometry conditions are also listed.

Cascade component	Theoretical Mass (Da)	Experimental Mass (Da)
CasA	55,972.4	55,972.2 ± 14.8
CasB with tag	21,260.4	21,261.5 ± 1.1
CasB without tag	19,204.0	19,201.9 ± 1.8
CasC	39,894.4	39,896.3 ± 1.3
CasD	25,208.9	25,210.4 ± 3.8
CasE	22,364.1	22,364.7 ± 1.1
crRNA	19,660.8	19,708[a]
Complex	**Theoretical Mass (Da)**	**Experimental Mass (Da)**
Cascade	405,095	405,365 ± 135
Cascade without CasA	349,122	349,399 ± 84
CasB dimer	42,521	42,524 ± 8

[a] The mass of crRNA was indirectly calculated via tandem mass spectrometry analysis of CasE-crRNA. Since only dissociation products for CasE were observed after collisional activation no standard deviation can be calculated for the mass of crRNA.

Table S3.2. List of masses for all Cascade (sub)complexes present in solution, and their dissociated products that are formed in the gas phase after collisional activation during tandem mass spectrometry experiments. In addition for each complex the theoretical mass (based on amino acid sequence and a crRNA mass of 19,662 Da) and stoichiometric information is given. (A=CasA, B=CasB, C=CasC, D=CasD, E=CasE, minus (-) indicates the lacking subunit, n.d.is not determined.

Mass of (sub) complexes in solution (Da)	Mass products (Da)	Theoretical mass (Da)	Annotation	Stoichiometry A B C D E crRNA
405,256		405,095	Cascade	1 2 6 1 1 1
	365,316	365,200	Cascade-CasC	1 2 5 1 1 1
	349,384	349,122	Cascade-CasA	0 2 6 1 1 1
	55,950	55,972	CasA	1 0 0 0 0 0
	39,941	39,894	CasC	0 0 1 0 0 0
349,333		349,122	Cascade-CasA	0 2 6 1 1 1
	324,389	323,914	Cascade-CasA-CasD	0 2 6 0 1 1
	309,644	309,228	Cascade-CasA-CasC	0 2 5 1 1 1
	270,031	269,334	Cascade-CasA-2•CasC	0 2 4 1 1 1
	39,946	39,894	CasC	0 0 1 0 0 0
	25,231	25,209	CasD	0 0 0 1 0 0
306,932		306,602	Cascade-CasA-2•CasB	0 0 6 1 1 1
	281,915	281,393	Cascade-CasA-2•CasB-CasD	0 0 6 0 1 1
	267,215	266,707	Cascade-CasA-2•CasB-CasC	0 0 5 1 1 1
	227,267	226,813	Cascade-CasA-2•CasB-2•CasC	0 0 4 1 1 1
	39,935	39,894	CasC	0 0 1 0 0 0
	25,250	25,209	CasD	0 0 0 1 0 0
267,076		266,707	Cascade-CasA-2•CasB-CasC	0 0 5 1 1 1
	241,503	241,498	Cascade-CasA-2•CasB-CasC-CasD	0 0 5 0 1 1
	226,673	226,813	Cascade-CasA-2•CasB-2•CasC	0 0 4 1 1 1
	39,928	39,894	CasC	0 0 1 0 0 0
	25,227	25,209	CasD	0 0 0 1 0 0
227,127		226,813	Cascade-CasA-2•CasB-2•CasC	0 0 5 1 1 1
	202,019	201,604	Cascade-CasA-2•CasB-2•CasC-CasD	0 0 4 0 1 1
	187,241	186,918	Cascade-CasA-2•CasB-3•CasC	0 0 3 1 1 1
	39,946	39,894	CasC	0 0 1 0 0 0
187,352		186,918	Cascade-CasA-2•CasB-3•CasC	0 0 3 1 1 1
	161,882	161,710	Cascade-CasA-2•CasB-3•CasC-CasD	0 0 3 0 1 1
	147,087	147,024	Cascade-CasA-2•CasB-4•CasC	0 0 2 1 1 1
	39,910	39,894	CasC	0 0 1 0 0 0
	25,218	25,209	CasD	0 0 0 1 0 0
147,210		147,024	Cascade-CasA-2•CasB-4•CasC	0 0 2 1 1 1
	121,869	121,815	Cascade-CasA-2•CasB-4•CasC-CasD	0 0 2 0 1 1
	n.d.	107,130	Cascade-CasA-2•CasB-5•CasC	0 0 1 1 1 1
	39,916	39,894	CasC	0 0 1 0 0 0
	25,219	25,209	CasD	0 0 0 1 0 0
107,431		107,130	Cascade-CasA-2•CasB-5•CasC	0 0 1 1 1 1
	84,696	84,766	Cascade-CasA-2•CasB-5•CasC-CasE	0 0 1 1 0 1
	22,375	22,364	CasE	0 0 0 0 1 0

Table S3.3. Strains, plasmids and primers used in this study

Strains	Description	Source
E.coli BL21(DE3)	F- ompT gal dcm lon hsdSB(rB -mB -) λ(DE3 [lacI lacUV5-T7 gene 1 ind1 sam7 nin5])	Novagen
E.coli BL21-Al	F- ompT gal dcm lon hsdSB(rB -mB -) araB::T7RNAP-tetA	Invitrogen
E.coli NEB5α	fhuA2 _(argF-lacZ)U169 phoA glnV44 Φ80_ (lacZ)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17	New England Biolabs
E.coli DH5α	F- endA1 glnV44 thi-1 recA1 relA1 gyrA96 deoR nupG Φ80dlacZΔM15 Δ(l*acZYA-argF*)U169, hsdR17(rK -mK +), λ–	

Plasmids	Description and order of genes (5'-3')	Restriction sites	Primers	Source
pET-52b(+)	T7 RNA polymerase based expression vector, Amp^R			Novagen
pRSF-1b	T7 RNA polymerase based expression vector, Kan^R			Novagen
pCDF-1b	T7 RNA polymerase based expression vector, Str^R			Novagen
pACYC duet-1	T7 RNA polymerase based expression vector, Cam^R			Novagen
pWUR381	cas3 in pET-52b with both Strep-tag II (N-term) and His_{10}-tag (C-term)	BamHI/NotI	BG2243 + BG2244	This study
pWUR388	*casA* in pET-52b with both Strep-tag II (N-term) and His10-tag (C-term)			(Brouns et al. 2008)
pWUR396	*E. coli* K12 CRISPR in pACYCDuet-1, see Fig S3.2			(Brouns et al. 2008)
pWUR397	*cas3* in pRSF-1b, no tags			(Brouns et al. 2008)
pWUR400	*casA-casB-casC-casD-casE* in pCDF-1b, no tags			(Brouns et al. 2008)
pWUR401	*casB-casC-casD-casE* in pCDF-1b, no tags			(Brouns et al. 2008)
pWUR402	*casC-casD-casE* in pCDF-1b, no tags			(Brouns et al. 2008)
pWUR403	*casD-casE* in pCDF-1b, no tags			(Brouns et al. 2008)
pWUR404	*casE* in pCDF-1b, no tags			(Brouns et al. 2008)
pWUR405	*casA-casB-casC-casD* in pRSF-1b, no tags			(Brouns et al. 2008)
pWUR406	*casA-casB-casC* in pRSF-1b, no tags			(Brouns et al. 2008)
pWUR407	*casA-casB* in pRSF-1b, no tags			(Brouns et al. 2008)
pWUR408	*casA* in pRSF-1b, no tags			(Brouns et al. 2008)
pWUR413	*casD* in pRSF-1b, no tags	NcoI/NotI	BG2466 + BG2482	This study
pWUR477	non targeting CRISPR in pACYCDuet-1 (N), see Fig S3.2			(Brouns et al. 2008)
pWUR478	template CRISPR in pACYCDuet-1 (T_{1-4}), see Fig S3.2			(Brouns et al. 2008)
pWUR479	coding CRISPR in pACYCDuet-1 (C_{1-4})			(Brouns et al. 2008)
pWUR480	*casB* with Strep-tag II (N-term)-*casC-casD* in pET52b			(Brouns et al. 2008)
pWUR482	*casE* in pET52-1b, with Strep-tag II (N-term)	BamHI/NotI	BG2586 + BG2253	This study
pWUR514	*casB* with Strep-tag II (N-term)-*casC-casD-CasE* in pET52b	Acc65I/NotI	BG2573 + BG2586	This study
pWUR547	*E. coli* R44 CRISPR, 7x spacer nr. 2, in pACYCDuet-1, see Fig S3.2	EcoNI/Acc65I		(Mojica et al. 2005), Geneart, Germany
pWUR553	*casB* in pET52-1b, with Strep-tag II (N-term)	Acc65I/NotI	BG2573 + BG2484	This study
pWUR554	*casC* in pCDF-1b, no tags	NcoI/NotI	BG2465 + BG2483	This study
pWUR555	*casC* in pET52-1b, with Strep-tag II (N-term)	BamHI/NotI	BG2249 + BG2483	This study

Table S3.3. *continued*

Experiment	Primer	Sequence (5'-3')	Description
Cloning	BG2243	GCGCGGGATCCTATGGAACCTTTTAAATATATATGCC	*cas3* + BamHI (fw)
	BG2244	GGCCCGCGGCCGCTTTGGGATTTGCAGGGATGACT	*cas3* + NotI (rv)
	BG2249	GCGCGGGATCCTATGTCTAACTTTATCAATATTCATGT	*casC* + BamHI (fw)
	BG2253	GCGCGGGATCCTATGTATCTCAGTAAAGTCATCATTG	*casE* + BamHI (fw)
	BG2316	GCGCGGGTACCAGATGAGATCTTATTTGATCTTGCGG	*casD* + Acc65I (fw)
	BG2465	GCGCGCCATGGCTATGTCTAACTTTATCAATATTCATGT	*casC* + NcoI (fw)
	BG2466	GCGCGCCATGGCTATGAGATCTTATTTGATCTTGCGG	*casD* + NcoI (fw)
	BG2482	GGCCCGCGGCCGCTTACTGAGATACATCCATACCTCC	*casD* + NotI (rv) + stopcodon
	BG2483	GGCCCGCGGCCGCTCACGCCTCGCCATTATTACGA	*casC* +NotI (rv) + stopcodon
	BG2484	GGCCCGCGGCCGCTTACGCATTTTTGTTTGTGGTCAAT	*casB* + NotI (rv) + stopcodon
	BG2573	GCGCGGGTACCAGATGGCTGATGAAATTGATGCAATG	*casB* + Acc65I (fw)
	BG2586	GGCCCGCGGCCGCCTGCAGTCACAGTGGAGCCAAAGA-TAGCAAG	*casE* + NotI (rv) + stopcodon
	BG3048	GCGCGGAATTCATGAGATCTTATTTGATCTTGCGG	*casD* + EcoRI (fw)
	BG3049	GGCCCCTGCAGTTACTGAGATACATCCATACCTCC	*casD* + PstI + stopcodon (rv)
EMSA	BG3009	AAGCTGATCGGCAAGCTCGAAAGC*ACG***ACGGTATTGT-TCAGATCCTGGCTTGCCAACAG**TGATTTGCTCAATTTTG-TAGATTGAAG	Target with R44 protospacer, and flanking sequences[1]
	BG3010	CTTCAATCTACAAAATTGAGCAAATCA**CTGTTGGCAAGC-CAGGATCTGAACAATACCGT***CGT*GCTTTCGAGCTTGC-CGATCAGCTT	Target with R44 protospacer, and flanking sequences[2]
	BG3011	AAGCTGATCGGCAAGCTCGAAAGC*ACG***TTGACTTGGAT-CACCGACCTTGGCGATTACGA**TGATTTGCTCAATTTTGTA-GATTGAAG	Target with scrambled R44 protospacer and flanking sequences[1]
	BG3012	CTTCAATCTACAAAATTGAGCAAATCA**TCGTAATCGCCA-AGGTCGGTGATCCAAGTCAA***CGT*GCTTTCGAGCTTGC-CGATCAGCTT	Target with scrambled R44 protospacer and flanking sequences[2]
	BG3032	**ACGGTATTGTTCAGATCCTGGCTTGCCAACAG**	Target with R44 protospacer, without flanking sequences[1]
	BG3033	**CTGTTGGCAAGCCAGGATCTGAACAATACCGT**	Target with R44 protospacer, without flanking sequences[2]
In vitro transcription (to generate RNA products for EMSA)	BG3030	CCATGGTAATACGACTCACTATAGGGCTTCAATCTA-CAAAATTGAGCAA	T7 promoter primer to generate a template for *in vitro* transcription[3]
	BG3031	AAGCTGATCGGCAAGCTCGAAAGC	Primer to generate a template for *in vitro* transcription[3]
	BG3079	CCATGGTAATACGACTCACTATAGGGAAGCTGATCG-GCAAGCTCGAAAG	T7 promoter primer to generate a template for *in vitro* transcription[3]
	BG3080	CTTCAATCTACAAAATTGAG	Primer to generate a template for *in vitro* transcription[3]
ESI-MS Native PAGE EM	BG3028	biotin-TEG-CTGTTGGCAAGCCAGGATCTGAACAATACCGT	R44 protospacer (A)

[1] DNA sequence corresponds to crRNA sequence, [2] DNA sequence complementary to crRNA sequence, [3] These primers were used to prepare material for *in vitro* transcription by PCR by adding on a T7 promoter, using either BG3009-BG3010 or BG3011-BG3012 as template. Restriction sites are underlined, Proto-spacer sequences are in bold, Location of PAM in italics.

Chapter 4

CRISPR interference requires a protospacer adjacent motif (PAM) and perfect basepairing at the PAM-side of the protospacer

Matthijs M. Jore, Edze R. Westra, John van der Oost, Stan J. J. Brouns

Manuscript In preparation

Abstract

CRISPR/Cas systems protect prokaryotes from invading nucleic acids. The CRISPRs contain spacer sequences that are derived from previously encountered invading nucleic acids, providing the basis for a sequence-specific defense system. In *E. coli* K12 long precursor CRISPR RNA is cleaved into mature crRNAs that remain bound to the Cas protein complex Cascade. Cascade loaded with crRNA and assisted by Cas3 can successfully inhibit virus proliferation. It has been demonstrated that Cascade-bound crRNA can recognize dsDNA and the crRNA can base pair to the complementary strand by strand displacement. It is currently unknown which nucleotide positions are important for base pairing and immunity. In addition to previously established anti-virus activity, we here demonstrate that the CRISPR/Cas system of *E. coli* can inhibit transformation of a plasmid containing a protospacer. We examined which nucleotide positions in the protospacer and flanking the protospacer are essential for inhibition of plasmid transformation. A mutant library was generated, transformed to a resistant host and escape mutants were selected. Sequence analysis showed that most escape mutants contain nucleotide substitutions in the protospacer adjacent motif (PAM) and in the PAM-side of the protospacer; in addition, deletion of the protospacer region has been observed. The PAM might be a criterion to verify that only invading DNA is being targeted. The PAM-side of the protospacer might play a crucial role during recognition, base pairing and putative cleavage of the target.

Introduction

Prokaryotes have developed several sophisticated mechanisms to defend themselves against invading nucleic acids that are potentially harmful. One of the recently discovered systems is based on clusters of regularly interspaced short palindromic repeats (CRISPR) (reviewed in (van der Oost et al. 2009; Horvath and Barrangou 2010; Karginov and Hannon 2010; Marraffini and Sontheimer 2010a). CRISPR loci contain short direct repeats that are separated by unique spacers which are identical to extrachromosomal elements such as phages and conjugative plasmids (the identical nucleic acid sequence is named the protospacer) (Bolotin et al. 2005; Mojica et al. 2005; Pourcel et al. 2005; Lillestøl et al. 2006; Semenova et al. 2009). In close proximity of CRISPR a set of conserved *cas* (CRISPR associated) genes can be found. Based on the composition of the *cas* gene cluster, eight different subtypes can be recognized (Haft et al. 2005) that are each associated with one type (or few types) of CRISPR repeat sequence (Kunin et al. 2007). The CRISPR/Cas system is an adaptive and heritable immune system (van der Oost et al. 2009), which appears to be frequently transferred by horizontal gene transfer (Haft et al. 2005; Godde and Bickerton 2006; Makarova et al. 2006; Horvath et al. 2009).

CRISPR-mediated defense comprises three stages. During the first stage fragments of invading nucleic acids are integrated as spacers in the CRISPR locus (Barrangou et al. 2007; Deveau et al. 2008; Horvath et al. 2008; van der Ploeg 2009). These studies revealed that a conserved protospacer adjacent motif (PAM) can be found in proximity of the protospacer sequences in the phage genome (Deveau et al. 2008; Horvath et al. 2008; van der Ploeg 2009). This had been independently observed before (Bolotin et al. 2005). The PAM seems to be a common theme among the different CRISPR types, although each CRISPR type uses a specific motif (Mojica et al. 2009). During the second stage, a CRISPR is unidirectionally transcribed (Brouns et al. 2008; Hale et al. 2008; Marraffini and Sontheimer 2008; Semenova et al. 2009); the reported bidirectional transcription in *Sulfolobus* spp. seems to be an exception to this rule (Lillestøl et al. 2006; Lillestol et al. 2009). The precursor CRISPR RNA (pre-crRNA) is subsequently cleaved into small CRISPR RNA (crRNA) by CasE in *E. coli* (Brouns et al. 2008) and Cas6 in *Pyrococcus furiosus* (Carte et al. 2008). In *E. coli* crRNAs remain bound by the Cas protein complex Cascade. In *P. furiosus* the crRNAs are further trimmed at their 3' end to generate mature crRNAs (also called psiRNAs) that form a stable interaction with the Cmr-complex (Hale et al. 2008; Hale et al. 2009). During the third stage, the CRISPR interference, crRNAs guide the Cas protein machinery to specifically target previously encountered genetic elements, nucleic acid fragments of which have previously been integrated in a CRISPR locus. The Cmr-complex from

P. furiosus has been shown to cleave complementary RNA *in vitro* (Hale et al. 2009). However, in *Staphylococcus epidermidis* it has been shown that DNA must be the target *in vivo* (Marraffini and Sontheimer 2008). DNA targeting has also been demonstrated in the case of *E. coli*; Cascade-bound crRNAs complementary to either the coding or the non-coding strand can prevent phage lambda infection, therein assisted by Cas3 (Brouns et al. 2008). This finding is supported by the observation that Cascade loaded with crRNA can sequence specifically bind to dsDNA by strand displacement (Chapter3).

Here we describe the analysis of nucleotide positions in the protospacer and PAM that are important for interference in *E. coli*. To address this question we constructed a plasmid containing a protospacer flanked by a PAM. Compared to an empty plasmid, the transformation efficiency dramatically decreased when we transformed this plasmid to cells overproducing Cas3, Cascade and complementary crRNA. The non-coding nature of the protospacer and PAM enabled us to create a library by error-prone PCR. Plasmids that are entering the cell and can escape the CRISPR-based immunity result in colony growth after plating. Sequence analysis of these escape mutants revealed several positions in the protospacer and PAM that are crucial for interference.

Results

To investigate whether the CRISPR/Cas system from *E. coli* K12 could prevent plasmid transformation we constructed a plasmid containing a KpnI-KpnI 1.5 kb fragment from phage λ (pUC-λ_{1500}) (Fig. 4.1A). We designed a CRISPR containing 5 identical spacers that have perfect complementarity to a fragment in the λ DNA compatible with the PAM sequence 5'-AWG-protospacer-3' (Mojica et al. 2009). Cascade, Cas3 and the λ targeting CRISPR (CRISPR λ) RNA were overproduced in *E. coli* BL21-AI, which lacks endogenous *cas* genes (Studier et al. 2009). These cells were made electrocompetent for plasmid transformation. The transformation efficiency of an empty pUC19 vector to the CRISPR λ carrying strain is 0.7 x 10^9 cfu/μg of plasmid DNA. This value is similar to the transformation efficiency of a control strain carrying a non-targeting CRISPR (CRISPR R44), which is 2.0 x 10^9 cfu/μg (Fig. 4.1B). The pUC-λ_{1500} plasmid transformed equally well to the non-targeting strain (0.7 x 10^9 cfu/μg), but showed a 3,000-fold reduction in transformation efficiency to the CRISPR λ-carrying strain (2.0 x 10^5 cfu/μg) (Fig. 4.1B). This clearly shows that the pUC-λ_{1500} plasmid is specifically being targeted and that the CRISPR/Cas system from *E. coli* K12 can prevent plasmid transformation.

To determine whether the pUC-λ_{1500} was affected by CRISPR/Cas, we selected four

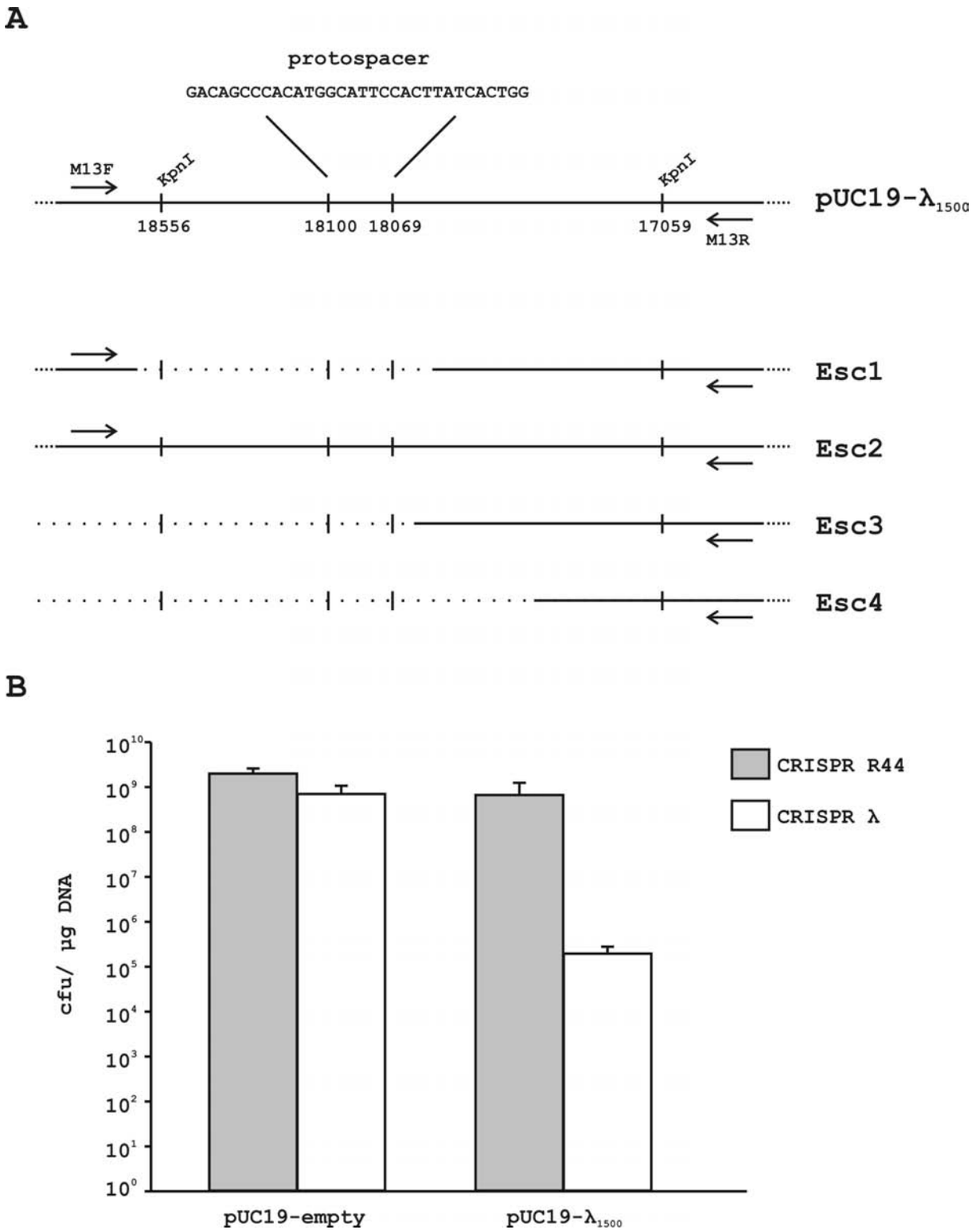


Figure 4.1. CRISR/Cas of the Cse-subtype inhibits plasmid transformation. (A) Schematic representation of the pUC-λ_{1500} plasmid. The genome coordinates of the lambda fragment and the protospacer sequence are given. The restriction enzyme KpnI was used for cloning the lambda fragment into pUC19. Primers M13F and M13R were used for sequencing of the insert. A schematic representation of four sequenced escape mutants (Esc1-4) is shown below. Observed deletions in the plasmids are represented by dashed lines. (B) Transformation efficiencies of pUC19 and pUC-λ_{1500} to *E. coli* BL21-AI overproducing Cascade, Cas3 and either CRISPR R44 (grey) or CRISPR λ (white). Error bars indicate the standard deviation.

clones from the transformation of the pUC-λ_{1500} to the CRISPR λ carrying strain. The plasmids were isolated and the pUC-λ_{1500} insert was sequenced using flanking primers. Three out of four clones (Esc1, 3 and 4) showed a major deletion that included the protospacer (Fig. 4.1A). The transformation efficiencies of these isolated plasmids were equal to that of empty pUC19 plasmids (data not shown). One clone sequence was unaffected (Esc2). When this isolated plasmid was transformed, again a low transformation efficiency was obtained, showing that the escape colony in the initial experiment is not caused by the pUC-λ_{1500} plasmid. An explanation for the successful escape might be an unidentified recombination event in any of the other plasmids carrying either *cas* genes or the CRISPR sequence, thus neutralizing CRISPR-mediated defense.

It has previously been shown that Cascade can bind dsDNA and that the crRNA can base pair with complementary DNA by strand displacement (Chapter 3). In order to reveal which positions in the protospacer and flanking sequence are essential for interference, a 350 basepair fragment containing the protospacer was cloned into a pUC19 vector (pUC-λ_{350}) (Fig. 4.2A). A mutant library was created by error-prone PCR and the amplicons cloned into pUC19. This library contains approximately 5,000 clones, of which 10 were sequenced to check the mutation content. Three out of ten clones contained one or two mutations or deletions, the other clones remained unchanged. Total plasmid DNA was isolated from this library and transformed to *E. coli* KRX overproducing Cascade, Cas3 and CRISPR λ. This strain is deficient in *recA and endA1* to enhance the stability of the transformed plasmid (Hartnett et al. 2006). The transformation efficiency of an empty pUC19 vector is 1.5 x 10^6 cfu/µg DNA. pUC-λ_{350} had a 10,000 fold lower transformation efficiency of 2.0 x 10^2 cfu/µg DNA (Fig. 4.2B). Next, plasmid DNA from the mutant library was transformed, showing an efficiency of ~2 x 10^3 (data not shown). Unfortunately, the library contained empty plasmids which resulted in false positives. Therefore, the plasmids were first checked for inserts by colony PCR. Approximately fifty escape mutants that contain an insert of expected size were selected and sequenced. Eight different mutations were observed (Fig. 4.2C). Two mutations occur in the PAM at two different positions (Esc5 and 6). Five mutations occur at the PAM-side in the protospacer (Esc7, 8, 9, 10 and 11). Finally, one deletion at position 20 was observed (counted from the PAM-side of the protospacer). For each mutation one clone was selected and transformed back to *E. coli* KRX overproducing Cascade, Cas3 and CRISPR λ to determine its transformation efficiency. The values were similar to the transformation efficiency of an empty pUC19 vector, indicating that they can successfully bypass CRISPR defense and that each mutated position is indeed essential for resistance.

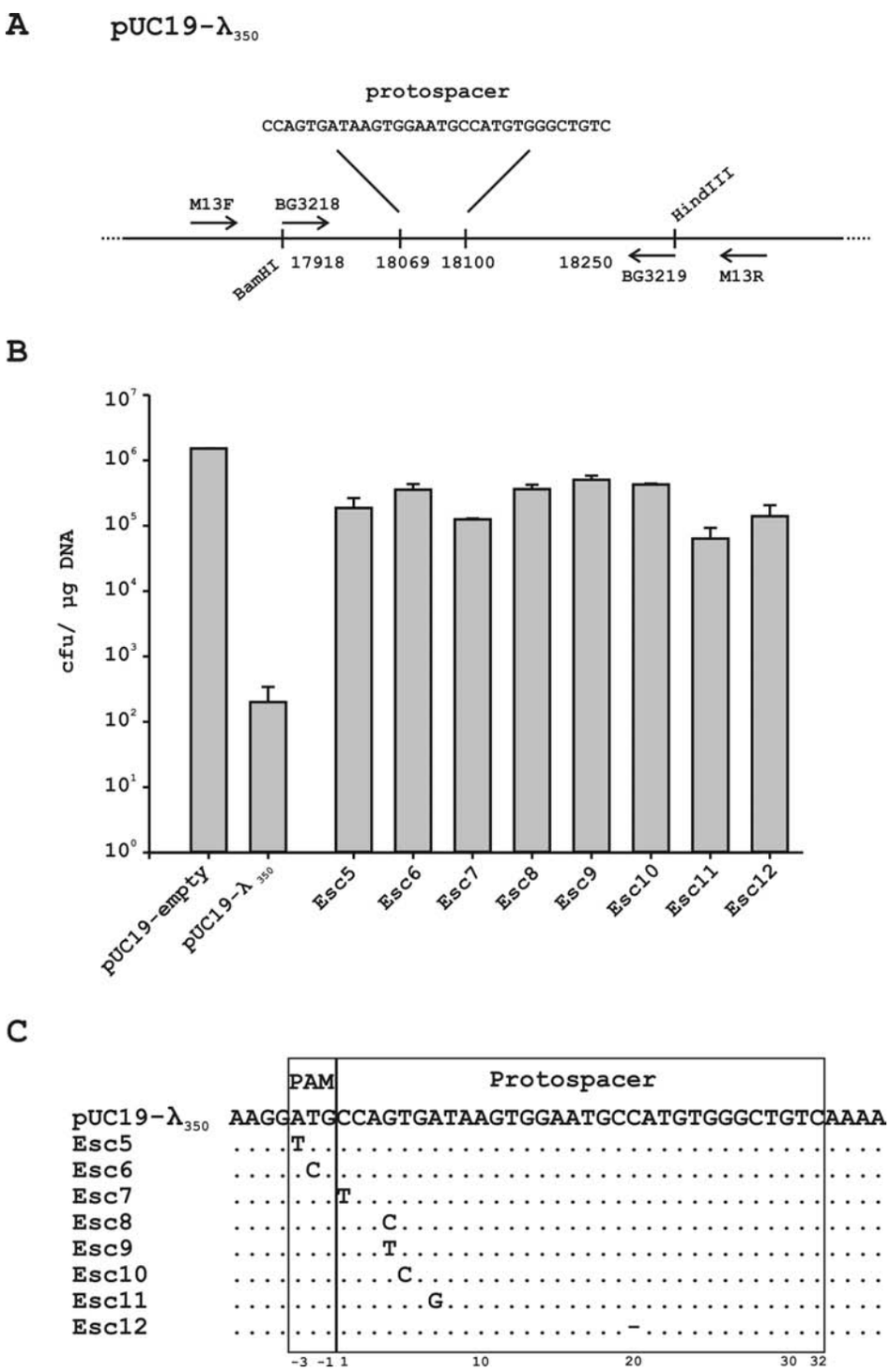


Figure 4.2. Escape mutants contain mutations in the PAM or protospacer. (A) Schematic representation of the pUC-λ350 plasmid. A PCR product of the lambda fragment was cloned into pUC19 using the BamHI and HindIII sites that were introduced along with primers BG3218 and BG3219. The genome coordinates of the lambda fragment and the protospacer sequence are given; please note that this lambda fragment is cloned in the opposite direction compared to pUC-λ350. The protospacer sequence is therefore reverse complement to the one given in Fig.4.1. In this orientation the PAM is located directly upstream the protospacer. For creation of the library the same cloning strategy was applied. Transformation efficiencies of pUC19, pUC-λ350 and the escape mutants to BL21-AI overproducing Cascade, Cas3 and CRISPR λ RNA are given in (B). Error bars indicate the standard deviation. (C) The sequence of the protospacer and flanking sequence of pUC-λ350 is given as a reference. Point mutations in the escape mutants are represented by the new base while deletions are shown as bars. Non-mutated positions are shown as dots.

Discussion

CRISPR/Cas inhibits plasmid transformation

CRISPR/Cas-mediated inhibition of transformation (and conjugation) has been previously shown in *Staphylococcus epidermidis* (Marraffini and Sontheimer 2008). The CRISPR/Cas system of the Cse-subtype has been shown to inhibit transduction of phage Lambda in *E. coli* (Brouns et al. 2008). Here we show that cells overproducing Cascade, Cas3 and crRNA can also inhibit transformation of a plasmid containing a protospacer and PAM. When this plasmid was transformed to *E. coli* BL21-AI overproducing Cas proteins and crRNA, we found that three out of four escape mutants contain major deletions in the transformed plasmid, including the protospacer and PAM. The fourth escape mutant most likely has encountered a recombination event in one of the plasmids carrying CRISPR or *cas* genes, since the escape is not caused by the transformed plasmid. Moreover, it is tempting to speculate that many, if not all, of the observed escape colonies carry deletions in either the plasmids that facilitate CRISPR interference or the transformed plasmid that is being targeted. However, more colonies should be screened to statistically found this hypothesis. If this were the case, the actual immunity against the protospacer containing plasmid is close to 100%. The three deletions that were identified occurred at different locations in the plasmid, indicating a non-specific deletion or recombination process. To reduce the number of deletions, further experiments were performed in *E. coli* KRX which lacks *recA* and *endA* and has partially defective restriction systems (*hsd* and e14). Interestingly, CRISPR interference in this strain was not affected and indirectly showed that RecA and EndA1 are not required for this defense mechanism.

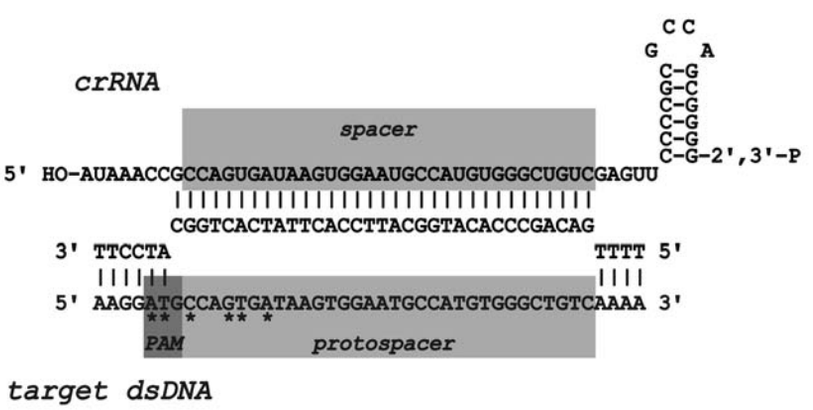


Figure 4.3. Base pairing of crRNA with target dsDNA. Schematic overview of a crRNA molecule (top) binding to one strand of the dsDNA molecule by strand displacement as described in Chapter 3. One base of the PAM is complementary to the 5' repeat sequence (5' handle) and can base pair, the other two cannot. Positions that are mutated in the escape mutants are marked with an asterisk. Please note that Cascade is not shown in this figure for simplicity.

The PAM is important for interference

Two out of eight escape mutants contain a mutation in the PAM. Based on sequence analysis of the protospacer regions, PAMs have been examined for several CRISPR-types (Bolotin et al. 2005; Deveau et al. 2008; Horvath et al. 2008; Mojica et al. 2009; Semenova et al. 2009; van der Ploeg 2009). The Cse-subtype, which is associated with the CRISPR-2 type (Kunin et al. 2007), contains an AWG motif (W = A or T) immediately upstream the protospacer (previously reported as CWT in the opposite strand) (Mojica et al. 2009) (Fig. 4.3). In our library we identified two escape mutations that result in an imperfect PAM; from ATG to TTG (Esc5) and from ATG to ACG (Esc6) (Fig. 4.2C). The PAM has previously been demonstrated to be important for interference, as phages were found to evade CRISPR-based immunity in *S. thermophilus* by mutations in the PAM (and protospacer) (Deveau et al. 2008). The role of the PAM is not fully understood but it may actually help the Cas protein machinery to distinguish between the CRISPR DNA (self) and the invading DNA element (non-self). Intact PAMs are present in invading DNA but absent from chromosomal CRISPR DNA thus preventing auto-immunity. An alternative model to prevent auto-immunity has been proposed for *S. epidermidis* (Marraffini and Sontheimer 2010b). The CRISPR/Cas system inhibits transformation of a plasmid containing a protospacer, unless it is flanked by a sequence that is substantially similar to the CRISPR sequence (Marraffini and Sontheimer 2010b). The potential of a flanking sequence to base pair with the CRISPR repeat thus determines whether a DNA sequence is subject to interference or not. In the case of *E. coli* one base of the PAM can base pair with the crRNA whereas the other two cannot (Fig. 4.3). A mutation from ATG to TTG (Esc5) does not influence the base pairing pattern, but does make a crucial difference for interference. this indicates that interference in *E. coli* is not only determined by the potential to base pair with the crRNA, but by the nucleotide sequence. However, to better understand the role of base pairing of the PAM with the crRNA and of the sequence itself during interference, the effect of all possible mutations in the PAM should be tested, ideally in combination with CRISPR mutagenesis, analogous to previous experiments (Marraffini and Sontheimer 2010b).

Perfect base pairing requirements for interference

Five out of eight escape mutants (Esc7-11) contain a point mutation at the PAM-side of the protospacer (Fig. 4.2C). It seems that base pairing at the PAM-side of the protospacer, and thus 5' end of the spacer is essential for interference. A similar phenomenon has been observed in the case of *S. thermophilus* (Deveau et al. 2008). Nonetheless, it cannot be ruled out that plasmids with mutations at other positions in the protospacer, which would lead to escape, are absent from the library and thus not

identified. To determine the importance of each position, a site directed mutagenesis approach should be taken. The role of the 5' end of the crRNA spacer could be crucial for base pairing with the target DNA. It might encompass a sequence where base pairing is initiated, similar to the seed sequence in miRNAs and siRNAs in eukaryotes (Lewis et al. 2005; Wang et al. 2009). Argonaute-bound miRNAs or siRNAs initiate base pairing at positions 2 - 8 from the 5' end (the seed sequence) and then 'zipper up' forming a duplex of nucleic acids (Wang et al. 2009). What happens after binding of the crRNA to complementary dsDNA remains obscure. It is known that Cas3 is not needed for the binding *in vitro* (Chapter 3), but is essential for immunity *in vivo* (Brouns et al. 2008). Cas3 comprises a HD-type nuclease domain fused to a DEAD/H box helicase domain (Makarova et al. 2006). The HD domain protein SSO2001 from *Sulfolobus solfataricus* has metal-dependent endonuclease activity on double-stranded oligonucleotides (Han and Krauss 2009). Possibly the Cas3 protein is recruited or activated upon duplex formation of crRNA and target DNA, and subsequently catalyze cleavage of the target DNA. Perfect base pairing at the 5' end of the spacer, as well as the presence of a perfect PAM, might be crucial to enable cleavage. Extensive stretches of mismatches at the 3' end of the spacer, which is the case in Esc12 (Fig. 4.2C), might destabilize the crRNA interaction with its target and might also inhibit the catalytic activity, leading to escape of the phage or plasmid. Whether the observed mutations in the protospacer inhibit base pairing between the spacer and the target DNA should be tested by means of Electrophoretic Mobility Shift Assays (EMSAs), as described previously (Chapter 3). This might provide more insight of the importance of the 5' end of the spacer for either base pairing and/ or target cleavage.

Materials and methods

Strains and plasmids used

E. coli NEB5α (New England Biolabs) was used for plasmid construction. *E. coli* BL21-AI (Invitrogen) and *E. coli* KRX (Promega) were used for transformation experiments. Construction of plasmids for overproduction of Cascade (pWUR400), Cas3 (pWUR397) and CRISPR R44 RNA (pWUR547) is described elsewhere (Chapter 2 and 3). The plasmid for overproduction of CRISPR λ RNA (pWUR564) was subcloned from a synthetic construct (GeneArt AG, Germany) into pACYCduet-1 (Novagen) with NcoI and Acc65I. The sequence of the insert is provided below. pUC-λ_{1500} (pWUR609) was constructed by digestion of λ DNA with KpnI, selection of the 1.5 kb fragment (17059-18556), and ligation of the fragment into a pUC19 vector (New England Biolabs) cleaved with kpnI. To generate pUC-λ_{350} (pWUR610) a fragment of ~350 bp (17918-18250) was amplified by PCR using primers BG3218 (5'-GGCCCGGATCCGTCGGGCGAGCGATGATGCG-3')

and BG3219 (5'-CGCGCAAGCTTCATCGGCGTTTCATTCCCGTTT-3') and cloned into pUC19 by digestion with BamHI and HindIII and subsequent ligation.

Library construction

The mutant library was generated using a GeneMorph II Random Mutagenesis Kit (Stratagene). The PCR reaction contained 34.5 µl MQ, 5 µl Mutazyme II Reaction Buffer (10x), 1 µl dNTP mix (40 mM), 1.25 µl BG3218 (100 ng/ µl), 1.25 µl BG3219 (100 ng/ µl), 6 µl pUC-λ_{1500} (220 ng/ µl) and 1 µl Mutazyme II (2.5 U/ µl). The program used is 2 min 95 °C, twenty cycles of the next three steps of 30 sec 95 °C, 30 sec 52 °C, 1 min 72 °C, and finally one step of 10 min 72 °C. The PCR product and pUC19 vector were cleaved with BamHI and HindIII, ligated and transformed to *E. coli* NEB5α (New England Biolabs).

Transformation and mutant selection

E. coli BL21-AI cells carrying pWUR397, pWUR400 and either pWUR564 or pWUR547 were grown in LB supplemented with the appropriate antibiotics at 37 °C till an OD_{600} of ~ 0.3. The cells were induced with 1 µM Isopropyl β-D-1-thiogalactopyranoside and 0.2 % arabinose, and grown for 45 min at 37 °C prior to harvesting. *E. coli* KRX containing pWUR397, pWUR400 and pWUR564 was grown and harvested according to the same protocol. Cells were made electrocompetent by washing twice with ice cold MQ and subsequently twice with ice cold 10 % glycerol. Typically 10 ng of plasmid DNA is transformed. Transformation efficiencies were calculated based on duplicate or triplicate experiments. Isolated plasmids from escape mutants were sequenced at Baseclear (The Netherlands) using M13F (5'-TTTCCCAGTCACGACGTTG-3') and M13R (5'-GGATAACAATTTCACACAGG-3') primers.

pWUR564 insert:

CCATGGAAACAAAGAATTAGCTGATCTTTAATAATAAGGAAATGTTACATTAAGGTTGGTGGGTTGTTTTTATGGGAAAAAATGCTTTAAGAAC
AAATGTATACTTCTAGAGAGTTCCCCGCGCCAGCGGGGATAAACCGCCAGTGATAAGTGGAATGCCATGTGGGCTGTCGAGTTCCCCGCGCCAG
CGGGGATAAACCGCCAGTGATAAGTGGAATGCCATGTGGGCTGTCGAGTTCCCCGCGCCAGCGGGGATAAACCGCCAGTGATAAGTGGAATGCC
ATGTGGGCTGTCGAGTTCCCCGCGCCAGCGGGGATAAACCGCCAGTGATAAGTGGAATGCCATGTGGGCTGTCGAGTTCCCCGCGCCAGCGGGG
ATAAACCGCAGCTCCCATTTTCAAACCCAGGTACC

Acknowledgements

We would like to thank our collaborators Ekaterina Semenova and Konstantin Severinov for helpful discussions. MMJ, JvdO and SJJB were financially supported by an NWO Vici grant to JvdO and a Veni grant to SJJB. ERW was financially supported by NWO Spinoza resources awarded to Willem M. de Vos.

Chapter 5

The antiviral Cas protein machinery is located at cellular poles

Matthijs M. Jore, Jasper Akerboom, Jianyong Tang, Sean A. McKinney, Loren L. Looger, Stan J.J. Brouns, John van der Oost

Manuscript In preparation

Abstract

The recently discovered Clusters of Regularly Interspaced Palindromic Repeats (CRISPRs) provide prokaryotes with an adaptive and inheritable immune system, which encompasses three distinct steps. Firstly, upon infection by previously non-encountered extrachromosomal DNA such as a virus, fragments are integrated into the genomic CRISPR locus. Secondly, the CRISPR is transcribed and cleaved into small CRISPR RNAs (crRNAs). Thirdly, crRNAs guide CRISPR associated (Cas) proteins to counteract invasion by previously encountered DNA elements that are stored in the CRISPR archive. Little is known about the subcellular location of the molecular actors participating in the CRISPR-mediated battle between the host and parasite. Here we used nanoscale resolution Photo-Activated Localization Microscopy (PALM) to determine the spatial distribution of the Cas proteins in *Escherichia coli*. We found that Cas proteins are preferentially located at the poles of the bacterial cell. This subcellular location of Cas proteins might be advantageous for the host, since phages frequently infect and replicate in the same area of the cell.

Introduction

Prokaryotes are constantly being attacked by extrachromosomal elements such as viruses and plasmids. To prevent infection of these foreign DNA entities, prokaryotes have developed several defense mechanisms, amongst which the recently discovered small RNA mediated CRISPR/Cas mechanism (for an overview, see Chapter 1). CRISPR arrays consist of short DNA repeats that are separated by unique spacers of similar length. The spacer sequences are derived from viral and plasmid DNA (Bolotin et al. 2005; Mojica et al. 2005; Pourcel et al. 2005; Lillestøl et al. 2006; Semenova et al. 2009). CRISPR-associated (*cas*) genes are generally located in close proximity to the CRISPR array. The encoded Cas protein machinery together with small CRISPR RNA that is complementary to the invading nucleic acid, confer specific resistance (see below). Several subtypes can be distinguished based on the composition of the *cas* gene cluster (Haft et al. 2005; Makarova et al. 2006; van der Oost et al. 2009), which are most likely acquired by horizontal gene transfer (Haft et al. 2005; Godde and Bickerton 2006; Makarova et al. 2006; Horvath et al. 2009).

CRISPR-mediated defense comprises three distinct stages. During the first stage, new spacers from invading nucleic acids are integrated in the CRISPR locus (Barrangou et al. 2007; Deveau et al. 2008; Horvath et al. 2009; van der Ploeg 2009). Not much is known about the spacer integration but it is hypothesized that the nucleases Cas1 and Cas2 are involved (Marraffini and Sontheimer 2009; van der Oost et al. 2009; Wiedenheft et al. 2009). During the second stage, the CRISPR locus is transcribed and subsequently cleaved (Brouns et al. 2008; Carte et al. 2008; Hale et al. 2008; Marraffini and Sontheimer 2008). The generated mature crRNAs are retained by a complex of Cas proteins. In *E. coli* K12, this complex consists of CasABCDE and is termed Cascade (Brouns et al. 2008). During the third and final stage, the crRNAs guide the Cas protein machinery to their specific targets: previously encountered invading DNA. Successful interference generates resistance of the host against the predator (Brouns et al. 2008; Marraffini and Sontheimer 2008).

In *E. coli* K12, Cascade loaded with crRNA binds the complementary strand from dsDNA by strand displacement (Chapter 3). However, *in vivo* studies revealed that, besides Cascade, the putative nuclease/helicase Cas3 is needed for resistance (Brouns et al. 2008). In *Saphylococcus epidermidis* the distinction between host DNA (the CRISPR itself) and invading DNA is determined by the flanking regions of the protospacer (this is the sequence in the extrachromosomal element that is identical to the spacer). If the flanking sequence is complementary to the repeat, no interference occurs, thus preventing the host from degrading its own DNA (Marraffini and Sontheimer 2010b). In

other subtypes, a conserved Protospacer Adjacent Motif (PAM), absent in the CRISPR, may determine whether the target is subject to interference (Deveau et al. 2008; Mojica et al. 2009).

Although knowledge of the CRISPR expression and target recognition is increasing, little is known about the sub-cellular localization of the Cas protein machinery. A genome-wide localization study in *E. coli* has been established, in which every single gene was cloned in a plasmid fused to the gene encoding a GFP marker (Kitagawa et al. 2005) (http://ecoli.naist.jp/GB6/search.jsp). This study revealed that CasC forms foci at the cellular pole, whereas Cas3, CasA, CasE and Cas1 show a more uniform distribution over the cell. The different distribution of CasC on the one hand, and CasA and CasE on the other, is in contrast with the previous finding that they are all part of the Cascade complex. No localization data are available for the remaining Cas proteins. Therefore we decided to study the spatial organization of the Cas proteins in *E. coli* more thoroughly. With the use of Photo-Activated Localization Microscopy (PALM) imaging (Betzig et al. 2006), we found that the single Cas1, Cas2 and Cas3 proteins as well as Cascade localize at the cellular poles. It has been previously reported that bacteriophages, such as phage lambda, infect and replicate *E. coli* at this same location (Edgar et al. 2008). Phage lambda infection can result in a release of 50-100 viroids per host cell within 30-35 minutes (Hendrix and Casjens, 2006), not only killing the host cell but also threatening neighboring cells. To rapidly counteract such a potentially devastating infection, it would make sense if the interference components of the CRISPR/Cas system would be localized close to the site of infection. Moreover, the Cas proteins are spatially separated from chromosomal DNA, which is located in the center of the cell; this might contribute to preventing the protein machinery from targeting its own chromosomal CRISPR DNA.

Results and Discussion

The *E. coli* K12 genome contains one *cas* gene cluster, encoding 8 Cas proteins, which is located upstream the CRISPR-I locus. The CasABCDE proteins form the Cascade complex that, loaded with crRNA and assisted by Cas3, targets invading DNA (Brouns et al. 2008) (Chapter 3). An earlier genome-wide localization study determined the spatial arrangement of individually overexpressed Cas proteins (Kitagawa et al. 2005). Since recent studies indicated that expression of *cas* genes is repressed under native conditions in *E. coli* (Pul et al. 2010) (Chapter 6) it is unlikely that the individually overexpressed Cas proteins are incorporated into a functional Cascade complex. This assumption is in agreement with the observation that CasC is spatially separated from both CasA and CasE (Kitagawa et al. 2005). In addition, the study mentioned above is

performed with a low resolution microscope. We therefore used PALM imaging as our spectroscopy method of choice to study the localization of both Cascade and Cas3. PALM employs photo-activatable fluorescent proteins that are sequentially being activated and recorded (Betzig et al. 2006). In this manner only a few fluorescent proteins are being activated and excited per image. This imaging cycle is repeated several times, until all proteins are photobleached. Finally the collected images are merged and the location of each fluorescent protein can be mathematically inferred with high accuracy (Betzig et al. 2006). PALM has also been used to study the cellular location of proteins involved in chemotaxis in *E. coli* (Greenfield et al. 2009).

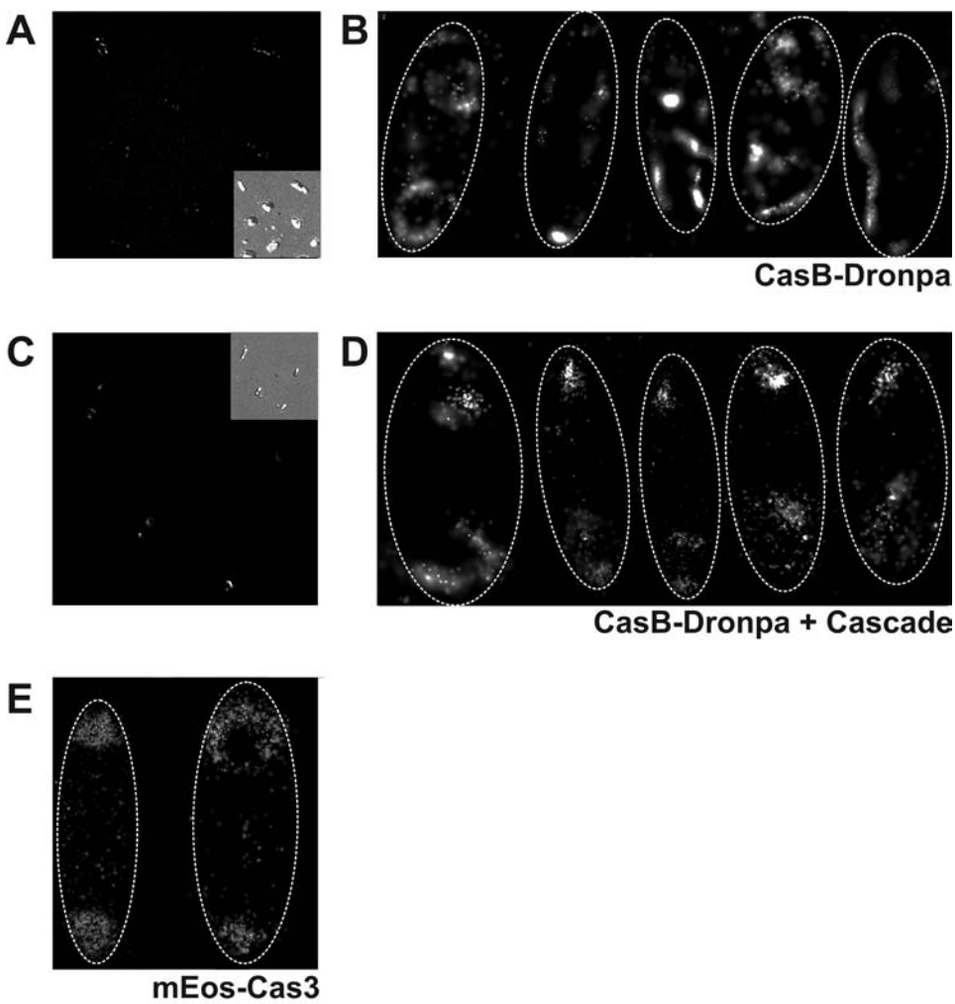


Figure 5.1. PALM images of the location of the Cascade proteins and Cas3. For full colour version see page 132. (A) An overview of CasB-Dronpa expressing cells visualized with PALM. A Differential Interference Contrast (DIC) image (inset) shows that cells look unhealthy. Several images of cells are selected, rotated and enlarged, as shown in (B). Most proteins are located in patches along the cell wall. (C) An overview of CasB-Dronpa expressing cells in the presence of Cascade is visualized with PALM. A DIC image (inset) shows that cells look natural. Several images of cells are selected, rotated and enlarged as shown in (D). In the presence of Cascade, CasB-Dronpa is differently distributed in the cell, mainly concentrated in one focus at one pole of the cell and in a ring-like structure at the other end. (E) PALM analysis of two mEos-Cas3 expressing cells (representing a large population of cells) shows that localization is very similar to that of CasB-Dronpa + Cascade (D), although the ring-like structure can be absent and replaced by a focused cluster. Each spot in the images represents one protein molecule. Cell borders are roughly indicated with dashed lines for clarity in (B), (D) and (E).

Cascade and Cas3 are located in clusters near the cellular poles

First we expressed CasB fused to a photo-activatable fluorescent protein (CasB-Dronpa) in *E. coli* BL21(DE3), a strain that lacks *cas* genes (Studier et al. 2009) and visualized the distribution with PALM (Fig. 5.1AB), in the absence of Cascade and CRISPRs. CasB-Dronpa was succesfully expressed, but the cells appeared to be very sick when viewed by Differential Interference Contrast (DIC) microscopy (see inset in Fig. 5.1A). The distribution of CasB-Dronpa is mainly located along the intracellular membrane. Previous attempts to affinity purify overexpressed StrepII-tagged CasB failed and polyacrylamide gel electrophoresis revealed that CasB was present in the insoluble fraction (data not shown), possibly due to the absence of the other Cascade subunits. The clustering of CasB might be explained by the formation of insoluble aggregates, or inclusion bodies. We therefore decided to co-express the Cascade protein complex, allowing CasB-Dronpa to integrate in the Cascade complex. The PALM images of CasB-Dronpa in the presence of Cascade show a striking difference to the images lacking Cascade (Fig. 5.1CD). CasB-Dronpa in the presence of Cascade is localized at the two poles of the *E. coli* cell, forming a focused cluster at one pole and a ring-like structure at the other pole (Fig. 5.1D).

It has been shown that Cascade, Cas3 and crRNA are required for immunity in *E. coli* (Brouns et al. 2008), suggesting they might be in close proximity of each other. However, Cas3 does not co-purify with Cascade when tagged Cascade is affinity purified from cell extracts containing both Cascade, Cas3 and crRNA (data not shown). This suggests that Cascade and Cas3 do not stably interact. In line with the hypothesis of co-localization, we find a similar distribution for Cas3 (Fig. 5.1E). To confirm this co-localization, a similar experiment should be performed with a strain that contains mEos-Cas3, CasB-Dronpa and Cascade; mEos and Dronpa are compatible photo-activatable fluorescent proteins with distinct spectral properties and can be visualized separately in a single experiment (Shroff et al. 2007).

Cas1 and Cas2 are located near one pole

Cas1 has endonuclease activity both on single stranded (ss) and double stranded (ds) DNA (Wiedenheft et al. 2009). The small Cas2 protein has RNase activity on ssRNA and preferentially cleaves in U-rich regions (Beloglazova et al. 2008). Both Cas1 and Cas2 are not essential for CRISPR expression and interference (Brouns et al. 2008) and are the only two Cas proteins that are present in all subtypes. It has been hypothesized that these proteins are involved in spacer integration (Makarova et al. 2006; Marraffini and Sontheimer 2009; van der Oost et al. 2009). To investigate if they are localized in the same region of the bacterial cell, we first expressed Cas1-Dronpa and Cas2-

mEos separately in different experiments. In analogy to the experiment described above, we used PALM to visualize the spatial distribution of both Cas proteins (Fig. 5.2). Both Cas1-Dronpa and Cas2-mEos are mainly located at one pole. Therefore, it would be interesting to see if they are located at the same pole. This could be tested in future experiments with a coexpression culture, since both Cas proteins are fused to distinct compatible photo-activatable fluorescent proteins. However, our finding that Cas2-mEos is located at mainly one pole is in contrast with a recent study that reports on bipolar organization of Cas2-mEos, which is shown using a newly developed 3D PALM technique (Tang et al. 2010). Follow-up studies revealed that protein expression for a longer time results in distribution to both poles. The distribution patterns of Cas1 and Cas2 in this study show similarity to the localization of misfolded proteins. Misfolded proteins can form large polar aggregates (LPAs) at one cellular pole that will be transferred to one daughter cell to increase the fitness of the other daughter cell (Rokney et al. 2009; Winkler et al. 2010). Alternatively, proteins in LPAs can be refolded, resulting in functional polypeptides. However, care should be taken when interpreting this data since overexpression of proteins might lead to aggregation which may not reflect the native situation.

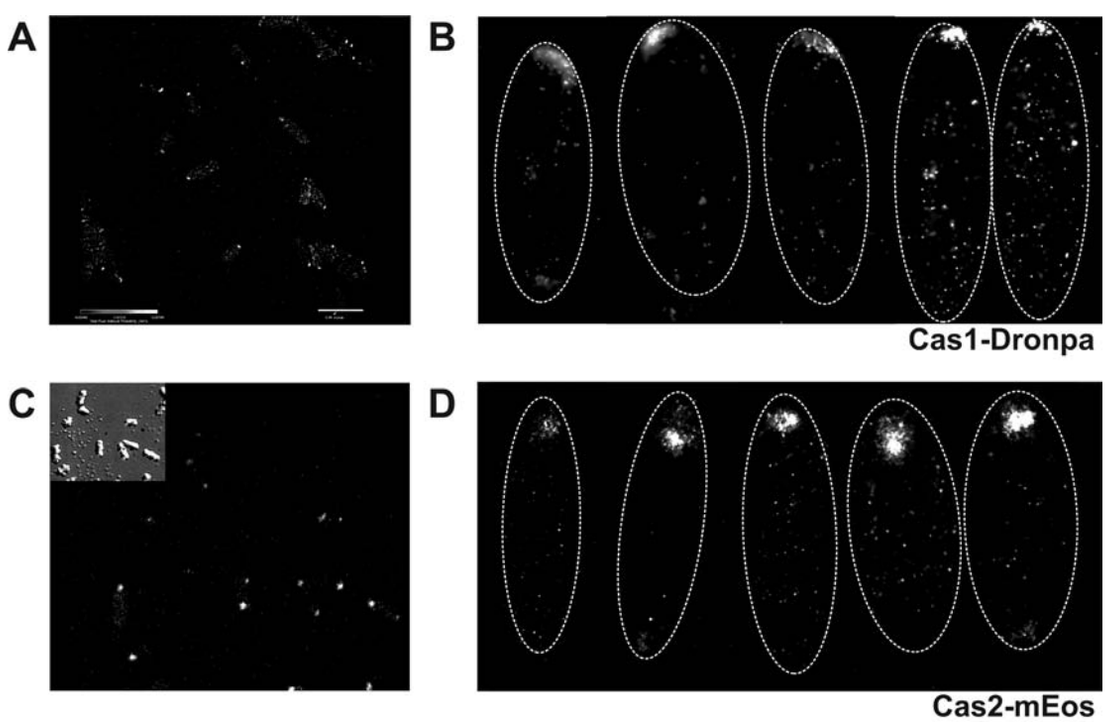


Figure 5.2. PALM images of the location of the Cas1 and Cas2 proteins. For full colour version see page 133. (A) An overview of Cas1-Dronpa expressing cells. Several images of cells are selected, rotated and enlarged as shown in (B). Some proteins localize uniformly over the cells, but most proteins form a cluster that localizes at one cellular pole close to the cell wall. (C) An overview of Cas2-mEos expressing cells. The inset shows a DIC image of the same cells. Several images of cells are selected, rotated and enlarged as shown in (D). As for Cas1-Dronpa, Cas2-mEos localizes mostly as one cluster at one pole of the cell. Each spot in the images represents one protein molecule. Cell borders are roughly indicated with dashed lines for clarity in (B) and (D).

Implications of polar localization

We have found that Cas1, Cas2, Cas3 and CasB-Donpa co-expressed with Cascade, can localize at either one or two poles of the cells, implicating that the majority of the Cas protein machinery is located in this area. Control experiments with Dronpa and mEos only showed uniform distribution over the cells (data not shown). This finding is in agreement with previous observations, where CasC was found to cluster in this area, but in contrast with CasA and CasE localization in the same study (Kitagawa et al. 2005) (http://ecoli.naist.jp/GB6/search.jsp). In this study we co-expressed Cascade and CasB-Dronpa which may result in a functional complex that localizes at a different location than the non-functional single Cascade subunits. This could be tested by affinity purification of CasB-Dronpa to determine whether the Cascade subunits copurify, and altogether form a full Cascade complex which is not affected by the fusion with Dronpa. In addition, an *in vivo* infection assay should be performed to test whether this complex is still functional.

It has been previously reported that several phages, such as phage *lambda*, preferably infect at the cellular poles of *E .coli* (Edgar et al. 2008). To inject its DNA, *lambda* requires the host protein ManY (Scandella and Arber 1974), which is located at the pole (Edgar et al. 2008). Replication of the lambda DNA takes place at the same location (Edgar et al. 2008). It seems to be a “smart decision” of *E. coli* to locate its antiviral Cas machinery in the vicinity of ManY. It could then immediately target lambda DNA that enters the cell and as such prevent the rapid and devastating effects of phage lambda infection. Clustering of Cas proteins and ManY could be confirmed in a future single experiment by labeling ManY and Cas proteins simultaneously. As an alternative to ManY, other proteins that are involved in lambda infection could be labeled. Another intriguing question concerns which factors determine where the Cas proteins go after synthesis. One model for protein localization in prokaryotes is the diffusion and capture model. According to this model proteins diffuse through the cell and are captured by other proteins (Shapiro et al. 2009). If this model applies to the Cas proteins, it remains to be seen to which other cytoplasmic or membrane proteins they attach.

A secondary advantage is perhaps the spatial separation of the Cas protein machinery and the host chromosome. The chromosome is located around the center of the cell (Robinow and Kellenberger 1994). The physical distance may prevent the Cas protein machinery from targeting the chromosome, which contains a CRISPR with partial complementarity to the crRNA. This physical separation might contribute to prevention of auto-immunity, next to sequence requirements outside the protospacer (Deveau et al. 2008; Marraffini and Sontheimer 2010b).

Materials and methods

Strains and plasmids used

Protein overexpression was performed in *E. coli* BL21 (DE3), a strain that lacks *cas* genes (Studier et al. 2009). The plasmids that were used include CasABCDE in pACYCduet, Cas1-Dronpa in pET52b, Cas2-mEos in pRSF1b, mEos-Cas3 in pET52b and CasB-Dronpa in pRSF1b. The construction of these plasmids will be described elsewhere (Jasper Akerboom).

Sample preparation

A fresh culture was inoculated from overnight cultures and grown in LB at 37 °C until the OD_{600} reached ~0.5. The protein expression was induced by the addition of 1 mM IPTG, and the culture was grown until it appeared fluorescent under a UV light (typically 1 – 2 hours). Cells were spun down and resuspended in 1/10 volume PBS. They were subsequently spotted on a hydrophobic glass slip and air-dried, together with nanosphere fiducials. Forty-nanometer gold nanospheres were used in the case of Dronpa fusions, and 100-nm gold nanospheres in the case of mEos fusions (both purchased from Microspheres – Nanospheres).

Sample analysis

Samples were analyzed under an Olympus IX81 inverted microscope, equipped with a DIC optics and a 100x, 1.65 NA objective. Laser beams were delivered to the object from a customized table. The Dronpa and mEos protein fusions were activated with a 405-nm laser. Dronpa fusion proteins were excited with a 488-nm laser whereas mEos were excited with a 561-nm laser. Data sets were collected with a shutter time of 100 ms or 50 ms for Dronpa and mEos, respectively, until cells were photobleached (typically after a few hours). The data analysis was performed as described elsewhere (Greenfield et al. 2009). The drift during data collection was corrected by tracking the fiducials.

Acknowledgments

The authors would like to thank Howard Hughes Medical Institute (HHMI) for technical and financial support. MMJ, JvdO and SJJB were financially supported by an NWO Vici grant to JvdO and a Veni grant to SJJB.

Chapter 6

H-NS mediated repression of CRISPR-based immunity in *Escherichia coli* K12 can be relieved by the transcription activator LeuO

Edze R. Westra*, Ümit Pul*, Nadja Heidrich, Matthijs M. Jore, Magnus Lundgren, Thomas Stratmann, Reinhild Wurm, Amanda Raine, Melina Mescher, Luc van Heereveld, Marieke Mastop, E. Gerhart H. Wagner, Karin Schnetz, John van der Oost, Rolf Wagner, Stan J. J. Brouns

*Contributed equally

Molecular Microbiology*, In press*

Abstract

The recently discovered prokaryotic CRISPR/Cas defense system provides immunity against viral infections and plasmid conjugation. It has been demonstrated that in *Escherichia coli* transcription of the Cascade genes (*casABCDE*) and to some extent the CRISPR array, is repressed by heat-stable nucleoid-structuring (H-NS) protein, a global transcriptional repressor. Here we elaborate on the control of the *E. coli* CRISPR/Cas system, and study the effect on CRISPR-based anti-viral immunity. Transformation of wildtype *E. coli* K12 with CRISPR spacers that are complementary to phage Lambda, does not lead to detectable protection against Lambda infection. However, when an H-NS mutant of *E. coli* K12 is transformed with the same anti-Lambda CRISPR, this does result in reduced sensitivity to phage infection. In addition, it is demonstrated that LeuO, a LysR-type transcription factor, binds to two sites flanking the *casA* promoter and the H-NS nucleation site, resulting in derepression of *casABCDE12* transcription. Over-expression of LeuO in *E. coli* K12 containing an anti-Lambda CRISPR leads to an enhanced protection against phage infection. This study demonstrates that in *E. coli* H-NS and LeuO are antagonistic regulators of CRISPR-based immunity.

Introduction

Invasions by viruses and conjugative plasmids pose a threat to microbial cells. To neutralize selfish DNA elements, bacteria and archaea have developed several defense strategies, such as receptor masking, restriction/modification and abortive infection (Hyman and Abedon, 2010; Labrie *et al.*, 2010). Recently it was discovered that Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) (Ishino *et al.*, 1987) and CRISPR associated (*cas*) genes (Jansen *et al.*, 2002) form a sophisticated immune system that uses small RNAs to target mobile genetic elements, reviewed by (Horvath and Barrangou, 2010; Karginov and Hannon, 2010; Marraffini and Sontheimer, 2010a; van der Oost *et al.*, 2009). CRISPRs consist of repeating sequences of approximately 30 nucleotides that are separated by unique sequences of similar size, called spacers (Mojica *et al.*, 2000). The spacer sequences are commonly derived from phages and plasmids (Bolotin *et al.*, 2005; Mojica *et al.*, 2005; Pourcel *et al.*, 2005), and new spacers can be added to the existing CRISPR array, expanding the invader repertoire (Barrangou *et al.*, 2007), in a process known as CRISPR adaptation. The presence of a spacer matching a viral or plasmid sequence confers resistance to invasion by these elements (Barrangou *et al.*, 2007; Brouns *et al.*, 2008; Marraffini and Sontheimer, 2008). The biochemical pathways underlying CRISPR defense are partially known and involve transcription of the array into a long precursor CRISPR RNA. This precursor is cleaved in the repeat sequences by a Cas endonuclease (CasE in *E. coli* (Brouns *et al.*, 2008), Cas6 in *Pyrococcus furiosus* (Carte *et al.*, 2008)), releasing small crRNAs that serve to guide the defense.

The *cas* genes encode the protein machinery that carries out the various steps of CRISPR defense. Approximately 45 families of *cas* genes have been identified (Haft *et al.*, 2005) that are classified in eight typical combinations or subtypes named after a representative organism, *e.g.* type E after *E. coli* (Haft *et al.*, 2005). The type E CRISPR/Cas immune system in *E. coli* K12 is composed of 8 *cas* genes (*cas1*, *cas2, cas3* and *casABCDE*) and a downstream CRISPR locus with type 2 repeats (Kunin *et al.*, 2007) containing 12 spacer-repeat units (CRISPR I) (Fig. 6.1A). An additional 6 spacer-containing CRISPR (CRISPR II) and a 2 spacer CRISPR (CRISPR III) with type 2 repeats, as well as a CRISPR with type 4 repeats (Kunin *et al.*, 2007) containing 1 spacer repeat unit (CRISPR IV) are located elsewhere on the genome (Diez-Villasenor *et al.*, 2010). In addition to a CRISPR containing an anti-invader sequence, only Cas3 and CasABCDE, forming the protein complex Cascade (CRISPR-associated complex for antiviral defense), are required for CRISPR interference (Brouns *et al.*, 2008). A recent study has demonstrated that in *E. coli* K12 transcription from the *casA* and CRISPR I promoters is repressed by heat-stable nucleoid-structuring protein (H-NS)

(Pul *et al.*, 2010), a global repressor of transcription in many Gram-negative bacteria. A microarray study indicates that transcription of *casABC* and *cas2* is elevated in an *E. coli* K12Δ*hns* strain compared to *wt E. coli* K12 (Hommais *et al.*, 2001). In addition, H-NS was shown to possess high binding affinity for the intergenic region between *cas3* and *casA* (Oshima *et al.*, 2006; Pul *et al.*, 2010). H-NS has a preference for binding AT-rich DNA sequences (Navarre *et al.*, 2006). After initial binding of H-NS to high affinity nucleation sites (Bouffartigues *et al.*, 2007; Lang *et al.*, 2007) repression of transcription is mediated by cooperative spreading along the DNA (defined as DNA stiffening (Liu *et al.*, 2010)) and by creating looped structures through formation of DNA-protein-DNA bridges (Dame *et al.*, 2005). Moreover, H-NS acts as a DNA structuring protein (Liu *et al.*, 2010; Stoebel *et al.*, 2008).

Overcoming H-NS mediated repression of *cas* gene transcription may be a key requirement for CRISPR/Cas functionality. Generally, H-NS repression can be relieved by a number of proteins, such as SlyA, VirB and others (Stoebel *et al.*, 2008). One of these proteins is the regulator LeuO (Chen and Wu, 2005; De la Cruz *et al.*, 2007), which belongs to the LysR family of transcription factors (Stoebel *et al.*, 2008) and is found in all proteobacteria, except the δ subdivision (Maddocks and Oyston, 2008). The *leuO* gene maps next to the *leuABCD* operon (Chen *et al.*, 2005; Chen and Wu, 2005; Hertzberg *et al.*, 1980), whose gene products are required for leucine synthesis (Vartak *et al.*, 1991). Recent data indicate that LeuO is involved in regulating transcription of many genes, often as an H-NS antagonist (Shimada *et al.*, 2009; Stoebel *et al.*, 2008). However, since under laboratory growth conditions the genomic *leuO* gene itself is repressed by H-NS (Klauck *et al.*, 1997; Majumder *et al.*, 2001) all LeuO regulation studies make use of plasmid encoded *leuO* under control of constitutive or ininducible promoters. In the host environment *leuO* is likely to be induced under certain conditions as for example virulence of *Salmonella enterica leuO* mutants is attenuated (Lawley *et al.*, 2006). A genomic screen for LeuO-binding DNA fragments in *E. coli* K12 revealed 12 gene clusters, including the *casA-cas2* operon (*ygcL*, *ygcK*, *ygcJ*, *ygcI*, *ygcH*, *ygbT*, *ygbF*) (Shimada *et al.*, 2009). When LeuO was over-expressed, increased expression of *casA* and *cas2* was observed in *E. coli* (Shimada *et al.*, 2009), and of *casA* (STY3070) in *Salmonella enterica* serovar Typhi (Hernandez-Lucas *et al.*, 2008). We therefore investigated whether LeuO can mediate H-NS derepression of *cas* gene and CRISPR transcription. In this study we demonstrate that LeuO counteracts H-NS-dependent repression of the *casA* promoter by reorganizing the DNA protein contacts within the transcription initiation region. The resulting change results in increased transcription of the Cascade genes, the limiting factor for CRISPR-based defense against phage infection in *E. coli* K12.

Results

LeuO activates cas gene expression

To study the effect of LeuO on *cas* gene expression, transcript levels of the *E. coli* K12 *cas* genes in mid-exponential growth phase were examined using a DNA microarray approach. RNA samples isolated from a *wt E. coli* K12 strain containing a *leuO* encoding plasmid were compared to RNA isolated from a strain containing the empty vector. In addition, RNA isolated from a Δ*leuO* mutant carrying the empty vector was analyzed. Comparison of *cas* gene transcription levels between the LeuO-expressing strain and the control strain revealed a significant upregulation of transcription of *casABCDE* and *cas1* and *cas2* transcription, showing a gradual decrease from *casA* (65-fold) to *cas2* (5-fold) (Table 6.1). No change in the transcription level of *cas3* was detected. These results are consistent with a polycistronic transcription of the *casABCDE* and possibly the *cas1*, *cas2* genes, with polar effects for the transcription of the more downstream genes. However, we did not observe significant differences in *cas* gene transcription in the Δ*leuO* mutant compared to the wildtype strain (Table 6.1), indicating that *leuO* is not expressed under the growth conditions used here.

To verify the observed increase in *cas* gene expression levels, quantitative PCR (qPCR) was performed on total RNA isolated from 3 strains during mid-exponential phase: *wt E. coli*, a Δ*hns* strain and a *wt* strain expressing *leuO* from a plasmid. This analysis showed that *casABCDE* displayed increased transcription in both *hns* knockout and *leuO* expressing strains (Fig. 6.1B). While the increase in *casABCDE* transcripts was

Table 6.1. Microarray analysis of activation of *cas* genes by LeuO

	pLeuO / wt		ΔleuO / wt	
gene	**fold-change[a]**	***p* value[b]**	**fold-change[a]**	***p* value[b]**
cas3	1.0	n. s. 0.93	1.1	n. s. 0.76
casA	65.4	< 0.05	-1.2	n. s. 0.08
casB	30.0	< 0.05	-1.1	n. s. 0.20
casC	24.8	< 0.05	1.0	n. s. 0.81
casD	17.5	< 0.05	1.1	n. s. 0.68
casE	15.4	< 0.05	-1.2	n. s. 0.31
cas1	8.8	< 0.05	1.2	< 0.05
cas2	5.4	< 0.05	1.1	n. s. 0.34

[a] The fold change of cas genes expression was determined by microarray analysis. pLeuO / wt indicates the ratio of *cas* transcripts detected upon overexpression of LeuO (using plasmid pKEDR13) as compared to wildtype *E.coli* K12 (transformed with the empty vector plasmid pKESK22). Δ*leuO* / wt indicates the ratio of cas transcripts detected in a $\Delta leuO_{FRT}$ mutant as compared to wildtype.

[b] n.s. is not significant

modest in *hns* knockout strains, on average 5-fold, the effect of introducing *leuO* was more dramatic, with an average increase of 236-fold after induction of *leuO* expression. An increase in *casABCDE* transcripts was also observed when *leuO* expression was not induced, due to leakage from the P*T5/lac* promoter. The *cas1* and *cas2* genes also displayed increased transcript abundance in *leuO* expressing strains, although at lower levels than *casABCDE*. Consistent with the microarray data, a trend of transcript fold change was observed, with polar effects downstream of *casA*, again suggesting a polycistronic mRNA of the *casABCDE12* operon. Compared to the effect on the other *cas* genes, only a small increase of *cas3* transcription was observed in *leuO* expressing strains.

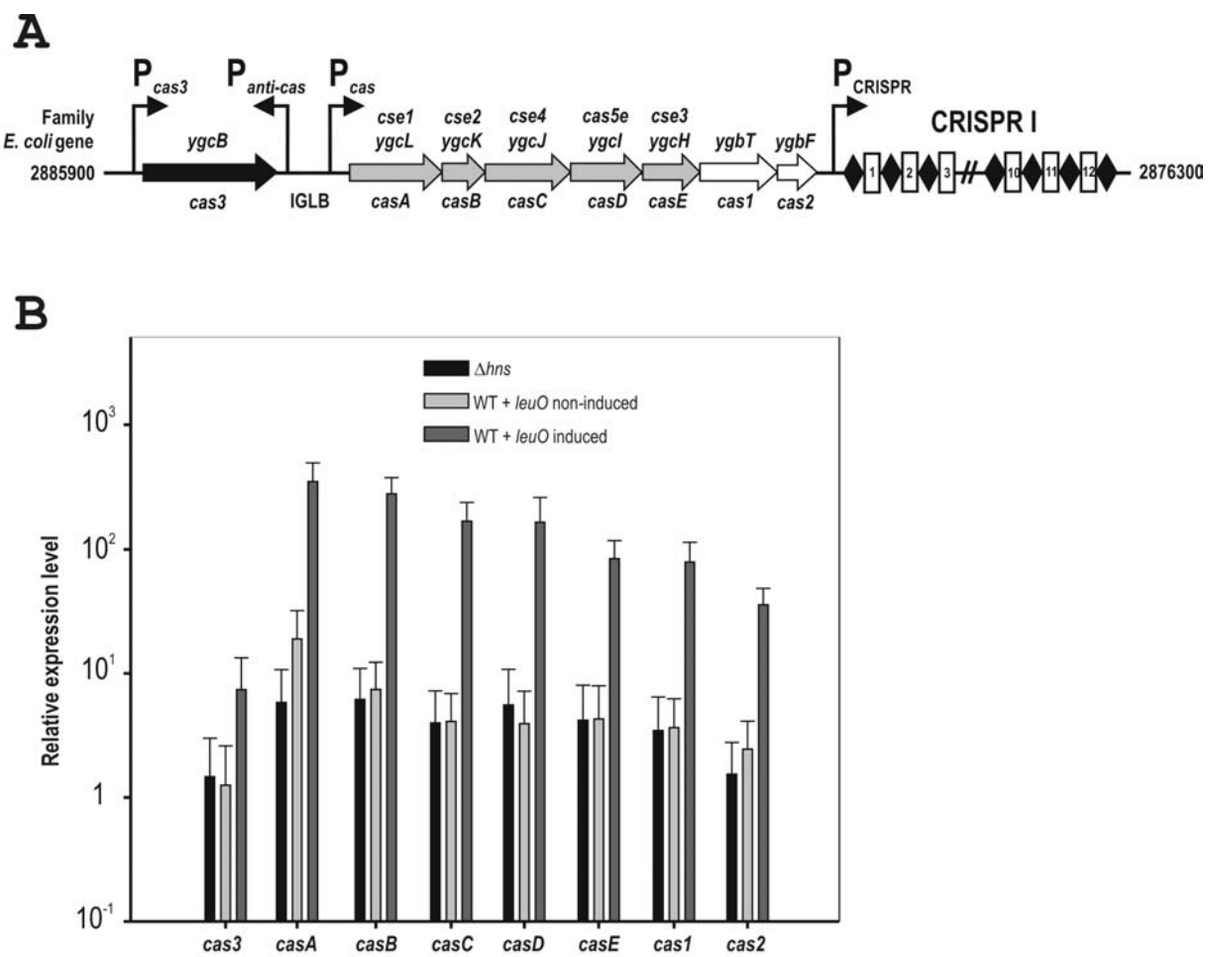


Figure 6.1. LeuO and H-NS regulate *cas* **gene expression.** A) Schematic illustration of the CRISPR/Cas locus in *E. coli* K12 that consists of 8 *cas* genes (*cas3* (*ygcB*), *casA* (*ygcL* or *cse1*), *casB* (*ygcK* or *cse2*), *casC* (*ygcJ* or *cse4*), *casD* (*ygcI* or *cas5e*), *casE* (*ygcH* or *cse3*), *cas1* (*ygbT*) and *cas2* (*ygbF*)) and a downstream CRISPR locus containing 12 spacers and 13 repeats (CRISPR I). The *cas3*, anti-cas3 (anti-P*cas*), casA (P*cas*) and CRISPR I promoter are indicated with an arrow (Pul et al., 2010). B) qPCR analysis of *cas* gene transcript abundance in *E. coli* Δ*hns* and *E. coli* W3110 expressing *leuO* (induced or non-uninduced). Fold changes are given as compared to *wt E. coli* W3110 expression levels. Error bars indicate one standard deviation.

To further evaluate the effects of H-NS and LeuO on transcription from the *casA* promoter (known as P*cas* (Pul *et al.*, 2010)), RNA samples from *wt* strains expressing *leuO* from a plasmid and strains lacking *hns* were compared by primer extension analysis. No *cas* transcripts were detected in *wt* cells containing an empty expression vector. Transcripts directed from P*cas* were only detected in cells expressing *leuO* from a plasmid or in *hns* knockout strains (Fig. S6.1), indicating that transcription of the *casABCDE12* operon is tightly controlled by H-NS and LeuO.

LeuO causes increased crRNA abundance

The CRISPR I locus is transcribed in *E. coli* K12 and the transcript is cleaved by the CasE subunit of Cascade into small crRNAs that subsequently remain bound by Cascade (Brouns *et al.*, 2008; Pul *et al.*, 2010). In K12 small crRNAs were virtually undetectable by Northern blot analysis (Brouns *et al.*, 2008) and (Fig. 6.2A). To investigate whether this was due to too low transcription levels of *casABCDE*, the *wt* strain was transformed with a plasmid encoding the Cascade protein components under control of an arabinose-inducible promoter. In the *wt* strain expressing *casABCDE* from a plasmid, crRNAs with a length of about 60 nt could be detected. The requirement for plasmid-encoded synthesis of Cascade for detection of small crRNAs indicates that the level of Cascade in *wt E. coli* is insufficient for generating and stabilizing mature crRNAs. Furthermore we analyzed the levels of crRNAs in an *hns* knockout

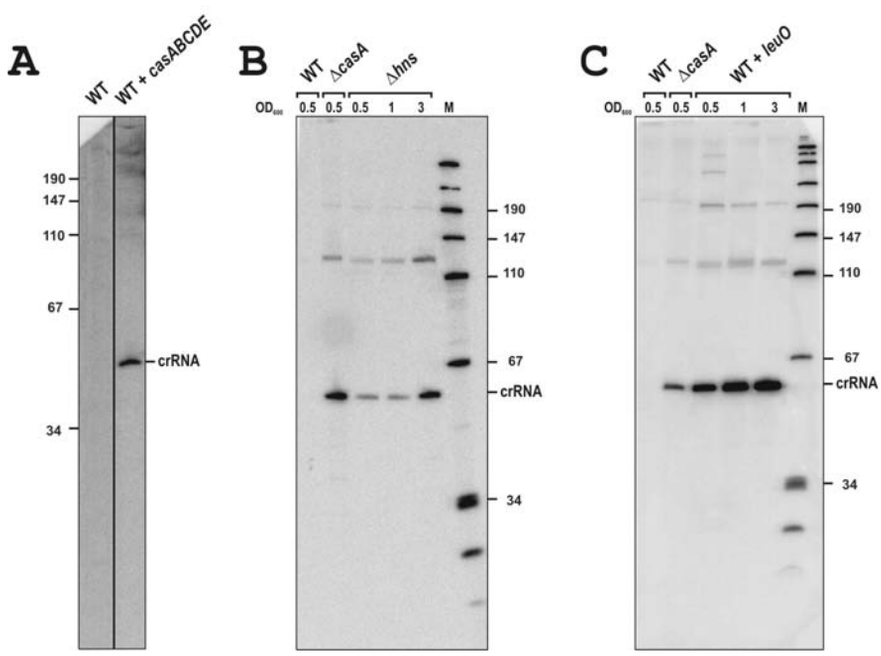


Figure 6.2. Formation of mature crRNA. A) Northern analysis of total RNA from *wt E. coli* K12 and *wt E. coli* K12 + *casABCDE* (pNH6) using the single stranded spacer sequence NH30 (Table S6.2) as a probe. B) Northern analysis as in (A) of total RNA from *E. coli* K12 Δ*hns* and C) *wt E. coli* K12 expressing *leuO* from a plasmid (pNH41) with an OD_{600} of 0.5, 1 and 3.0. M, size marker (pUC19/MspI ladder). *E. coli* K12 Δ*casA* (JW2730) serves as a control and marker for mature crRNA.

strain and in the *wt* strain expressing *leuO* constitutively from a plasmid. Both deletion of *hns* and over-expression of *leuO* caused significant crRNA accumulation, due to enhanced expression of Cascade in these two strains (Fig. 6.2B and C). The CasA knockout strain (JW2730) serves as a control and marker for the mature crRNA. It was previously shown that a CasA knockout strain generates elevated levels of mature crRNA (Brouns *et al.*, 2008), due to read-through of the downstream *cas* genes from the kanamycin resistance marker containing recombination cassette by which the *casA* gene is replaced (Pougach et al., in press).

Binding of LeuO and H-NS to the DNA sequence upstream of casA

The *casA-cas3* intergenic region (here denoted IGLB) contains P*cas*, for which H-NS has strong binding affinity as well as the divergently oriented anti-*cas3* (known as anti-P*cas*) promoter, that is located 80 bp upstream of P*cas* and gives rise to an antisense transcript of unknown function (Fig. 6.1A and Fig. 6.3C) (Pul *et al.*, 2010). Both LeuO and H-NS bind the IGLB fragment, as determined by Electrophoretic Mobility Shift Assay (EMSA) (Fig. 6.3A, lanes 2-4 and 5-7). Pre-bound LeuO impedes cooperative binding of H-NS to the IGLB fragment (Fig. 6.3A, lanes 9-11). In line with this, pre-bound H-NS is partly released from the DNA when LeuO is added to the complex (Fig. 6.3A lanes 12-15). In order to map the binding region of LeuO within the IGLB fragment DNase I footprint analysis was performed. Upon limited DNase I hydrolysis of the IGLB DNA fragment, H-NS causes an extended footprint (Fig. 6.3B), as shown before (Pul *et al.*, 2010). In addition, LeuO protects two sites (site 1 and site 2) within the IGLB fragment that flank the high affinity H-NS nucleation site (Fig. 6.3B and 6.3C). LeuO site 1 is located 20 bp downstream of P*cas* and LeuO site 2 spans the divergent anti-P*cas* (Fig. 6.3B and C). Interestingly, in the presence of LeuO the extended protection by H-NS is no longer visible (Fig. 6.3B, compare lanes 2 and 4), indicating that due to LeuO binding the DNA region containing the H-NS high-affinity binding site is no longer protected from DNase I cleavage, in agreement with decreased cooperative binding (Fig. 6.3A).

In order to analyze the effect of LeuO on RNA polymerase (RNAP) binding to the promoter sites, DNase I footprints were performed in the presence of RNAP and LeuO. Moreover, the effect on transcription initiation and RNAP open complex formation was analyzed by $KMnO_4$ footprints of stable initiation complexes. RNAP binds to the two promoters (P*cas* and anti-P*cas*) (Fig. 6.4A, lanes 3 and 3', indicated I (P*cas*) and II (anti-P*cas*)) (Pul *et al.*, 2010). Addition of the DNA binding proteins LeuO or H-NS alone does not cause changes in the $KMnO_4$ reactivity (Fig. 6.4A, lanes 2', 4' and 5'). Binding of LeuO abolishes the spreading of H-NS along the DNA, resulting in a lack of

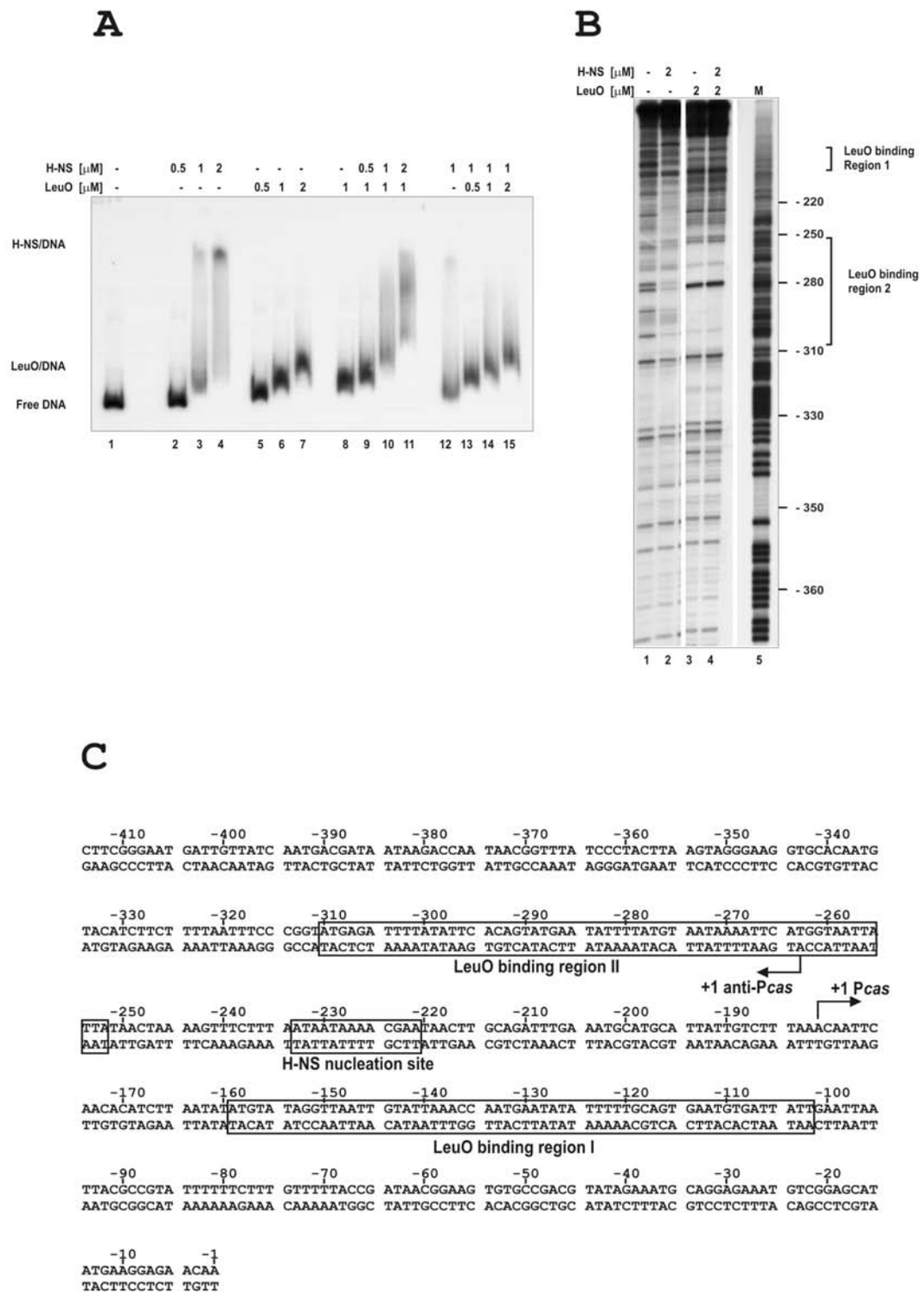


Figure 6.3. H-NS and LeuO binding to the DNA region upstream of casA (the IGLB fragment). A) Electrophoretic Mobility Shift Assay (EMSA) of the IGLB fragment with LeuO and H-NS, either alone (lanes 5-7) or with pre-bound LeuO and subsequent addition of H-NS (lanes 8-11) or pre-bound H-NS and subsequent addition of LeuO (lanes 12-15). B) DNase I footprint of IGLB in the presence of either H-NS or LeuO or both. LeuO was pre-incubated with the reaction mixture. The 2 main regions protected by LeuO are indicated. The ladder indicates the IGLB coordinates relative to the *casA* start codon, indicated in C) the sequence and coordinates of IGLB, with the H-NS and LeuO binding sites indicated with boxes. See also Fig. 6.4A for footprints showing the LeuO binding region I.

protection by H-NS in the region between positions -160 to -240 (Fig. 6.4A, compare lanes 2 and 5), as observed before (Fig. 6.3B). When RNAP binding was studied in the presence of both transcription factors it turned out that the order of addition to the DNA is crucial for the resulting footprint. RNAP binding was only affected when H-NS and/or LeuO were added to the DNA prior to RNAP. While prior binding of H-NS to the IGLB DNA fragment completely abolished RNAP-promoter interaction and open complex formation (Fig. 6.4A and B, compare lanes 3' and 6'), prior binding of LeuO had a repressive effect only on RNAP binding at anti-P*cas* (Fig. 6.4A and B, compare lanes 3' and 7'; Fig. 6.4E, lane 5'). This can also be seen on the retardation gels (Fig. 6.4C and D), where the DNA/RNAP complex II is lost in the presence of LeuO (Fig. 6.4C, lane 7). This complex remains stable when H-NS is added last (Fig. 6.4C, lane 9), in contrast to a sample with only H-NS or where H-NS is added before LeuO (Fig. 6.4C, lanes 6 or 8, respectively). Moreover, the change in nucleotide reactivities indicates that LeuO binding alters the architecture of the transcription initiation complex at the P*cas* promoter (compare Fig. 6.4B, lanes 7' with Fig. 6.4A, line 3' and Fig. 6.4E, lane 3' with lane 5'). Altogether these data indicate that LeuO plays an important role in the regulation of *casABCDE12* gene expression by antagonizing H-NS-dependent repression of P*cas*.

H-NS and LeuO regulate CRISPR-based immunity against phage infection

The effect of H-NS on CRISPR-based defense against phage infection was analyzed in *wt* and Δ*hns E. coli* strains, grown in Luria Bertani broth. Since none of the spacers of *E. coli* K12 target known bacteriophages, an artificial seven spacer CRISPR (J3) with the native promoter was designed containing one spacer that targets the template strand of the gene encoding the phage Lambda tail protein (J). A non-targeting (N) CRISPR (Brouns *et al.*, 2008) served as a negative control. Introducing the J3 CRISPR reduced the sensitivity to virulent phage Lambda (λ_{vir}) infection 4-fold in the Δ*hns* but not in the *wt* strain (Fig. 6.5A). Complementation of the Δ*hns* strain reversed the reduction in phage sensitivity, demonstrating that CRISPR-based defense is negatively regulated by H-NS (Fig. 6.5A). When cells were grown in richer media (2YT) until stationary growth phase, higher resistance levels were observed, up to a 10-fold reduced sensitivity compared to a non-targeting strain (data not shown). Moreover, plaques were much smaller in the *hns* knockout strains equipped with the J3 CRISPR when using 2YT. Further evidence that H-NS controls CRISPR-based immunity was obtained using genetically engineered strains (Table S6.4) in which the genomic CRISPR I locus was replaced by the J3 or a non-targeting CRISPR. *E. coli* K12Δ*CRISPRI::CRISPRJ3* (*E. coli* J3) was fully sensitive to infection by phage Lambda, despite the presence of a genomic J3 spacer (Fig. S6.2). However, when the dominant negative *hns*^{G113D} mutant

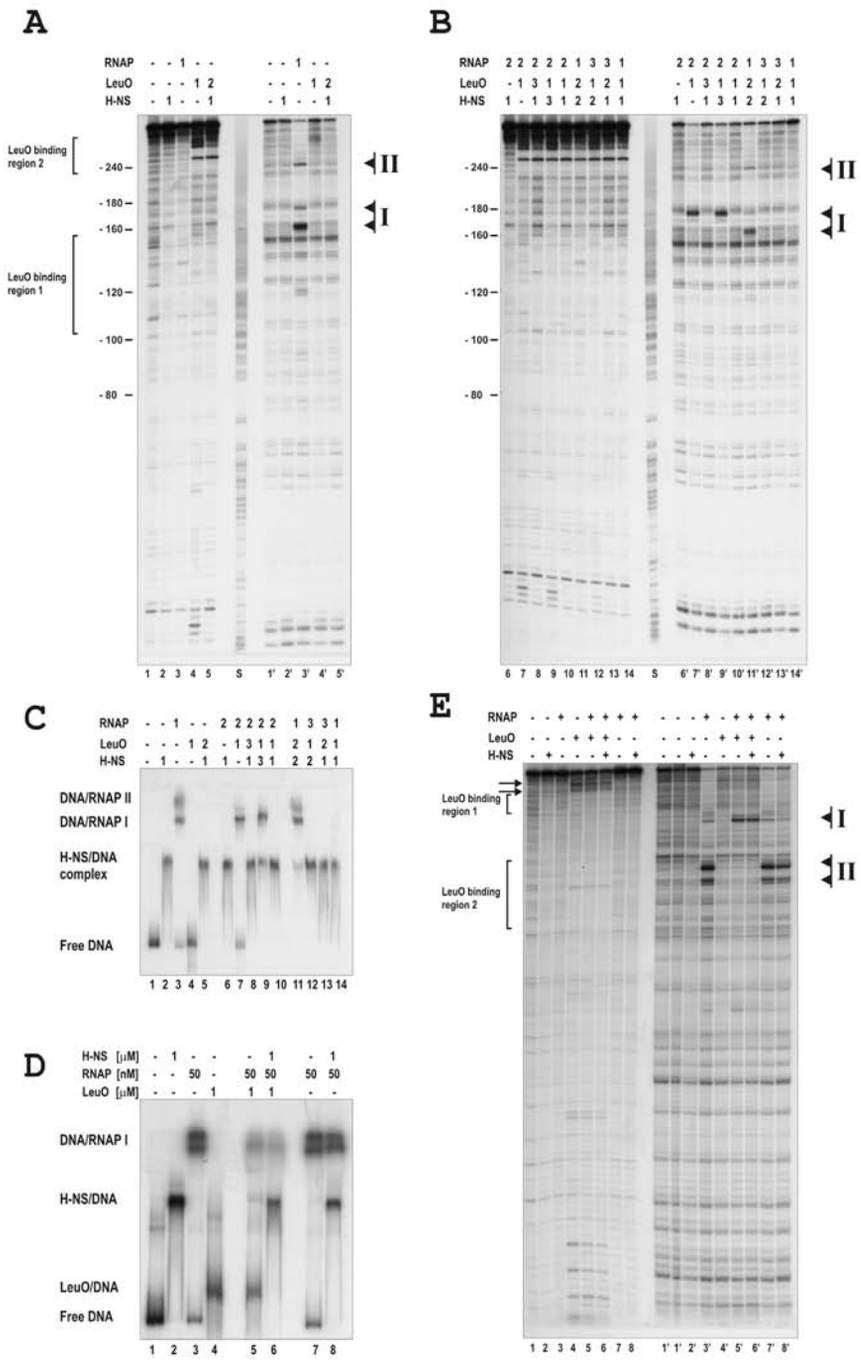


Figure 6.4. Effects of LeuO and H-NS on RNA polymerase open complex formation at the IGLB fragment. A) and B) show either DNase I (lanes 1-14) or KMnO4 (lanes 1'-14') footprint analyses of the non-template strand of IGLB after binding of RNA polymerase, H-NS or LeuO, either individually or in combination. The numbers above the lanes indicate the order of addition of the respective proteins. The *casA* promoter (P*cas*) (I) and the anti-*cas3* promoter (anti-P*cas*) (II) are indicated. The arrowheads indicate the nucleotides within open complexes at the promoter sites. C) and D) show the binding of RNA polymerase, H-NS or LeuO, either individually or in combination. C) shows Electrophoretic Mobility Shift Assays of the samples in (A) and (B). The numbers above the lanes indicate the order of addition of the respective proteins. D) shows Electrophoretic Mobility Shift Assays of the samples in (E). The positions of the free DNA and complexed DNA are given on the left. The retardation gels (C) and (D) are not standard retardation gels but controls for the footprint complexes to see if no major decomposition has occurred. Because complexes are kept at 30°C for some time under RNAP binding conditions these gels are sometimes different from standard retardation experiments (e.g. Fig. 6.3A). E) Shows a similar analysis as in (A) and (B) for the template strand of IGLB (DNase I footprint lanes 1-8; KMnO4 footprint lanes 1'-8'). The protein concentrations used are H-NS (1 μM), LeuO (1 μM) and RNA polymerase (50 nM).

was expressed from a plasmid, the sensitivity of *E. coli* J3 to phage λ_{vir} infection was reduced 3.6 fold (Fig. S6.2). This mutant still forms heterodimers with *wt* H-NS, but does not bind DNA and therefore interferes with H-NS mediated transcriptional repression resulting from the formation of higher-order DNA-protein complexes (Ueguchi *et al.*, 1996). This observation is consistent with the reported finding that expression of *hns* mutant G113D induces transcription from P*cas* in *wt E. coli* (Pul et al., 2010).

When plaque assays were performed in *E. coli* J3 over-expressing *leuO* from a plasmid, a ~6 fold reduced sensitivity to phage λ_{vir} infection was observed (Fig. 6.5B), demonstrating that LeuO activates CRISPR-based defense in *E. coli*. A 3-fold reduced sensitivity was observed when *leuO* expression was not induced, probably due to

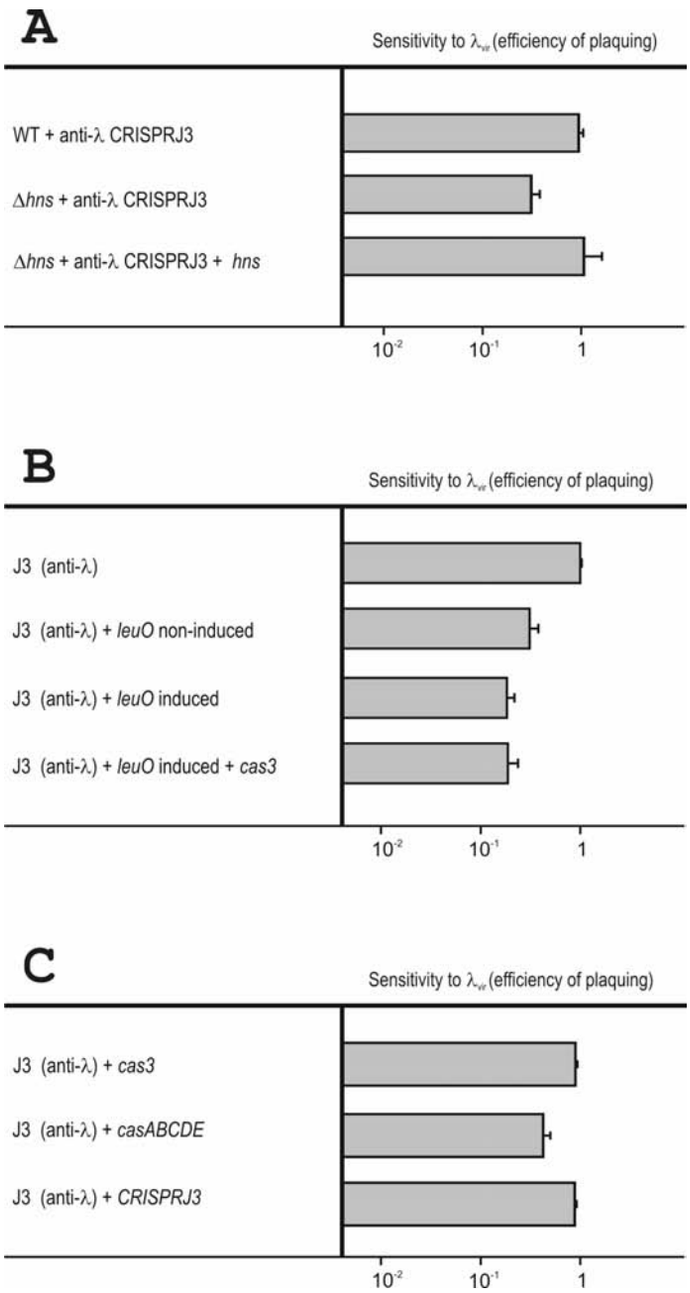


Figure 6.5. Effect of H-NS and LeuO on CRISPR-based resistance *in vivo*. A) A synthetic CRISPR with one spacer (J3) targeting phage Lambda on a plasmid (pWUR564) is introduced in *E. coli* K12Δ*hns*, *E. coli* K12, and complemented *E. coli* K12Δ*hns* expressing *hns* from the pHOP11 plasmid. Phage resistance is monitored by determining the efficiency of plaquing (EOP) after challenge with virulent Lambda phage. B) The effects of *leuO* (pKEDR13) and *cas3* (pWUR608) expression on phage resistance is monitored in *E. coli* K12Δ*CRISPRI::CRISPRJ3* (indicated as J3) C) The effect of introducing the *casABCDE12* (pWUR607), *cas3* or a CRISPR on phage resistance is monitored in *E. coli* K12Δ*CRISPRI::CRISPRJ3* (indicated as J3).

leakage from the P*tac* promoter. When *E. coli* J3 cells were grown to stationary phase in rich 2YT medium, an increased resistance level was observed with turbid and very small plaques in the *leuO* over-expressing strains containing a targeting CRISPR (data not shown), whereas plaques in the same strain containing a non-targeting CRISPR were clear and of normal size. Although *cas3* gene expression was not strongly induced when LeuO was expressed from a plasmid (Table 6.1 and Fig. 6.1B), the expression of *cas3* was not a limiting factor for resistance, since introduction of a *cas3* expression plasmid into *E. coli* J3 expressing *leuO* did not lead to elevated resistance levels (Fig. 6.5B). When a plasmid expressing *casABCDE12* was introduced in *E. coli* J3, a 2.5-fold reduced sensitivity to phage infection was observed (Fig. 6.5C), which was not observed when a CRISPR expression vector containing spacer J3 or a plasmid encoding *cas3* was introduced (Fig. 6.5C), indicating that expression of the genes encoding Cascade (*casABCDE*) is limiting for CRISPR-based defense in *wt E. coli* .

Discussion

The type E CRISPR/Cas system (Cse (Haft *et al.*, 2005)) is present in many proteobacteria and in some actinobacteria, firmicutes and methanogenic archaea. A recent analysis of a collection of natural isolates shows that this CRISPR/Cas subtype occurs in approximately 60% of the *E. coli* strains (Diez-Villasenor *et al.*, 2010). The study presented here provides experimental evidence for regulation of the type E CRISPR/Cas system in *E. coli* K12 by the antagonists H-NS and LeuO. These antagonistic DNA-binding proteins regulate the expression of several genes in *E. coli*, such as the *bgl* operon (utilization of β-glucosides) (Ueguchi *et al.*, 1998), the *yjjQ-bglJ* operon (virulence factor and activator of *bgl*, respectively) (Stratmann *et al.*, 2008) and the *Salmonella enterica* Serovar Typhi *ompS1* gene (outer membrane protein and pathogenicity determinant) (De la Cruz *et al.*, 2007).

We demonstrate that relieving H-NS-mediated repression of *cas* gene transcription is required for CRISPR-based immunity and that derepression is mediated by LeuO through direct binding of DNA sequences upstream of *casA*. The EMSA and footprint results (Fig. 6.3) support the finding that elevated amounts of LeuO counteract H-NS-mediated repression of *casABCDE12 in vivo*. Moreover, these data indicate that LeuO-induced activation of transcription from the *casA* promoter (P*cas*) does not simply result from a displacement of bound H-NS, since LeuO cannot facilitate the binding of RNA polymerase when H-NS is pre-bound. Instead, LeuO abrogates the cooperative spreading of H-NS upon binding to the *casA* promoter region. Interestingly, the transcript levels of the *casABCDE12* operon were higher in cells expressing *leuO* than in *hns* knockout strains (Fig. 6.1B), suggesting that either LeuO functions as an enhancer of

casABCDE12 transcription, or that derepression in K12Δ*hns* is incomplete. The latter could be due to additional repressors involved in silencing *casABCDE12*, or due to functional redundancy between suppressors of gene transcription. In particular, StpA has been reported to possess high binding affinity for P*cas* (Pul *et al.*, 2010). Although a K12Δ*stpA* strain showed similar *cas* gene transcript levels as a *wt* K12 strain (Pul *et al.*, 2010) it cannot be excluded that StpA-mediated repression of *cas* gene transcription takes place in the absence of H-NS.

Cells expressing *leuO* showed higher resistance levels compared to *hns* knockout strains, due to the higher expression of the *cas* genes and higher abundance of mature crRNA (Fig. 6.1B and Fig. 6.2BC). Compared to the CRISPR-based resistance levels to phage infection observed in *Streptococcus thermophilus* (Barrangou *et al.*, 2007) or *E. coli* BL21-AI over-expressing the *cas* genes and the CRISPR (Brouns *et al.*, 2008), the resistance levels of *wt E. coli* over-expressing *leuO* are relatively low. However, at present it is unknown whether a similar level of CRISPR-based immunity can be reached by *wt E. coli*, and if it can, under what conditions. Although we were able to show that CRISPR-based immunity is activated by overproducing LeuO, the natural growth conditions that induce CRISPR-based defense are still unknown. Our experiments show that a genomic anti-Lambda spacer alone does not provide resistance to phage infection in *wt E. coli* due to the absence of Cascade. We speculate that *leuO* expression levels under laboratory growth conditions are too low to induce derepression of P*cas*, and that phage exposure itself does not activate CRISPR-defense. Unaltered expression of *leuO* and the *cas* genes was also observed during infection with bacteriophage PRD1 (Poranen *et al.*, 2006).

Since H-NS is known to bind DNA of incoming phage or plasmid directly (Navarre *et al.*, 2006; Navarre *et al.*, 2007) this might result in redistribution of H-NS (Dillon *et al.*, 2010; Doyle *et al.*, 2007), allowing expression of the Cascade genes due to decreased local concentrations of the repressor. As *leuO* expression is negatively regulated by H-NS and positively by LeuO itself (Chen *et al.*, 2005; Hommais *et al.*, 2001), this would further amplify the activating signal for *cas* gene transcription. Interestingly, *leuO* expression levels are induced by the alarmone guanosine tetraphosphate (ppGpp) (Chen *et al.*, 2001; Fang *et al.*, 2000; Majumder *et al.*, 2001). ppGpp is involved in stress signaling cascades leading to the stringent response under nutrient limiting conditions. Since these conditions slow down phage proliferation dramatically (Schrader *et al.*, 1997), bacterial cells may then stand a better chance of surviving phage encounters, hence inducing CRISPR-based defense may be more beneficial. However, induction of the stringent response by amino-acid starvation, e.g. by serine hydroxamate (Tosa

and Pizer, 1971) neither increased the transcription from P*cas* nor the formation of mature crRNA (data not shown). Although under laboratory conditions CRISPR-based defense is suppressed, the diversity in spacer content in natural isolates of *E. coli* strongly suggests that the CRISPR/Cas system as a whole is active and functional in natural ecosystems (Diez-Villasenor *et al.*, 2010).

In an independent parallel study, it has been shown that an *E. coli hns* knockout strain containing an anti-Lambda spacer is less sensitive to phage infection (Pougach et al., in press), in agreement with the data presented here. It seems that the Δ*hns* strain containing the T3 spacer used in (Pougach et al., in press), shows higher levels of resistance than the Δ*hns* strain containing the J3 spacer that was used in this study. The T3 spacer has originally been described (Brouns *et al.*, 2008) as the spacer that confers the highest level of immunity of 8 different spacers tested. In BL21-AI over-expressing the *casABCDE* and *cas3* genes together with either the T3 CRISPR or the J3 CRISPR showed that the T3 CRISPR provides 10-fold more resistance (data not shown), indicating that the observed difference in immunity between (Pougach et al., in press) and this study is most likely resulting from a difference in the efficiency of the spacers used.

Although a number of studies involving H-NS and LeuO have been carried out in *E. coli* and *S. enterica* (Hernandez-Lucas *et al.*, 2008; Hommais *et al.*, 2001; Lucchini *et al.*, 2006; Navarre *et al.*, 2006; Shimada *et al.*, 2009), the outcome of these studies has never been interpreted in the light of CRISPR-based defense. Based on these genome-wide analyses we propose that the expression of the type E (Cse) *cas* genes from *Salmonella enterica* are likely to be regulated by H-NS and LeuO as well. For instance, in *S. enterica* Serovar Typhi transcription of *casA* (STY3070) appears to be affected by H-NS and LeuO (Hernandez-Lucas *et al.*, 2008), despite the poor conservation of the intergenic region between the divergently oriented *cas3* and *casA* genes in this strain. In *S. enterica* Serovar Typhimurium strain LT2 H-NS binding sites were found encompassing the translation start site of the *cas3* gene (Lucchini *et al.*, 2006). Another study showed that in this strain the transcription of *cas3*, *casB*, *casC* and *casD* is elevated in the absence of H-NS (Navarre *et al.*, 2006). Perhaps the *cas* genes are controlled by a single promoter in this strain, since the intergenic region between *cas3* and *casA* is only 12 nucleotides in length. Altogether, this study provides evidence that the type E CRISPR/Cas system in *E. coli* is regulated by the antagonists H-NS and LeuO, and we propose that this regulatory mechanism is conserved in *S. enterica* as well. The upcoming challenge will be to identify conditions that activate this sophisticated defense system to allow defense against invasion by foreign DNA.

Experimental procedures

Strains

The wildtype *E. coli* K12 W3110 (BW25113) strain and the *E. coli* K12 W3110 derivative Δ*hns* (JW1225) and Δ*casA* (JW2730) from the KEIO collection (Baba *et al.*, 2008), supplied by the American Type Culture Collection (ATCC), and *E. coli* K12 MC4100 (Peters *et al.*, 2003) derivative Δ*hns* (PD32) (Dersch *et al.*, 1993) were used throughout the study.

Gene cloning and recombination

A synthetic recombination cassette was designed corresponding to 400 bp flanking regions on each side of the CRISPR I locus separated by a kanamycin resistance gene flanked by FRT-sites (GAAGTTCCTATACTTTCTAGAGAATAGGAACTTC). The construct contained a *Pst*I site followed by 400 nt of the CRISPR I upstream region of the *E. coli* K12 W3110 genome (2875875-2876274), followed by a *Nco*I site, then the sequence AAACAAAGAATT, a *Kpn*I site, followed by an FRT-site, a *Sph*I site, a kanamycin resistance gene with a sequence corresponding to pJJDuet30 (2186-1276), a *Xho*I site, an FRT site, a *Not*I site, and 395 nt of the CRISPR I downstream region of the *E. coli* K12 W3110 genome (2877225-2877619) (GeneArt AG, Regensburg, Germany). A synthetic CRISPR sequence including leader sequence containing 7 spacers and 8 repeats was used (Table S6.1) (Geneart AG, Regensburg, Germany). This synthetic CRISPR was cloned between the flanking regions using the *Nco*I and *Kpn*I sites (Fig. S6.3). The *Nco*I and *EcoR*I sites in the leader and second spacer were used to exchange the first spacer sequence of the CRISPR; the constructs created were named J3 and R44 (Table S6.1). The other spacers in the CRISPR were sequences with no homology to phage Lambda. These constructs were used as recombination cassettes to replace the existing CRISPR I locus in the *E. coli* K12 W3110 genome, following a protocol described elsewhere (Datsenko and Wanner, 2000), with minor modifications. For recombination, the sequences were PCR-amplified using primers BG3017 and BG3019 (Table S6.2) with high fidelity *pfu*-turbo polymerase and transformed by electroporation into *E. coli* K12 W3110 containing pKD46, kindly provided by the ATCC. Transformants were grown at 30°C and plated on LB-Agar + kanamycin (50 μg ml^{-1}). The pKD46 plasmid has a temperature sensitive origin of replication, and was removed through growth at 37°C (Datsenko and Wanner, 2000). Recombination was validated by PCR and sequencing. The antibiotic resistance cassette was removed using Flp recombinase encoded on plasmid pCP20, and subsequent growth at 37°C, as described (Datsenko and Wanner, 2000). The Δ*leuO* mutant was constructed with the λred-gam system using oligonucleotides T209 and T210 (Table S6.2), as described

(Datsenko and Wanner, 2000). After deletion of *leuO* the resistance cassette used for selection was removed using Flp recombinase encoded on plasmid pCP20 (Datsenko and Wanner, 2000).

Plasmids and Vectors

Plasmid pWUR607 (Tet^R) contains the *casABCDE12* operon, which was PCR-amplified from *E. coli* K12 MG1655 genomic DNA using primers BG2173 and BG2174 (Table S6.2), and cloned into vector pACYC184 using the restriction sites *EcoR*I and *Nco*I. Plasmid pWUR608 (Cam^R) was constructed by cloning a *cas3* amplicon generated with primers BG2171 and BG2172 (Table S6.2) into pACYC184 using the restriction enzymes *BamH*I and *Sph*I. In the experiments where a CRISPR was introduced on a plasmid, the pACYCduet-1 vector (Cam^R) (Novagen) was used, using the *Nco*I and *Acc65*I restriction sites. pWUR477 containing the non-targeting CRISPR (N) was described previously (Brouns *et al.*, 2008). Expression of the CRISPR from this plasmid in K12 was under control of the leader sequence that contains the CRISPR I promoter (Pul *et al.*, 2010). pWUR564 is a derivative of pWUR477 that has the *Nco*I-*EcoR*I fragment (containing the leader sequence up to half of the second spacer) replaced with the *Nco*I-*EcoR*I fragment of construct J3 (Table S6.1). For expression of *wt hns* and hns^{G113D} the previously described pHOP11 and pHM52 plasmids were used, respectively (Pul *et al.*, 2010). The pCA24N plasmid from ASKA(-) clone JW0075 encodes *leuO* behind an P*T5/lac* promoter (IPTG inducible). pKEDR13, encoding *leuO* behind a P*tac* promoter (IPTG inducible), and the control vector pKESK22 were described earlier (Madhusudan *et al.*, 2005; Stratmann *et al.*, 2008). The IPTG inducible *leuO* expression plasmid pNH41 was constructed by cloning the *leuO* amplicon, generated using primers NH329 and NH330 (Table S6.2), into the 2.2 kb *Xba*I fragment of pZE12-luc, following a previously published protocol (Urban and Vogel, 2007). Plasmid pNH6 contains the *casABCDE* operon (PCR amplified with pre-phosphorylated primer NH193 and primer NH194 (Table S6.2)) inserted by blunt end and EcoRI cloning into vector pCU01 (pBAD-TOPO vector derivative), as described (Unoson and Wagner, 2008). Plasmid pUC18-IGLB was described before (Pul *et al.*, 2010).

Microarray

E. coli K12 MG1655 was transformed with plasmid pKEDR13 (Kan^R $lacI^q$ P*tac leuO*) (Stratmann *et al.*, 2008) for expression of LeuO or with control vector pKESK22 (Kan^R $lacI^q$ P*tac*). Exponential cultures were inoculated from fresh overnight cultures to an OD_{600} of 0.1 in LB supplemented with 25 µg ml^{-1} kanamycin. IPTG was added after 30 min of growth to a final concentration of 1 mM. After additional 60 minutes the bacteria were harvested using Qiagen RNAprotect and used for RNA isolation using the Qiagen

RNeasy MiniKit system. In brief, 1 ml of each culture (OD_{600} between 0.5 and 0.6) was used and processed according to the manufacturer's instructions including DNase I on-column treatment. RNA quality was assayed by denaturing urea-PAGE and by measuring the ratio of absorption at 260/280 nm in a GeneQuant II spectrophotometer (Amersham). RNA concentration was determined by measuring UV light absorption at 260 nm. The procedure was carried out four times with independent clones. Synthesis of cDNA (and cRNA) and hybridization of Affymetrix GeneChip® *E. coli* Genome 2.0 microarrays was carried out according to the manufacturer's instructions. In total, four independent RNA samples of each group (wildtype, *leuO* expressing and *leuO* deficient strains) were used. Data analysis was performed using Affymetrix Software. Fluorescence values were normalized to the GeneChip standard reference probes. Differential expression values were calculated as fold-change of *leuO* expressing samples compared to samples of LeuO-deficient control strains.

qPCR analysis of gene expression

qPCR analysis of *cas* gene transcript abundance was performed on cDNA synthesized using High Capacity Reverse Transcription Kit (Applied Biosystems) from RNA extracted by the hot-phenol method (Blomberg *et al.*, 1990) and DNase-treated using Turbo DNA-free kit (Ambion). 10 ml samples for RNA extraction were taken at OD_{600} ~0.5 from *E. coli* W3110, *E. coli* W3110 carrying pCA24N (*leuO*) and *E. coli* Δ*hns* (JW1225-2). When LeuO expression was induced, samples were taken 30 min after addition of 0.5 mM IPTG. The qPCR reactions were performed using Power SYBR green PCR master mix (Applied Biosystems) according to manufacturer's instructions, and primers were designed using Primer Express 3.0 (Applied Biosystems). For the complete list of primers used see Table S6.3. Two primer pairs were designed against *casA* as internal control. The PCR reactions were performed on a 7300 Real Time PCR System (Applied Biosystems) and analyzed using 7300 System SDS Software 1.3 (Applied Biosystems). Fold change of *cas* gene transcription was calculated using the relative quantification method with tmRNA as endogenous control and *E. coli* W3110 *cas* gene transcript abundance as calibrators. All PCR reactions were performed in six replicates. Control PCRs without template or without cDNA (produced by standard cDNA synthesis but excluding reverse transcriptase) were performed to monitor general contamination levels and genomic DNA contamination of RNA extracts, respectively.

Northern Blotting

Total RNA was extracted at the OD_{600} indicated using TRIZOL reagent (Invitrogen) according to the manufacturer's protocol. Expression of *casABCDE* from pNH6 was induced at an OD_{600} of 0.5 by adding 0.2% of arabinose for 15 min. 10 µg of total

RNA was denatured at 95°C with an equal volume of formamide loading dye, FD (90% formamide, 15 mM EDTA, 0.05% bromophenol blue and 0.05% xylene cyanol), and subsequently separated on an 8% denaturing polyacrylamide gel. A ^{32}P-labeled pUC19 DNA/*Msp*I ladder (Fermentas) was used as size marker. The RNA was electrotransferred to Nylon N+ membranes (GE Healthcare) at 10 V for 15h. Transfer was performed in a BIORAD blotting chamber in 1xTBE buffer at 4°C followed by drying of the membrane and UV-crosslinking. Prehybridization was carried out for 2–4 h at 42°C in 15 ml prehybridization buffer (5x SSC, 5x Dernhardt, 0.05 M sodium phosphate pH 6.7, 1% dextran sulphate, 0.1% SDS) together with 75 µl herring sperm DNA (20 mg ml^{-1}). Hybridization was carried out overnight at 42°C in the same buffer lacking herring sperm DNA but containing [γ-^{32}P]-ATP-labeled oligonucleotide probe NH30 (Table S6.2) specific for spacer 2 of the CRISPR1 locus. The probe was labeled with [γ-^{32}P]ATP (40 pmol DNA, 10x kinase buffer, T4 polynucleotide kinase (PNK, Ambion), [γ-^{32}P]ATP) by incubation at 37°C for 45 min. Prior to hybridization, the probe was purified over a G-50 column (GE Healthcare). Membranes were washed once for 20 min at 60°C in 2xSSC, 0.5% SDS and once for 20 min in 0.5x SSC, 0.5% SDS. Signals were quantified in a Molecular Dynamics PhosphorImager model 400S with ImageQuant software version 4.2a (Molecular Dynamics).

Electrophoretic Mobility Shift Assay

E. coli RNAP, LeuO and H-NS were purified according to published procedures (Pul *et al.*, 2010; Stratmann *et al.*, 2008). The IGLB fragment (position -1 to -414, relative to the first nucleotide of the *casA (ygcL)* start codon) was obtained by *Eco*RI/*Hinc*II or *BamH*I/*Sac*I digestion of plasmid pUC18-IGLB. Purified DNA fragments were end-labelled by Klenow (Promega) and [α-^{32}P]-dATP. Binding reactions with the indicated amounts of protein were performed in 50 mM Tris-HCl, pH 7.4, 70 mM KCl, 15 mM NaCl, 1 mM EDTA, 10 mM β-mercaptoethanol at a final heparin concentration of 20 ng µl^{-1}. Complexes were separated on native 5% (w/v) polyacrylamide gels and visualized by autoradiography as described (Pul *et al.*, 2010).

Footprint analyses

DNase I footprinting of free DNA and DNA-protein complexes was performed as described (Pul *et al.*, 2010). Formation of open RNAP-promoter complexes was analyzed by $KMnO_4$ modification of single-stranded nucleotides within the transcription bubble. 40 µl RNAP-DNA complexes were treated with 160 mM $KMnO_4$ for 2 min at 30°C. The reaction was stopped by addition of 5.3 µl β-mercaptoethanol and 5.3 µl 500 mM EDTA and the samples were ethanol precipitated after phenol/chloroform extraction. Pellets were dissolved in 10% piperidine and incubated at 90°C for 30 min.

After two rounds of washing with distilled water followed by lyophilizing, the pellets were dissolved in 50 µl distilled water and precipitated with ethanol. Cleavage products were separated on 10% denaturing polyacrylamid gels and visualized by autoradiography. The following protocol was used in footprint experiments with more than one protein: LeuO or the protein-free buffer, and H-NS or the respective buffer, were incubated with the template DNA for 5 min at 30°C. Next RNAP or the RNAP dilution buffer was added and incubated for another 10 min. Finally heparin was added to a final concentration of 200 ng μl^{-1} with a further incubation at 30°C for 5 min. An aliquot of this solution was loaded on a native gel to verify complex formation and the remaining solution was used for the different footprint experiments.

Primer extension analysis

Primer extension reactions with 25 µg total RNA hybridized to a radiolabeled specific *cas* primer oligonucleotide (5'-ATACAATTAATCTATACATATATTAAGATG-3') were performed with AMV reverse transcriptase (Promega) as described (Afflerbach *et al.*, 1998).

Phage Lambda infection studies

Host sensitivity to phage infection was tested using a virulent phage Lambda (λ_{vir}), as before (Brouns *et al.*, 2008). The host strains for infection were either *wt E. coli* K12 W3110, *E. coli* K12 W3110 Δ*hns*, or the engineered *E. coli* K12 W3110 strains (*E. coli* K12Δ*CRISPRI::CRISPRJ3* and *E. coli* K12Δ*CRISPRI::CRISPRR44*) (Table S6.4). The sensitivity of the host to infection was calculated as the efficiency of plaquing (the plaque count ratio of a strain containing an anti-Lambda CRISPR to that of the strain containing an non-targeting CRISPR) (Brouns *et al.*, 2008).

Acknowledgements

This work was financially supported by a Veni grant to S.J.J.B. (863.08.014) and a Vici grant to J.v.d.O. (865.05.001) from the Dutch Organization for Scientific Research (Nederlandse Organisatie voor Wetenschappelijk Onderzoek). E.R.W. was financially supported by Spinoza resources awarded to Willem M. de Vos. Ü.P. was supported by the Deutsche Forschungsgemeinschaft PU435/1-1. N.H. was supported by a postdoc fellowship from the Swedish Research Council (Vetenskapsrådet). M.L. was supported by the Wenner-Gren Foundations. T.S. and K.S. were supported by the Deutsche Forschungsgemeinschaft Schn 371 / 10-1. A.R. and E.G.H.W. were supported by the Swedish Research Council.

Supplementary tables can be found at http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2958.2010.07315.x/suppinfo

Supplementary figures

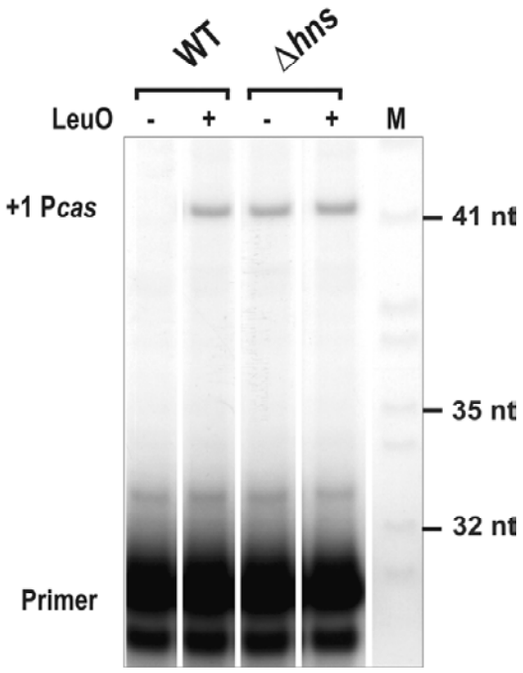


Figure S6.1. Antagonistic regulation of transcription from the casA promoter (Pcas) by H-NS and LeuO. The transcriptional activity of the *casA* promoter (P*cas*) was tested by primer extension analysis using 25 µg total RNA isolated from *wt* or Δ*hns* cells transformed with the control vector pKESK22 or the LeuO expression plasmid pKEDR13. RNA was isolated 2h after IPTG induction. In lane 5 a G+A-sequencing ladder was separated as size marker. The position of the resulting cDNA product and the primer are indicated on the left.

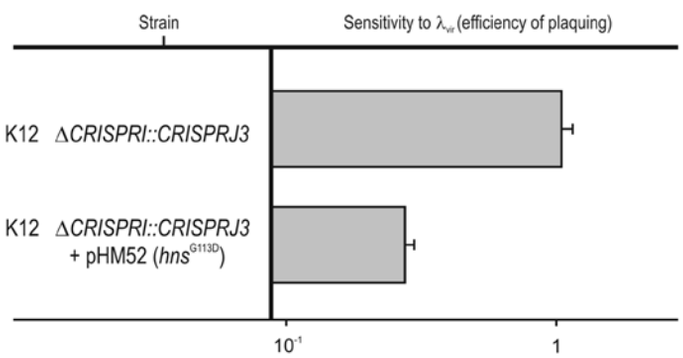


Figure S6.2. Expression of the dominant negative *hns* mutant G113D causes reduced sensitivity of *E. coli* K12Δ*CRISPRI::CRISPRJ3* to phage λ infection as measured by plaque assay. The efficiency of plaquing (EOP) is expressed as a ratio of plaque number with *E. coli* K12Δ*CRISPRI::CRISPRR44* (non-targeting CRISPR), as described before (Brouns et al., 2008). For these experiments bacteria were grown in 2YT medium.

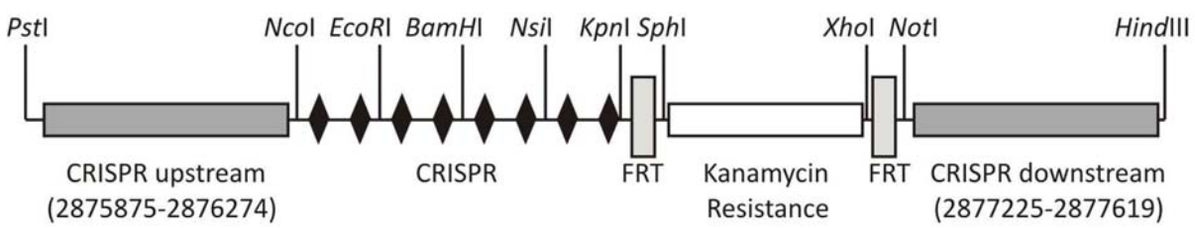


Figure S6.3. Recombination cassette used to generate strains *E. coli* K12Δ*CRISPRI::CRISPRJ3* and *E. coli* K12Δ*CRISPRI::CRISPRR44*, that differ in the first spacer of the CRISPR.

Chapter 7

Summary and general discussion

Summary

This thesis focuses on the antiviral and antiplasmid CRISPR/Cas (clusters of regularly interspaced short palindromic repeats/CRISPR associated) system of *Escherichia coli* K12, which belongs to the Cse-subtype. In total eight different subtypes have been identified in prokaryotes, four of which have been studied experimentally, as described in **Chapter 1.** In general, CRISPR-mediated defense comprises three distinct stages and shows clear analogies with RNAi in eukaryotes as described in **Chapter 1** (Fig. 7.1). During the first stage (adaptation) DNA fragments of invading viruses or plasmids are integrated as spacers into a CRISPR. During the second stage (expression), which is described in **Chapter 2**, CRISPR loci are transcribed into long precursor CRISPR RNA (pre-crRNA) which is subsequently processed by the metal-independent nuclease CasE. CasE cleaves within each repeat, generating mature crRNAs that all contain an eight nucleotide 5' handle of the repeat, a spacer and a part of the next repeat which includes the stemloop. CasE and the crRNA were shown to be members of the CRISPR-associated complex for antiviral defense (Cascade), which also contains the protein components CasA, B, C and D. In conjunction with Cas3, Cascade loaded with crRNA successfully inhibits phage lambda proliferation as was shown by plaque assays using a synthetic CRISPR complementary to the phage nucleic acids. Both crRNAs containing spacers complementary to the coding and the non-coding strand could inhibit phage proliferation suggesting that DNA is being targeted. A CasE mutant deficient in pre-crRNA processing abolishes immunity, showing that the formation of mature crRNAs is essential. The mechanism of recognition, which is at the basis of the third stage (interference), is described in **Chapter 3.** Cascade has the capacity to bind dsDNA. The spacer sequence of the crRNA can basepair with its target through ATP independent strand displacement, a process which is enhanced by CasA. This chapter also provides a more detailed mass and sequence analysis of crRNA which revealed that crRNAs are 61 nucleotides in length, corresponding to one spacer and one repeat, and contain a 2', 3' cyclic phosphate group. ESI-MS analysis of Cascade revealed that it has an unusual stoichiometry of $CasA_1B_2C_6D_1E_1$-$crRNA_{1.}$ EM and SAXS analysis showed that it adopts a seahorse-like shape in which the backbone is formed by 6 CasC subunits. Comparison with the stable subcomplexes $CasB_2C_6D_1E_1$-$crRNA_1$ and $CasC_6D_1E_1$-$crRNA_1$ revealed the location of the CasA and two CasB subunits. It was also shown that all the subunits and Cas3 are essential for immunity. **Chapter 4** describes how Cas proteins and an anti-lambda CRISPR could successfully inhibit transformation of a plasmid that contains a lambda fragment including a protospacer. A mutant library was created by error-prone PCR and transformed to cells resistant against the unmodified plasmid. Escape mutants were selected and the transformed

plasmid was sequenced. The sequence data of the escape mutants revealed three types of mutants. These mutants contained mutations in the protospacer and in the protospacer adjacent motif (PAM) as well as deletions of the entire protospacer region. The PAM might be a criterion to verify that only invading DNA is being targeted. Next to mutations in the PAM, mutations in the side of the protospacer that directly flanks the PAM can lead to escape. These data show that the PAM and perfect base pairing at the PAM-side of the protospacer are essential for immunity.

Since lambda phages enter their *E. coli* host at the cellular poles and rapidly replicate, we hypothesized that the antiviral Cas protein machinery might be co-located. In order to test this hypothesis we used photo-activatable localization microscopy (PALM) to determine the location of the Cas proteins at nanoscale resolution (**Chapter 5**). Indeed, at least some of the Cas proteins are located at the cellular pole, which might be advantageous to initiate a swift antiviral response. Additionally, it may yet be another mechanism of the host to prevent autoimmunity since the chromosome and thus CRISPRs are spatially separated from the actual antiviral machinery. **Chapter 6** finally focuses on transcriptional regulation of the system. H-NS was identified as a transcriptional repressor. The expression of most *cas* genes (with the exception of *cas3*), as well as the CRISPR were elevated in an H-NS knockout of *E. coli*. LeuO was identified as a transcriptional activator that functions as an antagonist of H-NS. Overexpression of LeuO in a wt *E. coli* background resulted in elevated *casABCDE12* transcription and crRNA levels. Plaque assays finally demonstrated that H-NS indeed represses the CRISPR/Cas system, contrary to LeuO that activates CRISPR/Cas mediated defense.

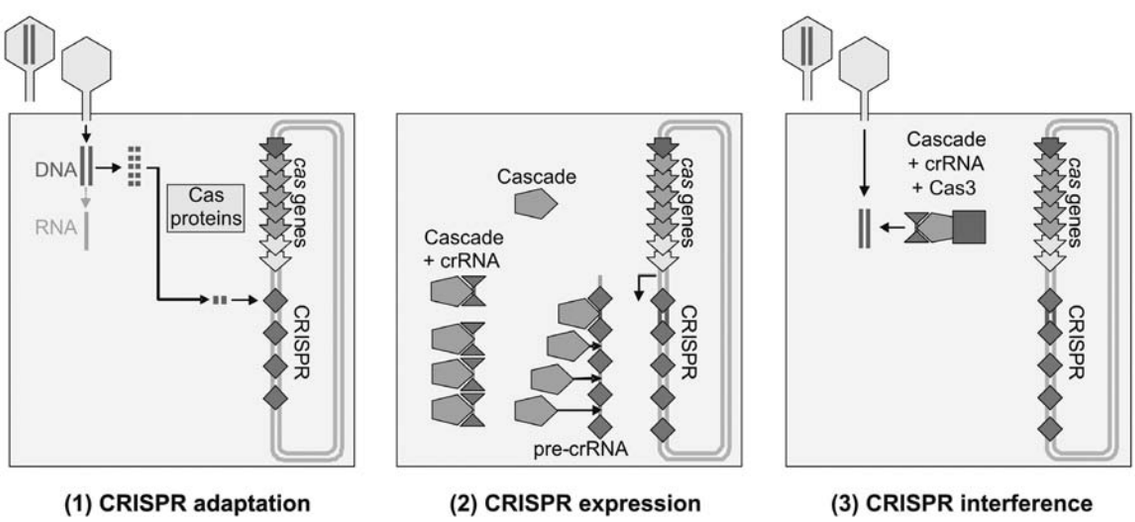


Figure 7.1. Three different stages in CRISPR-mediated antiviral defense. For full colour version see page 133. The different stages are explained in more detail in Chapter 1. Picture adapted from (van der Oost et al. 2009).

General discussion

Molecular mechanism

Spacer integration

Although the rapidly expanding research field on CRISPR/Cas results in a steady gain of insight, many questions about this fascinating defense system still remain to be answered. One poorly understood aspect is the mechanism of spacer integration during the adaptation stage. Spacer acquisition in the laboratory has only been observed in the lactic acid bacteria *Streptococcus thermophilus* (Barrangou et al. 2007; Deveau et al. 2008; Horvath et al. 2008) and *Streptococcus mutans* (van der Ploeg 2009). An interesting finding is that in most cases spacers were acquired at the leader side of the CRISPR. The CRISPR thus provides a chronological record of previously encountered phages and/or plasmids. This idea is supported by bioinformatics analyses that show that the variation between different similar CRISPRs mainly occurs at the leader proximal side (Lillestøl et al. 2006; Diez-Villasenor et al. 2010). A second interesting finding is that the protospacer fragment carried a conserved protospacer adjacent motif (PAM), which is essential for immunity during the interference stage (Chapter 4) (Deveau et al. 2008). If the PAM would only be important at this stage, as shown in Chapter 4, one would expect that when multiple spacers are simultaneously incorporated in the CRISPR locus only one spacer would match a protospacer with a PAM (the other spacer thus being dispensable). However all spacers, including those that were simultaneously incorporated, corresponded to a fragment carrying a PAM, albeit sometimes imperfect (Deveau et al. 2008). This suggests that the PAM is not only important during interference but might also determine whether fragments are selected for integration. A PAM sequence is not unique for invading DNA. It is also present in the host genome and therefore most likely not the only selection marker for integration. Moreover, careful analysis of spacer sequences from a large database revealed that protospacers can be found in bacteriophages that are distinct in infection characteristics and genome features (Mojica et al. 2009).

How do cells recognize invading DNA and incorporate this as spacer DNA into the CRISPR locus? Integration of spacers could be a random and rare event; the mutants that have acquired spacer-directed immunity then become dominant in the population by natural selection; Most cells that do not acquire spacers will most likely die. After recognition of the invading DNA, either specifically or randomly, it is most likely being cleaved into smaller fragments prior to integration. A prime candidate to be involved in this process is the Cas1 protein, since it is very well conserved among different

subtypes (Jansen et al. 2002; Haft et al. 2005; Makarova et al. 2006; Marraffini and Sontheimer 2009), it has non-specific nuclease activity (Wiedenheft et al. 2009), and it is not essential for crRNA processing or target interference (Brouns et al. 2008). The Cas1 endonuclease generates DNA fragments of ~80 basepairs (Wiedenheft et al. 2009), which should be further trimmed before they are incorporated. How these fragments are then integrated and why there is a preference for the leader proximal side of the CRISPR remains unknown, and should be investigated experimentally. Most likely motifs in the leader sequence play a role in this process. It is noteworthy that Cas2 has also been hypothesized to be involved in the process of spacer integration (van der Oost et al. 2009) since it is neither essential for crRNA processing nor target interference (Brouns et al. 2008) and is closely associated with Cas1 (Haft et al. 2005; Makarova et al. 2006). Cas2 has endonuclease activity on U-rich regions in ssRNA (Beloglazova et al. 2008). How this would fit in the model of spacer integration is a mystery. Cas2 could be involved in the primary cleavage of phage mRNA which might subsequently be converted into DNA through a reverse transcription step, prior to integration. However, this seems to be an elaborate way for the host to integrate DNA fragments from a DNA phage. Alternatively, Cas2 might process RNA derived from RNA phages, prior to reverse transcription and subsequent spacer integration. However, no spacers against RNA elements have been observed till now. To study the mechanism of spacer integration in *E. coli* K12 we have tried to obtain mutants that had acquired immunity to phage lambda or phage T3. Different approaches resulted in many resistant mutants, but none of them had integrated a spacer in the CRISPR locus against the phage (Brouns and Jore, *unpublished*). A possible explanation for this is the high background of other mutations that lead to resistance. One out of several strategies for *E. coli* to escape infection (reviewed in Labrie2010) is by mutating essential genes needed by the phage, such as phage receptors; phage lambda binds to the maltoporin receptor while phage T3 attaches to the lipopolysaccharides (LPS). Indeed, *E. coli* can mutate the maltoporin directly (Hofnung et al. 1976) and the LPS indirectly by mutations in genes involved in the LPS biosynthesis pathway (Qimron et al. 2006). An alternative explanation could be the low transcription levels of *cas* genes and CRISPRs due to transcriptional silencing by H-NS (Chapter 6) (Pul et al. 2010). To obtain higher rates of spacer integration other *E. coli* strains carrying the same Cse-subtype should be subjected to phage infections. Perhaps these strains have a more active CRISPR/Cas system. However, lactic acid bacteria such as *S. thermophilus* seem to be the ideal organisms to study the mechanism of spacer integration, since spacer integration has been observed in laboratory experiments.

CRISPR expression and processing

CRISPR-mediated defense is based on small CRISPR RNAs (crRNAs) . crRNAs have been detected in *Archaeoglobus fulgidus* (Tang et al. 2002), *Sulfolobus solfataricus* (Tang et al. 2005), *Sulfolobus acidocaldarius* (Lillestøl et al. 2006; Lillestol et al. 2009), *E. coli* (Brouns et al. 2008), *Pyrococcus furiosus* (Hale et al. 2008), *Xanthomonas oryzae* (Semenova et al. 2009) and *Staphylococcus epidermidis* (Marraffini and Sontheimer 2008). Transcription has been shown to be unidirectional from the AT-rich leader sequence (Jansen et al. 2002), *Sulfolobus acidocaldarius*. being an exception to this rule (Lillestøl et al. 2006; Lillestol et al. 2009). The biogenesis of mature crRNAs have only been analyzed in *E. coli* (Brouns et al. 2008) (Chapter 3) and *P. furiosus* (Carte et al. 2008; Hale et al. 2008). In the *E. coli* Cse-system the mature crRNAs are a product of a single endonucleolytic cleavage event by CasE, after which the crRNA is retained by Cascade. The 3' end of the crRNA carries a 2', 3' cyclic phosphate. This finding in Chapter 3 explains why we had previously cloned and sequenced a pool of crRNAs with a well defined 5' end and less well defined 3' ends (Chapter 2). During the cloning procedure we used poly(A) polymerase to add an adapter to the crRNA, but poly(A) polymerase cannot elongate a 2', 3' cyclic phosphate end. Most likely partly degenerated crRNAs were cloned instead. Mature crRNAs in *E. coli* differ in their 3' end from the ones that can be found *P. furiosus* (Hale et al. 2008). In *P. furiosus*, pre-crRNA is cleaved by Cas6, which is structurally similar to CasE, and subsequently remains bound to Cas6 as shown by *in vitro* assays. How these crRNAs are further processed at the 3' end to yield mature crRNAs of 39 and 45 nucleotides, and how these are transferred to the Cmr-complex is unknown. Perhaps one of the subunits of the Cmr-complex is responsible for further processing to mature crRNA. The anion exchange chromatography analysis of ribonucleoprotein (RNP) complexes comprising mature crRNAs by Hale and coworkers showed at least two distinct peaks (Hale et al. 2008). One peak contains dominant crRNAs of 39 and 45 nucleotides. The other peak contains crRNAs of 45 nucleotides and a slightly longer species. The complex containing the mature crRNAs of 39 and 45 nucleotides is the Cmr-complex, but the composition of the other complex has not yet been reported. Perhaps this complex is another variant of the Cmr-complex containing proteins that are involved in further processing of the intermediate crRNAs. Another interesting question concerns the role of the Cst proteins that are encoded in the same CRISPR/Cas locus. Do they form a complex containing mature crRNA products, initially cleaved by Cas6, analogous to the Cmr-complex? Alternatively, are the Cst proteins involved in generation of mature crRNA in the Cmr-complex? Further trimming of the 3' end has not only been observed in *P. furiosus*, but also in *S. epidermidis* (Marraffini and Sontheimer 2010b).

Despite the differences between crRNAs of the Cmr- and Cse-systems, analogies are also present; crRNAs are part of large RNPs and the crRNAs have a conserved eight nucleotide 5' handle, the latter has also been observed in *S. epidermidis* (Marraffini and Sontheimer 2008).

Target recognition and degradation

Some first light has been shed on how these mature crRNAs guide the Cas protein machinery to their complementary target in *E. coli* (Chapter 3) and *P. furiosus* (Hale et al. 2009). In *E. coli*, crRNA guides Cascade to complementary dsDNA and can base pair by strand displacement. The CasA subunit has been proposed to enhance target localization of Cascade, CasC forms the backbone of the complex and CasE cleaves the pre-crRNA. The role of CasB and CasD is currently unknown. The X-ray crystal structure of CasB revealed conserved basic patches on one side of its surface, suggesting that it might be involved in RNA or DNA binding (Agari et al. 2008). CasD belongs to the diverse Repeat-Associated Mysterious Protein (RAMP) superfamily (CasE also belongs to this family); members of this family are hypothesized to be involved in RNA binding (Makarova et al. 2006). CasD and CasE might each bind to one of the crRNA handles which are the common part of the crRNAs. More details of the function of each Cascade subunit are expected to be revealed in the near future. Inhibition of phage proliferation is most likely not achieved through just binding of the DNA, since overproduction of Cascade and crRNA are not sufficient for resistance *in vivo* (Chapter 2). In addition the Cas3 protein is needed for interference. Cas3 is composed of an HD-domain with putative nuclease function and a DEAD-box helicase domain. It is very well possible that the HD-domain is important for cleavage of the target, thus neutralizing the foreign DNA. Indeed the truncated HD-protein (SSO2001) from *S. solfataricus* has been shown to possess nuclease activity *in vitro* (Han and Krauss 2009). SSO2001 preferably cleaves GC rich regions. The repeats associated with this *cas* gene cluster of the Csa-subtype belong to the CRISPR-7 type (Kunin et al. 2007). This type is associated with a PAM sequence of NGG (Mojica et al. 2009). One cannot rule out the possibility that the G-nucleotides could be a recognition sequence for the HD nuclease. However, it is more likely that the activity and specificity of the HD nuclease (or other nuclease cleaving the target DNA) is determined by a Cascade-like RNP complex. Whether the PAM sequence carries the scissile phosphoester bond is not known, but the PAM is at least essential for interference (Chapter 4).

In contrast to DNA being targeted by the Cse-subtype in *E. coli* and Csm-subtype in *S. epidermidis* (Brouns et al. 2008; Marraffini and Sontheimer 2008), the Cmr-subtype of *P. furiosus* has been demonstrated to degrade complementary RNA. The

Cmr-complex cleaves complementary RNA following a molecular ruler mechanism, which is convincingly shown *in vitro* (Hale et al. 2009). An obvious question concerns which subunit has the nuclease activity. This could be easily determined by site-directed mutagenesis of conserved motifs in the subunits of the Cmr-complex. A good candidate is the Cmr2 protein which contains an HD domain that is fused to a predicted polymerase domain. Moreover, the presence of this potential polymerase in the Cmr-complex is interesting. What would be its role? Polymerases are present in three subtypes (van der Oost et al. 2009). Makarova and coworkers predicted that it might use the crRNA as a primer to amplify the targeted RNA molecule (Makarova et al. 2006). This dsRNA could subsequently be cleaved by other Cas proteins in the complex or RNaseIII domain containing proteins. The observation that the crRNAs are lacking a 3' handle that would inhibit polymerase activity is well in line with this proposed model. RNA cleavage has been shown *in vitro* but what is the biological function? Is the Cmr-complex targeting RNA from DNA viruses, which seems less efficient then targeting the DNA directly? Alternatively, are there RNA viruses out there that can be targeted? Unfortunately no infective viruses for *P. furiosus* have been described so far to address these questions. Moreover, thermophilic RNA viruses have not been identified so far. An alternative target for the Cmr complex might be mRNA from *P. furiosus* itself, which would come down to a role for the Cmr-subtype in gene regulation.

The Cmr-subtype has thus been demonstrated to target RNA (Hale et al. 2009), while the Csm-subtype targets DNA (Marraffini and Sontheimer 2008). Interestingly, the composition of the Cmr- and Csm-subtype are similar, each containing a putative polymerase and four or three RAMP proteins respectively (Haft et al. 2005). It will be interesting to study and compare the differences between these two subtypes. In general one can say that, although common themes between the CRISPR/Cas subtypes are present, the types substantially differ in composition and mechanism of action, reflecting the different targets they (might) have.

An ongoing battle between bacteria and phages

One of the most appealing aspects of phage research is the real-time monitoring of the ongoing battle between phages and bacteria. Phages can develop different pathways to bypass the defense systems of their host, to which the host respond inactivating these pathways resulting in resistance to this phage. This is beautifully illustrated by the ongoing war between *E. coli* and phage T4; *E. coli* adopted a restriction-modification system, the phage responded by incorporation of hydroxymethylcytosine (HMC) into its genome to bypass restriction, *E. coli* acquired systems that can specifically cleave HMC-containing DNA and so on (reviewed in (Labrie et al. 2010)). The ongoing war at

the CRISPR level has been demonstrated using metagenomic data. Viruses in a biofilm population undergo extensive gene shuffling to evade CRISPR-based immunity. Only the most recently acquired spacers matched viral genome sequences (Andersson and Banfield 2008). Another study of a metagenomic data set revealed silent mutations in viral sequences within protospacer elements, again illustrating the virus respons to escape the CRISPR-system (Heidelberg et al. 2009). Mutations in phages to bypass CRISPR/Cas immunity have also been observed in laboratory experiments. A first study revealed that the *Streptococcus* phages 858 and 2972 escaped CRISPR-mediated acquired immunity by mutating either the protospacer or the PAM sequence (Deveau et al. 2008). A mutation in the PAM of phage Xop411 might also explain the sensitivity of *Xantomonas oryzae* strain Xo21 to this phage, while carrying a perfect protospacer (Semenova et al. 2009). An analogous observation has been done in Chapter 4 where we describe the transformation of a plasmid to *E. coli* BL21-AI expressing the appropriate *cas* genes and a CRISPR targeting the plasmid. The colonies that are formed contain the plasmids that have recombined and contain major deletions, including the protospacer. Apparently the selection force for escape mutants is high enough to obtain these spontaneous mutants. A possible other mutation event to bypass CRISPR immunity can be found in Chapter 3. For designing a CRISPR with 7 identical spacers we selected a wildtype spacer from *E. coli* R44 that gave a perfect hit with the phage P7 genome (Mojica et al. 2005). The protospacer is flanked at the 5' end by ACG. Mojica and coworkers proposed AWG (alternatively CWT on the complementary strand) as the PAM for the Cse-subtype (Mojica et al. 2009). When we overexpressed the R44 CRISPR together with the required *cas* genes we did not observe any resistance towards phage P7, compared to strain carrying a different non-targeting CRISPR. This could be explained by the imperfect PAM sequence. Indeed, the transformation experiments described in Chapter 4, where we constructed a library of plasmids with single mutations, revealed that a mutation in the PAM from ATG to ACG caused a successful escape of a plasmid. We then analyzed the Upf62.1 gene of phage P7 which contains the protospacer. Blastp analysis of the Upf62.1 protein revealed two homologous proteins, UpfA from phage P1 and a hypothetical protein from a P1 prophage in *E. coli* O111:H-str. 11128. An alignment of these three 130 amino acid-long proteins showed in total 6 positions that were not completely conserved. One of these amino acids coincides with the PAM position. Indeed in the P1 and P7 phage this position is ACG, in contrast to an ATG sequence in the *E. coli* strain, the latter being a perfect PAM. UpfA and Upf62.1 have an arginine at this position while the hypothetical *E. coli* protein has an histidine at this position. It is tempting to speculate that the P1 and P7 phages have mutated the middle nucleotide of the PAM sequence to bypass CRISPR-mediated defence in *E. coli* R44 or a non-sequenced strain carrying

the same spacer. The *E. coli* O111:H-str. 11128 does not carry a CRISPR that contains this spacer, but it would be interesting to see if a CRISPR could target the protospacer and thus its own genome. The possible spatial separation of the Cas protein machinery and the chromosome may prevent this autoimmune reaction (Chapter 5).

Alternative roles for CRISPR/Cas?

Besides antiviral and antiplasmid defense other functions have been assigned to CRISPRs and *cas* genes. First, in 1995 Mojica and coworkers tested the effect of transformation of a plasmid containing a CRISPR (which they named TREP) to the archaeum *Haloferax volcanii* (Mojica et al. 1995). They observed diminished growth, lowered viability, and impaired genome segregation during cell division. The authors hypothesized that the CRISPRs were involved in replicon partitioning. Interestingly, we also observed an approximately two-fold slower growth when any CRISPR containing plasmid was transformed to *E. coli*, confirming the previously observed results. Perhaps replication of chromosomal CRISPR DNA is a delicate process which is highly affected by abundant non-chromosomal CRISPR elements. A second function was proposed by Makarova and coworkers. Before *cas* genes were found to be associated with CRISPRs, they stumbled on these regions of conserved gene contexts (Makarova et al. 2002). Bioinformatical analysis assigned putative nuclease, polymerase and helicase functions to these genes. These functional features made the authors propose that these genes are part of a previously undetected DNA repair system, which is largely specific to thermophilic prokaryotes. This hypothesis is supported by observed upregulation of one operon of *cas* genes in *P. furiosus* after exposure to ionizing gamma radiation (Williams et al. 2007). A third putative function was proposed when an infection study in *Pseudomonas aeroginosa* revealed that strains carrying a lysogenized DMS3 phage lost the ability to swarm and form biofilms (Zegans et al. 2009). This inhibition requires the genomic CRISPRs, suggesting that CRISPRs might play a role in biofilm formation. Finally, another study revealed that the *cas* genes in *Myxococcus xanthus* are cotranscribed with *dev* genes which are important for the development of spores in fruiting bodies (Viswanathan et al. 2007). Since the above proposed functions of CRISPR/Cas are yet poorly understood, more research will be needed to unravel the underlying mechanisms.

Applications

The adaptive and heritable antiviral and antiplasmid CRISPR/Cas system described in this thesis has many (potential) applications in research (*e.g.,* silencing of genes in prokaryotes), in medical diagnostics (*e.g.,* strain typing by comparison of CRISPR spacer sequences) and industry (*e.g.,* development of phage resistant strains) (Sorek

et al. 2008). Some of these applications will be described in more detail below.

Gene silencing

The Cmr-complex has been shown to target RNA *in vitro* (Hale et al. 2009). The Cmr-subtype or potentially other subtypes with RNA degrading activity that have not been characterized up till now might be a potential tool for gene silencing. Gene silencing in prokaryotes can be achieved through development of genomic disruptions of your gene of interest. However, this is labor intensive and for many prokaryotes a genetic system is lacking. An alternative strategy involves the antisense (as) RNA technology (reviewed in (Rasmussen et al. 2007)), which has been successfully applied in a few organisms, among which *E. coli* (Alessandra et al. 2008). However the use of asRNA is still limited due to the lack of a robust design strategy. Moreover, one asRNA is needed per mRNA. Introducing a Cmr-like system including a CRISPR targeting one or multiple mRNAs might thus provide a good alternative strategy. The recycling of the Cmr-like complex after target cleavage might enhance the silencing efficiency, a feature that is lacking in the antisense technology. An extra advantage of the use of CRISPRs for gene silencing is that multiple genes can be simultaneously silenced with one CRISPR construct, while asRNA technology requires one construct for each target mRNA.

No CRISPR/Cas has been demonstrated in eukaryotic genomes. However, an interesting question is whether the CRISPR/Cas system would be functional in eukaryotes either, at the RNA or DNA level. Gene silencing in eukaryotes is frequently performed by short interfering (si) RNA technology (reviewed in (Castanotto and Rossi 2009)). Implementing sequence specific silencing of invading RNA elements using CRISPR/Cas might provide a useful alternative. Moreover, silencing of DNA elements might open new opportunities. Achieving this goal depends on whether host factors are involved in CRISPR-mediated silencing in prokaryotes. The proposed model for transfer of CRISPR/Cas cassettes through horizontal gene transfer (Haft et al. 2005; Godde and Bickerton 2006; Makarova et al. 2006; Tyson and Banfield 2008; Chakraborty et al. 2009; Horvath et al. 2009) supports the idea that either host factors are not essential, or that only those proteins that are very well conserved among prokaryotes are involved. Cleavage of RNA by the Cmr-complex has been reconstituted *in vitro* (Hale et al. 2009), but it is not known whether non-Cas proteins are needed for crRNA maturation.

Strain differentiation

The CRISPR locus of *E. coli* was the first to be sequenced (Ishino et al. 1987). A few years later it became clear that CRISPR loci can vary between strains. Researchers

found that the CRISPR loci from 14 different *Mycobacterium tuberculosis* strains were polymorphic in composition and length (Hermans et al. 1991). The same research group subsequently developed a fast molecular method for both detection and strain differentiation of *M. tuberculosis*. The method combined PCR amplification of CRISPR loci and subsequent hybridization with probes containing different spacers and this new form of genotyping was named 'spoligotyping' (derived from spacer oligotyping) (Kamerbeek et al. 1997). This technique can be used for diagnostic and epidemiological studies and is widely used (Driscoll 2009). Direct sequencing of CRISPRs has been used for strain typing in other organisms, such as *Campylobacter jejuni* (Schouls et al. 2003) and *Thermotoga neapolitana* (DeBoy et al. 2006).

Industrial processes

The first study that observed CRISPR-based acquired immunity in *S. thermophilus* (Barrangou et al. 2007) indirectly demonstrated the potential applications for industry. Phages are a substantial problem in industry that relies on bacteria, such as fermentations in the dairy industry. The phages rapidly multiply in a bacterial culture, which can lead to a dramatic decrease or complete loss of desired products in a fermentation batch. To prevent phage infection, several techniques have been developed and used, such as superinfection exclusion, restriction-modification systems and adsorption interference (Sturino and Klaenhammer 2006). An alternative strategy might be the development of CRISPR-mediated resistant strains. This can either be achieved by selection of CRISPR mutants from cultures that are challenged with your phage of interest or by designing artificial CRISPRs against this phage (and possibly others). The CRISPR approach might be advantageous because multiple spacers can be incorporated against one or multiple phages. Having more spacers against one phage significantly decreases the chance of the evolution of escape mutants, since multiple simultaneous mutations are needed to bypass the CRISPR/Cas system. Knowledge about the functioning of CRISPR/Cas is a requirement, since elements like the PAM should also be taken into account. Researchers at Danisco filed a patent on CRISPRs as a tool to develop immune strains (Horvath et al. 2007).

Appendices

Colour figures

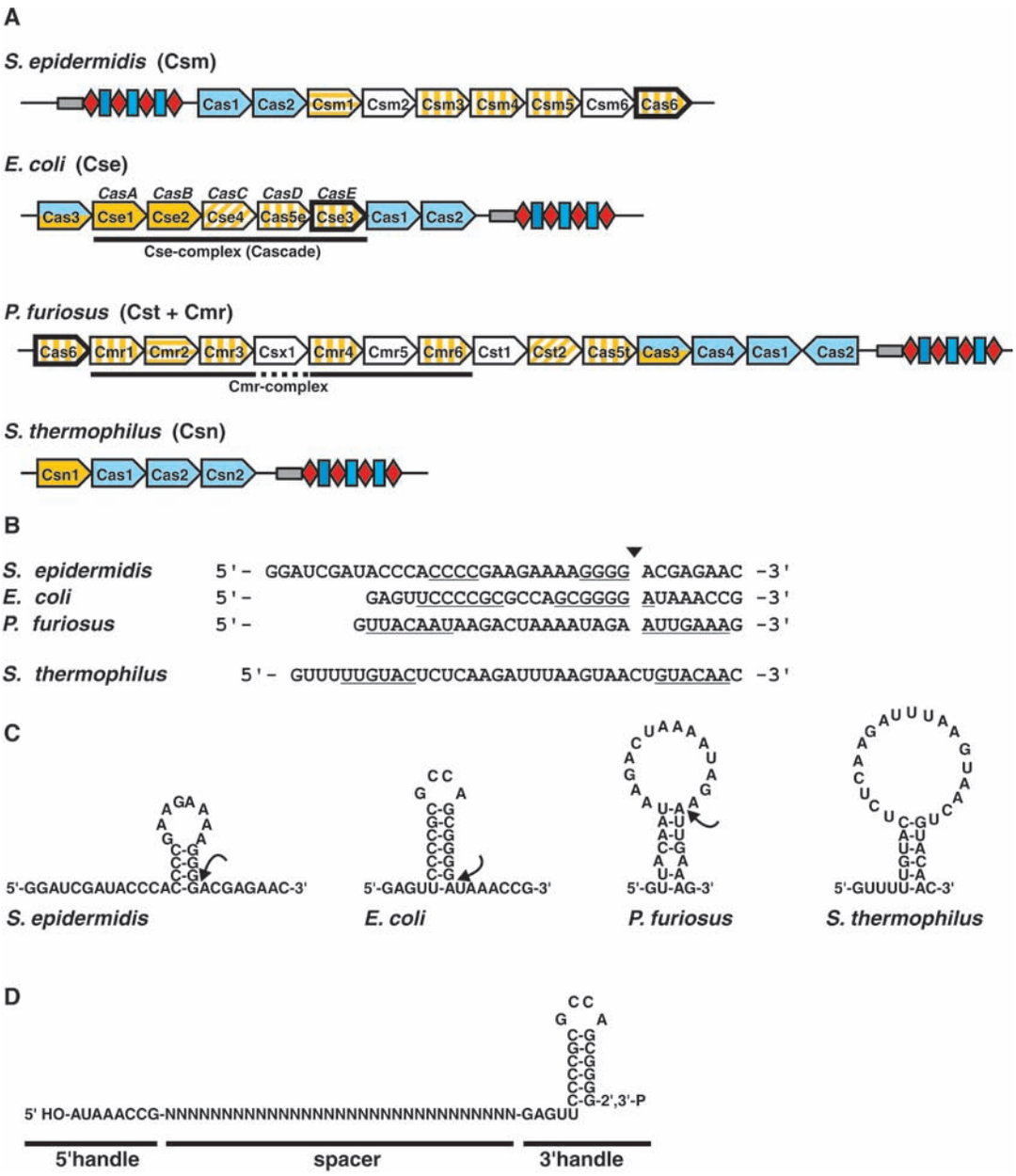


Figure 1.2. Overview of the four CRISPR/Cas subtypes that are described in this chapter. For an overview of all eight CRISPR/Cas subtypes, see (Haft et al. 2005; van der Oost and Brouns 2009). (A) *cas* gene neighborhoods in 4 experimentally studied organisms, each representing a different subtype indicated between brackets. CRISPRs consist of a leader (grey box), repeats (red diamonds) and spacers (blue boxes); only a fragment of the CRISPR is shown. Genes are indicated as arrows. Blue arrows indicate genes that are (possibly) involved in spacer acquisition. Yellow arrows indicate genes that are involved in CRISPR transcription and processing and target interference. The endonucleases that cleave pre-crRNA generating crRNA are highlighted as bold arrows. Hatching patterns indicate gene similarity: RAMP genes have vertical lines, polymerase genes have horizontal lines, CasC homologues have diagonal lines, and other genes that are not related to each other are filled. Genes that encode proteins from isolated complexes (Cse-complex from *E. coli* and Cmr-complex from *P. furiosus*) are underlined. (B) CRISPR RNA repeat sequences from each organism are given. The cleavage site is indicated by a triangle. Although the repeat sequences are different, all CRISPR RNA cleavage events generate an 8 nucleotide 5' handle. Please note that the cleavage site in *S. thermophilus* CRISPR RNA has not been determined. Palindromic sequences are underlined. (C) Predicted secondary structures of the different CRISPR RNA repeats. Cleavage sites are indicated with an arrow. As described previously by Kunin et al., the repeat of *P. furiosus* is not likely to form a stem loop (Kunin et al. 2007). (D) Mature crRNA from *E. coli* K12.

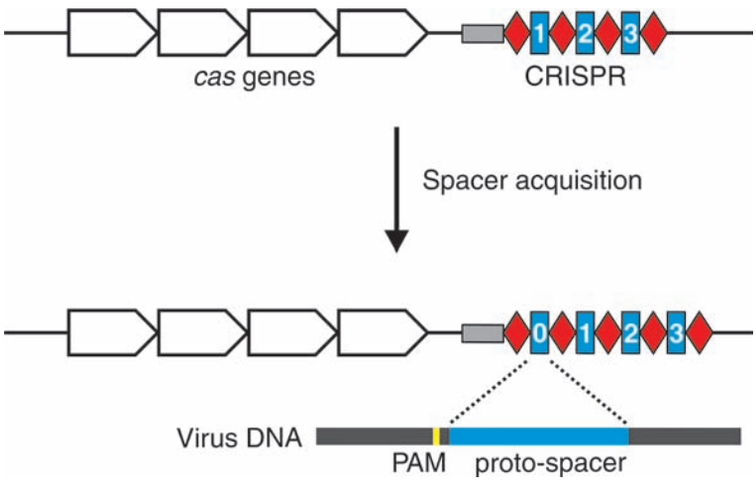


Figure 1.3. Integration of a new spacer. A new spacer is acquired at the leader proximal side of the CRISPR during virus infection, resulting in resistance. The CRISPR consist of a leader (grey box), repeats (red diamonds) and spacers (blue boxes). The newly acquired spacer is numbered 0 and matches the sequence of the virus (proto-spacer). The protospacer adjacent motif (PAM) is located upstream or downstream the protospacer.

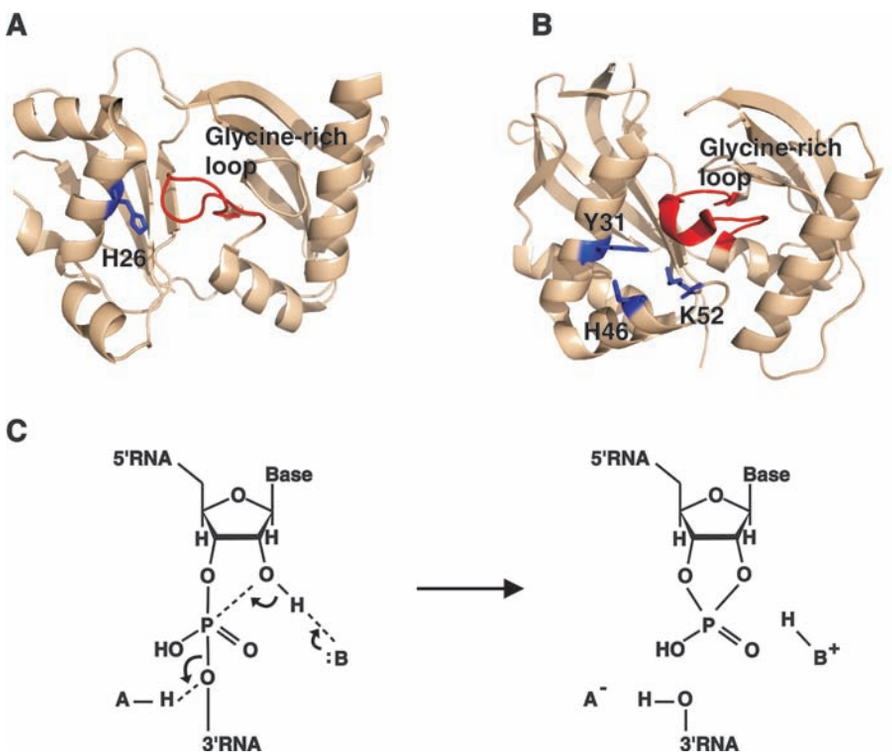


Figure 1.4. The catalytic sites of CasE and Cas6, and the proposed reaction mechanism of pre-crRNA cleavage. (A) Proposed catalytic site of CasE from *T. thermophilus* showing the conserved histidine residue (H26) and the glycine-rich C-terminal loop. The catalytic site of Cas6 from *P. furiosus* (B) contains a catalytic triad of tyrosine (Y31), histidine (H46) and lysine (K52) and a glycine-rich C-terminal loop. The loop and the overall duplicated ferredoxin fold are conserved among CasE and Cas6. Pre-crRNA cleavage might follow a general acid-base hydrolysis mechanism (D). A base (B) draws a proton from the 2'OH of the ribose ring. A subsequent nucleophilic attack on the phosphorus atom is simultaneously compensated by the acid (A) that donates a proton to the leaving 3'RNA. The tyrosine residue of Cas6 is proposed to be the base and the histidine the acid residue (Carte et al. 2008a). In CasE the histidine and a water molecule might be the catalytic residues. Pictures in (A) and (B) are generated with pymol (www.pymol.org), potential catalytic residues are depicted in blue; the glycine-rich loop is depicted in red. Coordinates were obtained from the Protein Data Bank (www.pdb.org).

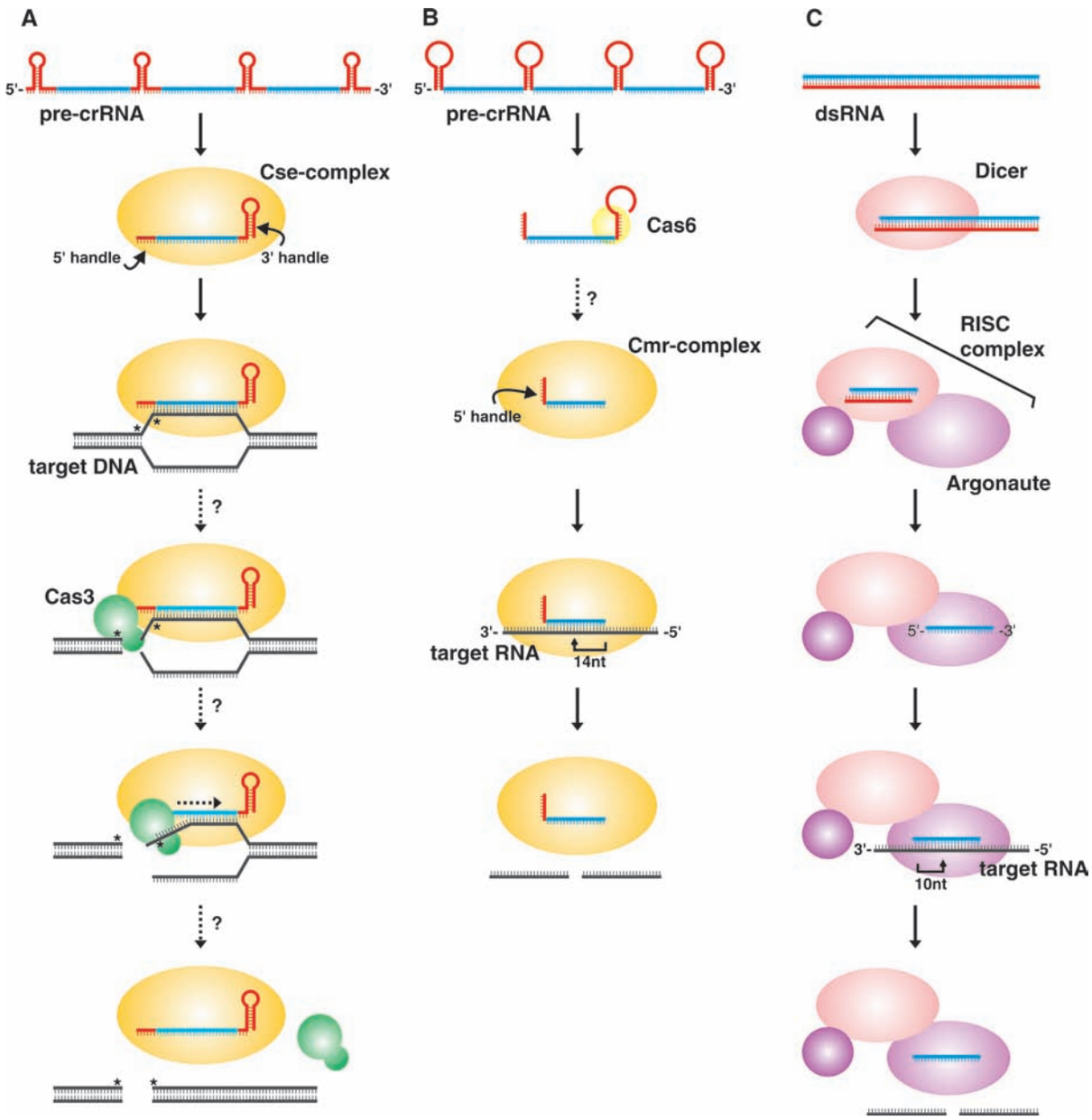


Figure 1.5. Antiviral DNA and RNA silencing pathways in prokaryotes and eukaryotes. (A) crRNA mediated DNA silencing pathway in *E. coli*. pre-crRNA is cleaved by the CasE subunit of Cascade (Cse-complex) and the mature crRNA remains bound to Cascade. When encountering viral dsDNA containing a sequence identical to the spacer sequence of the crRNA, it may basepair with the complementary DNA strand by a strand displacement event. The HD-domain of Cas3 is likely to be activated and cleave the viral DNA only when the 2 bases PAM on the viral DNA is present (marked with an asterisk). The helicase domain might subsequently separate the RNA:DNA duplex generating free Cascade that can be used in a next cleavage event. (B) crRNA mediated RNA silencing pathway in *P. furiosus*. Pre-crRNA is cleaved by Cas6 and then further trimmed to generate crRNAs of two different lengths. These crRNAs are bound by the Cmr-complex. This loaded Cmr-complex specifically binds viral RNA and cleaves the complementary strand 14 nucleotides away from the 3' end of the crRNA. This pathway shares functional analogies with siRNA mediated antiviral resistance in eukaryotes (C) siRNAs are generated from viral dsRNA by dicer. The first (random) cleavage event by dicer generates dsRNA with a 3' dinucleotide overhang. The second cleavage by dicer takes place 20-25 bases away from the overhang generating short dsRNAs. The dsRNA is transferred to the Argonaute protein of the RISC complex and the passenger strand is removed. The retained guide strand can basepair with a complementary viral mRNA molecule, followed by a cleavage of the scissile bond between the 10th and 11th base from the 3' end of the guide strand. The cleaved target RNA dissociates and the recycled RISC can be used in a second round of RNA binding and cleavage. Please note that dashed arrows indicate processes that are based on hypotheses.

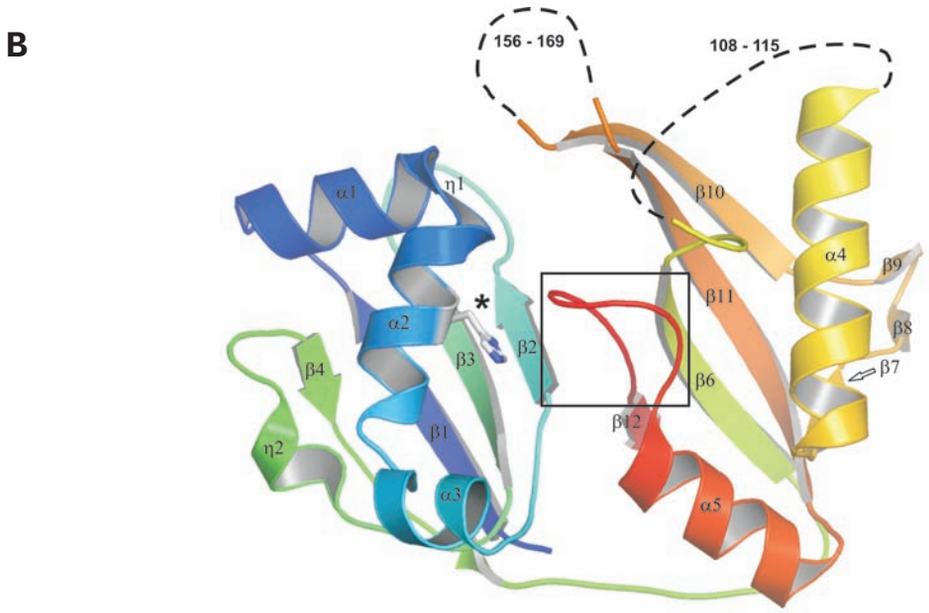


Figure S2.1

B. Ribbon diagram of the structure of TTHB192, a CasE homolog from *Thermus thermophilus* HB8 (PDB ID: 1WJ9) (Ebihara et al. 2006). Structural features are indicated as in **A**. Structurally disordered residues 108 to 115 and 156 to 169 are depicted by dashed lines. Note that the highly conserved glycine-rich loop between secondary structure elements α5 and β12 is spatially close to His26.

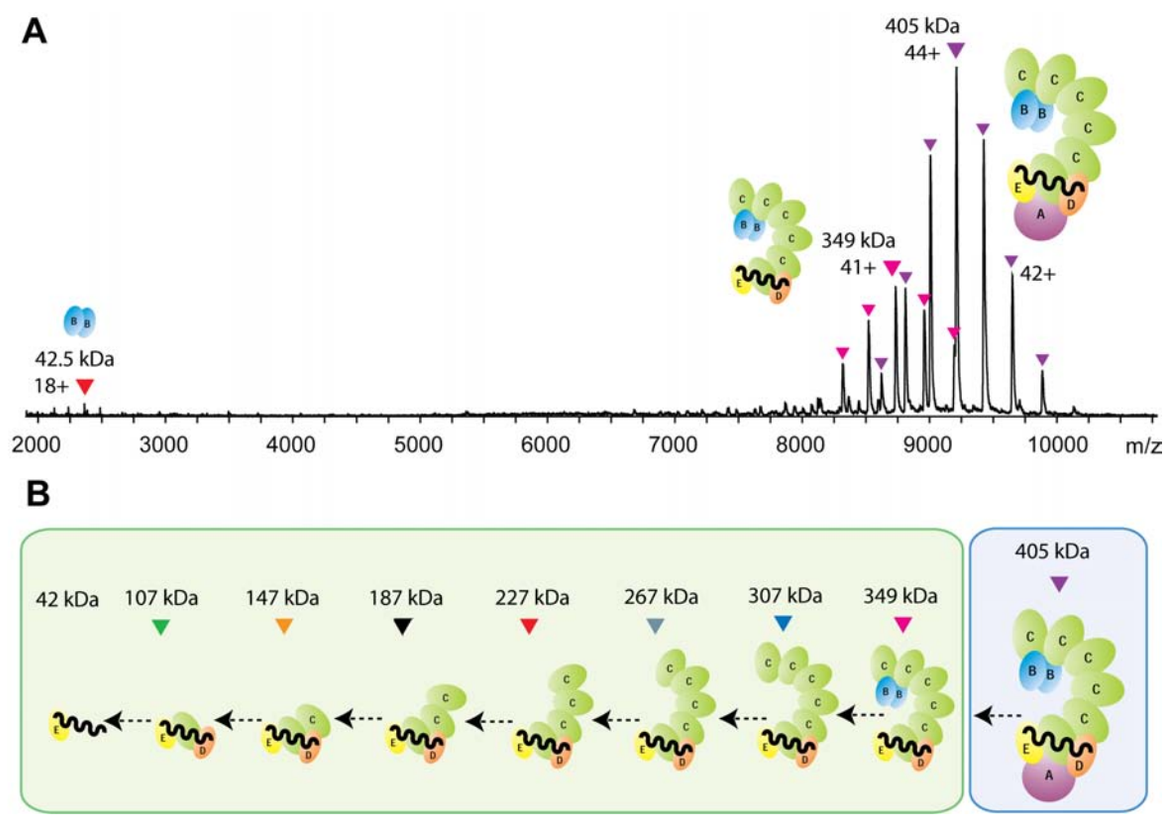


Figure 3.5. A) Native nano-ESI mass spectrum of Cascade. Two charge state distributions are present at high m/z values, corresponding to complexes of 405 kDa (purple) and 349 kDa (pink). The charge state distribution indicated in red indicates the CasB dimer. B) Cascade (sub-)complexes analyzed by native mass spectrometry. The sub-complexes were formed in solution after adding 5% 2-propanol to the buffer solution containing Cascade.

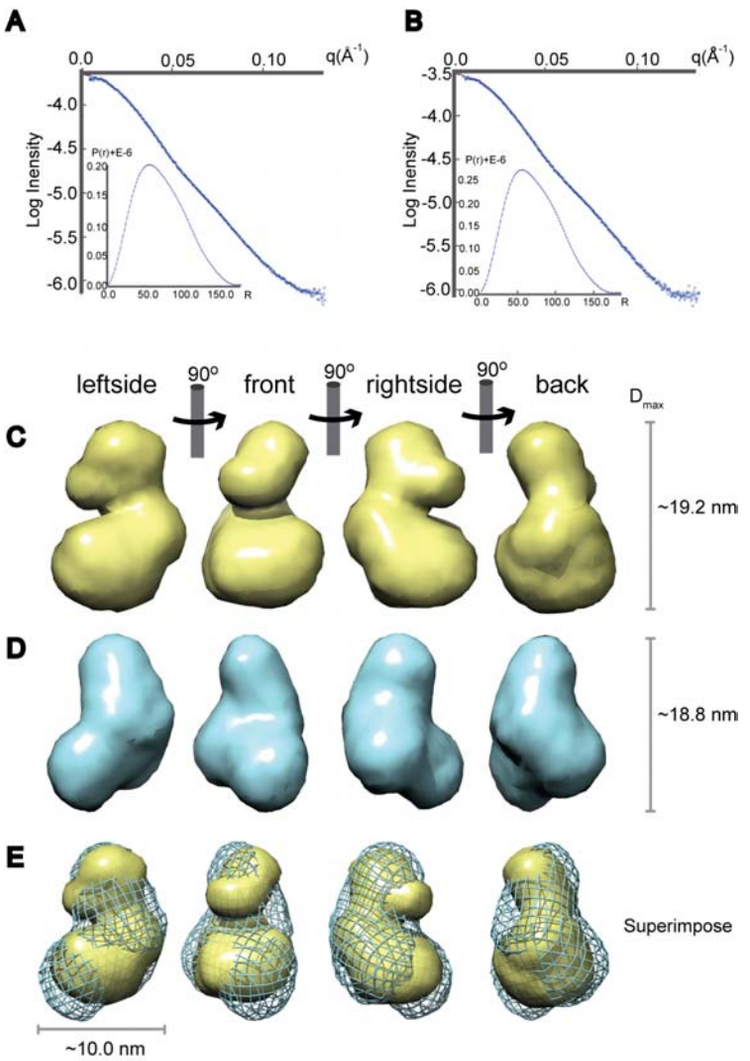


Figure 3.7. Solution scattering model of Cascade obtained with Small-Angle X-ray Scattering. Scattering data for Cascade were collected at 10 keV (1.24 Å λ) from two protein concentrations, and include scattering vectors (q), ranging from A) 0.015 $Å^{-1}$ to 0.127 $Å^{-1}$ for Cascade and B) 0.015 $Å^{-1}$ to 0.133 $Å^{-1}$ for Cascade bound to target DNA. The pair-distribution function (insert) indicates that the radius of gyration for both particles is ~5.6 nm. C) *Ab initio* reconstructions of Cascade reveal a seahorse shaped complex, consistent with EM imaging. D) DNA binding induces a conformational change in Cascade. E) Superposition of the solution structures of Cascade without (yellow) and with target DNA (mesh) suggest that regions of the complex assigned to the CasA and CasB are repositioned in the DNA bound state. Images have been rendered using Chimera (Goddard et al. 2005).

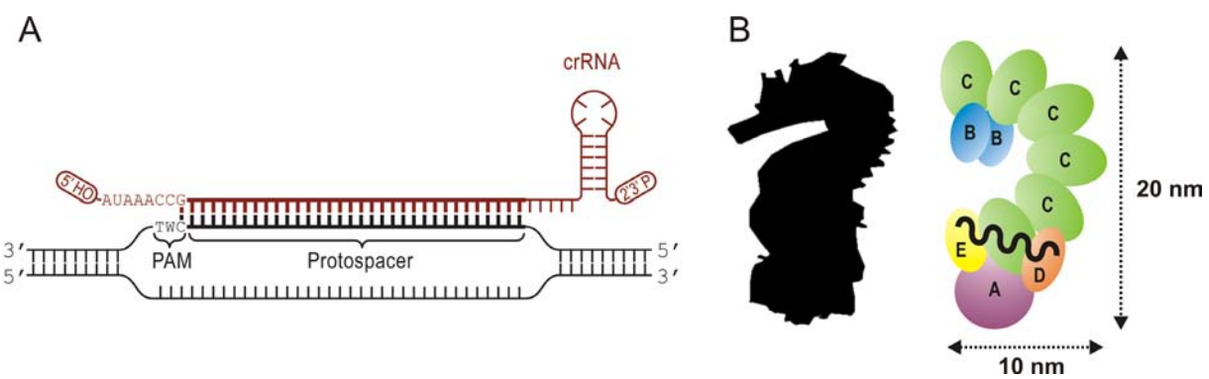


Figure 3.8. A) Schematic diagram of crRNA base paired to double stranded target DNA, indicating the local strand displacement, and the additional rG-dC basepair between the eighth base of the crRNA (rG) with the PAM (dC). B) Seahorse morphology and structural model of Cascade.

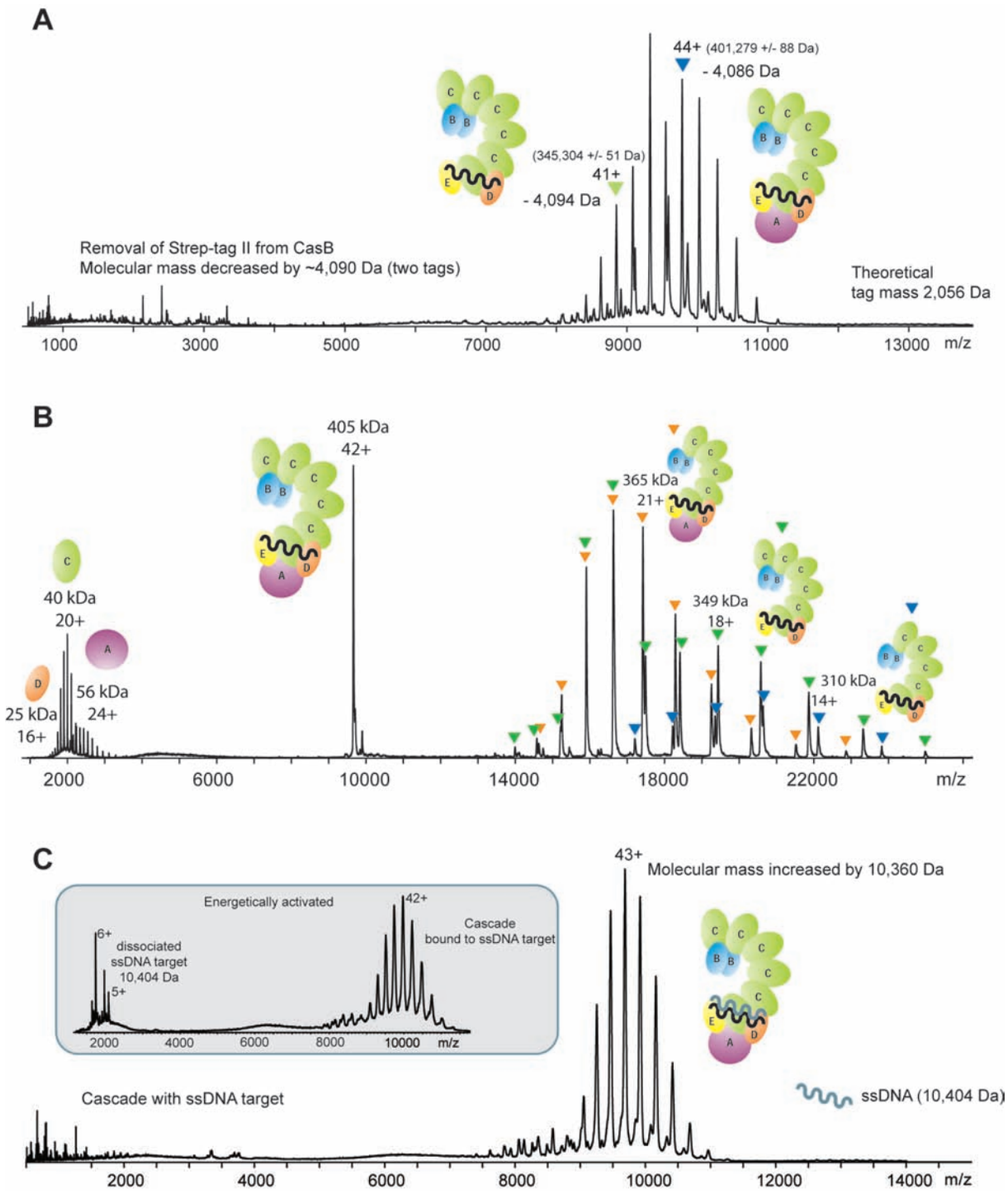


Figure S3.6. A) Native mass spectrum of Cascade after treatment with HRV3C protease. A dominant species with a mass of 401,279 Da (blue triangle) was observed, confirming the presence of two copies of CasB in the intact assembly. Indicated by the green triangles is the complex lacking CasA. B) Tandem mass spectrum of the 42+ ion of Cascade. Besides the dissociation of CasA (green) also CasC (orange) dissociated from the complex. The complex lacking CasA further expels a CasC subunit to form a 310 kDa Cascade sub-complex (blue). The low m/z region of the spectrum shows the dissociated CasA, CasC and CasD proteins. Overlapping peaks of two different complexes are indicated by two colours. C) Native mass spectrum of Cascade bound to the ssDNA-probe. The mass of the complex increased by 10,201 Da, indicating the presence of one crRNA per Cascade. The inset shows the same spectrum after energetically activating the Cascade-ssDNA probe complex. The charge state distribution for the ssDNA probe is centred around 2,000 m/z.

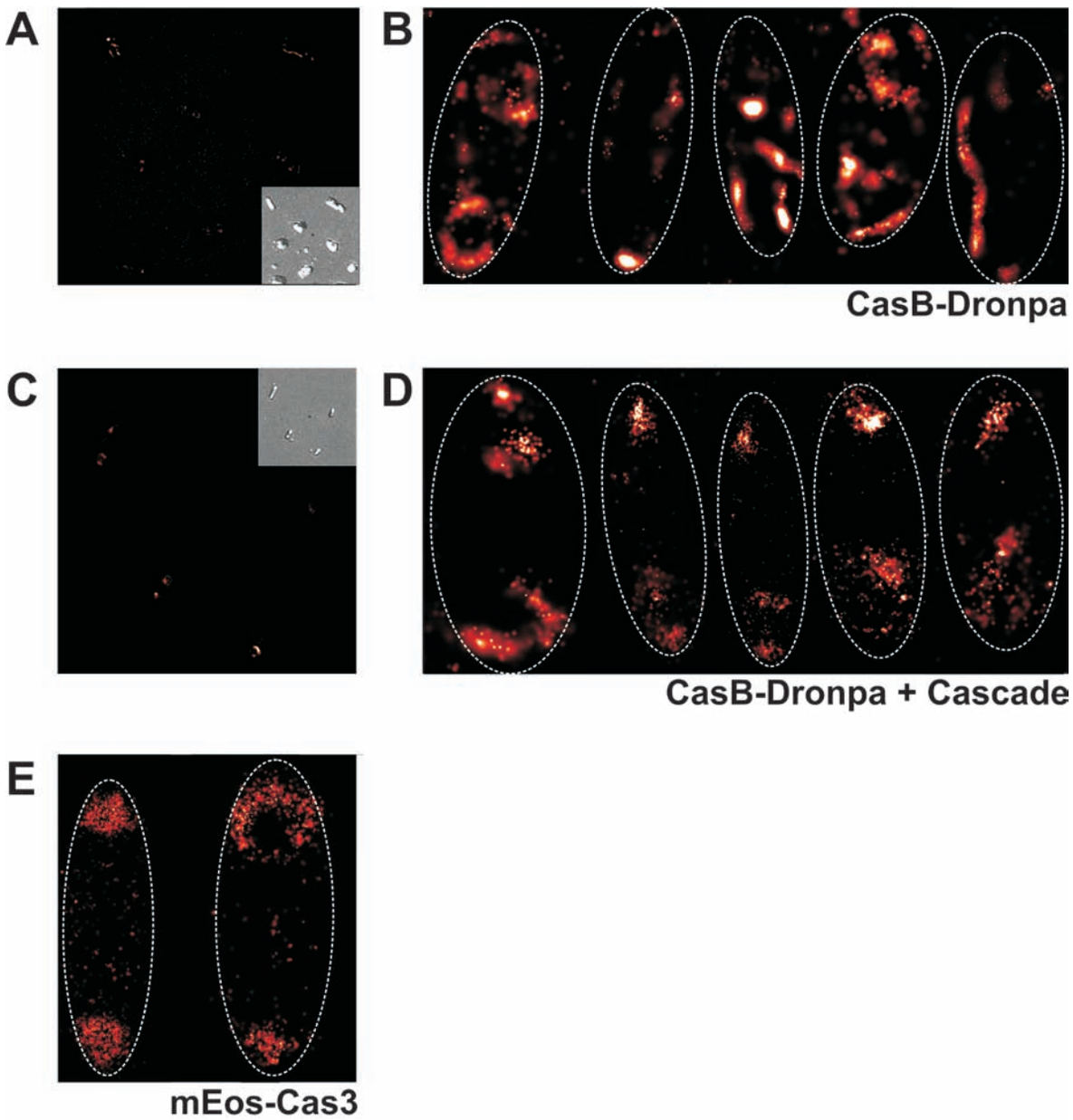


Figure 5.1. PALM images of the location of the Cascade proteins and Cas3. (A) An overview of CasB-Dronpa expressing cells visualized with PALM. A Digital Image Correlation (DIC) image (inset) shows that cells look unhealthy. Several images of cells are selected, rotated and enlarged, as shown in (B). Most proteins are located in patches along the cell wall. (C) An overview of CasB-Dronpa expressing cells in the presence of Cascade is visualized with PALM. A DIC image (inset) shows that cells look natural. Several images of cells are selected, rotated and enlarged as shown in (D). In the presence of Cascade, CasB-Dronpa is differently distributed in the cell, mainly concentrated in one focus at one pole of the cell and in a ring-like structure at the other end. (E) PALM analysis of two mEos-Cas3 expressing cells (representing a large population of cells) shows that localization is very similar to that of CasB-Dronpa + Cascade (D), although the ring-like structure can be absent and replaced by a focused cluster. Each spot in the images represents one protein molecule. Cell borders are roughly indicated with dashed lines for clarity in (B), (D) and (E).

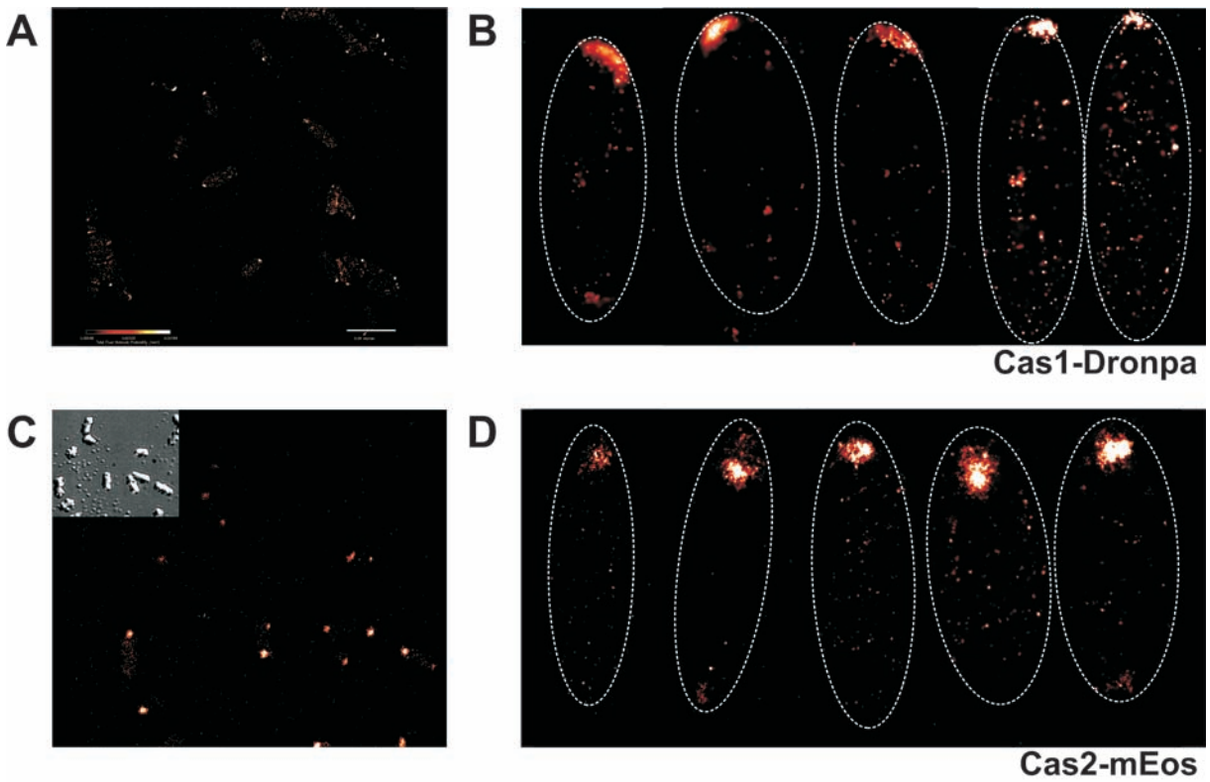


Figure 5.2. PALM images of the location of the Cas1 and Cas2 proteins. (A) An overview of Cas1-Dronpa expressing cells. Several images of cells are selected, rotated and enlarged as shown in (B). Some proteins localize uniformly over the cells, but most proteins form a cluster that localizes at one cellular pole close to the cell wall. (C) An overview of Cas2-mEos expressing cells. The inset shows a DIC image of the same cells. Several images of cells are selected, rotated and enlarged as shown in (D). As for Cas1-Dronpa, Cas2-mEos localizes mostly as one cluster at one pole of the cell. Each spot in the images represents one protein molecule. Cell borders are roughly indicated with dashed lines for clarity in (B) and (D).

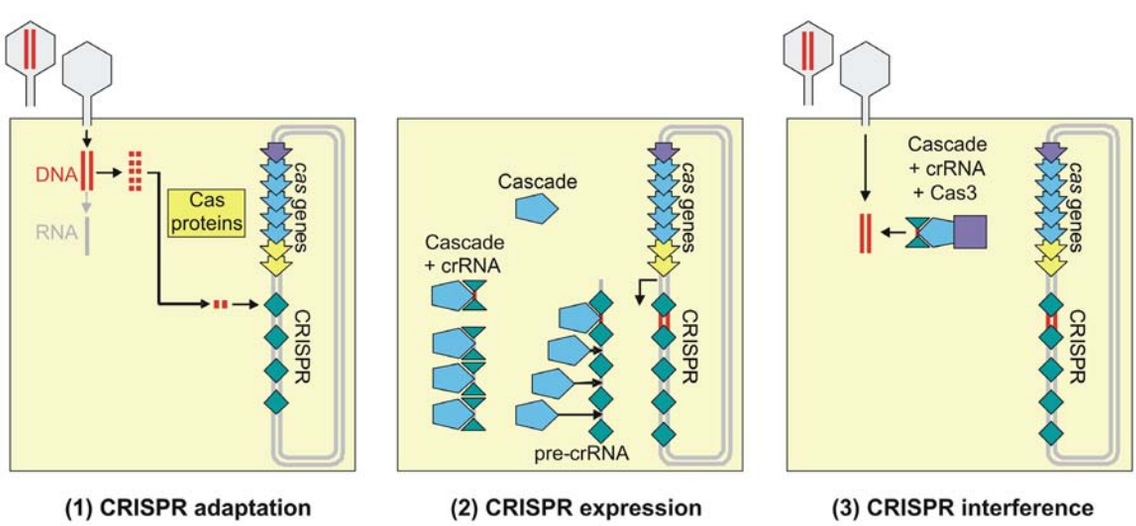


Figure 7.1. Three different stages in CRISPR-mediated antiviral defense. The different stages are explained in more detail in Chapter 1. Picture adapted from (van der Oost et al. 2009).

References

Afflerbach, H., Schroder, O., and Wagner, R. 1998. Effects of the Escherichia coli DNA-binding protein H-NS on rRNA synthesis in vivo. *Mol Microbiol* **28**(3): 641-653.

Agari, Y., Yokoyama, S., Kuramitsu, S., and Shinkai, A. 2008. X-ray crystal structure of a CRISPR-associated protein, Cse2, from Thermus thermophilus HB8. *Proteins* **73**(4): 1063-1067.

Alessandra, S., Alessandro, T., Flavio, S., and Alejandro, H. 2008. Artificial antisense RNAs silence lacZ in E. coli by decreasing target mRNA concentration. *BMB Rep* **41**(8): 568-574.

Andersson, A.F. and Banfield, J.F. 2008. Virus population dynamics and acquired virus resistance in natural microbial communities. *Science* **320**(5879): 1047-1050.

Aravind, L. and Koonin, E.V. 1998. The HD domain defines a new superfamily of metal-dependent phosphohydrolases. *Trends Biochem Sci* **23**(12): 469-472.

Baba, T., Huan, H.C., Datsenko, K., Wanner, B.L., and Mori, H. 2008. The applications of systematic in-frame, single-gene knockout mutant collection of Escherichia coli K-12. *Methods Mol Biol* **416**: 183-194.

Baba, T. and Mori, H. 2008. The construction of systematic in-frame, single-gene knockout mutant collection in Escherichia coli K-12. *Methods Mol Biol* **416**: 171-181.

Banfield, J.F. and Young, M. 2009. Microbiology. Variety--the splice of life--in microbial communities. *Science* **326**(5957): 1198-1199.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., and Horvath, P. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**(5819): 1709-1712.

Barrick, J.E., Yu, D.S., Yoon, S.H., Jeong, H., Oh, T.K., Schneider, D., Lenski, R.E., and Kim, J.F. 2009. Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature* **461**(7268): 1243-1247.

Baulcombe, D. 2004. RNA silencing in plants. *Nature* **431**(7006): 356-363.

Beloglazova, N., Brown, G., Zimmerman, M.D., Proudfoot, M., Makarova, K.S., Kudritska, M., Kochinyan, S., Wang, S., Chruszcz, M., Minor, W., Koonin, E.V., Edwards, A.M., Savchenko, A., and Yakunin, A.F. 2008. A novel family of sequence-specific endoribonucleases associated with the Clustered Regularly Interspaced Short Palindromic Repeats. *J Biol Chem*.

Bergh, O., Borsheim, K.Y., Bratbak, G., and Heldal, M. 1989. High abundance of viruses found in aquatic environments. *Nature* **340**(6233): 467-468.

Bernstein, E., Caudy, A.A., Hammond, S.M., and Hannon, G.J. 2001. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409**(6818): 363-366.

Betzig, E., Patterson, G.H., Sougrat, R., Lindwasser, O.W., Olenych, S., Bonifacino, J.S., Davidson, M.W., Lippincott-Schwartz, J., and Hess, H.F. 2006. Imaging intracellular fluorescent proteins at nanometer resolution. *Science* **313**(5793): 1642-1645.

Blomberg, P., Wagner, E.G., and Nordstrom, K. 1990. Control of replication of plasmid R1: the duplex between the antisense RNA, CopA, and its target, CopT, is processed specifically in vivo and in vitro by RNase III. *Embo J* **9**(7): 2331-2340.

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

Botero, L.M., D'Imperio, S., Burr, M., McDermott, T.R., Young, M., and Hassett, D.J. 2005. Poly(A) polymerase modification and reverse transcriptase PCR amplification of environmental RNA. *Appl Environ Microbiol* **71**(3): 1267-1275.

Bouffartigues, E., Buckle, M., Badaut, C., Travers, A., and Rimsky, S. 2007. H-NS cooperative binding to high-affinity sites in a regulatory element results in transcriptional silencing. *Nat Struct Mol Biol* **14**(5): 441-448.

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

Calvin, K. and Li, H. 2008. RNA-splicing endonuclease structure and function. *Cell Mol Life Sci* **65**(7-8): 1176-1185.

Carte, J., Wang, R., Li, H., Terns, R.M., and Terns, M.P. 2008. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes Dev* **22**(24): 3489-3496.

Castanotto, D. and Rossi, J.J. 2009. The promises and pitfalls of RNA-interference-based therapeutics. *Nature* **457**(7228): 426-433.

Chakraborty, S., Waise, T.M., Hassan, F., Kabir, Y., Smith, M.A., and Arif, M. 2009. Assessment of the evolutionary origin and possibility of CRISPR-Cas (CASS) mediated RNA interference pathway in. *In Silico Biol* **9**(4): 245-254.

Chen, C.C., Chou, M.Y., Huang, C.H., Majumder, A., and Wu, H.Y. 2005. A cis-spreading nucleoprotein filament is responsible for the gene silencing activity found in the promoter relay mechanism. *J Biol Chem* **280**(6): 5101-5112.

Chen, C.C., Fang, M., Majumder, A., and Wu, H.Y. 2001. A 72-base pair AT-rich DNA sequence element functions as a bacterial gene silencer. *J Biol Chem* **276**(12): 9478-9485.

Chen, C.C. and Wu, H.Y. 2005. LeuO protein delimits the transcriptionally active and repressive domains on the bacterial chromosome. *J Biol Chem* **280**(15): 15111-15121.

Chopin, M.C., Chopin, A., and Bidnenko, E. 2005. Phage abortive infection in lactococci: variations on a theme. *Curr Opin Microbiol* **8**(4): 473-479.

Dame, R.T., Luijsterburg, M.S., Krin, E., Bertin, P.N., Wagner, R., and Wuite, G.J. 2005. DNA bridging: a property shared among H-NS-like proteins. *J Bacteriol* **187**(5): 1845-1848.

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

De la Cruz, M.A., Fernandez-Mora, M., Guadarrama, C., Flores-Valdez, M.A., Bustamante, V.H., Vazquez, A., and Calva, E. 2007. LeuO antagonizes H-NS and StpA-dependent repression in Salmonella enterica ompS1. *Mol Microbiol* **66**(3): 727-743.

DeBoy, R.T., Mongodin, E.F., Emerson, J.B., and Nelson, K.E. 2006. Chromosome evolution in the Thermotogales: large-scale inversions and strain diversification of CRISPR sequences. *J Bacteriol* **188**(7): 2364-2374.

Dersch, P., Schmidt, K., and Bremer, E. 1993. Synthesis of the Escherichia coli K-12 nucleoid-associated DNA-binding protein H-NS is subjected to growth-phase control and autoregulation. *Mol Microbiol* **8**(5): 875-889.

Deveau, H., Barrangou, R., Garneau, J.E., Labonte, J., Fremaux, C., Boyaval, P., Romero, D.A., Horvath, P., and Moineau, S. 2008. Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *J Bacteriol* **190**(4): 1390-1400.

Dickman, M.J. and Hornby, D.P. 2006. Enrichment and analysis of RNA centered on ion pair reverse phase methodology. *RNA* **12**(4): 691-696.

Diez-Villasenor, C., Almendros, C., Garcia-Martinez, J., and Mojica, F.J. 2010. Diversity of CRISPR loci in Escherichia coli. *Microbiology*.

Dillon, S.C., Cameron, A.D., Hokamp, K., Lucchini, S., Hinton, J.C., and Dorman, C.J. 2010. Genome-wide analysis of the H-NS and Sfh regulatory networks in Salmonella Typhimurium identifies a plasmid-encoded transcription silencing mechanism. *Mol Microbiol*.

Doyle, M., Fookes, M., Ivens, A., Mangan, M.W., Wain, J., and Dorman, C.J. 2007. An H-NS-like stealth protein aids horizontal DNA transmission in bacteria. *Science* **315**(5809): 251-252.

Driscoll, J.R. 2009. Spoligotyping for molecular epidemiology of the Mycobacterium tuberculosis complex. *Methods Mol Biol* **551**: 117-128.

Ebihara, A., Yao, M., Masui, R., Tanaka, I., Yokoyama, S., and Kuramitsu, S. 2006. Crystal structure of hypothetical protein TTHB192 from Thermus thermophilus HB8 reveals a new protein family with an RNA recognition motif-like domain. *Protein Sci* **15**(6): 1494-1499.

Edgar, R., Rokney, A., Feeney, M., Semsey, S., Kessel, M., Goldberg, M.B., Adhya, S., and Oppenheim, A.B. 2008. Bacteriophage infection is targeted to cellular poles. *Mol Microbiol* **68**(5): 1107-1116.

Elbashir, S.M., Lendeckel, W., and Tuschl, T. 2001. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev* **15**(2): 188-200.

Fang, M., Majumder, A., Tsai, K.J., and Wu, H.Y. 2000. ppGpp-dependent leuO expression in bacteria under stress. *Biochem Biophys Res Commun* **276**(1): 64-70.

Franke, D. and Svergun, D.I. 2009. DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. *Journal of Applied Crystallography* **42**: 342-346.

Goddard, T.D., Huang, C.C., and Ferrin, T.E. 2005. Software extensions to UCSF Chimera for interactive visualization of large molecular assemblies. *Structure* **13**(3): 473-482.

Godde, J.S. and Bickerton, A. 2006. The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. *J Mol Evol* **62**(6): 718-729.

Gouet, P., Robert, X., and Courcelle, E. 2003. ESPript/ENDscript: Extracting and rendering sequence and 3D information from atomic structures of proteins. *Nucleic Acids Res* **31**(13): 3320-3323.

Greenfield, D., McEvoy, A.L., Shroff, H., Crooks, G.E., Wingreen, N.S., Betzig, E., and Liphardt, J. 2009. Self-organization of the Escherichia coli chemotaxis network imaged with super-resolution light microscopy. *PLoS Biol* **7**(6): e1000137.

Haft, D.H., Selengut, J., Mongodin, E.F., and Nelson, K.E. 2005. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput Biol* **1**(6): e60.

Hale, C., Kleppe, K., Terns, R.M., and Terns, M.P. 2008. Prokaryotic silencing (psi)RNAs in Pyrococcus furiosus. *RNA* **14**(12): 2572-2579.

Hale, C.R., Zhao, P., Olson, S., Duff, M.O., Graveley, B.R., Wells, L., Terns, R.M., and Terns, M.P. 2009. RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* **139**(5): 945-956.

Han, D. and Krauss, G. 2009. Characterization of the endonuclease SSO2001 from Sulfolobus solfataricus P2. *FEBS Lett* **583**(4): 771-776.

Han, D., Lehmann, K., and Krauss, G. 2009. SSO1450--a CAS1 protein from Sulfolobus solfataricus P2 with high affinity for RNA and DNA. *FEBS Lett* **583**(12): 1928-1932.

Hartnett, J., Jill, M.S., Gracyalny, B.S., and Slater, M.R. 2006. The single step (KRX): efficient cloning and high protein yields. *Promega Notes* **94**: 27-30.

Hayashi, K., Morooka, N., Yamamoto, Y., Fujita, K., Isono, K., Choi, S., Ohtsubo, E., Baba, T., Wanner, B.L., Mori, H., and Horiuchi, T. 2006. Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110. *Mol Syst Biol* **2**: 2006 0007.

Heck, A.J.R. 2008. Native mass spectrometry: a bridge between interactomics and structural biology. *Nature Methods* **5**(11): 927-933.

Heidelberg, J.F., Nelson, W.C., Schoenfeld, T., and Bhaya, D. 2009. Germ warfare in a microbial mat community: CRISPRs provide insights into the co-evolution of host and viral genomes. *PLoS One* **4**(1): e4169.

Held, N.L. and Whitaker, R.J. 2009. Viral biogeography revealed by signatures in Sulfolobus islandicus genomes. *Environ Microbiol* **11**(2): 457-466.

Hermans, P.W., van Soolingen, D., Bik, E.M., de Haas, P.E., Dale, J.W., and van Embden, J.D. 1991. Insertion element IS987 from Mycobacterium bovis BCG is located in a hot-spot integration region for insertion elements in Mycobacterium tuberculosis complex strains. *Infect Immun* **59**(8): 2695-2705.

Hendrix, R. and Casjens, S. 2006. The bacteriophages, second edition, Chapter 27, Oxford University Press

Hernandez-Lucas, I., Gallego-Hernandez, A.L., Encarnacion, S., Fernandez-Mora, M., Martinez-Batallar, A.G., Salgado, H., Oropeza, R., and Calva, E. 2008. The LysR-type transcriptional regulator LeuO controls expression of several genes in Salmonella enterica serovar typhi. *Journal of Bacteriology* **190**(5): 1658-1670.

Hertzberg, K.M., Gemmill, R., Jones, J., and Calvo, J.M. 1980. Cloning of an EcoRI-generated fragment of the leucine operon of Salmonella typhimurium. *Gene* **8**(2): 135-152.

Hofnung, M., Jezierska, A., and Braun-Breton, C. 1976. lamB mutations in E. coli K12: growth of lambda host range mutants and effect of nonsense suppressors. *Mol Gen Genet* **145**(2): 207-213.

Hommais, F., Krin, E., Laurent-Winter, C., Soutourina, O., Malpertuy, A., Le Caer, J.P., Danchin, A., and Bertin, P. 2001. Large-scale monitoring of pleiotropic regulation of gene expression by the prokaryotic nucleoid-associated protein, H-NS. *Molecular Microbiology* **40**(1): 20-36.

Horvath, P. and Barrangou, R. 2010. CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327**(5962): 167-170.

Horvath, P., Barrangou, R., Fremaux, C., Boyaval, P., and Romero, D. 2007. International Patent Application 2007025097.

Horvath, P., Coute-Monvoisin, A.C., Romero, D.A., Boyaval, P., Fremaux, C., and Barrangou, R. 2009. Comparative analysis of CRISPR loci in lactic acid bacteria genomes. *Int J Food Microbiol* **131**(1): 62-70.

Horvath, P., Romero, D.A., Coute-Monvoisin, A.C., Richards, M., Deveau, H., Moineau, S., Boyaval, P., Fremaux, C., and Barrangou, R. 2008. Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. *J Bacteriol* **190**(4): 1401-1412.

Houwing, S., Kamminga, L.M., Berezikov, E., Cronembold, D., Girard, A., van den Elst, H., Filippov, D.V., Blaser, H., Raz, E., Moens, C.B., Plasterk, R.H., Hannon, G.J., Draper, B.W., and Ketting,

R.F. 2007. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. *Cell* **129**(1): 69-82.

Hyman, P. and Abedon, S.T. 2010. Bacteriophage Host Range and Bacterial Resistance. *Adv Appl Microbiol* **70C**: 217-248.

Ishino, Y., Shinagawa, H., Makino, K., Amemura, M., and Nakata, A. 1987. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. *J Bacteriol* **169**(12): 5429-5433.

Jacob, F. and Wollman, E.L. 1954. *Ann Inst Pasteur (Paris)* **87**(6): 653-673.

Jansen, R., van Embden, J.D., Gaastra, W., and Schouls, L.M. 2002. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* **43**(6): 1565-1575.

Jinek, M. and Doudna, J.A. 2009. A three-dimensional view of the molecular machinery of RNA interference. *Nature* **457**(7228): 405-412.

Kamerbeek, J., Schouls, L., Kolk, A., van Agterveld, M., van Soolingen, D., Kuijper, S., Bunschoten, A., Molhuizen, H., Shaw, R., Goyal, M., and van Embden, J. 1997. Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. *J Clin Microbiol* **35**(4): 907-914.

Karginov, F.V. and Hannon, G.J. 2010. The CRISPR System: Small RNA-Guided Defense in Bacteria and Archaea. *Mol Cell* **37**(1): 7-19.

Kim, J.F., Jeong, H., and Lenski, R.E. Personal communication.

Kitagawa, M., Ara, T., Arifuzzaman, M., Ioka-Nakamichi, T., Inamoto, E., Toyonaga, H., and Mori, H. 2005. Complete set of ORF clones of Escherichia coli ASKA library (a complete set of E. coli K-12 ORF archive): unique resources for biological research. *DNA Res* **12**(5): 291-299.

Klauck, E., Bohringer, J., and Hengge-Aronis, R. 1997. The LysR-like regulator LeuO in Escherichia coli is involved in the translational regulation of rpoS by affecting the expression of the small regulatory DsrA-RNA. *Mol Microbiol* **25**(3): 559-569.

Kleywegt, G.J. and Jones, T.A. 1994. Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Crystallogr D Biol Crystallogr* **50**(Pt 2): 178-185.

Konarev, P.V., Volkov, V.V., Sokolova, A.V., Koch, M.H.J., and Svergun, D.I. 2003. PRIMUS: a Windows PC-based system for small-angle scattering data analysis. *Journal of Applied Crystallography* **36**: 1277-1282.

Koonin, E.V. and Wolf, Y.I. 2009. Is evolution Darwinian or/and Lamarckian? *Biol Direct* **4**: 42.

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

Labrie, S.J., Samson, J.E., and Moineau, S. 2010. Bacteriophage resistance mechanisms. *Nat Rev Microbiol* **8**(5): 317-327.

Lang, B., Blot, N., Bouffartigues, E., Buckle, M., Geertz, M., Gualerzi, C.O., Mavathur, R., Muskhelishvili, G., Pon, C.L., Rimsky, S., Stella, S., Babu, M.M., and Travers, A. 2007. High-affinity DNA binding sites for H-NS provide a molecular basis for selective silencing within proteobacterial genomes. *Nucleic Acids Res* **35**(18): 6330-6337.

Lawley, T.D., Chan, K., Thompson, L.J., Kim, C.C., Govoni, G.R., and Monack, D.M. 2006. Genome-wide screen for Salmonella genes required for long-term systemic infection of the mouse. *PLoS Pathog* **2**(2): e11.

Lewis, B.P., Burge, C.B., and Bartel, D.P. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**(1): 15-20.

Li, J., Yang, Z., Yu, B., Liu, J., and Chen, X. 2005. Methylation protects miRNAs and siRNAs from a 3'-end uridylation activity in Arabidopsis. *Curr Biol* **15**(16): 1501-1507.

Lillestol, R.K., Redder, P., Garrett, R.A., and Brugger, K. 2006. A putative viral defence mechanism in archaeal cells. *Archaea* **2**(1): 59-72.

Lillestol, R.K., Shah, S.A., Brugger, K., Redder, P., Phan, H., Christiansen, J., and Garrett, R.A. 2009. CRISPR families of the crenarchaeal genus Sulfolobus: bidirectional transcription and dynamic properties. *Mol Microbiol* **72**(1): 259-272.

Liu, Y., Chen, H., Kenney, L.J., and Yan, J. 2010. A divalent switch drives H-NS/DNA-binding conformations between stiffening and bridging modes. *Genes Dev* **24**(4): 339-344.

Llosa, M., Gomis-Ruth, F.X., Coll, M., and de la Cruz Fd, F. 2002. Bacterial conjugation: a two-step mechanism for DNA transport. *Mol Microbiol* **45**(1): 1-8.

Locker, J. 1979. Analytical and preparative electrophoresis of RNA in agarose-urea. *Anal Biochem* **98**(2): 358-367.

Lorenzen, K., Vannini, A., Cramer, P., and Heck, A.J. 2007. Structural biology of RNA polymerase III: mass spectrometry elucidates subcomplex architecture. *Structure* **15**(10): 1237-1245.

Lucchini, S., Rowley, G., Goldberg, M.D., Hurd, D., Harrison, M., and Hinton, J.C. 2006. H-NS mediates the silencing of laterally acquired genes in bacteria. *PLoS Pathog* **2**(8): e81.

Ma, J.B., Yuan, Y.R., Meister, G., Pei, Y., Tuschl, T., and Patel, D.J. 2005. Structural basis for 5'-end-specific recognition of guide RNA by the A. fulgidus Piwi protein. *Nature* **434**(7033): 666-670.

MacRae, I.J. and Doudna, J.A. 2007. Ribonuclease revisited: structural insights into ribonuclease III family enzymes. *Curr Opin Struct Biol* **17**(1): 138-145.

MacRae, I.J., Zhou, K., Li, F., Repic, A., Brooks, A.N., Cande, W.Z., Adams, P.D., and Doudna, J.A. 2006. Structural basis for double-stranded RNA processing by Dicer. *Science* **311**(5758): 195-198.

Maddocks, S.E. and Oyston, P.C. 2008. Structure and function of the LysR-type transcriptional regulator (LTTR) family proteins. *Microbiology* **154**(Pt 12): 3609-3623.

Madhusudan, S., Paukner, A., Klingen, Y., and Schnetz, K. 2005. Independent regulation of H-NS-mediated silencing of the bgl operon at two levels: upstream by BglJ and LeuO and downstream by DnaKJ. *Microbiology* **151**(Pt 10): 3349-3359.

Majumder, A., Fang, M., Tsai, K.J., Ueguchi, C., Mizuno, T., and Wu, H.Y. 2001. LeuO expression in response to starvation for branched-chain amino acids. *J Biol Chem* **276**(22): 19046-19051.

Makarova, K.S., Aravind, L., Grishin, N.V., Rogozin, I.B., and Koonin, E.V. 2002. A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res* **30**(2): 482-496.

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

Makarova, K.S., Wolf, Y.I., van der Oost, J., and Koonin, E.V. 2009. Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements. *Biol Direct* **4**: 29.

Marraffini, L.A. and Sontheimer, E.J. 2008. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* **322**(5909): 1843-1845.

Marraffini, L.A. and Sontheimer, E.J. 2009. Invasive DNA, chopped and in the CRISPR. *Structure* **17**(6): 786-788.

Marraffini, L.A. and Sontheimer, E.J. 2010a. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat Rev Genet* **11**(3): 181-190.

Marraffini, L.A. and Sontheimer, E.J. 2010b. Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature* **463**(7280): 568-571.

Mojica, F.J., Diez-Villasenor, C., Garcia-Martinez, J., and Almendros, C. 2009. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**(Pt 3): 733-740.

Mojica, F.J., Diez-Villasenor, C., Garcia-Martinez, J., and Soria, E. 2005. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* **60**(2): 174-182.

Mojica, F.J., Diez-Villasenor, C., Soria, E., and Juez, G. 2000. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Mol Microbiol* **36**(1): 244-246.

Mojica, F.J., Ferrer, C., Juez, G., and Rodriguez-Valera, F. 1995. Long stretches of short tandem repeats are present in the largest replicons of the Archaea Haloferax mediterranei and Haloferax volcanii and could be involved in replicon partitioning. *Mol Microbiol* **17**(1): 85-93.

Navarre, W.W., McClelland, M., Libby, S.J., and Fang, F.C. 2007. Silencing of xenogeneic DNA by H-NS-facilitation of lateral gene transfer in bacteria by a defense system that recognizes foreign DNA. *Genes Dev* **21**(12): 1456-1471.

Navarre, W.W., Porwollik, S., Wang, Y., McClelland, M., Rosen, H., Libby, S.J., and Fang, F.C. 2006. Selective silencing of foreign DNA with low GC content by the H-NS protein in Salmonella. *Science* **313**(5784): 236-238.

Noy, A., Perez, A., Marquez, M., Luque, F.J., and Orozco, M. 2005. Structure, recognition properties, and flexibility of the DNA.RNA hybrid. *J Am Chem Soc* **127**(13): 4910-4920.

Ochman, H. and Selander, R.K. 1984. Standard reference strains of Escherichia coli from natural

populations. *J Bacteriol* **157**(2): 690-693.
Oostergetel, G.T., Keegstra, W., and Brisson, A. 1998. Automation of specimen selection and data acquisition for protein electron crystallography. *Ultramicroscopy* **74**(1-2): 47-59.
Oshima, T., Ishikawa, S., Kurokawa, K., Aiba, H., and Ogasawara, N. 2006. Escherichia coli histone-like protein H-NS preferentially binds to horizontally acquired DNA in association with RNA polymerase. *DNA Research* **13**: 141-153.
Peters, J.E., Thate, T.E., and Craig, N.L. 2003. Definition of the Escherichia coli MC4100 genome by use of a DNA array. *J Bacteriol* **185**(6): 2017-2021.
Poirot, O., O'Toole, E., and Notredame, C. 2003. Tcoffee@igs: A web server for computing, evaluating and combining multiple sequence alignments. *Nucleic Acids Res* **31**(13): 3503-3506.
Poranen, M.M., Ravantti, J.J., Grahn, A.M., Gupta, R., Auvinen, P., and Bamford, D.H. 2006. Global changes in cellular gene expression during bacteriophage PRD1 infection. *J Virol* **80**(16): 8081-8088.
Pourcel, C., Salvignol, G., and Vergnaud, G. 2005. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151**(Pt 3): 653-663.
Pul, U., Wurm, R., Arslan, Z., Geissen, R., Hofmann, N., and Wagner, R. 2010. Identification and characterization of E. coli CRISPR-cas promoters and their silencing by H-NS. *Mol Microbiol*.
Qimron, U., Marintcheva, B., Tabor, S., and Richardson, C.C. 2006. Genomewide screens for Escherichia coli genes affecting growth of T7 bacteriophage. *Proc Natl Acad Sci U S A* **103**(50): 19039-19044.
Rand, T.A., Petersen, S., Du, F., and Wang, X. 2005. Argonaute2 cleaves the anti-guide strand of siRNA during RISC activation. *Cell* **123**(4): 621-629.
Rasmussen, L.C., Sperling-Petersen, H.U., and Mortensen, K.K. 2007. Hitting bacteria at the heart of the central dogma: sequence-specific inhibition. *Microb Cell Fact* **6**: 24.
Robinow, C. and Kellenberger, E. 1994. The bacterial nucleoid revisited. *Microbiol Rev* **58**(2): 211-232.
Rokney, A., Shagan, M., Kessel, M., Smith, Y., Rosenshine, I., and Oppenheim, A.B. 2009. E. coli transports aggregated proteins to the poles by a specific and energy-dependent process. *J Mol Biol* **392**(3): 589-601.
Sakamoto, K., Agari, Y., Agari, K., Yokoyama, S., Kuramitsu, S., and Shinkai, A. 2009. X-ray crystal structure of a CRISPR-associated RAMP superfamily protein, Cmr5, from Thermus thermophilus HB8. *Proteins* **75**(2): 528-532.
Sambrook, J., Fritsch, E.F., and Maniatis, T. 1989. Molecular Cloning. *Cold Spring Laboratory Press, New York, 2nd edition*.
Scandella, D. and Arber, W. 1974. An Escherichia coli mutant which inhibits the injection of phage lambda DNA. *Virology* **58**(2): 504-513.
Schouls, L.M., Reulen, S., Duim, B., Wagenaar, J.A., Willems, R.J., Dingle, K.E., Colles, F.M., and Van Embden, J.D. 2003. Comparative genotyping of Campylobacter jejuni by amplified fragment length polymorphism, multilocus sequence typing, and short repeat sequencing: strain diversity, host range, and recombination. *J Clin Microbiol* **41**(1): 15-26.
Schrader, H.S., Schrader, J.O., Walker, J.J., Wolf, T.A., Nickerson, K.W., and Kokjohn, T.A. 1997. Bacteriophage infection and multiplication occur in Pseudomonas aeruginosa starved for 5 years. *Can J Microbiol* **43**(12): 1157-1163.
Semenova, E., Nagornykh, M., Pyatnitskiy, M., Artamonova, II, and Severinov, K. 2009. Analysis of CRISPR system function in plant pathogen Xanthomonas oryzae. *FEMS Microbiol Lett* **296**(1): 110-116.
Shapiro, L., McAdams, H.H., and Losick, R. 2009. Why and how bacteria localize proteins. *Science* **326**(5957): 1225-1228.
Shimada, T., Yamamoto, K., and Ishihama, A. 2009. Involvement of the Leucine Response Transcription Factor LeuO in Regulation of the Genes for Sulfa Drug Efflux. *Journal of Bacteriology* **191**(14): 4562-4571.
Shroff, H., Galbraith, C.G., Galbraith, J.A., White, H., Gillette, J., Olenych, S., Davidson, M.W., and Betzig, E. 2007. Dual-color superresolution imaging of genetically expressed probes within individual adhesion complexes. *Proc Natl Acad Sci U S A* **104**(51): 20308-20313.
Snijders, A.P., Walther, J., Peter, S., Kinnman, I., de Vos, M.G., van de Werken, H.J., Brouns, S.J., van der Oost, J., and Wright, P.C. 2006. Reconstruction of central carbon metabolism in Sulfolobus solfataricus using a two-dimensional gel electrophoresis map, stable isotope labelling and DNA

microarray analysis. *Proteomics* **6**(5): 1518-1529.
Sorek, R., Kunin, V., and Hugenholtz, P. 2008. CRISPR - a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol* **6**(3): 181-186.
Stoebel, D.M., Free, A., and Dorman, C.J. 2008. Anti-silencing: overcoming H-NS-mediated repression of transcription in Gram-negative enteric bacteria. *Microbiology-Sgm* **154**: 2533-2545.
Stratmann, T., Madhusudan, S., and Schnetz, K. 2008. Regulation of the yjjQ-bglJ operon, encoding LuxR-type transcription factors, and the divergent yjjP gene by H-NS and LeuO. *J Bacteriol* **190**(3): 926-935.
Studier, F.W., Daegelen, P., Lenski, R.E., Maslov, S., and Kim, J.F. 2009. Understanding the differences between genome sequences of Escherichia coli B strains REL606 and BL21(DE3) and comparison of the E. coli B and K-12 genomes. *J Mol Biol* **394**(4): 653-680.
Sturino, J.M. and Klaenhammer, T.R. 2004. Bacteriophage defense systems and strategies for lactic acid bacteria. *Advances in Applied Microbiology, Vol 56* **56**: 331-378.
Sturino, J.M. and Klaenhammer, T.R. 2006. Engineered bacteriophage-defence systems in bioprocessing. *Nat Rev Microbiol* **4**(5): 395-404.
Svergun, D.I. 1992. Determination of the Regularization Parameter in Indirect-Transform Methods Using Perceptual Criteria. *Journal of Applied Crystallography* **25**: 495-503.
Tahallah, N., Pinkse, M., Maier, C.S., and Heck, A.J. 2001. The effect of the source pressure on the abundance of ions of noncovalent protein assemblies in an electrospray ionization orthogonal time-of-flight instrument. *Rapid Commun Mass Spectrom* **15**(8): 596-601.
Tang, J., Akerboom, J., Vaziri, A., Looger, L.L., and Shank, C.V. 2010. Near-isotropic 3D optical nanoscopy with photon-limited chromophores. *Proc Natl Acad Sci U S A* **107**(22): 10068-10073.
Tang, T.H., Bachellerie, J.P., Rozhdestvensky, T., Bortolin, M.L., Huber, H., Drungowski, M., Elge, T., Brosius, J., and Huttenhofer, A. 2002. Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus. *Proc Natl Acad Sci U S A* **99**(11): 7536-7541.
Tang, T.H., Polacek, N., Zywicki, M., Huber, H., Brugger, K., Garrett, R., Bachellerie, J.P., and Huttenhofer, A. 2005. Identification of novel non-coding RNAs as potential antisense regulators in the archaeon Sulfolobus solfataricus. *Mol Microbiol* **55**(2): 469-481.
Tock, M.R. and Dryden, D.T. 2005. The biology of restriction and anti-restriction. *Curr Opin Microbiol* **8**(4): 466-472.
Tomari, Y., Matranga, C., Haley, B., Martinez, N., and Zamore, P.D. 2004. A protein sensor for siRNA asymmetry. *Science* **306**(5700): 1377-1380.
Tosa, T. and Pizer, L.I. 1971. Biochemical bases for the antimetabolite action of L-serine hydroxamate. *J Bacteriol* **106**(3): 972-982.
Tyson, G.W. and Banfield, J.F. 2008. Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ Microbiol* **10**(1): 200-207.
Ueguchi, C., Ohta, T., Seto, C., Suzuki, T., and Mizuno, T. 1998. The leuO gene product has a latent ability to relieve bgl silencing in Escherichia coli. *J Bacteriol* **180**(1): 190-193.
Ueguchi, C., Suzuki, T., Yoshida, T., Tanaka, K., and Mizuno, T. 1996. Systematic mutational analysis revealing the functional domain organization of Escherichia coli nucleoid protein H-NS. *J Mol Biol* **263**(2): 149-162.
Unoson, C. and Wagner, E.G. 2008. A small SOS-induced toxin is targeted against the inner membrane in Escherichia coli. *Mol Microbiol* **70**(1): 258-270.
Urban, J.H. and Vogel, J. 2007. Translational control and target recognition by Escherichia coli small RNAs in vivo. *Nucleic Acids Res* **35**(3): 1018-1037.
van den Heuvel, R.H., van Duijn, E., Mazon, H., Synowsky, S.A., Lorenzen, K., Versluis, C., Brouns, S.J., Langridge, D., van der Oost, J., Hoyes, J., and Heck, A.J. 2006. Improving the performance of a quadrupole time-of-flight instrument for macromolecular mass spectrometry. *Anal Chem* **78**(21): 7473-7483.
van der Oost, J. and Brouns, S.J. 2009. RNAi: prokaryotes get in on the act. *Cell* **139**(5): 863-865.
van der Oost, J., Jore, M.M., Westra, E.R., Lundgren, M., and Brouns, S.J. 2009. CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends Biochem Sci* **34**(8): 401-407.
van der Ploeg, J.R. 2009. Analysis of CRISPR in Streptococcus mutans suggests frequent occurrence of acquired immunity against infection by M102-like bacteriophages. *Microbiology* **155**(Pt 6): 1966-1976.
van Heel, M., Gowen, B., Matadeen, R., Orlova, E.V., Finn, R., Pape, T., Cohen, D., Stark, H., Schmidt,

R., Schatz, M., and Patwardhan, A. 2000. Single-particle electron cryo-microscopy: towards atomic resolution. *Quarterly Reviews of Biophysics* **33**(4): 307-369.

Vartak, N.B., Liu, L., Wang, B.M., and Berg, C.M. 1991. A functional leuABCD operon is required for leucine synthesis by the tyrosine-repressible transaminase in Escherichia coli K-12. *J Bacteriol* **173**(12): 3864-3871.

Viswanathan, P., Murphy, K., Julien, B., Garza, A.G., and Kroos, L. 2007. Regulation of dev, an operon that includes genes essential for Myxococcus xanthus development and CRISPR-associated genes and repeats. *J Bacteriol* **189**(10): 3738-3750.

Volkov, V.V. and Svergun, D.I. 2003. Uniqueness of ab initio shape determination in small-angle scattering. *Journal of Applied Crystallography* **36**: 860-864.

Waghmare, S.P., Pousinis, P., Hornby, D.P., and Dickman, M.J. 2009. Studying the mechanism of RNA separations using RNA chromatography and its application in the analysis of ribosomal RNA and RNA:RNA interactions. *J Chromatogr A* **1216**(9): 1377-1382.

Wang, Y., Juranek, S., Li, H., Sheng, G., Wardle, G.S., Tuschl, T., and Patel, D.J. 2009. Nucleation, propagation and cleavage of target RNAs in Ago silencing complexes. *Nature* **461**(7265): 754-761.

Wiedenheft, B., Zhou, K., Jinek, M., Coyle, S.M., Ma, W., and Doudna, J.A. 2009. Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. *Structure* **17**(6): 904-912.

Williams, E., Lowe, T.M., Savas, J., and DiRuggiero, J. 2007. Microarray analysis of the hyperthermophilic archaeon Pyrococcus furiosus exposed to gamma irradiation. *Extremophiles* **11**(1): 19-29.

Winkler, J., Seybert, A., Konig, L., Pruggnaller, S., Haselmann, U., Sourjik, V., Weiss, M., Frangakis, A.S., Mogk, A., and Bukau, B. 2010. Quantitative and spatio-temporal features of protein aggregation in Escherichia coli and consequences on protein quality control and cellular ageing. *EMBO J* **29**(5): 910-923.

Wommack, K.E. and Colwell, R.R. 2000. Virioplankton: viruses in aquatic ecosystems. *Microbiol Mol Biol Rev* **64**(1): 69-114.

Wriggers, W., Milligan, R.A., and McCammon, J.A. 1999. Situs: A package for docking crystal structures into low-resolution maps from electron microscopy. *Journal of Structural Biology* **125**(2-3): 185-195.

Zaug, A.J., Linger, J., and Cech, T.R. 1996. Method for determining RNA 3' ends and application to human telomerase RNA. *Nucleic Acids Res* **24**(3): 532-533.

Zegans, M.E., Wagner, J.C., Cady, K.C., Murphy, D.M., Hammond, J.H., and O'Toole, G.A. 2009. Interaction between bacteriophage DMS3 and host CRISPR region inhibits group behaviors of Pseudomonas aeruginosa. *J Bacteriol* **191**(1): 210-219.

Zhou, M., Sandercock, A.M., Fraser, C.S., Ridlova, G., Stephens, E., Schenauer, M.R., Yokoi-Fong, T., Barsky, D., Leary, J.A., Hershey, J.W., Doudna, J.A., and Robinson, C.V. 2008. Mass spectrometry reveals modularity and a complete subunit interaction map of the eukaryotic translation factor eIF3. *Proc Natl Acad Sci U S A* **105**(47): 18139-18144.

Co-author affiliations

Stan J. J. Brouns, Magnus Lundgren[a], Edze R. Westra, Rik J. H. Slijkhuis, Marieke R. Beijer, Marieke Mastop, Luc van Heereveld, John van der Oost. Laboratory of Microbiology, Department of Agrotechnology and Food Sciences, Wageningen University, Dreijenplein 10, 6703 HB Wageningen, The Netherlands.

Ambrosius P. L. Snijders[b], Sakharam P. Waghmare, Mark J. Dickman. ChELSI Institute, Department of Chemical and Process Engineering, University of Sheffield, Mappin Street, Sheffield, S1 3JD, UK.

Kira S. Makarova, Eugene V. Koonin. National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.

Esther van Duijn, Arjan Barendregt, Albert J. R. Heck. Biomolecular Mass Spectrometry and Proteomics Group, Bijvoet Center for Biomolecular Research, Utrecht Institute for Pharmaceutical Sciences, and The Netherlands Proteomics Center, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands.

Jasper Akerboom, Jianyong Tang, Sean A. McKinney[c], Loren L. Looger. Janelia Farm Research Campus, Howard Hughes Medical Institute, Ashburn, VA 20147, USA.

Ümit Pul, Reinhild Wurm, Melina Mescher, Rolf Wagner. Institut für Physikalische Biologie, Heinrich-Heine-Universität Düsseldorf, Universitätsstr. 1, D-40225 Düsseldorf, Germany.

Thomas Stratmann, Karin Schnetz. Institute for Genetics, University of Cologne, Zülpicher Strasse 47, 50674 Cologne, Germany.

Jelle B. Bultema, Egbert J. Boekema. Department of Biophysical Chemistry, Groningen Biomolecular Sciences and Biotechnology institute, University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands.

Nadja Heidrich, Amanda Raine, E. Gerhart H. Wagner. Department of Cell and Molecular Biology, Uppsala University, Husargatan 3, SE-75124 Uppsala, Sweden.

Blake Wiedenheft, Kaihong Zhou, Jennifer A. Doudna. Howard Hughes Medical Institute, Department of Molecular and Cell Biology, University of California, and Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.

Present address:
[a]Department of Cell and Molecular Biology, Uppsala University, Husargatan 3, SE-75124 Uppsala, Sweden.
[b]MRC Clinical Sciences Centre, Faculty of Medicine, Imperial College London, Hammersmith Hospital Campus, Du Cane Road, London, W12 0NN, UK.
[c] The Stowers Institute, Kansas City, Missouri, USA.

Nederlandse samenvatting

Dit proefschrift beschrijft het onderzoek aan het antivirale en antiplasmide prokaryote CRISPR/Cas systeem uit *Escherichia coli*, behorend tot het Cse-subtype. Het CRISPR/Cas systeem bestaat uit clusters van repeterend DNA waartussen zich unieke stukjes DNA van een gelijke lengte (CRISPR) bevinden. In de buurt van dit repeterend DNA bevindt zich een groep van *cas* genen. Op basis van de compositie van de *cas* genen zijn er 8 subtypen te onderscheiden, waarvan er 4 experimenteel zijn onderzocht. Wat er tot nu toe bekend is over deze subtypen wordt beschreven in **Hoofdstuk 1**. De verschillende CRISPR/Cas systemen zijn variaties op een thema en werken volgens een mechanisme dat lijkt op RNA interferentie in eukaryoten wat ook wordt beschreven in dit hoofdstuk. Er zijn drie verschillende stadia te onderscheiden in CRISPR/Cas resistentie. Tijdens het eerste stadium wordt een stuk DNA van een virus of plasmide ingebouwd tussen twee repeterende eenheden DNA in de CRISPR (adaptatie). Tijdens het tweede stadium (expressie), wat beschreven wordt in **Hoofdstuk 2**, wordt CRISPR DNA afgeschreven in precursor CRISPR RNA (pre-crRNA), wat vervolgens verwerkt wordt door het metaal-onafhankelijke nuclease CasE. Laatstgenoemde knipt in elke repeterende eenheid van het RNA. Op deze wijze worden er functionele crRNAs gegenereerd die elk een 8 nucleotiden tellende 5' handle, een spacer en tot slot een haarspeldstructuur bevattende 3' handle bevatten. CasE en het crRNA zijn onderdeel van het Cascade complex waar ook CasA, B, C en D toe behoren. Cascade, geladen met crRNA en geassisteerd door Cas3, verhindert faag lambda infectie in *E. coli*. Met behulp van een kunstmatig gesynthetiseerde CRISPR met lambda fragmenten wordt dit in dit hoofdstuk gedemonstreerd in plaque assays. Zowel crRNAs complementair aan de coderende als aan de niet coderende streng van het lambda genoom zijn effectief. Dit suggereert dat DNA het doelwit is van dit systeem. Een mutatie in CasE, waardoor CasE geen pre-crRNA meer kan knippen, doet de resistentie teniet en laat zien dat crRNAs nodig zijn voor interferentie. Het mechanisme van herkenning van het doelwit, waarmee fase drie (interferentie) begint, wordt beschreven in **Hoofdstuk 3**. In dit hoofdstuk wordt gedemonstreerd dat Cascade dubbelstrengs DNA kan binden. De van oorsprong virale of plasmide sequentie van het crRNA baseparen kan vormen met het complementaire DNA door ATP-onafhankelijke strand displacement, een proces wat versterkt wordt door CasA. Dit hoofdstuk geeft ook een meer gedetailleerde massa- en sequentie-analyse van het crRNA. crRNA is 61 nucleotiden lang, wat in totaal overeenkomt met het van oorsprong virale stuk RNA en een repeterende eenheid RNA, en bevat een 2',3' cyclische fosfaat groep. De stoichiometrie van Cascade, $CasA_1B_2C_6D_1E_1$-$crRNA_1$, is bepaald met behulp van ESI-MS analyse. EM

en SAXS analyses laten zien dat Cascade de vorm van een zeepaard heeft, waarbij de ruggengraat gevormd wordt 6 CasC moleculen. Vergelijking van deze structuur met de structuren van de stabiele subcomplexen $CasB_2C_6D_1E_1$-$crRNA_1$ en $CasC_6D_1E_1$-$crRNA_1$ toont de locatie van de CasA en de twee CasB moleculen aan. Ten slotte wordt in dit hoofdstuk aangetoond dat alle subunits essentieel zijn voor resistentie. **Hoofdstuk 4** beschrijft hoe Cas eiwitten en een anti-lambda CRISPR de transformatie van een plasmide, met daarin dezelfde lambda sequentie als in de CRISPR, kunnen tegenhouden. Een bibliotheek van gemuteerde plasmiden, gemaakt met behulp van foutgevoelige PCR, is getransformeerd naar cellen die immuun waren voor het niet gemuteerde plasmide. De sequenties van de mutanten die het immuunsysteem konden omzeilen zijn te verdelen in drie klassen: de mutanten bevatten mutaties in het DNA complementair aan de crRNA, mutaties in het flankerende DNA motief (PAM) of deleties van het hele stuk wat aangevallen wordt. De PAM kan een herkenningspunt zijn voor het systeem zodat alleen 'vreemd' DNA wordt aangevallen. Mutaties in het complementaire stuk van het plasmide bevinden zich vooral aan de PAM-zijde. Deze data laat zien dat de PAM en perfecte baseparing aan de PAM-zijde in het aangevallen stuk DNA essentieel zijn voor resistentie.

Omdat lambda fagen *E. coli* vooral infecteren bij de polen en zich snel vermenigvuldigen, veronderstelden we dat het CRISPR/Cas afweer systeem hier ook gelokaliseerd zou zijn. Om dit te testen hebben we PALM gebruikt om op nanometerschaal te kijken waar de Cas eiwitten gelokaliseerd zijn (**Hoofdstuk 5**). Inderdaad bleken ten minste een aantal Cas eiwitten zich bij de celpool te bevinden, waardoor een snelle respons kan worden gestart. Daarbij kan de fysieke scheiding tussen de Cas eiwitten en de chromosomale CRISPR ook een mechanisme zijn om auto-immuniteit te voorkomen. **Hoofdstuk 6** ten slotte richt zich op de transcriptionele regulatie van het systeem. H-NS wordt geïdentificeerd als transcriptie- repressor; de expressie van de meeste *cas* genen (met uitzondering van *cas3*) en de CRISPR is verhoogd in een H-NS knockout mutant van *E. coli*. LeuO is juist een transcriptionele activator en antagonist van H-NS. Overexpressie van LeuO in een wildtype *E. coli* resulteert in verhoogde *casABCDE12* transcriptie en crRNA niveaus. Met behulp van plaque assays wordt tot slot aangetoond dat H-NS inderdaad het CRISPR/Cas systeem onderdrukt en LeuO juist dit systeem activeert.

Acknowledgements

En dan nu het leukste deel van het schrijven van dit boekje. Niet alleen is dit leuk omdat het betekent dat dit proces (bijna) afgerond is, maar vooral omdat ik iedereen, die direct dan wel indirect een aandeel heeft gehad in dit proefschrift, graag wil laten weten hoe dankbaar ik daarvoor ben.

Op de eerste plaats wil ik jou bedanken, John. Tijdens mijn stage in de VS stuurde je een mailtje met niet meer dan twee regels: “Wat ga je na je studie doen? Wij hebben misschien wel iets heel spannends voor je.” Ik twijfelde of ik na mijn studie wel in Wageningen wilde blijven. Maar zoals velen je enthousiasme niet konden weerstaan ging ook ik om. En “spannend” is het geworden! Dank voor je onovertroffen enthousiasme, je netwerken en het verschaffen van alle mogelijkheden voor het doen van leuk onderzoek. Het CRISPR project loopt door, met andere mensen en meer doorbraken in het verschiet. Ik denk er met veel plezier aan terug.

Stan, na al het succes wat we hebben behaald zou je bijna vergeten dat het begin van het project moeizaam verliep, het CRISPR systeem was eigenlijk een ‘black box’. Na een jaar begon het project te lopen, eigenlijk vanaf het moment dat we het Cascade complex ontdekten. Zonder jou zouden we nooit geweest zijn waar we nu staan. En ik waar ik nu sta. Je begeleiding, inzicht en ideeën maar ook gevoel voor humor, optimisme en gezelligheid zijn heel waardevol geweest en hebben een zeer belangrijk aandeel gehad in het succes van dit project. Voor ontspanning was ook ruimte, dat kon (en kan) door fietsen of bijvoorbeeld op een strandje chillen tijdens een zwaar congres;). Ik wens je alvast veel succes met je wetenschappelijke carrière, maar dat gaat helemaal goedkomen!

With so much to explore in the CRISPR field, the team has increased in size over time. Magnus, it was good having you here in the lab for two years. Great that you got all the assays to work. Besides, your enthusiasm and the way you enjoy life are very inspiring to me. Thinking about your cooking skills makes me feel hungry again. Going to the filmfestivals in Rotterdam was great fun. All the best in Uppsala! Edze, je inzichten, enthousiasme en doorzettingsvermogen zijn een aanwinst geweest voor het team. Ook je sambavoetbal op amper vier vierkante meter is een aanwinst voor de groep. Succes met het forceren van de volgende wetenschappelijke grote doorbraak! Faab, allereerst dank voor het tolereren van de foute Duitse grappen die ik (en anderen met mij) op je af konden vuren. Gelukkig was de humor zo nu en dan ook nog wel

van een minder bedenkelijk niveau. Je doorzettingsvermogen en kennis van eigenlijk alles wat met micro-organismen te maken heeft zijn bewonderenswaardig. Ondanks de zogenaamde rivaliteit tussen onze projecten heb je toch heel wat keren voor me uitgeplaat in de avonduren, waarvoor natuurlijk eeuwige dank. Gelukkig dat het zich nu terugbetaalt in een plek voor jou in het CRISPR team, daar komt vast iets moois uit. Tof dat je mijn paranimf wilt zijn. Ik hoop dat je het voor elkaar krijgt om in Engeland aan de slag te kunnen.

I would also like to thank all our collaborators, without you this project would definitely not have been as successful. Especially I would like to thank Jaapie and Loren for letting me work on the cellular localization of the CRISPR machinery in your “cool” laboratory. Jianyong and Sean, thanks for your PALM contributions. It’s great that the work on this aspect of CRISPR/Cas is still ongoing, and it will hopefully result in a nice paper. Kira and Eugene, thanks so much for your bioinformatical support and sharing your data before publication, which really helped the project. Bram, thanks for sequencing all the protein bands (it must have been more than 100 pieces in total…). And more than that, it is great that the collaboration is extended, for instance by sequencing the crRNA together with Sak and Mark. I am curious what the ongoing research will lead to. Thanks also to Ümit, Reini, Melina and Rolf for your expertise and experiments on the H-NS story. You really initiated the research on this topic in Wageningen, leading to a nice publication. I would also like to thank Nadja, Amanda and Gerhart from Sweden and Thomas and Karin from Germany for their nice contributions to this story. Blake, Kaihong and Jennifer, thanks for the nice collaboration on the crystallization and SAXS on Cascade. Although the SAXS worked out really nice, I hope that the crystallization efforts will someday succeed! Jelle en Egbert, dank voor alle EM werk zodat die `onzichtbare’ eiwitcomplexen nog meer vorm kregen. Esther, Arjan en Albert, tof dat jullie met een aantal creatieve methoden de stoichiometrie van dit complex hebben weten op te lossen.

AIO zijn was zeker ook leuk door het begeleiden van en samenwerken met studenten. Dank Marieke (B. en M.), Stineke, Tijmen en Man voor jullie inzet in dit project en de gezelligheid! Ook bedankt Luc, Rik, Sarah, Nadine en Kay voor jullie bijdragen. Veel succes voor jullie in de toekomst.

Being a PhD student in the Laboratory of Microbiology is such a nice experience. One of the definite highlights was participating in or even better, organizing the PhD trip. Thanks Thomas for keeping the cash flow go, John for feeling comfortable at the YMCA in Newark, Marcel for finding out where to go for bird watching and Faab for

asking all the 100+ questions at the laboratories that we visited. Arranging such a cool event together with you was a great experience! Having fun just happened on a daily basis in our office. Thanks Ans, Marke, Jasper W., Jaapie, Harmen, Pawel, Bart, Faab, Magnus, Stan and Edze for all the `gezelligheid'. Also thanks to all the other people at Bacgen: Servé, Mark M., Mark L., Rie (your sushi is amazing), John R, Nicolas, Marcel, Bram, Sjon, Katrin, Hao, Suzanne, Marco, Pierpaolo, Wouter, Sinan and all the other people who spent time here. But of course, thanks to all the other people in the laboratory of Microbiology: it has always been a great pleasure being in this group that next to building a good atmosphere in the lab and coffee corner, enthusiastically participates in all sorts of events such as the Veluweloop. Mijn dank gaat ook uit naar het ondersteunende team van Micro: Wim, Jannie, Nees, Anja en Renée, zonder jullie zou het op zijn zachtst gezegd een rommeltje worden bij Micro!

Afstand nemen van je werk is soms nodig. Ivo, in moeilijke tijden is het altijd goed een biertje met je te drinken! In goede tijden trouwens ook. Leuk dat je mijn paranimf wilt zijn. Dré, bedankt ook voor het zo nu en dan gezellige aanschuiven. Jessica, Ruud, Karin en Lidwien als 'AIO-lotgenoten' en Dennis, het dagelijks leven evalueren voor het spelletjesfestijn gewoonlijk aanving was waardevol en heeft voor de nodige pret gezorgd. Bart, gesprekken met jou resulteren over het algemeen in een gezonde dosis relativering. Mijn huisgenoten Linette, Sina, Rik, Klaske, Tim, Mieke en Eelke hebben ervoor gezorgd dat ik een leuke tijd heb gehad op de Hoogstraat, en dat het fijn thuiskomen was na mijn werk. Dank voor jullie gezelligheid en gesprekken van alle niveaus. Ook mijn oudhuisgenoten van de Rijnsteeg wil ik graag bedanken. Dat wat ons in eerste instantie bond bestaat niet meer, maar er is iets moois voor in de plaats gekomen. NS wandelingen en kerstdiners houden we wat mij betreft vol tot onze oude dag. Lieve Lobbers, dank voor de sportieve ontspanning en gezelligheid. In het bijzonder Jaap, Annemieke, Pauline, Peter, Bas, Susan, Tobias, Marjon, Nonja, Harmke, maar ook zeker alle anderen. En natuurlijk oud-Lobbers André, Douwe en Moniek, Het is leuk om elkaar nog steeds te zien, nu dan in een andere setting. Hans, het bierdrinken en luisteren naar stampende muziek was hoognodig en heeft de batterij op het laatst nog even opgeladen.

De ultieme manier om te onthaasten, en dat was nodig de laatste tijd, is een goede film draaien en kijken op Movie-W. Dolf en Joost, dank voor jullie inzet voor het mooiste en knuste filmhuis van Wageningen en verre omstreken. Ook alle andere Movie-W'ers, bedankt voor jullie bijdrage aan die goede sfeer!

Gelukkig heb ik een familie die wel houdt van een beetje gezelligheid, wat geïllustreerd wordt door trouwfeesten in een Noordhollands weiland en in Roemenië, verjaardagen en een jaarlijks weekendje weg. Dank voor al die broodnodige ontspanning, net als jullie interesse tijdens de afgelopen jaren. En Oma, gezellig al die telefoongesprekken, die houden we er in. Lieve overzeese familie, Ray en Ineke, Paul en Justine, en Jeff. Het onderzoek tijdens mijn stage in Seattle heeft ons weer dichter bij elkaar gebracht. Alle congressen in de VS die goed (of minder goed) te combineren waren met een bezoekje aan Seattle, dan wel een afspraak in Boston, hebben daar gelukkig goed aan bij kunnen dragen. Dank voor jullie steun, gehoor en pret. Paul en Elise, Michiel en Annemarie veel dank ook voor jullie belangstelling in het wel en wee van een AIO en de gezelligheid de afgelopen tijd.

Jan en Ans, jullie hebben mijn school- en studiegang altijd gesteund. Dat heeft ertoe geleid dat ik dit onderzoek kon doen. Leuk Jan, dat je de inhoud begrijpt en op waarde kan schatten. Ans, ondanks dat het onderzoek voor jou abracadabra is, heb je me altijd gesteund in wat ik deed. Ontzettend bedankt dat jullie altijd voor me klaar hebben gestaan!

Lieve Susanne, een promotietraject doorlopen en een relatie onderhouden gaan niet altijd goed samen, maar het is toch gelukt. En dat komt vooral jou! Fijn dat je me psychisch gesteund hebt en altijd klaarstond om te helpen: Je hebt op de leesversie-deadline-dag nog alle referenties ingevoerd en later nog een kaft bij elkaar gephotoshopt. Zo is het ook een beetje jouw boekje geworden. Samenzijn met jou is fijn, laten we daarvan genieten.

Matthijs

About the author

Matthijs Miklas Jore was born on the 15th of June, 1981, in Delft, The Netherlands. He started his education at the public primary school "'t Vlot" in Gouda. This was followed by secondary education at the Coornhert Gymnasium in the same town. After graduating in 1999, he chose the multidisciplinary study of Molecular Sciences at Wageningen University. His main field of interest was the biological part of the study. Therefore he did his major Master's thesis at the Laboratory of Nematology at Wageningen University, under the direction of Professor Arjen Schots. During this research he studied protein-protein-interactions of the resistance protein Rx in potato, using phage display. His minor Master's thesis followed, at the Laboratory of Microbiology under supervision of Professor John van der Oost. Here, he studied and characterized surface-layer proteins from thermophilic bacteria. To complete his studies, he did an internship at the University of Washington, Seattle, USA, at the group of Professor Wim Hol. During his internship, he crystallized proteins and protein complexes of *Trypanosoma brucei*, the causing agent of sleeping sickness. The aim was to design protein inhibitors based on the obtained crystal structures. After graduating, he started his PhD at the department of Microbiology in Wageningen, under supervision of Professor John van der Oost and Dr. Stan Brouns. During this research, first insights into the antiviral CRISPR system of *Escherichia coli* were obtained. Subsequently, he started as a post-doc in the same group. His aim: clarify the (possibly antiviral) function of the Argonaute protein in prokaryotes.

List of publications

Brouns, S.J.*, **Jore, M.M.***, Lundgren, M., Westra, E.R., Slijkhuis, R.J., Snijders, A.P., Dickman, M.J., Makarova, K.S., Koonin, E.V., van der Oost, J. 2008. Small CRISPR RNAs guide antiviral defense in prokaryotes. Science 321: 960-4.

van der Oost, J., **Jore, M.M.**, Westra, E.R., Lundgren, M., Brouns, S.J. 2009. CRISPR-based adaptive and heritable immunity in prokaryotes. Trends Biochem Sci. 34 (8): 401-7.

Jore, M.M., van der Oost, J., Brouns, S.J. 2010. RNA in Defense: CRISPRs protect prokaryotes against mobile genetic elements. RNA Worlds: From Life's Origins to Diversity in Gene Regulation. Cold Spring Laboratory Press. *In press.*

Westra, E.R.*, Pul, Ü.*, Heidrich, N., **Jore, M.M.**, Lundgren, M., Stratmann, T., Wurm, R., Raine, A., Mescher, M., van Heereveld, L., Mastop, M., Wagner, E.G.H., Schnetz, K., van der Oost, J., Wagner, R., Brouns, S.J.J. H-NS mediated repression of CRISPR-based immunity in *Escherichia coli* K12 can be relieved by the transcription activator LeuO. Molecular Microbiology. *In press.*

Jore, M.M.*, Lundgren, M.*, Van Duijn, E.*, Bultema, J.B.*, Westra, E.R., Waghmare, S.P., Wiedenheft, B., Ümit Pul, Reinhild Wurm, Rolf Wagner, Beijer, M.R., Barendregt, M., Zhou, K., Snijders, A.P., Dickman, M.J., Doudna, J.A., Boekema, E.J., Heck, A.J.,van der Oost, J., Brouns, S.J. Structural basis for CRISPR RNA-guided recognition of DNA by Cascade. *Submitted.*

Westra, E.R., **Jore, M.M.**, Sinan Al-Attar, Brouns, S.J., van der Oost, J. Small RNAs in Bacteria. *Submitted.*

*contributed equally

Overview of completed training activities

Discipline specific activities

Meetings:

- Annual Meeting Study Group Molecular Genetics. 2006. Lunteren, NL.**
- Annual Meeting Study Group Protein Research. 2006. Lunteren, NL.
- RNA Silencing: Mechanism, Biology and Application. 2007. Keystone CO, USA.**
- Annual Meeting Study Group Molecular Genetics. 2007. Lunteren, NL.
- Annual Meeting Study Group Protein Research and Nucleic Acids. 2007. Lunteren, NL.*
- Bijvoet Tutorial Symposium. 2008. Utrecht, NL.
- Molecular Genetics of Bacteria and Phages. 2008. Cold Spring Harbor NY, USA.*
- Annual Meeting Study Group Protein Research and Nucleic Acids. 2008. Veldhoven, NL.**
- General Meeting of the American Society for Microbiology. 2009. Philadelphia PA, USA.*
- Annual Meeting Protein Research, Nucleic Acids, and Lipids & Biomembranes. 2009. Veldhoven, NL.
- FASEB Conference on Nucleic Acid Enzymes. 2010. Saxtons River VT, USA.*

*Oral presentation
**Poster presentation

Courses:

- Guest scientist - Hands on course Photo-Activated Light Microscopy. 2009. Janelia Farm Research Campus, HHMI, Ashburn VA, USA.
- Protein Interactions and Networks. 2009. Wellcome Trust Genome Campus, Hinxton, UK.

General courses

- Supervising Undergraduate Students. 2006. Wageningen, NL.
- VLAG PhD Week. 2007. Bilthoven, NL.
- Philosophy and Ethics of Food Science and Technology. 2009. Wageningen, NL.
- Scientific writing. 2009. Wageningen, NL.

Optionals

- Preparing PhD Research Proposal.
- PhD trip 2006. California, USA.
- PhD trip 2009. Northeast Coast, USA.
- Bacterial Genetics Group Meetings (weekly).
- Laboratory of Microbiology PhD/PostDoc Meetings (biweekly).

The research described in this thesis was financially supported by a VICI-grant from the Netherlands Organization for Scientific Research to Prof. dr. John van der Oost (grant number 865.05.001)

Cover: Optical table of a Photo-Activatable Localization Microscope

Printed by: GVO Drukkers & Vormgevers BV, Ede

Printing of this thesis was financially supported by:
BaseClear BV, Leiden
GENEART AG, Regensburg, Germany