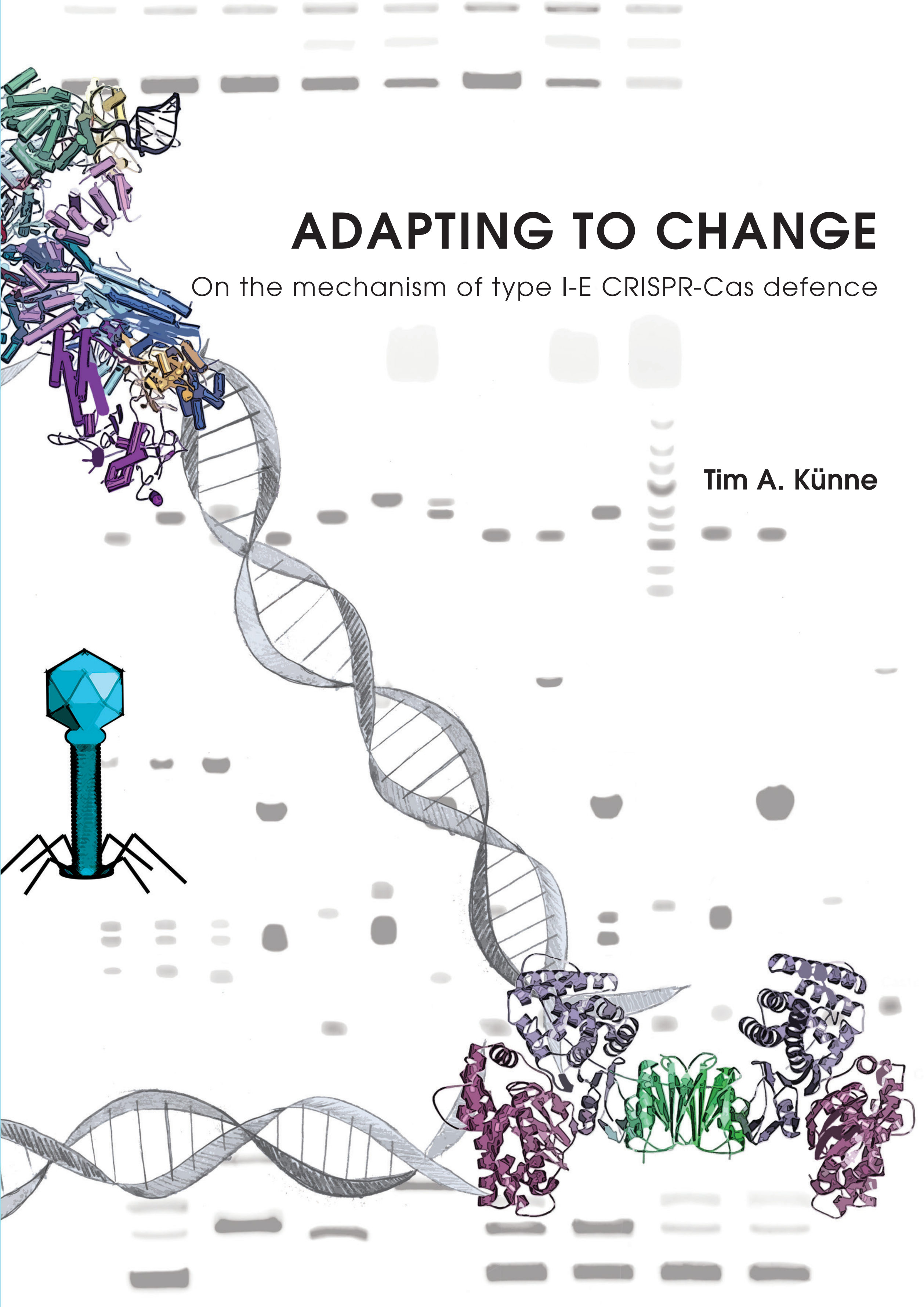


ADAPTING TO CHANGE

On the mechanism of type I-E CRISPR-Cas defence

Tim A. Künne

Adapting to Change

On the mechanism of type I-E CRISPR-Cas defence

Tim A. Künne

Thesis committee

Promotor

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

Co-promotor

Dr Stan J.J. Brouns
Associate Professor Molecular Microbiology
Delft University of Technology

Other members

Prof. Dr Dolf Weijers, Wageningen University & Research
Dr Blake Wiedenheft, Montana State University
Dr Chirlmin Joo, Delft University of Technology
Dr Rogier Louwen, Erasmus MC

This research was conducted under the auspices of the Graduate School VLAG (Advanced studies in Food Technology, Agrobiotechnology, Nutrition and Health Sciences).

Adapting to Change

On the mechanism of type I-E CRISPR-Cas defence

Tim A. Künne

Thesis

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Wednesday 4 October 2017
at 4 p.m. in the Aula.

Tim A Künne
Adapting to Change - On the mechanism of type I-E CRISPR-Cas defence,
240 pages.

PhD thesis, Wageningen University, Wageningen, NL (2017)
With references, with summary in English

ISBN: 978-94-6343-664-9
DOI: 10.18174/421025

Table of Contents

Chapter 1

General introduction and thesis outline

General introduction

Host pathogen interaction and co-evolution

Interactions between host organisms and their pathogens are among the most prevalent and evolutionary important biological interactions known today. Although a matter of life and death, these interactions are also drivers of evolution and have shaped the biodiversity of our planet (Filee et al., 2003; Koonin, 2016; Koonin and Dolja, 2013; Stern and Sorek, 2011). The most prevalent host-pathogen interaction in nature is the predation of prokaryotic microorganisms, especially bacteria, by their viruses. These so-called bacteriophages (phages) were discovered at the beginning of the last century by Twort in 1915 and independently by d'Herelle in 1917 (Summers, 2011). However, it took another 60-70 years to fully appreciate the abundance and the impact of these viruses. Throughout the 80s ever higher titres of phages were reported in natural environments, ranging from 10^4 to 10^7 ml^{-1} (Bitton, 1987; Torrella and Morita, 1979). Another ten years later phages were found to considerably outnumber bacteria in aquatic systems (10:1), where they cause significant prokaryotic death (Bergh et al., 1989; Lutze and Ewing, 1990; Proctor et al., 1988; Proctor and Fuhrman, 1990; Sieburth et al., 1988). Today, we consider phages the most abundant and diverse biological entities on the planet (Koonin and Dolja, 2013; Suttle, 2007). The remarkable diversity of viruses reflects their high mutation rates, which cause rapid evolution when facing selective pressure from antiviral barriers (Labrie et al., 2010). Viruses and their prokaryotic hosts have been found in almost every possible ecosystem, being involved in perpetual cycles of co-evolution (Samson et al., 2013). During this constant arms race, emerging virus-resistant hosts will become the dominant lineages, while viruses are forced to develop counter resistance. This reciprocal selection and development of prokaryotic defence mechanisms against their viral predators plays a key role in regulating populations in most ecological niches, rendering viral resistance a crucial survival phenotype.

Influence of mobile genetic elements

Next to phages, there are other mobile genetic elements (MGE) that can prove a burden to prokaryotic cells. Plasmids are transmissible between cells and require energy and resources for their replication. Often plasmids also encode for proteins, which increases their metabolic burden. Naturally occurring plasmids usually carry systems that ensure their stable replication within cells, such as toxin anti-toxin systems or antibiotic resistance systems (Jalasvuori and Koonin, 2015). Apart from being a burden on the cell, plasmids can also bring benefits. Antibiotic resistance genes on plasmids can be very beneficial for cells that are exposed to antibiotics in their environment. This has become particularly troublesome in today's medicine where the increasing antibiotic resistance reservoir has led to the development of microbial pathogens that persist in hospitals due to the resistance they gained to most known antibiotics.

Also viruses can be beneficial for their host organism. Temperate viruses insert their genetic material into the host genome (prophage/provirus), thereby transferring their genetic information to the host. Some viral genes can give a fitness advantage to the host, e.g. by protecting the host from additional viral infection (superinfection exclusion) or providing virulence factors to pathogenic hosts (Boyd and Brussow, 2002; Brussow et al., 2004; Labrie et al., 2010). Prophages are often stably maintained in bacterial lineages over long time spans, or, in case of inactivation of the provirus by mutations, the host has the opportunity to stably fix the newly acquired traits in its lineage (Canchaya et al., 2003). It has been shown for example that elimination of all prophages in *E. coli* reduces bacterial fitness (Wang et al., 2010).

In conclusion, MGEs can contribute to prokaryote evolution via two basic principles: the arms race that requires the cells to undergo major innovation cycles, or the direct transfer of innovative genetic information from the MGE to the cell (Koonin, 2016). Despite the occasional benefits, prokaryotes counteract the threat of mobile genetic elements through a set of fast evolving defence strategies that may act at virtually all stages of the invaders' life cycles. These defence systems encompass (i) blocking, modification and loss of phage receptors, (ii) production of extracellular matrix, (iii) superinfection exclusion (Sie) systems, (iv) restriction-modification (R-M) systems, (v) Argonaute-based (Ago) defence, (vi) toxin-antitoxin (TA) systems, (vii) abortive infection (Abi) systems, (viii) the bacteriophage exclusion (BREX) system, and (ix) the clustered regularly interspaced short palindromic repeats and associated genes (CRISPR-Cas) system (Goldfarb et al., 2015; Labrie et al., 2010; Swarts et al., 2014; Westra et al., 2012b). Most of these systems function by innate immunity, therefore being non-specific and non-adaptive. It was thought for a long time, that prokaryotes do not possess an adaptive and targeted immune system. This changed with the discovery of the CRISPR-Cas immune system, which provides bacteria and archaea with adaptive and heritable immunity against MGEs.

Brief History of CRISPR-Cas

The story of CRISPR began already in 1987 when Ishino and colleagues discovered the typical repeat array downstream of the *E. coli iap* gene (Ishino et al., 1987; Nakata et al., 1989). Subsequently, similar repeat arrays were discovered in a range of prokaryotes, both bacteria and archaea (Groenen et al., 1993; Hoe et al., 1999; Masepohl et al., 1996; Mojica et al., 1995; van Embden et al., 2000). However, there was only speculation on the function of these arrays and even the systematic discovery of similar arrays in many published genome sequences did not reveal a function (Jansen et al., 2002b; Mojica et al., 2000). The name CRISPR was first introduced by Jansen and colleagues, who also discovered that CRISPRs are often co-localizing with specific CRISPR associated (*cas*) genes (Jansen et al., 2002a). In 2005 three groups independently discovered that many of the spacer sequences in between the repeats matched the sequences of MGEs, suggesting they were derived from MGEs (Bolotin et al., 2005; Mojica et al., 2005; Pourcel et al., 2005). Finally, the detection of defined CRISPR array transcripts and the *in silico* prediction of *cas* gene functions led to the hypothesis that CRISPR-Cas acts as an RNA-based interference system similar to RNAi (Makarova et al., 2006). This hypothesis kicked off the molecular research into the function of CRISPR-Cas, which soon thereafter confirmed the function as a small RNA-based defence system (Barrangou et al., 2007; Brouns et al., 2008a).

CRISPR-Cas, an overview

CRISPR-Cas systems are best known for their function as an adaptive immune system, although recent studies suggest a number of non-defence roles (Vercoe et al., 2013; Westra et al., 2014). Non-defence roles include endogenous gene regulation, often coupled to virulence of bacterial pathogens (Gunderson and Cianciotto, 2013; Louwen et al., 2013; Louwen et al., 2014; Sampson et al., 2014; Sampson et al., 2013; Toledo-Arana et al., 2009), regulation of fruiting body formation in *Myoxococcus xanthus* (Viswanathan et al., 2007) and regulation of group behaviour, such as biofilm formation, in *Pseudomonas aeruginosa* (Zegans et al., 2009).

Today, CRISPR-Cas systems are still exclusively found in prokaryotes and are present in around 50 % of sequenced bacterial and 85 % of sequenced archaeal genomes (Grissa et al., 2007). However, a recent study analysing the genomes of unculturable microbes from environmental samples revealed a CRISPR-Cas abundance of only 9-10% in bacteria and archaea (Burstein et al., 2016). This might suggest that the real abundance of CRISPR-Cas systems in nature is lower than previously thought.

A typical CRISPR array consists of repeating sequences of usually around 30 bp length and these are interspaced by similarly sized spacer sequences (Figure 1). Often spacers can be mapped to MGEs, but spacers matching the host genome (self-targeting spacers) can also be observed at low frequencies (Stern et al., 2010). The number of spacers in CRISPR arrays varies from just a couple to dozens, or (rarely) several hundreds. The CRISPR array is preceded by a leader sequence. This AT rich sequence contains promotor elements to drive transcription of the array and it contains binding sites for other regulatory elements (Hale et al., 2012; Lillestol et al., 2009; Pougach et al., 2010; Pul et al., 2010). The CRISPR proximal end of the leader is involved in the spacer integration process, as will be discussed below.

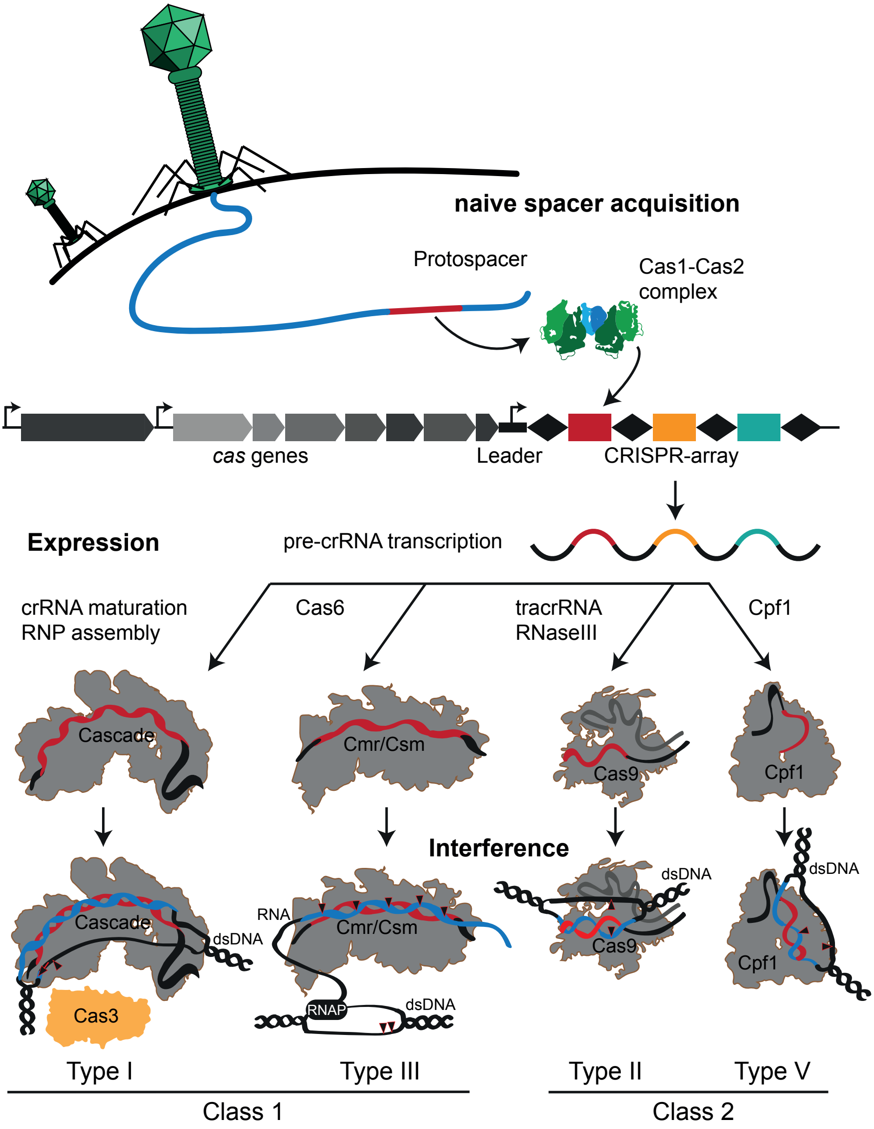


Figure 1 – Three stages of CRISPR-Cas defense. Schematic overview of the three stages in CRISPR-Cas immunity of the four most prominent systems (type I, type II, type III, type V). From top to bottom: 1) Phage or plasmid DNA enters the cell. A protospacer is selected by the acquisition machinery, including the Cas1-2 complex, and integrated as a new spacer in the CRISPR array. 2) CRISPR array is transcribed into pre-crRNA from the leader and *cas* genes are expressed. The pre-crRNA is processed into mature crRNA by Cas proteins and host factors as indicated, crRNA assembles with the effector proteins to form RNA guided effector complexes. 3) RNA guided effector proteins detect invaders by base pairing of the guide RNA with the target DNA. The target DNA gets cleaved (indicated by black and red triangles) by the effector complex or Cas3 (type I).

Briefly, CRISPR-Cas functions in three stages (Figure 1): (1) the adaptation stage, during which new memory is added to the CRISPR array in form of spacers; (2) the expression stage, when the *cas* genes are expressed and the CRISPR array is transcribed into precursor CRISPR RNA (pre-crRNA) that is processed into mature crRNAs by Cas proteins and/or host factors. The mature crRNA associates with the effector protein-(complex) to form an effector ribonucleoprotein complex (RNP); (3) the interference stage, during which the effector-RNP scans the nucleic acids in the cell for potential invaders. Upon a match of crRNA and target nucleic acid, the target is cleaved or degraded by the RNP or an accessory nuclease.

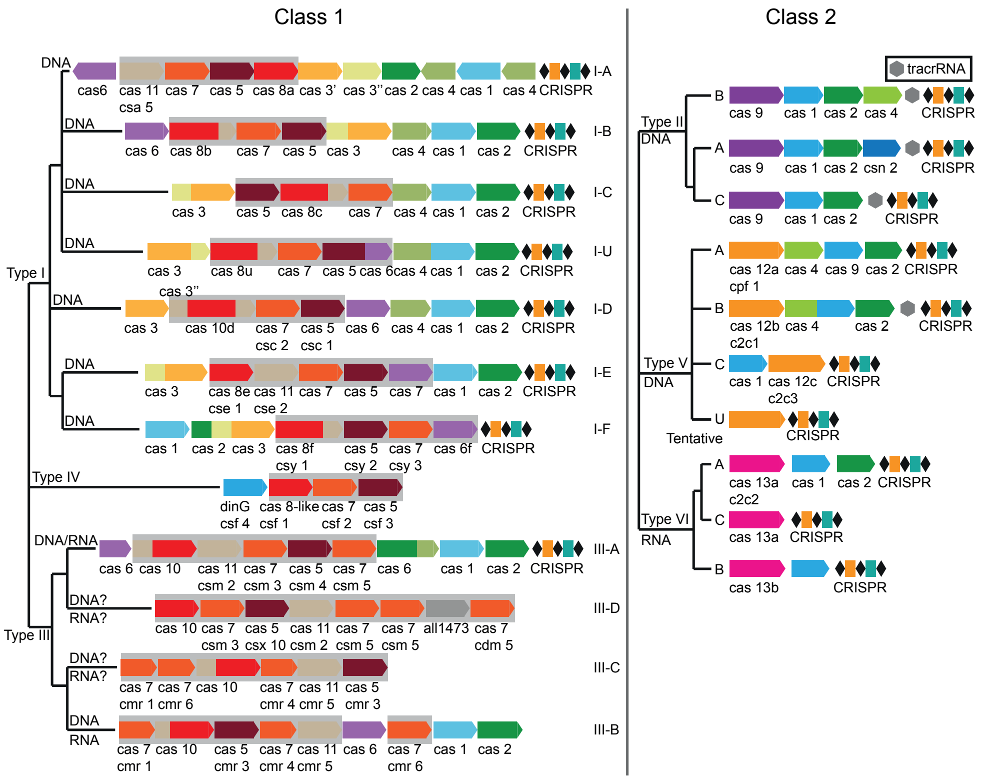

Class 1
Class 2
DNA
cas6 cas 11 cas 7 cas 5 cas 8a cas 3' cas 3'' cas 2 cas 4 cas 1 cas 4 CRISPR
csa 5
I-A
DNA
cas 6 cas 8b cas 7 cas 5 cas 3 cas 4 cas 1 cas 2 CRISPR
I-B
DNA
cas 3 cas 5 cas 8c cas 7 cas 4 cas 1 cas 2 CRISPR
I-C
DNA
cas 3 cas 8u cas 7 cas 5 cas 6 cas 4 cas 1 cas 2 CRISPR
I-U
Type I
cas 3''
DNA
cas 3 cas 10d cas 7 cas 5 cas 6 cas 4 cas 1 cas 2 CRISPR
csc 2 csc 1
I-D
DNA
cas 3 cas 8e cas 11 cas 7 cas 5 cas 7 cas 1 cas 2 CRISPR
cse 1 cse 2
I-E
DNA
cas 1 cas 2 cas 3 cas 8f cas 5 cas 7 cas 6f CRISPR
csy 1 csy 2 csy 3
I-F
Type IV
dinG cas 8-like cas 7 cas 5
csf 4 csf 1 csf 2 csf 3
DNA/RNA
cas 6 cas 10 cas 11 cas 7 cas 5 cas 7 cas 6 cas 1 cas 2 CRISPR
csm 2 csm 3 csm 4 csm 5
III-A
DNA?
RNA?
cas 10 cas 7 cas 5 cas 11 cas 7 cas 7 all1473 cas 7
csm 3 csx 10 csm 2 csm 5 csm 5 cdm 5
III-D
Type III
DNA?
RNA?
cas 7 cas 7 cas 10 cas 7 cas 11 cas 5
cmr 1 cmr 6 cmr 4 cmr 5 cmr 3
III-C
DNA
RNA
cas 7 cas 10 cas 5 cas 7 cas 11 cas 6 cas 7 cas 1 cas 2
cmr 1 cmr 3 cmr 4 cmr 5 cmr 6
III-B
tracrRNA
B
cas 9 cas 1 cas 2 cas 4 CRISPR
Type II
DNA
A
cas 9 cas 1 cas 2 csn 2 CRISPR
C
cas 9 cas 1 cas 2 CRISPR
A
cas 12a cas 4 cas 9 cas 2 CRISPR
cpf 1
B
cas 12b cas 4 cas 2 CRISPR
c2c1
Type V
DNA
C
cas 1 cas 12c CRISPR
c2c3
U
Tentative
CRISPR
A
cas 13a cas 1 cas 2 CRISPR
c2c2
Type VI
RNA
C
cas 13a CRISPR
B
cas 13b CRISPR

Figure 2 – Overview of known CRISPR-Cas systems. Adapted from (Makarova et al., 2017a, b). Schematic overview of the genomic loci of all currently identified CRISPR-Cas systems, showing the typical operon organization. The target nucleic acid of each system is indicated on the left. Genes are color-coded by homology and gene family names are given, in some cases the common name is given as well. Class 1 systems encode multisubunit effector protein complexes, while Class 2 systems encode single effector proteins. The beige background highlights genes encoding parts of the effector complexes. Genes with multiple colors are multidomain genes that can occur separately.

CRISPR-Cas classification

The *cas* genes are very diverse and form the basis of the classification system used to distinguish CRISPR-Cas systems. This classification evolves as new genomic information becomes available and new systems are being discovered. The most recent classification distinguishes class 1 and class 2 systems, which differ very much with respect to the architecture of their RNA-guided effector proteins (Figure 2) (Makarova et al., 2015; Makarova et al., 2017a, b). Class 1 systems utilize multiprotein complexes, while class 2 systems utilize single proteins. There are currently 6 types and >25 subtypes of CRISPR-Cas systems, with type I, type III and type IV being class 1 systems, and type II, type V and type VI being class 2 systems (Koonin et al., 2017). *Cas* genes can be divided by their functional association with the three stages of CRISPR-Cas. The adaptation module, including *cas1*, *cas2* and sometimes *cas4*, is the most uniform/conserved functional module between the diverse CRISPR-Cas systems (Jackson et al., 2017). The expression module is represented by *cas6* genes in type I and type III systems, which are essential for crRNA maturation. Type II systems make use of host RNase III, a trans-acting RNA (tracrRNA) and other unknown factors, while type V and type VI systems integrate crRNA processing into the effector protein. The interference module contains the effector genes. These are single genes in case of type II, type V and type VI systems (*cas9, cas12 (cpf1) and cas13 (c2c2)* respectively), or multigene operons for type I, type III and type IV systems. The latter operons encode the multiprotein effector complexes, which share striking similarities in their structural architecture, despite a low sequence similarity (e.g. Cascade in type I systems and CSM/CMR in type III systems) (Figure 3B). Common genes are *cas7*, which encodes the backbone protein of the effector complex, *cas5*, a small subunit and a large subunit. In some cases the *cas6* gene products are also part of the final effector complex, however, Cas6 function is often independent of the complex. Type I systems are the only systems with an additional effector gene, namely *cas3*, which is a helicase-nuclease responsible for target DNA degradation. The Cas3 protein is not part of the effector complex and is recruited *in trans* after target DNA recognition by Cascade. All other systems incorporate the nuclease functionality within their effector complex. The majority of CRISPR-Cas systems target DNA, however, type III systems target RNA sequence specifically and in addition degrade DNA non-specifically in an RNA-target dependent manner, while type VI systems target RNA only.

In the following paragraphs CRISPR-Cas mechanisms will be explained in more detail. The focus will be on type I systems and more specifically the type I-E system of *E. coli*. It is one of the best understood and most abundant CRISPR-Cas type in nature and is the focus of this thesis.

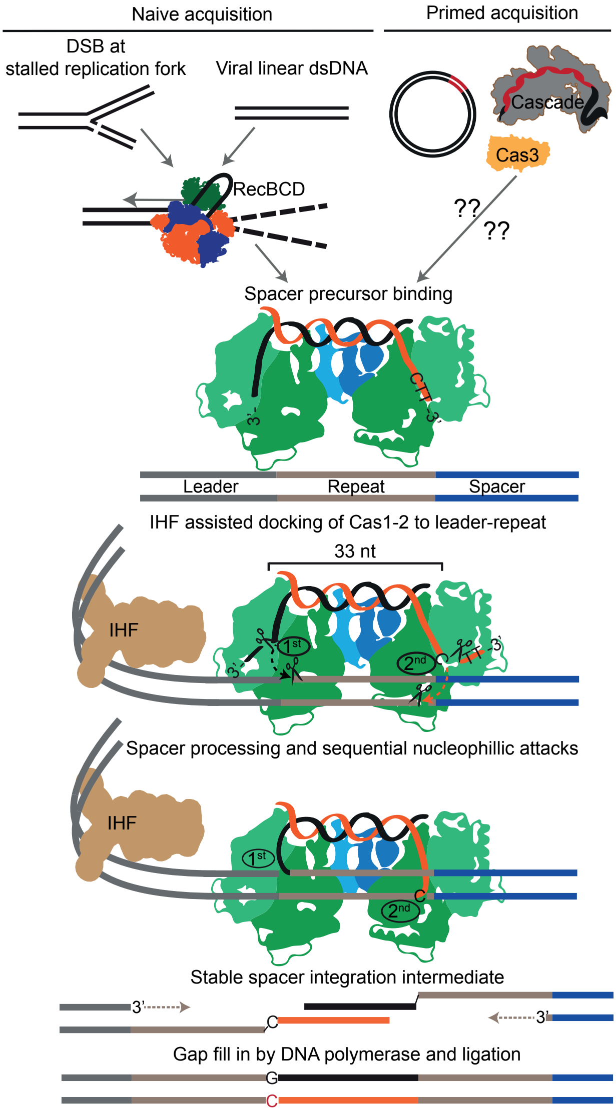


Figure 3 – Mechanism of spacer integration. Schematic of the current model of spacer integration in the type I-E system. The general mechanism likely applies to other systems as well. From top to bottom: Spacer precursors are generated via multiple pathways. During naïve acquisition precursors are thought to be generated dominantly by the RecBCD complex during DSB repair at stalled replication forks during degradation of linear dsDNA (e.g. phage DNA). During primed acquisition, precursors are thought to be generated by any combination of the Cascade complex, Cas3 and the Cas1-2 complex. Next, precursors are selected, bound and processed by the Cas1-2 complex to generate mature spacers containing a canonical PAM (CTT) in the 3' end of one strand. Integration host factor (IHF) binds to the leader and induces a bend, which allows proper docking of the Cas1-2 complex. The PAM on the precursor is cleaved to generate a C on the 3' end at some point during spacer integration. The 3'-OH of the spacer precursor carries out a Cas1-2 catalyzed nucleophilic attack on the first repeat of the array. This happens sequentially, first at the leader proximal end and then at the penultimate base of the leader distal end. This forms a stable spacer integration intermediate. Next, the remaining gaps are filled in by DNA polymerase and repaired by ligation.

CRISPR adaptation

The adaptation stage was for a long time the least understood, but recently a lot of progress has been made (Jackson et al., 2017). Adaptation is the integration of new spacer sequences into the CRISPR array during or after exposure to MGEs. Spacers are usually sampled from MGEs but occasionally also from the host genome. Integration occurs at the leader proximal end of the array and therefore spacers represent a chronological archive of encountered invaders (Barrangou et al., 2007; Garneau et al., 2010; Sternberg et al., 2016).

Self/non-self discrimination during adaptation

In type I, type II and type V systems, invader target sequences (protospacers) carry an additional adjacent short sequence motif (2-7 nt), called the protospacer adjacent motif (PAM) (Deveau et al., 2008; Garneau et al., 2010; Horvath et al., 2008b; Jinek et al., 2012; Mojica et al., 2009a; Sapranauskas et al., 2011; Zetsche et al., 2015). The PAM plays an important role in CRISPR adaptation and interference. During interference the PAM is used for authentication of the invader DNA target (in addition to base-pairing between crRNA and protospacer) and therefore allows self/non-self discrimination (Marraffini and Sontheimer, 2010b). Self-targeting is a potential risk, because the crRNA can also pair with the corresponding CRISPR spacer in the host genome. This is prevented due to the absence of the PAM sequence in the CRISPR array. Since the PAM is required during interference, only acquisition of spacers with a PAM is effective. Indeed, when naturally occurring spacer sequences can be mapped to target sites in MGEs, a PAM is very often observed (Horvath et al., 2008b; Mojica et al., 2009a). This can, however, also be the product of selection for cells with functional spacers due to increased chance of survival. Therefore, high throughput spacer acquisition experiments in the absence of selection pressure have been carried out and revealed PAM specific spacer acquisition (Heler et al., 2015; Savitskaya et al., 2013; Shmakov et al., 2014; Yosef et al., 2012; Yosef et al., 2013). The degree of PAM specificity varies between systems, producing only ~35 % of spacers with a correct PAM in the type I-E system of *E. coli* (Yosef et al., 2012), while producing more than 99 % of spacers with a correct PAM in the type II-A system of *Streptococcus pyogenes* (Heler et al., 2015). PAM selection during spacer acquisition has been shown to be an inherent feature of Cas1 in the type I-E system and this likely holds true for other type I systems as well (Wang et al., 2015; Yosef et al., 2012). Type II-A systems have been shown to require Cas9, trans activating crRNA (tracrRNA) and Csn2 for spacer acquisition in addition to Cas1-2 (Heler et al., 2015; Heler et al., 2017; Wei et al., 2015b). Specifically, Cas9 has been shown to be required for selection of protospacers with the correct PAM and this function is believed to be universal among type II systems. The involvement of Cas9 in PAM selection in type II systems therefore may suggest that Cas1 does not possess any

PAM specificity in these systems. Cas4, which is present in type I-ABCDU and II-B and V systems, is another protein predicted to be associated with spacer acquisition. The *cas4* gene is always localized together with *cas1* and *cas2* and even fused to *cas1* in type I-U and V-B (Makarova et al., 2015; Shmakov et al., 2017). The involvement of Cas4 in spacer acquisition has been shown experimentally for type I-B (Li et al., 2014).

The small size of the PAM allows for sufficient discrimination power to prevent auto-immunity, while retaining a large number of potential target sites in an invader genome. While the PAM prevents self-targeting of the CRISPR array, it does not prevent acquisition of spacers from the host genome. Acquisition of self-spacers is either lethal or coincides with a non-tolerated PAM, the loss/modification of the target sequence or inactivation of CRISPR effector genes to prevent cell death (Dy et al., 2013; Paez-Espino et al., 2013; Wei et al., 2015b). Concomitantly, the acquisition of self-spacers in *E. coli* and *S. thermophilus* has been shown to be increased when the effector complex was either absent or inactivated (Wei et al., 2015b; Yosef et al., 2012). However, even in the absence of interference and potential autoimmunity, spacers are sampled from MGEs with a much higher frequency than from the genome (Diez-Villasenor et al., 2013a; Nunez et al., 2014; Yosef et al., 2012). A mechanism for this has recently been described for spacer acquisition in *E. coli*, which is dependent on the active replication of the source DNA (Levy et al., 2015). Indeed, the ratio of self vs. non-self spacers is directly linked to the replication frequency of the host genome vs. MGEs, respectively. While the genome is only present in one copy and replicates once per cell cycle, plasmids and phages are more actively replicating and reach far higher copy numbers. The *E. coli* CRISPR-Cas system derives the spacers from DNA fragments that are produced during RecBCD-catalysed repair of double stranded breaks (DSB) at stalled replication forks (Levy et al., 2015). This form of spacer acquisition requires Cas1 and Cas2 as the only Cas proteins and is termed naïve acquisition (Fineran and Charpentier, 2012; Yosef et al., 2012). However, naïve acquisition is not completely abrogated in the absence of RecBCD, suggesting alternative mechanisms, or other possible sources for spacer molecules such as DNA fragments generated by RM systems (Dupuis et al., 2013).

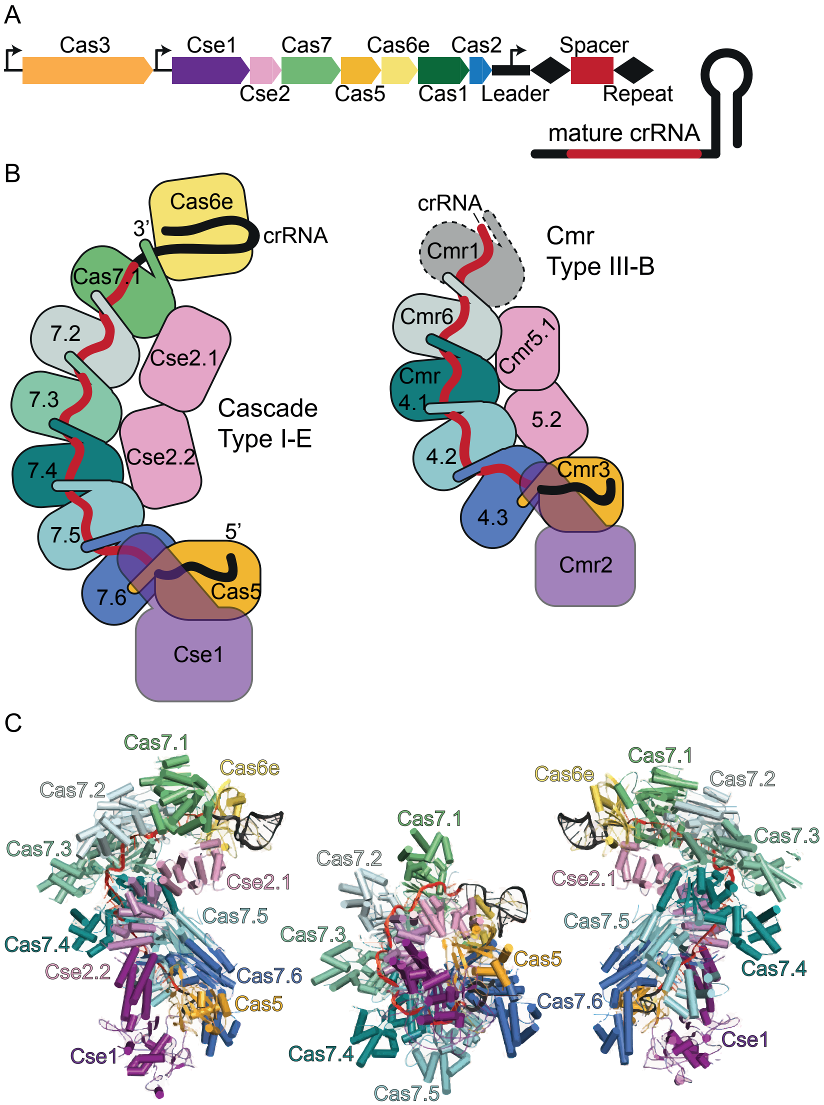


Figure 4 – Architecture of Cascade effector complex. Adapted from (Jackson et al., 2014; Jackson and Wiedenheft, 2015). A) Genomic locus architecture of the type I-E system in *E. coli*, showing the separate *cas3* gene, the operon encoding for the Cas proteins forming the Cascade complex (Cse1, Cse2, Cas7, Cas5, Cas6e), the *cas1* and *cas2* genes, the leader sequence and an array with two repeats and a spacer. Next to the array is a depiction of a mature crRNA. B) Schematics of the structures of the type I-E Cascade complex from *E. coli* and the type III-B Cmr complex from *Pyrococcus furiosus*. Both complexes show a similar architecture. Cas7 family subunits (Cas7, Cmr4) for a helical backbone around the crRNA which is capped at both ends by head and tail subunits. Both complexes contain additional subunits in the middle (Cse2, Cmr5) and both bind the 5' end of the crRNA in a subunit at the bottom (Cas5, Cmr3). In addition, both complexes contain a 'large subunit' at the bottom (Cse1, Cmr2). The short Cmr4 backbone in the Cmr complex is supplemented by structurally similar Cmr6 and Cmr1 subunits. Cascade contains the Cas6 subunit at the head of the complex, while Cmr contains no head subunit and has a trimmed crRNA 3'end. C) Three views of the crystal structure of the Cascade complex.

Mechanism of spacer integration

While there is still much uncertainty as to the origin and selection of new spacer molecules, our knowledge of the mechanism of spacer integration has vastly improved over the last few years. Most of this work has been done in the type I-E system of *E. coli*, but the general mechanism of spacer integration is likely universal (Figure 4). Cas1 and Cas2 are at the centre of this mechanism and have been shown to form a complex of two Cas1 dimers flanking a central Cas2 dimer in the type I-E system of *E. coli* (Nunez et al., 2014). Other systems likely also form complexes of Cas1-2, however they might differ in structure or carry additional protein subunits. Cas1 and Cas2 are both nucleases, however, it has been shown that the nuclease activity of Cas2 is dispensable for spacer acquisition (Babu et al., 2011). Spacer integration occurs at the leader end of a CRISPR array and includes the duplication of the first repeat. Both the leader and the first repeat have been shown to be essential and are interacting with either Cas1-2 or other integration factors (Moch et al., 2016; Wei et al., 2015a; Wright and Doudna, 2016; Yosef et al., 2012). Specifically, only 40 to 43 bp of the leader upstream of the first repeat are required for spacer acquisition (Diez-Villasenor et al., 2013a) and this region has been shown to bind the *E. coli* integration host factor (IHF) (Nunez et al., 2016; Yoganand et al., 2017). IHF binding induces a sharp bend in the DNA, which brings another upstream sequence motif, the integrase anchoring site (IAS), close to the first repeat and allows binding of the Cas1-2 complex to form a supercomplex (Yoganand et al., 2017). Binding of Cas1-2 is reinforced when it carries a spacer precursor molecule (Moch et al., 2016) and it has been shown that dsDNA is the preferred substrate (Nunez et al., 2015b; Rollie et al., 2015; Wang et al., 2015). Furthermore it has been shown that the ideal spacer precursor has 3’ overhangs and carries the PAM sequence in one of its 3’ termini. Spacer precursors are likely matured by the Cas1 subunits that cleave 3’ of the first C nucleotide of the CTT PAM and at a fixed distance on the other 3’end to create a 33 nt long spacer (Wang et al., 2015). The 3’-OH termini on the precursor are essential, since they each perform a Cas1-2 catalysed nucleophilic attack on one strand of the first repeat during integration. This process resembles that of retroviral integrases and transposases (Nunez et al., 2015b). Initially, a stable spacer integration intermediate was detected *in vivo*, which suggested a staggered cut at the two ends of the first repeat and subsequent integration of the double stranded spacer (Arslan et al., 2014). More recently, *in vitro* assays mimicking spacer integration have provided more details about the process (Nunez et al., 2015b; Rollie et al., 2015; Wright and Doudna, 2016). Although these assays are each limited in their reproduction of the situation *in vivo*, combined they have led to the following model. Cas1-2 first catalyses the nucleophilic attack of the 3’-OH of the spacer end that does not carry the CTT PAM. This initial nucleophilic attack occurs at the leader proximal end of the repeat. The second nucleophilic attack is carried out by the 3’-

OH of the C nucleotide left over from the PAM, which gets integrated at the leader distal end of the repeat. This second integration site is at the penultimate base of the original repeat, therefore replacing the terminal C nucleotide of the repeat with the incoming C from the spacer. The C of the original repeat then becomes the last nucleotide of the second repeat. After the spacer is coupled at both ends to one of the strands of the repeat, DNA polymerase I fills in the gaps to restore the repeats and cellular ligases likely repair the remaining nicks (Ivancic-Bace et al., 2015).

Primed acquisition

Next to naïve acquisition, primed acquisition was described for a number of type I systems (Datsenko et al., 2012; Li et al., 2014; Richter et al., 2014; Swarts et al., 2012; Vorontsova et al., 2015). Priming acquisition requires all *cas* genes and the presence of a spacer already targeting the MGE. Thus, priming serves as a positive feedback loop that strengthens the defence by facilitating additional spacer acquisition. This is also reflected by the fact that priming acquisition is around 50 times more frequent than naïve acquisition in *E. coli* (Datsenko et al., 2012). Priming in *E. coli* is a very robust process, tolerating up to 13 mutations in the priming protospacer (out of 32), as well as many possible PAM mutations (Fineran et al., 2014; Xue et al., 2015). Furthermore, priming results in a much higher percentage of spacers with the canonical PAM (>95 %) than naïve acquisition (37 %) (Shmakov et al., 2014; Yosef et al., 2013). Another hallmark of priming in type I-E systems of *E. coli* is the strand bias observed with newly acquired spacers. Around 90 % of newly acquired spacers target the same DNA strand as the original spacer that triggered priming (Datsenko et al., 2012; Shmakov et al., 2014). Naïve acquisition on the other hand results in a 50/50 strand distribution of spacers. The priming strand bias is, however, not universal across type I systems and this is related to mechanistic differences (Li et al., 2014; Richter et al., 2014). The mechanism of priming is still not fully understood and there is evidence supporting different models. The model that was first proposed for *E. coli* assumes that DNA degradation fragments generated during direct interference serve as spacer precursors during priming (Swarts et al., 2012). Another model proposes the formation of a Cas1-2-3 supercomplex, triggered by escape mutants, that scans the target DNA for potential new spacers (Redding et al., 2015). The mechanism of priming is the subject of Chapter 5 and is further discussed in Chapter 8 (general discussion).

Biogenesis of crRNA and effector complex formation

This process is often referred to as the expression stage of CRISPR-Cas defence, which refers to the expression of *cas* genes and transcription of the CRISPR array (Figure 1). Most important in this process is the biogenesis of the crRNA and the subsequent formation of the effector complexes. The mature crRNA which contains a single spacer sequence and varying repeat flanks is at the centre of CRISPR-Cas immune systems, providing the targeting information that makes these systems so specific and versatile. The CRISPR array is initially transcribed into one long precursor (pre-crRNA), which contains hairpin elements if the CRISPR repeats are palindromic (Brouns et al., 2008a; Charpentier et al., 2015; Jore et al., 2011a). After pre-crRNA transcription each system processes the RNA in a different way to create mature crRNAs of 30 – 65 nt length. Class 1 and class 2 systems differ fundamentally in their crRNA biogenesis. Class 1 systems make use of proteins from the Cas6 family of ribonucleases that recognize the repeat sequences and/or hairpin structures and cleave at a specific position (Brouns et al., 2008a; Carte et al., 2010; Carte et al., 2008). Some class 1 systems undergo a second step of maturation, but the mechanism is unknown (Charpentier et al., 2015). DNA-targeting class 2 systems have to be separated into type II systems and type V systems, which differ in crRNA maturation. Type II systems (Cas9) make use of trans-acting CRISPR RNAs (tracrRNA), which form partial duplexes with the pre-crRNA and direct housekeeping RNase III to cleave within the repeats (Deltcheva et al., 2011). Type V-A systems (Cpf1/Cas12a) do not require a tracrRNA; instead the Cpf1 protein has intrinsic RNase activity that allows it to process its own pre-crRNA to mature crRNAs (Fonfara et al., 2016; Swarts et al., 2017). Type V-B systems (C2c2/Cas12b) also have intrinsic guide processing, but also make use of tracrRNA (Abudayyeh et al., 2016; East-Seletsky et al., 2016). Still, both type II and type V-A require a second crRNA maturation step that trims the 5' end (type II) or the 3' end (type V). The mechanism and the involved components of this secondary processing are still unknown (Charpentier et al., 2015), but most likely involve a non-Cas exo-ribonuclease.

Cas6 family proteins are metal-independent endo-ribonucleases that contain two RRM-type RNA-binding domains. In type I systems, cleavage of the pre-crRNA typically occurs 8 nt upstream of the repeat-spacer boundary, creating an 8-nt 5' handle, followed by the spacer and a longer 3' handle that often forms a stem loop (Brouns et al., 2008a; Carte et al., 2008; Jore et al., 2011a). Type I systems typically encode the CRISPR-associated ribonucleoprotein (RNP) complex for antiviral defence (Cascade). This complex carries the crRNA and the RNA is likely to serve as a scaffold for complex assembly. Cas6 stays associated to the crRNA in type I-E and I-F systems and is a single-turnover enzyme, and a stably-associated subunit of the Cascade complex (Jore et al., 2011a; Wiedenheft et al., 2011a). In type I-A, I-B, I-D and type III systems, Cas6 is a multiple turnover enzyme that does not associate with the corresponding RNP complex after cleaving the crRNA (Plagens et al., 2014; Richter et al., 2013; Richter et al., 2012b; Scholz et al., 2013). The absence of Cas6 in the final RNP complex coincides with the lack of a stem loop in these systems. Type I-C systems lack a *cas6* gene entirely and its function is taken over by *cas5* (Nam et al., 2012a). There are many organisms with multiple different class I CRISPR-Cas systems, such as *Thermus thermophilus* and *Pyrococcus furiosus*. In these organisms, not all CRISPR loci contain the necessary *cas6* gene for crRNA processing. Instead, crRNAs can be provided in trans by other CRISPR loci or standalone *cas6* genes, even if these are part of a different subtype (Majumdar et al., 2015; Staals et al., 2014).

Structure of the Cascade complex

The type I-E Cascade complex consists of five different protein subunits in uneven stoichiometry (Cse1_1, Cse2_2, Cas7_6, Cas5_1, Cas6e_1) and a 61 nt crRNA (Figure 3) (Brouns et al., 2008a; Jore et al., 2011a). The complex structure resembles a seahorse shape, with the six Cas7 proteins forming the helical backbone (Wiedenheft et al., 2011a). The Cas7 proteins make direct contact with the 32 nt spacer part of the crRNA, which is arranged in short helical segments (Jackson et al., 2014; Mulepati et al., 2014; Zhao et al., 2014). These 6 nt helical segments are interrupted by a “thumb”-like structure of the Cas7 proteins that hold on to the crRNA. This interaction results in the flipping of the underlying RNA base, which therefore is not available for base-pairing with a DNA target. The head protein (Cas6e) makes contact with the conserved stem loop at the 3’ end of the crRNA, while the foot proteins (Cse1, Cas5) make contact with the conserved 5’ end, the PAM and the seed sequence of the crRNA. Although Cas6e is part of the final Cascade complex, it is not essential for activity. It was demonstrated that when mature crRNA is supplied in a Cas6e independent manner, Cascade activity is not impaired in the absence of Cas6 (Semenova et al., 2015). Specifically, Δcas6e-Cascade was able to trigger both direct interference as well as primed acquisition. Furthermore, in the absence of Cas6e, also the 3’ stem-loop of the crRNA is dispensable at least *in vitro*. Finally, the Cascade architecture has been shown to depend on the crRNA length. Lengthening or shortening the spacer part of the crRNA leads to Cascade complexes with extra or fewer Cas7 backbone subunits respectively (Gleditzsch et al., 2016; Kuznedelov et al., 2016; Luo et al., 2016). Elongated complexes are fully functional and even sensitive to target mismatches within the elongated part of the crRNA. Shortened complexes are fully functional as well, when the spacer part is at least 20 nt. A spacer of 14 nt appeared unable to drive direct interference, while still causing primed acquisition. Type I Cascade complexes with altered crRNA length and architecture have not been found in nature and the crRNA length in type I-E systems is largely fixed (Diez-Villasenor et al., 2010). However, type III complexes are naturally ambiguous in their architecture, because they are supplied with crRNA of various lengths (Hale et al., 2008; Hale et al., 2009; Staals et al., 2013; Tamulaitis et al., 2014). Since all class I complexes share remarkable structural similarities, it is likely that in all respective systems complexes are assembled around the crRNA and that the crRNA length determines the complex architecture (Jackson and Wiedenheft, 2015). Despite the important function of crRNA in complex assembly, crRNA independent complex formation has also been observed, although complex architecture was less defined (Beloglazova et al., 2015). This shows that the Cas proteins are capable of self-assembly and that the general complex architecture is a result of the protein interactions.

Direct Interference

Direct interference is the detection and destruction of invading nucleic acids by Cas nucleases. At the centre of this process are the RNP complexes that are programmed by their guide RNAs to target complementary sequences (Figure 1). Binding of a cognate target site leads to the formation of an R-loop, where one DNA strand is paired with the RNA while the other DNA strand is displaced. The identification and destruction of a canonical target depends on two main factors. The target sequence (protospacer) needs to be sufficiently complementary to the spacer portion of the crRNA, and, with the exception of RNA targeting type III systems, all systems require the presence of a short protospacer adjacent motif (PAM) (Barrangou et al., 2007; Brouns et al., 2008a; Garneau et al., 2010; Westra et al., 2012c; Zetsche et al., 2015). The PAM is the most important element of a potential target site. Even if the protospacer is perfectly matching the crRNA, an incorrect PAM will prevent interference activity in most systems. Some type I systems are more promiscuous towards the PAM sequence, tolerating degenerative PAM motifs. The type I-E system, for example, tolerates five different PAM sequences for direct interference, while nearly all PAM sequences can lead to primed acquisition (Fineran et al., 2014; Westra et al., 2012c). However, a PAM sequence identical to the sequence of the repeat that is adjacent to the spacer in the CRISPR array is not tolerated (Xue et al., 2015). This exemplifies the role of the PAM as a safety feature to prevent auto-immunity. In addition, the PAM is thought to function as a means to speed up the target search of the crRNP in the vast amount of cellular DNA (Künne et al., 2014; Redding et al., 2015; Sternberg et al., 2014). Continuously initiating base-pairing between crRNA and DNA in order to find a complementary site would significantly slow down the search, especially because the DNA strands need to be separated first. Instead, crRNPs have been shown to initially scan for PAM sequences and initiate base-pairing when a *bona fide* PAM has been encountered (Redding et al., 2015; Sternberg et al., 2014). Furthermore, the PAM-protein interaction has been shown to aid in the local unwinding of the dsDNA to initiate crRNA-DNA (target strand) base-pairing, which is important since crRNPs do not contain ATP-powered helicases. Despite this mechanism, PAM-independent target binding has also been demonstrated for type I-E Cascade (Blosser et al., 2015). This coincides with the fact that Cascade strongly favours negatively supercoiled (nSC) target DNA over relaxed or linear DNA and that the supercoiling energy aids in the local unwinding of the DNA (Westra et al., 2012c). Other crRNPs such as Cas9 strictly require the presence of a PAM for target binding and cleavage activity and do not prefer nSC DNA (Jinek et al., 2012). Cascade recognizes the PAM by direct protein-DNA interactions, specifically the minor grove of the double stranded PAM sequence is interacting with three structural features of the Cse1 subunit (Hayes et al., 2016). The promiscuity of Cascade towards the PAM sequence is explained by the inherent promiscuity of DNA minor grove recognition. After PAM recognition, a wedge is inserted that initiates directional DNA strand unwinding, followed by segmental base-pairing of crRNA and the target strand (Blosser et al., 2015; Hayes et al., 2016; Rutkauskas et al., 2015). The non-target strand is displaced and locked behind the Cse2 subunit dimer.

The seed sequence

The extent of mismatch tolerance between protospacer and crRNA varies between systems, however, type I and type II systems both contain a region that tolerates mutations to a far lesser extent than the rest of the protospacer/crRNA. This region is called “the seed” and usually encompasses seven base-pairs (type I) or seven to twelve base-pairs (type II and type V) at the PAM proximal end of the protospacer (Jinek et al., 2012; Künne et al., 2014; Semenova et al., 2011; Wiedenheft et al., 2011b; Zetsche et al., 2015). Recently, comprehensive studies of many different spacer sequences have revealed a flexibility in the definition of the seed based on the individual sequence and the experimental conditions (Cong et al., 2013; Jiang et al., 2013; Kuscu et al., 2014; O’Geen et al., 2015; Pattanayak et al., 2013; Wu et al., 2014a; Wu et al., 2014b; Xue et al., 2015). However, the seed is not only defined by its mutation intolerance, but also by structural or mechanistic features that allow for the initial interaction (base pairing) between crRNA and DNA target. Recent single molecule studies have shown that several RNPs initiate interactions with target DNA at the PAM site by protein-DNA interactions, this is followed by base-pairing of the PAM proximal parts (seed) and subsequently base-pairing continues in a zipper-like manner (Rutkauskas et al., 2015; Sternberg et al., 2014). This directional binding starting at the seed is likely what makes the seed so essential. Seed sequences have also been described for miRNA/siRNA and bacterial sRNA and their features have been extensively researched in miRNAs (Künne et al., 2014). Briefly, the seed sequence can be pre-ordered in a configuration resembling an A-form helix, which lowers the entropic cost of duplex formation. In addition, the seed can be better accessible for the DNA targets, while the rest of the RNA is more protected/shielded by the RNP. The seed sequence is the subject of chapter 2 and will be more thoroughly discussed in chapter 8 (general discussion).

Conformational effects

A number of single molecule studies on *E. coli* Cascade have demonstrated several conformational features that influence the immune response. First, cryo-EM structures and magnetic tweezer experiments have shown that binding of a cognate target site triggers a conformational change in the Cascade complex, leading to a more stable interaction (Rutkauskas et al., 2015; Wiedenheft et al., 2011a). This "locked" state has been shown to be required for recruitment and activation of the Cas3 helicase/nuclease. Single molecule imaging of DNA and Cas proteins has revealed that Cascade bound to canonical targets recruits Cas3, while Cascade bound to non-canonical targets does not. In the presence of Cas1-2, however, a nuclease-inactive Cas3 is recruited to non-canonical targets (Redding et al., 2015). Furthermore, a single molecule FRET study has demonstrated independently that Cascade exhibits two distinct binding modes for canonical and non-canonical targets (Blosser et al., 2015). While only the canonical binding mode leads to direct interference, the non-canonical binding mode is still able to trigger primed acquisition. In addition, the non-canonical binding mode is PAM- and seed-independent, suggesting a different mechanism of target binding. Finally, an additional single molecule FRET study has shown that upon DNA binding the Cse1 subunit of Cascade exists in an equilibrium between 'closed' and 'open' conformations. This equilibrium depends on the extend of target mutations and also correlates with the relative abundance of interference and priming respectively (Xue et al., 2016). The implications of Cascade binding modes and conformations on interference and priming are also described in chapter 5 and discussed in detail in chapter 8 (general discussion).

Cleavage of foreign DNA

In type I systems, DNA that is bound by Cascade complexes is ultimately degraded by Cas3. *E. coli* Cas3 is a two-domain protein that consists of an N-terminal HD domain and a C-terminal superfamily-2 (SF2) helicase domain. The HD domain exerts cobalt-dependent nuclease activity exclusively on single-stranded DNA (ssDNA) (Beloglazova et al., 2011; Mulepati and Bailey, 2011; Mulepati and Bailey, 2013; Sinkunas et al., 2011). The helicase domain exerts ATP- and magnesium-dependent DNA-DNA and DNA-RNA unwinding activity (Beloglazova et al., 2011; Howard et al., 2011). Cas3 is recruited to the R-loop by the Cse1 subunit of Cascade (Hochstrasser et al., 2014; Mulepati and Bailey, 2013; Westra et al., 2012c). Cas3 uses a DNA binding cleft in its C-terminal domain to bind to the single stranded non-target strand of the R-loop and nicks it using endonuclease activity of the HD domain (Gong et al., 2014; Huo et al., 2014; Mulepati and Bailey, 2013). After nicking, the binding cleft is transformed into a tunnel, ensuring stable binding of the DNA. The DNA is then threaded through the tunnel by the SF2 helicase domain, consuming ATP. This way, the helicase is feeding the DNA directly to the HD domain, resulting in 3’ to 5’ exo-nucleolytic cleavage (Mulepati and Bailey, 2013; Sinkunas et al., 2011; Westra et al., 2012c). The precise cleavage sites of Cas3 are the subject of chapter 5 and are further discussed in Chapter 8.

Only type I systems encode a *cas3* gene. Instead, all other systems have crRNPs that have integral nuclease domains/subunits to cleave the target (Makarova et al., 2015). However, while Cas3 activity leads to the complete degradation of a target, the other crRNPs only cleave at one or a few positions to disable the invader.

Class 2 systems employ single protein RNPs, rather than multi-subunit class 1 RNP complexes. Type II and type V-A/B systems code for Cas9 and Cpf1/C2c1, respectively. Cas9 contains two separate nuclease domains (HNH/RuvC), which cleave one DNA strand each to produce a blunt cut (Jinek et al., 2012). Cpf1 and C2c1 both contain a single nuclease domain (RuvC) that cleaves both strands and produces a staggered cut (Swarts et al., 2017; Yang et al., 2016; Zetsche et al., 2015). Interestingly, PAM sequences are located at opposite sides of the protospacer in Cas9 (5’ of target strand) and Cpf1/C2c1 (3’ of target strand). Furthermore, Cas9 and C2c1 require a tracrRNA in addition to the crRNA for activity, while Cpf1 only requires a short crRNA with a 5’ hairpin. Due to its compact, single protein architecture and easy programmability, Cas9 has been extensively applied for genome editing over the last years, and the newly discovered Cpf1 is emerging as a competitive alternative (Kim, 2016; Tóth et al., 2016; Wright et al., 2016).

Type III systems encode RNP complexes that share remarkable structural similarities to type I RNP complexes (Jackson et al., 2014; Jackson and Wiedenheft, 2015; Mulepati et al., 2014; Osawa et al., 2015; Rouillon et al., 2013; Spilman et al., 2013; Taylor et al., 2015). However, they show substantial functional differences. Both, the Csm complex (type III-A) and the Cmr complex (type III-B) were initially shown to target and cleave RNA (Hale et al., 2009; Staals et al., 2013; Zebec et al., 2014). More recently, it was demonstrated that both complexes are also transcription dependent DNA nucleases, therefore being able to cleave both DNA and RNA (Deng et al., 2013; Elmore et al., 2016; Estrella et al., 2016; Goldberg et al., 2014; Peng et al., 2015; Samai et al., 2015). The complexes first identify mRNA targets via base-pairing with their crRNA, after which both the RNA transcript as well as the transcribed DNA are being cleaved. Another major difference between type I and type III complexes is the mechanism of self/non-self discrimination. Unlike type I systems, type III systems have not been shown to require PAM sequences in their targets. Instead, auto immunity is prevented by the recognition of the repeat sequence in the host genome via base-pairing with the repeat part of the crRNA or via protein interactions (Marraffini and Sontheimer, 2010b; van der Oost et al., 2014). Type III systems therefore use a self-inactivating mechanism, while type I systems use a non-self activating mechanism for direct interference (PAM) and a self-inactivating mechanism for priming (the repeat).

Genome editing using crRNPs

Recently, the field of genome editing has been revolutionized by the introduction of RNA-guided endonucleases (RGENs), most prominently Cas9 from *Streptococcus pyogenes* (Cho et al., 2013; Cong et al., 2013; Jinek et al., 2013). RGENs are nucleases that make use of a guide RNA molecule that is used to identify *bona fide* target sequences via RNA-DNA base-pairing. Starting in 2013, Cas9 has been shown to efficiently create indels (frame-shifts) or stimulate homologous recombination of repair templates in a large range of cell types (Kim, 2016). Developments have been made to turn Cas9 into an antimicrobial or antiviral tool (Bikard et al., 2014; Ebina et al., 2013; Hu et al., 2014; Ramanan et al., 2015). Furthermore, catalytically dead Cas9 (dCas9) has been used as a targeting platform for a multitude of fused functional domains, such as transcriptional activators/repressors (Bikard et al., 2013; Gilbert et al., 2013; Konermann et al., 2013; Qi et al., 2013), fluorescent proteins as localization signals (Chen et al., 2013) and histone-modifying enzymes (Hilton et al., 2015; Kearns et al., 2015). More recently, another Cas effector protein, Cpf1, has been discovered as a very similar alternative to Cas9. Cpf1 has a few conceptual advantages, like its use of a single crRNA and its ability to process its own crRNA from CRISPR arrays (Fonfara et al., 2016; Zetsche et al., 2015). This could make it exceptionally well suited for multiplexed applications and this is beginning to be explored (Zetsche et al., 2017). Similar to the beginnings of Cas9, Cpf1 has already racked up an impressive amount of genome editing applications in a short time. Furthermore, since Cpf1 leaves sticky ends after cleavage, it is also being developed as a tool for *in vitro* DNA assembly (Lei et al., 2017; Li et al., 2016). At the moment, efforts are being made to characterize three other newly discovered Cas effector proteins, namely C2c1, C2c2, C2c3 (Liu et al., 2017a; Liu et al., 2017b; Shmakov et al., 2015; Yang et al., 2016). The unique features of C2c2 have already led to its development into a molecular diagnostics tool for the attomolar detection of RNA or DNA with single nucleotide specificity (Abudayyeh et al., 2016; Gootenberg et al., 2017). It is just a matter of time until more applications for these proteins will be developed (Lewis and Ke, 2017). In conclusion, RNA-guided proteins are of great interest and offer unprecedented potential for a diverse range of applications. It is therefore vital to extend the toolbox of these proteins and find the best solution to every problem.

Thesis outline

The overall aim of this thesis is to explore the type I-E CRISPR-Cas system of *Escherichia coli* and unravel mechanistic details of both interference and primed acquisition. We are especially interested in the dynamic relationship between these two processes.

Chapter 2 reviews a fundamental feature of many small RNA guided systems, the 'seed'. Small RNAs, such as RNAi associated RNAs, prokaryotic small regulatory RNAs and crRNAs rely on protein-assisted base pairing of the guide RNA with target mRNA or DNA to interfere with their transcription, translation or replication. All three groups identify their target sequence by base pairing after finding it in a pool of millions of other nucleotide sequences in the cell. In this complicated target search process, a region of 6 to 12 nucleotides of the small RNA termed the 'seed' plays a critical role.

Chapter 3 provides a description and protocol of the Electrophoretic Mobility Shift Assay (EMSA) and its use for studying crRNPs. EMSA is a straightforward and inexpensive method for the determination and quantification of protein–nucleic acid interactions. Protocols for two types of EMSA assays are described using the Cascade ribonucleoprotein complex from *Escherichia coli* as an example. The EMSA method and these protocols are applied throughout the other chapters of this thesis.

Chapter 4 focusses on the processes of interference and primed adaptation, specifically on their mutation tolerance. We provide a systematic analysis of the constraints of both direct interference and priming in *E. coli*. Our findings imply that even out-dated spacers containing many mismatches can induce a rapid primed CRISPR response against diversified or related invaders, giving microbes an advantage in the co- evolutionary arms race with their invaders.

In **Chapter 5** we elucidate the mechanism of priming. Specifically, we determine how new spacers are produced and selected for integration into the CRISPR array during priming. We show that priming is directly dependent on interference. We show that Cas3 couples CRISPR interference to adaptation by producing DNA breakdown products that fuel the spacer integration process in a two-step, PAM-associated manner. Our results highlight that the selection of PAM-compliant spacers during priming is enhanced by the combined sequence specificities of Cas3 and the Cas1-2 complex, leading to an increased propensity of integrating functional CTT-containing spacers.

In **Chapter 6** we look deeper into a nucleotide specific effect on priming that was discovered in Chapter 4. Immunity is based on the complementarity of host encoded spacer sequences with protospacers on the foreign genetic element. The efficiency of both direct interference and primed acquisition depends on the degree of complementarity between spacer and protospacer. We show that G substitutions have a profoundly negative effect on interference, while C substitutions are readily tolerated when in the same positions. Furthermore, we show that this effect is based on strongly decreased binding of the effector complex Cascade to G mutants, while C mutants only minimally affect binding.

Chapter 7 describes an attempt to develop the Cascade complex into a genome editing tool. Recently, RNA guided endonucleases (RGENs) such as Cas9 or Cpf1 have revolutionized genome editing. Here, we have explored the possibility to develop a new genome editing tool that makes use of the Cascade complex from *E. coli*. This RNA guided protein complex is fused to a FokI nuclease domain to sequence specifically cleave DNA. We validate the tool *in vitro* using purified protein and two sets of guide RNAs, showing specific cleavage activity. Unfortunately, we were not able to successfully apply the tool *in vivo* in eukaryotic cells.

Chapter 2

Planting the seed: target recognition of short guide RNAs

Tim Künne, Daan C. Swarts, Stan J.J. Brouns

Laboratory of Microbiology, Department of Agrotechnology and Food Sciences,
Wageningen University, Dreijenplein 10, 6703 HB Wageningen, The Netherlands.

This Chapter has been published as:

Künne T, Swarts DC, Brouns SJ. 2014. Planting the seed: target recognition of short guide RNAs. Trends Microbiol, 22(2):74-83.

Abstract

Small guide RNAs play important roles in cellular processes such as regulation of gene expression and host defense against invading nucleic acids. The mode of action of small RNAs relies on protein-assisted base pairing of the guide RNA with target mRNA or DNA to interfere with their transcription, translation or replication. Several unrelated classes of small non-coding RNAs have been identified including eukaryotic RNA silencing associated small RNAs, prokaryotic small regulatory RNAs and prokaryotic CRISPR (clustered regularly interspaced short palindromic repeats) RNAs. All three groups identify their target sequence by base pairing after finding it in a pool of millions of other nucleotide sequences in the cell. In this complicated target search process, a region of 6 to 12 nucleotides of the small RNA termed the 'seed' plays a critical role. Here we review the concept of seed sequences and discuss its importance for initial target recognition and interference.

Keywords

RNAi, CRISPR, Argonaute, Hfq, Cascade, Cas9

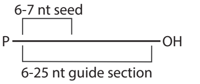
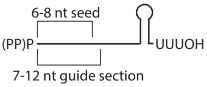
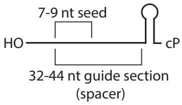
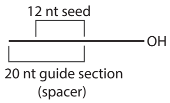
Small guide RNAs and their versatile roles

Due to the ever growing capacity to sequence cellular RNAs, small guide RNAs and their functions continue to be discovered. It has become increasingly more evident that small RNAs are key players in fine-tuning gene expression and protecting hosts from mobile genetic elements such as viruses and transposons. To date, three main classes of small guide RNAs have been described: (i) RNA interference (RNAi) based small RNAs, (ii) prokaryotic small regulatory RNAs and (iii) immune-associated CRISPR (clustered regularly interspaced short palindromic repeat) RNAs (crRNAs).

The first class of small RNAs is involved in regulating gene expression in eukaryotes through different RNAi pathways (reviewed in (Bartel, 2009; Juliano et al., 2011; Ketting, 2011; Siomi and Siomi, 2009)). RNA interference associated small RNAs are critical for cell development, adaptation to changing environmental conditions, resistance to viruses, repression of transposable elements, chromatin structuring and regulation of many other cellular processes. The second class of small RNAs consists of prokaryotic small regulatory RNAs (sRNA) and has been studied mostly in bacteria. This class also regulates gene expression, using distinct mechanisms (reviewed in (Storz et al., 2011; Waters and Storz, 2009)). sRNAs are involved in the response to changing environmental conditions. A third class of small RNAs, called CRISPR RNAs, guide an adaptive and heritable immune system in bacteria and archaea, which protects microbial cells from invading nucleic acids such as viral genomes and conjugative plasmids (Sorek et al., 2013b; Westra et al., 2012b).

Although these three classes of small guide RNAs are unrelated, they show similarities in the way their target nucleic acids are recognized. In all cases, the recognition of target nucleic acids occurs by Watson-Crick base pairing with the small RNA. Initial target recognition by these three small RNA classes is governed by only part of the small RNA guide, a 6-12 nucleotide (nt) segment referred to as the ‘seed’. This functionally essential segment is involved in initial pairing between guide and targets, and allows rapid probing of different regions of cellular nucleic acids. Seed sequences have been identified by one of four different methods including (i) *In vivo* loss of function assays; seed-target mismatches are poorly tolerated and lead to loss of a silencing or resistance phenotype. (ii) *In vitro* biochemical assays; the seed sequence of guide RNAs exhibits a higher affinity than the remainder of the guide RNA for the target nucleic acids. (iii) Structural observations; the seed sequence is structurally more accessible and/or in a configuration optimal for base pairing with the target. (iv) Computational analysis; seed sequences are conserved regions in RNA guides. In this review, we will compare the characteristics of seed sequences from RNAi-based small RNAs, sRNAs and crRNAs, and discuss how class-specific proteins assist in the target search process.

Table 1. Comparison of small guide RNA properties

	miRNA/siRNA	sRNA	crRNA	
			Type I	Type II
Schematics of short guide RNAs (5’ to 3’)				
Length of small RNA	20-25 nt	>50 nt	58-70 nt	35-45 nt
Length of guide section	6-25 nt	7-12 nt	32-44 nt	20 nt
Length of seed	6-7 nt	6-8 nt	7-9 nt	12 nt
Type of seed	Contiguous	Contiguous (with exceptions)	Non-contiguous	Contiguous
Position of seed	Position 2-7 or 2-8	Near 5’ end or internal	Position 9-19	Position 9-20
RNA 5’ end	P	PPP or P	OH	Not described
RNA 3’ end	OH, 2’-CH_3	OH	2’,3’–cyclic P or P	OH
Seed only pairing	Translation inhibition, no mRNAcleavage	Translation inhibition	Abolishes interference	Abolishes interference
Seed helical pre-ordering	Yes	No	Possibly	Possibly
Seed mismatch tolerance	Yes, via compensatory pairing	No	No	No
Target nucleic acid	mRNA	mRNA	Invader DNA	Invader DNA
Associated protein	Argonaute	Hfq	Cascade complex	Cas9
Refs	(Bartel, 2009; Brennecke et al., 2005; Chorn et al., 2010; Doench and Sharp, 2004; Elkayam et al., 2012; Grimson et al., 2007; Lewis et al., 2003; Nakanishi et al., 2012; Schirle and MacRae, 2012)	(Bandyra et al., 2012; Sauer, 2013; Storz et al., 2011)	(Jore et al., 2011a; Maier et al., 2013; Sinkunas et al., 2013; Wiedenheft et al., 2011a; Wiedenheft et al., 2011b)	(Deltcheva et al., 2011; Gasiunas et al., 2012; Jiang et al., 2013; Jinek et al., 2012)

RNAi based small RNAs

RNAi uses small RNA guides of 20-25 nt to regulate gene expression in eukaryotes (Table 1) (reviewed in (Bartel, 2009; Juliano et al., 2011; Ketting, 2011; Siomi and Siomi, 2009)). Three major types of RNAi-based small RNAs have been described: (i) microRNAs (miRNAs), which are derived from small hairpin structured RNAs encoded on the genome; (ii) small interfering RNAs (siRNAs), which originate from duplex

RNAs or longer RNA hairpins; and (iii) PIWI-interacting RNAs (piRNAs), of which the biogenesis is poorly understood. These small RNAs form complexes with proteins from the Argonaute family (Ago or PIWI proteins), the key protein component of the RNA-induced silencing complex (RISC). In these complexes, the RNA guides sequence specific binding of mRNA resulting in translational repression or mRNA cleavage. This review will focus on the seed of miRNA and siRNA, as no evidence exists for a seed in piRNAs (Vourekas et al., 2012).

Guide architecture

The term 'seed' was first used to describe nt 2-7 or 2-8 of miRNAs, as base pairing of only these nucleotides is sufficient to cause translational repression (Table 1) (Brennecke et al., 2005; Lewis et al., 2003). Furthermore, the guide consists of an anchor (nt 1), a central region (nt 9-12), a 3' supplementary region (nt 13-17) and tail region (nt 18-21), each with its own functional importance as described below (Wee et al., 2012).

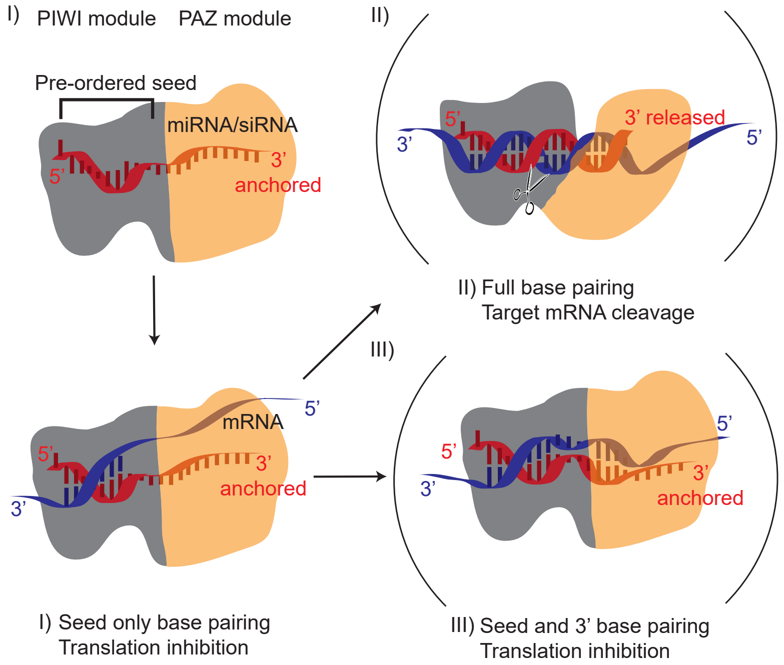


Figure 1 - Model of miRNA/siRNA-Argonaute complex and target binding. I) Guide RNA (red) is bound to Argonaute, stretching (5'- 3') from the PIWI module (grey) to the PAZ module (yellow). Nt 2-8 (seed) of the RNA are pre-ordered in a helical conformation, nt 1 is flipped into a binding pocket and is not available for base pairing. The 3' part of the guide is anchored to the PAZ domain. Recognition of the incoming mRNA (blue) is nucleated through pairing with seed nt 2-8. II) In case of extensive complementarity between guide and target, base pairing is extended starting from the seed. This leads to a conformational change of Argonaute. Progressive base pairing leads to the release of the 3' part of the guide from the PAZ domain, allowing guide and target to intertwine, possibly forming a complete helix. Complete binding, allows an active Argonaute to cleave the target mRNA between position 10 and 11 (indicated by scissors). III) In case of partial 3' complementarity, guide and target pair around nt 13-16, leaving a gap between seed and 3' pairing. Despite 3' pairing, the guide remains anchored to Argonaute. Figure adapted from (Bartel, 2009).

Translational repression by mRNA binding

Translational repression usually requires only pairing of the guide's seed with the target mRNA (Figure 1) (Brennecke et al., 2005; Chorn et al., 2010; Doench and Sharp, 2004; Grimson et al., 2007). Although the nucleotide composition of the seed affects binding affinity and the degree of translational repression (Ui-Tei et al., 2008), seed–target mismatches, bulges and G:U wobbles typically decrease target binding affinities much more than can be explained by thermodynamics of the seed-target match alone (Brennecke et al., 2005; Doench and Sharp, 2004; Schwarz et al., 2006; Wee et al., 2012). This indicates that the structure of the seed plays an important role during target binding. During perfect seed–target pairing, additional pairing of the 3' region of the guide called supplementary pairing can enhance translational repression (Figure 1) (Brennecke et al., 2005; Chorn et al., 2010; Doench and Sharp, 2004; Grimson et al., 2007; Wee et al., 2012). Supplementary pairing requires four contiguous Watson-Crick base pairs in the region of nt 13-16. If seed pairing is imperfect, 3' compensatory pairing at a minimum of nine nt positions may compensate for a single bulge or mismatches in the seed region (Bartel, 2009; Brennecke et al., 2005; Doench and Sharp, 2004; Grimson et al., 2007; Lewis et al., 2005; Ui-Tei et al., 2008; Wee et al., 2012; Yekta et al., 2004). Compensatory pairing can decrease the seed pairing requirement to as little as four base pairs (Brennecke et al., 2005). However, seed mismatches with compensatory sites are scarce and appear to be rarely conserved (Friedman et al., 2009; Lewis et al., 2005).

Target mRNA cleavage

Upon perfect pairing of the seed, central and 3' supplementary region target mRNA binding can lead to ATP-independent mRNA cleavage by the Argonaute protein when it has an RNase H –like PIWI domain with an intact active site (*i.e.* slicer Argonaute) (Ameres et al., 2007; Dahlgren et al., 2008; Elbashir et al., 2001; Haley and Zamore, 2004; Hutvágner and Zamore, 2002; Schwarz et al., 2006; Wee et al., 2012; Yekta et al., 2004) (Figure 1). Base pairing between the guide and the target initiates at the seed and propagates to the 3' end resulting in release of the 3' end of the guide to enable full base pairing (Ameres et al., 2007; Haley and Zamore, 2004). This induces a conformational change in Argonaute (Nakanishi et al., 2012) and results in cleavage of the target between nt 10 and 11 opposite of the guide (Elbashir et al., 2001).

Effect of mismatches on target cleavage

Single and double guide–target mismatches lower the activity of the slicer complex depending on their position (Ameres et al., 2007; Dahlgren et al., 2008; Haley and Zamore, 2004; Pusch et al., 2003; Schwarz et al., 2006). Mismatches in the seed and

3’ complementary region affect target cleavage efficiency, whereas mismatches in the central region result in a complete loss of cleavage (Elbashir et al., 2001; Pusch et al., 2003; Schwarz et al., 2006). Mismatches are only tolerated at positions 1 (the anchor) and 18-21 (the 3’ tail), as these nucleotides do not contribute to target binding (Wee et al., 2012) but instead are required for stable binding of the guide to Argonaute (Elkayam et al., 2012). The anchor is tightly bound in a pocket of the PIWI domain in which several amino acids strongly interact with the 5’-phosphate of the guide, preventing base pairing of the first nucleotide with the target (Elkayam et al., 2012; Nakanishi et al., 2012; Schirle and MacRae, 2012) (Table 1). Securing the anchor ensures proper guide binding while positioning the seed and cleavage site.

Structural basis for seed binding

The seed is bound in a narrow groove in which specific amino acids of the protein interact with backbone phosphates and 2’OH groups of the RNA (Elkayam et al., 2012; Nakanishi et al., 2012; Schirle and MacRae, 2012). Binding takes place in such a way that nt 2 to 6-8 are pre-ordered in an A-form helix with Watson-Crick faces pointed towards the solvent to promote target binding. The pre-ordered helix lowers the entropic cost when the seed of the guide RNA forms a stable duplex with the target (Bartel, 2004).

Differential guide loading

In most RNAi pathways, short duplex RNA molecules of 21-25 nt with 2 nt 3’ overhangs are the precursors for guides (Elbashir et al., 2001; Siomi and Siomi, 2009). Which one of the two strands becomes the guide is determined by both nucleotide composition and the presence of mismatches at either end of the double-stranded RNA (dsRNA) precursor. This differential guide loading process has direct consequences for what becomes the seed (Khvorova et al., 2003; Schwarz et al., 2003), as Argonaute is loaded with the strand that has the least thermostable 5’-end (nt 1-5) and 3’-mid region (nt 12-15). Both mismatches and G:U wobbles in these regions enhance the dsRNA unwinding and promote strand selection (Hibio et al., 2012; Kawamata et al., 2009; Khvorova et al., 2003; Schwarz et al., 2003; Ui-Tei et al., 2008; Yoda et al., 2009).

To summarize, RNAi-based small RNAs are relatively uniform and carry a 6-8 nt seed. The RNA is tightly associated with Argonaute and the seed is pre-ordered in an A-form helix. Target mRNA binding is sequential and starts at the seed. The extent of guide:target pairing and the type of Argonaute determines the fate of the target.

Prokaryotic sRNA

sRNAs in bacteria and archaea are morphologically a highly diverse group of molecules ranging from 50 nt to several kilobases that regulate gene expression through a variety of mechanisms (Georg and Hess, 2011; Storz et al., 2011). Although seed sequences have only been described for *trans*-acting sRNAs (*i.e.* transcribed from a locus other than their target), *cis*-acting antisense sRNAs (asRNA) also nucleate with their targets in specific regions and are therefore briefly discussed below.

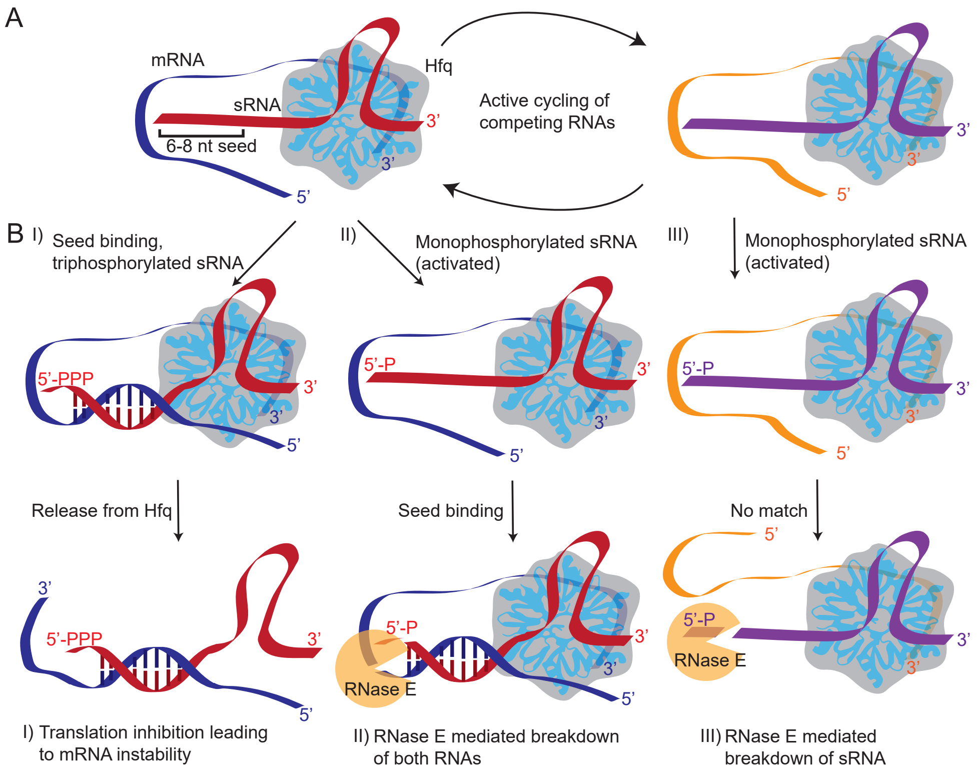

A
mRNA
sRNA
6-8 nt seed
Hfq
3'
3'
5'
Active cycling of competing RNAs
3'
3'
5'
B I) Seed binding, triphosphorylated sRNA
II) Monophosphorylated sRNA (activated)
III) Monophosphorylated sRNA (activated)
5'-PPP
3'
3'
5'
5'-P
3'
3'
5'
5'-P
3'
3'
5'
Release from Hfq
Seed binding
No match
3'
5'-PPP
3'
5'
5'-P
RNase E
3'
3'
5'
5'
5'-P
RNase E
3'
3'
I) Translation inhibition leading to mRNA instability
II) RNase E mediated breakdown of both RNAs
III) RNase E mediated breakdown of sRNA

Figure 2 - Model of interactions between sRNA/mRNA and Hfq. A) sRNA (red/purple) and mRNA (blue/yellow) are bound on opposite sides of the Hfq protein (grey-blue) (Bandyra et al., 2012; Vogel and Luisi, 2011). The sRNA seed is presented to mRNA for potential base pairing. Competing RNAs are rapidly cycled on Hfq and this is hypothesised to facilitate pair-matching. BI) 6-8 nt seed of sRNA binds to mRNA in the vicinity of the RBS and might in some cases be extended to 7-12 base pairs (Balbontín et al., 2010; Gottesman and Storz, 2011; Kawamoto et al., 2006; Vogel and Luisi, 2011; Waters and Storz, 2009). After duplex formation, Hfq releases the bound RNAs which leads to a stable inhibition of translation and ultimately degradation of the naked mRNA (Fender et al., 2010; Moll et al., 2003; Møller et al., 2002). BII) 5'-PPP of the sRNA is converted to 5'-P (activated) by an unknown pathway, leading to recruitment and allosteric activation of RNase E (yellow). After seed binding, the two RNA species are separately degraded by RNase E. BIII) 5'-PPP of the sRNA is converted to 5'-P (activated) by an unknown pathway, leading to recruitment and allosteric activation of RNase E. If the RNAs do not match, only the sRNA will be degraded by RNase E, releasing Hfq. Figure adapted from (Bandyra et al., 2012).

Trans-acting sRNA

Trans-acting sRNAs are typically 50-300 nt long (Table 1). Binding of the sRNA to an mRNA can lead to translational repression by occlusion of the ribosome binding site (RBS) or translation start site (Figure 2) (Bouvier et al., 2008; Gottesman, 2004). Alternatively, sRNA binding can cause mRNA instability by eliciting RNase E-mediated mRNA breakdown (Bandyra et al., 2012). *In vitro* studies with *E. coli* RNase E, Hfq and either MicC or an artificial sRNA suggest that RNase E-mediated breakdown of target mRNA depends on the conversion of the 5'-triphosphate of the sRNA into a 5'-monophosphate (Bandyra et al., 2012) (Figure 2). Apart from the fate of the target, the phosphorylation state of sRNA also determines its own turnover, as monophosphorylated sRNAs are broken down by RNase E (Figure 2). Another recent study shows that the position of stem loops within the sRNA protects against RNase E-mediated degradation (Shao et al., 2013). Interestingly, recent studies show that sRNA can also activate mRNA through seed pairing interactions by shielding it from RNase E (Frohlich et al., 2013; Papenfort et al., 2013).

Although there are examples of sRNAs with predicted pairing regions of 10-25 nt, most *trans*-acting sRNAs base pair over a stretch of only 7-12 nt with their target mRNAs (Bandyra et al., 2012). For some sRNAs base pairing of just 6-8 nt is sufficient to repress gene expression (Balbontín et al., 2010; Gottesman and Storz, 2011; Kawamoto et al., 2006; Waters and Storz, 2009). The minimal stretch of sRNA nucleotides that needs to base pair with the target to induce repression, and which does not tolerate mismatches, is generally referred to as the seed for prokaryotic sRNAs.

sRNA structure and seed

Most sRNAs show a remarkable variability in length, sequence and structure. Enterobacterial sRNAs, however, have a more defined architecture comprising a 5' seed, an A/U-rich binding site for the RNA chaperone Hfq (see below) and a structured 3' end containing a poly(U) stretch that binds Hfq. Both a poly(U) stretch and an occupied Hfq binding site help to confer resistance against exonucleases (Guillier and Gottesman, 2008; Papenfort et al., 2010; Storz et al., 2011) (Figure 2). The functional modularity of this type of sRNA has been demonstrated by showing that target recognition occurs when a seed is transplanted to an unrelated scaffold RNA (Papenfort et al., 2010; Pfeiffer et al., 2009). For most sRNAs the seed is predicted to be unstructured (Peer and Margalit, 2011) and it has been shown for the *E. coli* sRNA Spot 42 that only unstructured parts contribute to regulation (Beisel et al., 2012). However, some studies have reported sRNAs that rely on stem loops for their activity. For the enterobacterial CyaR, *Vibrio harveyi* Qrr sRNAs and several staphylococcal sRNAs the seed has been predicted or shown to be located within a stem loop (Bohn et al., 2010; Geissmann et al., 2009; Johansen et al., 2008; Papenfort et al., 2008; Shao et al., 2013).

Role of Hfq

Hfq is a hexameric ring-like protein required *in vivo* for the function of most sRNAs studied today (Hussein and Lim, 2011; Moon and Gottesman, 2011; Vogel and Luisi, 2011). Although it has a promiscuous affinity for many cellular RNAs, it preferably binds sRNA and mRNA molecules (Chao et al., 2012; Christiansen et al., 2006). Three RNA binding surfaces with distinct specificities have been described for Hfq: a proximal face of the ring with a preference for U-rich sequences such as sRNAs containing a 3' poly-U stretch, a distal face preferring A-rich sequences or ARN repeats found in mRNAs, and a lateral side (Sauer, 2013).

Available data supports a model in which Hfq is important for the initial formation of an sRNA–mRNA duplex by binding both RNA molecules on either face of the ring, bringing the seed in close proximity of a potential target (Figure 2). Although Hfq is an abundant protein, it has been proposed that both sRNAs and mRNAs are in constant competition for binding sites (Fender et al., 2010), causing active cycling of competing RNAs on Hfq (Figure 2). Moreover, Hfq creates free binding sites by releasing sRNA-mRNA duplexes, which then remain stable (Kawamoto et al., 2006; Moll et al., 2003; Møller et al., 2002). While the majority of studied sRNAs contain an Hfq binding site and are Hfq-dependent for their function, some sRNAs repress translation in the absence of Hfq. For example, the sRNA RyhB from *Escherichia coli* remains functional without an Hfq binding site in the absence of Hfq (Hao et al., 2011). Although most sRNAs carry a seed, there is no evidence for a direct structural association between the seed and Hfq that facilitates target binding, as is the case for RNA silencing based small RNAs with Argonaute. Instead, the seed of sRNAs seems to produce RNA duplexes based on predicted thermodynamic stability alone (Hao et al., 2011)i. The formation of these complexes is catalyzed by Hfq and is achieved by presenting the RNAs to each other, increasing the likelihood of pairing, reducing either the entropic penalty of duplex formation, or modifying sRNA structure for increased binding site exposure (Soper et al., 2011; Vogel and Luisi, 2011).

Cis-acting antisense sRNA

asRNA-mediated gene regulation employs a variety of mechanisms, many of which result in degradation of the asRNA–mRNA duplex by RNase III or RNase E (Brantl, 2007; Georg and Hess, 2011). Although asRNAs generally exhibit extensive complementarity to their targets, base pairing is often initiated in a 'kissing complex' in which transient pairing occurs between exposed loops of the asRNA and the mRNA (Brantl, 2007). Base pairing of these short stretches is subsequently extended to the flanking regions, but the extensive secondary structure of the asRNA often prevents complete duplex formation (Han et al., 2010). While asRNAs are generally not associated with Hfq, they are sometimes associated with other chaperones, protecting the asRNA from RNase E degradation and/or stabilizing the kissing complex (Brantl, 2007). Although the short stretches of the asRNA are generally not referred to as seed, the nucleation-type pairing mechanism bears some similarity to seed-containing small RNAs. Since the sequences of asRNAs co-evolve directly with their targets, the seed is usually difficult to identify by sequence conservation.

In summary, sRNAs are diverse in length and structure, but many employ a 6-8 nt seed for target recognition. The seed is not directly associated to protein and is in most cases free of secondary structures. The pairing of the seed to target mRNA is often catalyzed by Hfq by bringing sRNAs and target mRNAs in close proximity.

CRISPR RNA

crRNAs guide the prokaryotic CRISPR-Cas immune system that protects the cell from foreign nucleic acid invaders (Box 1). Mature crRNAs are 35-70 nt long and are only complementary to their target over 20-44 nt (Table 1). This segment is flanked, at the 5’ end, 3’ end or both, by ribonucleotides transcribed from the repeats (Box1). Similar to siRNA and miRNA which are bound by Argonaute, crRNAs are bound by Cas proteins to form ribonucleoprotein complexes. There is systematic evidence for a seed sequence in crRNAs from Type I and Type II CRISPR-Cas systems, while no evidence currently supports the existence of a seed in Type III systems (Gudbergsdottir et al., 2011; Millen et al., 2012). Therefore Type III systems are not further discussed.

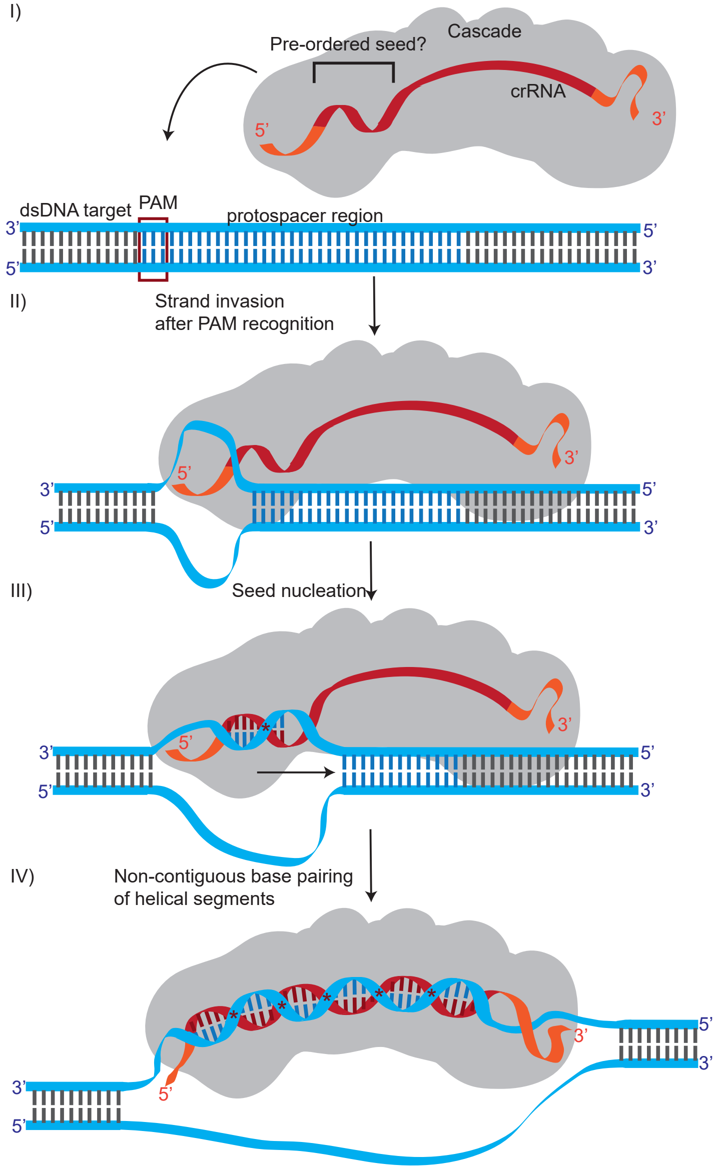


Figure 3 - Model of Cascade target binding. I) A Cascade complex (grey) with bound crRNA, consisting of repeat derived nucleotides (orange) and spacer derived nucleotides (red), containing the seed. The seed is potentially structurally pre-ordered in a helical conformation to promote base pairing. Cascade recognizes target site (protospacer, blue bases) in dsDNA (blue/grey) on the basis of a PAM sequence (red box) (Sashital et al., 2012). II) PAM recognition promotes strand invasion by Cascade. III) Base pairing nucleates at the seed (position 1-8), with the exception of position 6, which is unavailable for base pairing. The PAM is located directly next to the target site and is not involved in base-pairing with the crRNA (Westra et al., 2013). IV) Both ends of the crRNA remain bound to Cascade, so the target DNA strand cannot wrap around the crRNA and form a full duplex. Instead, the crRNA binds the target DNA in non-contiguous helical segments (interruptions are indicated by asterisks). The non-target strand of the dsDNA target is displaced, creating an R-loop. The steps involved in subsequent target cleavage are not included in this figure.

Seed of Type I CRISPR-Cas systems

An interrupted seed of 7 nt positions has been found for Type I-E and Type I-F crRNAs in *E. coli*, *Pseudomonas aeruginosa* and *Pectobacterium atrosepticum*, comprising positions 1-5 and 7-8 of the spacer segment of the crRNA (Cady et al., 2012; Semenova et al., 2011; Vercoe et al., 2013; Wiedenheft et al., 2011b) (Table 1 and Figure 3). Generation of point mutations in the seed region allows viruses and plasmids to escape CRISPR immunity, because these mutations greatly reduce the binding affinity of crRNA-loaded Cascade-like complexes for their DNA targets. Furthermore, compared to the remainder of the guide, the seed sequence has a higher affinity for target oligonucleotides (Wiedenheft et al., 2011b)|. This effect was abrogated when the Cas proteins were removed, indicating that the higher affinity is not an inherent property of the crRNA. Type I-B crRNAs from *Haloferax volcanii* carry 9 nt seed sequences and, strikingly, position 6 was again found not to be part of the seed (Maier et al., 2013). This points at a common structural conformation of the seed sequences from these Type I CRISPR-Cas systems in which position 6 is not base pairing with the target (Jore et al., 2011a), most likely due to a disruption of the crRNA-target DNA helix at this position (Wiedenheft et al., 2011a). Above phenomena suggest a structural pre-ordering of the seed, favoring duplex formation, as seen in RNAi based small RNAs.

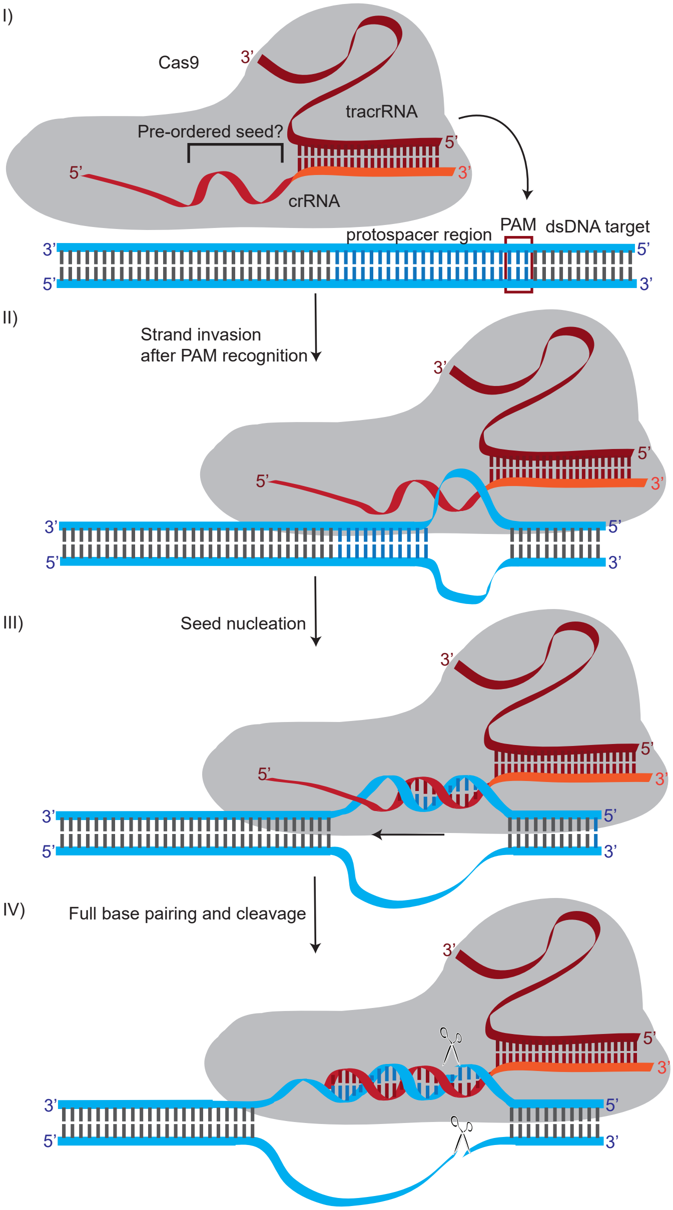


Figure 4 - Model of Cas9 target binding. I) *Streptococcus pyogenes* and *S. thermophilus* Cas9 (Gasiunas et al., 2012; Jinek et al., 2012) protein (grey) with bound crRNA, consisting of repeat derived nucleotides (orange) and spacer derived nucleotides (red) containing the seed. Cas9 furthermore carries a tracrRNA (dark red), which is paired to the repeat of the crRNA. The seed is potentially structurally pre-ordered in a helical conformation to facilitate base pairing. Cas9 likely recognizes the target site (protospacer, blue bases) in dsDNA (blue/grey) on the basis of a PAM sequence (red box). II) PAM recognition might promote strand invasion by Cas9. III) Base pairing nucleates at the seed (last 12 nt of spacer). The PAM is located on the DNA downstream of the seed and is not involved in base-pairing. IV) Base pairing is extended over the entire spacer region of the crRNA possibly through release of the 5' end of the crRNA, while the non-target strand of the dsDNA target is displaced, creating an R-loop. Target DNA cleavage then occurs in both strands by distinct nuclease domains of Cas9.

Seed of Type II CRISPR-Cas systems

Cas9 is guided by Type II crRNAs containing an uninterrupted 12 nt seed sequence at the 3’ end of the spacer segment (Jiang et al., 2013) (Table 1 and Figure 4). Seed-target DNA mismatches affect target DNA binding and result in loss of cleavage activity, which normally takes place within the seed, 3 nt from its 3’ end (Garneau et al., 2010; Gasiunas et al., 2012; Jinek et al., 2012). Especially mutations near the DNA cleavage site allow viruses and plasmids to escape from CRISPR immunity (Deveau et al., 2008; Jiang et al., 2013; Sapranauskas et al., 2011; Sun et al., 2013). Similar to Type I systems it was found that the affinity of the seed for target DNA is higher than other regions of the guide when associated with the Cas9 protein. Again, this might point to a structural pre-ordering of the seed. The RNA-directed DNA nuclease activity of Cas9 from a number of Bacterial species such as *Streptococcus pyogenes* has recently revolutionized genome editing in many eukaryotes (Cong et al., 2013; Mali et al., 2013; Ran et al., 2013b). The single nucleotide distinguishing feature of seed sequences could be harnessed in gene therapy strategies of dominant heterozygous mutations where cleavage of only the mutant chromosome is desired.

Protospacer adjacent motif and sequential target binding

Apart from the seed sequence, the protospacer adjacent motif (PAM) is another sequence element that typifies Type I and II CRISPR-Cas systems. The PAM is a conserved nucleotide sequence located near the seed-matching sequence in the target DNA, just outside the crRNA pairing region (Figure 3 and 4). During CRISPR interference, the PAM serves to differentiate target DNA from non-target DNA including the host CRISPR locus (Sashital et al., 2012; Westra et al., 2013). The short distance between the PAM and seed in Type I and II systems and the importance of the PAM for high affinity target DNA binding of Cascade and Cas9 (Gasiunas et al., 2012; Jinek et al., 2012; Semenova et al., 2011; Sinkunas et al., 2013) has led to the idea that PAMs might be recognized by Cas9 and Cascade in the initial phase of target search (Figure 3 and 4) (Jinek et al., 2012; Sashital et al., 2012). This is then followed by local dsDNA unwinding at the PAM, allowing strand invasion and base pairing of the crRNA seed, and progressing into base pairing of the 3’ remainder of the guide to form a full R-loop. PAM sequences therefore seem to be a necessary adaptation to efficiently recognize dsDNA targets by reducing the complexity of the target search process from checking all nucleotide sequences to checking for seed-matches next to fixed nucleotide sequences. A strikingly similar target search strategy is employed by Group II introns during retrohoming. Group II introns are mobile catalytic RNAs that form a complex with intron-encoded proteins (Lambowitz and Zimmerly, 2011). The complex scans dsDNA until it recognizes a short nucleotide motif (*e.g.* TNGAN$_{23}$T for *Lactococcus lactis* LtrA). These sites are then locally melted to allow two exposed RNA loops of the ribozyme to base pair over stretches of 4 to 9 nt with a target

DNA locus before the intron integrates. This mechanism resembles the target search strategy of Type I and II CRISPR systems and therefore represents an interesting case of convergent evolution of an RNA-guided dsDNA target search mechanism.

To summarize, the crRNA is tightly associated with Cas proteins. The seed is 7-12 nt long and possibly pre-ordered in a structure favoring duplex formation with target DNA. The PAM in the target DNA and seed cooperate in a sequential manner to maximize the efficiency of the target search process.

Role of seed in target search

To protect cells from mobile genetic invaders, siRNAs and crRNAs need to identify and neutralize their targets before they replicate and proliferate, while miRNAs and sRNAs must facilitate rapid response to changing environmental conditions. To achieve effective recognition of target nucleic acids, small guide RNAs make use of seed sequences of approximately 6-12 nt. By pairing only short regions, small guide RNAs can rapidly associate with and dissociate from potential targets. Too short pairing regions, however, are actually detrimental for association rates. In fact, it has been shown *in vitro* that a minimum of 7 contiguous base pairs is required for the rapid annealing of short DNA and RNA oligonucleotides (Cisse et al., 2012). Furthermore, advantageous kinetics through shortening of the pairing region comes at the cost of sequence specificity. This is important because guide RNAs will encounter many non-targets with partial sequence identity. The length of the seed therefore seems to be a trade-off between optimal kinetics during initial target scanning and sufficient sequence specificity. The scanning process itself likely involves sensing base pairing of the seed with a potential target site and hopping, jumping, or sliding to the next potential target site (Gorman and Greene, 2008). Structurally, the length of the seed in RNA interference based RNAs and crRNAs may be determined by topological constraints that result from fixing both ends of the guide RNA to the protein complex as this restricts helical duplex formation between the guide and target. The length of the seed may therefore depend on the number of nucleotides available for interacting with the target in the span of part of a single helical turn (Wiedenheft et al., 2011b). Longer uninterrupted seed pairing interactions would require strand twisting to form a helical duplex, which is difficult to envisage without the release of one end of the guide from the complex, and moreover, would seem inefficient during target scanning. Cascade deals with these topological constraints in a different way by pairing its guide in interrupted segments to the target. These segments allow pairing without twisting the guide and target strands thereby avoiding topological complications (Figure 3) (Wiedenheft et al., 2011a). sRNAs achieve a high efficiency of scanning in a completely different way by rapidly cycling sRNAs and mRNAs on the Hfq platform, bringing the seed in close proximity to many potential target RNAs.

Tolerance to mismatches in the seed

All three small RNA systems are highly sensitive to mismatches within the seed. Whereas miRNAs and siRNAs may overcome seed mismatches using compensatory pairing, crRNAs and sRNAs lose their function when mismatched at seed positions (Balbontín et al., 2010; Cady et al., 2012; Kawamoto et al., 2006; Maier et al., 2013; Papenfort et al., 2010; Sapranauskas et al., 2011; Semenova et al., 2011; Vercoe et al., 2013). In CRISPR-Cas, instead of directly interfering with this class of mutated invaders, seed mutants trigger priming: a process in which new spacers against the same target are integrated in the CRISPR array to restore immunity of the host (Datsenko et al., 2012).

Target site evolution

Base pairing of only the seed causes very different downstream effects in the three RNA guided systems. Whereas crRNAs and siRNAs require pairing of nearly the complete guide sequence to support target interference, miRNAs and sRNAs require only seed pairing to interfere with their targets. Contrary to crRNAs which are meant to contain a perfect and maximum degree of complementarity, seed-target interactions of miRNAs and sRNAs have co-evolved to form the desired degree of interaction. The seed is generally the most conserved part in miRNAs and in sRNAs. Moreover, the short length of the seed enables a single miRNA, siRNA, or sRNA to regulate a multitude of genes with conserved seed matching sites, reducing the number of small RNAs required to regulate a high number of genes (Beisel and Storz, 2010; Brennecke et al., 2005; Friedman et al., 2009; Guillier and Gottesman, 2008; Papenfort and Vogel, 2009). Genes that are part of the same pathway or that carry out related cellular functions can therefore be regulated simultaneously. Furthermore, a short conserved seed facilitates the process in which mRNAs are added or removed from small RNA regulons by target site mutagenesis (Papenfort et al., 2012). As such, guide RNAs contribute to the evolution of gene regulatory networks, which is thought to be vital for the evolution of biological complexity (Carroll, 2008; De Robertis, 2008; Shubin et al., 2009).

Concluding remarks

Although referred to by the same word, seed sequences of the three classes of small guide RNAs discussed in this review differ substantially in many ways. They not only have a different sequence and structure, associate with different proteins, and tolerate mismatches to a different degree, they also target different nucleic acid types. Yet, their analogous role of probing potential target nucleic acids for complementarity must be the shared basis of an efficient target search process in the crowded environment of a cell, allowing finding a needle in a haystack by planting the seed.

Acknowledgements

This work was financially supported by a KNAW Beijerinck premium and NWO Vidi grant to S.J.J.B. (864.11.005). We thank M. Vlot and J. van der Oost (Wageningen University), and L. Loeff and C. Joo (TU Delft) for critically reading the manuscript.

Glossary

Argonaute: Key protein in RNAi, utilizes small RNA guides to bind and/or cleave complementary RNA targets.

Bulge: Regions in which one strand of a helix has “extra” inserted bases with no counterparts in the opposite strand.

Cas9: crRNA-guided DNA endonuclease that occurs in Type II systems. It uses crRNA to find and bind double stranded target DNA to make a double stranded break.

Cascade: CRISPR associated complex for antiviral defense. These complexes in Type I CRISPR-Cas systems carry the crRNA and use it as a guide to find and bind complementary DNA. Base pairing of crRNA to DNA results in the formation of an R-loop.

CRISPR-Cas: Clustered regularly interspaced short palindromic repeats. Genomic array of conserved repeat sequences interspaced by unique, invader derived spacer sequences. Together with CRISPR-associated (*cas*) genes they form the CRISPR-Cas system, an adaptive immune system in prokaryotes against foreign nucleic acid invaders.

crRNA (CRISPR RNA): Small guide RNA transcribed from the CRISPR array which mediate target nucleic acid recognition and destruction.

Group II introns: Class of self-catalytic ribozymes encoded by mobile genetic elements that can splice in or out of the genomes.

Guide RNA: Small RNA that recognizes cognate target molecules by Watson-Crick base pairing and directs associated proteins to their target.

Hfq: RNA chaperone in bacteria with a preference for binding sRNA and mRNA.

Kissing complex: Unstable loop-loop interaction between sRNA and mRNA where base pairing between the two RNAs initiates.

miRNA (microRNA): Guides Argonaute proteins, derived from small hairpin structured RNAs.

PAM (Protospacer adjacent motif): Conserved nucleotide motif that is found next to target DNA sequences (protospacers) in Type I and II CRISPR-Cas systems.

piRNA (PIWI-interacting RNA): Interacts with PIWI proteins, biogenesis poorly understood.

R-loop: Formed upon base pairing of crRNA guided complexes with double stranded target DNA. The crRNA forms Watson-Crick base pairs with complementary strand of DNA, while the non-complementary DNA strand is displaced and remains single stranded.

RNAi (RNA Interference): Eukaryotic pathway in which small RNA molecules guide Argonaute family proteins to mediate mRNA binding and/or cleavage.

RNase E: Bacterial endoribonuclease that cleaves single stranded RNA in A- and U-rich regions.

RNase III: Ribonuclease family that cleaves dsRNA.

siRNA (small interfering RNA): Guides Argonaute proteins, derived from duplex RNAs or long RNA hairpins.

sRNA (Small regulatory RNA): Small RNA regulating gene expression by interacting with mRNA. The RNA can be *cis*- (same genetic locus as target) or *trans*-encoded (derived from another locus than the target).

Wobble base pair: Non-Watson-Crick base pairing in RNA-RNA or RNA-DNA duplexes (e.g. G-U pairing).

Box 1. Classification and mechanism of CRISPR-Cas systems

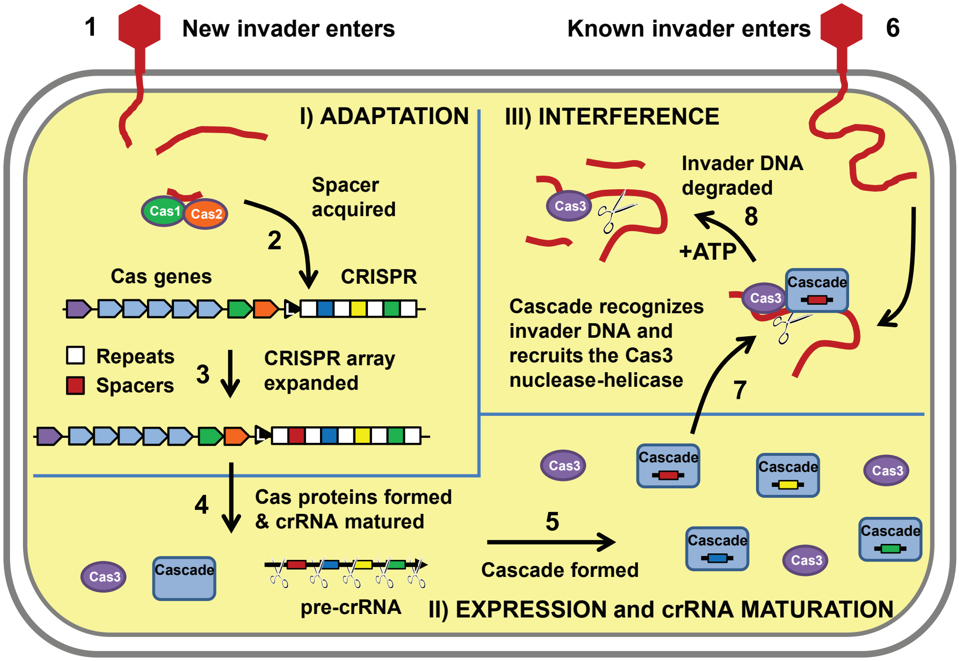


Figure I (in Box 1): Mechanism of CRISPR-Cas immunity depicted for Type I. In the adaptation stage the microbial host encounters a new virus or conjugative plasmid. Genetic material of the invader is recognized as foreign, processed, and somehow integrated into the CRISPR locus to form an additional memory unit. The expression stage involves transcription of the CRISPR locus into a long precursor CRISPR RNA (pre-crRNA), which is cleaved in the repeats by the dedicated endoribonucleases Cas6e. The mature crRNA ends up bound to Cascade to serve as a guide to recognize invader nucleic acids in the interference stage. When invader nucleic acids are identified through base pairing of the crRNA, they are cleaved by Cas3 preventing further virus or plasmid proliferation.

Clustered regularly interspaced short palindromic repeats (CRISPR) refers to genomic loci consisting of DNA repeats, interspaced by invader derived DNA sequences (termed spacers) which serve as a memory of an adaptive immune system. A leader sequence, containing promoter elements, is located upstream of the CRISPR array. CRISPR loci are often flanked by a set of CRISPR-associated (*cas*) genes, which encode the protein machinery of the immune system. Three substantially different types of CRISPR-Cas systems exist: Type I systems encode a Cascade-like multi-protein complex, which uses a small crRNA to specifically recognize target DNA by base pairing the crRNA to one of its strands (R-loop). After binding the target DNA is degraded by a recruited effector nuclease Cas3 (Westra et al., 2012c) (Figure I). Type II systems encode Cas9 which can bind and cleave target DNA molecules using a crRNA guide and a second small RNA called tracrRNA (Deltcheva et al., 2011; Gasiunas et al., 2012; Jinek et al., 2012). Type III systems encode a multi-Cas protein complex which utilizes crRNA to target DNA (Type IIIA) (Marraffini and Sontheimer, 2008) or RNA (Type IIIB) (Hale et al., 2009).

The molecular mechanism of CRISPR-Cas is divided into three functional stages (Figure I): adaptation of immune specificity, expression and maturation of crRNAs, and target interference. During the adaptation stage, pieces of invader DNA are incorporated at one end of the CRISPR locus (Barrangou et al., 2007). Although it is known that *cas1* and *cas2* are essential for adaptation (Datsenko et al., 2012; Yosef et al., 2012), the mechanism of spacer acquisition remains largely unknown. During the expression stage, the CRISPR array is transcribed into one long precursor crRNA, subsequently cleaved in the repeats by dedicated Cas proteins or RNase III and sometimes further trimmed to yield mature crRNAs (Brouns et al., 2008a; Deltcheva et al., 2011; Haurwitz et al., 2010; Sashital et al., 2011). These small guide RNAs are then loaded into Cas protein complexes to guide detection of memorized invaders through crRNA complementarity, resulting in target inactivation (Jinek et al., 2012; Westra et al., 2012c).

Box. Outstanding questions

- Differences in effectiveness of individual small RNAs can often be poorly explained. Could these differences be due to regions flanking target sites, and/or by differences in non-base pairing parts of guide RNAs?
- How do sRNAs regulate gene expression in prokaryotes lacking Hfq, including archaea?
- Is the seed pre-ordered in a helical conformation in crRNA-Cas complexes such as Cascade and Cas9?
- How do Type III CRISPR-Cas systems find their DNA or RNA targets in the absence of a PAM and/or seed?

Chapter 3

Electrophoretic mobility shift assay of DNA and CRISPR-Cas ribonucleoprotein complexes

Tim Künne[1], Edze R. Westra[2] and Stan J.J. Brouns[1]

[1]Laboratory of Microbiology, Department of Agrotechnology and Food Sciences,
Wageningen University, Dreijenplein 10, 6703 HB Wageningen, The Netherlands.
[2]Biosciences, University of Exeter, Penryn, TR10 9EZ, UK

This chapter has been published as:

Künne T, Westra ER, Brouns SJ 2015. Electrophoretic Mobility Shift Assay of DNA and CRISPR-Cas Ribonucleoprotein Complexes. Methods Mol Biol, 1311:171-84

Abstract

The Electrophoretic Mobility Shift Assay is a straightforward and inexpensive method for the determination and quantification of protein-nucleic acid interactions. It relies on the different mobility of free and protein-bound nucleic acid in a gel matrix during electrophoresis. Nucleic acid affinities of crRNA-Cas complexes can be quantified by calculating the dissociation constant (K_d). Here we describe how two types of EMSA assays are performed using the Cascade complex from *Escherichia coli* as an example.

Keywords

Electrophoretic mobility shift assay, EMSA, gel shift, binding assay, CRISPR, RNA guide, affinity, Cascade, protein-DNA interaction

Introduction

All CRISPR systems share the common feature of encoding a crRNA-guided ribonucleoprotein complex targeting complementary nucleic acids. Type I systems encode Cascade/crRNA complexes, Type II systems Cas9/crRNA complexes and Type III systems Cmr/crRNA or Csm/crRNA complexes (reviewed in (Westra et al., 2012b)). Invader detection by these complexes is a key step of the CRISPR-dependent immune response. The binding behaviour of these complexes is a key determinant of the activity and specificity of the respective systems and can reveal mechanistic features, such as the seed sequence (Hale et al., 2009; Jinek et al., 2012; Semenova et al., 2011). Examining the binding behaviour of proteins with nucleic acids can be done using various techniques, such as Surface Plasmon Resonance (Biacore), single molecule TIRF (total internal reflection microscopy), Microscale Thermophoresis or Electrophoretic Mobility Shift Assay (EMSA) (Helwa and Hoheisel, 2010; Jerabek-Willemsen et al., 2011; Monico et al., 2013).

Usually EMSA is the method of choice, as it is a relatively straightforward and inexpensive method that generally provides robust and easy to interpret data. EMSAs can be used for simple qualitative analysis, such as identifying target and non-target nucleic acids. Importantly, it can also be used for quantitative analysis, which can reveal binding stoichiometry and affinity (Fried, 1989). To this end, protein and target are brought to binding equilibrium over a range of molar ratios and separated by gel electrophoresis. Free nucleic acid generally migrates faster through the gel than protein bound nucleic acid. This shift in migration is dependent on the bulkiness of the protein and the combination of protein pI and electrophoresis conditions.

EMSAs are easy to perform and do not require specialized equipment. A wide range of conditions can be used, as long as they are compatible with electrophoresis. Furthermore, any nucleic acid can be used as substrate as long as it can be visualized after electrophoresis; the nucleic acid size range spans from single stranded short oligonucleotides to plasmids of several thousand nucleotides. Furthermore, EMSA can be combined with footprinting analyses or competition binding experiments (Fried and Daugherty, 1998; Jore et al., 2011a; Westra et al., 2010; Westra et al., 2012c).

Despite these advantages of EMSA over more specialized techniques, EMSA also has some disadvantages. The main drawback of EMSAs is the fact that the chemical environment of electrophoresis differs from the environment of equilibration. Hence, the binding equilibrium can change at electrophoresis conditions. However, when the sample enters the gel matrix, interactions are usually stabilized by a caging effect, preventing or slowing down further changes (Cann, 1989; Fried and Bromberg, 1997). Still, low-affinity interactions can be lost during electrophoresis while they are maintained in solution. This could lead to underestimation of binding affinity.

Here, we provide two different protocols for EMSA: Plasmid EMSA using agarose gel electrophoresis and short oligonucleotide EMSA using poly acrylamide gel electrophoresis (PAGE). The choice between these two protocols is determined mainly by the nature of the target nucleic acid. The use of plasmids better mimics biologically relevant conditions and allows one to address the influence of DNA topology on binding affinity. However, synthetic probes offer more experimental flexibility. Plasmids are best separated on agarose gels, while shorter nucleic acids are better separated on PAGE gels. When using intermediate sized nucleic acids, PAGE gels are preferred as they offer better resolution. Short probes usually need to be radio-labelled, since they cannot be sufficiently visualized by intercalating dyes. Here, ^{32}P 5’ end labelling is the most used technique, but internal or 3’ labelling is also possible. When using larger target molecules isotope labelling is generally not required, instead standard intercalating dyes are used, which are compatible with both agarose and PAGE gels.

Materials

Agarose EMSA

1. 37 ˚C incubator or water bath and microcentrifuge
2. 11-14 horizontal gel electrophoresis system (Biometra Horizon 11-14) or comparable
3. Electrophoresis power supply
4. UV imager (Syngene GBox or comparable)
5. Purified protein (1-10 mg/ml) (*see* **Note 1**)
6. Plasmid DNA (60 ng/µl)
7. 5x Equilibration buffer (100 mM HEPES pH 7.5, 375 mM NaCl, 5 mM DTT) (*see* **Note 3**)
8. Optional: 100 µM stabilizing probe (*see* **Note 4**)
9. Optional: Competitor nucleic acid (*see* **Note 5**)
10. 1x sodium boric acid (SB) buffer pH 8.3 (8.6 mM sodium borate, 45 mM boric acid) (*see* **Note 6**)
11. Agarose, molecular biology grade (Sigma)
12. 6x DNA loading dye (Thermo Scientific) and DNA size marker (Gene ruler 1kb,

Thermo Scientific)

13. SYBR safe (Thermo Scientific) or Ethidium bromide (Sigma)
14. Reagent grade water

PAGE EMSA

1. Isotope facilities
2. Programmable heat block or incubator and microcentrifuge
3. Vertical PAGE apparatus and casting setup, including glass plates, spacers, combs, clamps, casting stand and running unit (Bio-Rad or comparable)
4. Electrophoresis power supply
5. Phosphor screen (GE Healthcare) or autoradiography film (Kodak) or comparable
6. Phosphor Imager (Bio-Rad PMI)
7. Purified protein (1-10 mg/ml) (*see* **Note 1**)
8. DNA oligonucleotides (100 µM)
9. 5x Equilibration buffer (100 mM HEPES pH 7.5, 375 mM NaCl, 5 mM DTT) (*see* **Note 3**)
10. **Optional:** 100 µM stabilizing probe (*see* **Note 4**)
11. **Optional:** Competitor nucleic acid (*see* **Note 5**)
12. Polynucleotide kinase (PNK) (Thermo scientific). This comes with buffer A (forward reaction) and buffer B (exchange reaction)
13. Phenol, chloroform, isoamyl alcohol mix (25:24:1) (Roth)
14. Nucleotide removal kit (Qiagen) or Sephadex G50 columns (GE Healthcare)
15. ExoI enzyme (Thermo Scientific), supplied with 10x ExoI buffer
16. 5x Tris-borate-EDTA (TBE) buffer (445 mM Tris, 445 mM boric acid, 10 mM EDTA) (*see* **Note 6**)
17. 30% w/v acrylamide-bisacrylamide (29:1) stock solution (Sigma)
18. 10% Ammonium persulfate solution (APS), made by dissolving APS powder (Sigma) in reagent grade water
19. TEMED (N,N,N’,N’-Tetramethylethylenediamine) (Bio-Rad)
20. **Optional:** Gel dryer (Model 583 gel dryer, Bio-Rad or comparable)
21. Blotting paper (Whatman) or comparable

22. Plastic food wrap (Saran Wrap®)
23. Reagent grade water

Methods

Agarose EMSA

Protein-DNA equilibration

1. Set up a pipetting scheme as in Table 1. The amounts shown are based on Cascade (Mw = 405 kDa) and pUC-λ (Mw = 1739 kDa). The amount of DNA per reaction is fixed, while protein is titrated (*see* **Note 2**). For good visualization of the DNA, use 360 ng plasmid per reaction. The amount of protein is calculated based on the desired molar ratio. As a negative control, include a no-protein sample. Amounts can be modified according to the molecular weight of the ribonucleoprotein complex or the target plasmid, in order to keep the same molar ratio. Optional: Include a competitor nucleic acid by pre-mixing this with your target plasmid (*see* **Note 5**). The amount of 5x equilibration buffer in the final reaction is calculated to yield a final 1x concentration (taking into account the salts present in the protein solution). Reactions are brought to a total volume of 30 µl with reagent grade water.
2. Make fresh working stock dilutions of your protein (*see* **Note 1**) in 1x equilibration buffer to fit your requirements.
3. Pipette everything on ice, add water and buffer first, then add protein solution, and last add plasmid solution. Vortex the reaction mixture for 10 seconds and spin down in a microcentrifuge.
4. Incubate at 37 ˚C for 30 min to allow the reaction to equilibrate (*see* **Note 2**). In the meanwhile start preparing the gels.
5. **Optional:** Add 1 µl of a 100 µM stabilizing probe to the reaction **after** equilibration (*see* **Note 4**). Incubate for another 20 minutes at 37 ˚C.
6. **Optional:** At this point, additional procedures, such as enzymatic footprinting can be carried out (*see* **Note 7**).
7. Add 6 µl 6x DNA loading dye to each sample and store on ice until loading on gel (see below).

Table 1: Pipetting scheme examples (Top) Plasmid EMSA for agarose gel. (Bottom) Oligonucleotide EMSA for PAGE gel.

Molar ratio Cascade:DNA	**0**	**1**	**2**	**4**	**8**	**16**	**32**	**48**	**64**	**80**	**96**	**150**	**200**	**250**	**300**	**350**	**400**
Plasmid [µl]	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0
Cascade dilution factor		100x	100x	100x	20x	20x	4x	4x	4x	2x	2x	1x	1x	1x	1x	1x	1x
Cascade [ul]	0.0	1.9	3.7	7.5	3.0	6.0	2.4	3.6	4.8	3.0	3.6	2.8	3.7	4.7	5.6	6.5	7.5
5x equilibration buffer [µl]	6.0	5.6	5.3	4.5	5.4	4.8	5.5	5.3	5.0	5.4	5.3	5.4	5.3	5.1	4.9	4.7	4.5
Water [µl]	18.0	16.5	15.0	12.0	15.6	13.2	16.1	15.1	14.2	15.6	15.1	15.8	15.0	14.3	13.5	12.8	12.0
total volume [µl]	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0
Cascade stock [M]	1.1E-05																
Plasmid stock [M]	3.4E-08																

Molar ratio Cascade:DNA	**0**	**1**	**2**	**4**	**8**	**16**	**32**	**48**	**64**	**80**	**96**	**150**	**200**	**250**	**300**	**350**	**400**
Oligonucleotide [µl]	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Cascade dilution factor		1000x	1000x	1000x	100x	100x	20x	20x	20x	10x	10x	10x	10x	10x	10x	10x	10x
Cascade [ul]	0.0	3.2	6.3	12.6	2.5	5.0	2.0	3.0	4.0	2.5	3.0	4.7	6.3	7.9	9.5	11.0	12.6
5x equilibration buffer [µl]	6.0	5.4	4.7	3.5	5.5	5.0	5.6	5.4	5.2	5.5	5.4	5.1	4.7	4.4	4.1	3.8	3.5
Water [µl]	23.0	20.5	18.0	12.9	21.0	19.0	21.4	20.6	19.8	21.0	20.6	19.2	18.0	16.7	15.4	14.2	12.9
total volume [µl]	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0
Cascade stock [M]	1.1E-05																
Oligo stock (labelled) [M]	3.5E-08																

Preparing and running the gel

1. Prepare a 0.8% (*see* **Note 8**) SB buffer agarose gel by mixing 0.88g agarose with 110 ml 1x SB buffer (*see* **Note 6**). Dissolve the agarose by heating the solution in a microwave. Make sure the agarose is completely dissolved. Do **not** include an intercalating dye. Cast the gel in an 11-14 gel tray or comparable system. Use a comb for 20 µl slots.
2. Assemble the electrophoresis unit and fill the container with 1x SB buffer (*see* **Note 6**) until it covers the gel.
3. Load half of each sample (18 µl) in the slots and add the DNA size marker in the first and the last lane of the gel (this allows to check if the gel ran uniformly). Store the other half of the samples in the freezer (-20 ˚C) as backup.
4. Run the gel at 20 mA for 18 hours (based on Cascade binding to a ~3kb plasmid; less time is needed to separate smaller protein-DNA complexes).
5. Remove the gel from the electrophoresis chamber and put it in a plastic tray.
6. Stain the gel by covering it with 1x SB buffer containing 1:10000 SYBR safe or Ethidium Bromide for 30 min (**important:** mix the 1x SB buffer and SybR safe or Ethidium Bromide well before applying it on the gel. Poorly mixed solutions can yield stains on the gel).
7. Rinse the gel and destain it in dH_2O for 15 min
8. Visualize the DNA using a UV imager (see Table 2 for troubleshooting). Make sure not to saturate the signal anywhere in the gel. In case of quantification, safe the file in an appropriate format for your image analysis software (.sgd file for GBox or .tif file for cross platform analysis). Figure 1 is a typical example of a plasmid EMSA on agarose.
9. Continue to image analysis

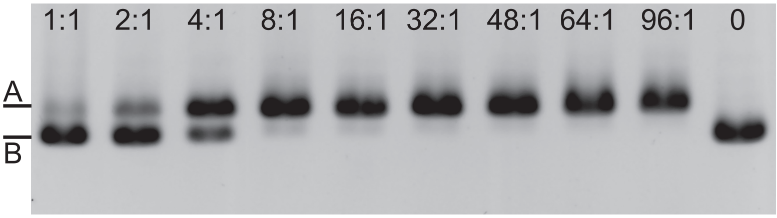


Figure 1. **Agarose EMSA of Cascade with plasmid DNA**. Cascade and plasmid have been incubated at indicated molar ratios to equilibrium and free plasmid (upper band (A)) has been separated from Cascade bound plasmid (lower band (B)) on a 0.8% sodium borate agarose gel. Each lane contains a total of 180 ng plasmid DNA and Cascade protein according to indicated molar ratios of Cascade:plasmid. Gel was run at 8 mA for 18h and post-stained with SYBR Safe for 30 min

PAGE EMSA

5’ ^{32}P Labelling of short oligonucleotide substrate

(**Note:** You can perform this protocol using cold ATP in parallel to conveniently measure DNA concentrations afterwards)

1. To prepare dsDNA oligonucleotides mix the following in a microcentrifuge tube (annealing mix):

1 µl	10 µM forward oligonucleotide
1 µl	10 µM reverse oligonucleotide
2 µl	10x PNK buffer (use either Buffer A, for non-phosphorylated oligonucleotides or buffer B, for phosphorylated oligonucleotide)
16 µl	H_2O

2. If you are using single stranded DNA, replace the reverse oligonucleotide volume with reagent grade water and **Skip** to step 5.
3. Heat to 95 °C for 5 min
4. Slowly cool down to 37 °C (>30min).
5. Add 1 µl PNK and 2 µl y-^{32}P ATP (**Caution**: Exposure to radiation is hazardous, follow safety procedures of your institution).
6. Incubate 1 h at 37 °C.
7. Clean up using nucleotide removal kit or a sephadex G-50 column (preferred method).
8. Elute in 44 µl reagent grade water.
9. Add 5 µl 10x ExoI buffer and 1 µl ExoI (**Important**: Skip this step when using a single stranded DNA substrate)
10. Incubate 30 min at 37 °C
11. Add 50 µl phenol-chloroform-isoamyl alcohol mix.
12. Vortex thoroughly, spin 3 min at maximum (>13000rpm) speed in a microcentrifuge
13. Recover the aqueous phase (usually the upper phase)
14. **Optional:** Determine DNA concentration of cold sample (if available), or hot sample (**Caution:** Only use specified equipment for the use with radiolabelled samples)
15. Store at -20 °C.

Protein-DNA equilibration

1. Set up a pipetting scheme as indicated in Table 1: The amounts shown are based on Cascade (Mw = 405 kDa) and a short dsDNA oligonucleotide. The amount of DNA per reaction is fixed, and protein is titrated. Use 1 µl of 4x diluted labelled oligonucleotide per reaction (The amount depends on the activity of the sample). The amount of protein is calculated based on the desired molar ratio (*see* **Note 2**). As a negative control include a no-protein sample. Amounts can be modified according to the molecular weight of the ribonucleoprotein complex or the target oligonucleotide, in order to keep the same molar ratio. Optional: Include an unlabelled competitor nucleic acid by pre-mixing this with your oligonucleotide (*see* **Note 5**). The amount of 5x equilibration buffer in the final reaction is calculated to yield a final 1x concentration (taking into account the salts in the protein solution). Reactions are brought to a total of 30 µl with reagent grade water.

2. Make fresh stock dilutions of your protein (*see* **Note 1**) in 1x reaction buffer to fit your requirements.

3. Pipette everything on ice, add water and buffer first, then add protein and add the oligonucleotide last. Vortex for 10 seconds and spin down in a microcentrifuge.

4. Incubate at 37 ˚C for 30 min to allow the reaction to equilibrate (*see* **Note 2**). In the meanwhile start preparing the gels.

5. **Optional:** Add 1 µl of a 100 µM stabilizing probe to the reaction after equilibration (*see* **Note 4**). Incubate for another 20 minutes at 37 ˚C.

6. **Optional:** At this point, additional procedures, such as enzymatic footprinting can be carried out (*see* **Note 7**).

7. Add 6 µl 6x DNA loading dye to each sample and store on ice until loading on gel (see below).

Preparing and running the gel

1. Prepare a 5% (*see* **Note 8**) TBE PAGE gel, preferably in a large format (e.g. 20x15):

8.3 ml	30% w/v acrylamide-bisacrylamide (29:1) stock solution
10 ml	5x TBE
31 ml	H_2O
0.8 ml	APS

Assemble the glass plates in a casting setup. Fill with water, to check for leakage. Remove water.

2. Add 50 µl TEMED to the gel solution and pour the gel. Insert the appropriate comb for 20 µl samples

3. Assemble the gel tray and wash the slots with running buffer (0.5x TBE)
4. Pre-run the gel for 20 min at 40mA
5. Load the samples and run for 15 min at 30 mA until the tracking dye has migrated into the gel
6. Run the gels for 3-4h at 20 mA or until the cyan blue dye reaches ¼ of the gel. Carefully rinse and dry the assembly, separate the glass plates and carefully transfer the gel to a blotting paper, wrap the gel and paper in plastic food wrap. Prevent air bubbles and wrinkles.
7. **Optional:** Dry the gel on a paper membrane using a gel-dryer (gives better resolution but there is a risk of breaking the gel)
8. Expose the gel to a phosphor screen or autoradiography film in an exposure cassette. Short time exposure (~2h) can be done at room temperature or 4 °C. Longer exposures of un-dried gels can be performed at -20 ˚C to prevent diffusion. Make sure not to saturate the signal anywhere on the phosphor screen or autoradiography film. Scan the image using a phosphor imager (see Table 2 for troubleshooting). Proceed to image analysis.

Image analysis and quantification

In many cases, EMSA results will need to be analysed quantitatively to obtain the affinity values (K_d) associated with a protein-DNA interaction. Such quantitative analysis requires that the intensities of the bands as well as the background intensity are quantified. To this end, import the picture in an image analysis software (e.g. Genetools for GBox .sgd files). Use the program to automatically or manually quantify the intensity of unshifted and shifted bands in the gel. Apply appropriate background correction for each lane to correct for uneven exposure of the gel. Once the intensities of shifted and unshifted bands have been obtained, the affinity of the interaction can be calculated as described below. Analysis of more complex binding behaviour can be done as described in (Fried and Daugherty, 1998). However, in the case of Cas ribonucleoprotein complexes binding to protospacer sequences, we can assume that each protein (complex) binds only one specific binding site per target molecule, which greatly simplifies the analysis. To obtain binding affinities using this assumption, calculate the fraction of bound substrate (*y*) and the free protein concentration (*x*) for each sample. ‘*y*’ is calculated by dividing the shifted band intensity by the total intensity present in that lane. ‘*x*’ is calculated by multiplying the total protein concentration (as added in the reaction) with the fraction of unbound DNA (1-*y*). Perform non-linear regression using the formula $y = x/(K_d+x)$ to determine the dissociation constant K_d.

Table 2: Troubleshooting

Problem	Possible cause	Potential solutions
No bands visible on gel	Too little or no nucleic acid in reaction	Check nucleic acid concentration, test sensitivity of applied visualization method
	Nucleic acid is degraded	Check substrate integrity on gel. Replace reagents when suspecting nuclease contamination. If possible exclude metal cations and include chelating agent (e.g. EDTA). If working with RNA, work RNase free or include commercial RNase inhibitors.
	Labelling failed or insufficient exposure time	Check functionality of labelling method. If necessary, adapt protocol. Increase exposure time.
No shifted band present	Protein concentration too low	Verify protein concentration, check protein for purity. Increase protein concentration in EMSA.
	Protein is inactive or co-factor missing	Check protein on SDS-PAGE for integrity. Re-purify protein, possibly adapt protocol to get more active preparation. Test co-factors, e.g. divalent metal ions.
	Protein bound nucleic acid migrates at same speed as free nucleic acid	Check migration of protein alone in gel. Use different pH in equilibration and electrophoresis buffer or use a different electrophoresis buffer. If protein is very small compared to nucleic acid, use smaller nucleic acid.
Bands are generally smeared	Gel overheating	Check gel concentration and running buffer conductivity. Lower the above or use lower voltage during electrophoresis.
	High sample conductivity	Reduce salt content in samples
	Bad gel quality	Check even polymerization/solidification. Use fresh, clean components.
		Degas PAGE gels before polymerization.
Only protein bound band is smeared	Too high conductivity in protein solution	Reduce salt concentration in protein stock solution, or concentrate protein and add less volume to reactions.
	Complex dissociates during electrophoresis	Minimize time of sample in loading well. Start with higher voltage to run samples into the gel. Minimize overall electrophoresis time.

Notes

1. Protein: Use purified Cas ribonucleoprotein complexes. Make sure to accurately determine the protein concentration. If additives are required for protein storage that are undesired in the EMSA equilibration, exchange the buffer beforehand by dialysis or a buffer exchange column. Alternatively, keep additive concentration low and protein concentration high to minimize the volume that is added in the reactions. Ideally the protein is dissolved in a buffer very similar

to the equilibration buffer. Make working stock dilutions of the protein in 1x Equilibration buffer.

2. Equilibration: Test a range of molar ratios of protein:DNA (e.g. 0.5:1 up to 400:1). In later experiments, choose ratios that cover the whole dynamic range of binding (i.e. ranging from all DNA in the unbound state to all DNA protein-bound). Test the incubation time needed to reach equilibrium for each protein and each different type of substrate (plasmid, short oligonucleotides etc.). Do this by testing several time points (e.g. 5 min, 30 min, 60 min); when there is no change in the bound fraction between two time points, equilibrium is reached. Typically an incubation time of 30 min at 37 ˚C is used.

3. Equilibration buffer: Choose a buffer for protein-DNA complex equilibration that gives efficient complex formation, is relatively close to physiological conditions and is compatible with the buffer of your protein solution. Do not use salt concentrations that are considerably higher than in your electrophoresis buffer, as this leads to a higher conductivity of the sample compared to the gel and electrophoresis buffer, which will lead to distorted bands. If additives are required for the function of the protein (e.g. metal ions), these should be added to the equilibration buffer. Most common buffers can be used (e.g. Tris, MOPS, HEPES, Phosphate) and total salt concentration is typically around 100 mM.

4. Stabilizing probe: When working with an R-loop forming complex (i.e. the situation where the crRNA base pairs with the target DNA strand, while the non-target strand is displaced and remains single stranded), stabilizing DNA oligonucleotides can be added after equilibration to prevent changes during subsequent steps (Westra et al., 2012a). These probes are complementary to the crRNA as well as to the displaced strand. By binding to the displaced strand these probes prevent re-annealing of the target and non-target strands and thereby inhibit complex dissociation. Furthermore free complex is inactivated by binding of the probe to the crRNA. This step is advisable when additional steps, such as enzymatic footprinting, are performed, or when the interaction of the complex and the target is unstable and tends to dissociate during electrophoresis (e.g. Cascade binding to relaxed plasmid DNA (Westra et al., 2012c)). Choose to add an amount of probe resulting in a 10-fold excess of probe, compared to target- or protein molecules.

5. Competitor nucleic acid: In case of a high nonspecific nucleic acid binding affinity of the protein (independent of crRNA sequence), which might obscure specific binding, it is advisable to add unlabelled competitor nucleic acid. It should have the same nonspecific binding affinity to the protein as the target, while the target should have a higher specific binding affinity to the protein. Concentrations of competitor have to be optimized empirically to find a ratio

that allows clear discrimination of specific and nonspecific binding. If the gel is non-selectively stained by an intercalating dye, make sure the competitor has a different size than the target, to distinguish the target and the competitor.

6. Electrophoresis conditions: Do not leave samples in loading slots over a longer period of time, as complexes might dissociate. The choice of electrophoresis buffer can have an effect on the relative mobility of the protein-DNA complex and the free nucleic acid. Especially with large nucleic acid targets in agarose EMSAs it is possible that the mobility of free- and protein-bound nucleic acid is very similar, yielding poor resolution. In this case it is important to choose electrophoresis conditions such that the protein is not negatively charged. Negative protein charges lead to a co-migration of protein with nucleic acid to the anode, decreasing resolution. Hence, the resolution can be improved by changing to a running buffer with a lower pH. Commonly used buffers are TAE and TBE, while we successfully applied sodium borate buffer (Brody and Kern, 2004), improving the resolution of the shift. Make sure to use the same buffer in all involved steps. Although long electrophoresis times increase the risk of complex dissociation, we got the best results with overnight runs at low currents. We have not seen differences in bound fractions between short or long electrophoresis, while longer runs produced sharper bands.

7. Footprinting is generally an independent alternative to EMSA, but can be used in concert with it. Particularly useful, in the case of crRNA-Cas complexes, is the use of ssDNA specific Nuclease P1 (Jore et al., 2011a) or an endonuclease cleaving in the known binding site (protospacer) on the target DNA (Westra et al., 2012c). Footprinting allows detection and quantification of protein-DNA interactions in solution, without relying on their stability in subsequent electrophoresis

8. An agarose gel percentage of 0.8% is typically used and will give satisfactory results with most large nucleic acid targets. A PAGE gel percentage of 5% is typically used, this can be adjusted to yield the best resolution with the chosen nucleic acid target. Be aware that higher gel percentages might not allow large proteins to enter the gel matrix, trapping protein-DNA complexes in the loading slots or leading to dissociation.

Table 3: DNA size separation by gel percentage

DNA Size Range (Base Pairs)	Acrylamide (%)
100 - 1,000	3.5
80 - 500	5.0
60 - 400	8.0
40 - 200	12.0
10 - 100	20.0
DNA Size Range (Base Pairs)	**Agarose (%)**
1,000 - 30,000	0.5
800 - 12,000	0.7
500 - 10,000	1.0
400 - 7,000	1.2

Acknowledgements

This work was financially supported by a KNAW Beijerinck premium and NWO Vidi grant to S.J.J.B. (864.11.005). ERW received funding from the People Programme (Marie Curie Actions) of the European Union’s Seventh Framework Programme (FP7/2007-2013) under REA grant agreement nº [327606].

Chapter 4

Degenerate target sites mediate rapid primed CRISPR adaptation

Peter C. Fineran[1,2], Matthias J.H. Gerritzen[1], María Suárez-Diez[3], Tim Künne[1], Jos Boekhorst[4], Sacha A.F.T. van Hijum[4,5] , Raymond H.J. Staals[1] and Stan J.J. Brouns[1]

[1]Laboratory of Microbiology, Wageningen University, 6703 HB Wageningen, Netherlands.
[2]Department of Microbiology and Immunology, University of Otago, PO Box 56, Dunedin 9054, New Zealand.
[3]Laboratory of Systems and Synthetic Biology, Wageningen University, 6703 HB Wageningen, Netherlands.
[4]NIZO Food Research, 6718 ZB Ede, Netherlands.
[5]Center for Molecular and Biomolecular Informatics, Radboud University Medical Center, 6500 HB Nijmegen, Netherlands.

This chapter has been published as:

Fineran PC, Gerritzen MJ, Suárez-Diez M, Künne T, Boekhorst J, van Hijum SA, Staals RH, Brouns SJ 2014. Proc Natl Acad Sci U S A, 111(16):E1629-38

Abstract

Prokaryotes encode adaptive immune systems called CRISPR-Cas to provide resistance against mobile invaders such as viruses and plasmids. Host immunity is based on incorporation of invader DNA sequences in a memory locus (CRISPR), the formation of guide RNAs from this locus and the degradation of cognate invader DNA (protospacer). Invaders can escape Type I-E CRISPR-Cas immunity in *Escherichia coli* K12 by making point mutations in the seed region of the protospacer or its adjacent motif (PAM), but hosts quickly restore immunity by integrating new spacers in a positive feedback process termed priming. Here, by using a randomized protospacer and PAM library and high-throughput plasmid loss assays we provide a systematic analysis of the constraints of both direct interference and subsequent priming in *E. coli*. We have defined a high-resolution genetic map of direct interference by Cascade and Cas3, which includes five positions of the protospacer at 6 nt intervals that readily tolerate mutations. Importantly, we show that priming is an extremely robust process capable of utilizing degenerate target regions with up to thirteen mutations throughout the PAM and protospacer region. Priming is influenced by the number of mismatches, their position and is nucleotide dependent. Our findings imply that even out-dated spacers containing many mismatches can induce a rapid primed CRISPR response against diversified or related invaders, giving microbes an advantage in the co-evolutionary arms race with their invaders.

Significance Statement

Bacteria are constantly exposed to foreign elements, such as bacteriophages and plasmids. The CRISPR-Cas adaptive immune systems provide heritable sequence-specific protection against these invaders. To develop immunity, bacteria add segments of foreign nucleic acid to their CRISPR memory. However, phage and plasmid mutants can evade CRISPR-Cas recognition by altering their targeted sequence. CRISPR-Cas responds to evasion by quickly generating immunity by acquiring new pieces of invader genome. We determined that this rapid generation of resistance is promiscuous, with recognition of highly diverged or related elements eliciting new immunity. Our results demonstrate that CRISPR-Cas systems are more robust than previously thought and, not only have a highly-specific resistance memory, but also have a broad ability to identify divergent genetic elements.

Introduction

Bacteria and Archaea are regularly exposed to bacteriophages and other mobile genetic elements, such as plasmids. To control the competing effects of horizontal gene transfer, a spectrum of resistance strategies have evolved in prokaryotes (Samson et al., 2013). One of the most widespread and well-characterised are the CRISPR-Cas (**c**lustered **r**egularly **i**nterspaced **s**hort **p**alindromic **r**epeats-**C**RISPR-**as**sociated) systems, which provide bacterial 'adaptive immunity' (Barrangou, 2013; Marraffini and Sontheimer, 2010a; Richter et al., 2012a; Samson et al., 2013; Sorek et al., 2013b; Terns and Terns, 2011; Westra et al., 2012b; Wiedenheft et al., 2012). Simply, CRISPR-Cas functions in three major steps. Firstly, in a process termed adaptation, short sequences are derived from the invading element and incorporated into a CRISPR array (Fineran and Charpentier, 2012). CRISPR arrays are composed of short repeats that are separated by the foreign-derived sequences, termed spacers. Secondly, CRISPRs are transcribed into a pre-crRNA, which is then processed into short crRNAs, which encompass portions of the repeat(s) and most, or all of the spacer. Finally, as part of a Cas ribonucleoprotein complex, the crRNAs guide a sequence-specific targeting of complementary nucleic acids (for recent reviews see (Barrangou, 2013; Marraffini and Sontheimer, 2010a; Richter et al., 2012a; Samson et al., 2013; Sorek et al., 2013a; Westra et al., 2012b; Wiedenheft et al., 2012)).

CRISPR-Cas systems are divided into three major Types (I-III) and further categorized into subtypes (e.g. I-A to I-F) (Makarova et al., 2011). The mechanisms of both crRNA generation and interference differ between the types and there are even significant differences between closely related subtypes. However, Cas1 and Cas2 are the only two Cas proteins completely conserved across all CRISPR-Cas systems and they are crucial for adaptation in *E. coli* (Datsenko et al., 2012; Makarova et al., 2011; Yosef et al., 2012). The acquisition of new spacers is the most poorly understood stage in CRISPR-Cas immunity, mainly hindered by the paucity of robust laboratory assays to monitor this process (reviewed in (Fineran and Charpentier, 2012)). *Streptococcus thermophilus* is highly proficient at spacer acquisition and provided much of the early insight into adaptation, showing that new spacers are typically acquired at one end of the CRISPR array from either phages (Barrangou et al., 2007; Deveau et al., 2008; Horvath et al., 2008a) or plasmids (Garneau et al., 2010). Recently, spacer acquisition has been detected in a variety of other systems (Cady et al., 2012; Datsenko et al., 2012; Erdmann and Garrett, 2012; Li et al., 2014; Swarts et al., 2012; Yosef et al., 2012). Adjacent to the expanding end of the array is the leader region, which harbours the promoter for pre-crRNA expression and sequences important for spacer acquisition (Diez-Villasenor et al., 2013b; Yosef et al., 2012). Recent studies in *E. coli* in the Type I-E system have shown that spacer acquisition can occur from phages and plasmids when either the Cas1 and Cas2 proteins are overexpressed or if the native *cas* genes are upregulated, due to deletion of *hns* (Datsenko et al., 2012; Diez-Villasenor et al., 2013b; Savitskaya et al., 2013; Swarts et al., 2012; Yosef et al., 2012). The DNA targets (termed protospacers) of newly acquired spacers are consistently flanked by protospacer adjacent motifs (PAMs), with the *E. coli* Type I-E consensus 5'-protospacer-CTT-3'. PAMs were originally identified computationally (Mojica et al., 2009b) and were shown to play a role in interference in an early study (Deveau et al., 2008). The importance of PAMs in the recognition and selection of precursor-spacers (pre-spacers) during adaptation was demonstrated unequivocally using assays that were independent of interference (Diez-Villasenor et al., 2013b; Yosef et al., 2012). The simple overexpression of Cas1 and Cas2, in the absence of other *cas* genes, demonstrated these are the only Cas proteins essential for adaptation and are likely to recognize PAMs (Yosef et al., 2012).

Adaptation consists of two related stages, termed naïve and primed (Fineran and Charpentier, 2012). Naïve adaptation occurs when a bacterium harbouring a CRISPR is infected by a new foreign element that it has not previously encountered. Although the acquisition of a new spacer can result in effective protection from the element, point mutations within the protospacer or PAM allow the element to 'escape' CRISPR-Cas targeting (Deveau et al., 2008; Semenova et al., 2011; Vercoe et al., 2013). This had been viewed as a weakness of CRISPR-Cas interference, but recent studies show that a positive feedback loop called priming occurs, which enables one or more new spacers to be acquired (Datsenko et al., 2012; Savitskaya et al., 2013; Swarts et al., 2012). Specifically, single mutations within either the PAM or the seed region of the protospacer, although inactive for interference, promote the rapid acquisition of new spacers from the same target (Datsenko et al., 2012). Priming is proposed to allow an effective response against viral or plasmid escapees, through the incorporation of new spacers. Unlike naïve adaptation, priming is more complex, and in Type I-E systems requires Cas1, Cas2, crRNA, the targeting complex termed Cascade (**C**RISPR **as**sociated **c**omplex for **a**ntiviral **de**fence – composed of Cse1, Cse2, Cas7, Cas5 and Cas6e (Brouns et al., 2008a; Jore et al., 2011a)) and the Cas3 nuclease/helicase (Datsenko et al., 2012). Interestingly, the vast majority of spacers acquired through priming are derived from the same DNA strand as the original priming spacer (Datsenko et al., 2012; Savitskaya et al., 2013; Swarts et al., 2012). In addition, priming in *E. coli* was abolished by two mutations in the protospacer and PAM region (Datsenko et al., 2012).

In this study, we generated a mutagenic variant library of a protospacer and PAM region and used both individual high-throughput plasmid loss assays and next generation sequencing to determine the limits of both direct interference and indirect interference through priming. Our results demonstrate that direct interference tolerates mutations mostly at very specific positions in the protospacer, whereas priming tolerates extensive mutation of the PAM and protospacer region. The results have wide evolutionary consequences for primed acquisition and could explain the retention of multiple 'older' spacers in CRISPR arrays.

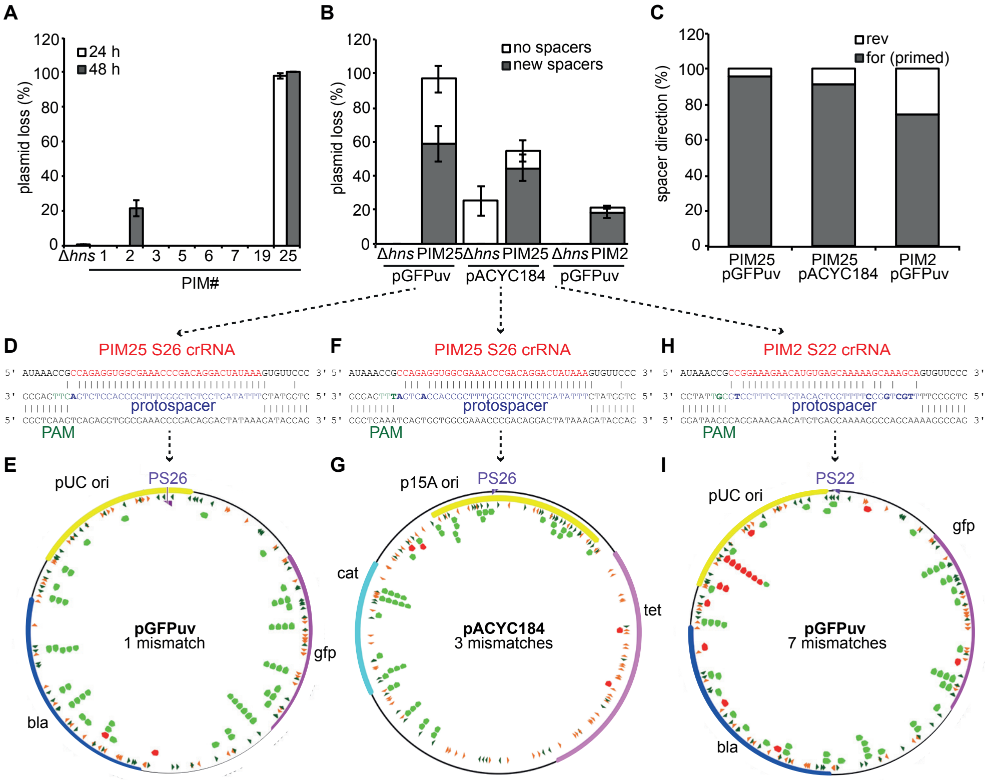


Figure 1 - Preexisting spacers with up to 7 mismatches promote priming. (*A*) Loss of plasmid pGFPuv from Δ*hns* and various pRSF-1b PIM derivatives (Swarts et al., 2012). (*B*) Plasmid loss in Δ*hns*, PIM25 and PIM2 backgrounds for plasmids pGFPuv and pACYC184. The percentage of plasmid-free clones containing no spacers (white) or at least one new spacer (grey) is shown. (*C*) Percentage of spacers derived from forward (priming) or reverse strands of the plasmids from *B*. (*D*) Match of PIM25 S26 crRNA to the protospacer in pGFPuv. (*E*) Mapping of new spacers acquired by PIM25 following loss of pGFPuv. (*F*) Match of PIM25 S26 crRNA to the protospacer in pACYC184. (*G*) Mapping of new spacers acquired by PIM25 following loss of pACYC184. (*H*) Match of PIM2 S22 crRNA to the protospacer in pGFPuv. (*I*) Mapping of new spacers acquired by PIM2 following loss of pGFPuv. In *D*, *F* and *H* the spacer (red), protospacer (blue), PAM (green) and mismatches (bold) are indicated. In *E*, *G* and *I* the protospacer (PS) region is indicated in purple, new forward (primed) spacers in pale green, new reverse spacers in red and all the respective consensus PAMs are shown in green (forward) and orange (reverse).

Results

Plasmid insensitive mutants lose unrelated plasmids via priming

Previously, *E. coli* strain Δ*hns* was shown to acquire spacers from plasmid pRSF-1b when cultured over ~1-2 weeks in the absence of antibiotic selection for plasmid maintenance (Swarts et al., 2012). Naïve spacer acquisition and plasmid loss were not robustly reproducible and the requirement for prolonged cultivation was unclear. Therefore, we tested the ability of the Δ*hns* strains that had acquired new pRSF-1b-derived spacers to lose an alternative plasmid. Eight **p**lasmid **i**nsensitive **m**utants (PIMs) previously isolated after acquiring spacers against pRSF-1b (Swarts et al., 2012) were transformed with pGFPuv, an unrelated plasmid with a different antibiotic resistance marker and origin of replication. Plasmid retention was consistent with previous observations that Δ*hns* did not lose plasmids over two days (Swarts et al., 2012). However, PIM2 and PIM25, displayed different levels of plasmid loss (Fig. 1*A*). PIM2 had 20% plasmid loss, whereas PIM25 had almost 100% plasmid loss by two days.

We were interested why *E. coli* PIM25 displayed heightened plasmid loss. PIM25 previously acquired five new spacers from pRSF-1b, three in CRISPR2.1 and two in CRISPR2.3 (Swarts et al., 2012). Potential targets for these five spacers were assessed against pGFPuv and surprisingly, spacer 26 (spacer two from the CRISPR2.3 leader) matched 31 of 32 bases in pGFPuv and a consensus PAM was present (5'-protospacer-CTT-3'). The mismatch was at +1 in the seed sequence of the protospacer (Fig. 1*D*), a mutation previously shown to enable priming (Datsenko et al., 2012). Most plasmid loss in PIM25 (~60%) was associated with spacer acquisition in CRISPR2.1 and/or CRISPR2.3 and the remaining plasmid-free clones had not acquired spacers (Fig. 1*B*). For 40 PIMs, 54 newly-acquired spacers were sequenced and the protospacer locations in pGFPuv, their orientation relative to the original spacer (S26), and the presence of PAM sequences were determined (Table S1). The vast majority of spacers (52/54; 96%) mapped to the same DNA strand as the S26 spacer, indicative of priming (Fig. 1*C* and 1*E*) (Datsenko et al., 2012; Swarts et al., 2012). Therefore, strains that have acquired spacers targeting a plasmid may have increased loss of an unrelated plasmid. This data was consistent with a priming model, where the single mismatch between S26 (derived from pRSF-1b) and the protospacer in pGFPuv promoted the accelerated acquisition of spacers in a strand-specific manner (Datsenko et al., 2012; Swarts et al., 2012).

At least seven mismatches are tolerated for priming

Datsenko et al. demonstrated that a single mutation in the PAM (-1 position) or in the seed (+1 position) resulted in primed spacer incorporation, but a double mutation

(-1 PAM and +1 seed) abolished priming in the *E. coli* Type I-E system (Datsenko et al., 2012). To examine if the results observed using PIM25 were due to priming, we tested the ability of PIM25 to lose pACYC184 (Rose, 1988). S26 from PIM25 matches pACYC184 but has three mismatches (-1 PAM, +1, +5 in seed; Fig. 1*F*), whereas S29 has 6 mismatches. Since the S26 spacer:protospacer match contained the exact two mutations previously observed to abolish priming (Datsenko et al., 2012), we expected no priming. Surprisingly, ~55% of the PIM25 strains had lost pACYC184, 80% of which showed CRISPR expansion (Fig. 1*B*). In contrast, ~25% of Δ*hns* clones had lost pACYC184, yet none had acquired new spacers (Fig. 1*B*). Of 44 new PIM25-derived strains, the new spacers were strand-specific (43/47; 91%) indicative of priming (Fig. 1*C*, 1*G* and Table S2). We cannot conclude which original PIM25 spacer caused priming since they both are predicted to pair with the same DNA strand. However, it is clear that a minimum of three spacer:protospacer mutations within the PAM and seed region enable primed spacer acquisition.

The question arose why PIM2 showed increased loss of pGFPuv (Fig. 1*A*). Compared with Δ*hns*, PIM2 contains two spacers derived from pRSF-1b, one in each CRISPR (Swarts et al., 2012). The closest spacer:protospacer match to pGFPuv is S22 (the first spacer in CRISPR2.1), which has seven mutated positions (-2 PAM, +2 seed and 5 positions outside of seed; Fig. 1*H*). PIM2 had increased CRISPR-dependent pGFPuv loss compared with Δ*hns* (Fig. 1*B*). Spacers acquired by 37 PIM2 derivatives showed strand-specific features of priming (46/62; 74%; Fig. 1*C*, 1*I* and Table S3), albeit less pronounced than for PIM25. Since PIM2 differs from Δ*hns* by two spacers, it was possible that either spacer was promoting the priming phenotype. To test if S22 was responsible for priming, CRISPR2.1 was replaced by an array with three spacers bearing no homology to pGFPuv. This strain, which lacked S22 but still contained S11 in CRISPR2.3, showed no acquisition during multiple plasmid loss assays, demonstrating that S22 in CRISPR2.1 was required for priming. Therefore, spacers with up to seven mismatches to a protospacer and PAM region enable priming.

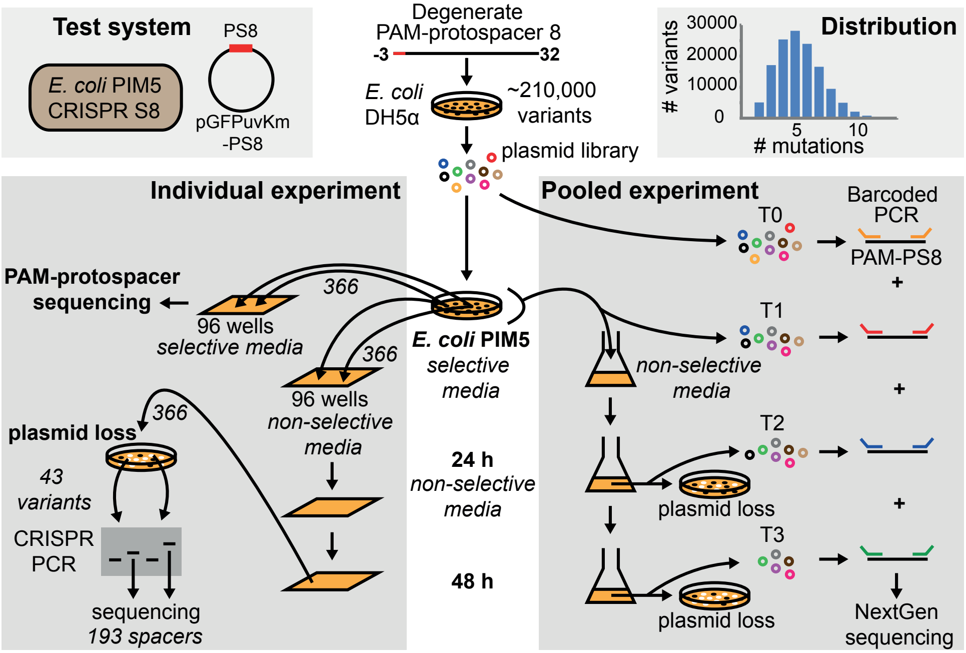


Figure 2 - Experimental design for high-throughput individual and pooled plasmid loss experiments. A *test system* of *E. coli* PIM5 containing protospacer (PS) 8 was selected that would target a pGFPuv-Km plasmid containing a consensus PAM and PS8. A *degenerate PAM-protospacer 8* library of variants was generated in pGFPuv-Km with an average *distribution* of 5 mutations per insert (histogram). For the *individual experiment*, the plasmid library was transformed into PIM5, individual colonies were sequenced, plasmid loss experiments without selection performed and a subset of variants checked for spacer acquisition by PCR and sequencing. In the *pooled experiment*, plasmid DNA was prepared for the original library (T0), which was then transformed into PIM5 (T1) and then passaged for plasmid loss without selection for 24 h (T2) and 48 h (T3). For T0-T3, samples were amplified with barcoded primers, pooled and sequenced.

Individual high-throughput assay reveals that up to 11 mutations support priming

The observation that up to 7 mismatches between a crRNA and protospacer-PAM target region promoted priming, led us to develop a randomized screen to test the limits of priming (Fig. 2). A strain that contained an additional spacer in CRISPR2.1 compared with Δ*hns* was selected (i.e. PIM5) and a library of variant PAM-protospacers (PS8; protospacer 8) with 85% WT nt and 5% alternative nt at each position was generated in pGFPuv (pGFPuv-Km-mPS8). This yielded on average 5 mutations per PAM-protospacer (Fig. 2). Control plasmids were generated that contained either no protospacer (pGFPuv-Km; negative control) or PS8 with a +1 seed mutation (pGFPuv-Km-PS8; priming positive control). Approximately 210,000 transformants of *E. coli* DH5α were grown and a plasmid library prepared. Transformation bias was avoided by using *E. coli* DH5α which lacks any spacers targeting the PAM-protospacer plasmid. The library and control plasmids were introduced into *E. coli* PIM5 and those directly targeted were expected to be eliminated. After 48 h of culturing in non-selective media, the positive priming control showed 82% (±25%) average plasmid loss and the negative control 0.3% (±1.2%) loss.

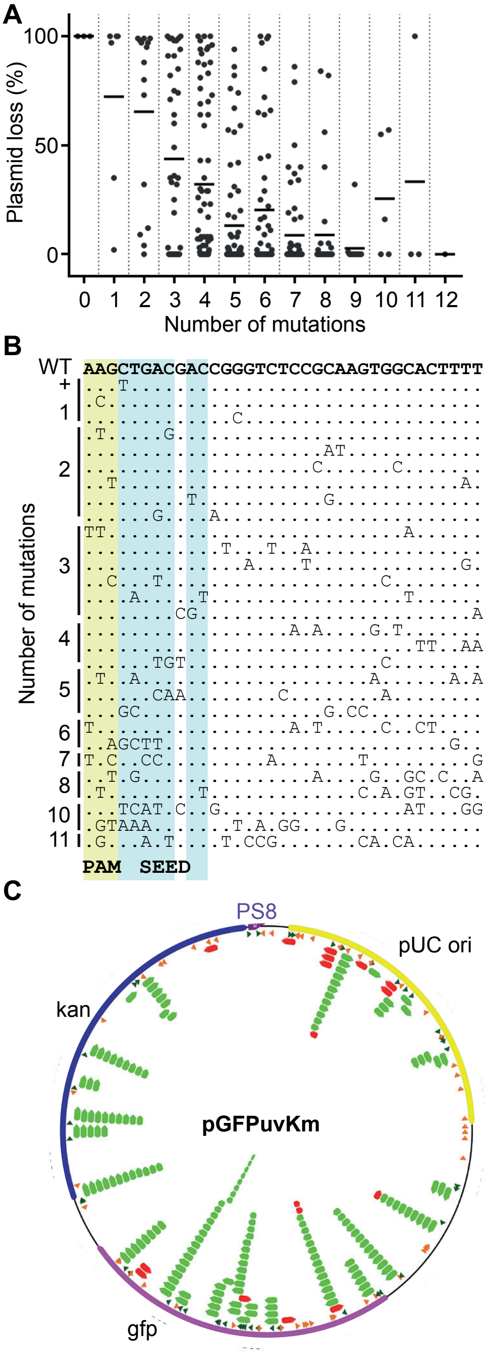


Figure 3 - Up to 11 mutations within the protospacer and/or PAM region promote priming. (*A*) Plasmid loss at 48 h for 366 individual loss experiments relative to the number of mutations in each protospacer and PAM (lines represent average loss for variants with that number of mutations). (*B*) Twenty variants with between 1-11 mutations that displayed plasmid loss via priming. The spacers acquired by these variants are shown in (*C*) and details are in Datasets S1, S2 and Table S4.

A total of 366 individual PAM-protospacer variants were sequenced and each tested for stability upon growth for 48 h without antibiotic selection (Fig. 2 and Dataset S1). Up to 12 mutations were detected and plasmid loss ranged from 0-100% (Fig. 3*A*). On average, the percentage loss decreased with increasing mutations (Fig. 3*A*). Analysis of 20 PIMs from the positive control demonstrated priming (20/24; 83% of spacers analyzed). Colonies (n=4 to 16) from 43 variants exhibiting >10% plasmid loss were checked by PCR for spacer acquisition, which revealed that 88% (38/43) of the variants had acquired new spacers. Twenty eight variants with a range of 1-11 mutations were selected (Fig. 3*B*) and the new spacers sequenced (Table S4 and Dataset S2). The resulting spacers were mapped to pGFPuvKm-PS8 and demonstrated the expected strand bias of priming (175/193; 91%; Fig. 3*C* and Table S4). Remarkably, one of these protospacers deviated from the original PAM and protospacer region by 11 mutations, yet still provoked priming. In the various mutants the locations of these mismatches were throughout the protospacer and PAM region, including the seed sequence (Fig. 3*B*). In conclusion, extensive protospacer mutations, even in the PAM and seed, enable the acquisition of new spacers through priming.

High-throughput overview of the entire dataset

The individual high-throughput assay demonstrated that up to 11 mutations promoted priming (Fig. 3*B*). This screen also enabled us to follow the loss, spacer acquisition, target strand and location for individual mutants (n=366). However, this only represented a small proportion of the mutant library generated. Therefore, we transformed the entire library into *E. coli* PIM5 and performed high-throughput pooled plasmid loss experiments and followed plasmid abundance by deep sequencing (Fig. 2). Plasmids were prepared from i) the initial non-targeting *E. coli* DH5α strain (total library; denoted T0), ii) *E. coli* PIM5 immediately following growth of transformants (i.e. variants surviving direct interference; denoted T1), iii) *E. coli* PIM5 after 24 h of non-selective growth (T2) and iv) *E. coli* PIM5 after 48 h of non-selective growth (T3). Following plasmid preparation, the variant protospacers were amplified by barcoded PCR, the different samples (T0-T3) were pooled and sequenced. In the total library, the distribution of mutations was close to the theoretical prediction with an average of 5 mutations (Fig. 2).

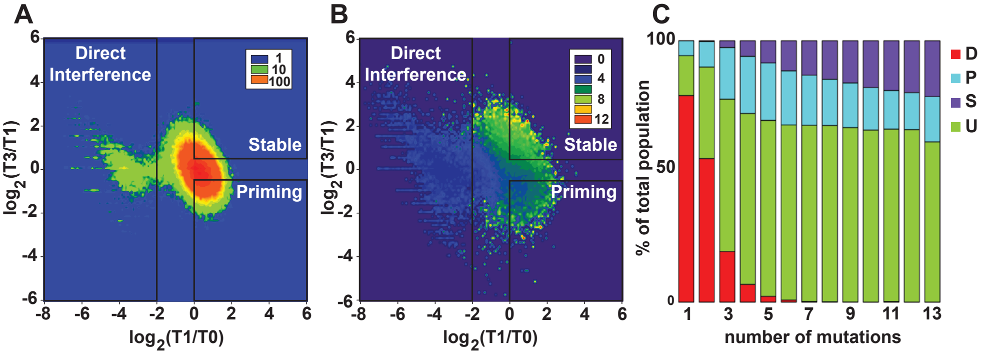


Figure 4 - Classification of functional behavior of individual PAM-protospacer variants in the pooled loss experiment. (*A*) Contour map of the number of variants. The depletion ratio of the number of reads at T1 and T0 (revealing direct interference) is plotted on a double log$_2$ scale vs the ratio of reads at T3 and T1 (revealing priming). Variants were binned in 0.1 by 0.1 bins and the number of sequences per bin is shown in a color range of blue to red (1-100 variants). Black boxes indicate the boundaries of the three different functional categories (*Direct interference* (X<-2, n=8,792), *Priming* (X>0, Y<-0.5, n=26,842), and *Stable* (X>0, Y>0.5, n=12,066)) and a group of *Unclassified* (n=86,395) variants. (*B*) Contour map as in (A) showing the average number of mutations of each bin in a color range of purple to red (0-12 mutations). (*C*) Percentage distribution of the functional categories at increasing numbers of mutations (D, *Direction interference*; P, *Priming*; S, *Stable* and U, *Unclassified*).

The protospacer-PAM plasmids were analyzed based on changes in abundance over the total time course of the experiment using plots depicting the ratio of reads at T3/T1 versus T1/T0 (Fig. 4*A*). This allowed us to classify the behavior of the sequences into three functional categories termed *Direct interference*, *Priming* and *Stable*. The *Direct interference* group was defined as protospacers that decreased in abundance between the initial library and following transformation into PIM5 (T0 to T1); the *Priming* group was defined as protospacers that did not decrease in abundance after transformation into PIM5, and only decreased in abundance following 48 h of culturing (T1 to T3) and the *Stable* group showed no decrease in abundance after transformation and prolonged culturing (T0 to T1, and T1 to T3). A number of other protospacers were not classified (*Unclassified*) due to the stringent criteria we applied to define the groups (Fig. 4 legend). When the average number of mutations was plotted of each local region of the graph, a clear link became apparent between the functional categories and the number of mutations (Fig. 4*B*). As expected, the number of mutations increased going from *Direct interference* to *Priming* to *Stable*. Although this was the general trend, some clusters were also evident of priming variants with high numbers of mutations. Plotting the percentage of each of the different groups at increasing numbers of mutations revealed that *Direct interference* drops rapidly with more mutations (Fig. 4*C*). Priming is the most dominant behavior at five mutations and still occurs at 13 mutations (Fig. 4*C*). The *Stable* group steadily increases in abundance (Fig. 4*C*). Each category is addressed in detail in the following sections.

Direct interference

Previously in *E. coli*, cases were shown of protospacers carrying 4 or 5 mutations within the protospacer that would still lead to interference (Semenova et al., 2011). In the same study, Semenova and colleagues demonstrated, using single point mutations, that perfect pairing in the seed (position 1-5, 7-8) was essential for interference. The importance of the PAM sequence in interference has also been demonstrated for Type I-E (Semenova et al., 2011; Sinkunas et al., 2013; Westra et al., 2013). However, a detailed understanding of the “rules of interference” are not available for any Type I system. Our high-throughput approach allowed us to directly assess which protospacer and PAM mutations abolish interference. Firstly, we noted that the distribution of the number of mutations in the *Direct interference* group (n=8,792 variants) was maximal at 3 mutations per protospacer (Fig. 5*A*). This means that of the initial library, plasmids with fewer mutations were likely to be targets of direct interference. Secondly, by plotting the percentage of mutations at each PAM or protospacer position we observe which positions tolerate mutation (high abundance after introduction into PIM5) and those that do not (low abundance after introduction into PIM5) (Fig. 5*B*). As expected, mutations within the PAM (-3 to -1) and within the seed (1-5, 7-8) do not readily tolerate mutation and therefore inhibit direct interference. In addition, our analysis reveals that seed positions 5 and 8 tolerate mutations better than position 1, 2, 3, 4 and 7.

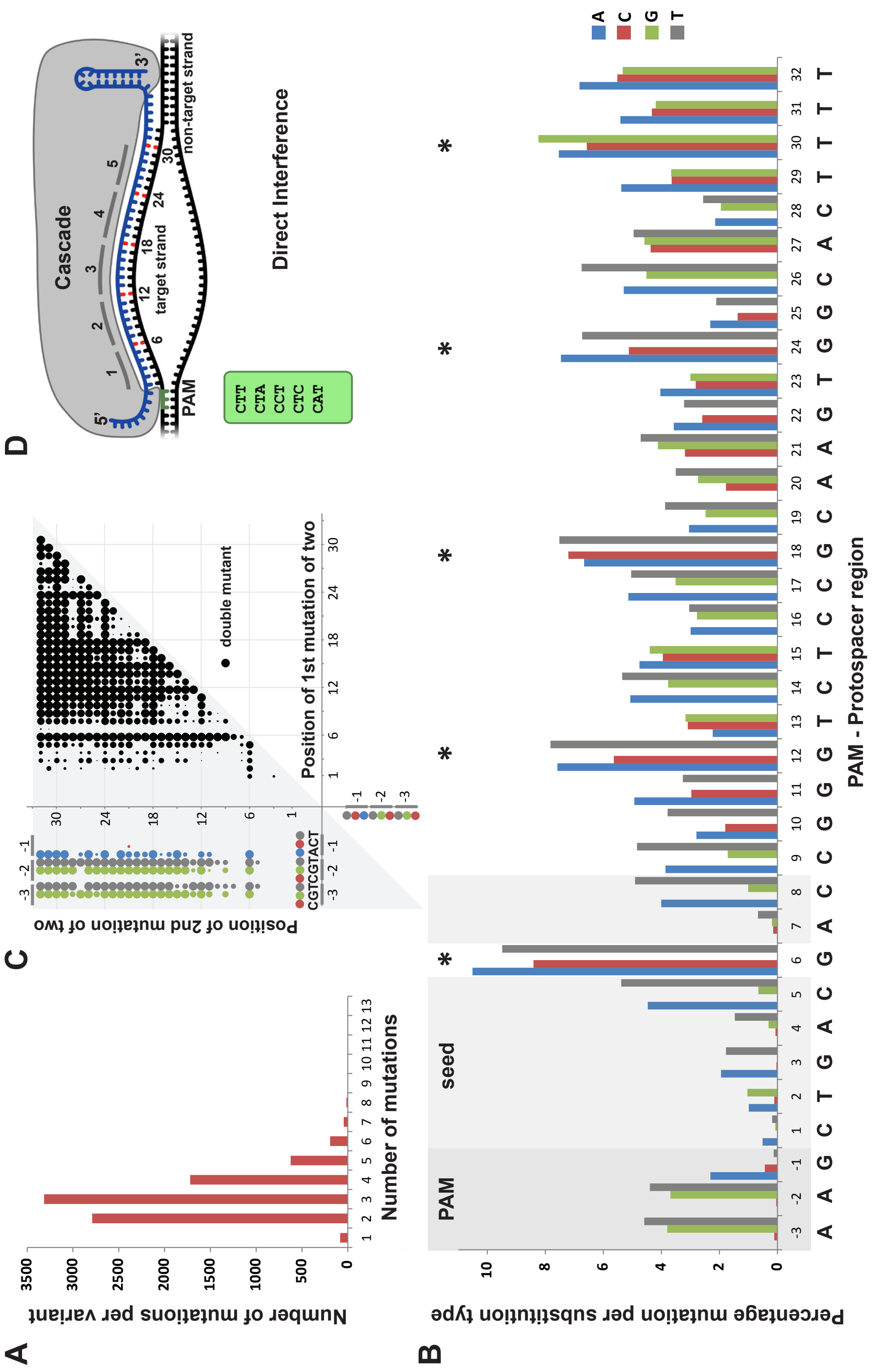

A
Number of mutations per variant
0
500
1000
1500
2000
2500
3000
3500
1 2 3 4 5 6 7 8 9 10 11 12 13
Number of mutations
B
Percentage mutation per substitution type
0
2
4
6
8
10
PAM
seed
*
A
C
G
T
-3 -2 -1 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
A A G C T G A C G A C C G G G T C T C C G C A A G T G G C A C T T T T
PAM - Protospacer region
C
Position of 2nd mutation of two
Position of 1st mutation of two
double mutant
1
6
12
18
24
30
-1
-2
-3
CGTCGTACT
D
Cascade
1
2
3
4
5
5'
3'
6
12
18
24
30
target strand
non-target strand
PAM
CTT
CTA
CCT
CTC
CAT
Direct Interference

Figure 5 - Analysis of variants displaying direct interference. (*A*) Distribution of the number of mutations per variant. (*B*) Percentage mutation per substitution type per position. For example, ~10% of all G to A mutations at position 6 end up in the *Direct interference* class. (*C*) Analysis of the mutation position of all double mutants (n=2,791 sequences) found in the *Direct interference* class. Protospacer mutations were scored regardless of the identity of the nucleotide, whereas PAM nucleotides were scored taking the nucleotide identity into account. Bubble size is indicative of the fraction of variants with mutations at a certain combination of positions displaying *Direct interference* behavior. The absence of a bubble indicates that a certain combination of mutations leads to escape from direct interference. (*D*) Schematic representation of five PAM sequences on the target strand from 5' to 3' supporting *Direct interference* (see also Table S7). The five pinch point positions of Cas7 where mutations are readily tolerated for direct interference (red base pairs numbered 6, 12, 18, 24 and 30), as well as the five helical segments within the Cascade R-loop (grey lines, numbered 1-5) are indicated.

To interrogate the data in an alternative manner, we plotted the percentage of different nucleotide variants with two mutations at different positions in the entire PAM and protospacer region (Fig. 5*C*). This clearly high lights the significance of the PAM and seed, and shows seed discontinuity at position 6. The greater importance of seed positions 1, 2 and 7 becomes evident as well as they rarely co-occur in the *Direct interference* group together with mutations elsewhere in the protospacer. Particular double mutant variants are synergistic, and escape direct interference (i.e. they have decreased black circles in Fig 5*C*) This synergy for evasion of direct interference is observed between position 28 with 19-23, 25-27, 31 and 32 (Fig. 5*C* and 5*D* segments 4 and 5; see below). Likewise, a similar synergy is detected with position 25 with 19-23, 26-27, 29 and 31 (Fig. 5*C* and 5*D* segments 4 and 5). Other clusters of low abundance double mutants occur, in particular at positions 19-23 and also in positions 9 and 10 with positions 13-16 (Fig. 5*C* and 5*D* segments 2 and 3). It is probable that two mismatches in close proximity in certain locations are more likely to disrupt the required crRNA:DNA base pairing.

Most striking from the data in Figure 5*B* is the six nt periodicity in which mutations are allowed (i.e. position 6, 12, 18, 24 and 30). Variants with up to seven mutations may still display direct interference, while carrying typically half of their mutations at these positions. The positions where mutations are tolerated accurately map five pinch points of the backbone subunit Cas7 in the cryo-TEM structure of Cascade (Wiedenheft et al., 2011a), where the crRNA:target DNA heteroduplex is not base paired (Fig. 5*D*). These interruptions are key in avoiding topological complications associated with base pairing Cascade-bound crRNA to a double stranded target DNA, and create five segments of five nt and a final two nt region in which the crRNA can base pair to the target DNA (Fig. 5*D*).

To determine the influence of the PAM on direct interference, we focused on sequences which only contained mutations in the PAM. Out of the 64 possible PAMs (with no mutations in the protospacer), 40 were present in our dataset. Of these, only 5 PAMs allowed direct interference, indicating that the PAM requirement for this particular behavior is rather strict (Fig. 5*D*). In summary, analysis of the transformation efficiency of a large protospacer and PAM library has reinforced the importance of the PAM and seed regions in direct interference and revealed with high-accuracy the 5 pinch point mutations permitted with a periodicity of 6 nt.

Priming

In the *Priming* group (n=26,842 variants), the distribution of the number of mutations peaks at 5 mutations per PAM-protospacer region (Fig. 6*A*), which is more than for the *Direct interference* group, and is consistent with the individual priming data (Fig. 3). Based on the published data (Datsenko et al., 2012), we predicted that protospacers that enable priming might have a preference for mutations within the PAM or seed that allow escape from direct interference. Indeed, analysis of the position of mutations of *Priming* protospacers revealed a subtle increase in mutations within the PAM and seed that promoted priming (Fig. 6*B*). This was very pronounced when protospacers with three mutations were examined (Fig. 6*C*). The analysis revealed that mutants with two mutations in the PAM-seed region (position -2 to 5, 7, 8) displayed significant *Priming* behavior regardless of the position of a third mutation (Fig. 6*C*). Also a mutation at position 9 which is adjacent to the seed, shows enhanced priming (Fig. 6*C*). Strikingly, almost 35% of all mutations at position 28 ends up in the *Priming* category (Fig. 6*B*). Even in combination with two other mutations, a mutation at position 28 significantly leads to priming (Fig. 6*C*). Further inspection of the plot showed that position 16 also contributes significantly to priming in a number of triple mutants. Apart from a number of positions that stimulate priming, a decreased abundance of mutations in positions 11, 12, 22, 24 and 30 were associated with *Priming* (Fig. 6*B* and 6*C*, also see below).

With regard to the influence of PAM mutations, the number of PAMs that allow *Priming* is much greater compared with *Direct interference* (Fig. 6*D*). We observed that 22 out of the 40 PAM mutants in the dataset resulted in priming, suggesting that the majority of PAMs promote either *Direct interference* of an invader (5 out 40), or indirect interference by *Priming* (22 out of 40). Especially a CGT PAM, which is obtained by point mutation of the -2 position of the PAM, appeared to have a dominant priming effect on the behavior of many triple mutants (Fig. 6*C* and 6*D*). Overall, priming is a robust host response that can deal with a myriad and large number of mutations in protospacer and PAM regions.

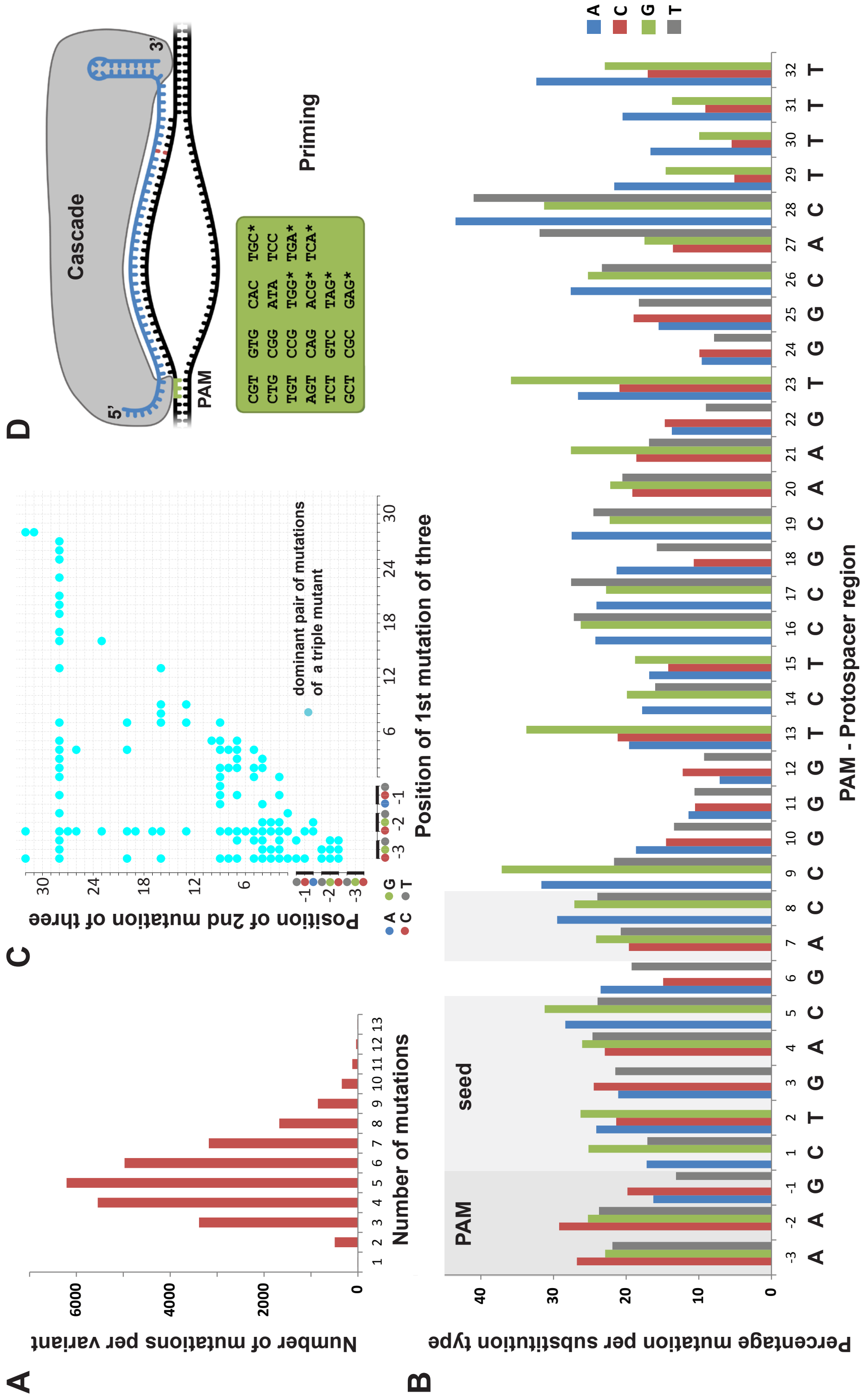

A
Number of mutations per variant
6000
4000
2000
0
1 2 3 4 5 6 7 8 9 10 11 12 13
Number of mutations
C
Position of 2nd mutation of three
30
24
18
12
6
-1
-2
-3
Position of 1st mutation of three
-3 -2 -1
6
12
18
24
30
dominant pair of mutations of a triple mutant
A
C
G
T
D
Cascade
5'
3'
PAM
Priming
CGT GTG CAC TGC*
CTG CGG ATA TCC
TGT CCG TGG* TGA*
AGT CAG ACG* TCA*
TCT GTC TAG*
GCT CGC GAG*
B
Percentage mutation per substitution type
40
30
20
10
0
PAM
seed
-3 -2 -1 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
A A G C T G A C G A C C C G G G T C T C C G C A A G T G G C A C T T T T
PAM - Protospacer region
A
C
G
T

Figure 6 - Analysis of variants displaying priming. (*A*) Distribution of the number of mutations per variant. (*B*) Percentage mutation per substitution type per position. For example, ~40% of all C to A mutations at position 28 end up in the *Priming* class. (*C*) Analysis of pairs of mutations in triple mutants that significantly ($p < 0.05$) contribute to priming behavior (n=17,109 triple mutants sequences). Protospacer mutations were scored regardless of the identity of the nucleotide, whereas PAM nucleotides were scored taking the nucleotide identity into account. The absence of a point indicates that a certain combination of mutations does not significantly lead to priming. (*D*) Schematic representation of the Cascade R-loop indicating PAM sequences on the target strand from 5' to 3' supporting priming (see also Table S7). PAM sequences indicated with an asterisk (*) were computationally inferred from the analysis of the behavior of the sequences containing the given PAM and either one or two additional mutations (see Table S7). Protospacer position 28, which is highly associated with priming, is shown in red.

Stable

The *Stable* group had the highest number of mutations of all the groups, with on average 6 substitutions per PAM-protospacer region (Fig. 7*A*). This was expected, since this group avoids both *Direct interference* and subsequent *Priming*. Mutations at position 10, 11, 12, 18, 22, 24 and 25 are all overrepresented in the *Stable* group (Fig. 7*B*), and surprisingly, these positions are all guanosines. This suggests that rG-dC (i.e. riboG-deoxyriboC) base pairing between the crRNA and the target DNA should not be disrupted for priming to occur, which is consistent with the fact that rG-dC basepairs are the strongest basepair known (Hall and McLaughlin, 1991; Roberts and Crothers, 1992). The importance of rG is again reflected in the analysis of triple mutants (Fig. 7*C*) where mutations of guanosines at position 11, 24 and 25 mutations are frequently found in *Stable* triple mutants. A mutation from T to C is strongly overrepresented in the *Stable* group at position 29, 30 and 31, and to some degree also position 15 and 32 (Fig. 7*B*). These mutations all result in rU-dG wobble basepairs between the crRNA and target DNA (Sugimoto et al., 2000), which appear to disrupt the priming process and result in stable plasmids. Mutations in the PAM alone are not sufficient to promote *Stable* behavior. However in triple mutants, some PAMs significantly lead to *Stable* variants (i.e. CCC, ATC, ACT, AAT), while ATT PAMs yield a dominant stable behavior in combination with mutations at eight different positions of the protospacer (i.e. 10, 11, 14, 24, 25, 29, 30 and 31) (Fig. 7*C* and *D*, also see below).

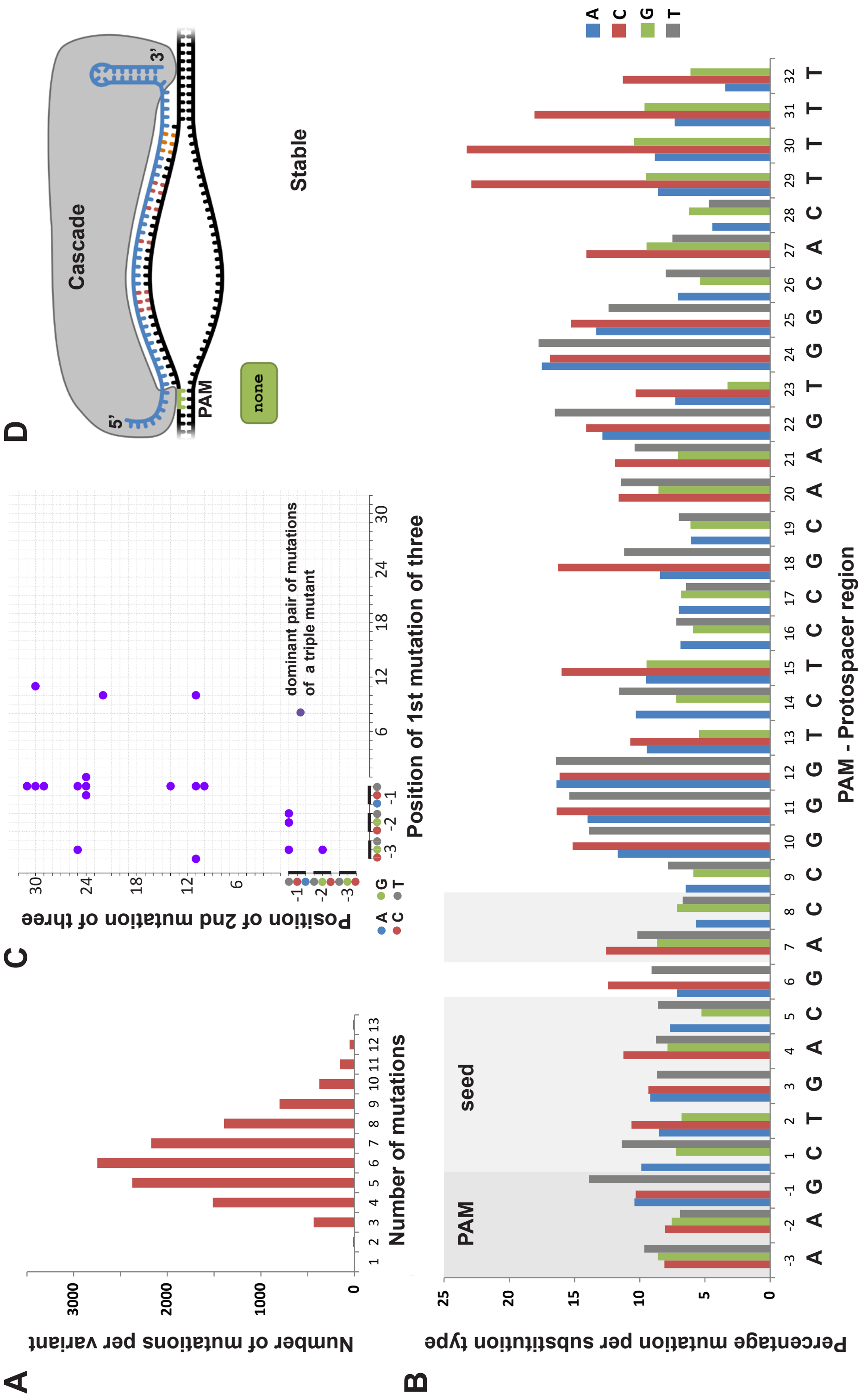

A
Number of mutations per variant
0
1000
2000
3000
1 2 3 4 5 6 7 8 9 10 11 12 13
Number of mutations
B
Percentage mutation per substitution type
0
5
10
15
20
25
PAM
seed
-3 -2 -1 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
A A G C T G A C G A C C G G G T C T C C G C A A G T G G C A C T T T T
PAM - Protospacer region
A
C
G
T
C
Position of 2nd mutation of three
-3
-2
-1
6
12
18
24
30
Position of 1st mutation of three
dominant pair of mutations of a triple mutant
A
C
G
T
D
Cascade
5'
3'
PAM
none
Stable

Figure 7 - Analysis of stable variants. (*A*) Distribution of the number of mutations per variant. (*B*) Percentage mutation per substitution type per position. For example, ~20% of all T to C mutations at positions 29, 30 and 31 end up in the *Stable* class. (*C*) Analysis of pairs of mutations in triple mutants that significantly ($p < 0.05$) contribute to stable behavior (n=17,109 triple mutants sequences). Protospacer mutations were scored regardless of the identity of the nucleotide, whereas PAM nucleotides were scored taking the nucleotide identity into account. The absence of a point indicates that a certain combination of mutations does not significantly lead to priming. (*D*) Schematic representation of the Cascade R-loop indicating that no PAM sequences are sufficient to cause stable behavior. Positions of variants overrepresented in the *Stable* class are shown in red (10-12, 18, 22, 24, 25) and orange (29-31).

Nucleotide dependent effects

To investigate nucleotide specific effects in more detail, we analyzed the behavior of variants containing increasing numbers of specific mismatches in the protospacer (Fig. 8). Mutations in the protospacer DNA resulting in all types of mismatches with specific RNA nucleotides (rA, rC, rG, or rU) reduce direct interference (Fig. 8*A*). However, priming is very differently affected by each type of mismatched ribonucleotide. While mismatched rG nucleotides are detrimental to priming and lead to more *Stable* behavior as was observed for mutations at position 10, 11, 12, 18, 22, 24 and 25 (Fig 7*B* and 8*A*), mismatched rC nucleotides on the contrary strongly promote priming (Fig. 8*A* and 7*B*). In addition, mismatched rA and rU nucleotides do not strongly effect the behavior.

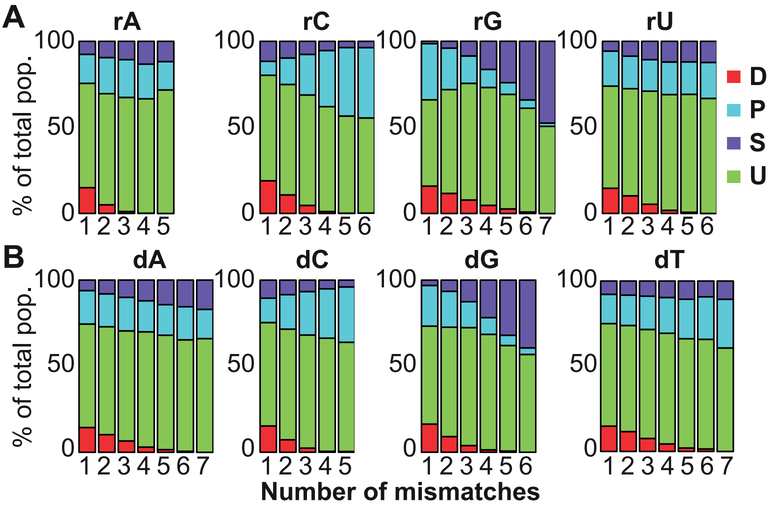


Figure 8 - Effect of mismatches between the crRNA spacer sequence and the targeted strand of the protospacer. Variants with mutations in the PAM were excluded in this analysis (remaining group n=83,655). (*A*) Mismatched ribonucleotides in the crRNA spacer. (*B*) Mismatched deoxyribonucleotides in the targeted strand of the protospacer DNA.

We next repeated the analysis from a target point of view by looking at mismatched DNA nucleotides (dA, dC, dG and dT) in the targeted strand of the protospacer (Fig. 8*B*). Although this analysis is very different from the one described above (i.e. analysis of mutation to a particular nucleotide (Fig. 8*B*; e.g. dA, dG, or dT to dC) compared with mutations of specific nucleotides to any nucleotide (Fig. 8*A*; e.g. dG to dA, dC, or dT)), the results were strikingly similar. Again, the increased numbers of mismatched dC and to some extent also dT nucleotides promoted priming, whereas increased numbers of dG abolished priming. Mismatched dA nucleotides in the targeted strand of the protospacer appeared to have no strong effect on priming behavior. Analysis of 10 priming variants with 12 mutations in the protospacer indeed revealed that on average each these highly mutated variants carry mostly mutations promoting priming.

The opposing effects of rC and rG mismatches argue against a role for the thermostability of the protospacer DNA duplex in affecting priming, as the decrease in stability of double stranded DNA on average is similar when introducing mutations causing an rC mismatch (dG to dA, dC, or dT) or rG mismatch (dC to dA, dG, dT). The same is true for dC mismatches (dA, dG, or dT to dC), and dG mismatches (dA, dC, or dT to dG) both of which on average increase the duplex stability. Instead, the reason for the unequal effects of particular mutations might reside in differences in stabilities or conformation between individual mismatched RNA:DNA nucleotide pairs (Sugimoto et al., 2000) in the context of the Cascade R-loop.

All in all the rules for priming appear to be a complex combination of the number of mutations (Fig. 6*A* and 7*A*), position dependent effects, such as in the PAM and seed (Fig. 6*C* and 7*C*), and nucleotide dependent effects (Fig. 8). As a rule of thumb, however, mismatched rG or dG nucleotides lower the chances of priming, while mismatched rC and dC promote priming.

Discussion

A perceived Achilles heel of the CRISPR-Cas adaptive immune systems, which was detected in early studies (Deveau et al., 2008), is the ability of phages and plasmids to escape immunity through mutation of their PAM or protospacer (Sapranauskas et al., 2011; Sun et al., 2013). Recently, it was shown that despite these mutations allowing avoidance of direct interference, they led to an enhanced positive feedback process of new spacer acquisition, termed priming (Datsenko et al., 2012; Swarts et al., 2012). This suggests a further evolved function of the CRISPR-Cas systems in the arms race between bacteria and their invading mobile genetic elements. In the current study, we have investigated the requirements of both direct interference and priming by the Type I-E system of *E. coli*. We have revealed that direct interference readily tolerates mutations at specific positions in the protospacer (6, 12, 18, 24, 30), and can cope with 2 or 3 more mutations in the non-seed region of the protospacer. Protospacer sequences that are not directly targeted, including sequences with mutations in the PAM and seed, enhance the acquisition of new spacers even when the existing spacer has many mismatches. Therefore, priming is an incredibly flexible and promiscuous process that may provide a major, if not the main, route to adaptation in CRISPR-Cas systems.

We observed that up to 5 mutations in the PAM and seed (of 11 total mutations) still facilitated primed spacer acquisition, in contrast to Datsenko et al. (Datsenko et al., 2012), who reported that a double mutant of the PAM and seed abolished the process. The difference between these findings may be due to the use of either phages or plasmids, and/or the time to allow priming to occur (0-8 h vs 24-48 h). Our finding that high numbers of mismatches in the PAM and protospacer region stimulate primed spacer acquisition, challenges and expands the concept of adaptation in CRISPR-Cas systems. Further, we show that the number, position and kind of mutation greatly influence the priming process. For example, mismatched cytosine ribonucleotides in the crRNA promote priming, whereas mismatched guanine ribonucleotides abolish priming. It would therefore seem advantageous for a host to select and retain C-rich spacers in their CRISPR arrays to be capable of a better primed response, while discarding outdated G-rich spacers.

During the preparation of this manuscript, priming was demonstrated in the Type I-B system of *Haloarcula hispanica* against an archaeal virus (Li et al., 2014). Interestingly, the pre-existing spacer responsible for priming contained incomplete complementarity (Li et al., 2014). Analysis of a number of other reports suggests that priming might occur in other CRISPR-Cas systems. Firstly, the most active CRISPR-Cas system for adaptation is the Type II in *S. thermophilus* (Barrangou et al., 2007; Deveau et al., 2008; Garneau et al., 2010; Horvath et al., 2008a; Magadan et al., 2012; Paez-Espino et al., 2013). Previously, upon testing the CRISPRTarget protospacer finding tool, we identified partial matches between existing spacers and phage used in challenge experiments and proposed that this might have resulted in priming if such a process would exist in *S. thermophilus* (Biswas et al., 2013). In another report, *gfp* mRNA and plasmid DNA were reduced in *E. coli* with a Type I-E system by a spacer that only had an ~10 bp stretch of complementarity (Perez-Rodriguez et al., 2011). Although this degree of complementarity falls well outside the limits we observed, all Cas proteins, including Cas1 and Cas2, and the CRISPR locus were required for *gfp* loss, which might indicate that priming was involved. More recently, *Sulfolobus solfataricus* was shown to have remarkable flexibility for protospacer recognition. Up to 15 mismatches still enabled interference, albeit at roughly 50% efficiency (Manica et al., 2013). Again, the interesting possibility arises that these divergent protospacers might be able to trigger priming and subsequent interference. Whether these, and other, studies are attributable to priming will require further research. It is clear however, that when looking for targets of crRNA, those with partial matches, or lacking PAMs, can no longer be discarded as non-functional. Tools such as CRISPRTarget enable easy detection of these degenerate crRNA targets and their flanking sequences (Biswas et al., 2013).

It was proposed that priming enables a rapid response to genetic elements that have acquired a point mutation to escape CRISPR-Cas direct interference (Datsenko et al., 2012). Our data supports this concept, since mutations that led to escape from direct interference promoted priming. The ability of highly-variant PAM-protospacer regions to be recognized as foreign, suggests that even diverse sequences can elicit immunity. This might enable CRISPR-Cas to remember invasions that occurred more distantly in evolutionary time and still mount a response. Likewise, this promiscuous immunity may provide a broad spectrum resistance to a range of mobile genetic elements possessing that protospacer (i.e. a family of related phages or plasmids). By virtue of this loose selectivity, CRISPR-Cas can apparently detect unrelated elements that share only weak sequence identity. Therefore, it is conceivable that priming might be the favored route to resistance, which is in agreement with the robustness of this process relative to the apparent intractable nature of naïve acquisition (Fineran and Charpentier, 2012; Li et al., 2014). In theory, longer CRISPR arrays that contain a greater number of spacers would have an improved chance of expressing a crRNA that facilitates priming. Therefore, the evolutionary selection acting on spacers, in particular, the older ones, might function at two levels; immediate protection and primed protection. This might provide a rationale for the maintenance of spacers in long CRISPRs, for which the immediate invader is no longer a threat, and fits with the low turnover of spacers in some CRISPR-Cas systems (Diez-Villasenor et al., 2010; Touchon et al., 2011; Touchon and Rocha, 2010).

Various models have been proposed to explain priming. In the *E. coli* Type I-E system, all *cas* genes and crRNA are required and new spacers are integrated in a strand-specific manner (Datsenko et al., 2012; Swarts et al., 2012). A sliding model was proposed in which Cascade-crRNA weakly binds protospacers with mismatches, then slides along the DNA until a PAM is reached. Spacers are then acquired from the strand of the original priming protospacer in a process requiring Cas1, Cas2 and Cas3 (Datsenko et al., 2012). The sliding model requires a bias of new spacers acquired from nearby the priming protospacer; however, new spacers do not reveal this distribution (Savitskaya et al., 2013). Since short sequence motifs other than PAMs can influence spacer acquisition efficiency (Yosef et al., 2013), the interpretation of spacer distribution becomes more challenging. Indeed, we observed that four different spacers accounted for a quarter of all newly acquired spacers and contained the recently identified AA nt motif (Yosef et al., 2013) at the 3’ end of the spacer (Table S4). This suggests that this AA nt motif influences acquisition efficiency during both primed and naïve adaptation. As an alternative model to ‘sliding’, Cascade-crRNA was proposed to cause the exposure of extended regions of ssDNA, potentially mediated by Cas3, and that Cas1 and Cas2 might have a preference for single-stranded DNA substrates (Savitskaya et al., 2013). This is similar to the original priming model, where we proposed that single-stranded substrates, possibly preferred by Cas1 and Cas2, are generated due to Cascade-crRNA and Cas3-dependent targeting or weak targeting, resulting in spacer acquisition (Swarts et al., 2012). Which of these models proves to be correct remains subject of future studies.

To date, the PAM has been well characterized in a number of Type I and Type II systems and the effect of mutations in the protospacer has been documented (Deveau et al., 2008; Künne et al., 2014; Mojica et al., 2009b; Shah et al., 2013; Sorek et al., 2013a). However, only few high-throughput random-mutagenesis studies of the effects of PAM and protospacer mutations have been reported. Recently, studies of the Type II/Cas9 system have used high-throughput approaches to investigate PAM and protospacer requirements for direct interference. Jiang et al. generated a variant 5 nt PAM library in *Streptococcus pneumoniae* which revealed that NGG PAMs support interference (Jiang et al., 2013). Investigation of the effects of protospacer mutations on direct interference revealed a seed region of nt 1 to 12 (Jiang et al., 2013), in line with previous studies (Jinek et al., 2012). In another study, different PAM specificities of three different Cas9s were demonstrated (Esvelt et al., 2013). Our approach allowed an in depth assessment of the Type I-E sequence requirements for both direct interference and priming. We revealed five PAMs for direct interference and 22 PAMs for priming. These direct interference PAMs included a new CTA PAM and the remaining ones were consistent with the four previously observed PAMs for interference (Westra et al., 2013; Westra et al., 2012c). There were no PAM mutants which resulted in the stable phenotype, indicating that mutating the PAM alone does not allow complete escape from Type I-E CRISPR-Cas systems. The results demonstrated the critical role of the PAM and the seed sequence, in agreement with previous work (Semenova et al., 2011; Westra et al., 2013), and revealed that base pairing at position 1, 2 and 7 within the seed is of greater importance than base pairing at other seed positions. Semenova et al. showed cases where interference tolerated up to 4 or 5 mutations outside the seed or PAM and that additional single non PAM/seed mutations could lead to escape (Semenova et al., 2011). Similarly, in the Type I-F system, in the presence of four mismatches, further mutations inside or outside of the seed or PAM were shown to disrupt interference (Cady et al., 2012). The Cryo-EM structure of Cascade revealed that the pinch points of the backbone subunit Cas7 disrupted 1-2 nt of base pairing that were separated by 4-5 nt helical segments (Wiedenheft et al., 2011a). Here, the pinch points of Cas7 were mapped with nucleotide precision, resulting in five helical segments of 5 nt. The importance of position 1, 2 and 7 fits with the hypothesis that crRNA:target DNA base pairing starts from the PAM end of the protospacer (Sashital et al., 2012; Semenova et al., 2011) as position 1 is the first nucleotide of the first segment to be involved in base pairing with the target DNA (Westra et al., 2013), and 7 is the first nucleotide of the second segment. Remarkably, the distance of 6 nucleotides between pinch points is identical to the interval at which Type III-B CRISPR-Cas systems cleave their target (Hale et al., 2009; Staals et al., 2013), suggesting that the backbone subunits of Type I and III complexes bind their crRNA and target nucleic acid with the same periodicity.

An apparent paradox also emerged from the data. Apart from being overrepresented in the *Direct interference* dataset, mutations at some pinch point positions (i.e. 12, 18, and 24) were also overrepresented in the *Stable* group. This would suggest that although these positions are not engaged in base pairing with the target during direct interference, they do base pair with the target during priming. This finding raises the possibility that there is a conformational change in Cascade:crRNA during priming that alters the position of the crRNA, freeing these positions for base pairing thereby allowing the detection of weak targets.

We are only beginning to understand the intricate interaction between the CRISPR-Cas systems and their invaders. This study highlights the incredible flexibility of these systems to respond to a rapidly evolving, or diverse mobile genetic elements and generate new resistance. We propose that promiscuous priming is an extremely important feature of the CRISPR-Cas mechanism and is not restricted to the Type I-E system, but is a feature of other CRISPR-Cas Types.

Materials and Methods

Bacterial strains and growth conditions. The *E. coli* K12 W3110 derivative, Δ*hns* was generated previously by removal of the kanamycin resistance cassette from JW1225 (Westra et al., 2010). *E. coli* strains were grown at 37°C in Luria Broth (LB; 5 g L^{-1} NaCl, 5 g L^{-1} yeast extract and 10 g L^{-1} tryptone) at 180 rpm or on LB-agar plates containing 1.5% (w v^{-1}) agar. When required, medium was supplemented with the following: ampicillin (Ap; 100 µg ml^{-1}), chloramphenicol (Cm; 25 µg ml^{-1}) or kanamycin (Km; 50 µg ml^{-1}). Bacterial growth was measured at 600 nm (OD_{600}).

Molecular biology and DNA sequencing. Oligonucleotides used in this study are listed in Table S5, Table S6 contains read counts, barcodes and scaling factors of the large scale sequencing experiment and the PAM sequence analysis is shown in Table S7. All strains and plasmids were confirmed by PCR and sequencing (GATC-Biotech, Konstanz, Germany). Plasmids were prepared using GeneJET Plasmid Miniprep Kits (Thermo Scientific) and DNA from PCR and agarose gels was purified using the Thermo Scientific GeneJET PCR Purification and Gel Extraction Kits.

Replacement of the PIM2 CRISPR2.1 locus. Previously, a synthetic recombination cassette was generated that corresponded to 400 bp flanking regions on each side of the CRISPR 2.1 locus separated by a kanamycin resistance gene flanked by FRT sites

(Westra et al., 2010). This construct includes a synthetic CRISPR sequence with the leader, eight repeats and seven spacers, the first of which (J3) targets bacteriophage lambda (Westra et al., 2010). These spacers were compared with pGFPuv using CRISPRTarget (Biswas et al., 2013) with the cut off score lowered to 0 and only spacer 3 'matched' (score 7, 14/35 mismatches and 7 bp longest continuous match). Therefore, this CRISPR 2.1 replacement cassette should not promote priming. Recombineering was performed using a protocol described elsewhere (Datsenko and Wanner, 2000), with minor modifications. *E. coli* PIM2 was transformed with the recombineering functions on plasmid pKD46 and grown at 30°C (plasmid pKD46 is unstable at temperatures > 37°C). The CRISPR 2.1 J3 replacement sequence was amplified by PCR using primers BG3113 and BG3114 with *Pfu* polymerase (Thermo Scientific). The PCR product was purified, digested with DpnI to remove any remaining plasmid template DNA, re-purified and transformed by electroporation into *E. coli* PIM2 containing pKD46. Note that electrocompetent PIM2 pKD46 were prepared with 0.2% (w v^{-1}) L-arabinose for expression of the lambda red proteins. Transformants were recovered for 2.5 h in LB at 30°C, plated out onto LBA with kanamycin (50 µg ml^{-1}) and incubated at 30°C. Plasmid pKD46 was cured by growth at 37°C and recombination was validated by PCR and sequencing.

Plasmid loss experiments. Plasmids were introduced into *E. coli* by either electroporation or via heat shock and were pGFPuv (Clontech; pUC origin, ~500/cell, Ap^R, *gfp*, 3337 bp), pACYC184 ((Rose, 1988), p15A origin, 20-30/cell, Cm^R, Tc^R, 4245 bp) and pACYCDuet^{TM}-1 (Novagen; p15A origin, 20-30/cell, Cm^R, *lacI*, 4008 bp). *E. coli* with plasmids of interest were grown for 24 h in 10 ml LB in 50 ml tubes (Greiner) at 37°C with shaking at 180 rpm. For further passaging, 100 µl of culture was subcultured into 10 ml LB in 50 ml tubes for a further 24 h at 37°C at 180 rpm. When indicated, further periods of incubation were performed using the same conditions. Dilutions were plated on LBA and pGFPuv loss detected under UV and GFP +/- colonies identified. For plasmids pACYC184 and pACYCDuet^{TM}-1, individual colonies were patched onto LBA with or without the appropriate antibiotics to identify colonies that had lost the plasmid.

Colony PCR and spacer sequencing. Plasmid-free colonies were screened for spacer integration by colony PCR using DreamTaq Green DNA polymerase (Fermentas) (Swarts et al., 2012). Briefly, acquisition of spacers in CRISPR2.1 was detected by PCR using primers BG3474, which binds in the leader and primer BG3475, which anneals in spacer 4. New spacers in CRISPR2.3 were detected using BG3414, which anneals in the leader and BG3415, which binds in spacer 3. PCR products were visualized on 2% agarose gels and stained with SYBR-safe (Invitrogen). CRISPR2.1 and 2.3 were sequenced with BG3474 and BG3414, respectively and analyzed as follows. Firstly, the sequence was uploaded into CRISPRFinder (Grissa et al., 2007) and spacers and repeats were manually extracted. Spacer lists were generated in

FASTA format and then aligned with the target plasmids by using Geneious (v6.0.5) (Kearse et al., 2012) and CRISPRTarget (Biswas et al., 2013).

Generation of control plasmids for priming experiments. Plasmids were generated that either lacked a protospacer target (pGFPKm; negative control) or contained a single protospacer with a +1 seed mutation (pGFPKm-PS8; positive priming control). The Km^R gene was amplified by PCR from pRSF-1b with BG4225 as the forward primer and either BG4226 (no protospacer) or BG4227 (protospacer 8 and PAM with +1 seed mutation) as reverse primers. The products were digested with EcoRI and NcoI and ligated to pGFPuv, previously cut with EcoRI and PagI (BspHI), which removed the *bla* gene. This strategy enabled positive selection with Km for plasmids containing the inserts into the 1.8 kb EcoRI/PagI backbone of pGFPuv.

Generation of a library of protospacer and PAM variants. To generate a pool of plasmids with variant protospacer and PAM sequences, Km^R gene was amplified by PCR from pRSF-1b with BG4225 as the forward primer and BG4228 (protospacer 8 and PAM with 85% WT nt and 5% chance of each alternative nt at each position) as the reverse primer. This ratio provided variants with an average of 5 mutations in the 35 nt PAM and protospacer sequence. Products were digested with EcoRI and NcoI and ligated to pGFPuv, previously cut with EcoRI and PagI. Five libraries were produced from independent PCRs and used to transform chemically-competent *E. coli* NEB5a. A total of ~210,000 transformants were isolated after plating on LBA containing Km. For each library, colonies were pooled into LB by scraping the bacteria from the plates and plasmids were extracted.

Individual high-throughput priming assays of protospacer and PAM variants. Libraries of pGFPKm-mPS8 were transformed into *E. coli* PIM5 by electroporation and plated onto LBA with Km. Single colonies were picked into 200 μl of LB in 2 ml 96-well culture plates (Greiner), in addition to positive and negative controls, and the plates were incubated for 24 h with shaking at 750 rpm at 37°C. Five μl was subcultured into 195 μl of LB in a further 96-well plate as above and cultures were grown for 24 h. At 24 h and 48 h, samples were plated on LBA. Total colonies were counted, GFP +/- colonies were assessed under UV and spacer acquisition determined by PCR.

High-throughput plasmid loss assays and Illumina sequencing. Plasmid DNA was pooled for the 5 libraries (~210,000 protospacer 8 and PAM transformants; T0), transformed in triplicate into *E. coli* PIM5 by electroporation. Positive (PIM5 pGFPKm-PS8) and negative (pGFPKm) controls were included to enable tracking of priming and plasmid stability, respectively. Approximately 2×10^7 colonies were pooled for each transformation (~100-fold excess over the estimated library size of 2×10^5) and resuspended in 10 ml LB. The OD_{600} was measured and adjusted to an OD_{600}=4 and 100 μl was used to inoculate 10 ml LB without antibiotics, divided equally over 3 replicates of 16 wells of a 2 ml 96-well microtiter plate. Plates were incubated at 37°C with shaking at 750 rpm and every 24 h, each replicate of 16 wells was pooled and subcultured 1:40 into 200 µl fresh LB. This was performed for 48 h. After transformation (T1), 24 h (T2) and 48 h (T3) growth, plasmids were prepared from the pooled 10 ml culture. Positive (pGFPKm-PS8) and negative (pGFPKm) controls were also analysed for plasmid loss as described earlier. Plasmid loss of the total library was monitored by the agar plate method as described for the individual high-throughput assay. Spacer integration during these assays was verified by PCR as described earlier.

These DNA pools from pre-PIM5 transformation (T0, replicate A,B,C), post-PIM5 transformation (T1, replicate A, B, C), 24 h (T2, replicate A, B, C) and 48 h (T3, replicate A, B, C) subculturing were amplified by PCR using Phusion DNA polymerase (Thermo Scientific) and a primer pair selected from BG4325 to BG4356 flanking the protospacer and PAM. Each primer contained a unique 6-nucleotide barcode differing in at least two positions from another barcode, which enabled sorting of the different samples (Table S5). The 125 base pair PCR amplicons were separated by agarose gel electrophoresis, excised from gel and purified using the Zymoclean DNA recovery kit (Zymo Research) and eluted in 30 µl H_2O. The purified fragments were quantitated by Qubit fluorometic quantitation (Invitrogen) and equal quantities were pooled. Purity of the sample was analyzed on a Bioanalyzer using the High Sensitivity DNA Assay Kit (Agilent). PCR amplicons were prepared for Illumina sequencing using the TruSeq SBS Kit (v3), 2 × 100 bp (Paired-End) and sequenced on a Hi-Seq (FC-401-3001, Illumina) at the Imagif, Centre for Molecular Genetics, CNRS, France.

Analysis of plasmid loss next generation sequence data. A total of 82,221,629 read pairs were obtained. Data analysis consisted of the following steps. As the 100 bp paired end reads of the 125 bp amplicon fully overlapped in the 35 bp PAM-protospacer region, we first assembled both paired end reads into a single consensus sequence (CS) with the merger application from the EMBOSS package (Rice et al., 2000):. Only those CSs where there was complete agreement between both paired reads of the 35 bp PAM-protospacer sequences were taken for further analysis. CSs were categorized based on the barcodes introduced at either end. A total of 49,198,699 sequences were accepted with a full hit of both the 5’ and 3’ barcodes (Table S6) and without insertions or deletions in the consensus sequence. Next, the 35 bp fragments were extracted from the 49,198,699 contigs by using the PAM-protospacer flanking sequences. A text file was created specifying the nucleotide sequence of the region of interest of the CS (i.e. the PAM-protospacer) and its counts across all replicates of all time points (12 replicate in total). As the sequence reaction contained equal amounts of DNA from all samples, the sequence counts were scaled and rounded by using scaling factors based on the sample with the highest number of sequences (i.e. T2B) (Table S6). Next, a filtering step was applied to select only PAM-protospacer sequences that had at least 20 reads in any three sequenced samples. This resulted in 134,095 unique PAM-protospacer sequences which were selected for further analysis.

Computational analysis of PAM sequences and combinatorial mutants leading to stable or priming classes. Significance scores (p-values) for the frequencies of the categorizations were computed by comparing the obtained frequencies with the frequencies in a randomly selected subset of a background set. The background set was selected in each case to have the same characteristics (number of mutations) as the set under study. The randomizations were performed 1000 times to estimate the p-values. The p-values are given in Table S7.

Acknowledgements

We thank M.M. Jore for helpful discussions. This work was supported by a Rutherford Discovery Fellowship from the Royal Society of NZ to PCF, the Voeding, Levensmiddelentechnologie, Agrobiotechnologie en Gezondheid to PCF and John van der Oost, an NWO-VIDI grant to SJJB (864.11.005) and the Beijerinck premium from the Royal Netherlands Academy of Arts and Sciences to SJJB.

Supplementary tables

Table S1. pGFPuv PIMs derived from PIM25

PIM	CRISPR locus	spacer position	spacer #	repeat (5'-3')	spacer (5'-3')	spacer length (nt)	target nt	target	F/R[a]	PAM[b]
25.01	2.3	-1	S100	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TGGTCCTGCAACTTTATCCGCCTCCATCCAGT	32	2156-2187	bla	F	CTT
25.02	2.1	-1	S101	GAGTTCCCCGCGCCAGCGGGGATAAACCG	ATATAGTGCGTTCCTGTACATAACCTTCGGGC	32	0552-0583	gfp	F	CTT
25.03	2.3	-1	S102	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TTGGCCGCAGTGTTATCACTCATGGTTATGGC	32	1906-1937	bla	F	CTT
25.04	2.1	-1	S103	GAGTTCCCCGCGCCAGCGGGGATAAACCA	TCCGCTCATGAGACAATAACCCTGATAAATGC	32	1479-1510	bla pro	R	TAC
25.05	2.1	-1	S104	GAGTTCCCCGCGCCAGCGGGGATAAACCG	TGGTCCTGCAACTTTATCCGCCTCCATCCAGT	32	2156-2187	bla	F	CTT
25.05	2.3	-1	S105	GTGTTCCCCGCGCCAGCGGGGATAAACCG	CATTGAACACCATAAGAGAAAGTAGTGACAAG	32	0466-0497	gfp	F	CTT
25.06	2.3	-1	S106	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TGTTGGCCATGGAACAGGTAGTTTTCCAGTAG	32	0434-0465	gfp	F	CTT
25.11	2.3	-2	S107	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GGCCGAGCGCAGAAGTGGTCCTGCAACTTTAT	32	2171-2202	bla	F	CTT
25.11	2.3	-1	S108	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TGTTGGCCATGGAACAGGTAGTTTTCCAGTAG	32	0434-0465	gfp	F	CTT
25.12	2.3	-1	S109	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TTGGCCGCAGTGTTATCACTCATGGTTATGGC	32	1906-1937	bla	F	CTT
25.13	2.3	-1	S110	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TCATTCTGAGAATAGTGTATGCGGCGACCGAG	32	1801-1832	bla	F	CTT
25.14	2.3	-2	S111	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TTAGCTTTGATTCCATTCTTTTGTTTGTCTGC	32	0748-0779	gfp	F	CTT
25.14	2.3	-1	S112	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GACCATGTGGTCACGCTTTTCGTTGGGATCTT	32	0914-0945	gfp	F	CTT
25.15	2.3	-2	S113	GTGTTCCCCGCGCCAGCGGGGATAAACCG	CTGGGCTGTGTGCACGAACCCCCCGTTCAGCC	32	2881-2912	ori	F	CTT
25.15	2.3	-1	S114	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GACAGGGCCATCGCCAATTGGAGTATTTTGTT	32	0836-0867	gfp	F	CTT
25.16	2.3	-2	S115	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TATATATGAGTAAACTTGGTCTGACAGTTACC	32	2393-2424	bla	F	CTT
25.16	2.3	-1	S116	GTGTTCCCCGCGCCAGCGGGGATAAACCG	CCAGCCCCGACACCCGCCAACACCCGCTGACGC	33	1206-1238	Bb	F	CTT
25.17	2.3	-1	S117	GTGTTCCCCGCGCCAGCGGGGATAAACCG	ATATAGTGCGTTCCTGTACATAACCTTCGGGC	32	0552-0583	gfp	F	CTT
25.18	2.1	-1	S118	GAGTTCCCCGCGCCAGCGGGGATAAACCG	GGCCGAGCGCAGAAGTGGTCCTGCAACTTTAT	32	2171-2202	bla	F	CTT
25.18	2.3	-1	S119	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GCGGTAATACGGTTATCCACAGAATCAGGGGA	32	3232-3263	ori	F	CTT
25.19	2.3	-1	S120	GTGTTCCCCGCGCCAGCGGGGATAAACCG	ATCCTTTGATCTTTTCTACGGGGTCTGACGCT	32	2532-2563	ori	F	CTT
25.20	2.3	-1	S121	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GCGAGTTACATGATCCCCCATGTTGTGCAAAA	32	1982-2013	bla	F	CTT
25.21	2.3	-1	S122	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TTGGTAATGGTAGCGACCGGCGCTCAGTTGGA	32	1011-1042	Bb	F	CTT
25.22	2.1	-1	S123	GAGTTCCCCGCGCCAGCGGGGATAAACCG	TTGGCCGCAGTGTTATCACTCATGGTTATGGC	32	1906-1937	bla	F	CTT
25.22	2.3	-2	S124	GTGTTCCCCGCGCCAGCGGGGATAAACCT	CGCCCCGAAGAACGTTTTCCAATGATGAGCAC	32	1711-1742	bla	R	AAA
25.22	2.3	-1	S125	GTGTTCCCCGCGCCAGCGGGGATAAACCT	GGTCCTGCAACTTTATCCGCCTCCATCCAGT	31	2156-2186	bla	F	ACT
25.23	2.1	-1	S126	GAGTTCCCCGCGCCAGCGGGGATAAACCG	GCGAGTTACATGATCCCCCATGTTGTGCAAAA	32	1982-2013	bla	F	CTT
25.23	2.3	-1	S127	GTGTTCCCCGCGCCAGCGGGGATAAACCG	CTGTGACCGTCTCCGGGAGCTGCATGTGTCAG	32	1130-1161	Bb	F	CTT
25.24	2.3	-2	S128	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TTGGCCGCAGTGTTATCACTCATGGTTATGGC	32	1906-1937	bla	F	CTT
25.24	2.3	-1	S129	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GATCTTCACCTAGATCCTTTTAAATTAAAAAT	32	2447-2478	Bb	F	CTT

25.25	2.3	-1	S130	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TATATATGAGTAAACTTGGTCTGACAGTTACC	32	2393-2424	bla	F	CTT
25.26	2.1	-1	S131	GAGTTCCCCGCGCCAGCGGGGATAAACCG	TTTTCCGTATGTTGCATCACCTTCACCCTCTC	32	0380-0411	gfp	F	CTT
25.27	2.1	-1	S132	GAGTTCCCCGCGCCAGCGGGGATAAACCG	ATGCTTTTCTGTGACTGGTGAGTACTCAACCA	32	1835-1866	bla	F	CTT
25.27	2.3	-1	S133	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TTGGTAATGGTAGCGACCGGCGCTCAGTTGGA	32	1011-1042	Bb	F	CTT
25.28	2.3	-3	S134	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TTGGTAATGGTAGCGACCGGCGCTCAGTTGGA	32	1011-1042	Bb	F	CTT
25.28	2.3	-2	S135	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TTGGTAATGGTAGCGACCGGCGCTCAGTTGGA	32	1011-1042	Bb	F	CTT
25.28	2.3	-1	S136	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GGCAGATTGTGTCGACAGGTAATGGTTGTCTG	32	0875-0906	gfp	F	CTT
25.29	2.3	-2	S137	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TATATATGAGTAAACTTGGTCTGACAGTTACC	32	2393-2424	bla	F	CTT
25.29	2.3	-1	S138	GTGTTCCCCGCGCCAGCGGGGATAAACCG	CTGTGACCGTCTCCGGGAGCTGCATGTGTCAG	32	1130-1161	Bb	F	CTT
25.30	2.3	-1	S139	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GATCTTACCGCTGTTGAGATCCAGTTCGATGT	32	1667-1698	bla	F	CTT
25.31	2.3	-2	S140	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TGCTCATCATTGGAAAACGTTCTTCGGGGCGA	32	1710-1741	bla	F	CTT
25.31	2.3	-1	S141	GTGTTCCCCGCGCCAGCGGGGATAAACCG	ATATAGTGCGTTCCTGTACATAACCTTCGGGC	32	0552-0583	gfp	F	CTT
25.32	2.1	-1	S142	GAGTTCCCCGCGCCAGCGGGGATAAACCG	TGTTGGCCATGGAACAGGTAGTTTTCCAGTAG	32	0434-0465	gfp	F	CTT
25.32	2.3	-1	S143	GTGTTCCCCGCGCCAGCGGGGATAAACCG	AACAGTATTTGGTATCTGCGCTCTGCTGAAGC	32	2684-2715	ori	F	CTT
25.33	2.1	-1	S144	GAGTTCCCCGCGCCAGCGGGGATAAACCG	CTTGGCGTAATCATGGTCATAGCTGTTTCCTG	32	0205-0236	Bb	F	CTT
25.34	2.1	-1	S145	GAGTTCCCCGCGCCAGCGGGGATAAACCG	TTGGTAATGGTAGCGACCGGCGCTCAGTTGGA	32	1011-1042	Bb	F	CTT
25.34	2.3	-1	S146	GTGTTCCCCGCGCCAGCGGGGATAAACCG	ATATAGTGCGTTCCTGTACATAACCTTCGGGC	32	0552-0583	gfp	F	CTT
25.35	2.3	-1	S147	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TGCTCATCATTGGAAAACGTTCTTCGGGGCGA	32	1710-1741	bla	F	CTT
25.36	2.3	-1	S148	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TTAGCTTTGATTCCATTCTTTTGTTTGTCTGC	32	0748-0779	gfp	F	CTT
25.37	2.3	-1	S149	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TTGGTAATGGTAGCGACCGGCGCTCAGTTGGA	32	1011-1042	Bb	F	CTT
25.38	2.3	-2	S150	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TTAGCTTTGATTCCATTCTTTTGTTTGTCTGC	32	0748-0779	gfp	F	CTT
25.38	2.3	-1	S151	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TTAGCTTTGATTCCATTCTTTTGTTTGTCTGC	32	0748-0779	gfp	F	CTT
25.39	2.1	-1	S152	GAGTTCCCCGCGCCAGCGGGGATAAACCG	GGCAGATTGTGTCGACAGGTAATGGTTGTCTG	32	0875-0906	gfp	F	CTT
25.40	2.1	-1	S153	GAGTTCCCCGCGCCAGCGGGGATAAACCG	ATCCTTTGATCTTTTCTACGGGGTCTGACGCT	32	2532-2563	ori	F	CTT

[a] F/R refers to the direction of the new spacer relative to the original 'priming' spacer, where F is the same (or primed) direction and R is the reverse direction.

[b] PAM is defined as 5'-protospacer-PAM-3' on the targeted strand.

Table S2. pACYC184 PIMs derived from PIM25

PIM	CRISPR locus	spacer position	spacer #	repeat (5'-3')	spacer (5'-3')	spacer length (nt)	target nt	target	F/R[a]	PAM[b]
25.41	2.3	-1	S200	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TGCGTCGGGTGATGCTGCCAACTTACTGATTT	32	0436-0476	Bb	F	CTT
25.42	2.1	-1	S201	GAGTTCCCCGCGCCAGCGGGGATAAACCG	CATTCTGCCGACATGGAAGCCATCACAAACGG	32	3853-3884	cat	F	CTT
25.42	2.3	-1	S202	GTGTTCCCCGCGCCAGCGGGGATAAACCG	CACACGGTCACACTGCTTCCGGTAGTCAATAA	32	3610-3641	Bb	F	CTT

25.43	2.3	-1	S203	GTGTTCCCCGCGCCAGCGGGGATAAACCG	CACCGCCGGACATCAGCGCTAGCGGAGTGTAT	32	0566-0597	ori	F	CTT
25.44	2.1	-1	S204	GAGTTCCCCGCGCCAGCGGGGATAAACCG	CGGTCCTCGCCGAAAATGACCCAGAGCGCTGC	32	2268-2299	tet	R	CTT
25.45	2.3	-1	S205	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TGCTTCATGTGGCAGGAGAAAAAAGGCTGCAC	32	0639-0670	ori	F	CTT
25.46	2.1	-2	S206	GAGTTCCCCGCGCCAGCGGGGATAAACCG	CGCCCTGCACCATTATGTTCCGGATCTGCATC	32	3420-3451	Bb	F	CTT
25.46	2.1	-1	S207	GAGTTCCCCGCGCCAGCGGGGATAAACCG	GGCACCAATAACTGCCTTAAAAAAATTACGCC	32	3780-3811	cat	F	CTT
25.46	2.3	-1	S208	GTGTTCCCCGCGCCAGCGGGGATAAACCG	CGCTAACCGTTTTTATCAGGCTCTGGGAGGCA	32	3503-3534	Bb	F	CTT
25.47	2.1	-1	S209	GAGTTCCCCGCGCCAGCGGGGATAAACCG	CCATCACAAACGGCATGATGAACCTGAATCGC	32	3872-3903	cat	F	CTT
25.48	2.1	-1	S210	GAGTTCCCCGCGCCAGCGGGGATAAACCG	CATCACGAAATCTGACGCTCAAATCAGTGGTG	32	0876-0907	ori	F	CTT
25.49	2.1	-1	S211	GAGTTCCCCGCGCCAGCGGGGATAAACCG	ATCACTTCGCAGAATAAATAAATCCTGGTGTC	32	0373-0404	Bb	R	CTT
25.50	2.3	-1	S212	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TTGGCCCAGGGCTTCCCGGTATCAACAGGGAC	32	0344-0375	Bb	F	CTT
25.51	2.1	-1	S213	GAGTTCCCCGCGCCAGCGGGGATAAACCG	TTGTCCATATTGGCCACGTTTAAATCAAAACT	32	3971-4002	cat	F	CTT
25.51	2.3	-1	S214	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GCTAAACTGAAAGGACAAGTTTTGGTGACTGC	32	1236-1267	ori	F	CTT
25.52	2.1	-1	S215	GAGTTCCCCGCGCCAGCGGGGATAAACCG	TTGTAATTCTCATGTTTGACAGCTTATCATCG	33	1490-1521	ori	F	CTT
25.53	2.3	-1	S216	GTGTTCCCCGCGCCAGCGGGGATAAACCG	CACACGGTCACACTGCTTCCGGTAGTCAATAA	32	3610-3641	Bb	F	CTT
25.54	2.3	-1	S217	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GCCGGATAAAACTTGTGCTTATTTTTCTTTAC	32	0047-0078	cat	F	CTT
25.55	2.1	-1	S218	GAGTTCCCCGCGCCAGCGGGGATAAACCG	CGCCCTGCACCATTATGTTCCGGATCTGCATC	32	3420-3451	Bb	F	CTT
25.56	2.3	-1	S219	GTGTTCCCCGCGCCAGCGGGGATAAACCG	CACACGGTCACACTGCTTCCGGTAGTCAATAA	32	3610-3641	Bb	F	CTT
25.57	2.3	-1	S220	GTGTTCCCCGCGCCAGCGGGGATAAACCG	CGCTCATGAGCCCGAAGTGGCGAGCCCGATCT	32	1961-1992	tet	R	CTT
25.58	2.3	-1	S221	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GCCGTAATATCCAGCTGAACGGTCTGGTTATA	32	0092-0123	cat	F	CTT
25.59	2.1	-1	S222	GAGTTCCCCGCGCCAGCGGGGATAAACCG	CCATCACAAACGGCATGATGAACCTGAATCGC	32	3872-3903	cat	F	CTT
25.60	2.1	-1	S223	GAGTTCCCCGCGCCAGCGGGGATAAACCG	GCCGGATAAAACTTGTGCTTATTTTTCTTTAC	32	0047-0078	cat	F	CTT
25.60	2.3	-1	S224	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TGCTTCATGTGGCAGGAGAAAAAAGGCTGCAC	32	0639-0670	ori	F	CTT
25.61	2.1	-1	S225	GAGTTCCCCGCGCCAGCGGGGATAAACCG	ATACTTAACAGGGAAGTGAGAGGGCCGCGGCA	32	0808-0839	ori	F	CTT
25.62	2.1	-1	S226	GAGTTCCCCGCGCCAGCGGGGATAAACCG	GGTGAACACTATCCCATATCACCAGCTCACCG	32	4197-4228	cat	F	CTT
25.63	2.1	-1	S227	GAGTTCCCCGCGCCAGCGGGGATAAACCG	AGTTGGTAGCTCAGAGAACCTTCGAAAAACCG	32	1299-1330	ori	F	CTT
25.64	2.1	-1	S228	GAGTTCCCCGCGCCAGCGGGGATAAACCG	CGCTAACCGTTTTTATCAGGCTCTGGGAGGCA	32	3503-3534	Bb	F	CTT
25.65	2.1	-1	S229	GAGTTCCCCGCGCCAGCGGGGATAAACCG	CACCACTGGCAGCAGCCACTGGTAATTGATTT	32	1169-1200	ori	F	CTT
25.66	2.3	-1	S230	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GCTGCACCGGTGCGTCAGCAGAATATGTGATA	32	0664-0695	ori	F	CTT
25.67	2.1	-1	S231	GAGTTCCCCGCGCCAGCGGGGATAAACCG	TGAGAGGGCCGCGGCAAAGCCGTTTTTCCATA	32	0824-0855	ori	F	CTT
25.68	2.3	-1	S232	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GCACCAATAACTGCCTTAAAAAAATTACGCCC	32	3781-3812	cat	F	CCT
25.69	2.3	-1	S233	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GCCGGATAAAACTTGTGCTTATTTTTCTTTAC	32	0047-0078	cat	F	CTT
25.70	2.3	-1	S234	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GCTGCACCGGTGCGTCAGCAGAATATGTGATA	32	0664-0695	ori	F	CTT
25.71	2.3	-1	S235	GTGTTCCCCGCGCCAGCGGGGATAAACCG	CCGTTTTTCCATAGGCTCCGCCCCCCTGACAA	32	0843-0874	ori	F	CTT
25.72	2.1	-1	S236	GAGTTCCCCGCGCCAGCGGGGATAAACCG	CACCGCCGGACATCAGCGCTAGCGGAGTGTAT	32	0566-0597	ori	F	CTT
25.73	2.1	-1	S237	GAGTTCCCCGCGCCAGCGGGGATAAACCG	GCCGGATAAAACTTGTGCTTATTTTTCTTTAC	32	0047-0078	cat	F	CTT
25.74	2.1	-1	S238	GAGTTCCCCGCGCCAGCGGGGATAAACCG	GCCGTAATATCCAGCTGAACGGTCTGGTTATA	32	0092-0123	cat	F	CTT
25.75	2.1	-1	S239	GAGTTCCCCGCGCCAGCGGGGATAAACCG	TTTTGGTGACTGCGCTCCTCCAAGCCAGTTAC	32	1255-1286	ori	F	CTT
25.76	2.1	-1	S240	GAGTTCCCCGCGCCAGCGGGGATAAACCG	TTGGCAGCATCACCCGACGCACTTTGCGCCGA	32	0425-0456	Bb	R	CTT

25.78	2.3	-1	S241	GTGTTCCCCGCGCCAGCGGGGATAAACCG	CCATCACAAACGGCATGATGAACCTGAATCGC	32	3872-3903	cat	F	CTT
25.79	2.3	-1	S242	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GCCGTAATATCCAGCTGAACGGTCTGGTTATA	32	0092-0123	cat	F	CTT
25.81	2.1	-1	ND	ND	ND	ND	ND	ND	ND	ND
25.81	2.3	-1	S243	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GCGGTTTTTTCGTTTTCAGAGCAAGAGATTAC	32	1340-1371	ori	F	CTT
25.82	2.3	-1	S244	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GCCGTAATATCCAGCTGAACGGTCTGGTTATA	32	0092-0123	cat	F	CTT
25.83	2.3	-1	S245	GTGTTCCCCGCGCCAGCGGGGATAAACCG	AGTTGGTAGCTCAGAGAACCTTCGAAAAACCG	32	1299-1330	ori	F	CTT
25.84	2.3	-1	S246	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GCCGGATAAAACTTGTGCTTATTTTTCTTTAC	32	0047-0078	cat	F	CTT

[a]F/R refers to the direction of the new spacer relative to the original ‘priming’ spacer, where F is the same (or primed) direction and R is the reverse direction.

[b]PAM is defined as 5’-protospacer-PAM-3’ on the targeted strand.

Table S3. pGFPuv PIMs derived from PIM2

PIM	CRISPR locus	spacer position	spacer #	repeat (5’-3’)	spacer (5’-3’)	spacer length (nt)	target nt	target	F/R[a]	PAM[b]
2.01	2.1	-1	S300	GAGTTCCCCGCGCCAGCGGGGATAAACCG	CTGGGCTGTGTGCACGAACCCCCCGTTCAGCC	32	2881-2912	ori	F	CTT
2.01	2.3	-1	S301	GTGTTCCCCGCGCCAGCGGGGATAAACCG	AACATGTGAGCAAAAGGCCAGCAAAAGGCCAG	32	3187-3218	Bb/ori	F	CTT
2.02	2.1	-1	S302	GAGTTCCCCGCGCCAGCGGGGATAAACCG	GTAACTGGCTTCAGCAGAGCGCAGATACCAAA	32	2677-2708	ori	R	CTT
2.03	2.1	-1	S303	GAGTTCCCCGCGCCAGCGGGGATAAACCG	ATATAGTGCGTTCCTGTACATAACCTTCGGGC	32	0552-0583	gfp	F	CTT
2.04	2.3	-1	S304	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GGCAGATTGTGTCGACAGGTAATGGTTGTCTG	32	0875-0906	gfp	F	CTT
2.05	2.1	-1	S305	GAGTTCCCCGCGCCAGCGGGGATAAACCG	AGAAAGTAGTGACAAGTGTTGGCCATGGAACA	32	0450-0481	gfp	F	CTT
2.05	2.3	-1	S306	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TTTTCCGTATGTTGCATCACCTTCACCCTCTC	32	0380-0411	gfp	F	CTT
2.06	2.1	-2	S307	GAGTTCCCCGCGCCAGCGGGGATAAACCA	TGAGTAAACTTGGTCTGACAGTTACCAATGCT	32	2387-2418	bla	F	TAT
2.06	2.1	-1	S308	GAGTTCCCCGCGCCAGCGGGGATAAACCG	GTAACTGGCTTCAGCAGAGCGCAGATACCAAA	32	2677-2708	ori	R	CTT
2.06	2.3	-1	S309	GTGTTCCCCGCGCCAGCGGGGATAAACCG	ATGGATCCGTTCAACTAGCAGACCATTATCAA	32	0806-0837	gfp	R	CTT
2.07	2.3	-1	S310	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GAATAAGGGCGACACGGAAATGTTGAATACTC	32	1539-1570	bla	F	CCT
2.08	2.1	-1	S311	GAGTTCCCCGCGCCAGCGGGGATAAACCG	CCATACCAAACGACGAGCGTGACACCACGATG	32	2045-2076	bla	R	CTT
2.08	2.3	-2	S312	GTGTTCCCCGCGCCAGCGGGGATAAACCG	CGGAAGAGCGCCCAATACGCAAACCGCCTCTC	32	3332-0026	Bb	R	CTT
2.08	2.3	-1	S313	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GTAACTGGCTTCAGCAGAGCGCAGATACCAAA	32	2677-2708	ori	R	CTT
2.09	2.3	-1	S314	GTGTTCCCCGCGCCAGCGGGGATAAACCG	CCCTCCCGTATCGTAGTTATCTACACGACGGG	32	2293-2324	bla	R	CTT
2.09	2.3	-2	S315	GTGTTCCCCGCGCCAGCGGGGATAAACCT	TCTGCTATGTGGCGCGGTATTATCCCGTATTG	32	1752-1783	bla	R	ACT
2.10	2.1	-1	S316	GAGTTCCCCGCGCCAGCGGGGATAAACCG	TTGGCCGCAGTGTTATCACTCATGGTTATGGC	32	1906-1937	bla	F	CTT
2.10	2.3	-2	S317	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GGCCGAGCGCAGAAGTGGTCCTGCAACTTTAT	32	2171-2202	bla	F	CTT
2.10	2.3	-1	S318	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TTGGTAATGGTAGCGACCGGCGCTCAGTTGGA	32	1011-1042	Bb	F	CTT

2.11	2.1	-1	S319	GAGTTCCCCGCGCCAGCGGGGATAAACCG	TCAGAGGTGGCGAAACCCGACAGGACTATAAA	32	3074-3105	ori	F	CTT
2.11	2.3	-1	S320	GTGTTCCCCGCGCCAGCGGGGATAAACCG	ATCCTTTTTGATAATCTCATGACCAAAATCCC	32	2476-2507	Bb	R	CTT
2.12	2.1	-1	S321	GAGTTCCCCGCGCCAGCGGGGATAAACCG	GGCAGATTGTGTCGACAGGTAATGGTTGTCTG	32	0875-0906	gfp	F	CTT
2.12	2.1	-2	S322	GAGTTCCCCGCGCCAGCGGGGATAAACCG	TATATATGAGTAAACTTGGTCTGACAGTTACC	32	2393-2424	bla	F	CTT
2.12	2.3	-1	S323	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GGTAAGTTTTCCGTATGTTGCATCACCTTCAC	32	0386-0417	gfp	F	CTT
2.13	2.1	-1	S324	GAGTTCCCCGCGCCAGCGGGGATAAACCG	CATTTATCAGGGTTATTGTCTCATGAGCGGAT	32	1509-1478	bla pro	F	CTT
2.13	2.1	-2	S325	GAGTTCCCCGCGCCAGCGGGGATAAACCG	GGTAAGTTTTCCGTATGTTGCATCACCTTCAC	32	0386-0417	gfp	F	CTT
2.14	2.3	-1	S326	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GCCAGGAACCGTAAAAAGGCCGCGTTGCTGGC	32	3160-3191	ori	F	CTT
2.14	2.3	-2	S327	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GCGAGTTACATGATCCCCCATGTTGTGCAAAA	32	1982-2013	bla	F	CTT
2.15	2.1	-1	S328	GAGTTCCCCGCGCCAGCGGGGATAAACCG	CTGTGACCGTCTCCGGGAGCTGCATGTGTCAG	32	1130-1161	Bb	F	CTT
2.15	2.3	-1	S329	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TGCTCATCATTGGAAAACGTTCTTCGGGGCGA	32	1710-1741	bla	F	CTT
2.16	2.3	-1	S330	GTGTTCCCCGCGCCAGCGGGGATAAACCG	AAACCATTATTATCATGACATTAACCTATAAA	32	1365-1396	Bb	F	CTT
2.17	2.3	-1	S331	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GACCATGTGGTCACGCTTTTCGTTGGGATCTT	32	0914-0945	gfp	F	CTT
2.18	2.3	-1	S332	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GACCATGTGGTCACGCTTTTCGTTGGGATCTT	32	0914-0945	gfp	F	CTT
2.19	2.1	-1	S333	GAGTTCCCCGCGCCAGCGGGGATAAACCG	GGAATAAGGGCGACACGGAAATGTTGAATACT	32	1540-1571	bla	F	CTT
2.20	2.1	-1	S334	GAGTTCCCCGCGCCAGCGGGGATAAACCG	GCCGCCTGATGCGGTATTTTCTCCTTACGCAT	32	1307-1338	Bb	F	CTT
2.20	2.3	-1	S335	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GATCTTCTTGAGATCCTTTTTTTCTGCGCGTA	32	2561-2592	ori	R	CTT
2.21	2.1	-1	S336	GAGTTCCCCGCGCCAGCGGGGATAAACCG	TCGTGTCTTACCGGGTTGGACTCAAGACGATA	32	2824-2855	ori	R	CTT
2.22	2.1	-1	S337	GAGTTCCCCGCGCCAGCGGGGATAAACCG	TGGTCCTGCAACTTTATCCGCCTCCATCCAGT	32	2156-2187	bla	F	CTT
2.23	2.1	-1	S338	GAGTTCCCCGCGCCAGCGGGGATAAACCG	TATATATGAGTAAACTTGGTCTGACAGTTACC	32	2393-2424	bla	F	CTT
2.24	2.3	-2	S339	GTGTTCCCCGCGCCAGCGGGGATAAACCG	AAACCATTATTATCATGACATTAACCTATAAA	32	1365-1396	Bb	F	CTT
2.24	2.3	-1	S340	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TATATATGAGTAAACTTGGTCTGACAGTTACC	32	2393-2424	bla	F	CTT
2.25	2.3	-1	S341	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GTAACTGGCTTCAGCAGAGCGCAGATACCAAA	32	2677-2708	ori	R	CTT
2.26	2.1	-2	S342	GAGTTCCCCGCGCCAGCGGGGATAAACCG	TAGTTCGCCAGTTAATAGTTTGCGCAACGTTG	32	2093-2124	bla	F	CTT
2.26	2.1	-1	S343	GAGTTCCCCGCGCCAGCGGGGATAAACCG	AAACCATTATTATCATGACATTAACCTATAAA	32	1365-1396	Bb	F	CTT
2.26	2.3	-1	S344	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TTAGCTTTGATTCCATTCTTTTGTTTGTCTGC	32	0748-0779	gfp	F	CTT
2.28	2.3	-1	S345	GTGTTCCCCGCGCCAGCGGGGATAAACCG	AGAAAGTAGTGACAAGTGTTGGCCATGGAACA	32	0450-0481	gfp	F	CTT
2.29	2.3	-1	S346	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GTAACTGGCTTCAGCAGAGCGCAGATACCAAA	32	2677-2708	ori	R	CTT
2.30	2.1	-1	S347	GAGTTCCCCGCGCCAGCGGGGATAAACCG	TCATGCCGTTTCATGTGATCCGGATAACGGGA	32	0502-0533	gfp	F	CTT
2.31	2.1	-1	S348	GAGTTCCCCGCGCCAGCGGGGATAAACCG	GGCAGATTGTGTCGACAGGTAATGGTTGTCTG	32	0875-0906	gfp	F	CTT
2.32	2.1	-2	S349	GAGTTCCCCGCGCCAGCGGGGATAAACCG	ATCCTTTGATCTTTTCTACGGGGTCTGACGCT	32	2532-2563	ori	F	CTT
2.32	2.1	-1	S350	GAGTTCCCCGCGCCAGCGGGGATAAACCG	GGCAGATTGTGTCGACAGGTAATGGTTGTCTG	32	0875-0906	gfp	F	CTT
2.32	2.3	-1	S351	GTGTTCCCCGCGCCAGCGGGGATAAACCG	AACAGTATTTGGTATCTGCGCTCTGCTGAAGC	32	2684-2715	ori	F	CTT
2.33	2.1	-1	S352	GAGTTCCCCGCGCCAGCGGGGATAAACCG	TTTTCCGTATGTTGCATCACCTTCACCCTCTC	32	0380-0411	gfp	F	CTT
2.33	2.3	-2	S353	GTGTTCCCCGCGCCAGCGGGGATAAACCG	CTTGGCGTAATCATGGTCATAGCTGTTTCCTG	32	0205-0236	Bb	F	CTT
2.33	2.3	-1	S354	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TGCCACCTGACGTCTAAGAAACCATTATTATC	32	1383-1414	Bb	F	CTT
2.34	2.3	-2	S355	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TGGTGGCCTAACTACGGCTACACTAGAAGAAC	32	2713-2744	ori	F	CTT
2.34	2.3	-1	S356	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TGCTCATCATTGGAAAACGTTCTTCGGGGCGA	32	1710-1741	bla	F	CTT

2.35	2.3	-1	S357	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GATCTTCTTGAGATCCTTTTTTTCTGCGCGTA	32	2561-2592	ori	R	CTT
2.36	2.3	-3	S358	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GTAACTGGCTTCAGCAGAGCGCAGATACCAAA	32	2677-2708	ori	R	CTT
2.36	2.3	-2	S359	GTGTTCCCCGCGCCAGCGGGGATAAACCG	TCATTCTGAGAATAGTGTATGCGGCGACCGAG	32	1801-1832	bla	F	CTT
2.36	2.3	-1	S360	GTGTTCCCCGCGCCAGCGGGGATAAACCG	CTGTGACCGTCTCCGGGAGCTGCATGTGTCAG	32	1130-1161	Bb	F	CTT
2.37	2.3	-1	S361	GTGTTCCCCGCGCCAGCGGGGATAAACCG	GTAACTGGCTTCAGCAGAGCGCAGATACCAAA	32	2677-2708	ori	R	CTT

[a] F/R refers to the direction of the new spacer relative to the original ‘priming’ spacer, where F is the same (or primed) direction and R is the reverse direction.

[b] PAM is defined as 5’-protospacer-PAM-3’ on the targeted strand.

Table S4. pGFPuvKm-mPS8 PIMs derived from PIM5 in high-throughput individual plasmid loss experiments

spacer variant	spacer (5'-3')	Length (nt)	target nt	F/R[a]	PAM[b]	#spacers found
S400	AAATGCATAAACTTTTGCCATTCTCACCGGAT	32	1660-1691	F	CTT	5
S401	AACAGTATTTGGTATCTGCGCTCTGCTGAAGC	32	2193-2162	R	CTT	1
S402	AACTCTGTAGCACCGCCTACATACCTCGCTCT	32	2230-2261	F	CTT	1
S403	AACTTTTCACTGGAGTTGTCCCAATTCTTGTT	32	0305-0336	F	CTT	6
S404	AATATCCTGATTCAGGTGAAAATATTGTTGAT	32	1453-1484	F	CTT	9
S405	AATGGAATCAAAGCTAACTTCAAAATTCGCCA	32	0763-0794	F	CTT	2
S406	ACACGACTTATCGCCACTGGCAGCAGCCACTG	32	2307-2276	R	CTT	2
S407	ACGATAGTTACCGGATAAGGCGCAGCGGTCGG	32	2328-2359	F	CTT	1
S408	ACGCGTGCTGAAGTCAAGTTTGAAGGTGATAC	32	0610-0641	F	CTT	1
S409	AGCGCCCAATACGCAAACCGCCTCTCCCCGCG	32	0001-0032	F	CTT	1
S410	AGTATGAGCCATATTCAACGGGAAACGTCTTG	32	1116-1147	F	CTT	9
S411	AGTGCCATGCCCGAAGGTTATGTACAGGAACG	32	0544-0575	F	CTT	3
S412	AGTTGGTAGCTCTTGATCCGGCAAACAAACCA	32	2142-2111	R	CTT	3
S413	ATCCTTTGATCTTTTCTACGGGGTCTGACGCT	32	2041-2010	R	CTT	1
S414	ATGACGGGAACTACAAGACGCGTGCTGAAGTC	32	0593-0624	F	CTT	2
S415	ATGGAAACATTCTCGGACACAAACTCGAGTAC	32	0686-0717	F	CTT	5
S416	ATGGATCCGTTCAACTAGCAGACCATTATCAA	32	0806-0837	F	CTT	16
S417	CATTTTATCCGTACTCCTGATGATGCATGGTT	32	1374-1405	F	CTT	3
S418	CCCGATGCGCCAGAGTTGTTTCTGAAACATGG	32	1257-1288	F	CTT	5
S419	CCGTTTCTGTAATGAAGGAGAAAACTCACCGA	32	1856-1825	R	CTT	1
S420	CGAGTGATTTTGATGACGAGCGTAATGGCTGG	32	1606-1637	F	CAT	1
S421	CGGAAGAGCGCCCAATACGCAAACCGCCTCTC	32	2810-0026	F	CTT	2
S422	CGGCAGGGTCGGAACAGGAGAGCGCACGAGGG	32	2493-2524	F	CTT	2
S423	CGGGCAGTGAGCGCAACGCAATTAATGTGAGT	32	0085-0116	F	CTT	9
S424	CGTGACCACATGGTCCTTCTTGAGTTTGTAAC	32	0931-0962	F	CTT	8
S425	CGTTGCCAATGATGTTACAGATGAGATGGTCA	32	1298-1329	F	CTA	1
S426	CTAACTTCAAAATTCGCCACAACATTGAAGAT	32	0776-0807	F	CTT	8
S427	CTCACATTAATTGCGTTGCGCTCACTGCCCGC	32	0115-0084	R	TTA	1
S428	CTGTTGAACAAGTCTGGAAAGAAATGCATAAA	32	1639-1670	F	GCC	1
S429	CTTGCATGCCTGCAGGTCGACTCTAGAGGATC	32	0237-0268	F	CTT	13
S430	GAAGAGTATGAGCCATATTCAACGGGAAACGT	32	1112-1143	F	CTT	3
S431	GAGTGCCATGCCCGAAGGTTATGTACAGGAACG	33	0543-0575	F	CTT	1
S432	GCGCAGCGGTCGGGCTGAACGGGGGGTTCGTG	32	2347-2378	F	CTT	1

S433	GCGGACAGGTATCCGGTAAGCGGCAGGGTCGG	32	2473-2504	F	CTT	3
S434	GGCAGATTGTGTCGACAGGTAATGGTTGTCTG	32	0906-0875	R	CTT	2
S435	GGTAAGTTTTCCGTATGTTGCATCACCTTCAC	32	0417-0386	R	CTT	1
S436	GGTATCACCTTCAAACTTGACTTCAGCACGCG	32	0642-0611	R	CTT	1
S437	GTAACTGGCTTCAGCAGAGCGCAGATACCAAA	32	2155-2186	F	CTT	12
S438	GTAGCGTTGCCAATGATGTTACAGATGAGATG	32	1294-1325	F	CTT	8
S439	GTAGCGTTGCCAATGATGTTACAGATGAGATGG	33	1294-1326	F	CTT	1
S440	GTATTGATTTTAAAGAAGATGGAAACATTCTC	32	0668-0699	F	CTT	1
S441	GTGATACCCTTGTTAATCGTATCGAGTTAAAA	32	0635-0666	F	CTT	10
S442	GTGATGCAACATACGGAAAACTTACCCTTAAA	32	0392-0423	F	CTT	14
S443	TCATGCCGTTTCATGTGATCCGGATAACGGGA	32	0533-0502	R	CTT	1
S444	TCGTGTCTTACCGGGTTGGACTCAAGACGATA	32	2302-2333	F	CTT	3
S445	TCTGGAAAGAAATGCATAAACTTTTGCCATTC	32	1651-1682	F	CTT	1
S446	TGGTGGCCTAACTACGGCTACACTAGAAGAAC	32	2222-2191	R	CTT	1
S447	TTTGAAGGTGATACCCTTGTTAATCGTATCGA	32	0628-0659	F	CTT	4
S448	TTTTCCGTATGTTGCATCACCTTCACCCTCTC	32	0411-0380	R	CTT	2

[a] F/R refers to the direction of the new spacer relative to the original 'priming' spacer, where F is the same (or primed) direction and R is the reverse direction.

[b] PAM is defined as 5'-protospacer-PAM-3' on the targeted strand.

4

Table S5. Oligonucleotides used in this study

Name	Sequence (5'-3')[a]	Description	Restriction site
BG3113	GCTGGAGAAATACAACCGCCGGCCCCACCT	F CRISPR2.1 J3 replacement	
BG3114	CTGAGGCAGTAAGGAAATTAACGCGCGACA	F CRISPR2.1 J3 replacement	
BG3414	GGTAGATTTTAGTTTGTATAGAG	F CRISPR2.3 leader	
BG3415	CAACAGCAGCACCCATGAC	R CRISPR2.3 spacer 3	
BG3474	AAATGTTACATTAAGGTTGGTG	F CRISPR2.1 leader	
BG3475	GAAATTCCAGACCCGATCC	R CRISPR2.1 spacer 4	
BG4225	TTT**GAATTC**GCGCTGCATGCCTATTTG	F Km^R in pRSF-1b	EcoRI
BG4226	TTTT**CCATGG**TTAGAAAAACTCATCGAGCATC	R Km^R in pRSF-1b with PIM5 PS8	NcoI
BG4227	TTTT**CCATGG***AAAAGTGCCACTTGCGGAGACCCGGTCGTCAA**CTT***TTAGAAAAACTCATCGAGCATC	R Km^R in pRSF-1b with PIM5 primed PS8	NcoI
BG4228	TTTT**CCATGG***AAAAGTGCCACTTGCGGAGACCCGGTCGTCAG**CTT***TTAGAAAAACTCATCGAGCATC	R Km^R in pRSF-1b with PIM5 mutant PS8 (85% WT nt 5% each other nt in PS + PAM)	NcoI
BG4325	aactaaTAAATTGCAGTTTCATTTGATG	Barcoded F primer for pGFPKm-mPS8 T0A	
BG4326	acacttTAAATTGCAGTTTCATTTGATG	Barcoded F primer for pGFPKm-mPS8 T0B	
BG4327	acgcatTAAATTGCAGTTTCATTTGATG	Barcoded F primer for pGFPKm-mPS8 T0C	
BG4328	acgttcTAAATTGCAGTTTCATTTGATG	Barcoded F primer for pGFPKm-mPS8 T1A	

BG4329	actaacTAAATTGCAGTTTCATTTGATG	Barcoded F primer for pGFPKm-mPS8 T1B
BG4330	actctcTAAATTGCAGTTTCATTTGATG	Barcoded F primer for pGFPKm-mPS8 T1C
BG4331	agactcTAAATTGCAGTTTCATTTGATG	Barcoded F primer for pGFPKm-mPS8 T1D
BG4332	agcaacTAAATTGCAGTTTCATTTGATG	Barcoded F primer for pGFPKm-mPS8 T2A
BG4333	agcagaTAAATTGCAGTTTCATTTGATG	Barcoded F primer for pGFPKm-mPS8 T2B
BG4334	agtcagTAAATTGCAGTTTCATTTGATG	Barcoded F primer for pGFPKm-mPS8 T2C
BG4335	agtgcgTAAATTGCAGTTTCATTTGATG	Barcoded F primer for pGFPKm-mPS8 T3A
BG4336	atcacgTAAATTGCAGTTTCATTTGATG	Barcoded F primer for pGFPKm-mPS8 T3B
BG4337	atcgttTAAATTGCAGTTTCATTTGATG	Barcoded F primer for pGFPKm-mPS8 T3C
BG4340	caccgcTAAATTGCAGTTTCATTTGATG	Barcoded F primer for pGFPKm-mPS8 +Cntr
BG4341	cactgtGTGGAACGAAAACTCACG	Barcoded R primer for pGFPKm-mPS8 T0A
BG4342	cagtagGTGGAACGAAAACTCACG	Barcoded R primer for pGFPKm-mPS8 T0B
BG4343	catgatGTGGAACGAAAACTCACG	Barcoded R primer for pGFPKm-mPS8 T0C
BG4344	catgccGTGGAACGAAAACTCACG	Barcoded R primer for pGFPKm-mPS8 T1A
BG4345	ccagttGTGGAACGAAAACTCACG	Barcoded R primer for pGFPKm-mPS8 T1B
BG4346	ccggaaGTGGAACGAAAACTCACG	Barcoded R primer for pGFPKm-mPS8 T1C
BG4347	cctaccGTGGAACGAAAACTCACG	Barcoded R primer for pGFPKm-mPS8 T2A
BG4348	cctagaGTGGAACGAAAACTCACG	Barcoded R primer for pGFPKm-mPS8 T2B
BG4349	cctgtaGTGGAACGAAAACTCACG	Barcoded R primer for pGFPKm-mPS8 T2C
BG4350	taatagGTGGAACGAAAACTCACG	Barcoded R primer for pGFPKm-mPS8 T3A
BG4351	tactctGTGGAACGAAAACTCACG	Barcoded R primer for pGFPKm-mPS8 T3B
BG4352	tcacagGTGGAACGAAAACTCACG	Barcoded R primer for pGFPKm-mPS8 T3C
BG4356	ttgataGTGGAACGAAAACTCACG	Barcoded R primer for pGFPKm-mPS8 +Cntr

[a]Restriction sites are in bold, protospacers are in italics, PAMs are in bold italics and barcodes are in lowercase

Table S6. Read counts, barcodes and scaling factors of the large scale sequencing experiment

Sample	Barcode 5’-3’	Primer	Read count	Total reads	Scaling factor
T0A	*AACTAA*	forward	1,455,741	4,425,417	0.642264
	CACTGT	reverse	2,970,335		
T0B	*ACACTT*	forward	1,195,931	3,866,736	0.561182
	CAGTAG	reverse	2,671,231		
T0C	*ACGCAT*	forward	1,043,464	2,618,215	0.379984
	CATGAT	reverse	1,575,124		
T1A	*ACGTTC*	forward	845,894	2,260,308	0.32804
	CATGCC	reverse	1,414,824		
T1B	*ACTAAC*	forward	1,284,315	3,504,841	0.50866
	CCAGTT	reverse	2,220,967		
T1C	*ACTCTC*	forward	953,688	3,004,487	0.436043
	CCGGAA	reverse	2,051,229		

T2A	*AGACTC*	forward	1,812,468	5,718,939	0.829994
	CCTACC	reverse	3,907,707		
T2B	*AGCAAC*	forward	2,478,705	6,890,338	1
	CCTAGA	reverse	4,412,609		
T2C	*AGCAGA*	forward	1,902,481	4,761,649	0.691062
	CCTGTA	reverse	2,859,905		
T3A	*AGTCAG*	forward	1,981,100	5,249,299	0.761835
	TAATAG	reverse	3,269,185		
T3B	*AGTGCG*	forward	1,138,619	3,067,987	0.445259
	TACTCT	reverse	1,929,819		
T3C	*ATCACG*	forward	1,558,125	3,830,483	0.555921
	TCACAG	reverse	2,273,260		
Total				49,198,699	

Table S7. PAM sequence analysis[a]

PAM	# mut in PAM	No additional mutations	Behavior based on original assignment	Comp. inferred behavior	# seq with 1 additional mutation	D fraction	U fraction	P fraction	S fraction	p-value for D enrichment	p-value for U enrichment	p-value for P enrichment	p-value for S enrichment	# seq with 1 additional mutation	D fraction	U fraction	P fraction	S fraction	p-value for D enrichment	p-value for U enrichment	p-value for P enrichment	p-value for S enrichment
CTT	0	TRUE	D	D	97	0.825	0.134	0.041	0.000	0.000	0.984	0.924	1.000	4328	0.597	0.323	0.079	0.002	0.050	0.828	0.860	0.160
ATT	1	TRUE	U	U	97	0.000	0.835	0.134	0.031	1.000	0.005	0.699	0.160	845	0.000	0.769	0.169	0.062	1.000	0.051	0.555	0.154
CAT	1	TRUE	D		93	0.720	0.194	0.086	0.000	0.001	0.982	0.921	1.000	418	0.347	0.486	0.167	0.000	0.360	0.623	0.555	1.000
CCT	1	TRUE	D	D	82	0.646	0.244	0.098	0.012	0.007	0.929	0.924	0.163	321	0.368	0.467	0.159	0.006	0.360	0.623	0.555	0.157
CGT	1	TRUE	P	P	96	0.000	0.417	0.563	0.021	1.000	0.562	0.012	0.161	661	0.000	0.536	0.445	0.020	1.000	0.375	0.019	0.146
CTA	1	TRUE	D	D	95	0.737	0.179	0.084	0.000	0.001	0.985	0.919	1.000	424	0.361	0.441	0.193	0.005	0.366	0.622	0.558	0.150
CTC	1	TRUE	D	D	82	0.683	0.220	0.098	0.000	0.007	0.924	0.924	1.000	285	0.375	0.418	0.204	0.004	0.367	0.623	0.259	0.162
CTG	1	TRUE	P		95	0.000	0.674	0.326	0.000	1.000	0.131	0.163	1.000	630	0.000	0.662	0.319	0.019	1.000	0.170	0.082	0.160
GTT	1	TRUE	U	U	96	0.010	0.854	0.125	0.010	0.963	0.006	0.706	0.166	721	0.014	0.785	0.178	0.024	0.973	0.053	0.548	0.160
TTT	1	TRUE	U	NA	96	0.469	0.458	0.073	0.000	0.113	0.577	0.921	1.000	638	0.169	0.650	0.171	0.009	0.851	0.175	0.556	0.156
AAT	2	TRUE	U	S	12	0.000	0.750	0.083	0.167	1.000	0.038	0.921	0.014	48	0.000	0.771	0.083	0.146	1.000	0.053	0.857	0.013
ACT	2	TRUE	U	S	15	0.000	0.533	0.333	0.133	1.000	0.313	0.172	0.012	45	0.000	0.689	0.244	0.067	1.000	0.178	0.254	0.157
AGT	2	TRUE	P	P	17	0.000	0.647	0.353	0.000	1.000	0.129	0.174	1.000	47	0.000	0.638	0.340	0.021	1.000	0.175	0.084	0.155
ATA	2	TRUE	P	P/S	16	0.000	0.563	0.438	0.000	1.000	0.318	0.050	1.000	55	0.000	0.582	0.309	0.109	1.000	0.371	0.082	0.012
ATC	2	TRUE	U	S	17	0.000	0.588	0.059	0.353	1.000	0.318	0.925	0.000	58	0.000	0.741	0.034	0.224	1.000	0.055	0.865	0.001
ATG	2	TRUE	U	U	17	0.000	0.588	0.412	0.000	1.000	0.314	0.052	1.000	47	0.000	0.660	0.298	0.043	1.000	0.170	0.261	0.155
CAA	2	TRUE	U	U	11	0.000	0.818	0.182	0.000	1.000	0.007	0.698	1.000	33	0.000	0.727	0.273	0.000	1.000	0.053	0.257	1.000
CAC	2	TRUE	P	P	6	0.000	0.333	0.667	0.000	1.000	0.781	0.002	1.000	35	0.000	0.571	0.429	0.000	1.000	0.372	0.020	1.000
CAG	2	TRUE	P	NA	11	0.000	0.455	0.455	0.091	1.000	0.565	0.052	0.165	35	0.000	0.571	0.371	0.057	1.000	0.383	0.086	0.155
CCA	2	TRUE	U	P	7	0.000	0.571	0.429	0.000	1.000	0.311	0.054	1.000	27	0.000	0.519	0.481	0.000	1.000	0.371	0.020	1.000
CCC	2	TRUE	U	P/S	3	0.000	0.333	0.333	0.333	1.000	0.783	0.171	0.000	10	0.000	0.400	0.500	0.100	1.000	0.825	0.020	0.154

CCG	2	TRUE	P	P	12	0.000	0.417	0.583	0.000	1.000	0.561	0.012	1.000	33	0.000	0.485	0.485	0.030	1.000	0.614	0.019	0.157
CGA	2	TRUE	U	P	12	0.000	0.583	0.417	0.000	1.000	0.317	0.049	1.000	38	0.000	0.526	0.421	0.053	1.000	0.371	0.020	0.161
CGC	2	TRUE	P	P	10	0.000	0.200	0.800	0.000	1.000	0.985	0.000	1.000	38	0.000	0.500	0.500	0.000	1.000	0.622	0.019	1.000
CGG	2	TRUE	P	P	14	0.000	0.429	0.571	0.000	1.000	0.564	0.012	1.000	38	0.000	0.553	0.447	0.000	1.000	0.372	0.020	1.000
GAT	2	TRUE	U	U	10	0.000	0.800	0.100	0.100	1.000	0.039	0.925	0.164	43	0.000	0.791	0.140	0.070	1.000	0.054	0.550	0.160
GCT	2	TRUE	P	P	21	0.000	0.619	0.333	0.048	1.000	0.136	0.176	0.163	53	0.000	0.660	0.321	0.019	1.000	0.166	0.083	0.157
GGT	2	TRUE	U	P	23	0.000	0.435	0.565	0.000	1.000	0.564	0.011	1.000	69	0.000	0.565	0.391	0.043	1.000	0.379	0.085	0.157
GTA	2	TRUE	U	U	15	0.000	0.733	0.267	0.000	1.000	0.039	0.393	1.000	45	0.000	0.689	0.222	0.089	1.000	0.170	0.251	0.157
GTC	2	TRUE	P	P/S	12	0.000	0.583	0.333	0.083	1.000	0.321	0.165	0.161	35	0.000	0.600	0.286	0.114	1.000	0.377	0.255	0.010
GTG	2	TRUE	P		14	0.000	0.500	0.500	0.000	1.000	0.559	0.050	1.000	47	0.000	0.532	0.383	0.085	1.000	0.385	0.084	0.158
TAT	2	TRUE	U	U	11	0.091	0.727	0.182	0.000	0.965	0.038	0.704	1.000	41	0.024	0.707	0.244	0.024	0.977	0.052	0.248	0.156
TCT	2	TRUE	P	P	22	0.000	0.545	0.409	0.045	1.000	0.315	0.053	0.158	46	0.000	0.630	0.348	0.022	1.000	0.167	0.084	0.156
TGT	2	TRUE	P	P	15	0.000	0.467	0.533	0.000	1.000	0.554	0.010	1.000	51	0.000	0.588	0.412	0.000	1.000	0.376	0.019	1.000
TTA	2	TRUE	U	U	16	0.000	0.813	0.188	0.000	1.000	0.006	0.698	1.000	56	0.000	0.804	0.196	0.000	1.000	0.011	0.554	1.000
TTC	2	TRUE	U	U	9	0.000	0.889	0.111	0.000	1.000	0.006	0.698	1.000	39	0.000	0.769	0.205	0.026	1.000	0.056	0.251	0.152
TTG	2	TRUE	U	NA	24	0.000	0.625	0.375	0.000	1.000	0.134	0.169	1.000	69	0.000	0.609	0.391	0.000	1.000	0.163	0.085	1.000
AAA	3	FALSE	NA	U	1	0.000	1.000	0.000	0.000	1.000	0.001	1.000	1.000	2	0.000	1.000	0.000	0.000	1.000	0.002	1.000	1.000
AAC	3	FALSE	NA	U	1	0.000	1.000	0.000	0.000	1.000	0.001	1.000	1.000	3	0.000	0.667	0.333	0.000	1.000	0.169	0.091	1.000
AAG	3	FALSE	NA	S	1	0.000	1.000	0.000	0.000	1.000	0.100	1.000	1.000	2	0.000	0.500	0.000	0.500	1.000	0.627	1.000	0.000
ACA	3	FALSE	NA	U	0	NA	NA	NA	NA	NA	NA	NA	NA	2	0.000	1.000	0.000	0.000	1.000	0.001	1.000	1.000
ACC	3	FALSE	NA	U	2	0.000	1.000	0.000	0.000	1.000	0.000	1.000	1.000	3	0.000	1.000	0.000	0.000	1.000	0.001	1.000	1.000
ACG	3	FALSE	NA	P	0	NA	NA	NA	NA	NA	NA	NA	NA	2	0.000	0.500	0.500	0.000	1.000	0.614	0.021	1.000
AGA	3	FALSE	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA	3	0.000	0.333	0.333	0.333	1.000	0.833	0.081	0.000
AGC	3	FALSE	NA	U	0	NA	NA	NA	NA	NA	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA
AGG	3	FALSE	NA	U	0	NA	NA	NA	NA	NA	NA	NA	NA	2	0.000	1.000	0.000	0.000	1.000	0.001	1.000	1.000
GAA	3	FALSE	NA	U	1	0.000	1.000	0.000	0.000	1.000	0.001	1.000	1.000	2	0.000	1.000	0.000	0.000	1.000	0.002	1.000	1.000
GAC	3	FALSE	NA	U	1	0.000	1.000	0.000	0.000	1.000	0.001	1.000	1.000	7	0.000	0.714	0.286	0.000	1.000	0.052	0.262	1.000
GAG	3	FALSE	NA	P	0	NA	NA	NA	NA	NA	NA	NA	NA	2	0.000	0.500	0.500	0.000	1.000	0.611	0.018	1.000

GCA	3	FALSE	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA
GCC	3	FALSE	NA	U	0	NA	NA	NA	NA	NA	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA
GCG	3	FALSE	NA	U	0	NA	NA	NA	NA	NA	NA	NA	NA	3	0.000	1.000	0.000	0.000	1.000	0.001	1.000	1.000
GGA	3	FALSE	NA	U/S	1	0.000	1.000	0.000	0.000	1.000	0.000	1.000	1.000	4	0.000	0.250	0.250	0.500	1.000	0.941	0.263	0.000
GGC	3	FALSE	NA	U	0	NA	NA	NA	NA	NA	NA	NA	NA	2	0.000	1.000	0.000	0.000	1.000	0.001	1.000	1.000
GGG	3	TRUE	U	NA	4	0.000	0.500	0.500	0.000	1.000	0.563	0.054	1.000	6	0.000	0.667	0.333	0.000	1.000	0.174	0.085	1.000
TAA	3	FALSE	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA
TAC	3	TRUE	U	U	2	0.000	1.000	0.000	0.000	1.000	0.001	1.000	1.000	5	0.000	0.600	0.400	0.000	1.000	0.368	0.086	1.000
TAG	3	FALSE	NA	P	0	NA	NA	NA	NA	NA	NA	NA	NA	1	0.000	0.000	1.000	0.000	1.000	1.000	0.000	1.000
TCA	3	FALSE	NA	P	2	0.000	0.500	0.500	0.000	1.000	0.563	0.051	1.000	2	0.000	0.500	0.500	0.000	1.000	0.623	0.021	1.000
TCC	3	TRUE	P	P/S	1	0.000	0.000	1.000	0.000	1.000	1.000	0.000	1.000	2	0.000	0.000	0.500	0.500	1.000	1.000	0.017	0.000
TCG	3	FALSE	NA	U	0	NA	NA	NA	NA	NA	NA	NA	NA	3	0.000	0.667	0.333	0.000	1.000	0.162	0.084	1.000
TGA	3	FALSE	NA	P	1	0.000	0.000	1.000	0.000	1.000	1.000	0.000	1.000	4	0.000	0.500	0.500	0.000	1.000	0.617	0.019	1.000
TGC	3	FALSE	NA	P	1	0.000	0.000	1.000	0.000	1.000	1.000	0.000	1.000	3	0.000	0.000	1.000	0.000	1.000	1.000	0.000	1.000
TGG	3	FALSE	NA	P	1	0.000	0.000	1.000	0.000	1.000	1.000	0.000	1.000	1	0.000	0.000	1.000	0.000	1.000	1.000	0.000	1.000

[a] All PAM-only mutants were extracted from the dataset and their behavior scored (37 out 64 possible PAMs found). The behavior of PAMs not present as a PAM-only mutant was predicted computationally by analyzing the distribution of categories in sequences with up to two additional non-PAM mutations. Grey cells indicate statistical significance based on p-values < 0.05.

Chapter 5

Cas3-derived target DNA degradation fragments fuel primed CRISPR adaptation

Tim Künne[1], Sebastian N. Kieper[1], Jasper W. Bannenberg[1], Anne I.M. Vogel[1,4], Willem R. Miellet[1], Misha Klein[2], Martin Depken[2], Maria Suarez-Diez[3], Stan J.J. Brouns[1,2]

[1]Laboratory of Microbiology, Wageningen University, 6708 WE Wageningen, Netherlands.
[2]Kavli Institute of Nanoscience and Department of BioNanoscience, Delft University of Technology, 2629 HZ, Delft, The Netherlands
[3]Laboratory of Systems and Synthetic Biology, Wageningen University, 6708 WE Wageningen, Netherlands.
[4]Current address: Department of Biotechnology, NTNU, N-7491 Trondheim, Norway

This chapter has been published as:

Künne T, Kieper SN, Bannenberg JW, Vogel AIM, Miellet WR, Klein M, Depken M, Suarez-Diez M, Brouns SJ 2016. Mol Cell, 63(5):852-64

Abstract

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

Keywords

CRISPR-Cas; Priming; Interference; adaptive immunity; Phage resistance; Cascade; Cas1; Cas2; Cas3; Spacer acquisition

Introduction

Priming is a mechanism by which immune systems provide an improved immune response to parasite exposure. In vertebrates, priming of adaptive immunity can occur upon first contact of a T or B cell with a specific antigen and causes epigenetic changes as well as cell differentiation into effector T or B cells, producing high levels of antibodies (Bevington et al., 2016). More recently, immune priming has been observed in invertebrates, where it provides increased resistance to previously encountered pathogens (Kurtz and Franz, 2003; Schmid-Hempel, 2005). In plants, priming refers to a state in which the plant can activate its defense responses more rapidly and strongly when challenged by pathogenic microbes, insects, or environmental stress (Conrath et al., 2015). In microbes, priming is a mechanism in which cells can update their immunological memory to provide protection against previously encountered but slightly changed viruses or conjugative plasmids (Datsenko et al., 2012; Li et al., 2014; Richter et al., 2014; Swarts et al., 2012; Vorontsova et al., 2015). Microbial adaptive immune systems do this by integrating short fragments of invader DNA sequences (called spacers) into Clusters of Regularly Interspaced Short Palindromic Repeats (CRISPR). These spacers are transcribed and processed into small CRISPR RNAs and guide Cas (CRISPR-associated) surveillance complexes such as Cascade, Cas9, Cpf1, Csm and Cmr to their DNA or RNA target sequences, resulting in target cleavage and neutralization of the invading threat (Carter and Wiedenheft, 2015; Charpentier et al., 2015; Makarova et al., 2015; Marraffini, 2015; Reeks et al., 2013).

For many years, the acquisition of new spacers was the least understood process in CRISPR-Cas defense, but recent advances have begun to change this (Amitai and Sorek, 2016; Fineran and Charpentier, 2012; Heler et al., 2014; Sternberg et al., 2016). In the Type I-E system of *E. coli*, Cas1 and Cas2 form a complex that binds, processes and integrates DNA fragments into the CRISPR array to form spacers (Arslan et al., 2014; Nunez et al., 2014; Nunez et al., 2015b; Rollie et al., 2015; Wang et al., 2015). Apart from priming, spacers can also be acquired in a naïve manner. During naïve acquisition the host acquires spacers from an invading DNA element that has not been catalogued in the CRISPR array yet. This process is dependent on DNA replication of the invading DNA element (Levy et al., 2015) and requires only *cas1* and *cas2* genes (Yosef et al., 2012). In type I CRISPR-Cas systems, primed acquisition makes use of pre-existing spacers that partially match an invading DNA element. Therefore, primed acquisition of spacers is important to rapidly counter invaders that escape immunity by mutating their target site (Cady et al., 2012; Datsenko et al., 2012; Fineran et al., 2014; Semenova et al., 2011; Xue et al., 2015). Priming allows new spacers from such an ‘escaper’ to be rapidly acquired, leading to renewed immunity. Priming is especially advantageous for a host because the process quickly generates a population of bacteria with different spacers against the same virus, efficiently driving the virus extinct (van Houte et al., 2016). In addition to Cas1-2, all remaining Cas proteins are required for priming, including the crRNA effector complex Cascade and the nuclease-helicase Cas3 (Datsenko et al., 2012; Richter et al., 2014). Despite knowing the genetic requirements for priming, the exact role of these proteins during priming remains unknown. Several models that explain parts of the priming process have been proposed.

In the Cascade-sliding model, Cascade moves along the DNA until a PAM is encountered, which marks the DNA for acquisition of a new spacer (Datsenko et al., 2012). A second model was proposed in which a Cas1:Cas2-3 complex translocates away from the primed protospacer marked by the crRNA-effector complex until a new PAM is encountered (Richter et al., 2014). This new site is then used to acquire a new spacer from. Recently, supporting evidence for this hypothesis has been obtained. Single molecule studies have suggested that Cascade bound to a priming protospacer recruits Cas1-2, which in turn recruit a nuclease inactive Cas3 (Redding et al., 2015). A complex of Cas1-3 may then translocate along the DNA to select new spacers. While these models describe the biochemistry and movement of the proteins involved in priming, it has remained unknown how actual DNA fragments from an invading element are obtained to drive the priming process. We have previously put forward a model in which we propose that DNA breakdown products of Cas3 provide the positive feedback needed to fuel the priming process (Swarts et al., 2012). Similar models were proposed for priming in I-B and I-F systems (Li et al., 2014; Vorontsova et al., 2015). In line with that hypothesis, it has recently been suggested that during naïve acquisition spacer precursors are generated during DNA repair at double stranded breaks (Levy et al., 2015). These breaks are frequently formed at stalled replication forks during DNA replication and are repaired by the RecBCD complex. RecBCD unwinds the DNA strands with its helicase activity, while degrading the subsequent single stranded stretches using exonuclease activity. The resulting DNA oligomers have been proposed to form precursors for Cas1-2 to produce new spacers. Similar to RecBCD, Cas3 is also a nuclease-helicase that degrades dsDNA by unwinding, with the difference that Cas3 has been shown to degrade one strand at a time (Gong et al., 2014; Huo et al., 2014; Mulepati and Bailey, 2013; Sinkunas et al., 2013; Westra et al., 2012c). This leads to the hypothesis that Cas3 also produces substrates for Cas1-2 mediated spacer acquisition during priming.

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

Results

Previous studies have shown that direct interference in Type I CRISPR-Cas systems (*i.e.* the breakdown of Cascade-flagged invading DNA by Cas3) is relatively sensitive to mutations in the PAM and seed sequence of the protospacer (Künne et al., 2014; Semenova et al., 2011; Wiedenheft et al., 2011b; Xue et al., 2015). Priming on the other hand is an extremely robust process capable of dealing with highly mutated targets with up to 13 mutations. Priming is influenced by a complex combination of the number of mutations in a target, the position of these mutations, and the nucleotide identity of the mutation. Furthermore, the degree of tolerance of mutations in a protospacer during interference and priming depends on the spacer choice (Xue et al., 2015).

Timing of plasmid loss and spacer acquisition reveals distinct underlying processes

In order to find the molecular explanation for why some mutants with equal numbers of mutations show priming while others do not, we performed detailed analysis of a selected set of target mutants obtained previously (Fineran et al., 2014). From the available list we chose the *bona fide* target (WT) and 30 mutants carrying an interference permissive PAM (*i.e.* 5’-CTT-3’). The mutants had between 2 and 5 effective mutations (i.e. mutations outside the kinked positions, 6, 12, 18, 24, 30 (Fineran et al., 2014; Jackson et al., 2014; Mulepati et al., 2014; Zhao et al., 2014)) (Figure S1). We used *E. coli* strain KD263 with inducible expression of *cas3* and *cascade-cas1-2* genes (Shmakov et al., 2014) to test both direct interference and priming in a plasmid loss setup. Plasmid loss curves of individual mutants (Figure S2) showed four distinct behaviors that led us to classify these target mutants into four groups: mutants capable of only direct interference (D^+P^-), mutants capable of direct interference and priming (D^+P^+), mutants capable of only priming (D^-P^+), and mutants incapable of both direct interference and priming (D^-P^-) (Figure 1A, B).

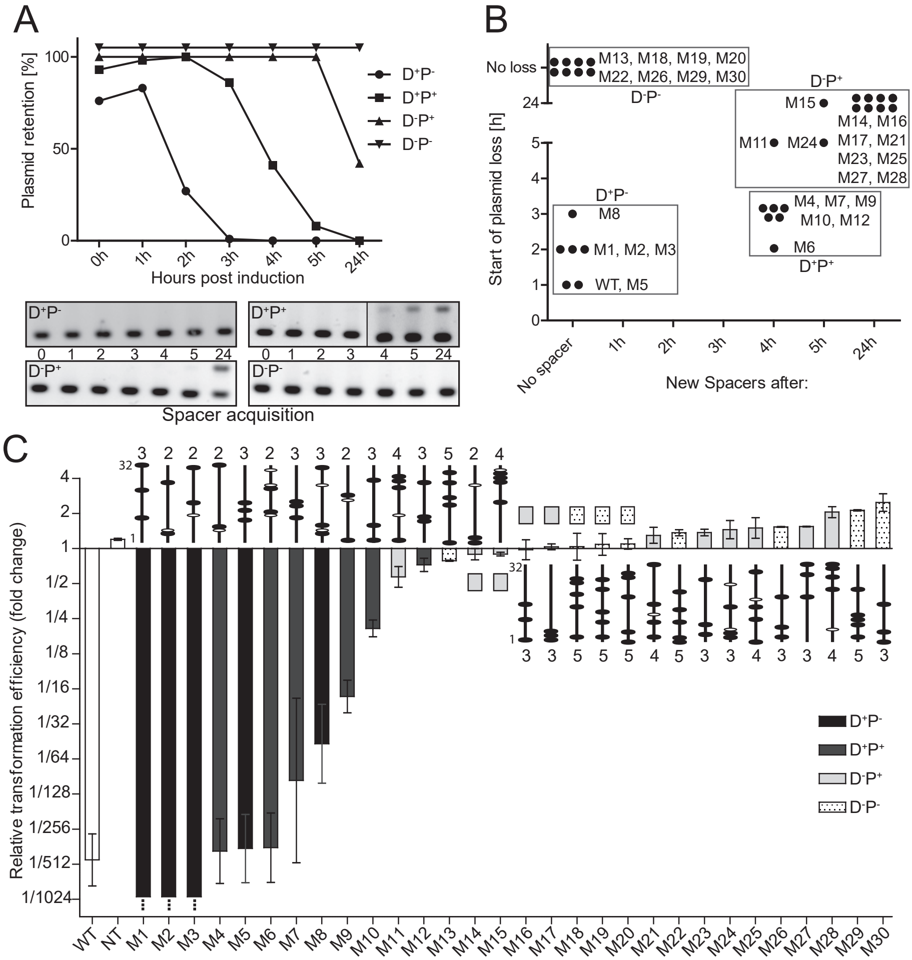


Figure 1: Plasmid loss and transformation assay. Plasmid loss was assessed by plating cells and scoring for the GFP signal at various time points after induction of *cas* genes. Individual assays can be seen in Figure S2. The *bona fide* target is abbreviated as WT. A) Example curves and CRISPR PCR of four different types of plasmid behaviors that were observed: Rapid plasmid loss without spacer integration (D^+P^-), delayed plasmid loss and spacer integration (D^+P^+), strongly delayed plasmid loss and spacer integration (D^-P^+), and no plasmid loss with no spacer integration (D^-P^-). B) Summary of plasmid behavior of all mutants, showing timing of first plasmid loss and time of first observable spacer integration. C) The relative transformation efficiency is plotted for all mutant plasmids (fold change compared to co-transformed non-target plasmid, log2 scale). Bars are color coded based on plasmid behavior classification. Error bars represent the standard error of the mean of triplicate experiments. The positions of mutations are indicated schematically for each mutant (Pos1: Bottom, Pos32: Top). Open ovals represent mutations on positions 6, 12, 18, 24, 30. Closed ovals represent mutations outside of those positions (effective mutations). The amount of effective mutations is indicated above or below the schematic. For a more detailed overview of the mutations, see Figure S1.

As expected, rapid plasmid loss was observed for the *bona fide* target, but also for five mutant targets. These target variants (D^+P^-) showed plasmid loss within 2 hours post induction (hpi), reaching complete loss after 3 hpi (Figure 1B bottom left cluster), and did not incorporate new spacers. The $D^{+}P^{+}$ group of mutants showed a slower decrease in plasmid abundance (starting ~3 hpi) and this decrease was accompanied by incorporation of new spacers 4 hpi (Figure 1B bottom right cluster). The $D^{-}P^{+}$ group of mutants showed more strongly delayed plasmid loss (>5 hpi), and this loss was preceded or directly accompanied by spacer acquisition (Figure 1B top right cluster). Therefore, these mutants could not be cleared from the cells by direct interference initially, but after primed spacer acquisition the plasmid was rapidly lost. No spacer incorporation was observed for $D^{-}P^{-}$ targets and these variants did not show any plasmid loss within 48 hpi, similar to a non-target plasmid (Figure 1B top left cluster). This group exemplifies that no naïve acquisition had occurred within 48 h in our experimental setup and that all spacer integration events observed in P^{+} groups were due to priming. To validate that spacer acquisition occurred by priming, we sequenced the newly incorporated spacers for a representative set of clones, especially including mutants with late acquisition. We did indeed observe the 9:1 strand bias of new spacers that is typical for priming (Datsenko et al., 2012; Savitskaya et al., 2013; Swarts et al., 2012). Taken together, we found that priming is facilitated by slow or delayed direct interference ($D^{+}P^{+}$), but that it does not strictly require direct interference as exemplified by the $D^{-}P^{+}$ group.

Moderate direct interference activity facilitates the priming process

To verify that rapid plasmid loss indeed results from direct interference, we performed plasmid transformation assays of the target plasmid set into *E. coli* KD263 and compared the transformation efficiency to a co-transformed control plasmid (Almendros and Mojica, 2015). While the *bona fide* target plasmid exhibited a relative transformation efficiency that was 512x lower than the control plasmid (1/512), also mutants with up to two effective mutations gave rise to strongly decreased transformation efficiencies (1/16 to 1/512) (Figure 1C). This means that these target variants still triggered an efficient direct interference response. Triple mutants showed a range of relative transformation efficiencies from full direct interference (*i.e.* 1/512) to no direct interference (~1), suggesting a dominant role for the position of the mutations in the protospacer. Mutants with 4 or 5 effective mutations transformed as efficient as the reference plasmid and displayed no direct interference. When we mapped the classification of all the mutants onto the relative transformation efficiency data, the same trend was observed that target variants with the highest direct interference showed no priming. Instead, intermediate levels of direct interference lead to rapid spacer acquisition, while low levels or the absence of direct interference lead to delayed spacer acquisition. This also confirms that late plasmid loss in the $D^{-}P^{+}$ group is indeed not caused by direct interference with the original spacer, but by primed spacer acquisition followed by direct interference.

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

The average number of effective mutations in a protospacer increases gradually over the groups D^+P^-, D^+P^+, D^-P^+, and D^-P^- (Figure S1). While D^+P^- and D^+P^+ had either 2 or 3 effective mutations, the D^-P^+ mutants had 3 or 4 mutations and the D^-P^- mutants carried 3 or 5 effective mutations in the protospacer. In order to quantify how significant the shifts in the average number of mutations are, we used empirical bootstrapping to test against the hypothesis that the classification does not depend on the number of mutations. Our analysis showed that the D^+P^- and D^+P^+ groups have significantly fewer mutations than would be expected if the classification did not correlate with the number of mutations (>95% and >68% confidence respectively), while D^-P^- has significantly more mutations (>95% confidence) (Figure S3A). We next looked in detail at the number of mutations in each segment, and the position of mutations in each five-nucleotide segment. As has been observed for the seed sequence (Semenova et al., 2011; Wiedenheft et al., 2011b), this showed a significantly lower than average number of mutations in segment 1 for D^+P^- and D^+P^+ groups (both 95% confidence, Figure S3B). Surprisingly, the analysis also revealed that groups showing direct interference (D^+P^-, D^+P^+) had no mutations at the third position of each segment (significantly lower than expected, 95% confidence), whereas D^-P^+ and D^-P^- groups were enriched for mutations at this position (>68% and >95% confidence respectively, Figure S3C). This observation therefore suggests that pairing of the middle nucleotide of the segment is somehow important for direct interference. The third nucleotide of each segment could represent a tipping point in the directional pairing of the crRNA to the DNA. This may occur during canonical, PAM-dependent target DNA binding, which leads to R-loop locking, efficient Cas3 recruitment and target DNA degradation (Blosser et al., 2015; Huo et al., 2014; Rutkauskas et al., 2015).

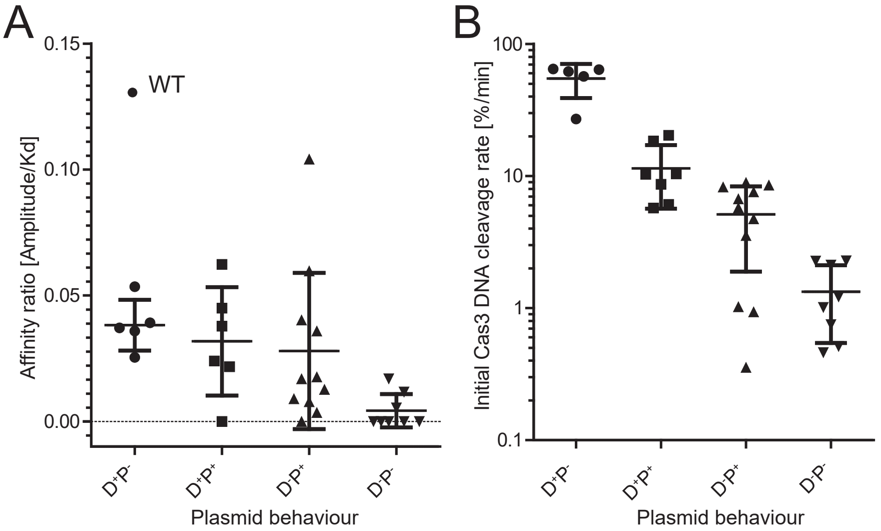


Figure 2: EMSA and Cas3 activity assay. A) Electrophoretic mobility shift assay (EMSA) of the mutant plasmid set. The affinity ratio (Amplitude/K$_d$) is plotted for each mutant (see Table S3 for more details). Mutants are separated by the previously made plasmid behavior classification. The mean and standard deviation for each group are indicated. The *bona fide* target is abbreviated as WT. B) Cas3 DNA degradation activity assay of mutant plasmid set. The initial Cas3 DNA cleavage rate [%/min] is plotted for each mutant. Mutants are classified according to previously identified plasmid behavior. The mean and standard deviation for each group are indicated. Individual gels for all activity assays can be found in Figure S4.

Cascade-plasmid binding is required for interference and priming

To determine the biochemical basis of priming, we first asked what determines if a mutant target can prime or not, and we hypothesized that the affinity of Cascade for a target plasmid would determine its fate. To test this, we performed plasmid based mobility shift assays with purified Cascade complexes (Künne et al., 2015). While the *bona fide* target and most of the mutant targets were bound to completion at increasing Cascade concentrations, some mutant target plasmids were only partially bound (Table S3), as has been observed before (Hochstrasser et al., 2014). By calculating an affinity ratio (Amplitude/K$_d$) and using it as an index for the binding strength, we were able to directly compare the binding properties of all target mutants (Figure 2A). The results show that the *bona fide* target plasmid had the highest affinity ratio (0.31 nM^{-1}), while the mutants cover a range of ratios ranging from very weak binding (>0.008 nM^{-1}) to almost the same levels as the *bona fide* target (<0.1 nM^{-1}). D$^-$P$^-$ mutants all cluster together with low ratios (<0.02 nM^{-1}), and 5 out of 8 show no measurable Cascade binding. This suggests that a minimal level of target plasmid binding by Cascade is required for both direct interference and priming. However, the affinity ratio alone does not predict direct interference and/or priming behavior of a target plasmid.

Cas3 DNA cleavage activity determines plasmid fate

Next, we analyzed if the catalytic rate of target DNA degradation by Cas3 would be related to direct interference and priming. Target DNA degradation is required for direct interference and might be required for priming as well, since all *cas* genes are required for priming in *E. coli* (Datsenko et al., 2012). To test this, we performed Cas3 activity assays with the same panel of target plasmids (Figure 2B, Figure S4). This showed that there is a strong dependence between plasmid fate and Cas3 activity. Mutants capable of only direct interference (D^+P^-) display 5 to 10 times higher activity than priming mutant classes (D^+P^+, D^-P^+), while stable mutants (D^-P^-) show the lowest Cas3 activity. Furthermore, D^+P^+ mutants show a higher average activity than D^-P^+ mutants, although there is overlap between the two groups. The difference between the Cascade affinity and the Cas3 activity plots shows that Cas3 activity is not a simple reflection of Cascade affinity, but is likely influenced by other factors such as conformational differences or the dynamics of Cascade binding. Taken together, there is a link between the Cas3 activity on a target, and target plasmid fate. Direct interference requires the highest Cas3 activity, while priming requires a level of target degradation and occurs at a broad range of intermediate or low Cas3 activities. Finally, it is striking that higher Cas3 activities seem to result in faster priming (D^+P^+ vs D^-P^+), while very high Cas3 activities (D^+P^-) do not lead to priming.

Cas3 produces degradation fragments of near-spacer length

After establishing a connection between plasmid degradation (direct interference) and primed spacer acquisition, we sought to analyze whether the degradation fragments created by Cas3 could serve as spacer precursors. To this end, we performed Cascade-mediated plasmid degradation assays with Cas3 and plasmids containing the *bona fide* target or M4 target. Agarose gel electrophoresis showed that both target plasmids were degraded into similar sized products smaller than 300 nt. Further biochemical analysis of the products revealed that the products were of double stranded nature and contained phosphates at their 5' end (Figure S5A, B). Based on the unidirectional unwinding and single stranded DNA cleavage mechanism of Cas3 (Gong et al., 2014; Huo et al., 2014; Mulepati and Bailey, 2013; Sinkunas et al., 2013; Westra et al., 2012c), we had expected to find single stranded DNA. However, it appeared that complementary fragments had re-annealed to form duplexes, most likely generating annealed products with both 3' and 5' overhangs.

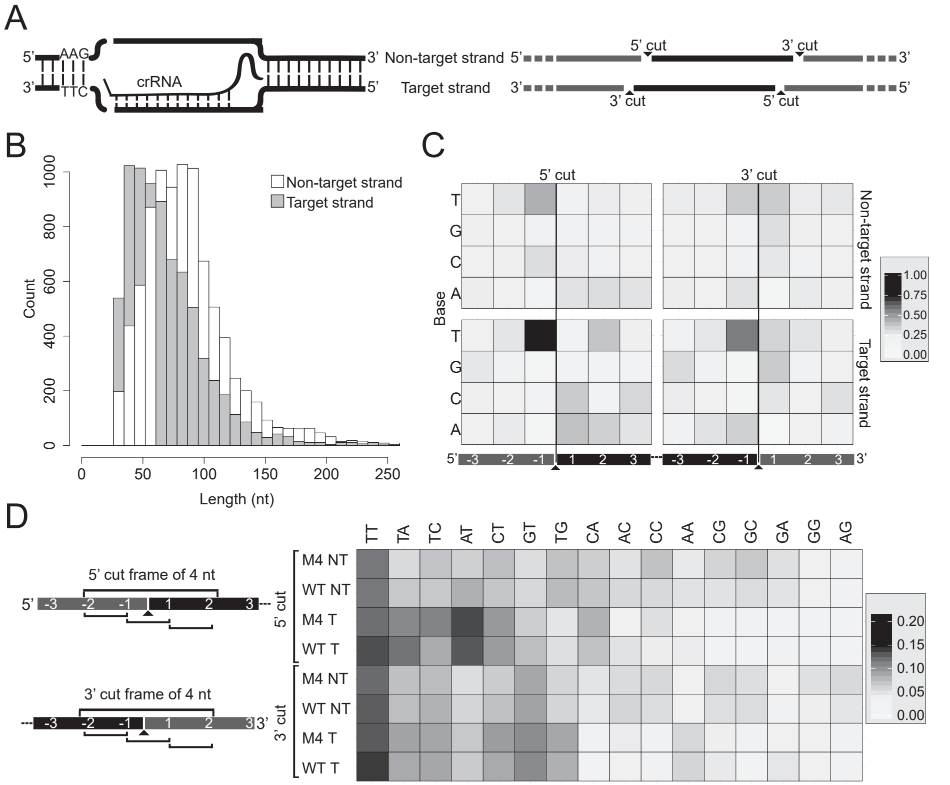


Figure 3: Next generation sequencing analysis of Cas3 DNA degradation products. A) Left: Schematic of R-loop formed by binding of Cascade to dsDNA target. Right: Schematic showing the four distinct Cas3 cleavage sites in dsDNA target. B) Length distribution of Cas3 DNA degradation fragments of M4 target. C) Heat map of nucleotide frequencies around cleavage sites. The cleavage site is between position -1 and 1. Positions indicated in black are on the fragments, positions indicated in grey are outside of fragments. D) Heat map of dinucleotide frequencies around cleavage sites. Abundance of dinucleotides was measured in a shifting frame within 4 nucleotides around the cleavage sites.

In order to determine the exact cleavage patterns of target plasmids by Cas3, we isolated DNA cleavage products from gel and sequenced them using the Illumina MiSeq platform. Analysis of the length of the DNA degradation products from the *bona fide* and M4 target revealed that the majority of fragments from the target strand had a size of around 30-70 nt (Figure 3B, Figure S6A). The non-target strand displayed a shifted distribution with most fragments being 60-100 nt long. Instead of cleaving the target DNA randomly, Cas3 produces fragments with a distinct length profile. Furthermore, the length of the main fraction, especially in the target strand, is close to the length of a spacer molecule (*i.e.* 32/33 nucleotides), supporting the idea that these fragments might be used as spacer precursor molecules.

Cas3 cleavage is sequence specific for thymine stretches

In order to see if Cas3 cleaves the target DNA in a sequence specific manner, we analyzed the region encompassing the cleavage site. This revealed a preference for Cas3 to cleave in thymine-rich sequences for both the *bona fide* and the M4 target, preferably cleaving 3’ of a T nucleotide (Figure 3C,D and Figure S6B). The same pattern was also observed for single stranded m13mp8 DNA cleaved in the absence of Cascade, indicating that T-dependent cleavage specificity is an inherent feature of the HD domain of Cas3. The cleavage specificity of Cas3 leaves one or multiple T nucleotides on the 3’ ends of DNA degradation products. This enriches the 3’ ends of the fragments for NTT sequences, including the PAM sequence CTT. A considerable proportion of degradation fragments therefore satisfies the requirement of Cas1-2 for having CTT sequences in the 3’ ends of spacer precursors in order for these to be correctly integrated into the CRISPR array (Shipman et al., 2016; Wang et al., 2015). Interestingly, C/T-associated cleavage has previously been shown for *Streptococcus thermophilus* Cas3 cleaving oligo nucleotides (Sinkunas et al., 2013), suggesting that this cleavage specificity may be common for HD-domains of Cas3 proteins.

Cas1-2 integrate Cas3-derived degradation fragments

To find out if Cas3 degradation products can indeed serve as spacer precursors, we reconstituted spacer integration *in vitro* using purified Cas proteins. Two types of spacer integration assays were performed (Figure 4A): the first assay used all Cas proteins simultaneously (Cascade, Cas3, Cas1-2) to degrade a target plasmid and integrate the resulting fragments into a plasmid carrying a leader and single CRISPR repeat (pCRISPR). The second assay used DNA degradation products from a separate Cascade-Cas3 reaction. These products were incubated with Cas1-2 and pCRISPR, as described (Nunez et al., 2015b). We noticed a pronounced Cas1-2-dependent shift of the degradation fragments in the gel, suggesting the fragments are bound by Cas1-2 (Figure 4B, left panel). Interestingly, when Cas1-2 was present in the reaction we observed twice as much nicking of plasmid pCRISPR, suggesting half site integration of DNA fragments into pCRISPR had occurred (Figure 4B, right panel) (Nunez et al., 2015b). The same pCRISPR nicking activity was observed using purified Cas3 degradation products (integration assay 2) indicating the integration reaction was not dependent on Cascade or Cas3.

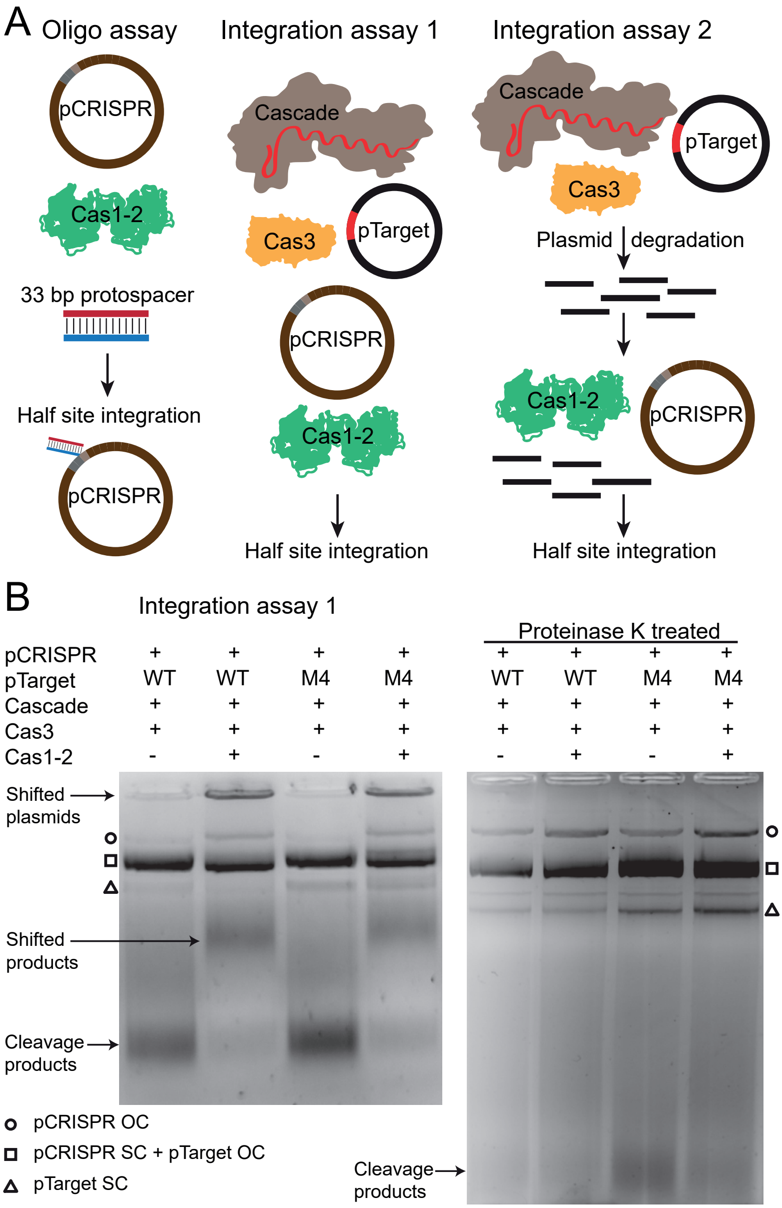


Figure 4: *In vitro* spacer acquisition assays. A) Illustration of the three types of assays performed. In the oligo assay, pCRISPR is incubated with Cas1-2 and a spacer oligo (BG7415/6), leading to half site integration. In assay 1, pTarget and pCRISPR are incubated with Cascade, Cas3 and Cas1-2 for simultaneous degradation of pTarget and half site integration into pCRISPR. In assay 2, pTarget is incubated with Cascade and Cas3 and the resulting DNA degradation products are then separately incubated with pCRISPR and Cas1-2. B) Gel electrophoresis of integration assay 1. The *bona fide* target is abbreviated as WT. Left gel, untreated; right gel, Proteinase K treated. Cas1-2 presence causes upwards shift of DNA. Original plasmids are supercoiled (SC), half site integration causes nicking of pCRISPR, resulting in the open circular conformation (OC).

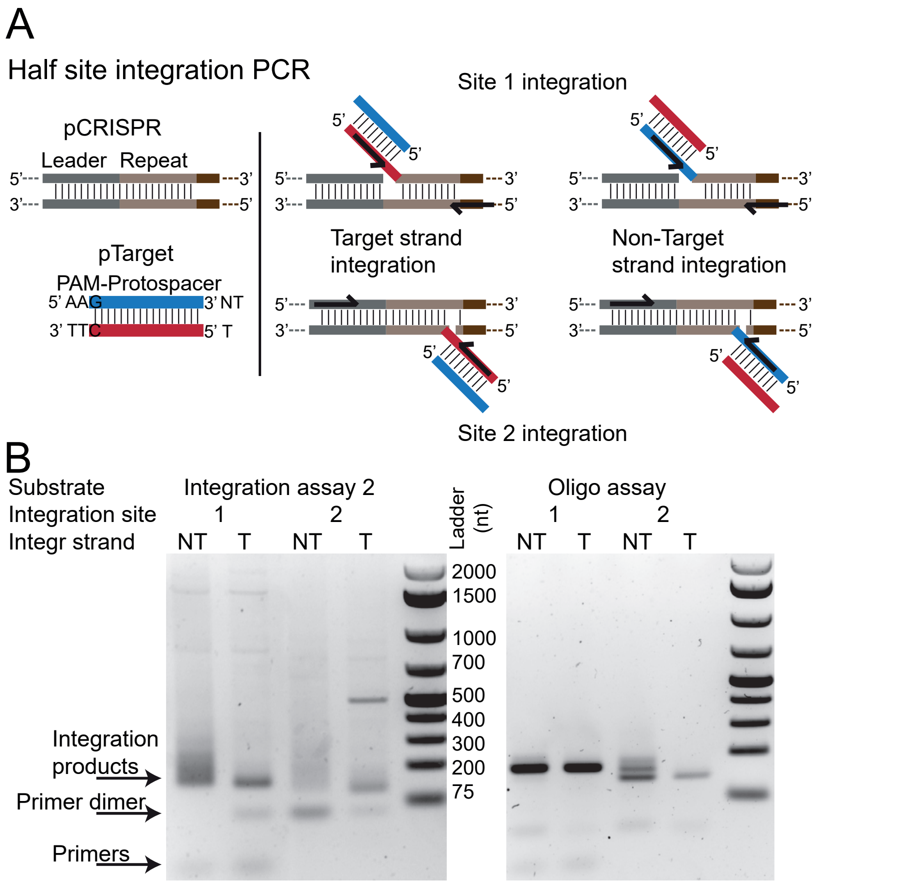


Figure 5: Half site integration PCR. A) Illustration of the half site integration PCR. Primer sets are chosen to show integration into site 1 (leader-proximal repeat end) and site 2 (leader distal repeat end), and to see both possible orientations of the integrated spacer. Primer sequences were chosen based on frequently incorporated spacers (hotspots) *in vivo* (Fineran et al., 2014). B) Gel electrophoresis of half site integration PCR based on integration assay 2 (left) and oligo assay (right). PCR products representing integrations are indicated with an arrow. PCR products were specific to reactions containing all components. Lower running PCR products are primer dimers (verified by sequencing).

To verify that spacer half-site integration had taken place and not just pCRISPR nicking, we gel-isolated the nicked pCRISPR band for PCR analysis. Since we did not know the sequence of the integrated fragments, we selected three primer pairs that would amplify frequently incorporated spacers from the plasmid *in vivo* (Fineran et al., 2014). Two of the three tested primers gave a PCR product of the expected size and we chose one of the primers for more detailed analysis. It has previously been shown that the first half-site integration may occur at the boundary of the leader and repeat in the sense strand (*i.e.* site 1), or at the penultimate base of the repeat in the antisense strand (*i.e.* site 2) (Nunez et al., 2015b; Rollie et al., 2015). Furthermore, fragments can be integrated in two different orientations. We performed PCR amplification reactions to test for all four different situations (Figure

5A). This showed that integration of Cas3-derived degradation products occurs sequence specifically at both site 1 and site 2, and in both orientations (Figure 5B).

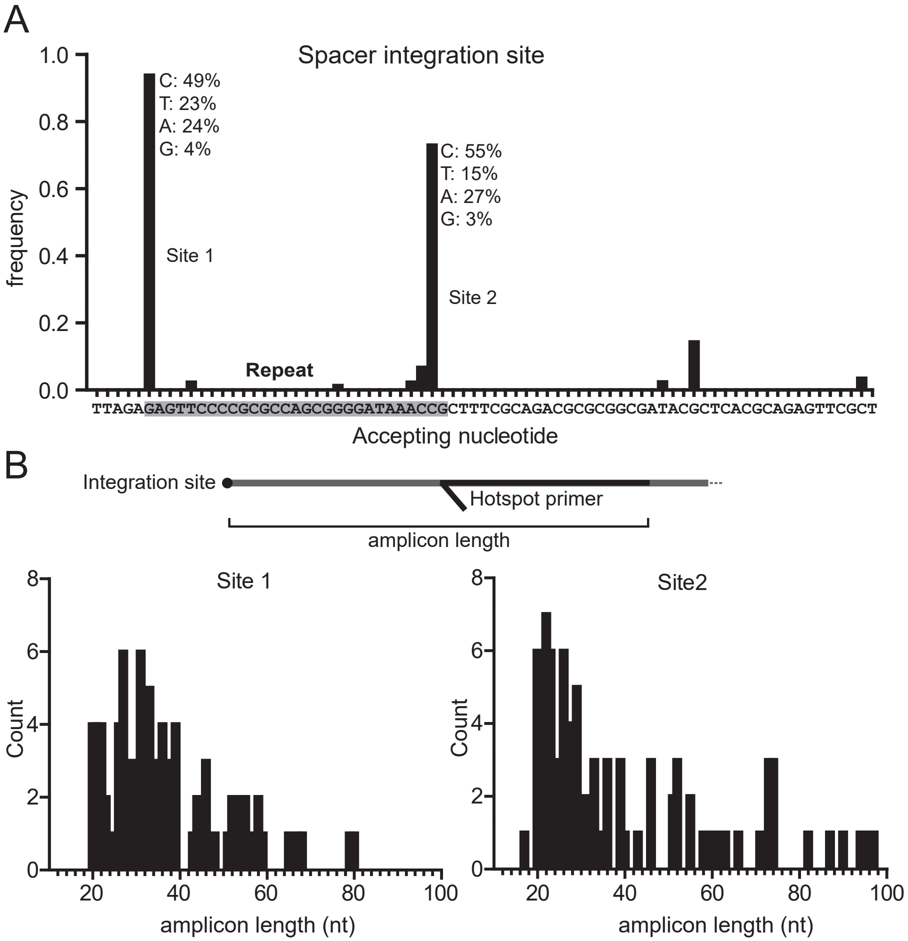


Figure 6: Sequencing analysis of spacer integration. A) Frequencies of exact integration locations for integration at site 1 (grey bars) and site 2 (black bars) as determined by sequencing. X-axis gives the backbone nucleotide to which the spacer is coupled. Frequencies of coupled spacer nucleotides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

Integration of fragments in the repeat is nucleotide and position specific

In order to obtain more insight into the accuracy of integration, we sequenced 48 clones for each of the four primer sets. The results confirm that fragments from the target and non-target strands are integrated at both site 1 and site 2 of the repeat. Integration is very specific to the correct positions in the repeat. At site 1, 94% of the integrated fragments were coupled correctly to the first nucleotide of the sense strand of the repeat, while at site 2, 73% of integrated fragments were coupled correctly to the penultimate nucleotide of the antisense strand of the repeat, replacing the last nucleotide of the repeat in the process (Figure 6A). In line with previous findings (Nunez et al., 2015b; Rollie et al., 2015), both integration sites show a preference for coupling incoming C nucleotides; 49% and 55% for site 1 and site 2 respectively (Figure 6A). Considering that Cas3 DNA degradation fragments have T nucleotides on their 3' ends, this suggests that precursors have been pre-processed by Cas1-2 before integration, as has been demonstrated for artificial substrates (Wang et al., 2015). The majority of the integration amplicons had a length of only 20 to 40 nucleotides (Figure 6B), indicating that the integration reaction prefers short to long substrates. Altogether, we show that the integration of PAM-containing spacers in the repeat during priming is enhanced by the combined sequence specificities of two Cas enzymes: (1) Cas3 which leaves thymines in the 3'-end of DNA fragments, enriching the fragment ends for CTT, and (2) Cas1-2 which prefer CTT carrying substrates and process and couple the 3' cytosine specifically to both integration sites of the repeat.

Discussion

A remaining gap in our understanding of Type I CRISPR-Cas mechanisms is how new spacers are selected and processed before being incorporated into the CRISPR array. In this work we demonstrate that Cas3 produces spacer precursors for primed adaptation of the CRISPR array. These spacer precursors are 30-100 nt long partially double stranded DNA molecules formed by fragmentation of the target DNA. Cas3 DNA degradation fragments fulfill all criteria for spacer precursors that can be deduced from recent studies of the Cas1-2 complex (Figure 7). Ideal spacer precursors in *E. coli* are partially double stranded duplexes of at least 35 nucleotides containing splayed single stranded 3' ends with a CTT PAM sequence on one of the 3' overhangs (Nunez et al., 2015a; Rollie et al., 2015; Shipman et al., 2016; Wang et al., 2015). We have shown that Cas3 DNA degradation products are mainly double stranded *in vitro*. This is most likely due to re-annealing of the single stranded products that are produced by the nuclease-helicase activity of Cas3. It is possible that *in vivo* other proteins are involved in the formation of duplexes after degradation. In fact, it has been shown that Cas1 from *Sulfolobus solfataricus* can facilitate the annealing of oligonucleotides (Han and Krauss, 2009). These re-annealed duplexes likely contain a mix of 3' and 5' overhangs, because the two DNA strands of the target are degraded independently. This also results in slightly shorter fragments for the target strand. Despite these differences in fragment size, both strands are cleaved by Cas3 with the same specificity, enriching the 3' ends of the fragments for stretches of thymines. Contrary to the CTT requirements for spacer integration, it is known that Cascade tolerates five different PAM sequences (*i.e.* CTT, CTA, CCT, CTC, CAT) for direct interference (Fineran et al., 2014; Leenay et al., 2016). However, the vast majority of new spacers (97%) resulting from primed acquisition carry CTT PAM sequences (Shmakov et al., 2014). This further supports the idea that spacer precursors with CTT-ends are selected non-randomly by the Cas1-2 complex from pools of Cas3 breakdown fragments and further trimmed to a 3’ C (Wang et al., 2015). These are then coupled to the repeat by nucleophilic attack of the 3’-OH (Nunez et al., 2014; Rollie et al., 2015). The T-dependent target DNA cleavage specificity of Cas3 further enhances the production of precursors that fit the requirements of new spacers by creating a pool of DNA fragments with the correct size and correct 3’ ends. The interference phase of CRISPR immunity is therefore effectively coupled to the adaptation phase, providing positive feedback about the presence of an invader.

It was previously reported that a dinucleotide motif (AA) at the 3’ end of a spacer increases the efficiency of naïve spacer acquisition (Yosef et al., 2013). We did not observe this motif at the expected distance from the end in the Cas3 DNA degradation fragments, suggesting that Cas3 does not take the AA motif into account when generating spacer precursors.

We found that the integration reaction is very precise for the two correct integration sites in the repeat (site 1 and site 2), and we observed that the integrated fragments most often were the result of a 3’ cytosine coupling reaction. *In vivo*, however, only the integration of a CTT-containing fragment at site 2 would lead to a functional spacer targeting a protospacer with PAM (Figure 7), while half site integrations initiating at site 1 would result in ‘flipped’ spacers (Shmakov et al., 2014). Using a selective PCR strategy, we detected primed spacer acquisition events at both integration sites, and we identified that DNA fragments from both the target and non-target strand of the plasmid could be used for integration. In Type I-E CRISPR-Cas systems, primed spacer acquisitions display a typical 9:1 strand bias for the acquisition of spacers targeting the same strand of DNA as the spacer causing priming (Datsenko et al., 2012; Swarts et al., 2012). This suggests that *in vivo*, other factors might be involved in further increasing the accuracy of functional spacer integration. This includes the formation of supercomplexes between various Cas proteins (*i.e.* Cascade, Cas3, Cas1-2) (Plagens et al., 2012; Redding et al., 2015; Richter et al., 2014), and the involvement of non-Cas host proteins such as PriA, RecG and IHF (Ivancic-Bace et al., 2015; Nunez et al., 2016). IHF ensures that the first integration event takes place at the leader-proximal end of the repeat (site 1) and would be involved in ensuring that the PAM cytosine gets integrated at the leader-distal end (site 2). Supercomplex formation during precursor generation may lead to the selection of fragments from the target strand containing a CTT PAM at the 3’ end. Although the length of the observed integration amplicons is centered around 20-40 nt, we also find amplicons of up to 100 nt. *In vivo*, *E. coli* integrates fragments of 33 nt length. We speculate that trimming of the precursor to 33 nt length occurs after half-site integration and before formation of the stable integration intermediate (Figure 7). Despite the mechanisms that lower erroneous integration of new spacers, it is likely that natural selection of functional spacers *in vivo* also plays a role in the spacers that end up being part of the first population of bacteria following a priming event.

It was surprising that that the *bona fide* target and several D^+P^- mutants did not show priming despite providing Cas3 degradation products. Furthermore, the degradation fragments of the *bona fide* target were very similar to the fragments of the M4 target (D^+P^+), which cannot explain the difference in priming behavior. We propose that these targets are degraded and cured from the cell too rapidly, giving the acquisition machinery insufficient time to generate new spacers. However, a low level of spacer integration might be taking place at undetectable levels even for the *bona fide* target, as has been observed previously (Swarts et al., 2012; Xue et al., 2015). In this case, cells with additional spacers do not have a selective growth advantage over cells without new spacers as the plasmid is already effectively cleared from cells without new spacers. Mutant targets with intermediate levels of direct interference however, are replicated and subject to interference over a longer time period, thereby providing more precursors, more time for spacer acquisition to occur, and therefore a greater selective growth advantage. Low levels of direct interference lead to a slow priming response due to the scarcity of spacer precursor molecules. While this paper was under review, another study showed that perfectly matching protospacers with canonical PAMs can indeed stimulate priming and that plasmid targeting is the stimulating factor (Semenova et al., 2016). In line with our findings, the authors further propose that priming is usually not observed with fully matching protospacers because these targets are degraded too rapidly.

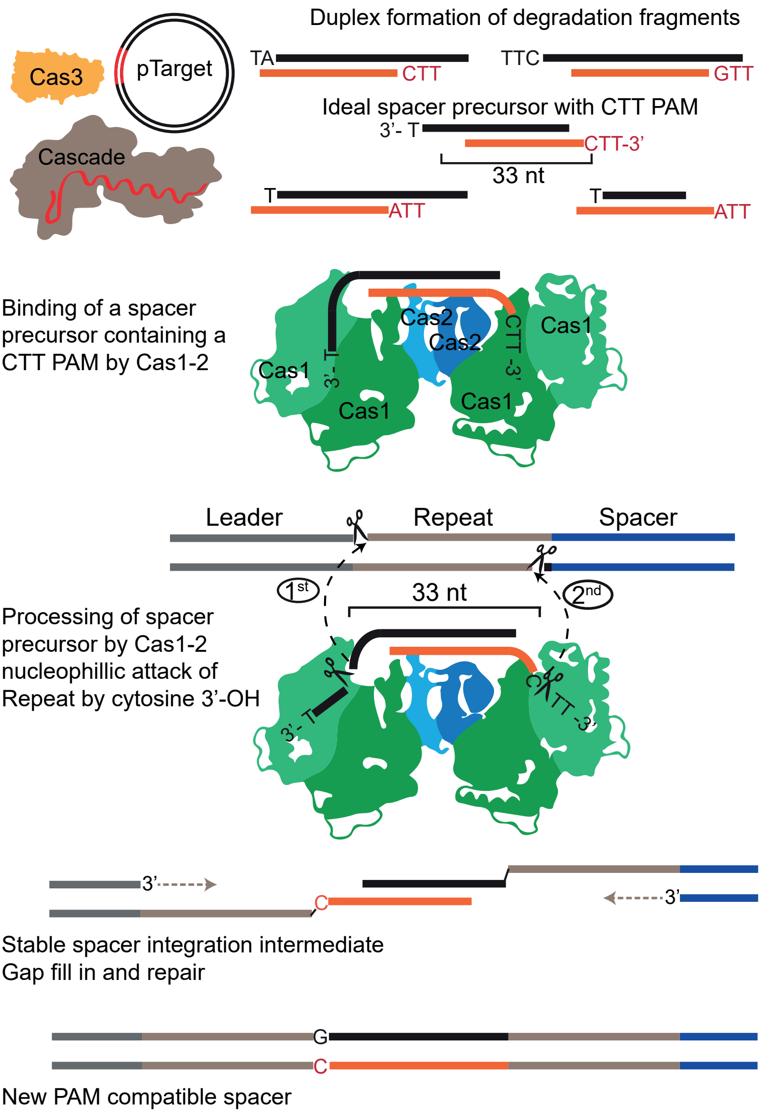

Duplex formation of degradation fragments
Cas3
pTarget
Cascade
TA
CTT
TTC
GTT
Ideal spacer precursor with CTT PAM
3'- T
CTT-3'
33 nt
T
ATT
T
ATT
Binding of a spacer precursor containing a CTT PAM by Cas1-2
Cas2
Cas2
Cas1
Cas1
Cas1
Cas1
3'- T
CTT -3'
Leader
Repeat
Spacer
1st
2nd
33 nt
Processing of spacer precursor by Cas1-2 nucleophillic attack of Repeat by cytosine 3'-OH
3'- T
CTT -3'
3'
C
3'
Stable spacer integration intermediate
Gap fill in and repair
G
C
New PAM compatible spacer

Figure 7: Model of primed spacer acquisition. Cleavage of a targeted plasmid during direct interference by Cascade and Cas3. Cleavage products are near-spacer length and reanneal to form duplexes with 5' and/or 3' overhangs. The fragments are enriched for NTT sequences on their 3' ends. A fraction of the duplexes fulfils spacer precursor requirements: 3' overhangs, CTT at one 3' end and a 33 nt distance between the C and the opposite 3' overhang. Cas1-2 binds spacer precursors with a preference for ideal duplexes as described above (Nunez et al., 2015a; Wang et al., 2015). The precursor is processed by Cas1-2 to a length of 33 nt with 3' cytosine. In parallel to processing, 3' ends of the precursor perform a Cas1-2 catalyzed nucleophilic attack on the two integration sites of the repeat (Nunez et al., 2015b; Rollie et al., 2015). Integration at the leader-repeat junction occurs first (Nunez et al., 2016), subsequently the PAM derived 3' cytosine is integrated to assure correct orientation and production of a functional spacer. A Stable spacer integration intermediate is formed (Arslan et al., 2014). The gaps are filled in and repaired by the endogenous DNA repair systems, including DNA polymerase I (Ivancic-Bace et al., 2015).

Cut-paste spacer acquisition

We have shown that priming reuses target DNA breakdown products as precursors for new spacers, providing support for a cut and paste mechanism of spacer selection (Wang et al., 2015). Compatible models have recently been proposed for naïve spacer acquisition (Levy et al., 2015). It was shown that CRISPR adaptation is linked to double stranded DNA breaks that form at stalled DNA replication forks. Invading genetic elements often go through a phase of active DNA replication when they enter a host cell, and a replication dependent mechanism therefore helps the host to primarily select spacers from the invading element. The RecBCD complex is key in this process as it repairs double stranded breaks by first chewing back the ends of the DNA creating fragments of tens to thousands of nucleotides (Amitai and Sorek, 2016). These fragments are thought to reanneal and serve as precursors for new spacers. Other studies have shown the direct involvement of crRNA-effector complexes in spacer selection. In the Type I-F CRISPR-Cas system of *Pseudomonas aeruginosa* the Csy complex is required for naïve spacer acquisition (Vorontsova et al., 2015). Also Cas9 in Type II systems has a direct role in spacer acquisition (Heler et al., 2015; Wei et al., 2015b). Both systems incorporate spacers very specifically from canonical PAM sites, suggesting that the Csy complex and Cas9 are directly involved in PAM recognition during spacer sampling.

Mutations in the protospacer

In this study we have focused on the effect of mutations in the protospacer on direct interference and priming, while maintaining the dominant interference permissive PAM CTT. Apart from underscoring the importance of the number of mutations and existence of a seed sequence (Semenova et al., 2011; Künne et al., 2014; Wiedenheft et al., 2011b; Xue et al., 2015), we uncover that for direct interference pairing of the middle nucleotide in each 5-nucleotide segment of the protospacer is disproportionately important, and may represent a tipping point in the binding of a target. None of the 30 mutants showing direct interference carried mutations at these middle positions. Also in a previously obtained list of approximately 3,300 triple mutants showing direct interference (Fineran et al., 2014), mutations at this position were underrepresented (Figure S3D). This suggests that pairing at the middle position of each segment may be important for continuation of the directional zipping process. This process starts at the PAM and leads to the formation of a canonical locked R-loop, which is required for Cas3 recruitment and target DNA degradation (Blosser et al., 2015; Redding et al., 2015; Rutkauskas et al., 2015; Sashital et al., 2012; Semenova et al., 2011; Szczelkun et al., 2014). We stress that we have used variants with CTT PAMs only, which can be engaged by Cascade in the canonical PAM-dependent binding mode (Blosser et al., 2015; Hayes et al., 2016; Redding et al., 2015; Rutkauskas et al., 2015; Sashital et al., 2012; Semenova et al., 2011), and can also trigger priming. It has become clear, however, that targets with mutations in the PAM display a broad spectrum of distinct characteristics depending on the chosen PAM, including a range of efficiencies of direct interference (Westra et al., 2013) and the reluctance to trigger efficient Cas3 target DNA degradation (Blosser et al., 2015; Hochstrasser et al., 2014; Mulepati and Bailey, 2013; Redding et al., 2015; Rutkauskas et al., 2015; Xue et al., 2015). In many cases these PAMs still support the priming process (Datsenko et al., 2012; Fineran et al., 2014; Xue et al., 2015). Targets with highly disfavored PAMs (Hayes et al., 2016) are likely engaged in the non-canonical PAM-independent binding mode (Blosser et al., 2015) and may require recruitment and translocation events of Cas1-2 and Cas3 proteins to initiate the target degradation needed to acquire new spacers.

Conclusion

The findings presented here, showcase the intricate PAM-interplay of all Cas proteins in type I systems to update the CRISPR memory when receiving positive feedback about the presence of an invader. The robustness of priming is achieved by three components that co-evolved to work with PAM sequences: Cas3 producing spacer precursors enriched for correct PAM ends, Cas1-2 selecting PAM-compliant spacer precursors and Cascade efficiently recognizing targets with PAMs. This process stimulates the buildup of multiple spacers against an invader, preventing the formation of escape mutants (Datsenko et al., 2012; Richter et al., 2014; Swarts et al., 2012). When the original spacer triggers sufficiently strong interference, priming acquisition does not frequently occur. This prevents the unnecessary buildup of spacers and keeps the CRISPR array from getting too long. Any subsequent reduction in effectivity of the immune response by further mutations of the invader will in turn allow priming acquisition, restoring immunity.

Materials and Methods

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

Molecular Biology and DNA Sequencing. All oligonucleotides are listed in Table S1. All plasmids are listed in Table S2. All strains and plasmids were confirmed by PCR and sequencing (GATC-Biotech). Plasmids were prepared using GeneJET Plasmid Miniprep Kits (Thermo Scientific). DNA from PCR and agarose gels was purified using the DNA Clean and Concentrator and Gel DNA Recovery Kit (Zymo Research). The library of pGFPuv sp8 mutants was available from a previous study (Fineran et al., 2014). pMAT MBP-Cas3 was a kind gift from Scott Bailey lab (Mulepati and Bailey, 2013).

Transformation assay. Transformation assays were carried out in *E. coli* KD263. Cells were grown to OD600 ~0.4, induced with 0.2% L-arabinose and 0.5 mM IPTG and allowed to grow for 1h. Cells were then made chemically competent for heat shock transformation using the $RuCl_2$ method. Cells were co-transformed with 10 ng target plasmid (pWUR836-868, Kan^R) and 10 ng control plasmid (pWUR835, Amp^R) simultaneously (Almendros and Mojica, 2015). Dilutions of transformants were then plated on LBA plates with Amp and LBA plates with Kan. The transformation efficiency of mutated target plasmids was normalized against the transformation efficiency of the control plasmid.

Plasmid loss assay. *E. coli* KD263 cells were transformed with the target plasmids (pWUR836-868) by heat shock. Individual colonies were picked in triplicate and grown overnight in 5 ml LB supplemented with 2% glucose to repress *cas* gene expression. The next day, cultures were transferred 1:100 into induced medium (0.2% L-Arabinose, 0.5 mM IPTG) and plasmid loss was monitored. Samples were taken every hour until 5h, and then again at 24h and 48h. Dilutions were plated on non-selective plates and plasmid loss was counted based on loss of fluorescence using a Syngene G-box imager. Plasmid-free colonies were screened for spacer integration by colony PCR using DreamTaq Green DNA polymerase (Thermo Scientific). Acquisition of spacers was detected by PCR using primers BG5301 and BG5302. PCR products were visualized on 2% agarose gels and stained with SYBR-safe (Invitrogen). PCR products were sequenced using Sanger sequencing at GATC (Konstantz, Germany) using primer BG5301.

EMSA assays. Purified Cascade complex with spacer8 crRNA was incubated with plasmid at a range of molar ratios (1:1-100:1, Cascade:DNA) in buffer A (20 mM HEPES pH7.5, 75 mM NaCl, 1 mM DTT) for 30 min. Reactions were run on 1% native agarose gels for 18h at 22 mA in 8 mM sodium-borate buffer. Gels were post stained with SYBR Safe (Invitrogen). Shifted (Cascade bound DNA) and unshifted (free DNA) bands were quantified using the GeneTools software (Syngene) and total Cascade concentration (X) was plotted against the fraction of bound DNA (Y). The curves were fitted with the following formula: Y = (amplitude * X)/(K_d + X) (van Erp et al., 2015). The amplitude is the maximum fraction of bound DNA. Since the amplitude is not always 1, we cannot directly compare K_d values, instead the 'affinity ratio' was calculated as: amplitude/K_d (i.e. normalizing the K_d against the variable amplitude).

Cas3 DNA degradation assays. Cas3 DNA degradation activity was routinely tested by incubating 500 nM Cas3 with 4 nM M13mp8 single stranded circular DNA in buffer R (5 mM HEPES, pH8, 60 mM KCl) supplemented with 100 µM Ni^{2+} at 37 °C for 1 h. Plasmid-based assays were performed by incubating 70 nM Cas3 with 70 nM Cascade, 3.5 nM plasmid DNA in buffer R (+ 10 µM $CoCl_2$, 10 mM $MgCl_2$, 2 mM ATP) at 37 °C for 10-60 minutes unless indicated otherwise. For quantifying Cas3 activity, assays were run at normal conditions and samples were taken at 0 min, 1 min, 10 min and 30 min. Samples were immediately quenched with 6x DNA loading dye (Thermo scientific) on ice. Samples were run on agarose gels and supercoiled plasmid bands were quantified using the GeneTools software (Syngene). The DNA degradation was plotted (X: time [min]; Y: Intact Plasmid [%]) and the initial activity of Cas3 [%/min] calculated from the initial slope of the curve.

Protein purification. All proteins were expressed in Bl21-AI cells. Cascade was purified as described earlier (Jore et al., 2011a). MBP-Cas3 was purified as described in (Mulepati and Bailey, 2013). The Cas1-2 complex was purified as follows. The Cas1-2 operon was PCR amplified with primers BG4556/7 and cloned into pET52b (SmaI/SacI) to make pWUR871. The Cas1-2 complex was purified using the N-terminal StrepII tag on Cas1. Briefly, cells were grown to an OD_{600} of 0.4, cooled on ice for 30 minutes and induced with 0.5 mM IPTG and 0.2% l-arabinose. Protein was expressed at 20 °C overnight. Cells were collected by centrifugation and lysed in buffer L (20mM HEPES pH 7.5, 75mM NaCl, 1mM DTT, 5% glycerol, 0.1% Triton X100) using a Stansted pressure cell homogenizer. The lysate was cleared by centrifugation and filtration. The cleared lysate was incubated with Strep-tactin beads (IBA) for 30 minutes at 4 °C and loaded into a gravity column. The column was washed with buffer A (20mM HEPES pH 7.5, 300mM NaCl, 1mM DTT, 5% glycerol) and the proteins eluted in buffer B (20mM HEPES pH 7.5, 75mM NaCl, 1mM DTT, 5% glycerol, 2.5 mM biotin). The presence and purity of the Cas1-2 complex was checked via Tris-tricing SDS PAGE (10-20%). The final complex was snap frozen in liquid nitrogen and stored at -80 °C.

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

Statistical testing against the null hypothesis. We used a version of the empirical bootstrap method (Dekking, 2005) to test our data against the null hypothesis that observed behaviors (D±P±) do not correlate with a particular sequence property. To establish the confidence with which the null hypothesis can be disregarded, we construct randomized mock behavioral groups by repeatedly (10^5 times, resulting in an accuracy in the significance intervals of about $1/\sqrt{(10^5)} \approx 0.3\%$) drawing a random selection (allowing repetitions) of sequences from the complete set of 31 protospacers (including the *bona fide* spacer). The average property of interest is then calculated for the generated mock behavioral groups, giving histograms showing the distribution over the mock sets. The above procedure is performed for the total number of effective mismatches, and the number of mutations within segment 1, and the number of mutations on position 3 within all segments combined.

***In vitro* acquisition assay**. Two types of assays were performed. 1) Cas3 plasmid DNA degradation assays were carried out as described above, the reaction products were incubated with Cas1-2 and pWUR869 in buffer R for 60 min. 2) Target plasmid, Cascade, Cas3, Cas1-2 and pWUR869 were incubated in buffer R for 60 min. Component concentrations for assay 1 and 2 were as follows: 70 nM Cascade, 70 nM Cas3, 300 nM Cas1-2, 3.5 nM target plasmid, 5 nM pWUR869 (pCRISPR). Reaction products of both assays were run on a 1.8% TAE-agarose gel. To verify half-site integration of spacers in the CRISPR array as described in (Nunez et al., 2015b), nicked pWUR869 was isolated from gel and analyzed by PCR. PCR was performed with forward primer BG5301 (site2) or BG7522 (site1) and reverse primers BG7415/6 (control) or BG6713-15 (3 hotspots) or BG7215/6 (fw/rv of hotspot3). These primers match spacers that are frequently incorporated *in vivo* (Fineran et al., 2014). To verify and analyze integration, PCR products were cloned into a pGEMT-easy vector (Promega) and individual clones were sequenced.

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

NGS Data analysis. Sequencing data was deposited at the European Nucleotide Archive under the accession number PRJEB13999. Samples were de-multiplexed using their barcodes. All pair-end reads were mapped to their originating sequences (pWUR836/853, m13mp8) using BLAST and allowing for up to one mismatch. Reads for which both ends could not be aligned to the reference sequence were discarded. For the cleavage sites, distinct start/end positions were analyzed independently (see Table S4 and Table S5 for details). For the duplets a sliding window around the cut point was used. For the duplets the following positions were considered: (-2,-1), (-1,1) and (1,2). In this notation the cut point is between -1 and 1, positive positions are inside the considered fragment and negative positions are outside. Enrichment analysis was performed using a hypergeometric probability distribution to model the background probability density associated to the originating sequence. R packages stats (R-Development-Core-Team, 2008) and ggplot2 (Wickham, 2009) were used for these computations and to generate corresponding graphics.

Author contributions

T.K. and S.J.J.B. designed research; T.K., S.N.K., J.W.B., A.I.M.V., W.R.M performed research; T.K., M.S.-D., S.N.K, M.K., M.D. and S.J.J.B. analyzed data; and T.K. and S.J.J.B. wrote the paper with input from all authors.

Acknowledgements

This work was financially supported by an LS6 ERC starting grant (639707) and NWO VIDI grant to S.J.J.B. (864.11.005). S.J.J.B and M.D. were supported by a TU Delft start up grant. This work was supported by the Netherlands Organization for Scientific Research (NWO/OCW), as part of the Frontiers in Nanoscience program. We thank Sebastian Hornung for depositing the DNA sequencing dataset.

Supplementary Figures

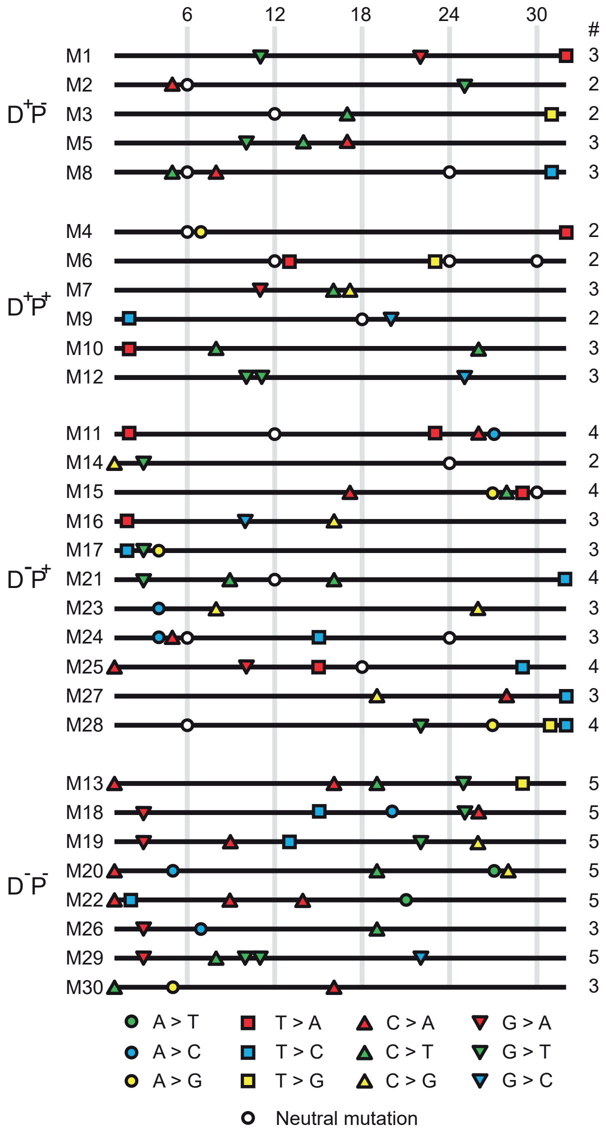
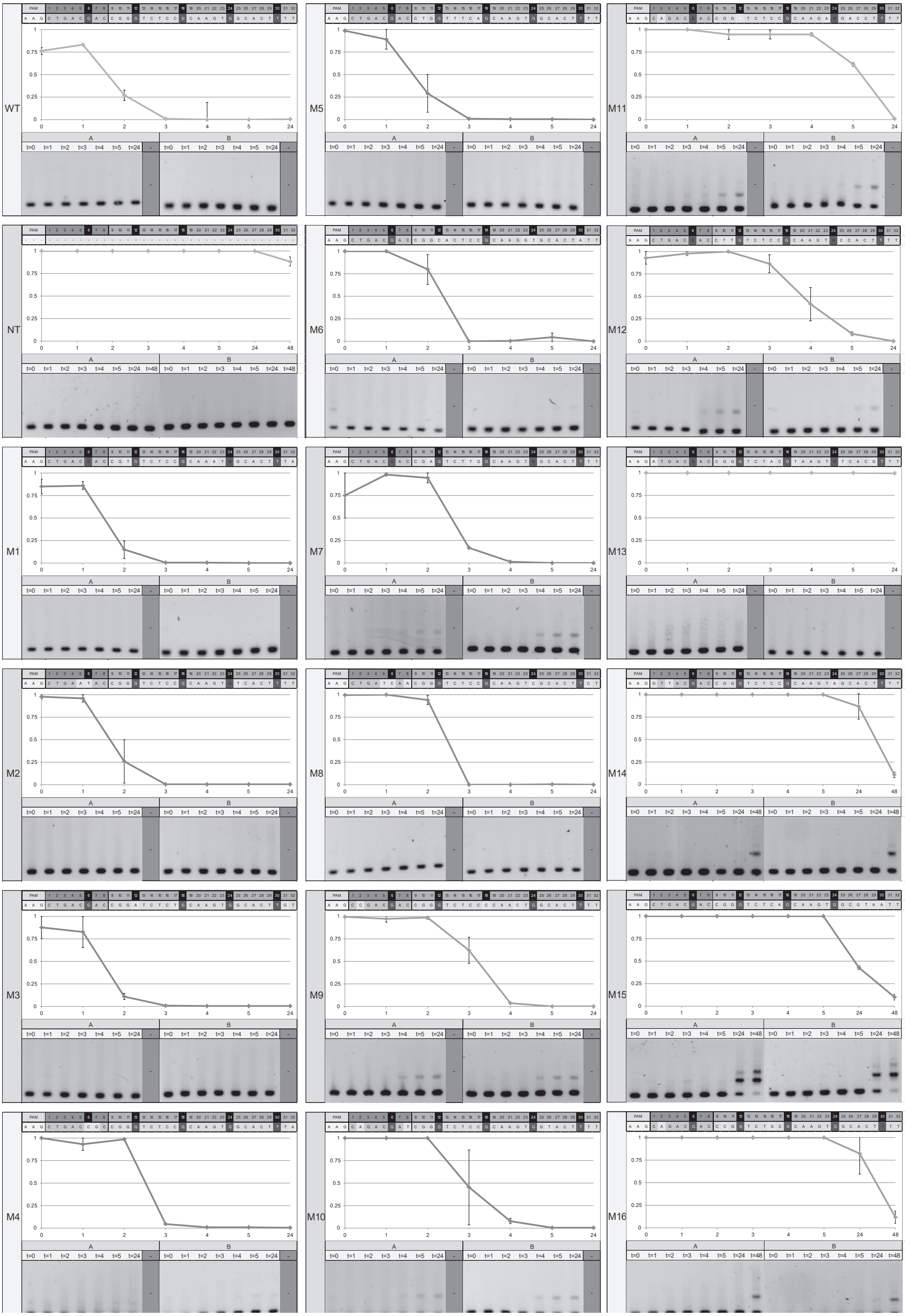


Figure S1. Related to Figure 1: Overview of Protospacer8 mutants. 30 mutants of protospacer8 containing either 3 or 5 total mutations were used throughout the study. Mutations on positions 6, 12, 18, 24, 30 (empty circles) are not participating in base-pairing and are therefore not considered as effective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


WT
M5
M11
NT
M6
M12
M1
M7
M13
M2
M8
M14
M3
M9
M15
M4
M10
M16
PAM
A
B
t=0
t=1
t=2
t=3
t=4
t=5
t=24
t=48
-
0
0.25
0.5
0.75
1
0 1 2 3 4 5 24 48

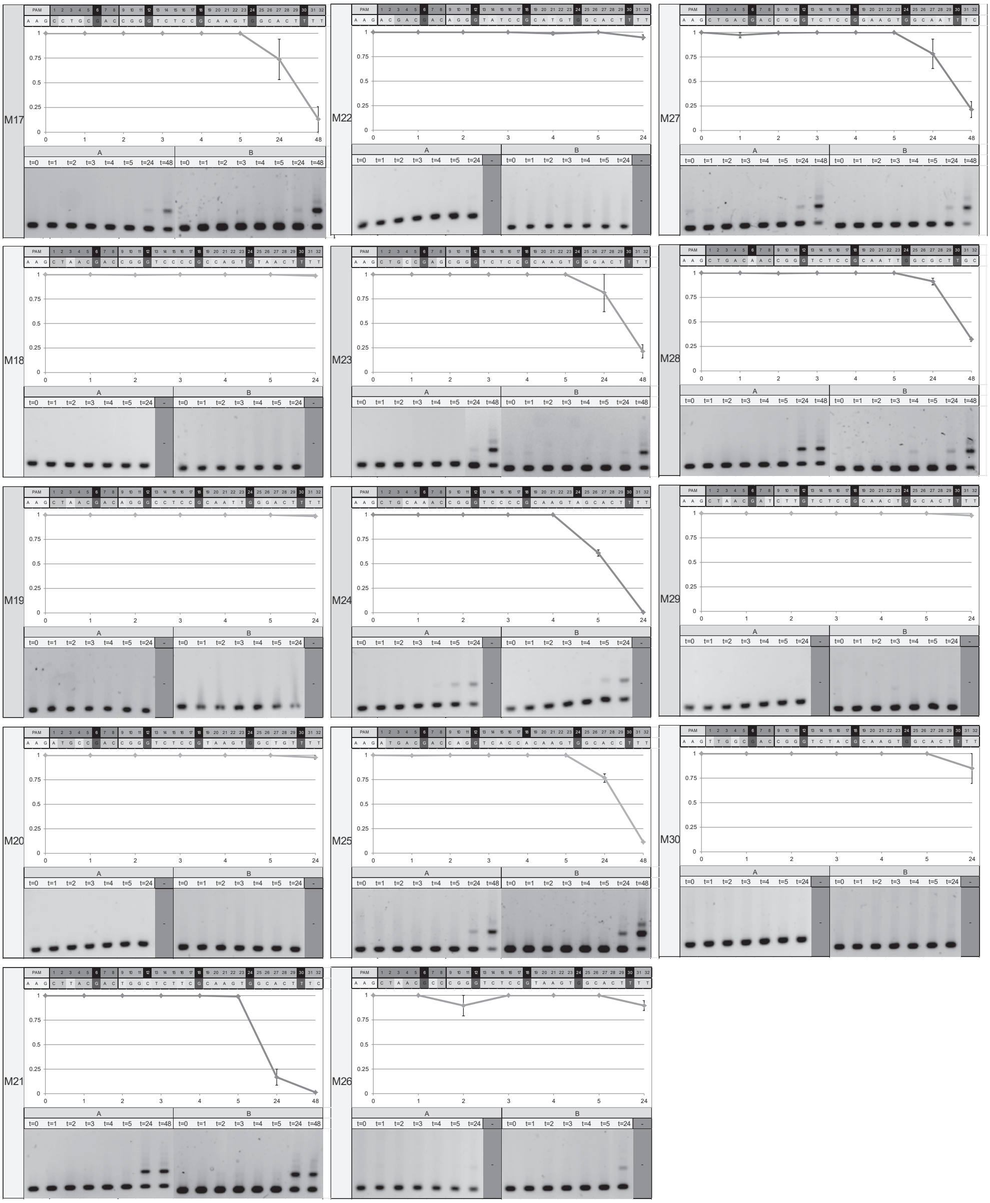


Figure S2. Related to Figure 1: Individual plasmid loss assays. Panels for each plasmid mutant with (top to bottom): Sequence with indicated mutations, plasmid loss curves from 0 h to 24 h or 48 h, duplicate of CRISPR PCR showing spacer acquisition. The bottom bands in the PCR gels represent the unextended array, higher bands represent the array with an extra spacer. Error bars in plasmid loss graphs represent the standard deviation of replicate experiments. The *bona fide* target is abbreviated as WT.

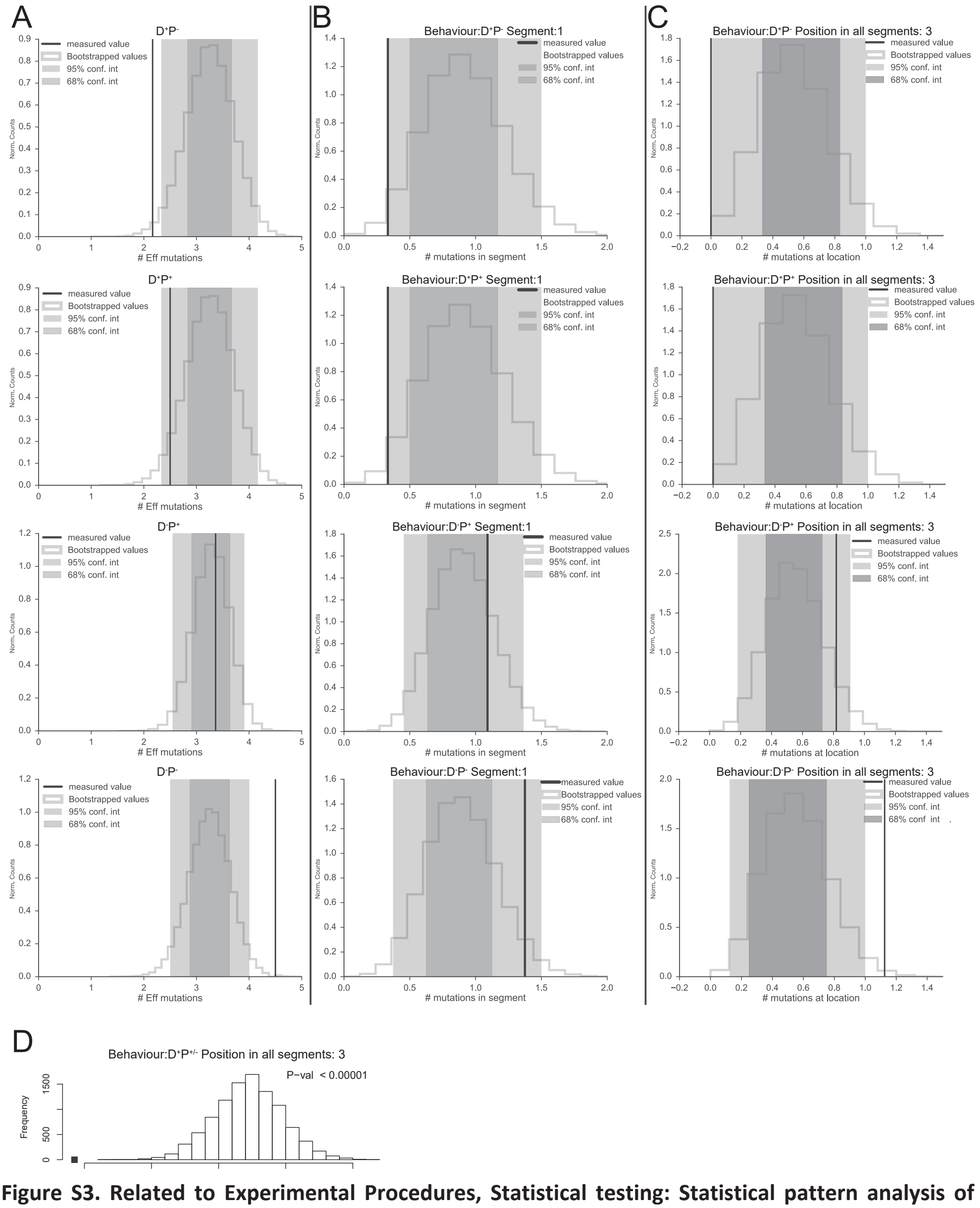


Figure S3. Related to Experimental Procedures, Statistical testing: Statistical pattern analysis of 30 mutants set. Three properties were analyzed separately for each group of plasmid behavior. The average of each behavioral group is indicated by the yellow vertical line. To test if the plasmid behavior depends on a certain property, for each property a distribution was made based on empirical bootstrapping of the whole set of 30 mutants (blue line). The 95% and 68% confidence intervals of each distribution are indicated by the light and dark grey boxes respectively. A) Average number of effective mutations. B) Average number of mutations in segment 1. C) Average number of mutations on position 3 within all segments combined. D) Average number of mutations on position 3 within all segments combined but the analysis was performed on a previously published large dataset (Fineran et al., 2014). From this dataset, mutants with 3 mutations (all canonical PAM) were analyzed. The average of the direct interference group is indicated by the red square.

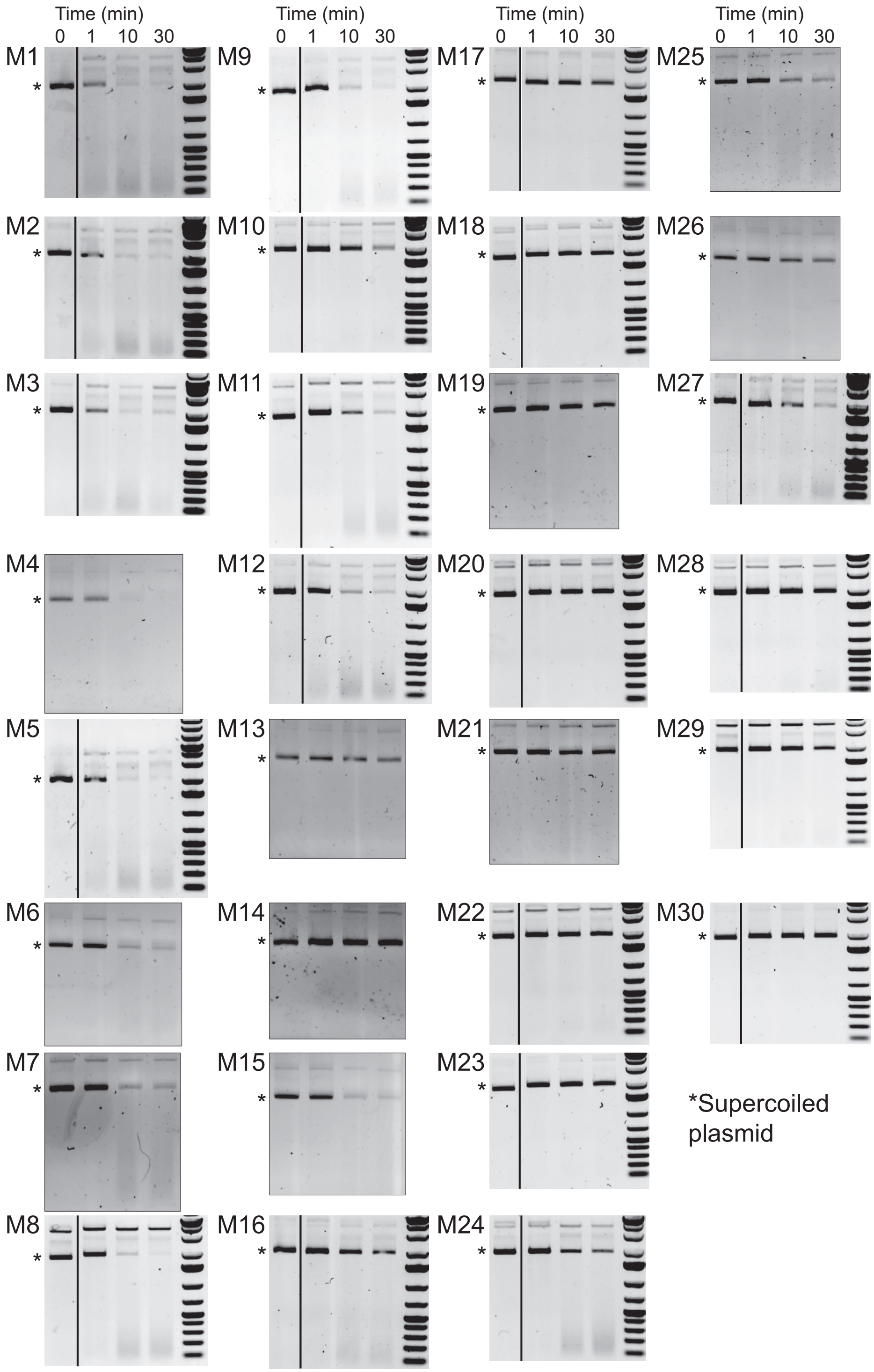

Time (min)
0 1 10 30
Time (min)
0 1 10 30
Time (min)
0 1 10 30
Time (min)
0 1 10 30
M1
M2
M3
M4
M5
M6
M7
M8
M9
M10
M11
M12
M13
M14
M15
M16
M17
M18
M19
M20
M21
M22
M23
M24
M25
M26
M27
M28
M29
M30
*
*Supercoiled plasmid

Figure S4. Related to Figure 2B. Representative gels of Cas3 activity assays. Individual gels for each mutant showing Cas3 plasmid degradation reactions at time points 0, 1, 10, 30 minutes. Vertical black lines indicate removal of 3 gel lanes with irrelevant samples. Supercoiled plasmid is indicated with an asterisk, gel lanes above are linearized and nicked plasmids, which are not considered in quantification.

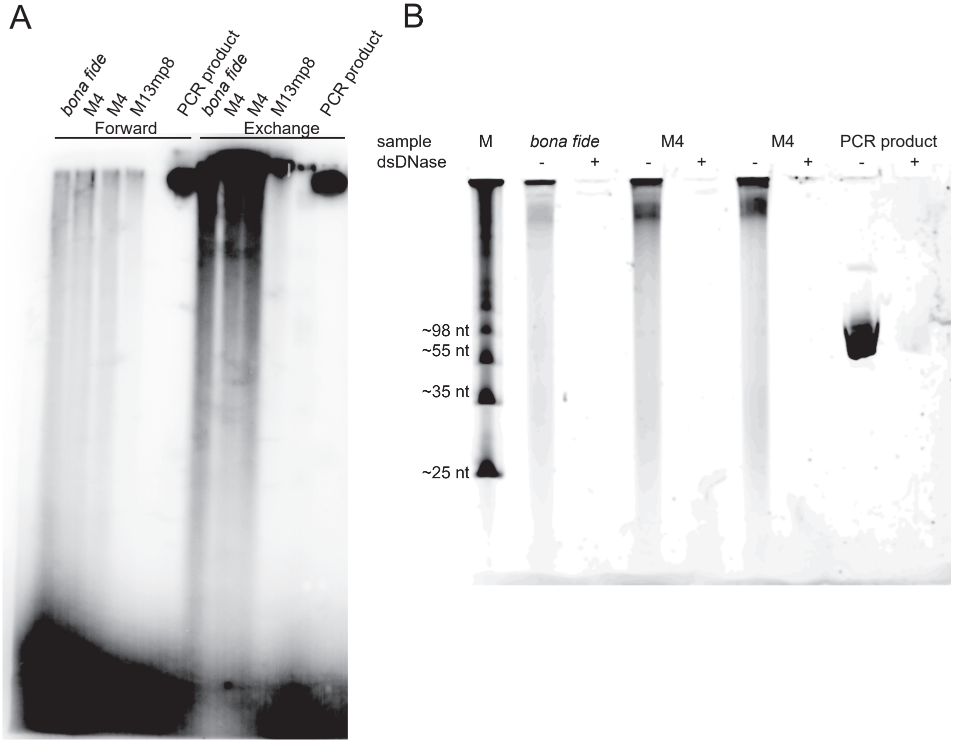


Figure S5. Related to Figure 3: Biochemical analysis of Cas3 DNA degradation fragments. A) ^{32}P PNK labeling of degradation fragments from *bona fide* target plasmid, M4 target plasmid and m13mp8 single stranded plasmid. Forward reaction can only label non-phosphorylated 5'ends, exchange reaction can label both phosphorylated and non-phosphorylated 5' ends. Non-phosphorylated PCR product for reference. B) dsDNase incubation with degradation fragments of *bona fide* target plasmid and M4 target plasmid. dsDNase is a double stranded DNA specific endonuclease with no activity on single stranded DNA.

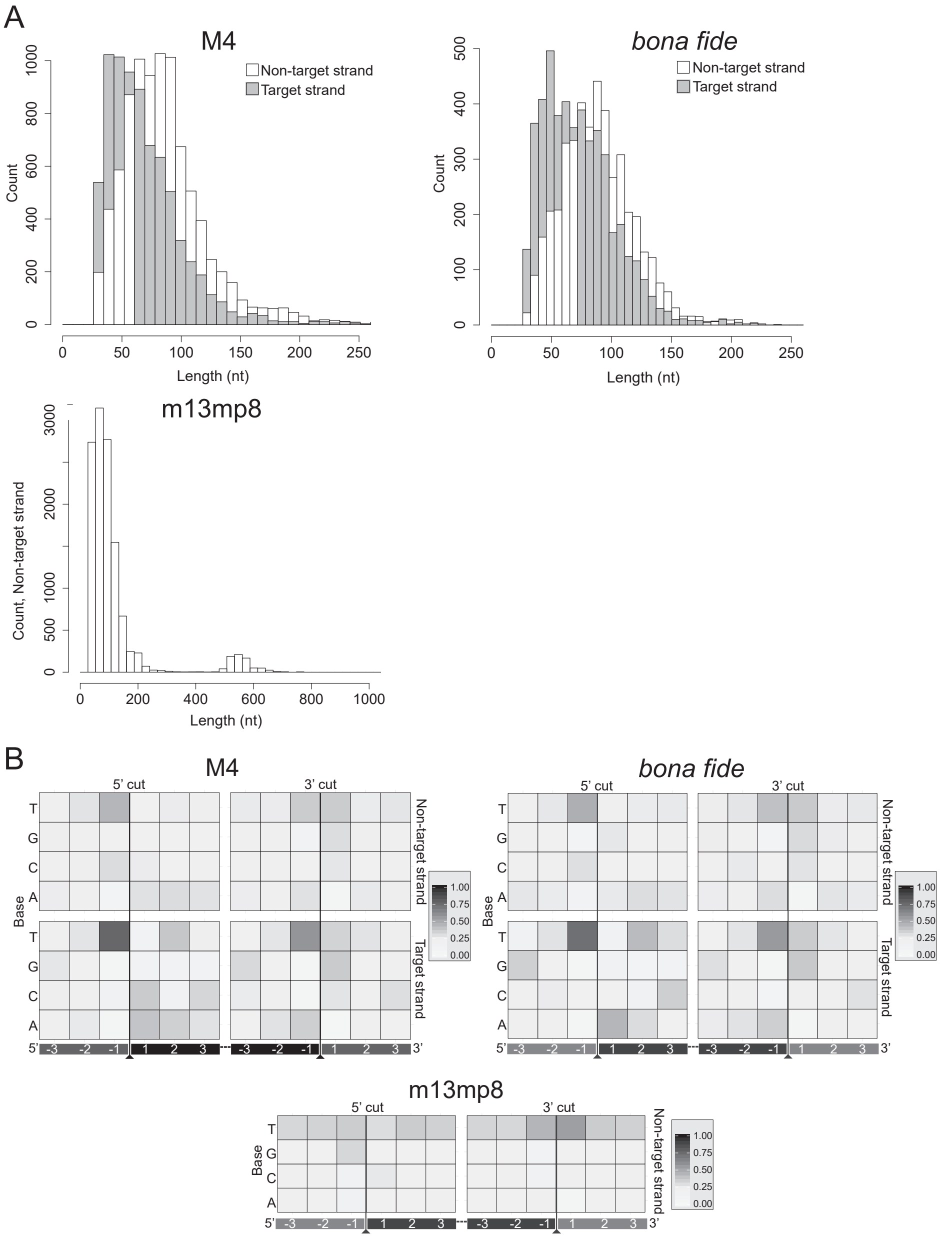


Figure S6. Related to Figure 3: Next generation sequencing analysis of Cas3 DNA degradation products. A) Length distribution bar charts for Cas3 DNA degradation products of *bona fide* target plasmid, M4 target plasmid and m13mp8 single stranded plasmid. B) Heat maps of nucleotide frequencies around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

Table S1. Related to Figures 1-6: Oligo nucleotides used in this study

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

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

Plasmid	Description (positions of all mutations)	Name in paper	source
pWUR835	pGFP-UV Amp	-	(Fineran et al., 2014)
pWUR836	pGFP-UV Km protospacer8 WT	pTarget bona fide	(Fineran et al., 2014)
pWUR837	pGFP-UV Km protospacer8 mutant pos. 1, 3, 24	pTarget M14	(Fineran et al., 2014)
pWUR838	pGFP-UV Km protospacer8 mutant pos. 10, 11, 25	pTarget M12	(Fineran et al., 2014)
pWUR839	pGFP-UV Km protospacer8 mutant pos. 1, 4, 16	pTarget M30	(Fineran et al., 2014)
pWUR840	pGFP-UV Km protospacer8 mutant pos. 2, 3, 4	pTarget M17	(Fineran et al., 2014)
pWUR841	pGFP-UV Km protospacer8 mutant pos. 3, 7, 19	pTarget M26	(Fineran et al., 2014)
pWUR842	pGFP-UV Km protospacer8 mutant pos. 4, 8, 26	pTarget M23	(Fineran et al., 2014)
pWUR843	pGFP-UV Km protospacer8 mutant pos. 2, 10, 16	pTarget M16	(Fineran et al., 2014)
pWUR844	pGFP-UV Km protospacer8 mutant pos. 2, 18, 22	pTarget M9	(Fineran et al., 2014)
pWUR845	pGFP-UV Km protospacer8 mutant pos. 10, 14, 17	pTarget M5	(Fineran et al., 2014)

pWUR846	pGFP-UV Km protospacer8 mutant pos. 11, 16, 17	pTarget M7	(Fineran et al., 2014)
pWUR847	pGFP-UV Km protospacer8 mutant pos. 11, 22, 32	pTarget M1	(Fineran et al., 2014)
pWUR848	pGFP-UV Km protospacer8 mutant pos. 5, 6, 25	pTarget M2	(Fineran et al., 2014)
pWUR850	pGFP-UV Km protospacer8 mutant pos. 2, 8, 26	pTarget M10	(Fineran et al., 2014)
pWUR851	pGFP-UV Km protospacer8 mutant pos. 19, 27, 32	pTarget M27	(Fineran et al., 2014)
pWUR852	pGFP-UV Km protospacer8 mutant pos. 12, 17, 31	pTarget M3	(Fineran et al., 2014)
pWUR853	pGFP-UV Km protospacer8 mutant pos. 6, 7, 32	pTarget M4	(Fineran et al., 2014)
pWUR854	pGFP-UV Km protospacer8 mutant pos. 1, 10, 15, 18, 29	pTarget M25	(Fineran et al., 2014)
pWUR855	pGFP-UV Km protospacer8 mutant pos. 1, 16, 19, 25, 29	pTarget M13	(Fineran et al., 2014)
pWUR856	pGFP-UV Km protospacer8 mutant pos. 1, 4, 19, 27, 28	pTarget M20	(Fineran et al., 2014)
pWUR857	pGFP-UV Km protospacer8 mutant pos. 2, 12, 23, 26, 27	pTarget M11	(Fineran et al., 2014)
pWUR859	pGFP-UV Km protospacer8 mutant pos. 3, 8, 10, 11, 22	pTarget M29	(Fineran et al., 2014)
pWUR860	pGFP-UV Km protospacer8 mutant pos. 3, 15, 20, 25, 26	pTarget M18	(Fineran et al., 2014)
pWUR859	pGFP-UV Km protospacer8 mutant pos. 3, 9, 13, 22, 26	pTarget M19	(Fineran et al., 2014)
pWUR860	pGFP-UV Km protospacer8 mutant pos. 5, 6, 8, 24, 31	pTarget M8	(Fineran et al., 2014)
pWUR861	pGFP-UV Km protospacer8 mutant pos. 4, 5, 6, 15, 24	pTarget M24	(Fineran et al., 2014)
pWUR862	pGFP-UV Km protospacer8 mutant pos. 1, 2, 9, 14, 21	pTarget M22	(Fineran et al., 2014)
pWUR863	pGFP-UV Km protospacer8 mutant pos. 6, 22, 27, 31, 32	pTarget M28	(Fineran et al., 2014)
pWUR864	pGFP-UV Km protospacer8 mutant pos. 12, 13, 23, 24, 30	pTarget M6	(Fineran et al., 2014)
pWUR866	pGFP-UV Km protospacer8 mutant pos. 3, 9, 12, 16, 32	pTarget M21	(Fineran et al., 2014)
pWUR867	pGFP-UV Km protospacer8 mutant pos. 17, 27, 28, 29, 30	pTarget M15	(Fineran et al., 2014)
pWUR868	pGFP-UV Km non-target	pTarget NT	(Fineran et al., 2014)
pWUR748	pMAT11-MBP-Cas3		(Mulepati and Bailey, 2013)
pWUR868	pACYC poly spacer8 CRISPR array		This study
pWUR514	cse2 with Strep-tag II (N-term)-cas7-cas5-cas6e in pET52b		(Jore et al., 2011a)
pWUR408	cse1 in pRSF-1b, no tags		(Brouns et al., 2008b)
pWUR477	pACYC with artificial CRISPR array		(Brouns et al., 2008b)
pWUR872	pWUR477 with only one repeat	pCRISPR	This study
pWUR871	Cas1-Cas2 operon with Strep-tag II (N-term) in pET52b		This study

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

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

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

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

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

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

Chapter 6

Opposite effects of guanine and cytosine protospacer mismatches in direct CRISPR interference and priming

Tim Künne[1], Yifan Zhu[1], Fausia da Silva[1], Nico Konstantinides[1], John van der Oost[1] and Stan J.J. Brouns[1,2]

[1]Laboratory of Microbiology, Wageningen University, 6708 WE Wageningen, The Netherlands.
[2]Department of Bionanoscience, Kavli Institute of Nanoscience, Delft University of Technology, 2629 HZ Delft, The Netherlands.

Manuscript in preparation:

Künne T, Zhu Y, da Silva F, Konstantinides N, van der Oost J, and Brouns SJ

Abstract

Prokaryotes use the CRISPR-Cas system to remove unwanted nucleic acid invaders from the cell. Immunity is based on the complementarity of host-encoded spacer sequences with protospacers on the foreign genetic element. Matching protospacers are detected by RNA-guided effector complexes and flagged for destruction. Invaders can evade this interference by acquiring mutations in the protospacer or in the protospacer adjacent motif (PAM). The type I-E system in *Escherichia coli* can tackle these invaders using a primed acquisition process that rapidly restores immunity by incorporating new spacers. The efficiency of both direct CRISPR interference and primed acquisition depends on the degree of complementarity between spacer and protospacer. Previous studies focused on the number and positions of mutations in the protospacer, not on the identity of the substituted nucleotide. We previously detected that nucleotide-dependent effects rule priming, showing a positive effect of C mismatches and a negative effect of G mismatches. Here we show that these substitutions in the protospacer dictate the efficiency of interference and therefore determine the efficiency of interference-dependent priming. We show that G substitutions in the target strand of the protospacer are detrimental to interference, while C substitutions are readily tolerated. Furthermore, we show that this effect is based on strongly decreased binding affinity of the effector complex Cascade for G mismatches, while C mismatches only minimally affect binding. This effect has strong implications for the mutations that mobile genetic elements can introduce to escape CRISPR systems more effectively.

Introduction

Clustered regularly interspaced short palindromic repeats (CRISPR) together with CRISPR-associated (Cas) proteins provide immunity against foreign nucleic acids in prokaryotes (Barrangou et al., 2007; Brouns et al., 2008a). The constant battle between prokaryotes and their viruses is one of the oldest and most prominent predator-prey interactions on our planet (Breitbart and Rohwer, 2005; Rohwer and Thurber, 2009). The CRISPR array consists of identical repeat units separated by unique spacers. In many cases spacer sequences are derived from foreign genetic elements although 'self'-derived spacers can also be found (Bolotin et al., 2005; Makarova et al., 2006; Mojica et al., 2005; Pourcel et al., 2005; Stern et al., 2010). CRISPR-Cas systems are currently divided into class 1 and class 2 systems, with class 1 consisting of type I, III and IV, and class 2 consisting of type II, V and VI (Makarova et al., 2015; Makarova et al., 2017a, b). Each type (except IV) contains a number of subtypes. Type I systems contain the universally conserved *cas1* and *cas2* genes, the hallmark *cas3* helicase-nuclease and a set of genes encoding for the Cascade-like effector complexes. The mechanism of CRISPR-Cas defence is divided into three stages: Adaptation, expression and interference (van der Oost et al., 2014). First, a new spacer is acquired from an invader DNA that has not previously been encountered and is incorporated into the CRISPR array by the Cas1-2 complex (adaptation) (Barrangou et al., 2007; Jackson et al., 2017). Next, the whole array is transcribed from the AT-rich leader sequence into long pre-CRISPR RNA (pre-crRNA) and subsequently processed into mature crRNAs that each carry one spacer (expression) (Brouns et al., 2008a; Charpentier et al., 2015). The latter assembles with Cas proteins to form surveillance complexes that make up the core of all CRISPR systems (Makarova et al., 2015). In the last stage (interference), these surveillance complexes scan the cell for complementary targets and flag them for destruction, leading to immunity (Garneau et al., 2010; Hale et al., 2009; Marraffini and Sontheimer, 2008; Westra et al., 2012c). Invaders can escape immunity by acquiring mutations in their recognition sequence (protospacer) or PAM, which implies that the host has to acquire a new spacer in order to regain immunity. Several type I systems possess a primed acquisition mechanism that leads to rapid acquisition of new spacers when escape protospacers are detected (Datsenko et al., 2012; Li et al., 2014; Richter et al., 2014; Swarts et al., 2012; Vorontsova et al., 2015). Unlike naïve acquisition, which requires only *cas1* and *cas2*, primed acquisition requires all *cas* genes and a targeting spacer (Datsenko et al., 2012; Yosef et al., 2012). A number of studies have described the effect of mutations on interference and priming in the type I-E system of *E. coli*. Two early studies have shown on a small scale that interference tolerates only few mutations in the protospacer and no mutations in the seed and PAM, while priming is slightly more tolerant (Datsenko et al., 2012; Semenova et al., 2011). Our previous work has extended this knowledge on a large scale, showing that interference tolerates mutations in the seed to a low degree and that priming is extremely robust against mutations in the entire protospacer (Chapter 4) (Fineran et al., 2014). More recently, it was shown that mutation tolerance of the different immune responses is highly dependent on the primary sequence of the spacer/protospacer (Xue et al., 2015).

These studies mainly focussed on the number and position of the mutations, not on the individual nucleotides, thus we know little about the effect of different types of nucleotide substitutions. Interestingly, however, in a previous study we did observe a nucleotide bias that affected priming acquisition (Fineran et al., 2014). Having more cytosine substitutions in the target strand of the protospacer resulted in a positive correlation with the ability to induce priming, while an increasing number of guanine substitutions negatively correlated with the induction of priming. In contrast, adenine and thymine did not show any significant effect. In this analysis, the number of C or G substitutions was scored irrespective of any other mutations present in a particular mutant. We therefore set out to test whether this holds true for individual mutants with only C or G mutations. More specifically, we were wondering whether the behavior of a mutant (priming or stable) could be reversed by switching C mutations to G mutations in the same positions or vice versa. Moreover, we analyzed whether C mutations, on an individual level, actually promote priming or rather repress it not as strong as other mutations. Finally, attempts were made to reveal the mechanism that causes this opposing behavior of C and G mutations.

Here we show that C and G mutations affect priming indirectly, by altering the efficiency of interference (target degradation), which in turn has an effect on priming. We show that, while the overall effect is strongly dependent on the position of the mutations, C mutations repress interference only moderately compared to G mutations at the same positions. The effect on priming is more complex, not correlating directly with the type of mutation. Instead, priming is stimulated by intermediate interference rates, while high or low interference appears to repress priming. This is in agreement with previous studies that have shown the same dependence of priming on interference (Künne et al., 2016; Semenova et al., 2016; Staals et al., 2016). Furthermore we show that this behavior is caused by a higher mismatch penalty for G’s in the target strand of the protospacer compared to C’s, resulting in lower Cascade binding affinities for mutant targets containing G substitutions.

Results

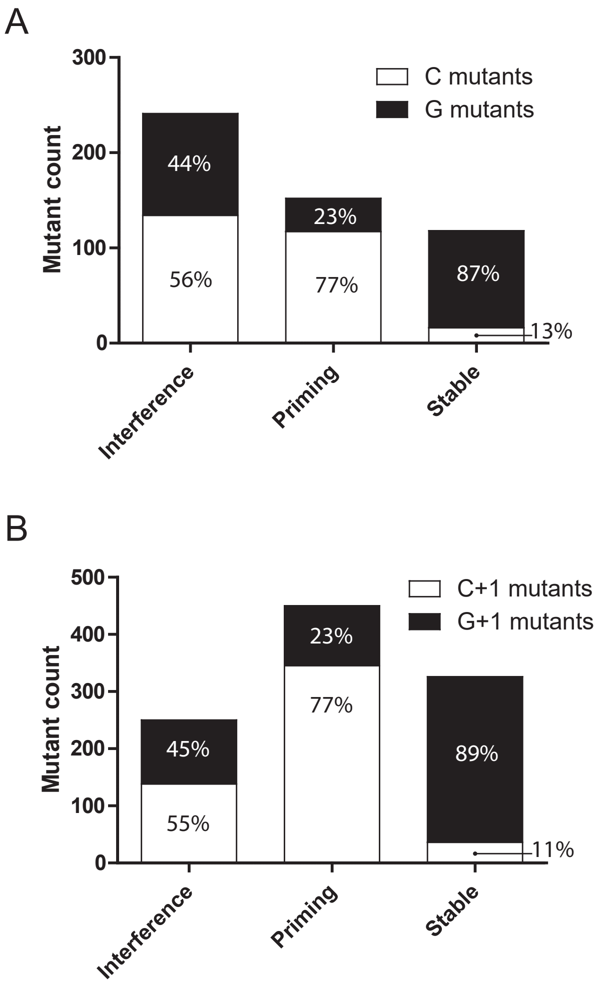


Figure 1 – Statistics of G/C bias. Statistical analysis of the high throughput plasmid loss dataset from (Fineran et al., 2014). Only effective mutations, thus excluding positions 6, 12, 18, 24, 30, are considered. A) Mutants with only C or G mutations (> 2) are counted for each group of immune responses (Interference, Priming, Stable). B) Same as in (A) but in addition to the C or G mutations, one random mutation is allowed when there are at least 3 C or G mutations.

Statistical scoring of C and G mutants

In chapter 4, we performed a high throughput plasmid loss assay with a large library of PAM/protospacer mutants, which lead to their classification as causing either (i) interference, (ii) priming, or (iii) stable plasmid maintenance. Analysis showed the aforementioned nucleotide bias, but mutants were scored for the number of C or G mutations irrespective of the presence of any other mutations (Fineran et al., 2014). To verify that the observed effect is purely based on the C and G mutations, we re-analyzed the original dataset, but this time we selected mutants with only C or G mutations to exclude the influence of other mutations. All analyses were done using effective mutations, thus excluding positions 6, 12, 18, 24, 30 (pinch points), which do not participate in base pairing (Jackson et al., 2014; Mulepati et al., 2014; Zhao et al., 2014). We included mutants with at least 2 effective mutations, since single mutants are unlikely to have a nucleotide specific effect and most single mutants lead to direct interference. The pure C_n or G_n mutants (n≥2) were grouped according to their classifications and counted (Figure 1A). Due to the relatively small resulting sample set, we repeated the same analysis, this time allowing one additional random mutation in mutants with at least 3 C or G mutations (n≥3; Figure 1B). The results of the two analyses are almost identical, showing the same effect that was observed previously (Fineran et al., 2014). The priming group in both analyses contains mainly C mutants (117 out of 152, 77%; 345 out of 450, 77%), while the stable group contains mostly G mutants (94 out of 108, 87%; 290 out of 326, 89%). This confirms that C mutations generally stimulate priming, while G mutations generally repress priming. Interference has only a slight preference for C mutants over G mutants (56 %/ 44 % for C/G respectively). This suggests either that interference is largely unaffected by the type of mutation or that the effect of the mutations on interference is hard to detect because the majority of mutants in this group carry only 2 effective mutations which likely do not show a strong effect. It is considered very likely that interference is indeed also affected by the type of mutations, because priming is directly dependent on interference (Künne et al., 2016).

Table 1: Overview of mutant set

Original mutant	Original classification	Conversion mutant	Predicted classification
C1	Priming	G1	Predicted stable
C2	Priming	G2	Predicted stable
C3	Priming	G3	Predicted stable
C4	Priming	G4	Predicted stable
G5	Stable	C5	Predicted priming
G6	Stable	C6	Predicted priming
G7	Stable	C7	Predicted priming
G8	Stable	C8	Predicted priming
G9	Stable	C9	Predicted priming

To address the question to what extent priming and interference are influenced by the type of mutations, and to analyze the effect of the mutations in more detail, we selected four priming protospacers from the dataset with only C mutations (C1-4) and five stable protospacers with only G mutations (G5-9) (Table 1, Table S1). The mutants were selected based on two restrictions: (i) the mutations had to be effective mutations, meaning they could not be on a pinch point, and (ii) the original nucleotide must be A or T, so that we can switch the mutations from C to G or vice versa without reverting to WT. After selecting the mutants, we designed the respective conversion mutants (G1-4, C5-9).

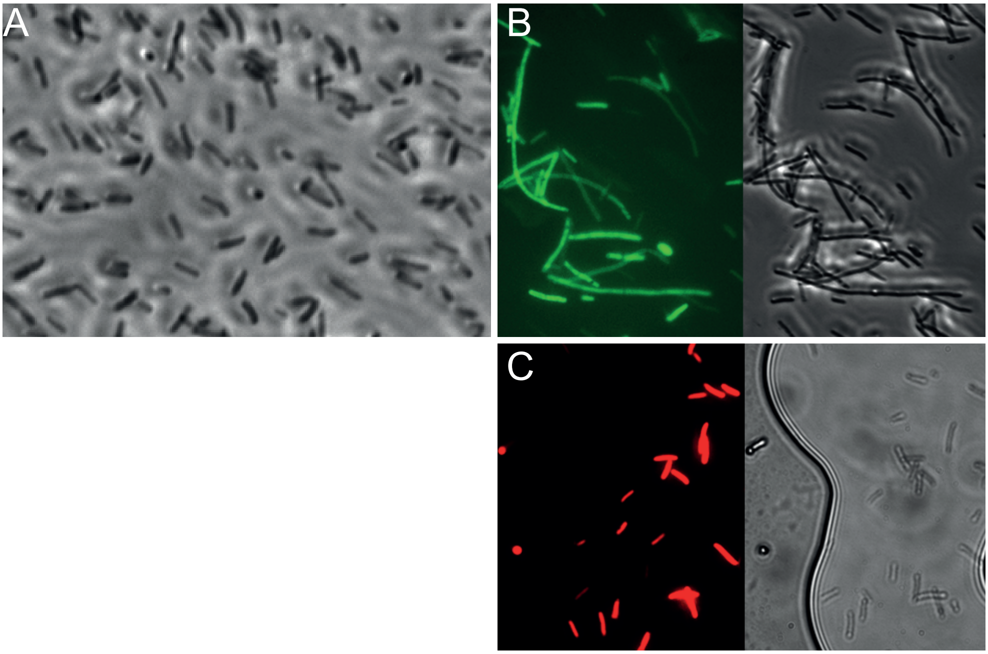


Figure 2 – Phenotypes of *E. coli* with reporter plasmids. Microscopy pictures of *E. coli* KD263 cells used in this study during growth with different reporter plasmids. A) Without reporter plasmid. B) With pGFP-UV plasmid (constitutive GFP expression) used in earlier studies, showing elongated cells C) With pBex plasmid (Rhamnose inducible RFP expression).

RFP-reporter plasmids for optimized in vivo assays

The selected protospacers and their corresponding conversion mutants were cloned into plasmids carrying an inducible RFP reporter. We used the inducible RFP reporter instead of the constitutive GFP reporter from our previous studies, because we observed an aberrant growth phenotype in *E. coli* cells expressing GFP (Figure 2B). Cells expressing GFP were elongated and this might lead to an increased selective advantage/growth rate for cells that lost the plasmid over cells that carry the plasmid. RFP expression does not show this effect (Figure 2C) and the inducible promoter further reduces the energetic burden and thereby growth disadvantage of plasmid carrying cells. Although we did not observe any problems with the previously used GFP-based system, the RFP system further reduced the background of CRISPR-independent plasmid loss to undetectable levels in the timeframe of our experiments (≤48 h). Moreover, the new system reduces the potential overestimation of plasmid loss, due to the reduced growth rate advantage of cells without plasmid.

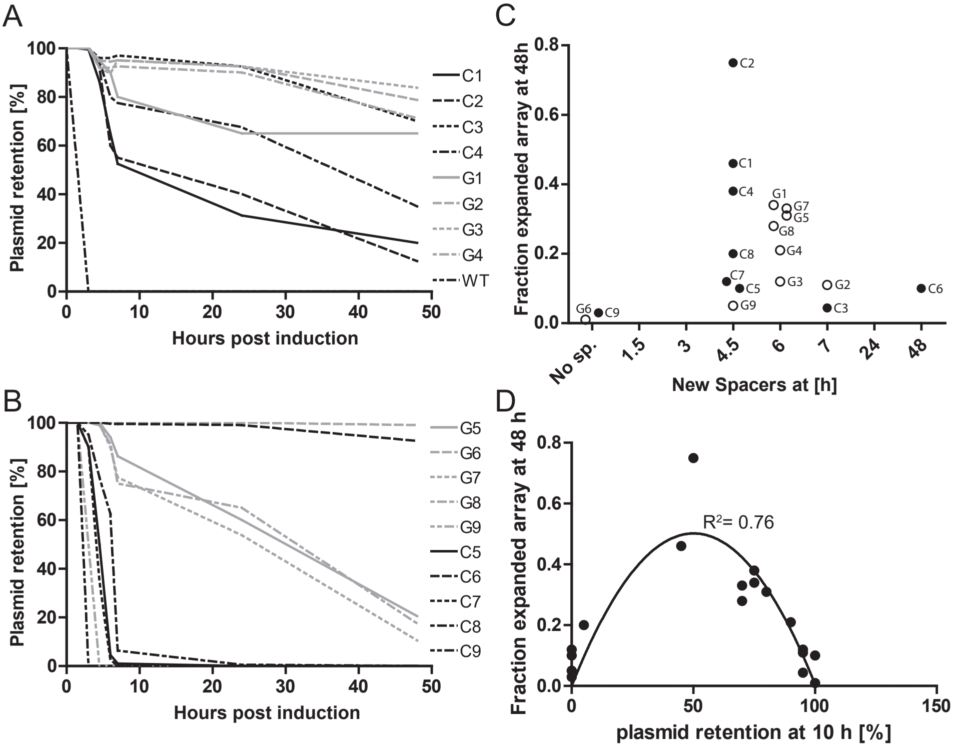


Figure 3 – Plasmid loss and spacer acquisition. Plasmid loss and spacer acquisition assays of individual mutant plasmids. Two independent assays were carried out, each in duplicate. The standard error of the mean of the measurements can be found in figure S1. A) Plasmid loss curves of C/G pairs 1-4 and the WT. B) Plasmid loss curves of C/G pairs 5-9. C) Analysis of priming during plasmid loss assays, indicating the first occurrence of visible CRISPR array expansion (X-axis) and the extent of priming (fraction of population with expanded array after 48 h, Y-axis). D) Plot showing the correlation of speed of plasmid loss (represented by the remaining plasmid after 10 h) with the extent of priming (represented by the fraction of the population with an expanded array after 48 h. Distribution is fitted with a parabola, indicating optimal array expansion at intermediate plasmid loss speeds.

G mutants strongly inhibit direct interference

First we performed plasmid loss assays to accurately determine and quantify the ability of the mutant protospacers to trigger direct interference and priming. No plasmid loss and no spacer acquisition was observed with a non-target plasmid after 48h, showing that CRISPR-independent plasmid loss and naïve acquisition do not occur at detectable levels in this timeframe (not shown). When comparing the respective pairs of C and G mutants, we observed that the C mutants consistently showed more rapid plasmid loss than the G mutants (Figure 3AB, Figure S1). Especially, some mutants that were switched from G to C (C5, C7, C8) drastically increased their speed of plasmid loss to almost WT levels. Two pairs of mutants show only small differences between the C and G version (C6/G6, G9/C9). The original G9 mutant already shows rapid plasmid loss, which simply cannot be increased much more in the C9 mutant. The original G6 mutant is stable and the C6 mutant is only able to show strongly delayed priming. This is likely an effect of the positions of the mutations, which are detrimental for interference/priming regardless of nucleotide identity. In many of the mutants, significant plasmid loss was observed within 5 hours, indicating that this was caused by direct interference rather than priming (Figure 3AB). This is supported by the analysis of spacer acquisition that showed observable spacer acquisition initiated after the onset of plasmid loss (Figure 3C). Spacer acquisition also initiated consistently earlier in the C mutants than their respective G mutants. The extent of priming, i.e. the fraction of the population that acquired new spacers, on the other hand is not consistent with the type of mutations. For example, the G9/C9 mutant pair where the interference is already very high in the G mutant (and even higher in the C mutant) show opposite behavior with respect to their priming response. Here, the G mutant shows a low level of early priming, while the C mutant shows no priming. We observe that the extent of priming is the highest when plasmid loss is occurring at intermediate speeds, while rapid or slow plasmid loss leads to a low extent of priming (Figure 3D). This is very well in line with the model proposed in our and others previous work, i.e. that priming is not only dependent on interference, but also on persistence of the invader in the host cell in order to provide sufficient time for spacer acquisition (Künne et al., 2016; Semenova et al., 2016; Severinov et al., 2016; Staals et al., 2016).

In conclusion, G mutants strongly inhibit interference while C mutants are much more tolerated. The effect on priming that has been observed in the original dataset therefore results from the effects of mismatched C/G bases on direct interference rates.

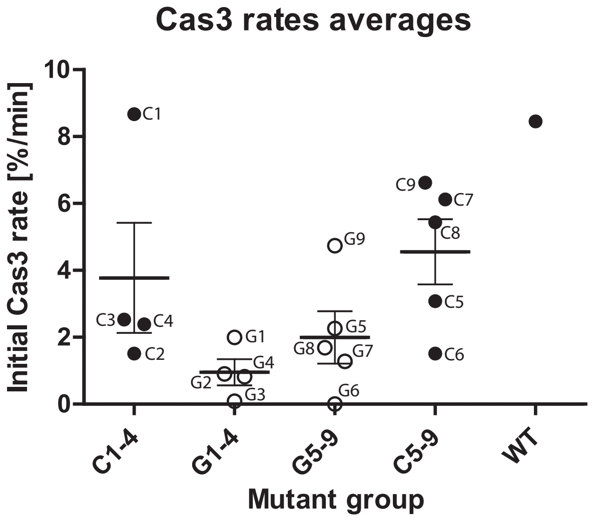
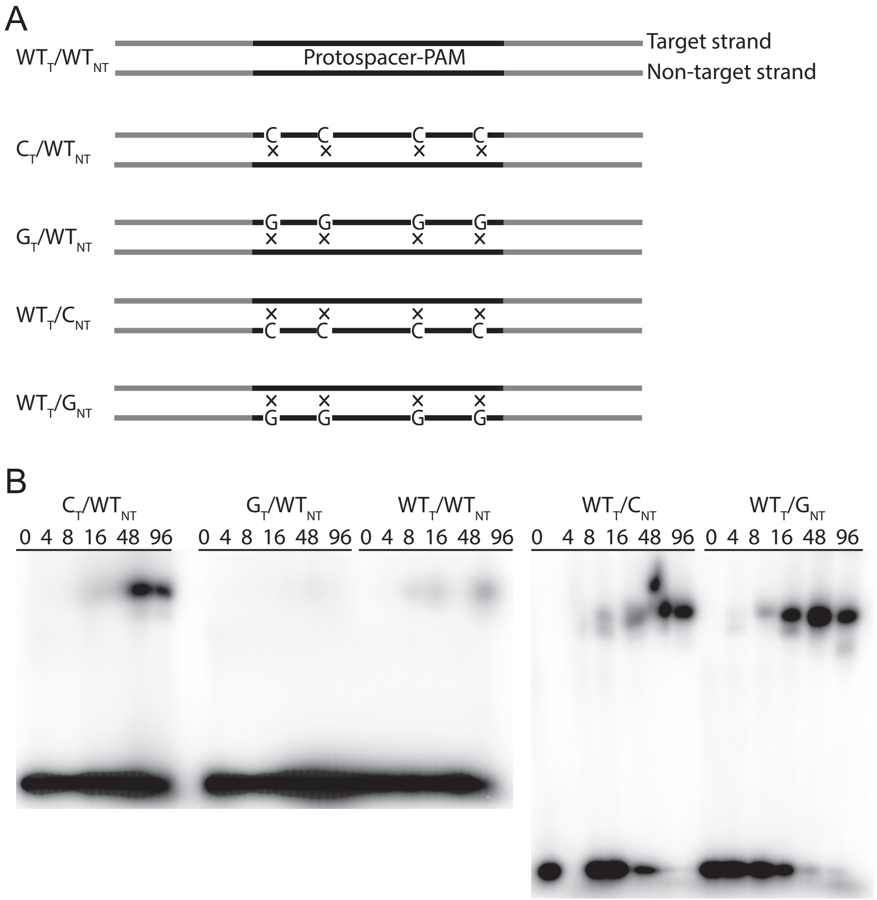


Figure 4 – *In vitro* Cas3 activity assays. The initial Cas3 reaction rate (over first 10 minutes) is plotted for each mutant plasmid. The rate is measured as the percentage of plasmid degradation per minute. The rates are the average of two independent experiments. The plasmids are split into groups based on their type of mutations. C mutants are indicated by full circles, G mutants are indicated by empty circles. The mean rate and standard error of the mean (SEM) are indicated for each group. Individual Cas3 activity graphs can be found in figure S2.

G mutants inhibit Cas3 degradation rate

To elucidate the molecular basis of this difference in interference of C and G mutants, we initially tried to measure the binding affinity of Cascade for the mutant plasmids. Unfortunately, the used EMSA approach did not allow for detecting binding in the majority of the used mutants, indicating that the interaction is too weak to be detected by EMSA analysis. Instead, we performed Cas3 activity assays with the set of plasmids. We have previously shown that Cas3 activity *in vitro* is a very good indicator for the level of direct interference and consequently for priming (Künne et al., 2016). Cas3 assays consistently show a higher average activity of C mutants over G mutants (Figure 4, Figure S2). Furthermore, looking at the individual C/G mutant pairs, we see consistently higher activity of the C mutants compared to their corresponding G mutants. This confirms that the more rapid plasmid loss of C mutants is indeed caused by a more efficient plasmid degradation by Cas3.



6

Figure 5 – Oligo EMSA. A) Overview of oligonucleotide duplex combinations tested. Each combination contains one WT strand and one mutated strand. B) EMSA of Cascade with ^{32}P labelled oligo duplexes. Cascade to DNA molar ratios are indicated above each lane. Top bands are Cascade bound DNA and bottom bands are free DNA.

G mutations in the target strand disrupt Cascade binding

Since the plasmid EMSAs were unsuccessful, we wondered whether the difference in Cas3 activity is caused by affinity differences of Cascade for the mutant protospacers, by conformational differences of the Cascade/R-loop complexes or just by differences in Cas3 activity. Furthermore, it is currently unknown whether the effect is caused only by the mismatches between the crRNA and the DNA target strand, or whether the non-target strand plays a role as well. The non-target strand has been proposed to make interactions with the Cse2 dimer and might therefore have an effect on overall R-loop stability (Jackson et al., 2014; Nam et al., 2012b; Wiedenheft et al., 2011a). To address these matters, we designed oligo nucleotides for each strand carrying a protospacer with either the WT sequence, C mutations or G mutations. The sequences were chosen based on the mutant pair C7/G7. The oligos were annealed in certain combinations that allow investigation of the effect of the mutations in either strand separately (Figure 5A). We performed EMSA assays to measure the binding affinity of Cascade with the 5 oligo duplex combinations (Figure 5B). We observed the highest binding affinity for the C and G mutants of the non-target strand with a WT target strand (WT_T/C_{NT}, WT_T/G_{NT}). These two mutants have near identical binding, showing that the mutations in the non-target strand have no effect on the affinity. The C mutant in the target strand (C_T/WT_{NT}) has a considerably higher binding affinity than the G mutant in the target strand (G_T/WT_{NT}). The latter shows almost no detectable binding, while the former shows binding comparable to the full WT (WT_T/WT_{NT}). The full WT oligos (WT_T/WT_{NT}) show lower binding affinity than the two non-target strand mutants with a WT target strand, probably due to the fact that the mismatches between the two strands in the mutants lower the energetic barrier of strand invasion and R-loop formation. Cascade strongly favors these destabilized duplexes, which agrees with the observation that it prefers negatively supercoiled target plasmids (Westra et al., 2012c). Overall, these affinity measurements show that G mutations in the target strand disrupt Cascade binding to a much greater extent than C mutations. The bases in the non-target strand have no effect on the Cascade affinity.

Discussion

In this study, we have shown that the type I-E CRISPR-Cas system readily tolerates cytosine mutations in the target strand of the protospacer, while guanine mutations severely reduce the efficiency of direct interference. This difference is caused by a strong reduction of Cascade binding affinity towards targets with guanine substitutions, while the binding affinity is hardly affected in case of cytosine substitutions. Thus, these mutations do not directly influence the priming process as originally thought (Fineran et al., 2014), but instead alter target degradation rates and consequently the interference-dependent priming process. The fact that we did not observe a nucleotide-dependent effect on interference in the original dataset (Fineran et al., 2014), may have two reasons: the conservative classification in the high throughput assay of that study, leaving many mutants unclassified, and the fact that the mutants in the direct interference category carry very low numbers of mutations which likely does not produce the nucleotide specific effect. The direct effect on interference and the indirect effect on priming of the C/G mutants is also shown by the fact that, although C mutants in all cases lead to earlier priming than G mutants, the extent of priming in the whole population is not directly related to the type of mutation. Instead, the extent of priming follows the model that was conceived in an early study (Swarts et al., 2012) and established in three recent studies. These studies showed that the priming process is directly dependent on direct interference and that in fact direct interference produces the precursor molecules for new spacers during priming (Künne et al., 2016; Semenova et al., 2016; Staals et al., 2016). However, next to requiring interference for the production of precursors, priming is also dependent on sufficient time of persistence of the invader in the cell. Only prolonged persistence gives sufficient opportunity for spacer capture and integration. Thus, a very high rate of direct interference, such as for a WT target or the C9/G9 mutant pair in this study, on the one hand leads to a very early onset of priming, but on the other hand to a very low extent of priming. Mutants with low rates of direct interference lead to late onset of priming and a low extent of priming due to the lack of precursor generation. Mutants with an intermediate rate of direct interference, however, result in relatively early and a high extent of priming, due to prolonged persistence and simultaneous degradation of the invader.

This study confirms that priming in protospacer mutants (not PAM mutants) is dominantly occurring via the interference-dependent pathway (Künne et al., 2016; Semenova et al., 2016; Staals et al., 2016) and not the interference-independent pathway (Redding et al., 2015).

A possible explanation as to why G mutations are so much more detrimental for protospacer binding by Cascade than C mutations may be found in the bulkiness of mismatched C and G nucleotides and their effects on R-loop progression and stability. Guanine, being a purine, is bulkier than cytosine which is a pyrimidine. However, in this case we would expect to see the same effect for adenine (purine) and thymine (pyrimidin). The difference between C and G mutations also cannot be explained by pure thermodynamics of RNA-DNA mismatches for several reasons. First, the effect of mismatches on the total duplex free energy is highly context dependent. Thus both the identity of the opposing RNA base as well as both neighboring base pairs influence the free energy (Huang et al., 2009; Sugimoto et al., 2000; Watkins et al., 2011; Wu et al., 2002; Zhu and Wartell, 1999). This would mean that the effect of dC or dG mismatches should average out, rather than show the clear opposing effect that we observed. Second, the experimental analysis of many RNA/DNA mismatch combinations has shown that some combinations containing a dG mismatch are actually more stable than combinations containing a dC mismatch (e.g. rU·dG ≈ rG·dG > rA·dG ≈ rA·dC > rU·dC>rC·dC) (Sugimoto et al., 2000; Watkins et al., 2011). The complete estimation of duplex free energies remains difficult, because only around 72 of the 240 possible RNA-DNA mismatches (dinucleotide nearest-neighbor model) have been experimentally measured (Farasat and Salis, 2016). Another complicating factor in predicting the mismatch energies might be found in the distorted nature of the crRNA:DNA duplex which does not perfectly resemble an A-form helix (Hayes et al., 2016). Thus it is likely that the effect we observe is caused by the differential tolerance to G and C mutations in the context of the Cascade associated R-loop.

The detrimental effect of G mismatches also has consequences for the success of viral escape mutants. Although viruses have been shown in a number of systems to preferably mutate the PAM or seed to escape CRISPR-Cas immunity (Box et al., 2015; Deveau et al., 2008; Kupczok and Bollback, 2014; Paez-Espino et al., 2015; Semenova et al., 2011), systems capable of priming can rapidly regain immunity against these mutants (Datsenko et al., 2012; Fineran et al., 2014; Li et al., 2014; Richter et al., 2014). Thus, only mutants with sufficient mutations to completely escape immunity have increased long-term survival. For these escape viruses it would be very beneficial to accumulate G mutations in the targeted strand of their protospacers to maximize the chances of escape. Since the CRISPR-Cas system can target either strand, viruses cannot simply prefer G over C mutations in general. Instead, the selective pressure on the viruses by the type I-E system should lead to a overrepresentation of G mutations in protospacers of viruses in natural ecosystems.

Materials and Methods

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

Molecular Biology and DNA Sequencing. All oligonucleotides are listed in Table S2. All plasmids are listed in Table S3. All strains and plasmids were confirmed by PCR and sequencing (GATC-Biotech). Plasmids were prepared using GeneJET Plasmid Miniprep Kits (Thermo Scientific). DNA from PCR was purified using the DNA Clean and Concentrator and Gel DNA Recovery Kit (Zymo Research). The protospacer plasmid set was constructed by cutting pWUR925 with XbaI and SacI, removing the kanamycin resistance marker, and ligating a PCR product containing the streptomycin resistance marker and the desired protospacer (primers: BG7167/7395-7 for controls, BG7167/8393-8410 for mutant set).

Plasmid loss assay. The assay was carried out in *E. coli* KD263 cells, which have inducible *cas* gene expression. Expression was induced with 0.2 % L-arabinose and 0.5 mM IPTG where appropriate. *E. coli* KD263 cells were transformed with the target plasmids (pWUR926-946) by heat shock. Individual colonies were picked in duplicate and grown overnight in 5 ml LB supplemented with 2% glucose to repress *cas* gene expression. The next day, cultures were transferred 1:100 into induced medium (0.2% L-Arabinose, 0.5 mM IPTG) and plasmid loss was monitored. Samples were taken at the time of induction and 1.5 h, 3 h, 4.5 h, 6 h, 7 h, 24 h and 48 h post induction (HPI). Dilutions were plated on non-selective plates containing 0.2 % rhamnose and plasmid loss was quantified based on loss of red color. Liquid culture samples were screened for spacer integration by colony PCR using OneTaq (NEB). Acquisition of spacers was detected by PCR using primers BG5301 and BG5302. PCR products were visualized on 2% agarose gels and stained with SYBR-safe (Invitrogen). PCR products were sequenced using Sanger sequencing at GATC (Konstantz, Germany) using primer BG5301.

Protein purification. All proteins were expressed in Bl21-AI cells. Cascade was purified as described earlier (Jore et al., 2011b). MBP-Cas3 was purified as described in (Mulepati and Bailey, 2013).

Oligo annealing and labelling. Complementary oligo nucleotides (BG9069-9074) were mixed (1:1) in a tris-sodium buffer, heated to 95 °C and slowly cooled to room temperature. Duplexes were checked on a native 20% acrylamide gel for residual single stranded oligo. The non-target substrate was PCR amplified from pWUR928 using BG9141/2. Duplexes were then labelled with γ-^{32}P-ATP using T4 PNK (NEB) and free label was removed using a G25 column.

EMSA assays. Purified Cascade complex with spacer8 crRNA was incubated with plasmid or oligos at a range of molar ratios (1:1-96:1, Cascade:DNA) in buffer A (20 mM HEPES pH7.5, 75 mM NaCl, 1 mM DTT) for 30 min. Plasmid reactions were run on 1% native agarose gels for 18h at 22 mA in 8 mM sodium-borate buffer. Gels were post stained with SYBR Safe (Invitrogen). Oligo reactions were run on 5 % native acrylamide gels at 4 mA for 18 h. Gels were exposed to a phosphor screen (GE Healthcare) and scanned using a phosphor imager (Bio-Rad PMI). Shifted (Cascade bound DNA) and unshifted (free DNA) bands were quantified using the GeneTools software (Syngene) or ImageJ and free Cascade concentration (X) was plotted against the fraction of bound DNA (Y). The curves were fitted with the following formula: $Y = (\text{amplitude} * X)/(K_d + X)$ (van Erp et al., 2015). The amplitude is the maximum fraction of bound DNA.

Cas3 DNA degradation assays. Plasmid-based assays were performed by incubating 70 nM Cas3 with 100 nM Cascade and 3.5 nM plasmid DNA. Oligo-based assays were performed by incubating 110 nM or 220 nM Cascade with 75 nM Cas3 and 5.5 nM oligo. Reactions were incubated in buffer R (5 mM HEPES, pH8, 60 mM KCl) supplemented with 10 µM $CoCl_2$, 10 mM $MgCl_2$, 2 mM ATP at 37 °C for the indicated amount of time. Plasmid samples were immediately quenched on ice with 6x DNA loading dye (Thermo scientific), oligo samples were quenched on ice in 2x RNA loading dye (NEB). Plasmid samples were run on 0.8 % agarose gels at 100 V for 40 minutes and supercoiled plasmid bands were quantified using the GeneTools software (Syngene).

Author contributions

T.K. and S.J.J.B. designed research; T.K., Y.Z., F.dS. and N.K performed research; T.K., N.K., J.vdO. and S.J.J.B. analyzed data; and T.K. and S.J.J.B. wrote the paper with input from all authors.

Acknowledgements

This work was financially supported by an LS6 ERC starting grant (639707) and NWO VIDI grant to S.J.J.B. (864.11.005). S.J.J.B was supported by a TU Delft start up grant.

Supplementary information

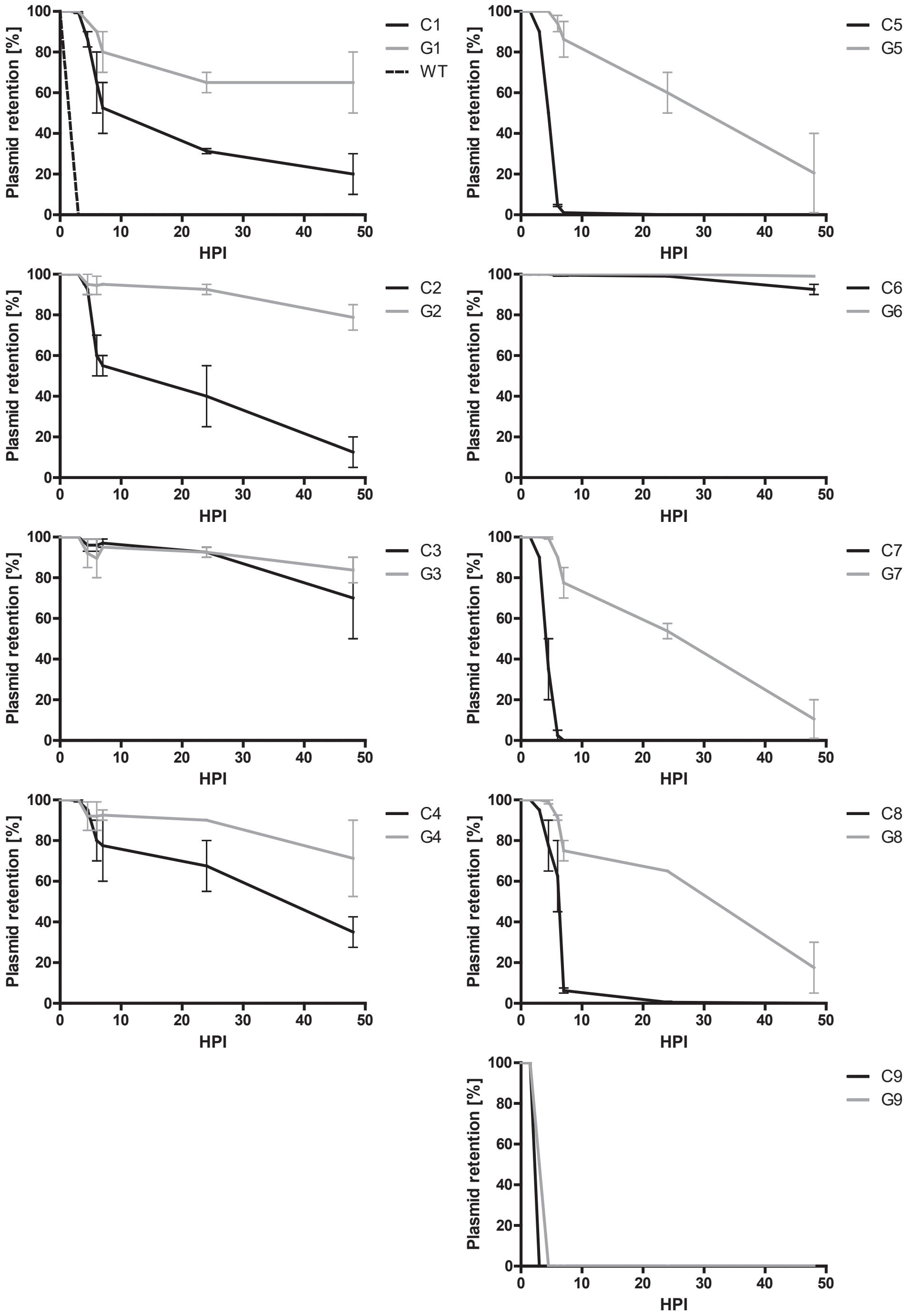


Figure S1 – Individual plasmid loss assays. Individual plasmid loss graphs of C/G mutant pairs. Graphs show standard error of the mean (SEM) of two independent experiments

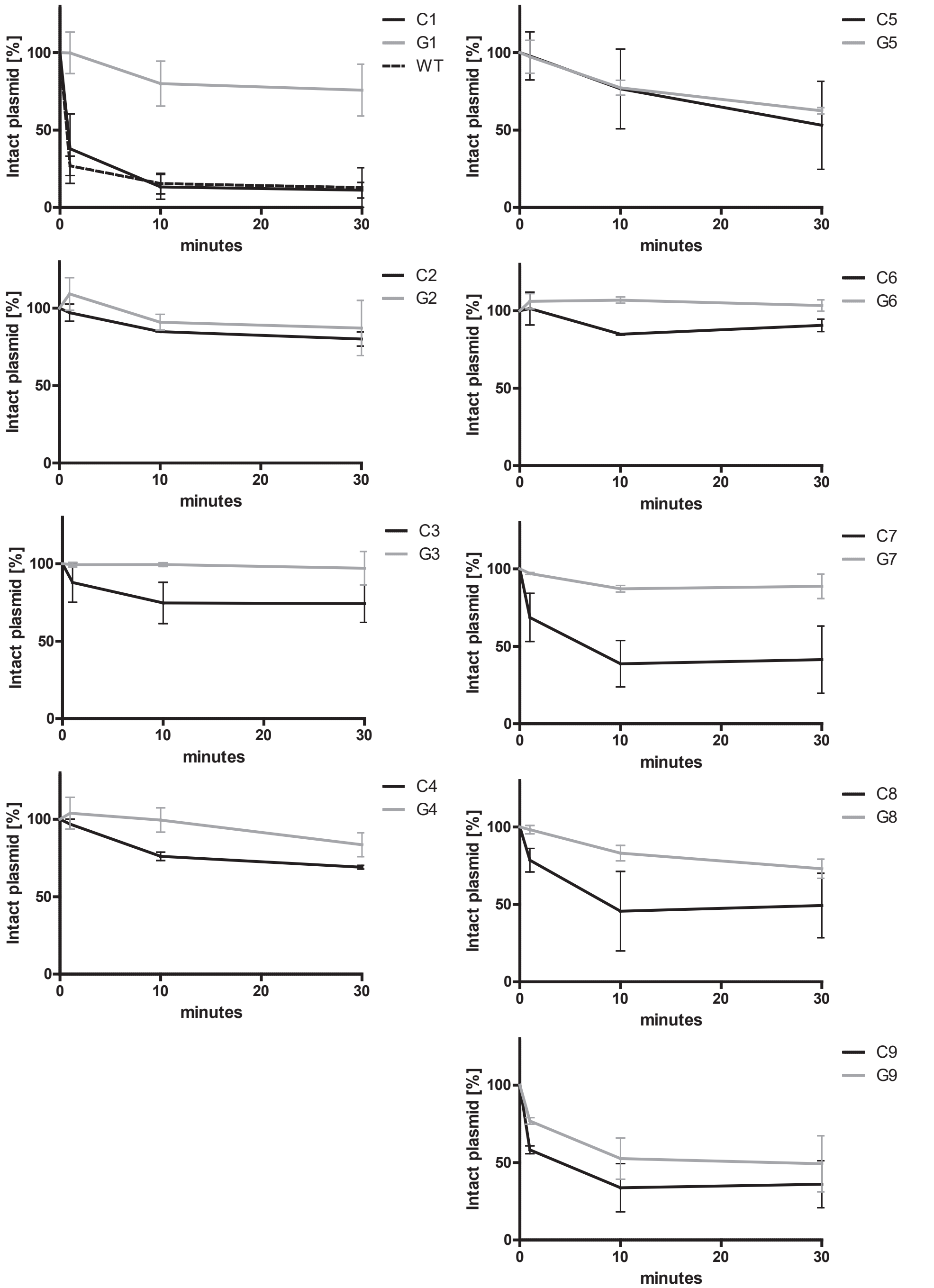


Figure S2 – Individual Cas3 activity assays. Individual graphs of C/G mutant pairs showing plasmid degradation by Cas3. The standard error of the mean of two independent experiments is indicated.

Table S1: Detailed list of tested mutants

Mutant	Target-strand sequence (Protospacer-PAM)																																		
	A	A	A	A	G	T	G	C	C	A	C	T	T	G	C	G	G	A	G	A	C	C	C	G	G	T	C	G	T	C	A	G	C	T	T
C1		C								C								C								C									
C2			A	C																						C					C				
C3	C									C		C																							
C4													C					C		C															
G1		G								G								G								G									
G2			A	G																						G					G				
G3	G									G		G																							
G4													G					G		G															
G5	G											G	G		G			G																	
G6				G					G									G		G						G									
G7		G				G														G	A								G						
G8				G																G						G									
G9						G						G														G									
C5	C											C	C		C			C																	
C6				C					C									C		C						C									
C7		C				C														C	A								C						
C8				C																C						C									
C9						C						C														C									

*grey columns represent pinch point positions that do not base pair and are ignored

Table S2: Oligo nucleotides used in this study

Name	Sequence	Description
BG5301	AAGGTTGGTGGGTTGTTTTTATGG	Primer fw for CRISPR array PCR
BG5302	GGATCGTCACCCTCAGCAGCG	Primer rv for CRISPR array PCR
BG6577	ATGTCATTGCGCTGCCATTC	Sequencing primer for protospacers (StrepR)
BG7157	TTTGAGCTCTTATTTGCCGACTACCTTGGTGATCTC	StrepR fw SacI
BG7395	TTTTCTAGAAAAAGTGCCACTTGCGGAGACCCGGTCGTCAGCTTACATTCAAATATGTATCCGCTC	PCR primer StrepR+sp8 protospacer
BG7396	TTTTCTAGAAAAAGTGCCACTTGCGGAGACCCGGTCGTCAGCGTACATTCAAATATGTATCCGCTC	PCR primer StrepR+mutPAMsp8 protospacer
BG7397	TTTTCTAGACAACCGGGGCAGATCATTAGTTGCACCGCGATTCGACATTCAAATATGTATCCGCTC	PCR primer StrepR+scrambled protospacer
BG8393	TTTTCTAGAaagctgacgGccgggtcGccgcaagGggcacttGtACATTCAAATATGTATCCGCTC	StrepR rv XbaI + mutC1 sp8 protospacer
BG8394	TTTTCTAGAaagcGgacgGccgggtctccgcaagtggcacGAttACATTCAAATATGTATCCGCTC	StrepR rv XbaI + mutC2 sp8 protospacer
BG8395	TTTTCTAGAaagctgacgaccgggtctccgcTGgGggcactttGACATTCAAATATGTATCCGCTC	StrepR rv XbaI + mutC3 sp8 protospacer
BG8396	TTTTCTAGAaagctgacgaccgggGcGccgcGagtggcacttttACATTCAAATATGTATCCGCTC	StrepR rv XbaI + mutC4 sp8 protospacer
BG8397	TTTTCTAGAaagctgacgCccgggtcCccgcaagCggcacttCtACATTCAAATATGTATCCGCTC	StrepR rv XbaI + mutG1 sp8 protospacer
BG8398	TTTTCTAGAaagcCgacgCccgggtctccgcaagtggcacCAttACATTCAAATATGTATCCGCTC	StrepR rv XbaI + mutG2 sp8 protospacer
BG8399	TTTTCTAGAaagctgacgaccgggtctccgcTCgCggcactttCACATTCAAATATGTATCCGCTC	StrepR rv XbaI + mutG3 sp8 protospacer
BG8400	TTTTCTAGAaagctgacgaccgggCcCccgcCagtggcacttttACATTCAAATATGTATCCGCTC	StrepR rv XbaI + mutG4 sp8 protospacer
BG8401	TTTTCTAGAaagctgacgaccgggtcCccCcCCgtggcactttCACATTCAAATATGTATCCGCTC	StrepR rv XbaI + mutG5 sp8 protospacer
BG8402	TTTTCTAGAaagctgacgCccgggCcCccgcaagtCgcacCtttACATTCAAATATGTATCCGCTC	StrepR rv XbaI + mutG6 sp8 protospacer
BG8403	TTTTCTAGAaagctgCcgaccggACctccgcaagtggcCcttCtACATTCAAATATGTATCCGCTC	StrepR rv XbaI + mutG7 sp8 protospacer
BG8404	TTTTCTAGAaagctgacgCccgggCctccgcaagtggcacCtttACATTCAAATATGTATCCGCTC	StrepR rv XbaI + mutG8 sp8 protospacer
BG8405	TTTTCTAGAaagctgacgCccgggtctccgcaCgtggcCcttttACATTCAAATATGTATCCGCTC	StrepR rv XbaI + mutG9 sp8 protospacer
BG8406	TTTTCTAGAaagctgacgaccgggtcGccGcGGgtggcactttGACATTCAAATATGTATCCGCTC	StrepR rv XbaI + mutC5 sp8 protospacer
BG8407	TTTTCTAGAaagctgacgGccgggGcGccgcaagtGgcacGtttACATTCAAATATGTATCCGCTC	StrepR rv XbaI + mutC6 sp8 protospacer
BG8408	TTTTCTAGAaagctgGcgaccggAGctccgcaagtggcGcttGtACATTCAAATATGTATCCGCTC	StrepR rv XbaI + mutC7 sp8 protospacer
BG8409	TTTTCTAGAaagctgacgGccgggGctccgcaagtggcacGtttACATTCAAATATGTATCCGCTC	StrepR rv XbaI + mutC8 sp8 protospacer
BG8410	TTTTCTAGAaagctgacgGccgggtctccgcaGgtggcGcttttACATTCAAATATGTATCCGCTC	StrepR rv XbaI + mutC9 sp8 protospacer
BG9069	ACTCCAAGCGACCGGAGGCTTTTGACTTGATCGGCACGTAAGAGTCTAGAaagctgCcgaccggACctccgcaagtggcCcttCtACATTCAAATATGTATCCGCTCATGAGACAATAACCCTGATAAATGCTTC	mutG7 sp8 non-target strand
BG9070	GAAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATGTaGaagGgccacttgcggagGTccggtcgGcagcttTCTAGACTCTTACGTGCCGATCAAGTCAAAAGCCTCCGGTCGCTTGGAGT	mutG7 sp8 target strand

BG9071	ACTCCAAGCGACCGGAGGCTTTTGACTTGATCGGCACGTAAGAGTCTAGAaagctgGcgaccggAGctccgcaagtggcGcttGtACATTCAAATATGTATCCGCTCATGAGACAATAACCCTGATAAATGCTTC	mutC7 sp8 non-target strand
BG9072	GAAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATGTaCaagCgccacttgcggagCTccggtcgCcagcttTCTAGACTCTTACGTGCCGATCAAGTCAAAAGCCTCCGGTCGCTTGGAGT	mutC7 sp8 target strand
BG9073	ACTCCAAGCGACCGGAGGCTTTTGACTTGATCGGCACGTAAGAGTCTAGAaagctgacgaccgggtctccgcaagtggcacttttACATTCAAATATGTATCCGCTCATGAGACAATAACCCTGATAAATGCTTC	Sp8-WT non-target strand
BG9074	GAAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATGTaaaagtgccacttgcggagacccggtcgtcagcttTCTAGACTCTTACGTGCCGATCAAGTCAAAAGCCTCCGGTCGCTTGGAGT	Sp8-WT target strand
BG9141	GAAGCATTTATCAGGGTTATTG	pBex_50nt_flank-protospacer_fw
BG9142	AAAGCCTCCGACCGGAGG	pBex_50nt_flank-protospacer_rv

Table S3: Plasmids used in this study

Plasmid	Description	Name in paper	source
pWUR925	Rhamnose inducible promotor – RFP, ColE1, KanR	pBEX05	(Lee et al., 2011)
pWUR926	pBex05 backbone, sp8 protospacer, StrepR	pBEX06	This study
pWUR927	pBex05 backbone, mutPAM sp8 protospacer, StrepR	pBEX07	This study
pWUR928	pBex05 backbone, scrambled protospacer, StrepR	pBEX08	This study
pWUR929	pBex05 backbone, mutC1 sp8 protospacer, StrepR	pBEX-C1	This study
pWUR930	pBex05 backbone, mutC2 sp8 protospacer, StrepR	pBEX-C2	This study
pWUR931	pBex05 backbone, mutC3 sp8 protospacer, StrepR	pBEX-C3	This study
pWUR932	pBex05 backbone, mutC4 sp8 protospacer, StrepR	pBEX-C4	This study
pWUR933	pBex05 backbone, mutG1 sp8 protospacer, StrepR	pBEX-G1	This study
pWUR934	pBex05 backbone, mutG2 sp8 protospacer, StrepR	pBEX-G2	This study
pWUR935	pBex05 backbone, mutG3 sp8 protospacer, StrepR	pBEX-G3	This study
pWUR936	pBex05 backbone, mutG4 sp8 protospacer, StrepR	pBEX-G4	This study
pWUR937	pBex05 backbone, mutG5 sp8 protospacer, StrepR	pBEX-G5	This study
pWUR938	pBex05 backbone, mutG6 sp8 protospacer, StrepR	pBEX-G6	This study
pWUR939	pBex05 backbone, mutG7 sp8 protospacer, StrepR	pBEX-G7	This study
pWUR940	pBex05 backbone, mutG8 sp8 protospacer, StrepR	pBEX-G8	This study
pWUR941	pBex05 backbone, mutG9 sp8 protospacer, StrepR	pBEX-G9	This study
pWUR942	pBex05 backbone, mutC5 sp8 protospacer, StrepR	pBEX-C5	This study
pWUR943	pBex05 backbone, mutC6 sp8 protospacer, StrepR	pBEX-C6	This study
pWUR944	pBex05 backbone, mutC7 sp8 protospacer, StrepR	pBEX-C7	This study
pWUR945	pBex05 backbone, mutC8 sp8 protospacer, StrepR	pBEX-C8	This study
pWUR946	pBex05 backbone, mutC9 sp8 protospacer, StrepR	pBEX-C9	This study
pWUR748	pMat11-MBP-Cas3		(Mulepati and Bailey, 2013)
pWUR868	pACYC poly spacer8 CRISPR array		(Künne et al., 2016)
pWUR514	cse2 with Strep-tag II (N-term)-cas7-cas5-cas6e in pET52b		(Jore et al., 2011a)
pWUR408	cse1 in pRSF-1b, no tags		(Brouns et al., 2008b)

Chapter 7

Development of Cascade into a programmable RNA-guided nuclease

Tim Künne[1], Lieuwe Biewenga[1,3], Sebastian N. Kieper[1,2], Selma van Esveld[1,4], Emiel B.J. ten Buren[1,5], John van der Oost[1] and Stan J.J. Brouns[1,2]

[1]Laboratory of Microbiology, Wageningen University, 6708 WE Wageningen, The Netherlands.
[2]Current address: Department of Bionanoscience, Kavli Institute of Nanoscience, Delft University of Technology, 2629 HZ Delft, The Netherlands.
[3]Current address: Department of Chemical and Pharmaceutical Biology, Groningen Research Institute of Pharmacy, University of Groningen, Groningen, The Netherlands
[4]Current address: Radboud University Medical Centre, CMBI, 6525GA Nijmegen, The Netherlands.
[5]Current address: Institute of Laboratory Animal Science , University of Zurich, 8952 Schlieren, Switzerland

Abstract

The development of genome editing tools has made major leaps in the last decade. Recently, RNA guided endonucleases (RGENs) such as Cas9 and Cpf1 have revolutionized genome editing. These RGENs are the hallmark proteins of class 2 CRISPR-Cas systems. In this study we have explored the possibility to develop a new genome editing tool that makes use of the class 1 CRISPR-associated complex for antiviral defense (Cascade) from *E. coli*. This RNA guided protein complex is fused to a FokI nuclease domain to sequence-specifically cleave DNA. We validate the tool *in vitro* using purified protein and two sets of guide RNAs, showing specific cleavage activity. The tool requires two target sites of 32 nt each at a distance of 30-40 nt and inward facing three nucleotide flexible PAM sequences. Like class 1 systems, the guide RNA sequence can easily be designed. Furthermore, we show that an additional RFP can be fused to FokI-Cascade, allowing visualization of the complex in target cells.

Keywords

CRISPR-Cas; genome editing, Cas9

Introduction

The ability to edit genomes in a precise and controlled manner is of high importance for both fundamental biological research as well as applications such as therapy of genetic diseases and crop improvement (Ma et al., 2016; Paul and Qi, 2016; Porteus, 2016). Genome editing harnesses targetable nucleases with tailored specificity to induce a double-stranded break (DSB) at a DNA target site. In a wide variety of cell types, these DSBs are then usually repaired via either one of two predominant endogenous repair pathways: non-homologous end-joining (NHEJ) or homology-directed repair (HDR) (Kim, 2016). Repair via the error-prone NHEJ pathway usually results in insertions or deletions (indels) of varying size, which is convenient when aiming for inactivation of genes or disruption of transcription factor binding sites (Weterings and Chen, 2008). HDR-mediated repair can be utilized for seamless introduction of a piece of DNA when it is exogenously supplied as donor template. This has multiple applications like the introduction of precise point mutations or desired coding sequences (Tóth et al., 2016). Targetable nucleases generally consist of two components: a target DNA binding module and an dsDNA cleavage (endonuclease) module. Until recently, the repertoire of genome editing tools consisted of proteins that bind DNA via protein-DNA interactions. These include the natural meganucleases (homing enzymes), and the synthetic zinc finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs) (Kim, 2016). Using protein engineering, the latter two can be customized to recognize any DNA sequence, however this process is complex and laborious. ZFNs have the additional problem of considerable levels of cytotoxicity, likely caused by off-target cleavage (Kim et al., 2009). Recently, the field of genome editing has been revolutionized by the introduction of RNA-guided endonucleases (RGENs), most prominently Cas9 from *Streptococcus pyogenes* (Cho et al., 2013; Cong et al., 2013; Jinek et al., 2013), but also Cpf1 from *Acidaminococcus* sp. (Zetsche et al., 2015). RGENs are nucleases that make use of a guide RNA molecule that is used to identify *bona fide* target sequences via RNA-DNA base-pairing. Cleavage occurs after full base-pairing and formation of an R-loop structure. Only few mismatches between guide RNA and DNA target are tolerated, leading to highly specific cleavage and low cytotoxicity. Cas9 is part of class 2, type II CRISPR-Cas systems (Clustered Regularly Interspaced Short Palindromic Repeats, CRISPR-associated), which are found almost exclusively in bacteria (Makarova et al., 2015). CRISPR-Cas systems are best known for their function as an adaptive immune system, although recent studies suggest a number of non-defence roles (Louwen et al., 2013; Louwen et al., 2014; Vercoe et al., 2013; Westra et al., 2014). The CRISPR-Cas system provides resistance to mobile genetic elements by storing short memory sequences of invaders in the CRISPR array on the genome (Mohanraju et al., 2016). These short sequences can be transcribed into pre-crRNA, and after processing to mature crRNA guides assembled with Cas proteins to form CRISPR-ribonucleoprotein (crRNP) surveillance complexes. These complexes can either be a single protein (class 2 systems) or a multiprotein complex (class 1 systems) (Makarova et al., 2015). Class 2 surveillance complexes, such as Cas9, contain in almost all cases nuclease domains that are capable of cleaving the target, making them ready-to-use RGENs. Class 1 systems are most prominently represented by type I and type III systems, the latter of which also contain surveillance complexes with nuclease functionality. Type I complexes, however, do not possess nuclease functionality. Instead, they are dependent on the recruitment of nucleases such as Cas3 to degrade the target after binding.

In this study we developed a new type of RGEN, based on the type I-E CRISPR associated complex for antiviral defence (Cascade) protein complex (Brouns et al., 2008a). Cascade is a ribonucleoprotein complex, consisting of 6 different subunits in uneven stoichiometry (Cse1_1, Cse2_2, Cas7_6, Cas5_1, Cas6e_1) and a single 61 nt CRISPR RNA (crRNA). The crRNA consist of a 32 nt spacer sequence (targeting sequence) flanked by an 8 nt 5’ handle and a 21 nt 3’ stem loop (Jore et al., 2011a). In CRISPR immunity, Cascade plays a similar role to Cas9, it is responsible for finding and binding invader DNA. However, unlike Cas9, Cascade does not possess any nuclease functionality to destroy the invader DNA, but recruits an external nuclease, Cas3, to degrade the target (Westra et al., 2012c). To turn Cascade into an RGEN, we made use of the well-established FokI nuclease domain, by fusing it to the Cse1 subunit of Cascade. To allow for generating a DSB in target DNA, two FokI-Cascade complexes are required to bind in close proximity via their guide-RNA to allow for dimerization of the FokI monomers. This approach is reminiscent of ZFNs, TALENs and the FokI-Cas9 system, and promises high specificities.

Results

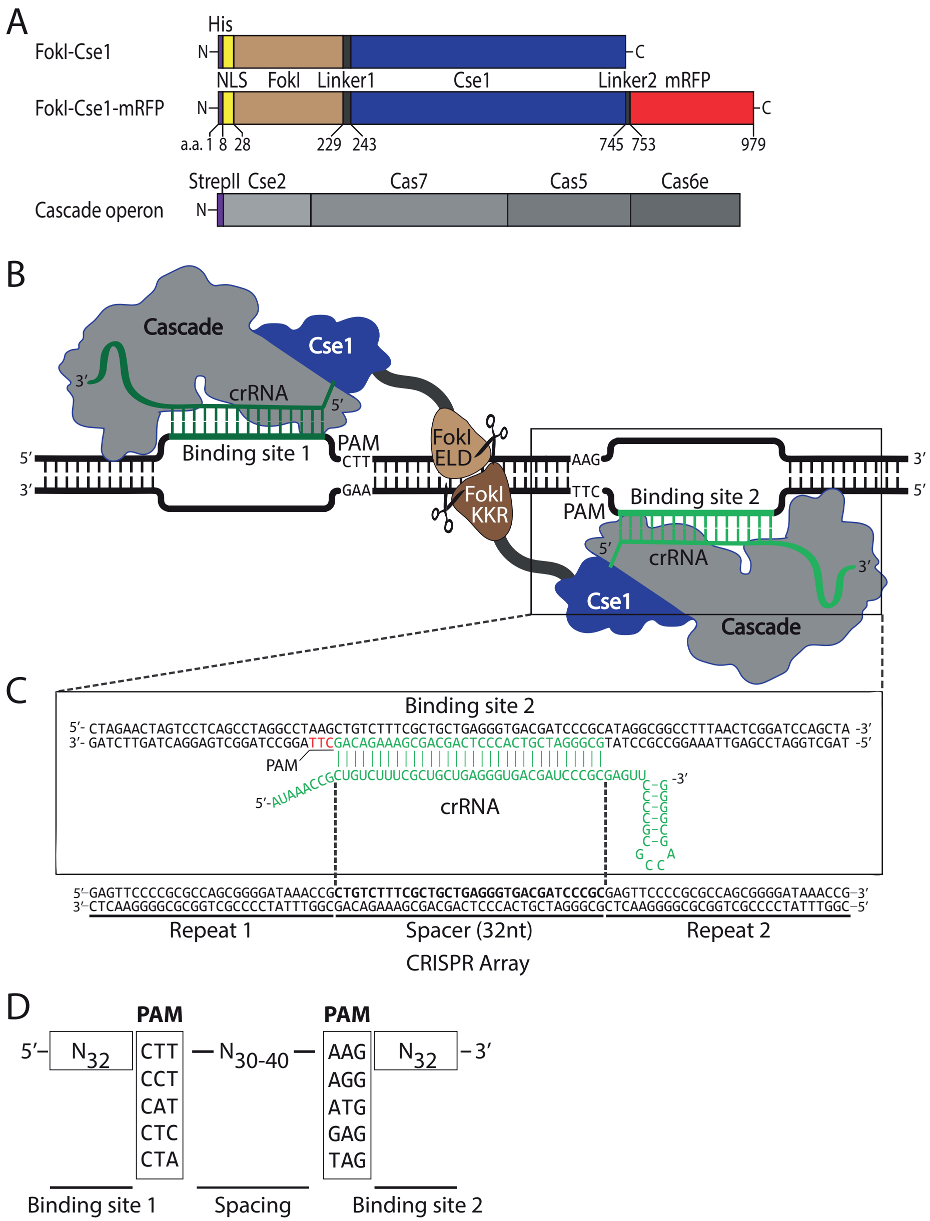


Figure 1 – Design and strategy of FokI-Cascade genome editing. A) Schematic representation of *fokI-cse1* (top), *fokI-cse1-mRFP* (middle) and tagged Cascade operon (bottom) constructs. Top and middle, from left to right: 6x Histidine tag (HIS), dual monopartite SV40 nuclear localization signal (NLS), *fokI* nuclease domain (Sharkey and KKR/ELD mutations), linker from natural *cse1-cas3* fusion in *Streptomyces griseus* (linker1), *cse1* gene from *E. coli*, glycine or proline linker (linker2), *mRFP1* gene. Amino acid positions of individual parts are indicated below. Bottom from left to right: StrepII-tag, cas operon consisting of cse2, Cas7, Cas5, Cas6e. B) Schematic of DNA cleavage strategy. Two FokI-Cascade complexes with negative (ELD) and positive (KKR) dimerization interface bind on opposite

DNA strands, facing each other with their FokI domains. DNA binding is achieved by base-pairing of the crRNA with the target DNA strand and PAM recognition. The flexible linker allows dimerization of the FokI monomers resulting in cleavage of the DNA. C) Sequence representation of DNA binding site 2 with corresponding crRNA (in green) and PAM (in red). Below is a minimal CRISPR array with one spacer and two repeats, encoding for the crRNA. D) Summary of target site criteria. Two target sites of 32 nt each have to be chosen, both need to be flanked by one of the permitted PAMs (facing each other). PAMs are listed by order of preference (top=best). The distance between the target sites can be between 30-40 nt.

Design of the tool

To develop a novel RNA guided nuclease, the well-established FokI nuclease domain was fused to the N-terminus of the Cse1 subunit of Cascade (Figure 1A). For this, we made use of a naturally occurring linker that connects Cse1 with Cas3 in *Streptomyces griseus* and has been previously applied in *E. coli* to fuse Cse1 with Cas3 (Westra et al 2012). Fusing the FokI nuclease domain to a DNA binding protein has been widely used with ZFNs, TALENs and recently Cas9 (REF). This strategy depends on the dimerization of two opposing FokI nuclease monomers which are only active as a dimer (Figure 1B). We used 'sharkey' mutants of the FokI nuclease domain which have enhanced activity (Guo et al., 2010). In addition, the two FokI domains carry opposite charges in their dimerization interface (noted as FokI^{KKR}-Cascade and FokI^{ELD}-Cascade), requiring hetero-dimerization for activity, which has been shown to reduce off-target activity (Doyon et al., 2011). We included an N-terminal HIS tag and a dual monopartite SV40 nuclear localization signal (NLS) on FokI for flexibility with affinity purification and efficient nuclear localization, respectively. As Cas9, FokI-Cascade can be reprogrammed to target any sequence by changing its short crRNA sequence (Brouns et al., 2008a). Target sequences can be any 32 nt sequences that are flanked by a three nt PAM sequence. To achieve a double stranded break, two target sequences on opposite strands with PAM flanks facing each other have to be chosen (Figure 1B). Cascade tolerates a set of 5 different PAMs with relatively high efficiency, allowing flexibility in the choice of target site (Figure 1D). For application, the two individual FokI-Cascade complexes are expressed separately in *E. coli* and purified. There are separate expression cassettes for *fokI-cse1*, the tagged *cas* operon and the CRISPR array. The CRISPR array contains four identical spacers which serve as the guide RNA sequences (Figure 1D) and the crRNA-expression plasmid can simply be exchanged in order to obtain complexes targeting the desired sequence. After purification the active protein can be used *in vitro,* or injected or transfected (via electroporation or iTOP technology (D'Astolfo et al., 2015)) into the cell type of choice.

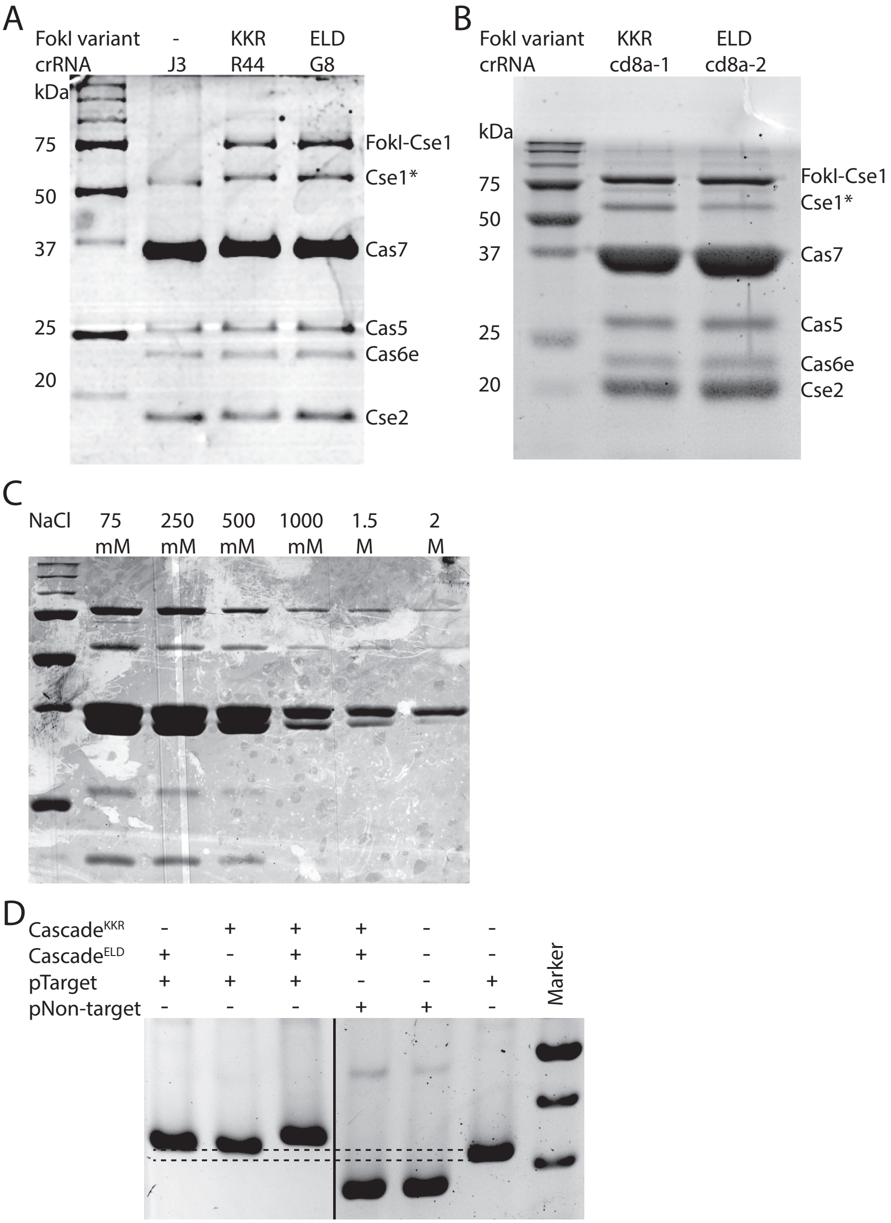


Figure 2 – Fusion protein purification and DNA binding. A) SDS PAGE showing the stoichiometry of native Cascade with J3 crRNA, FokI^{KKR} -Cascade with R44 crRNA and FokI^{ELD} -Cascade with G8 crRNA after single step affinity purification using Streptactin. Bands in native Cascade are from top to bottom: Cse1, Cas7, Cas5, Cas6e, Cse2. FokI-Cascades show the FokI-Cse1 fusion band and an additional band representing Cse1 with a small part of FokI as a result of proteolytic degradation. B) SDS PAGE of FokI-Cascade complexes with cd8a targeting crRNAs. The complexes show minimal proteolytic degradation of FokI-Cse1. C) SDS PAGE showing FokI-Cascade elutions after washing with increasing salt concentrations (75 mM – 2 M). Although lower protein bands are not visible at high salt concentrations, the presence of Cas7 and FokI-Cse1 in the elution confirms that complexes are intact, since the streptag is on Cse2. D) Electrophoretic mobility shift assay (EMSA) of FokI-Cascade

complexes with target and non-target plasmid (molar ratio 20:1 protein:DNA). Binding of a plasmid by FokI-Cascade causes an upward shift in the gel. Single complexes result in upward shift of target DNA, while both complexes result in a double shift of target DNA. Non-target DNA is not shifted.

Protein complex purification

The FokI$^{KKR/ELD}$-Cascade complexes with their respective crRNAs were separately produced in *E. coli* and isolated using one step affinity purification. We made use of the StrepII-tag on the Cse2 subunit of Cascade, which has proven to yield high purity Cascade in earlier studies (Jore et al., 2011a; Westra et al., 2012c). SDS PAGE analysis shows a high purity as well as correct subunit stoichiometry of the complex (Figure 2A). However, next to the expected band for FokI-Cse1, a band is visible that runs slightly above the Cse1 protein of the native Cascade control. Mass spectrometry analysis shows this to be Cse1 with the linker and part of the C-terminal end of FokI (Figure S1). The ratio of intact and shortened fusion protein varied between purifications, but optimized expression and isolation resulted in 65-75% of intact FokI-Cse1 (Figure 2B). Degradation of FokI-Cse1 continues after purification at 37 °C, leading to near complete cleavage of the fusion complex within a few hours (data not shown). Protein was therefore snap frozen in liquid nitrogen and stored at -80 °C. Unfortunately, we were not yet successful in utilizing the His-tag on FokI to perform a two-step purification that should assure only fully intact complexes to be purified.

Since Cascade is a multiprotein complex, we questioned if it remains intact at elevated salt concentrations. High salt concentrations are often used during purification to achieve higher purities. Furthermore, a novel method of protein transfection into mammalian cells (iTOP) exposes the protein to NaCl concentrations above 1 M (D'Astolfo et al., 2015). To test the salt stability of the complex, we used increasing salt concentrations during the washing step of the affinity purification. Increasing salt concentrations resulted in a decrease of overall protein yield from 500 mM NaCl onwards, however, the subunit stoichiometry remained constant (Figure 2C). This shows that the complexes stay intact even at very high salt concentrations (2 M) and that the lower yield is likely caused by a lower binding strength of the strep tag at elevated salt concentrations.

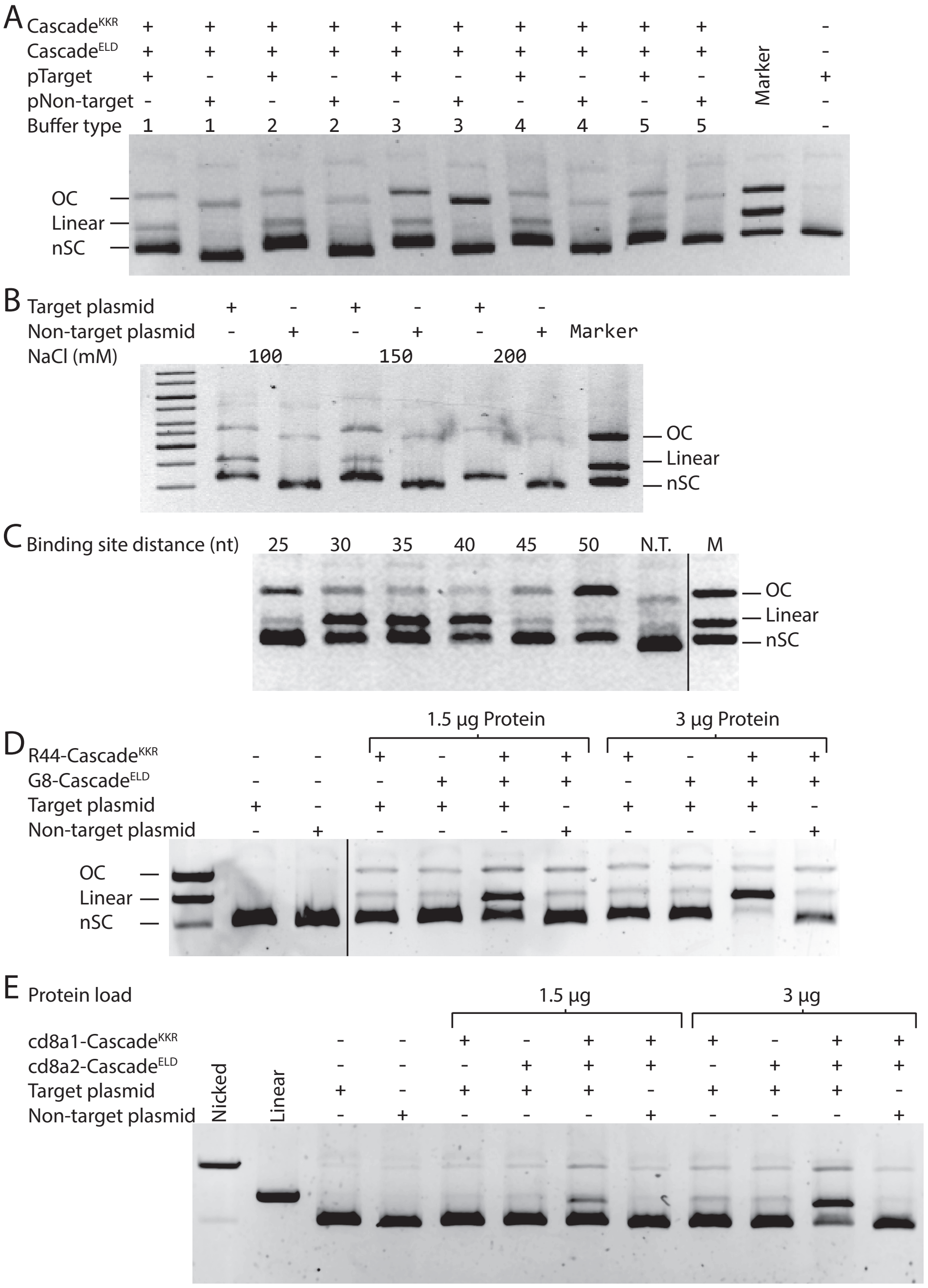


Figure 3 – In vitro activity assays. FokI-Cascade complex activity is tested on plasmid DNA. Unless otherwise stated reactions are run for 30 minutes at 37 °C in a buffer containing 50mM Tris-HCl (pH 7.5), 100 mM NaCl, 10 mM $MgCl_2$ and 0.1 mg/ml BSA. For reference, a marker is included consisting of the three possible plasmid topologies (from top to bottom: Open circular (OC), linear, negatively supercoiled (nSC)). Vertical black lines indicate removal of gel lanes. A) Comparison of 5 buffers for optimal cleavage activity. Buffer 1 (10 mM Tris-HCl (pH 7.5), 10 mM $MgCl_2$, 50 mM NaCl, 0.1 mg/

ml BSA), buffer 2 (50 mM Tris-HCl (pH 7.5), 10 mM MgCl2, 100 mM NaCl, 0.1 mg/ml BSA), buffer 3 (10 mM Tris-HCl (pH 7.5 at 37°C), 10 mM MgCl2, 0.1 mg/ml BSA), buffer 4 (10 mM Tris-HCl (pH 8.5), 10 mM MgCl2, 100 mM KCl, 0.1 mg/ml BSA), buffer 5 (50 mM KAc, 20 mM Tris-Ac (pH 7.9), 10 mM MgAc, 1 mM DTT). B) pTarget35 is incubated with FokI$^{KKR/ELD}$-Cascade at a range of salt concentrations as indicated. C) pTarget with different distances between binding sites (25-50 nt) is incubated with FokI$^{KKR/ELD}$-Cascade. D) pTarget35 is incubated with different amounts of either or with both FokI$^{KKR/ELD}$-Cascade complexes. A non-target plasmid is incubated with both complexes. E) Activity assay of cd8a targeting complexes.

In vitro cleavage activity

To assess the functionality of the FokI-Cascade complexes, we performed *in vitro* plasmid binding and cleavage assays. To this end, plasmids were created carrying the G8 and R44 protospacer at different distances (25-50 nt). These plasmids were incubated with FokI$^{KKR/ELD}$-Cascade carrying anti R44 and G8 crRNA respectively. Plasmids containing only one or no binding sites served as controls.

DNA binding is not affected by FokI-fusion

Initially, we tested binding of the complexes to plasmid DNA. This was done by omitting Mg^{2+} ions from the incubations, preventing DNA cleavage but allowing binding of the FokI-Cascade complexes to the plasmids. We tested the binding using electrophoretic mobility shift assays (EMSA) using similar protein to DNA molar ratios that lead to full DNA binding with native Cascade complexes (Künne et al., 2016; Künne et al., 2015; Westra et al., 2012c). Either complex alone caused a single upwards shift of the target DNA, while adding both complexes caused a double upwards shift (Figure 2D). Non-target DNA was not shifted. This shows that the addition of the FokI domain and peptide linker do not interfere with the DNA binding ability of the Cascade complex.

FokI-Cascade cleaves specifically at physiological conditions

First we determined the spectrum of reaction buffers that result in the best cleavage performance. To this end we tested a range of commercial restriction enzyme buffers (Figure 3A). All tested buffers resulted in the generation of double-stranded breaks in a target plasmid, with only slightly different efficiencies. However, the buffers resulted in different degrees of non-specific nicking of a non-target plasmid. Especially Buffer 3 resulted in strong non-specific nicking activity. Buffer 3 has the lowest salt concentration and thereby likely promotes non-specific weak binding of the FokI-Cascade complexes to DNA. To test whether an increase in salt concentration can further improve specificity of the system, we performed reactions with Buffer 2 (100 mM NaCl) and increased the salt concentration to 150 and 200 mM (Figure 3B). Increasing the salt concentration had only a mildly positive effect on the specificity but also reduced the overall cleavage activity on the target plasmid. No cleavage was observed at 200 mM. FokI-Cascade is therefore able to cleave specifically in salt concentrations between 50 mM (Buffer 1) and 150 mM. This is well in agreement with physiological salt concentrations in most cell types and suggests that the system can produce double stranded breaks at physiological conditions with a low risk of off-target cleavage.

FokI-Cascade tolerates flexible target site distances

Next we determined the optimal distance between the two binding sites in the DNA that allow binding of the FokI$^{KKR/ELD}$-Cascade complexes and dimerization of the FokI domains for efficient cleavage. We tested distances from 25 to 50 nt in 5 nt increments (Figure 3C). DSBs are generated efficiently when binding sites were separated by 30, 35 and 40 nucleotides. At 25, 45 and 50 nucleotides, the target plasmid was mostly nicked. A distance of 30-40 nt is thus ideal to allow correct dimerization of the FokI domains and formation of DSBs. Shorter or longer distances likely interfere with correct dimerization or binding of the FokI domains to the DNA, leading to incomplete cleavage. This flexibility in distance between target sites strongly increases the chance to find properly spaced target sites in any given locus.

Ideal conditions result in efficient and highly specific DSB formation

Finally, we tested the system in ideal conditions at different protein concentrations. Incubation with either one of the complexes resulted in negligible activity, while incubation with both complexes efficiently produced double stranded breaks (Figure 3D). In addition, the activity on the non-target plasmid is comparable to reactions with either of the protein complexes, showing a low degree of non-specific activity. Near complete cleavage was achieved after 30 minutes with 3 µg of protein which corresponds to a 36 fold molar excess of protein to DNA. In addition to the R44/G8 targeting complexes we also produced complexes targeting the cd8a gene of zebrafish. These complexes exhibited comparable levels of activity and specificity *in vitro* (Figure 3E).

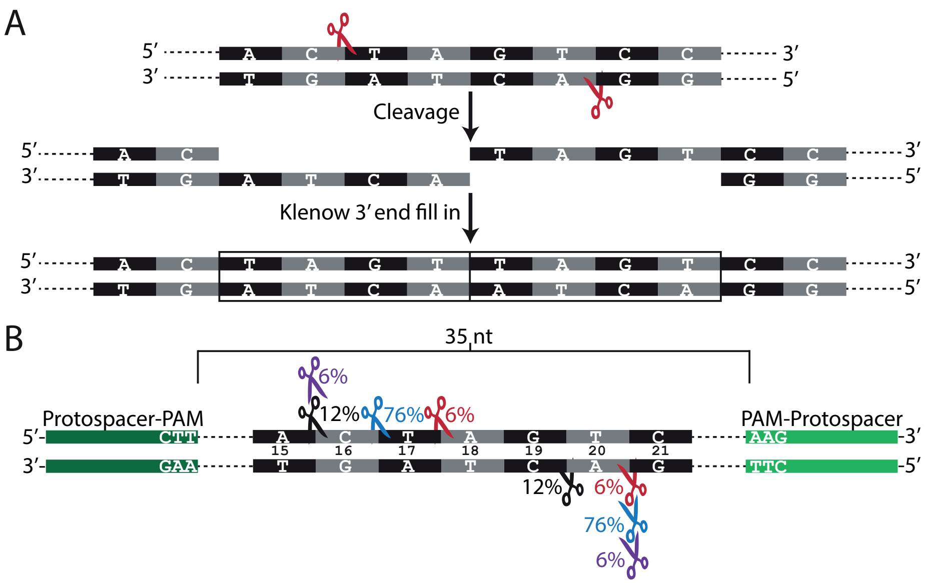
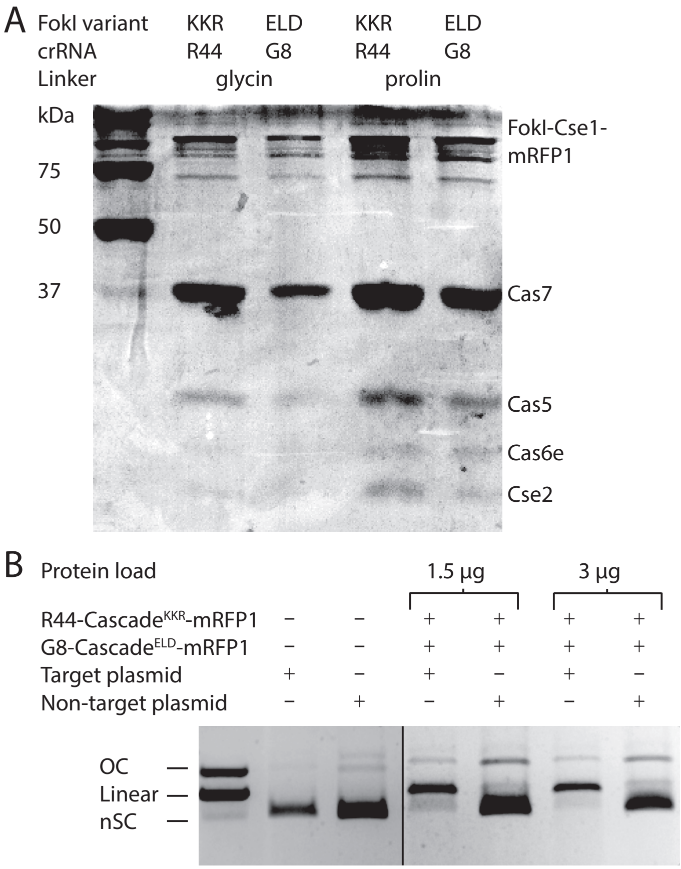


Figure 4 – Cleavage site determination. A) Linear plasmid products after in vitro assay were isolated, ends filled in with Klenow polymerase, religated and cloned. Klenow fill in introduces extra nucleotides, identical to the overhang created by FokI cleavage. B) Cloned plasmids were sequenced and mapped to original sequence to reveal extra nucleotides and thereby the cleavage sites. Identically coloured scissors represent staggered cut sites in both DNA strands with respective frequencies. Numbers in between basepairs represent position in between the two binding sites.

Cleavage site determination

After showing specific cleavage activity of the system, we set out to determine the cleavage site. We chose the target plasmid with a 35 nt spacing and isolated the linear cleavage products from an agarose gel after a typical cleavage reaction. We assumed that FokI cleavage would result in recessed 3' ends as for the native FokI enzyme. We filled in the recessed 3' ends with the Klenow fragment of the *E. coli* DNA polymerase to create blunt ends (Figure 4A). The linear vector was then self-ligated, transformed and sequenced. Filling in of recessed 3' ends and re-ligation will lead to extra nucleotides in the sequence that represent the overhang left by FokI cleavage. Figure 4B shows the original sequence of pTarget35, with indicated top- and bottom-strand cleavage sites. We sequenced seventeen clones and these all showed cleavage around the centre in between the two target sites, creating varying overhangs between 3 and 5 nt. Overhangs of 4 nt are most abundant (cumulatively 88%), while overhangs of 3 nt and 5 nt occur only once (6% each). This shows that all observed plasmid cleavage is likely specific to the target region and that the majority of the cleavage products have the expected 4 nt overhang.



FokI variant	KKR	ELD	KKR	ELD
crRNA	R44	G8	R44	G8
Linker	glycin		prolin	

Protein load			1.5 µg		3 µg	
R44-Cascade^{KKR}-mRFP1	−	−	+	+	+	+
G8-Cascade^{ELD}-mRFP1	−	−	+	+	+	+
Target plasmid	+	−	+	−	+	−
Non-target plasmid	−	+	−	+	−	+

Figure 5 – FokI-Cascade-mRFP1. mRFP1 was fused to the C-terminus of Cse1 via a glycin or prolin linker. A) SDS-PAGE of FokI-Cascade-mRFP1 showing presence of fusion protein as well as partial degradation. B) In vitro activity assay with FokI-Cascade-mRFP1 and plasmid DNA. Target and non-target plasmids have been incubated with both FokI $^{KKR/ELD}$-Cascade-mRFP1 complexes.

mRFP1 fusion for visualization

One of the envisaged applications will be to inject or transfect the protein directly into prokaryotic and eukaryotic cells. Being able to visualize the protein would allow observing and quantifying transfection efficiency, and (in case of eukaryotes) nuclear import of the protein. To achieve this, we fused *mRFP1* to the C-terminus of *cse1*, using either a glycine linker or a proline linker (Figure 1A). Both constructs yielded intact protein complexes with correct subunit stoichiometry (Figure 5A). Furthermore, the complexes retained full activity and specificity (Figure 5B). This fusion is also a proof of principle, showing that fusions of proteins with alternative functionalities can be made to the C-terminus of *Cse1* without disrupting Cascade stoichiometry and function.

Preliminary *in vivo* testing

To test whether our tool is functional *in vivo*, we tested protein injections into zebrafish embryos. We used the cd8a targeting complexes in order to induce indel mutations in the cd8a gene. Furthermore, we used the R44/G8 targeting complexes as a negative control. The protein was injected into the egg yolk of single cell embryos. Initially, we tested three different protein concentrations: 21 ng, 10 ng and 2.5ng. Unfortunately, after the first round of injections all injected embryos died, while non-injected embryos showed normal survival rates. We performed a second round of injections with the buffer flow through of the buffer exchange from the original elution buffer to the injection buffer. We did this to test for toxicity of buffer components or other possible toxic contaminants in our protein preparations. Injections resulted in 80 % mortality, while control injections with pure injection buffer resulted in 40-50 % mortality. While the control injections still result in unexpectedly high mortality, the purification buffer caused a significantly higher mortality. Unfortunately, we were not able to test any further injections and the results remain inconclusive at this point.

Discussion

The rapid development and widespread application of Cas9 has revolutionized genome editing in many areas of research. Other programmable nucleases such as ZFNs and TALENs can target and cleave a specific DNA sequence, however, CRISPR-Cas9/Cpf1 is favored due to the simplicity of design and workflow and its high efficiency and specificity.

In this study, we developed a novel RNA-guided programmable nuclease for eukaryotic genome editing. We made use of the Cascade complex from the type I-E CRISPR-Cas system of *E. coli* (Brouns et al., 2008a; Jore et al., 2011a), and employed the dual FokI strategy similar to ZFNs and TALENs. A set of FokI-Cascade complexes has to bind in close proximity of each other (30-40 nt apart) and the target sites need to be flanked by inward facing PAMs (Figure 1D). Only when these criteria are met, hetero-dimerization of the FokI monomers occurs and the DNA is cleaved. We validated the tool *in vitro* and observed specific cleavage activity with two different pairs of guide RNAs.

The base-pairing segment of the guide RNA of the Cascade complex is 32 nt as compared to 18-20 nt for Cas9. Moreover, Cascade has been shown to bind much more stably to DNA than Cas9 in single molecule studies (Szczelkun et al., 2014). Although Cascade tolerates mutations during CRISPR immunity in *E.coli*, and has positions in the crRNA that are not involved in base pairing (Fineran et al., 2014; Hayes et al., 2016), the binding affinity of Cascade towards mutated targets is strongly reduced (Künne et al., 2016; Semenova et al., 2011; Westra et al., 2012c). Therefore, the use of two target sites with increased length lowers the range of potential off-target sites, likely leading to reduced off-target activity. Cascade tolerates 5 different PAMs for CRISPR interference and therefore high binding affinity, while many other PAMs are bound with reduced affinity (Fineran et al., 2014; Westra et al., 2013; Xue et al., 2015). Whether this promiscuity affects specificity remains to be seen. However, the flexibility in PAM choice and the flexibility in distance between target sites creates a large pool of potential target sites, potentially enabling the accurate editing of any locus on a genome.

The double FokI fusion strategy has recently been successfully applied to Cas9 (Guilinger et al., 2014; Tsai et al., 2014). At the same time, Cas9 has been applied in a double nickase strategy, where two Cas9 monomers nick opposite strands independent of each other (Ran et al., 2013a; Shen et al., 2014). The FokI-Cas9 strategy resulted in a lower frequency of undesired deletions, insertions and base-pair substitutions than other techniques (Tsai et al., 2014). This is likely due to the obligate dimerization, while Cas9 nickases can cleave independently. Additionally, FokI-Cas9 is more stringent towards PAM orientation and target distance to allow for proper dimerization, which reduces the range of potential off-target sites (Ran et al., 2013a). These findings suggest that the dual FokI strategy can be superior to monomer-based approaches even for the appraised Cas9 tool. Taken together, it could therefore be interesting to have tools in addition to Cas9 in the form of other programmable RNA-guided proteins as the DNA binding module.

For practical application in genome editing, the FokI-Cascade system should be expressed and purified from *E. coli*. Purified protein can then be delivered into target cells by a range of techniques, including electroporation, microinjection, protein transfection and induced micropinocytosis (iTOP) (D'Astolfo et al., 2015). Studies on Cas9 have shown that direct introduction of protein into cells instead of expression constructs leads to highly efficiency editing (Kim et al., 2014; Lin et al., 2014; Liu et al., 2015; Ramakrishna et al., 2014; Zuris et al., 2015). Furthermore, they revealed that off-target effects are lower when introducing protein, due to the limited lifetime of protein in the cell and better control over the dosage. Furthermore, introducing pre-assembled complexes prevents the risk of formation of guide free complexes, which in case of Cas9 have been found to cause DNA damage (Van der Oost, personal communication, June 18 2017). While expression constructs will lead to a permanent high level of Cas9 in the cell, which increases the chance for unwanted activity, injected protein will be broken down in the cells more rapidly, reducing the chance for off-target activity. The expression and purification of FokI-Cascade is simple and fast although the proteolytic cleavage of the FokI domain still has to be addressed. An alternative linker sequence might be sufficient to achieve this, since there is no evidence for instability of the FokI domain itself. Minimal

CRISPR arrays carrying the desired guide RNA sequence can be made synthetically, cloned into an expression vector and included for protein expression.

Cascade has previously been shown to be capable of inducing stable, long-lasting and multiplexed gene silencing when binding to the promoter region of a bacterial gene (Chang et al., 2016; Luo et al., 2015; Rath et al., 2014). Furthermore, a type I-A Cascade-like complex from *Sulfolobus islandicus* has been used to target its own genome to select for recombinants after genome editing by homologous recombination (Li et al., 2015). This shows that Cascade is able to stably and lastingly bind to genomic DNA at least in prokaryotes.

Another interesting application for this tool could be *in vitro* DNA assembly and editing. Recently, Cpf1 has been used to develop a new standard for DNA assembly called C-Brick (Li et al., 2016) and a method for efficient editing of large DNA constructs called “Cpf1-assisted cutting and Taq DNA ligase-assisted ligation” (CCTL) (Lei et al., 2017). Both methods make use of the ability of Cpf1 to introduce a staggered cut at any desired target site. The resulting sticky ends enable easy *in vitro* DNA assembly with minimal scars. This method is superior to restriction enzyme based methods as it does not require the removal of extra restriction sites from constructs. Similarly FokI-Cascade could be employed to cleave DNA parts *in vitro*, creating staggered ends for assembly or sub-cloning. One disadvantage, however, would be the need for sufficient flanking sequence on both sides of the desired cut site to allow binding of both FokI-Cascade complexes.

In conclusion, we developed a novel programmable RGEN that introduces staggered cuts with high specificity *in vitro*. Attempts are currently ongoing to see whether this tool can be successfully applied in genome editing or *in vitro* applications.

Materials and Methods

Strains used in this study. *E. coli* BL21-AI/*E. coli* T7-express (protein expression), *E. coli* DH5α (cloning)

Plasmid construction. pWUR811/812 were constructed by replacing the Cas3 sequence in pWUR657 (NcoI/BamHI) (Westra et al., 2012c) with an *E. coli* codon optimized FokI sequence, carrying the sharkey mutations and either the ELD or KKR dimerization interface mutations (Synthesized at GeneArt). Poly His-tag and dual monopartite SV40 NLS were added to make pWUR813/814 by digesting pWUR811/812 with NcoI and ligating annealed oligonucleotides carrying NcoI compatible overhangs (BG4112/4113) in front of the FokI sequence. pWUR815/816 (glycine linker) and pWUR922/923 (proline linker) were made by PCR amplifying the FokI-Cse1 sequence lacking the stop codon (BG4899/5096 or BG4899/5098) and ligating it together with the PCR amplified mRFP1 sequence (BG5097/4970 or BG5099/4970) into pET52b (PscI/SalI/AvrII). pWUR 817-822 were synthesized at GeneArt, the non-target plasmid was made by digesting pT25 with EcoRI and SpeI to remove the target sites, blunting with Klenow polymerase and religation. pWUR918-921 were made by cloning a synthetic CRISPR array with 4 identical spacers (R44, G8, cd8a-1, cd8a-2 respectively) into pACYC (NcoI/XhoI). The cd8a target plasmid was constructed by PCR amplifying the cd8a gene from zebrafish genomic DNA using BG4905/4906 and cloning it into pUC19 (SmaI).

Preparation of Proteins. FokI-Cse1(-mRFP1), the strep-tagged Cas operon and one CRISPR array were co-expressed in fresh transformants of *E. coli* strain BL21 AI or *E. coli* strain T7 express. Complexes form *in vivo* and were subsequently affinity purified. Briefly, cells were collected by centrifugation and re-suspended in ice cold Buffer A (20 mM HEPES (pH 7.5), 75 mM NaCl, 1 mM DTT). Cells were lysed in a pre-cooled French press and kept on ice at all times. The lysate was cleared by centrifugation and filtering (0.45 µm pores). The clear lysate was incubated with strep tactin beads (IBA) at 4 °C for 20 minutes. The beads were collected by gentle centrifugation and loaded onto a gravity column. The beads were washed with Buffer A (then with 300 mM NaCl) and eluted in buffer B (20 mM HEPES (pH 7.5), 75 mM NaCl, 1 mM DTT, 2.5 mM Desthiobiotin). Elutions were pooled and buffer exchanged/concentrated into Buffer A using Amicon Ultra filters (100 kDa MWCO). All proteins were flash frozen in liquid nitrogen and stored at -80 °C.

Proteolytic cleavage site determination. Discrete protein bands were cut from SDS-PAGE and analyzed by MS-MS as previously described (Jore et al., 2011).

In vitro activity assays. Activity assays were performed with the following conditions unless otherwise noted: Reactions were performed in a buffer containing 50 mM Tris-HCl (pH 7.5), 10 mM MgCl2, 100 mM NaCl, 0.1 mg/ml BSA for 30 minutes at 37 °C. 1.5 µg of each FokI-Cascade complex was added to 200ng DNA (20:1 molar ratio protein:DNA). Reactions were stopped on ice and immediately processed by PCI extraction to separate protein and DNA. DNA was analyzed on a 0.8% agarose gel stained with SYBR Safe (Thermo Scientific), the gel was visualized using a UV-imager (Syngene/Bio-Rad).

Cleavage site determination. After regular activity assays, the linearized plasmid fraction was isolated from an agarose gel (GeneJet gel isolation Kit, Thermo Scientific) and treated with the Klenow fragment of DNA polymerase I (Thermo Scientific) according to manufacturer’s protocol to fill in the recessed 3’ ends. Blunt ends were re-ligated using T4 DNA ligase (Thermo Scientific) according to manufacturer’s protocol. Re-circularized plasmids were transformed into *E. coli* and individual clones were sequenced (GATC Biotech). Sequences were mapped to the original plasmid to reveal the extra nucleotides incorporated by the Klenow fragment.

Electrophoretic mobility shift assay. Plasmids and purified protein complexes were incubated for 30 minutes at 37 °C to reach binding equilibrium. Reactions were performed in a buffer containing 20 mM HEPES (pH 7.5), 75 mM NaCl and 1 mM DTT. Samples were run on a native 0.8% agarose gel for 18 h at 20 mA in sodium borate buffer. After electrophoresis the gel was stained with SYBR Safe (Thermo Scientific) for 20 minutes, rinsed with demineralized water and visualized using a UV-imager (Syngene/Bio-Rad).

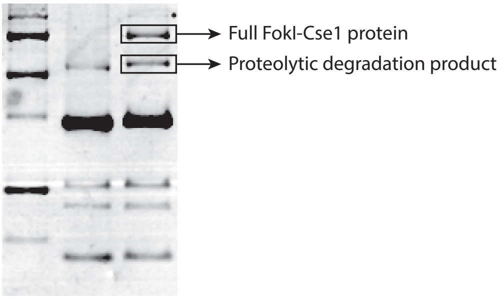
Zebrafish injections. Purified protein was buffer exchanged into egg buffer (60 µg/ml Instant Ocean® Sea Salt) and concentrated using Amicon Ultra filters (100 kDa MWCO). The buffer flow through of the buffer exchange/concentration step was collected and used as a buffer toxicity control for injections. The protein solution was supplemented with phenol red as tracking dye and 1x Tango buffer (Thermo Scientific). 1 nl of the final solution was microinjected into the yolk of single cell zebrafish embryos.

Author contributions

T.K., S.J.J.B and J.vd.O. designed research; T.K., L.B., S.N.K., S.v.E. and E.B.J.t.B. performed research; T.K. and S.J.J.B. analyzed data; and T.K. and S.J.J.B. wrote the paper with input from all authors.

Supplementary Information

A



B

Full FokI-Cse1 protein

FokI

MAQLVKSELEEKKSELRHKLKYVPHEYIELIEIARNPTQDRILEMKVMEFFMKVYGYRGEHLG
GSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMERYVEENQTRDKHLNPNEWW
KVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEV
RRKFNNGEINF**ADPTNRAKGLEAVSVAS**MNLLIDNWIPVRPRNGGKVQIINLQSLYCSRDQ

Cse1

WRLSLPRDDMELAALALLVCIGQIIAPAKDDVEFRHRIMNPLTEDEFQQLIAPWIDMFYLNHA
EHPFMQTKGVKANDVTPMEKLLAGVSGATNCAFVNQPGQGEALCGGCTAIALFNQANQAPG
FGGGFKSGLRGGTPVTTFVRGIDLRSTVLLNVLTLPRLQKQFPNESHTENQPTWIKPIKSNES
IPASSIGFVRGLFWQPAHIELCDPIGIGKCSCCGQESNLRYTGFLKEKFTFTVNGLWPHPHSP
CLVTVKKGEVEEKFLAFTTSAPSWTQISRVVVDKIIQNENGNRVAAVVNQFRNIAPQSPLELI
MGGYRNNQASILERRHDVLMFNQGWQQYGNVINEIVTVGLGYKTALRKALYTFAEGFKNKDF
KGAGVSVHETAERHFYRQSELLIPDVLANVNFSQADEVIADLRDKLHQLCEMLFNQSVAPYA
HHPKLISTLALARATLYKHLRELKPQGGPSNG

Proteolytic degradation product

FokI

MAQLVKSELEEKKSELRHKLKYVPHEYIELIEIARNPTQDRILEMKVMEFFMKVYGYRGEHLG
GSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMERYVEENQTRDKHLNPNEWW
KVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEV
RRKFNNGEINF**ADPTNRAKGLEAVSVAS**MNLLIDNWIPVRPRNGGKVQIINLQSLYCSRDQ

Cse1

WRLSLPRDDMELAALALLVCIGQIIAPAKDDVEFRHRIMNPLTEDEFQQLIAPWIDMFYLNHA
EHPFMQTKGVKANDVTPMEKLLAGVSGATNCAFVNQPGQGEALCGGCTAIALFNQANQAPG
FGGGFKSGLRGGTPVTTFVRGIDLRSTVLLNVLTLPRLQKQFPNESHTENQPTWIKPIKSNES
IPASSIGFVRGLFWQPAHIELCDPIGIGKCSCCGQESNLRYTGFLKEKFTFTVNGLWPHPHSP
CLVTVKKGEVEEKFLAFTTSAPSWTQISRVVVDKIIQNENGNRVAAVVNQFRNIAPQSPLELI
MGGYRNNQASILERRHDVLMFNQGWQQYGNVINEIVTVGLGYKTALRKALYTFAEGFKNKDF
KGAGVSVHETAERHFYRQSELLIPDVLANVNFSQADEVIADLRDKLHQLCEMLFNQSVAPYA
HHPKLISTLALARATLYKHLRELKPQGGPSNG

Figure S1 – MS analysis of FokI-Cse1 fusion protein. The full FokI-Cse1 protein band and the smaller extra band were cut from an SDS PAGE gel and analyzed by MS-MS. Mapping of peptides detected in MS-MS (yellow) on FokI-Cse1 fusion protein sequence. The linker sequence is in bold and underlined.

Table 1: Plasmids used in this study

Plasmid	Backbone	Description	Source
pWUR656	pCDF1b	Cas operon without Cse1 (Cse2-Cas5-Cas7-Cas6e), N-terminal strep-tag on Cse2.	(Jore et al., 2011a)
pWUR811	pWUR657	pET52b-FokI-CseI (sharkey, KKR)	This study
pWUR812	pWUR657	pET52b-FokI-CseI (sharkey, ELD)	This study
pWUR813	pWUR811	pET52b-HIS-NLS-FokI-CseI (KKR) 6xHIS+sv40NLS	This study
pWUR814	pWUR812	pET52b-HIS-NLS-FokI-CseI (ELD) 6xHIS+sv40NLS	This study
pWUR815	pET52b	His-NLS-FokI-Cse1-mRFP1 (KKR, glycine linker)	This study
pWUR816	pET52b	His-NLS-FokI-Cse1-mRFP1 (ELD, glycine linker)	This study
pWUR817	pMA-T	pT25, P7(R44)/M13(G8) target sites, 25nt spacing	This study
pWUR818	pMA-T	pT30, P7(R44)/M13(G8) target sites, 30nt spacing	This study
pWUR819	pMA-T	pT35, P7(R44)/M13(G8) target sites, 35nt spacing	This study
pWUR820	pMA-T	pT40, P7(R44)/M13(G8) target sites, 40nt spacing	This study
pWUR821	pMA-T	pT45, P7(R44)/M13(G8) target sites, 45nt spacing	This study
pWUR822	pMA-T	pT50, P7(R44)/M13(G8) target sites, 50nt spacing	This study
pWUR823	pMA-T	pNon-target, Target sites removed	This study
pWUR918	pACYC	CRISPR array with 4x R44 spacer	This study
pWUR919	pACYC	CRISPR array with 4x G8 spacer	This study
pWUR920	pACYC	CRISPR array with 4x cd8a-1 spacer	This study
pWUR921	pACYC	CRISPR array with 4x cd8a-2 spacer	This study
pWUR922	pET52b	His-NLS-FokI-Cse1-mRFP1 (KKR, proline linker)	This study
pWUR923	pET52b	His-NLS-FokI-Cse1-mRFP1 (ELD, proline linker)	This study
pWUR924	pUC19	Cd8a target plasmid	This study

Table 2: Oligonucleotides used in this study

Name	Sequence	Description
BG4112	CATGcatcaccatcatcaccacCCGAAAAAAAAGCGCAAAGTG-GATCCGAAGAAAAAACGTAAAGTTGAAGATCCGAAAGA	HIS tag+SV40 NLS fw
BG4113	CATGTCTTTCGGATCTTCAACTTTACGTTTTTTCTTCGGATC-CACTTTGCGCTTTTTTTTCGGgtggtgatgatggtgatg	HIS tag+SV40 NLS rv
BG4899	AGTACATGTTGCATCACCATCATCACCACCCGAA	Fok1-Cse1(HIS/NLS) fw PscI
BG4905	TTTATTAGGCATTCAGCATGAAATA	Cd8 target fw
BG4906	TGCAGACATGGTCAGTTTTTCT	Cd8 target rv
BG4970	Gcgcctagggttattaagcaccggtggagtga	mRFP1 rv
BG5096	ccagtcgacccgccgccaccagaGCCATTTGATGGCCCTCCTT	Fok1-Cse1 reverse (glycine-SalI)
BG5097	cgggtcgactggtATGGCTTCCTCCGAAGA	mRFP1 forward (glycine-salI)
BG5098	ggggtcgacggggttggtgtGCCATTTGATGGCCCTCCTTG	Fok1-Cse1 reverse (proline-SalI)
BG5099	cccgtcgaccccaATGGCTTCCTCCGAAGA	mRFP1 forward (proline-salI)

Chapter 8

Thesis summary

Host-pathogen interactions are among the most prevalent and evolutionary important interactions known today. The predation of prokaryotes by their viruses is happening on an especially large scale and had a major influence on the evolutionary history of prokaryotes. Since most viruses are lytic at some point in their life-cycle, there is a high selection pressure for prokaryotes to develop defense mechanisms. As described in **Chapter 1**, the CRISPR-Cas system is a relatively recently discovered defense system and is also the first adaptive defense system discovered in prokaryotes. CRISPR-Cas systems are widespread, occurring in the majority of archaea and also a considerable fraction of bacteria. This diversity is also reflected in the diversity of different types of CRISPR-Cas systems, currently being divided into 6 major types with a large number of subtypes. The type I-E system of *Escherichia coli* is a well-studied model system and of high relevance, since it is a major subtype of type I systems which make up around 50 % of all discovered CRISPR-Cas systems. CRISPR-Cas systems basically comprise the CRISPR array, made up of repeats and foreign derived spacers, and a set of *cas* genes. Immunity is commonly divided into three functional stages, adaptation, expression and interference. Adaptation is the acquisition of new spacers from the foreign nucleic acid and its incorporation into the CRISPR array. During expression, the CRISPR array is transcribed, processed and assembled with Cas proteins into CRISPR RNA (crRNA) guided ribonucleoprotein complexes (crRNP). Interference is the detection, binding and destruction of foreign nucleic acids by the crRNP and in type I systems the Cas3 nuclease. The type I-E system contains another function, called primed adaptation. Primed adaptation is a more rapid and efficient version of regular (naïve) adaptation. In addition to the adaptation machinery, primed adaptation also requires the interference machinery.

Chapter 2 describes and compares a fundamental feature of most, if not all, CRISPR-Cas systems and also many other small RNA based systems. The mode of action of small RNAs relies on protein-assisted base pairing of the guide RNA with target mRNA or DNA to interfere with their transcription, translation or replication. Several unrelated classes of small non-coding RNAs have been identified including eukaryotic RNA silencing associated small RNAs, prokaryotic small regulatory RNAs and prokaryotic CRISPR (clustered regularly interspaced short palindromic repeats) RNAs. All three groups identify their target sequence by base pairing after finding it in a pool of millions of other nucleotide sequences in the cell. In this complicated target search process, a region of 6 to 12 nucleotides of the small RNA termed the 'seed' plays a critical role. The seed is often a structurally pre-ordered region that increases accessibility and lowers the energy barrier of RNA-DNA duplex formation. Furthermore, the length of the seed is optimally chosen to allow rapid probing and also rejection of potential target sites. The seed is a perfect example of parallel evolution, showing that nature comes up with the same strategy independently multiple times.

Chapter 3 provides a description and protocol of the Electrophoretic Mobility Shift Assay (EMSA) and its use for studying crRNPs. EMSA is a straightforward and inexpensive method for the determination and quantification of protein–nucleic acid interactions. It relies on the different mobility of free and protein-bound nucleic acid in a gel matrix during electrophoresis. Nucleic acid affinities of crRNPs can be quantified by calculating the dissociation constant (K_d). Protocols for two types of EMSA assays are described using the Cascade ribonucleoprotein complex from *Escherichia coli* as an example. One protocol uses plasmid DNA as substrate, while the other uses short linear oligonucleotides. Plasmids can be easily visualized with traditional DNA staining, while oligos have to be radioactively labelled using the ^{32}Phosphate isotope. The EMSA method and these protocols are applied throughout the other chapters of this thesis.

Chapter 4 focusses on the processes of interference and primed adaptation, specifically on their tolerance of mutations. Invaders can escape Type I-E CRISPR-Cas immunity in *E. coli* by making point mutations in the protospacer (especially in the seed) or its adjacent motif (PAM), but hosts quickly restore immunity by integrating new spacers in a positive feedback process termed priming. Here, we provide a systematic analysis of the constraints of both direct interference and subsequent priming in *E. coli*. We have defined a high-resolution genetic map of direct interference by Cascade and Cas3, which includes five positions of the protospacer at 6 nt intervals that readily tolerate mutations. Importantly, we show that priming is an extremely robust process capable of utilizing degenerate target regions with up to at least eleven mutations throughout the PAM and protospacer region. Priming is influenced by the number of mismatches, their position and is nucleotide dependent. Our findings imply that even out-dated spacers containing many mismatches can induce a rapid primed CRISPR response against diversified or related invaders, giving microbes an advantage in the co- evolutionary arms race with their invaders.

In **Chapter 5** we elucidate the mechanism of priming. Specifically, we determine how new spacers are produced and selected for integration into the CRISPR array during priming. We show that priming is directly dependent on interference. Rapid priming occurs when the rate of interference is high, delayed priming occurs when the rate of interference is low. Using *in vitro* assays and next generation sequencing, we show that Cas3 couples CRISPR interference to adaptation by producing DNA breakdown products that fuel the spacer integration process in a two-step, PAM-associated manner. The helicase-nuclease Cas3 pre-processes target DNA into fragments of about 30–100 nt enriched for thymine-stretches in their 3’ ends. By reconstituting the spacer integration process *in vitro*, we show that the Cas1-2 complex further processes these fragments and integrates them sequence- specifically into CRISPR repeats by coupling of a 3’ cytosine of the fragment. Our results highlight that the selection of PAM-compliant spacers during priming is enhanced by the combined sequence specificities of Cas3 and the Cas1-2 complex, leading to an increased propensity of integrating functional CTT-containing spacers.

In **Chapter 6** we look deeper into a nucleotide specific effect on priming that was discovered in Chapter 4. Immunity is based on the complementarity of host encoded spacer sequences with protospacers on the foreign genetic element. The efficiency of both direct interference and primed acquisition depends on the degree of complementarity between spacer and protospacer. Previous studies focused on the amount and positions of mutations, not the identity of the substituted nucleotide. In Chapter 4, we describe a nucleotide bias, showing a positive effect on priming of C substitutions and a negative effect on priming of G substitutions in the basepairing strand of the protospacer. Here we show that these substitutions rather directly influence the efficiency of interference and therefore indirectly influence the efficiency of interference dependent priming. We show that G substitutions have a profoundly negative effect on interference, while C substitutions are readily tolerated when in the same positions. Furthermore, we show that this effect is based on strongly decreased binding of the effector complex Cascade to G mutants, while C mutants only minimally affect binding. In Chapter 5 we showed a connection between the rate of interference and the time of occurrence of priming. Here, we also quantify the extent of priming and show that priming is very prevalent in a population that shows intermediate levels of interference, while high or low levels of interference lead to a lower prevalence of priming.

Chapter 7 describes an attempt to make use of our knowledge about the Cascade complex and develop it into a genome editing tool. The development of genome editing tools has made major leaps in the last decade. Recently, RNA guided endonucleases (RGENs) such as Cas9 or Cpf1 have revolutionized genome editing. These RGENs are the hallmark proteins of class II CRISPR-Cas systems. Here, we have explored the possibility to develop a new genome editing tool that makes use of the Cascade complex from *E. coli*. This RNA guided protein complex is fused to a FokI nuclease domain to sequence specifically cleave DNA. We validate the tool *in vitro* using purified protein and two sets of guide RNAs, showing specific cleavage activity. The tool requires two target sites of 32 nt each at a distance of 30-40 nt and inward facing three nucleotide flexible PAM sequences. Cleavage occurs in the middle between the two binding sites and primarily creates 4 nt overhangs. Furthermore, we show that an additional RFP can be fused to FokI-Cascade, allowing visualization of the complex in target cells. Unfortunately, we were not able to successfully apply the tool *in vivo* in eukaryotic cells.

Chapter 9

General discussion

The CRISPR hype

CRISPR-Cas systems have just been discovered around 15 years ago and their research has already made an amazing journey. Everything started with just a handful of people, excited by these hypothetical prokaryotic defence systems and driving forward the fundamental understanding of their function and mechanisms. Initially, the field only grew slowly as more groups joined in. However, in 2013 the field exploded due to the discovery that Cas9 can be used for precision genome editing. However, despite the genome editing hype, fundamental research into the mechanisms of CRISPR-Cas for all known systems remains important. In this thesis, I elucidated molecular mechanisms of the type I-E CRISPR-Cas system in *Escherichia coli* and I developed the Cascade complex into a Cas9-like RNA guided nuclease for genome editing purposes.

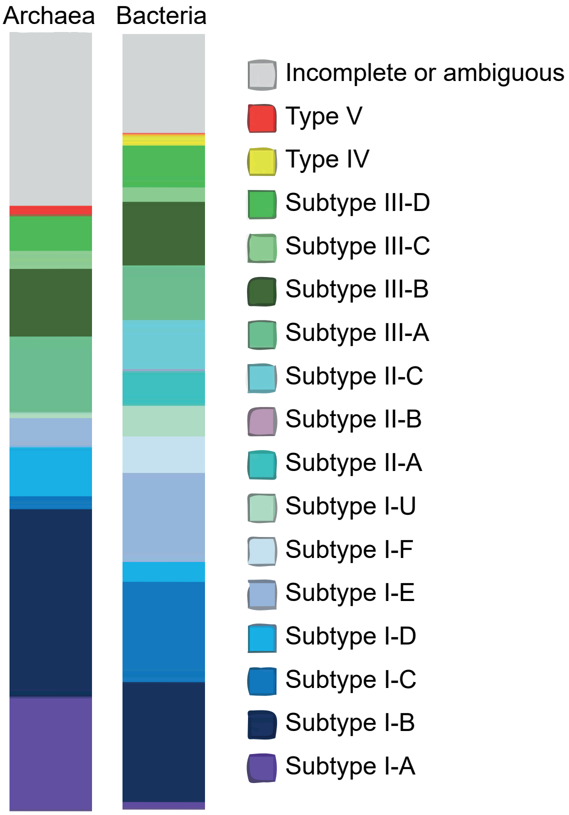


Figure 1 – Distribution of CRISPR-Cas systems. Adapted from (Makarova et al., 2015). Chart depicting the proportions of most currently known CRISPR-Cas systems, including subtypes, among archaea (left) and bacteria (right).

Regulation and ecological significance of CRISPR-Cas

Many organisms contain highly active CRISPR-Cas systems, that are either constitutively expressed or upregulated upon phage infection (Agari et al., 2010; Barrangou et al., 2007; Deltcheva et al., 2011; Shinkai et al., 2007). In addition, several systems are regulated by quorum sensing, leading to upregulation at high cell densities (Hoyland-Kroghsbo et al., 2017; Patterson et al., 2016). Few additional regulatory pathways have been discovered, showing a lack of understanding of CRISPR-Cas regulation (Patterson et al., 2017).

Type I-E in E. coli

The type I-E system in *E. coli* on the other hand, is tightly regulated and *cas* gene expression is almost completely repressed under laboratory conditions (Pul et al., 2010; Westra et al., 2010). The global regulatory protein HNS is responsible for this repression and de-repression is possible via the transcriptional activators LeuO and BaeS (Perez-Rodriguez et al., 2011; Westra et al., 2010). The latter is part of the BaeSR system that is involved in sensing membrane stress. This might suggest that the CRISPR-Cas system in *E. coli* can be upregulated by sensing membrane stress following phage infection. Another gene product required for an efficient immune response is the chaperone HtpG, which is essential for Cas3 stability and activity (Yosef et al., 2011). Cas3 has been shown to be a limiting factor in type I-E immunity in *E. coli* and HtpG can be induced by phage infection (Majsec et al., 2016; Poranen et al., 2006). Interestingly, expression of the CRISPR array seems to be constitutive in *E. coli*, although it can be modulated by regulatory proteins (Pul et al., 2010).

Despite the identification of a number of regulators, to date no growth conditions have been identified that lead to active expression of *cas* genes in *E. coli*. Instead, overexpression has to be conducted either with ΔHNS strains or with strains in which the native *cas* promoter has been substituted by a constitutive/ inducible counterpart. This raises the question whether the system is active at all in nature and in how far experiments using artificial expression levels reflect the natural situation. One way to analyse the activity of the system in nature is to look at the diversity of spacers in different *E. coli* isolates, which should mirror the spacer acquisition activity. A few studies have analysed the spacer diversity in reference strains and natural isolates, revealing a very low diversity in spacer content (Diez-Villasenor et al., 2010; Savitskaya et al., 2016; Touchon et al., 2011). Moreover these studies showed that the spacer content between strains was either highly similar or completely different, suggesting only rare but radical turnover of spacers rather than gradual acquisition. This suggests that the type I-E system in *E. coli* might not function like a traditional CRISPR-Cas immune system. However, the system is functionally conserved and when expressed, it functions like a traditional

CRISPR-Cas immune system. Therefore, there must be a certain selection pressure in the natural environments of *E. coli* to maintain the functionality of the type I-E system.

The type I-E system is not limited to *E. coli*, it is rather wide spread among bacteria (Figure 1) (Makarova et al., 2015). For example, *Streptococcus thermophilus* and *Thermus thermophilus* both contain type I-E systems, in addition to a number of other systems, and both have been shown to up-regulate type I-E expression during phage infection (Agari et al., 2010; Young et al., 2012). Furthermore, the high prevalence of type I systems in prokaryotes, the mechanistic similarities between type I systems, and the importance of *E. coli* as a model system, have made the type I-E system a very well-studied CRISPR-Cas model.

Prokaryotes possess a multitude of defence strategies against phage infection, most of which are less complex than CRISPR-Cas. Examples of more simple strategies include receptor masking or loss, or blocking of phage DNA injection (Westra et al., 2012b). This brings up the question of the relative impact of each of these strategies on cell survival and on population dynamics. Undoubtedly, some of these strategies can work together as subsequent layers of defence, cumulatively reducing the risk of infection. But the relative ecological significance of the different defence strategies, especially CRISPR-Cas, remains largely unstudied. Recently, efforts have been made to study the fitness cost, selective advantage and ecological significance of the type I-F system in *Pseudomonas aeruginosa* (van Houte et al., 2016; Westra et al., 2015). The authors compared the evolution of constitutive defence strategies such as receptor modification with the inducible CRISPR-Cas defence. These systems have a constitutive and induced fitness cost for the host, respectively (Agari et al., 2010; Lenski, 1988; Quax et al., 2013; Young et al., 2012). Initially, theoretical modelling and experimental evolution experiments revealed that constitutive defence is strongly favoured in high resource environments (>95%), while inducible defence is strongly favoured in low resource environments (>95%) (Westra et al., 2015). The authors then linked this behaviour to the differences in cell and phage density in the different environments. High resource environments produce high cell densities and high virus titres, which creates a high infection risk, while low resource environments produce low cell densities and low virus titres, which creates a low infection risk. Finally, increasing virus concentrations were shown to gradually decrease the prevalent evolution of CRISPR-Cas defence in low resource environments and instead lead to the evolution of receptor modification. Furthermore, cells that carry receptor modifications show a fitness advantage over cells that carry CRISPR-Cas defence, outcompeting them in a high phage environment. This shows that, while the inducible CRISPR-Cas system can be an energy efficient alternative to constitutive defence, it suffers from high relative fitness costs at high virus pressures. In conclusion, CRISPR-Cas defence is likely only of ecological significance in environments with relatively low cell and phage densities. This might explain the higher relative prevalence of CRISPR-Cas systems in archaea compared to bacteria, since archaea typically occur in low density environments with less phage exposure (Weinberger et al., 2012).

In this context, it would be very interesting to study the evolution of CRISPR-Cas defence in *E. coli* in response to phage exposure, granting that laboratory conditions can be found that allow activation of the system. The natural environment of *E. coli*, the gut, has the highest bacterial density of any known environment, with likely high phage prevalence (Actis, 2014). The CRISPR-Cas system might therefore not be the defence strategy of choice for *E. coli*, which might explain why it seems to be inactive. This still leaves us with the question, why the system is functionally conserved in *E. coli* and supports the idea that it might have alternative functions. With the advent of human microbiome studies in recent years, a lot of metagenomics data is becoming available that includes *E. coli* and phage sequences. This data might help with our understanding of the phage-host interaction and ecological relevance of the CRISPR-Cas system in the natural environment of *E. coli*.

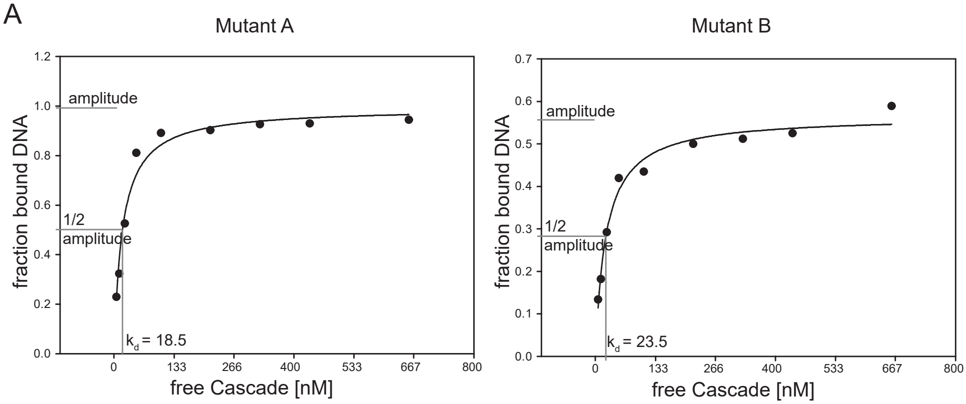
Seed sequences in crRNA guided effector complexes

The increasing number of crystal structures of crRNPs with or without their nucleic acid targets enabled a more detailed description of the seed sequences in CRISPR-Cas systems (see also Chapter 2). To date, seed sequences have been identified in type I, type II and type V systems, which are also the systems that make use of PAM sequences (Jinek et al., 2012; Semenova et al., 2011; Swarts et al., 2017; Wiedenheft et al., 2011b). In addition a potential seed was recently described for the RNA-targeting type VI system (Abudayyeh et al., 2016). The RNPs in all these systems have very diverse architectures and also the crRNA conformations vary strongly. While the crRNA in the type I-E system is tightly bound to the protein by its 5’ end, its 3’ end and several positions throughout the crRNA (Jackson et al., 2014; Mulepati et al., 2014; Zhao et al., 2014), type II and type V systems only anchor the crRNA at its 3’ end or 5’ end, respectively (Dong et al., 2016; Gao et al., 2016; Hirano et al., 2016; Jinek et al., 2014; Nishimasu et al., 2015; Nishimasu et al., 2014; Swarts et al., 2017; Yamada et al., 2017; Yamano et al., 2016). This also has consequences for the mechanism of target DNA binding and seed sequence identity. The crRNA in Cas9 only has a 3’ repeat handle that is duplexed with the anti-repeat part of the tracrRNA and tightly bound to the protein. At the 3’ end of the spacer, next to the repeat handle, the ~10 nt seed is located that is pre-ordered in an A-form helix, while the 5’ part of the spacer is flexible and not tightly associated to the protein. Target DNA binding is achieved by PAM recognition in the DNA, subsequent strand separation and base-pairing with the pre-ordered seed (Hayes et al., 2016; Sternberg et al., 2014; Szczelkun et al., 2014). Base pairing then continues throughout the rest of the spacer to form a complete duplex. This is achieved by wrapping of the flexible 5' end of the crRNA around the DNA target strand. Cpf1 employs a similar mechanism, although PAM-protospacer are mirrored compared to Cas9 and details between Cpf1 homologs vary. While *Francisella novicida* Cpf1 uses a preordered seed for DNA binding (Swarts et al., 2017), *Lachnospiraceae bacterium* and *Acidaminococcus sp.* Cpf1 both contain disordered seeds (Dong et al., 2016; Gao et al., 2016). However, in these two cases the seed is speculated to become ordered upon PAM binding (Gao et al., 2016).

The type I-E Cascade complex on the other hand contains a 5' and a 3' repeat handle that are both tightly bound to the protein (Wiedenheft et al., 2011a). Furthermore the crRNA extensively interacts with a thumb-domain of the Cas7 backbone subunits at every 6th base (Jackson et al., 2014; Mulepati et al., 2014; Zhao et al., 2014). The 5 nucleotide segments in between these positions are all preordered in a near-A form helix, while the interacting bases are flipped and not available for base pairing (see also Chapter 4). Although this may suggest that all segments are potential seed sequences, only the first 2 segments are easily accessible for target DNA, while the other segments are partially obstructed by the Cse2 dimer (Zhao et al., 2014). In addition, like in Cas9 and in Cpf1, the seed in Cascade is adjacent to the PAM and target binding has been shown to initiate at the PAM and continue with the seed (Redding et al., 2015; Szczelkun et al., 2014). An interesting question remains: why does the seed include positions 1-5 and 7-8 and not the entire first 2 segments or only the first segment? This might be explained by the structure as well, since bases 7 and 8 point in a similar direction as bases 1-5 (Zhao et al., 2014). Another explanation is that 7 base pairs have been shown to be the optimal length to produce a thermodynamically stable interaction while still being short enough to allow rapid association and dissociation (Cisse et al., 2012). In addition it has been shown for human Argonaute that 7 base pairs make a stable interaction, while 6 base pairs rapidly dissociate again (Chandradoss et al., 2015). This behaviour cannot be explained by thermodynamics alone and is actually influenced by amino acid residues of the Argonaute protein. As a consequence of the rigid crRNA structure in Cascade, complete base pairing with the target DNA is not possible. Instead, the non-contiguous segments of 5 bases each pair with the target DNA, without RNA and DNA forming a double helix (as in Cas9 and Cpf1) .

For Cascade also PAM and seed-independent binding was described (Blosser et al., 2015). While this binding is shorter lived than canonical binding, it still is a specific interaction. It remains unknown how and where the base-pairing interaction is initiated in the absence of a PAM and seed match. The PAM has been described to be an essential mechanistic element in dsDNA binding, since it is required for initial strand invasion and local unwinding of the DNA strands to allow pairing with the crRNA (Hayes et al., 2016). It is therefore unclear how PAM-independent dsDNA binding of Cascade is achieved. Cascade prefers negatively-supercoiled (nSC) target DNA, and it has been shown that the supercoiling energy helps with local unwinding of the DNA, making it more accessible (Westra et al., 2012c). This feature might be especially relevant in PAM-independent binding, by allowing the crRNA to pair with the target strand without prior PAM assisted strand invasion.

The definition of the seed is very clear from a structural point of view. But a number of studies have shown that the seed is not clearly defined based on its mutation tolerance. Initially the seed was believed to be completely intolerant to mutations and was therefore defined to a certain length (Jinek et al., 2012; Semenova et al., 2011; Wiedenheft et al., 2011b). Recently, the plethora of genome editing applications of Cas9 and a systematic study of many different spacer sequences in Cascade have shown that the mutation tolerance of the seed is dependent on the individual spacer sequence and possibly the conditions of the experiment (Hsu et al., 2013; Xue et al., 2015). It is possible that a GC-rich seed can tolerate mutations better than an AT-rich seed due to higher thermodynamic stability, although experimental evidence to support this is currently lacking. In conclusion, the seed plays a primary role in facilitating the rapid target search, by allowing fast probing and fast rejection of potential target sites near canonical PAMs. This effect is accomplished by the structural accessibility and the pre-ordering of the seed that decreases the entropic penalty of dsDNA unwinding and R-loop formation.



B

$Y = X / (k_d + X)$ simplified formula assuming amplitude of 1

$Y = \text{amplitude} * X / (k_d + X)$ complete formula including variable amplitude

Figure 2 – Cascade binding saturation. A) Examples of two binding curves, one with an amplitude of ~1, and one with an amplitude of <1. Reference points for K_d determination are indicated. B) Two formulas for regression analysis of Cascade-DNA binding curves. In the past, the simplified formula was often used, because the amplitude was assumed to be 1.

DNA binding property of Cascade

One intriguing observation that we and others made when performing EMSAs with Cascade, is that some protospacer mutants do not reach full binding/saturation with protein-guide complexes (Chapter 5/6) (Hochstrasser et al., 2014). In some cases, even at very high protein to DNA ratios, less than 50 % of the DNA was bound by protein (Figure 2A). The shape of the binding curve in these cases also suggests that maximal binding is reached. Usually the K_d is considered as the protein concentration at which half of the DNA is bound and therefore the amplitude is assumed to be 1 (see Chapter 3). But technically the K_d is the concentration where half of the maximal binding is reached. Since we are dealing with a non-constant amplitude (maximum fraction of bound DNA), this adds another variable in the binding formula (Figure 2B). Therefore, if the maximum is not constant, we cannot fairly compare the K_d values with each other, because they have different reference points. This may lead to overestimation of the affinity when the maximum binding is low (Figure 2A, right graph). To circumvent this problem, we decided to represent the affinity by a ratio of the K_d and the amplitude rather than the K_d alone. No matter how the affinity is represented in the end, both parameters (K_d and amplitude) need to be considered to allow fair comparison of affinities.

There is currently no biochemical explanation that might explain why some protospacer mutants cannot be saturated to 100% binding even at high protein concentrations. It is, however, possible that this is an artefact of the EMSA and that these targets are saturated in solution. This could be connected to the different binding modes of Cascade with targets, which will be discussed in detail in the following paragraphs. Briefly, Cascade likely binds targets in a mixture of different binding modes, dependent on the mutations (Blosser et al., 2015; Rutkauskas et al., 2015; Xue et al., 2016). While the canonical binding mode is very stable, non-canonical binding modes are short-lived. Thus, during EMSA, non-canonical binding potentially dissociates, leaving only a fraction of canonically bound targets intact. Whether this is true remains to be tested.

Priming acquisition

Ecological importance of naïve and primed adaptation

Naïve spacer acquisition in *E. coli* is very slow and inefficient, usually being virtually undetectable after 48 h both in ΔHNS as well as in the overexpression strains used in this thesis. This is in stark contrast to the rapid naïve spacer acquisition in many other organisms, especially with type II systems (Fineran and Charpentier, 2012). Interestingly, type II systems require Cas9 in addition to Cas1-2 for naïve acquisition, although Cas9 does not need to carry a spacer targeting the invader (Heler et al., 2015; Wei et al., 2015b). The mechanism and activity of spacer acquisition in all systems is likely tuned for an optimal balance between providing efficient immunity and reducing the risk of lethal self-spacer acquisition. Maybe type II systems have found a way to increase their naïve acquisition activity without increasing the risk for self-spacer acquisition using Cas9. It is also possible that in organisms with type II systems the benefits of a highly active naïve acquisition outweigh the cost of self-spacer acquisition in light of a high virus burden. Several type I systems seem to have solved this problem by employing a low activity naïve acquisition mechanism that has a minimal risk for self-spacer acquisition. Naive acquisition is then supplemented by the highly active and highly target specific priming process, which leads to a rapid diversification of the spacer content. It is possible that the virus burden on organisms with type I systems is generally lower, making them prioritize low self-spacer acquisition over rapid immunization. While a single spacer might be sufficient to provide initial immunity against a particular virus or plasmid, escape mutants will quickly arise. The priming process is therefore vital to rapidly restore immunity against these escapers. But priming is not only important as a rapid response mechanism to escape mutants. By integrating additional spacers from a group of invaders that differ between host cells, a highly diverse immune repertoire is created. A recent study has shown that spacer diversity is a major determinant of population resistance to virus infection (van Houte et al., 2016). Bacterial populations with a high diversity in spacer content are able to drive viruses to extinction, while populations with low spacer diversity allow for the virus to coevolve and persist. The high-diversity-generating priming process therefore greatly reduces the success of viral escape by simple mutations. Consequently, viruses are under strong selection pressure to evolve sophisticated escape strategies, such as anti-CRISPR proteins (Bondy-Denomy et al., 2013; Pawluk et al., 2014; Pawluk et al., 2016), extensive genome recombination (Paez-Espino et al., 2015), or hijacked CRISPR-Cas systems (Seed et al., 2013).

Self-regulation and energy efficiency

The priming process in *E. coli* has been shown to be extremely robust and active against invaders, tolerating up to 13 mutations in their primed protospacer (Chapter 4). Thus, priming is potentially active against invaders with only 60% sequence similarity to the spacer. However, priming can also be triggered by perfectly matching protospacers (Semenova et al., 2016; Staals et al., 2016; Swarts et al., 2012). Actually, it was shown that, when correcting for the copy number of an invader, perfectly matching protospacers trigger priming more efficiently than escape mutants (Semenova et al., 2016). However, when the copy number is considered, escape mutants trigger priming more efficiently. This is because escape mutants are degraded slower by the interference machinery and therefore persist in the cell for a longer time, while perfect targets are rapidly cleared. Combined with the continuous replication in parallel to degradation, escape targets therefore present much more potential spacer substrates and more time for the acquisition machinery (Semenova et al., 2016; Severinov et al., 2016; Staals et al., 2016) (see also Chapter 5). Thus priming acquisition is active against invaders that have recently donated a spacer (perfect match), against invaders that had time to accumulate some or many escape mutations, and even against related invaders that share sufficient sequence similarity. At the same time, the extent of priming is controlled by a combination of interference efficiency and target copy number/replication. This results in a system that is perfectly balanced to preserve energy by not acquiring extra spacers from targets that are already efficiently cleared from the cell, while maintaining the ability to rapidly acquire spacers if targeting is insufficient, for example when the target interference is reduced due to mutations or increased copy numbers.

Two distinct models of priming

It remains difficult to formulate a unified model of priming acquisition in *E. coli*. A number of studies all supply small parts of the puzzle that add up to our current understanding (Jackson et al., 2017). In the beginning, some of these studies seemed slightly contradicting, offering alternative models, but eventually lead us to believe that priming is a more complex process than initially anticipated. Currently, these contradictions are solved by a model that accepts the presence of two separate mechanisms of primed spacer acquisition, one being interference associated and one being interference independent (Figure 3). The interference-dependent model was originally proposed by Swarts et al in our lab (Swarts et al., 2012). This model has since been supported in *E. coli* by the work presented in this thesis (Chapter 5; (Künne et al., 2016)) and by others (Semenova et al., 2016); in addition, a similar mechanism has been demonstrated in the type I-F system of *Pectobacterium atrosepticum* (Staals et al., 2016). These studies show a direct quantitative connection between direct interference and resulting primed spacer acquisition, and we have shown for *E. coli* that Cas3-derived DNA degradation products can directly be used by Cas1 and Cas2 for spacer integration (Chapter 5). This model holds true for protospacer mutants, but not PAM mutants. Many PAM mutants have been shown to be unable to trigger direct interference, because they are unable to trigger Cas3-mediated DNA degradation (Blosser et al., 2015; Hochstrasser et al., 2014; Mulepati and Bailey, 2013; Rutkauskas et al., 2015; Xue et al., 2015). These mutants are, however, still able to trigger priming acquisition (interference-independent priming).

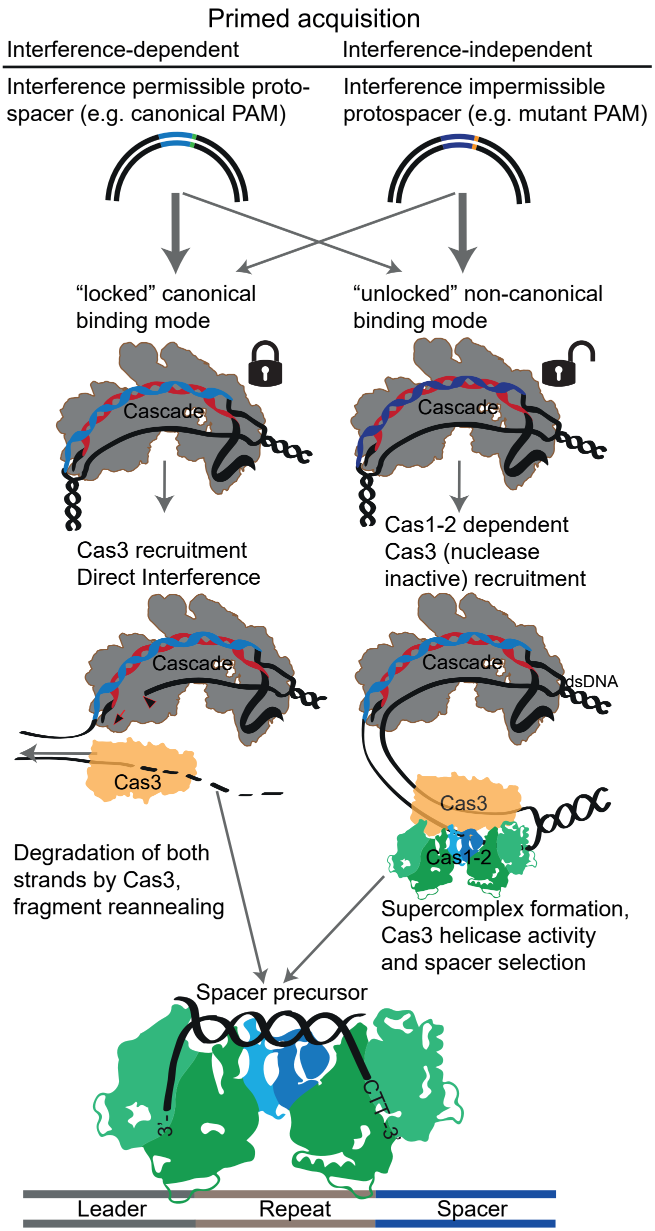

Primed acquisition
Interference-dependent
Interference-independent
Interference permissible proto-spacer (e.g. canonical PAM)
Interference impermissible protospacer (e.g. mutant PAM)
“locked” canonical binding mode
“unlocked” non-canonical binding mode
Cascade
Cascade
Cas3 recruitment
Direct Interference
Cas1-2 dependent
Cas3 (nuclease inactive) recruitment
Cascade
Cascade
dsDNA
Cas3
Cas3
Cas1-2
Degradation of both strands by Cas3, fragment reannealing
Supercomplex formation, Cas3 helicase activity and spacer selection
Spacer precursor
3'-
CTT-3'
Leader
Repeat
Spacer

Figure 3 – Two modes of priming. Schematic illustration showing the two current models of primed spacer acquisition and how they may be connected. We distinguish interference-dependent and independent priming. The former is triggered by any target that still causes (partial) interference, while the latter is dominantly triggered by PAM mutants that do not trigger interference. Interference permissible protospacers are dominantly forming a “locked” or canonical binding mode, while interference impermissible protospacer are dominantly forming “unlocked” or non-canonical binding modes. However, protospacers exist in an equilibrium of both binding modes with varying relative contributions. Canonical binding modes trigger Cas3 recruitment and target DNA degradation (direct interference). During DNA degradation, spacer precursors are generated by Cas3 that are captured by Cas1-2 and processed into spacers. Non-canonical binding modes trigger Cas1-2 dependent Cas3

recruitment, however, Cas3 is nuclease inactive. A hypothetical Cas1-2-3 supercomplex is formed and scans the DNA for PAMs/protospacers using Cas3 helicase activity. Either Cas3 or Cas1-2 directly cleave the spacer precursor from the target DNA.

Therefore, another priming model has been proposed in a single molecule study (Redding et al., 2015). The study shows that PAM mutants are only able to recruit Cas3 in the presence of Cas1-2 and that Cas3 is only displaying helicase activity, therefore being nuclease inactive. Their model predicts a supercomplex of Cas1-2-3, as originally suggested by (van der Oost et al., 2009), that is using the Cas3 helicase to move along the DNA while Cas1-2 is looking for potential new spacers. We initially hypothesized that this model might apply specifically to PAM mutants, since these are mechanistically different from protospacer mutants. However, recent studies have since provided evidence that suggests a more dynamic relationship between mutations (in PAM and protospacer) and the resulting immune response.

Conformational effects

Especially the combination of three single molecule studies strongly suggests that there is a dynamic equilibrium of different Cascade (mainly the Cse1 subunit) conformations leading to different binding modes that determine the outcome (Blosser et al., 2015; Rutkauskas et al., 2015; Xue et al., 2016). A canonical, perfectly matching target, will almost exclusively lead to the 'closed' Cse1 conformation; this canonical binding mode triggers rapid direct interference. However, any mutation might, to a varying degree, shift this balance towards the 'open' Cse1 conformation (non-canonical binding mode). It seems that PAM mutants are especially powerful at shifting this balance, followed by seed mutants and at last non-seed protospacer mutants. The extent to which each mutant skews the balance between binding modes likely also depends on the relative binding affinities of the modes for each particular mutant. Blosser et al. have shown that the non-canonical binding mode is much shorter lived when compared to the canonical binding mode (Blosser et al., 2015). This shows that the canonical binding mode is intrinsically preferred when the target allows it. This makes sense, since the primary objective is to degrade the invader and only if this process is insufficient, priming becomes the priority.

Mix of immune responses

This balance of binding modes shows that any target likely produces a mixture of different responses depending on the balance of conformations or binding modes. This is biologically relevant especially for the defence against multicopy invaders, which will trigger different responses in each host cell. Both plasmids and viruses are typically replicating to copy numbers of a few to several hundred copies per cell, in a few cases even up to a few thousand copies (Brock, 1990; Shao and Wang, 2009). Studies have always classified mutants as either triggering interference or not. Later priming was added as another option. However, in reality all responses are occurring simultaneously with different relative contributions and efficiencies. Just because we classify a particular mutant as an escape mutant in a plaque assay or transformation assay, does not mean that the virus or plasmid is not being targeted and degraded at all. In other words, just because we consider a mutant not to trigger interference, does not mean it does not trigger interference at all. Thus, the definition of interference can be very misleading. Still, for the sake of continuity we have continued to define interference as the successful removal of the invader, while we considered target degradation separately.

In chapter 5, we have found that the Cas3 activity matches very well with both interference and priming. More Cas3 activity leads to better interference (faster plasmid loss), which in turn results in faster priming. Thus, in our case all mutants seem to dominantly trigger the interference-dependent priming process, since the interference-independent process would show priming in the absence of Cas3 activity. We therefore suggested that this is the result of keeping the canonical PAM for all mutants that we tested. We do, however, include seed mutants, which have been reported to skew the balance of conformations towards the 'open' mode that is thought to trigger interference independent priming (Xue et al., 2016). It is possible, that the amount of seed- or overall mutations is relatively low in our study and therefore the mutants still retain a large fraction of 'closed' conformations leading to interference.

In conclusion, the actual outcome of an immune response is determined by many factors such as the target copy number and replication speed, the binding affinities of Cascade to the target, the balance of binding modes and the efficiency in Cas3 recruitment and activity. All these will influence the overall time the cell needs to get rid of the invader, thereby influencing both interference-dependent and independent priming processes.

Spacer precursor generation

The least well-described part of spacer acquisition is still the generation of spacer precursors. The crystal structures of the Cas1-2 complex, binding studies and *in vitro* spacer acquisition experiments have revealed the basic architecture of the spacer precursor molecules in *E. coli* (Nunez et al., 2015a; Nunez et al., 2014; Nunez et al., 2015b; Rollie et al., 2015; Wang et al., 2015). These precursors have to be partially double stranded, contain single stranded 3' ends, and carry a PAM sequence on one 3' end. It is not really clear yet whether there is a length restriction for the precursors, but Cas1-2 might favour precursors of near-spacer length (Chapter 5) (Fagerlund et al., 2017). How are precursors with such specific requirements produced? The first proposed model suggested that, for naïve acquisition, spacer precursors are left-over products generated by RecBCD during DNA repair of DSBs at stalled replication forks (Levy et al., 2015). However, there is still no direct evidence for this and acquisition is still present, although reduced, in the absence of RecBCD. RecBCD degrades both DNA strands next to the DSB and degradation fragments are thought to serve as spacer precursors. This model assumes that ssDNA degradation fragments reanneal to form partial dsDNA duplexes and that PAMs will be present by chance. Furthermore RecBCD must degrade the DNA in fragments of at least spacer length. The DNA fragment length after RecBCD activity *in vivo* is not known and fragment lengths *in vitro* are strongly dependent on experimental conditions (Wigley, 2013). Since this process is proposed to be responsible for the majority of naïve spacer acquisition, taken together this might explain why naïve acquisition is so slow/inefficient in *E. coli*.

In chapter 5 we have validated our early theory on the production of spacer precursors during primed adaptation (Swarts et al., 2012). Here, precursors are generated directly during direct interference, thus DNA degradation by Cas3. Our data suggests that Cas3 by itself is able to generate DNA fragments that fulfil the requirements of spacer precursors. The single stranded DNA degradation products of Cas3 only need to anneal back together with fragments from the other strand such that 3’ overhangs are created. This part of the proposed mechanism is dependent on chance though, as Cas3 is thought to process each of the two strands independently. This is also true for RecBCD, but Cas3 has the additional ability to enrich for PAMs at the 3’ ends of fragments and Cas3 produces fragments of near-spacer length (Chapter 5). These features, and the proposed formation of Cas1-2-3- supercomplexes (Fagerlund et al., 2017; Rollins et al., 2017), likely contribute to the high efficiency and high PAM specificity of priming. It remains possible that reannealing of precursor fragments is actually assisted or controlled by the Cas1-2 complex. Cas1 has been shown to catalyse annealing of oligos and more recently Cas1 was shown to associate with non-double-stranded spacer-sized DNA during primed adaptation *in vivo* (Babu et al., 2011; Musharova et al., 2017). These non-double-stranded spacer-sized fragments are reported to be excised from longer non-double stranded fragments produced by Cas3. This suggests that Cas1-2 might initially bind to longer single stranded spacer precursors directly generated by Cas3, trim these fragments to near-spacer length (possibly assisted by distinct nucleases) and then find fitting complementary fragments and actively participate in duplex formation. This could increase the chance for the formation of proper sized duplexes with single stranded 3’ ends. Furthermore, this mechanism would be universally applicable to both primed adaptation and naïve adaptation, since also RecBCD produces single stranded DNA fragments that are likely not of exact spacer length. Despite all these recent studies, we still do not know whether Cas1-2 in the type I-E system forms a supercomplex with Cas3 during primed acquisition or just scavenges the single stranded intermediates from its environment. We do know, however, that Cas1-2 is required for Cas3 recruitment during interference-independent priming and that type I-F systems carry fusions of Cas2-3 (Redding et al., 2015; Makarova et al., 2017; van der Oost et al., 2009). Thus, direct interaction between Cas1-2 and Cas3 also in interference-dependent priming is very likely.

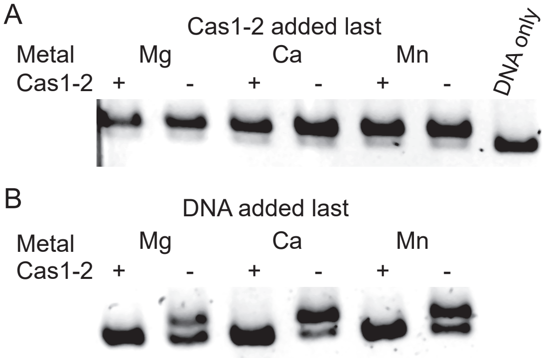


Figure 4 – Cas1-2 interacts with Cascade. EMSA assay of plasmid with protospacer and Cascade with matching crRNA in the presence of different metals and in the presence or absence of Cas1-2 as indicated. Upwards shifted DNA is bound by Cascade. A) Cascade and plasmid were incubated first to allow R-loop formation, Cas1-2 was added last. B) Cascade and Cas1-2 were pre-incubated, then DNA was added. No shift is observed in the presence of Cas1-2.

During the development of our *in vitro* acquisition assays we observed that when Cascade and Cas1-2 were incubated together before the addition of target DNA (PAM mutant), no binding of the DNA by Cascade could be detected with EMSA. Cascade-DNA binding was unimpaired when incubated before the addition of Cas1-2 (Figure 4). This shows that Cas1-2 can either block Cascade from binding to DNA, or possibly Cas1-2 can skew the binding mode balance towards the non-canonical mode that may not be detectable in EMSA. This would suggest that Cas1-2 can actively control the immune response towards interference independent priming. Whether this is true or not, it clearly implies an interaction between Cascade and Cas1-2.

Finally, CRISPR-Cas function is usually described as three consecutive stages, adaptation, expression and interference. The priming process forms a feedback connection from interference to adaptation, therefore creating a circle of immunity (Figure 5).

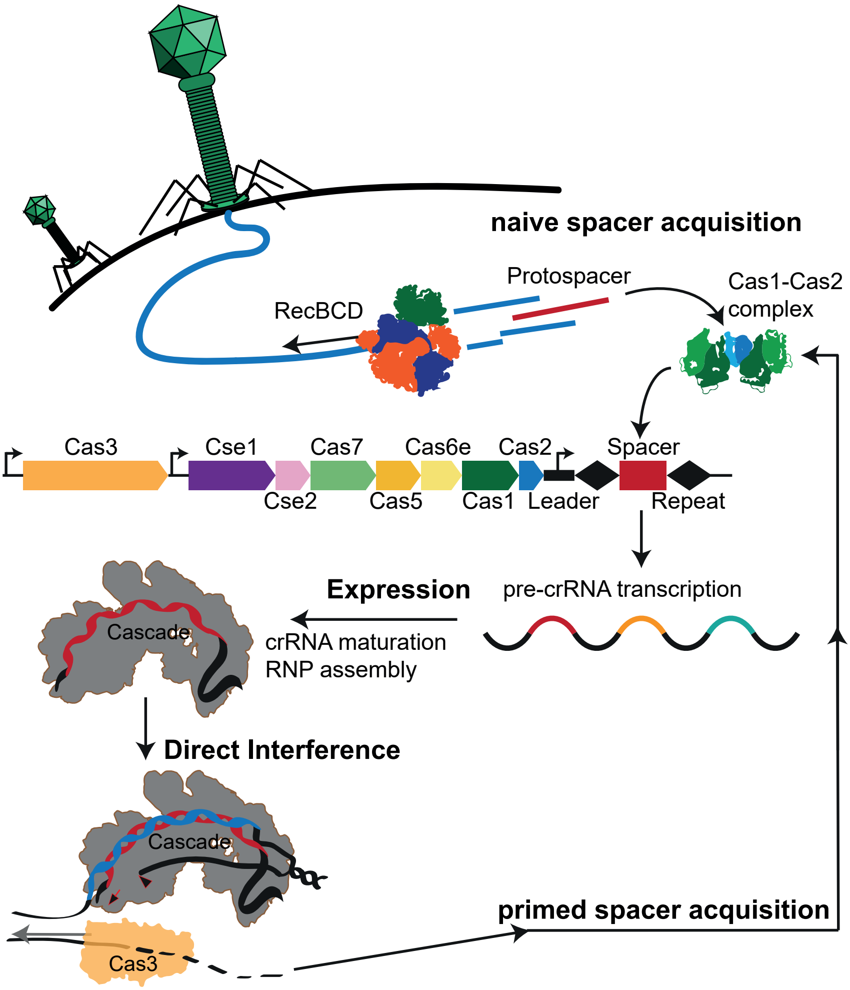


Figure 5 – The circle of CRISPR-Cas defence. Schematic overview of the three stages in CRISPR-Cas immunity and the priming process, forming a circle. From top to bottom: 1) Phage or plasmid DNA enters the cell. A protospacer is selected by the acquisition machinery, including the Cas1-2 complex, and integrated as a new spacer in the CRISPR array. 2) CRISPR array is transcribed into pre-crRNA from the leader and *cas* genes are expressed. The pre-crRNA is processed into mature crRNA and assembles with the effector proteins to form the Cascade complex. 3) Cascade detects invaders by base pairing of the guide RNA with the target DNA. The target DNA gets degraded by Cas3 (indicated by black and red triangles). Degradation fragments are fed back to the acquisition machinery (Cas1-2) creating new spacers.

Strand bias of priming

Based on current literature, there is still no mechanistic model that can explain the origin of the strand bias of new spacers. The strand bias requires the fragment carrying the CTT PAM to be selected from the target strand. We could not reproduce the *in vivo*-observed strand bias (Swarts et al., 2012) in *in vitro* acquisition assays.

This might be caused by two possible issues with these bulk biochemical assays, (i) the bi-directional integration of spacers that cause targeting of either strand, or (ii) the erroneous production and processing of precursors. We do actually see a difference in fragment sizes produced by Cas3 between target and non-target strand, suggesting that Cas3 degrades the two strands differently. However, how this can play a role in creating a strand bias is unclear. Furthermore, bulk assays are likely less specific due to high concentrations of the components. High expression levels of Cas1-2 for example negatively impact the PAM specificity during naïve spacer acquisition *in vivo* (Staals et al., 2016). Additional insights into this mechanism could be gained by a more thorough understanding of the Cas3 mechanism, and by Cas1-2-3 interaction studies. One of the open questions regarding the Cas3 mechanism is for example whether it moves along the DNA freely, or rather stays attached to Cascade and actively pulls the DNA strand towards its nuclease site. A recent single molecule study has indeed seen indications of the latter process (including looping of the other strand) at least in a fraction of cases (Redding et al., 2015), and recent unpublished work using single molecule FRET shows this in great detail (L. Loeff, personal communication, June 13, 2017). In this case, the non-target strand is pulled through Cas3 in steps of 3 nt and degraded into pieces of defined length (~90 nt). This size matches well with the size of the *in vitro* generated spacer precursors in chapter 5. The target strand is looped out next to the Cascade-Cas3 supercomplex. The looped out ssDNA might be a substrate for the same Cas3 molecule, another free Cas3 molecule or a co-localized Cas1-2 complex. Either way, cleavage or selection of fragments from the target strand would have to be PAM associated, while the primary degradation of the non-target strand could be PAM independent.

Type I-F and type I-B systems are both capable of primed acquisition. Both do produce a strand bias of new spacers, but it differs from the type I-E system (Li et al., 2014; Richter et al., 2014). In type I-E, >90% of new spacers target the same strand as the spacer that triggered priming and protospacers are distributed throughout the whole target. However, spacer acquisition experiments using larger phages instead of plasmids also revealed a preference for new spacer selection close to the original protospacer. Spacer sampling is most active 5’ of the original protospacer on the non-target strand. Type I-F/I-B systems produce spacers in equal amounts from both strands. However, spacers are distributed strand specifically around the original spacer. This means, that they preferably acquire spacers from protospacers that are located close to the original protospacer and that acquisition is unidirectional in a 3’ to 5’ direction on both strands. The single strand bias in *E. coli* suggests that the degradation/acquisition machinery only moves 3’ to 5’ on the non-target strand. As mentioned before, it seems obvious that the complex interactions and mechanisms of Cascade, Cas3 and Cas1-2 during priming cannot be elucidated in bulk biochemical assays, but that this rather requires a number of detailed studies to generate a more complete and unified model.

Details about the timing of spacer precursor trimming are also scarce. Our data and recent literature suggest that the size of Cas3-generated fragments is bigger than the spacer length (Chapter 5)(Musharova et al., 2017). Spacers have to be trimmed to both the correct size as well as the correct PAM terminal end. To this end the PAM containing 3' end of the precursor is cleaved such that the cytosine of the CTT PAM remains as the terminal nucleotide (Wang et al., 2015). The other side of the precursor is then likely cleaved according to a ruler mechanism, ensuring the correct length. A similar ruler mechanism was recently demonstrated for Cas1-2 which is responsible for the correct choice of integration sites in the array, and as such for the repeat length of the newly generated repeat after spacer integration (Goren et al., 2016). In Chapter 5 we analysed half-site integration products by sequencing. Unfortunately we could not determine the actual length of the integrated fragments, due to the unintegrated end being cut off by the PCR primer. Still, we found that the sizes of a considerable fraction of the fragments were bigger than the spacer size. This suggests that these spacers are not yet properly processed by Cas1-2 into mature spacers. We therefore suggested that processing of the unintegrated site might only occur at full integration. Thus, whatever missing factor is limiting the *in vitro* assays to half-site integration, likely also limits precursor processing to one half. This short coming of these assays also explains why we observe half site integrations at both possible integration sites and in both possible orientations, while correct integration initiates at the leader-repeat boundary with the non-PAM containing end of the precursor (Nunez et al., 2016). Some (but not all) type I, type II and type V systems actually encode an additional *cas* gene, *cas4*, in their adaptation module (i.e. associated to *cas1* and *cas2*) (Makarova et al., 2017a, b). In addition *cas4* is sometimes fused to *cas1* (van der Oost et al., 2009). This shows that *cas4* is very likely involved in the spacer acquisition process. Indeed, it has been shown that *cas4* is required for primed spacer acquisition in the type I-B system in *Haloarcula hispanica* and that Cas4 interacts with the Cas1-2 complex in the type I-A system of *Thermoproteus tenax* (Li et al., 2014; Plagens et al., 2012). Although the RecB-like Cas4 proteins seem to be functionally diverse, they are generally characterized as exo- or endonucleases on ssDNA or dsDNA (Lemak et al., 2013; Lemak et al., 2014; Zhang et al., 2012). Consequently Cas4 has been hypothesized to be involved in precursor generation or processing. In the systems that do not contain a *cas4* gene, this function must then be either contained in Cas1-2 or carried out by other unknown (RecB-like, or analogous) host factors. If the latter is the case, this would explain the failure of *in vitro* integration assays to produce complete integration.

Nucleotide bias of priming

In chapter 6, we describe a surprising nucleotide bias that strongly affects the efficiency of direct interference and therefore interference-dependent priming. Cytosine substitutions in the base pairing strand of the protospacer hardly affect the immune response while Guanine substitutions have a detrimental effect. As mentioned in chapter 6, this cannot be explained by simple base pairing thermodynamics. Instead this effect is likely influenced by interactions with the Cascade protein, such as direct interactions with amino acid residues or steric effects. This bias also has consequences for the success of viral escape from the type I-E system. Obviously, viruses have no use of preferring G mutations over C mutations, since these occur together in dsDNA and spacers can target either strand. But at least viruses with G substitutions in the target strand of a protospacer should have a selective advantage and should therefore be overrepresented in a viral population. Therefore, we are currently analyzing metagenomic data, which allows us to analyze spacer sequences of host organisms and protospacer sequences of their phages from one environment, assuming that host and phage have co-evolved. We will extract spacer sequences from type I-E systems and map them to potential protospacers allowing mismatches. Here, we anticipate to find an overrepresentation of G substitutions in the base pairing strand of the viral protospacers.

Genome editing

The history of genome editing has seen many steps of evolution and revolution (Kim, 2016). After the initial discovery that endogenous DNA repair pathways can be used to edit genomes (Rudin and Haber, 1988), the first evolution was the application of meganucleases to induce DSBs which greatly stimulated editing at the target locus (Rouet et al., 1994). Meganucleases, however, are not programmable with respect to target sequence specificity and therefore of very limited value for widespread application. The first revolution was the utilization of zinc-finger domains for programmable DNA targeting, combined with the FokI nuclease domain to create ZFNs (Kim et al., 1996). Although technically challenging, the freely programmable ZFNs allowed genome editing to be applied in many cell types and in any locus (Bibikova et al., 2003; Bibikova et al., 2002; Desjarlais and Berg, 1992; Rebar and Pabo, 1994). ZFNs were then outcompeted by TALENs which were slightly easier to reprogram, although still labour intense (Briggs et al., 2012; Cermak et al., 2011; Kim et al., 2013; Miller et al., 2011; Reyon et al., 2012). Furthermore TALENs had a higher specificity, reducing toxicity and off-target editing. The development of Cas9 into a genome editing tool was the next big revolution (Cho et al., 2013; Cong et al., 2013; Jinek et al., 2013; Mali et al., 2013). Reprogramming Cas9 proteins is very easy and fast, and these CRISPR-associated nucleases have shown great efficiency and specificity in a vast range of cell types and organisms (Kim, 2016). The massive boost of genome editing research utilizing Cas9 is proof that this tool has really made genome editing feasible and accessible. But after every revolution there must be evolution, since nothing is without flaws. Evolution came in the form of protein modifications to improve and fine tune Cas9 (Fu et al., 2014; Guilinger et al., 2014; Kleinstiver et al., 2016; Slaymaker et al., 2016; Tsai et al., 2014), in the form of novel Cas9 proteins (Hou et al., 2013; Kleinstiver et al., 2015), or in the form of alternative CRISPR-effector proteins such as Cpf1 (Kim et al., 2016; Zetsche et al., 2015; Zetsche et al., 2017). Having a large pool of tools to choose from is very beneficial, since every tool might have advantages and disadvantages for a given application. Although the CRISPR class 2 RGENs have conceptual advantages over Cascade, such as being much smaller and a single protein, Cascade might still prove superior in other aspects due to its radically different architecture and stronger DNA binding characteristics (Szczelkun et al., 2014). Potential advantages and disadvantages of Cascade have already been discussed in Chapter 7, but recent developments have indicated more potential problems with Cas9. A genome editing study in mice has found a plethora of Cas9 induced SNPs throughout the entire genome by whole genome sequencing. The majority of these sites are not predicted by off-target site prediction software, as they carry no or insufficient homology (Schaefer et al., 2017). This study is, however, being heavily criticized for a poor execution and missing controls. Furthermore, a role of Cas9 has recently been implicated in the virulence of bacterial pathogens, where it can cause DNA damage in the host cell. It is thought that Cas9 does so by scavenging cellular RNAs of the host as guide RNAs instead of using crRNA, however, this problem can be solved by saturating the nuclease with an appropriate crRNA guide (Van der Oost, personal communication, June 18 2017). These findings indicate that this new technology is not yet perfect, which further strengthens the need for alternatives. Finally, it is obvious that the development of a new tool requires significant manpower and could therefore not be completed in this thesis. Additionally, there is always the risk that a tool might not be functional *in vivo*. Likewise, screenings for new Cas9 and Cpf1 variants have shown that only part of the tested proteins were successful in genome editing *in vivo* (Ran et al., 2015; Zetsche et al., 2015).

The success of the FokI-Cascade tool developed in this thesis is currently limited by the instability of the linker and/or the protein fusion. The FokI subunit is hydrolysed or being cleaved off from the complex, likely somewhere in its C-terminus. Complexes lacking FokI can block the binding sites on the target DNA, preventing cleavage. In addition, loose FokI subunits, in case they remain intact, could lead to unwanted random DNA cleavage. While we did not observe the latter *in vitro*, it remains possible that this causes problems *in vivo*. FokI is a widely used restriction enzyme, and its nuclease domain has been applied in genome editing as part of ZFNs, TALENs and even Cas9. Even though the break does not occur within the linker peptide, the linker might still be responsible for the instability of the adjacent FokI residues. The sequence of the linker peptide has been taken from a naturally occurring Cas3-Cse1 fusion in *Streptomyces griseus* (Westra et al., 2012c), and might therefore have unknown effects on the linked proteins. It is conceivable that the linker might have conformational effects on Cas3 in *S. griseus*, since Cas3 nuclease activity needs to be controlled to prevent unwanted DNA degradation until a bona fide target is detected. Changing the linker to a well-characterized one, such as glycine or proline-threonine linkers, is therefore the first priority in the future of this project. The instability of the fused FokI domain might also be caused by incomplete folding of the protein. This could be prevented by codon optimizing the linker (bad codons) to slow down translation and allow proper folding of the FokI domain.

Another important experiment would be to use standalone Cascade *in vivo* in eukaryotic cells to bind to promotor sequences and block gene expression. This could validate the ability of Cascade to remain functional in eukaryotic cells and to stably bind to genomic DNA. This has already been successfully done in prokaryotic cells, such as *E. coli* and *Salmonella typhimurium* (Chang et al., 2016; Luo et al., 2015; Rath et al., 2014). Cascade could be used as a DNA binding platform for other functional proteins, such as epigenetic factors (DNA or histone modifying enzymes), transcriptional regulators or simply fluorescent proteins/probes for locus visualization.

Ultimately, Cas proteins can be used in many applications that require specific interactions with nucleic acids, be it DNA or RNA. Possible functions include the detection, manipulation, regulation, visualization of specific targets in a pool of nucleic acid. These abilities are not limited to fundamental research and the curing of genetic diseases, but can be applied in many more areas such as crop improvement or trace pathogen detection in hospitals or even in the field (Gootenberg et al., 2017; Khatodia et al., 2016). In essence, the CRISPR-Cas toolbox is to us now, what hammer and knife were to our ancestors. The versatility of these tools opens up new ways to manipulate our environment. However, with new possibilities come new risks and responsibilities. Just like a hammer and knife, CRISPR-Cas may be abused or turned into a weapon (DiEuliis and Giordano, 2017). Thus, proper regulation and oversight is essential to protect us from this technology.

Concluding remark

Our understanding of CRISPR-Cas based defence in prokaryotes has greatly improved over the last decade. Extensive knowledge was gathered on function, mechanisms, protein structures and more. Sometimes it seemed as if we were reaching a point where there is little new to describe, but the constant discovery of new systems, new mechanisms and even entirely new functions of CRISPR-Cas systems make this field more exciting than ever. We know now that these systems are not only functioning as immune systems, but are implicated in many additional processes such as virulence in pathogens. Furthermore, genome editing will not remain the only practical application of Cas effector proteins, especially with the discovery of new proteins with versatile functions. Finally, CRISPR-Cas systems exemplify that the central dogma of ‘survival of the fittest’ should really be ‘survival of the most adaptive’.

Appendices

References

About the author

List of publications

Overview of completed training activities

Acknowledgments

References

Abudayyeh, O.O., Gootenberg, J.S., Konermann, S., Joung, J., Slaymaker, I.M., Cox, D.B., Shmakov, S., Makarova, K.S., Semenova, E., Minakhin, L., *et al.* (2016). C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. Science *353*, aaf5573.

Actis, G.C. (2014). The gut microbiome. Inflammation & allergy drug targets *13*, 217-223.

Agari, Y., Sakamoto, K., Tamakoshi, M., Oshima, T., Kuramitsu, S., and Shinkai, A. (2010). Transcription profile of Thermus thermophilus CRISPR systems after phage infection. J Mol Biol *395*, 270-281.

Almendros, C., and Mojica, F.J. (2015). Exploring CRISPR Interference by Transformation with Plasmid Mixtures: Identification of Target Interference Motifs in Escherichia coli. Methods in molecular biology *1311*, 161-170.

Ameres, S.L., Martinez, J., and Schroeder, R. (2007). Molecular basis for target RNA recognition and cleavage by human RISC. Cell *130*, 101-112.

Amitai, G., and Sorek, R. (2016). CRISPR-Cas adaptation: insights into the mechanism of action. Nature reviews. Microbiology *14*, 67-76.

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

Babu, M., Beloglazova, N., Flick, R., Graham, C., Skarina, T., Nocek, B., Gagarinova, A., Pogoutse, O., Brown, G., Binkowski, A., *et al.* (2011). A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. Mol Microbiol *79*, 484-502.

Balbontín, R., Fiorini, F., Figueroa-Bossi, N., Casadesús, J., and Bossi, L. (2010). Recognition of heptameric seed sequence underlies multi-target regulation by RybB small RNA in Salmonella enterica. Molecular Microbiology *78*, 380-394.

Bandyra, K.J., Said, N., Pfeiffer, V., Górna, M.W., Vogel, J., and Luisi, B.F. (2012). The seed region of a small RNA drives the controlled destruction of the target mRNA by the endoribonuclease RNase E. Molecular Cell *47*, 943-953.

Barrangou, R. (2013). CRISPR-Cas systems and RNA-guided interference. Wiley Interdisciplinary Reviews-Rna *4*, 267-278.

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*, 1709-1712.

Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281-297.

Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. Cell *136*, 215-233.

Beisel, C.L., and Storz, G. (2010). Base pairing small RNAs and their roles in global regulatory networks. FEMS microbiology reviews *34*, 866-882.

Beisel, C.L., Updegrove, T.B., Janson, B.J., and Storz, G. (2012). Multiple factors dictate target selection by Hfq-binding small RNAs. The EMBO journal *31*, 1961-1974.

Beloglazova, N., Kuznedelov, K., Flick, R., Datsenko, K.A., Brown, G., Popovic, A., Lemak, S., Semenova, E., Severinov, K., and Yakunin, A.F. (2015). CRISPR RNA binding and DNA target recognition by purified Cascade complexes from Escherichia coli. Nucleic Acids Res *43*, 530-543.

Beloglazova, N., Petit, P., Flick, R., Brown, G., Savchenko, A., and Yakunin, A.F. (2011). Structure and activity of the Cas3 HD nuclease MJ0384, an effector enzyme of the CRISPR interference. *30*, 4616-4627.

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

Bevington, S.L., Cauchy, P., Piper, J., Bertrand, E., Lalli, N., Jarvis, R.C., Gilding, L.N., Ott, S., Bonifer, C., and Cockerill, P.N. (2016). Inducible chromatin priming is associated with the establishment of immunological memory in T cells. EMBO J *35*, 515-535.

Bibikova, M., Beumer, K., Trautman, J.K., and Carroll, D. (2003). Enhancing gene targeting with designed zinc finger nucleases. Science *300*, 764.

Bibikova, M., Golic, M., Golic, K.G., and Carroll, D. (2002). Targeted chromosomal cleavage and mutagenesis in Drosophila using zinc-finger nucleases. Genetics *161*, 1169-1175.

Bikard, D., Euler, C.W., Jiang, W., Nussenzweig, P.M., Goldberg, G.W., Duportet, X., Fischetti, V.A., and Marraffini, L.A. (2014). Exploiting CRISPR-Cas nucleases to produce sequence-specific antimicrobials. Nat Biotechnol *32*, 1146-1150.

Bikard, D., Jiang, W., Samai, P., Hochschild, A., Zhang, F., and Marraffini, L.A. (2013). Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. Nucleic Acids Res *41*, 7429-7437.

Biswas, A., Gagnon, J.N., Brouns, S.J.J., Fineran, P.C., and Brown, C.M. (2013). CRISPRTarget: Bioinformatic prediction and analysis of crRNA targets. Rna Biology *10*, 817-827.

Bitton, G. (1987). Fate of bacteriophages in water and wastewater treatment plants. Wiley Interscience, New York, 181-195.

Blosser, T.R., Loeff, L., Westra, E.R., Vlot, M., Künne, T., Sobota, M., Dekker, C., Brouns, S.J., and Joo, C. (2015). Two distinct DNA binding modes guide dual roles of a CRISPR-Cas protein complex. Mol Cell *58*, 60-70.

Bohn, C., Rigoulay, C., Chabelskaya, S., Sharma, C.M., Marchais, A., Skorski, P., Borezée-Durant, E., Barbet, R., Jacquet, E., and Jacq, A. (2010). Experimental discovery of small RNAs in Staphylococcus aureus reveals a riboregulator of central metabolism. Nucleic acids research *38*, 6620-6636.

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*, 2551-2561.

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

Bouvier, M., Sharma, C.M., Mika, F., Nierhaus, K.H., and Vogel, J. (2008). Small RNA binding to 5′ mRNA coding region inhibits translational initiation. Molecular Cell *32*, 827-837.

Box, A.M., McGuffie, M.J., O'Hara, B.J., and Seed, K.D. (2015). Functional Analysis of Bacteriophage Immunity through a Type I-E CRISPR-Cas System in Vibrio cholerae and Its Application in Bacteriophage Genome Engineering. J Bacteriol *198*, 578-590.

Boyd, E.F., and Brussow, H. (2002). Common themes among bacteriophage-encoded virulence factors and diversity among the bacteriophages involved. Trends Microbiol *10*, 521-529.

Brantl, S. (2007). Regulatory mechanisms employed by cis-encoded antisense RNAs. Current opinion in microbiology *10*, 102-109.

Breitbart, M., and Rohwer, F. (2005). Here a virus, there a virus, everywhere the same virus? Trends in Microbiology *13*, 278-284.

Brennecke, J., Stark, A., Russell, R.B., and Cohen, S.M. (2005). Principles of MicroRNA–Target Recognition. PLoS Biol *3*, e85.

Briggs, A.W., Rios, X., Chari, R., Yang, L., Zhang, F., Mali, P., and Church, G.M. (2012). Iterative capped assembly: rapid and scalable synthesis of repeat-module DNA such as TAL effectors from individual monomers. Nucleic Acids Res *40*, e117.

Brock, T.D. (1990). The emergence of bacterial genetics (Cold Spring Harbor Laboratory Press Cold Spring Harbor, NY).

Brody, J.R., and Kern, S.E. (2004). Sodium boric acid: a Tris-free, cooler conductive medium for DNA electrophoresis. BioTechniques *36*, 214-217.

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. (2008a). Small CRISPR RNAs guide antiviral defense in prokaryotes. Science *321*, 960-964.

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

Brussow, H., Canchaya, C., and Hardt, W.D. (2004). Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. Microbiology and molecular biology reviews : MMBR *68*, 560-602, table of contents.

Burstein, D., Sun, C.L., Brown, C.T., Sharon, I., Anantharaman, K., Probst, A.J., Thomas, B.C., and Banfield, J.F. (2016). Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. *7*, 10613.

Cady, K.C., Bondy-Denomy, J., Heussler, G.E., Davidson, A.R., and O'Toole, G.A. (2012). The CRISPR/Cas adaptive immune system of Pseudomonas aeruginosa mediates resistance to naturally occurring and engineered phages. Journal of bacteriology *194*, 5728-5738.

Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillmann, M.L., and Brussow, H. (2003). Phage as agents of lateral gene transfer. Curr Opin Microbiol *6*, 417-424.

Cann, J.R. (1989). Phenomenological theory of gel electrophoresis of protein-nucleic acid complexes. Journal of Biological Chemistry *264*, 17032-17040.

Carroll, S.B. (2008). Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell *134*, 25-36.

Carte, J., Pfister, N.T., Compton, M.M., Terns, R.M., and Terns, M.P. (2010). Binding and cleavage of CRISPR RNA by Cas6. RNA *16*, 2181-2188.

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*, 3489-3496.

Carter, J., and Wiedenheft, B. (2015). SnapShot: CRISPR-RNA-guided adaptive immune systems. Cell *163*, 260-260 e261.

Cermak, T., Doyle, E.L., Christian, M., Wang, L., Zhang, Y., Schmidt, C., Baller, J.A., Somia, N.V., Bogdanove, A.J., and Voytas, D.F. (2011). Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. Nucleic Acids Res *39*, e82.

Chandradoss, Stanley D., Schirle, Nicole T., Szczepaniak, M., MacRae, Ian J., and Joo, C. (2015). A Dynamic Search Process Underlies MicroRNA Targeting. Cell *162*, 96-107.

Chang, Y., Su, T., Qi, Q., and Liang, Q. (2016). Easy regulation of metabolic flux in Escherichia coli using an endogenous type I-E CRISPR-Cas system. Microbial Cell Factories *15*, 195.

Chao, Y., Papenfort, K., Reinhardt, R., Sharma, C.M., and Vogel, J. (2012). An atlas of Hfq-bound transcripts reveals 3[prime] UTRs as a genomic reservoir of regulatory small RNAs. EMBO J *31*, 4005-4019.

Charpentier, E., Richter, H., van der Oost, J., and White, M.F. (2015). Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. FEMS Microbiol Rev *39*, 428-441.

Chen, B., Gilbert, L.A., Cimini, B.A., Schnitzbauer, J., Zhang, W., Li, G.W., Park, J., Blackburn, E.H., Weissman, J.S., Qi, L.S.*, et al.* (2013). Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. Cell *155*, 1479-1491.

Cho, S.W., Kim, S., Kim, J.M., and Kim, J.-S. (2013). Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *31*, 230-232.

Chorn, G., Zhao, L., Sachs, A.B., Flanagan, W.M., and Lim, L.P. (2010). Persistence of seed-based activity following segmentation of a microRNA guide strand. Rna *16*, 2336-2340.

Christiansen, J.K., Nielsen, J.S., Ebersbach, T., Valentin-Hansen, P., Søgaard-Andersen, L., and Kallipolitis, B.H. (2006). Identification of small Hfq-binding RNAs in Listeria monocytogenes. Rna *12*, 1383-1396.

Cisse, I.I., Kim, H., and Ha, T. (2012). A rule of seven in Watson-Crick base-pairing of mismatched sequences. Nature structural & molecular biology *19*, 623-627.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A.*, et al.* (2013). Multiplex genome engineering using CRISPR/Cas systems. Science *339*, 819-823.

Conrath, U., Beckers, G.J., Langenbach, C.J., and Jaskiewicz, M.R. (2015). Priming for enhanced defense. Annual review of phytopathology *53*, 97-119.

D’Astolfo, D.S., Pagliero, R.J., Pras, A., Karthaus, W.R., Clevers, H., Prasad, V., Lebbink, R.J., Rehmann, H., and Geijsen, N. (2015). Efficient intracellular delivery of native proteins. Cell *161*, 674-690.

Dahlgren, C., Zhang, H.-Y., Du, Q., Grahn, M., Norstedt, G., Wahlestedt, C., and Liang, Z. (2008). Analysis of siRNA specificity on targets with double-nucleotide mismatches. Nucleic acids research *36*, e53-e53.

Datsenko, K.A., Pougach, K., Tikhonov, A., Wanner, B.L., Severinov, K., and Semenova, E. (2012). Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. Nature Communications *3*.

Datsenko, K.A., and Wanner, B.L. (2000). One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proceedings of the National Academy of Sciences of the United States of America *97*, 6640-6645.

De Robertis, E. (2008). Evo-devo: variations on ancestral themes. Cell *132*, 185-195.

Dekking, M. (2005). A modern introduction to probability and statistics : understanding why and how (London: Springer).

Deltcheva, E., Chylinski, K., Sharma, C.M., Gonzales, K., Chao, Y., Pirzada, Z.A., Eckert, M.R., Vogel, J., and Charpentier, E. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature *471*, 602-607.

Deng, L., Garrett, R.A., Shah, S.A., Peng, X., and She, Q. (2013). A novel interference mechanism by a type IIIB CRISPR-Cmr module in Sulfolobus. Mol Microbiol *87*, 1088-1099.

Desjarlais, J.R., and Berg, J.M. (1992). Redesigning the DNA-binding specificity of a zinc finger protein: a data base-guided approach. Proteins *12*, 101-104.

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*, 1390-1400.

DiEuliis, D., and Giordano, J. (2017). Why Gene Editors Like CRISPR/Cas May Be a Game-Changer for Neuroweapons. Health security.

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

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

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

Doench, J.G., and Sharp, P.A. (2004). Specificity of microRNA target selection in translational repression. Genes & development *18*, 504-511.

Dong, D., Ren, K., Qiu, X., Zheng, J., Guo, M., Guan, X., Liu, H., Li, N., Zhang, B., Yang, D*., et al.* (2016). The crystal structure of Cpf1 in complex with CRISPR RNA. Nature *532*, 522-526.

Doyon, Y., Vo, T.D., Mendel, M.C., Greenberg, S.G., Wang, J., Xia, D.F., Miller, J.C., Urnov, F.D., Gregory, P.D., and Holmes, M.C. (2011). Enhancing zinc-finger-nuclease activity with improved obligate heterodimeric architectures. *8*, 74-79.

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

Dy, R.L., Pitman, A.R., and Fineran, P.C. (2013). Chromosomal targeting by CRISPR-Cas systems can contribute to genome plasticity in bacteria. Mobile genetic elements *3*, e26831.

East-Seletsky, A., O'Connell, M.R., Knight, S.C., Burstein, D., Cate, J.H., Tjian, R., and Doudna, J.A. (2016). Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. Nature *538*, 270-273.

Ebina, H., Misawa, N., Kanemura, Y., and Koyanagi, Y. (2013). Harnessing the CRISPR/Cas9 system to disrupt latent HIV-1 provirus. Sci Rep *3*, 2510.

Elbashir, S.M., Martinez, J., Patkaniowska, A., Lendeckel, W., and Tuschl, T. (2001). Functional anatomy of siRNAs for mediating efficient RNAi in Drosophila melanogaster embryo lysate. The EMBO journal *20*, 6877-6888.

Elkayam, E., Kuhn, C.-D., Tocilj, A., Haase, A.D., Greene, E.M., Hannon, G.J., and Joshua-Tor, L. (2012). The structure of human Argonaute-2 in complex with miR-20a. Cell *150*, 100-110.

Elmore, J.R., Sheppard, N.F., Ramia, N., Deighan, T., Li, H., Terns, R.M., and Terns, M.P. (2016). Bipartite recognition of target RNAs activates DNA cleavage by the Type III-B CRISPR-Cas system. Genes Dev *30*, 447-459.

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

Estrella, M.A., Kuo, F.T., and Bailey, S. (2016). RNA-activated DNA cleavage by the Type III-B CRISPR-Cas effector complex. Genes Dev *30*, 460-470.

Esvelt, K.M., Mali, P., Braff, J.L., Moosburner, M., Yaung, S.J., and Church, G.M. (2013). Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. Nature methods *10*, 1116-1121.

Fagerlund, R.D., Wilkinson, M.E., Klykov, O., Barendregt, A., Pearce, F.G., Kieper, S.N., Maxwell, H.W.R., Capolupo, A., Heck, A.J.R., Krause, K.L*., et al.* (2017). Spacer capture and integration by a type I-F Cas1–Cas2-3 CRISPR adaptation complex. Proceedings of the National Academy of Sciences.

Farasat, I., and Salis, H.M. (2016). A Biophysical Model of CRISPR/Cas9 Activity for Rational Design of Genome Editing and Gene Regulation. PLoS computational biology *12*, e1004724.

Fender, A., Elf, J., Hampel, K., Zimmermann, B., and Wagner, E.G.H. (2010). RNAs actively cycle on the Sm-like protein Hfq. Genes & development *24*, 2621-2626.

Filee, J., Forterre, P., and Laurent, J. (2003). The role played by viruses in the evolution of their hosts: a view based on informational protein phylogenies. Res Microbiol *154*, 237-243.

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

Fineran, P.C., Gerritzen, M.J., Suarez-Diez, M., Künne, T., Boekhorst, J., van Hijum, S.A., Staals, R.H., and Brouns, S.J. (2014). Degenerate target sites mediate rapid primed CRISPR adaptation. Proceedings of the National Academy of Sciences of the United States of America *111*, E1629-1638.

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

Fried, M.G. (1989). Measurement of protein-DNA interaction parameters by electrophoresis mobility shift assay. Electrophoresis *10*, 366-376.

Fried, M.G., and Bromberg, J.L. (1997). Factors that affect the stability of protein-DNA complexes during gel electrophoresis. Electrophoresis *18*, 6-11.

Fried, M.G., and Daugherty, M.A. (1998). Electrophoretic analysis of multiple protein-DNA interactions. Electrophoresis *19*, 1247-1253.

Friedman, R.C., Farh, K.K.-H., Burge, C.B., and Bartel, D.P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. Genome research *19*, 92-105.

Frohlich, K.S., Papenfort, K., Fekete, A., and Vogel, J. (2013). A small RNA activates CFA synthase by isoform-specific mRNA stabilization. EMBO J *32*, 2963-2979.

Fu, Y., Sander, J.D., Reyon, D., Cascio, V.M., and Joung, J.K. (2014). Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. Nat Biotechnol *32*, 279-284.

Gao, P., Yang, H., Rajashankar, K.R., Huang, Z., and Patel, D.J. (2016). Type V CRISPR-Cas Cpf1 endonuclease employs a unique mechanism for crRNA-mediated target DNA recognition. Cell Res *26*, 901-913.

Garneau, J.E., Dupuis, M.-E., Villion, M., Romero, D.A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadan, A.H., and Moineau, S. (2010). The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature *468*, 67-+.

Gasiunas, G., Barrangou, R., Horvath, P., and Siksnys, V. (2012). Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proceedings of the National Academy of Sciences *109*, E2579-E2586.

Geissmann, T., Chevalier, C., Cros, M.-J., Boisset, S., Fechter, P., Noirot, C., Schrenzel, J., François, P., Vandenesch, F., and Gaspin, C. (2009). A search for small noncoding RNAs in Staphylococcus aureus reveals a conserved sequence motif for regulation. Nucleic acids research *37*, 7239-7257.

Georg, J., and Hess, W.R. (2011). cis-antisense RNA, another level of gene regulation in bacteria. Microbiology and Molecular Biology Reviews *75*, 286-300.

Gilbert, L.A., Larson, M.H., Morsut, L., Liu, Z., Brar, G.A., Torres, S.E., Stern-Ginossar, N., Brandman, O., Whitehead, E.H., Doudna, J.A.*, et al.* (2013). CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell *154*, 442-451.

Gleditzsch, D., Muller-Esparza, H., Pausch, P., Sharma, K., Dwarakanath, S., Urlaub, H., Bange, G., and Randau, L. (2016). Modulating the Cascade architecture of a minimal Type I-F CRISPR-Cas system. Nucleic Acids Res *44*, 5872-5882.

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

Goldfarb, T., Sberro, H., Weinstock, E., Cohen, O., Doron, S., Charpak-Amikam, Y., Afik, S., Ofir, G., and Sorek, R. (2015). BREX is a novel phage resistance system widespread in microbial genomes. EMBO J *34*, 169-183.

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

Gootenberg, J.S., Abudayyeh, O.O., Lee, J.W., Essletzbichler, P., Dy, A.J., Joung, J., Verdine, V., Donghia, N., Daringer, N.M., Freije, C.A.*, et al.* (2017). Nucleic acid detection with CRISPR-Cas13a/C2c2. Science.

Goren, M.G., Doron, S., Globus, R., Amitai, G., Sorek, R., and Qimron, U. (2016). Repeat Size Determination by Two Molecular Rulers in the Type I-E CRISPR Array. Cell reports *16*, 2811-2818.

Gorman, J., and Greene, E.C. (2008). Visualizing one-dimensional diffusion of proteins along DNA. Nat Struct Mol Biol *15*, 768-774.

Gottesman, S. (2004). The small RNA regulators of Escherichia coli: roles and mechanisms*. Annu. Rev. Microbiol. *58*, 303-328.

Gottesman, S., and Storz, G. (2011). Bacterial small RNA regulators: versatile roles and rapidly evolving variations. In Cold Spring Harb Perspect Biol.

Grimson, A., Farh, K.K.-H., Johnston, W.K., Garrett-Engele, P., Lim, L.P., and Bartel, D.P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Molecular Cell *27*, 91-105.

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

Groenen, P.M., Bunschoten, A.E., van Soolingen, D., and van Embden, J.D. (1993). Nature of DNA polymorphism in the direct repeat cluster of Mycobacterium tuberculosis; application for strain differentiation by a novel typing method. Mol Microbiol *10*, 1057-1065.

Gudbergsdottir, S., Deng, L., Chen, Z., Jensen, J.V., Jensen, L.R., She, Q., and Garrett, R.A. (2011). Dynamic properties of the Sulfolobus CRISPR/Cas and CRISPR/Cmr systems when challenged with vector-borne viral and plasmid genes and protospacers. Molecular Microbiology *79*, 35-49.

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

Guillier, M., and Gottesman, S. (2008). The 5′ end of two redundant sRNAs is involved in the regulation of multiple targets, including their own regulator. Nucleic acids research *36*, 6781-6794.

Gunderson, F.F., and Cianciotto, N.P. (2013). The CRISPR-associated gene cas2 of Legionella pneumophila is required for intracellular infection of amoebae. mBio *4*, e00074-00013.

Guo, J., Gaj, T., and Barbas Iii, C.F. (2010). Directed Evolution of an Enhanced and Highly Efficient FokI Cleavage Domain for Zinc Finger Nucleases. Journal of Molecular Biology *400*, 96-107.

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

Hale, C.R., Majumdar, S., Elmore, J., Pfister, N., Compton, M., Olson, S., Resch, A.M., Glover, C.V., 3rd, Graveley, B.R., Terns, R.M.*, et al.* (2012). Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. Mol Cell *45*, 292-302.

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*, 945-956.

Haley, B., and Zamore, P.D. (2004). Kinetic analysis of the RNAi enzyme complex. Nature structural & molecular biology *11*, 599-606.

Hall, K.B., and McLaughlin, L.W. (1991). THERMODYNAMIC AND STRUCTURAL-PROPERTIES OF PENTAMER DNA-DNA, RNA-RNA, AND DNA-RNA DUPLEXES OF IDENTICAL SEQUENCE. Biochemistry *30*, 10606-10613.

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

Han, K., Kim, K.-s., Bak, G., Park, H., and Lee, Y. (2010). Recognition and discrimination of target mRNAs by Sib RNAs, a cis-encoded sRNA family. Nucleic acids research *38*, 5851-5866.

Hao, Y., Zhang, Z.J., Erickson, D.W., Huang, M., Huang, Y., Li, J., Hwa, T., and Shi, H. (2011). Quantifying the sequence–function relation in gene silencing by bacterial small RNAs. Proceedings of the National Academy of Sciences *108*, 12473-12478.

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

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

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

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

Heler, R., Wright, A.V., Vucelja, M., Bikard, D., Doudna, J.A., and Marraffini, L.A. (2017). Mutations in Cas9 Enhance the Rate of Acquisition of Viral Spacer Sequences during the CRISPR-Cas Immune Response. Mol Cell *65*, 168-175.

Helwa, R., and Hoheisel, J. (2010). Analysis of DNA–protein interactions: from nitrocellulose filter binding assays to microarray studies. Anal Bioanal Chem *398*, 2551-2561.

Hibio, N., Hino, K., Shimizu, E., Nagata, Y., and Ui-Tei, K. (2012). Stability of miRNA 5′terminal and seed regions is correlated with experimentally observed miRNA-mediated silencing efficacy. In Sci. Rep. (Macmillan Publishers Limited. All rights reserved).

Hilton, I.B., D'Ippolito, A.M., Vockley, C.M., Thakore, P.I., Crawford, G.E., Reddy, T.E., and Gersbach, C.A. (2015). Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. Nat Biotechnol *33*, 510-517.

Hirano, H., Gootenberg, J.S., Horii, T., Abudayyeh, O.O., Kimura, M., Hsu, P.D., Nakane, T., Ishitani, R., Hatada, I., Zhang, F.*, et al.* (2016). Structure and Engineering of Francisella novicida Cas9. Cell *164*, 950-961.

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

Hoe, N., Nakashima, K., Grigsby, D., Pan, X., Dou, S.J., Naidich, S., Garcia, M., Kahn, E., Bergmire-Sweat, D., and Musser, J.M. (1999). Rapid molecular genetic subtyping of serotype M1 group A Streptococcus strains. Emerg Infect Dis *5*, 254-263.

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

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

Hou, Z., Zhang, Y., Propson, N.E., Howden, S.E., Chu, L.F., Sontheimer, E.J., and Thomson, J.A. (2013). Efficient genome engineering in human pluripotent stem cells using Cas9 from Neisseria meningitidis. Proceedings of the National Academy of Sciences of the United States of America *110*, 15644-15649.

Howard, J.A.L., Delmas, S., Ivančić‒Baće, I., and Bolt, E.L. (2011). Helicase dissociation and annealing of RNA-DNA hybrids by Escherichia coli Cas3 protein. Biochemical Journal *439*, 85-95.

Hoyland-Kroghsbo, N.M., Paczkowski, J., Mukherjee, S., Broniewski, J., Westra, E., Bondy-Denomy, J., and Bassler, B.L. (2017). Quorum sensing controls the Pseudomonas aeruginosa CRISPR-Cas adaptive immune system. Proceedings of the National Academy of Sciences of the United States of America *114*, 131-135.

Hsu, P.D., Scott, D.A., Weinstein, J.A., Ran, F.A., Konermann, S., Agarwala, V., Li, Y., Fine, E.J., Wu, X., Shalem, O.*, et al.* (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol *31*, 827-832.

Hu, W., Kaminski, R., Yang, F., Zhang, Y., Cosentino, L., Li, F., Luo, B., Alvarez-Carbonell, D., Garcia-Mesa, Y., Karn, J.*, et al.* (2014). RNA-directed gene editing specifically eradicates latent and prevents new HIV-1 infection. Proceedings of the National Academy of Sciences of the United States of America *111*, 11461-11466.

Huang, Y., Chen, C., and Russu, I.M. (2009). Dynamics and stability of individual base pairs in two homologous RNA-DNA hybrids. Biochemistry *48*, 3988-3997.

Huo, Y., Nam, K.H., Ding, F., Lee, H., Wu, L., Xiao, Y., Farchione, M.D., Jr., Zhou, S., Rajashankar, K., Kurinov, I.*, et al.* (2014). Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. Nat Struct Mol Biol *21*, 771-777.

Hussein, R., and Lim, H.N. (2011). Disruption of small RNA signaling caused by competition for Hfq. Proceedings of the National Academy of Sciences *108*, 1110-1115.

Hutvágner, G., and Zamore, P.D. (2002). A microRNA in a Multiple-Turnover RNAi Enzyme Complex. Science *297*, 2056-2060.

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. Journal of bacteriology *169*, 5429-5433.

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

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

Jackson, R.N., and Wiedenheft, B. (2015). A Conserved Structural Chassis for Mounting Versatile CRISPR RNA-Guided Immune Responses. Mol Cell *58*, 722-728.

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

Jalasvuori, M., and Koonin, E.V. (2015). Classification of prokaryotic genetic replicators: between selfishness and altruism. Ann. N.Y. Acad. Sci. *1341*, 96-105.

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

Jansen, R., Embden, J.D.v., Gaastra, W., and Schouls, L. (2002b). Identification of a novel family of sequence repeats among prokaryotes. OMICS *6(1)*, 23-33.

Jerabek-Willemsen, M., Wienken, C.J., Braun, D., Baaske, P., and Duhr, S. (2011). Molecular interaction studies using microscale thermophoresis. Assay and drug development technologies *9*, 342-353.

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

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science *337*, 816-821.

Jinek, M., East, A., Cheng, A., Lin, S., Ma, E., and Doudna, J. (2013). RNA-programmed genome editing in human cells. eLife *2*, e00471.

Jinek, M., Jiang, F., Taylor, D.W., Sternberg, S.H., Kaya, E., Ma, E., Anders, C., Hauer, M., Zhou, K., Lin, S*., et al.* (2014). Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. Science *343*, 1247997.

Johansen, J., Eriksen, M., Kallipolitis, B., and Valentin-Hansen, P. (2008). Down-regulation of outer membrane proteins by noncoding RNAs: unraveling the cAMP-CRP-and sigmaE-dependent CyaR-ompX regulatory case. Journal of molecular biology *383*, 1.

Jore, M.M., Lundgren, M., van Duijn, E., Bultema, J.B., Westra, E.R., Waghmare, S.P., Wiedenheft, B., Pul, Ü., Wurm, R., and Wagner, R. (2011a). Structural basis for CRISPR RNA-guided DNA recognition by Cascade. Nature structural & molecular biology *18*, 529-536.

Jore, M.M., Lundgren, M., van Duijn, E., Bultema, J.B., Westra, E.R., Waghmare, S.P., Wiedenheft, B., Pul, U., Wurm, R., Wagner, R*., et al.* (2011b). Structural basis for CRISPR RNA-guided DNA recognition by Cascade. Nature Structural & Molecular Biology *18*, 529-U141.

Juliano, C., Wang, J., and Lin, H. (2011). Uniting germline and stem cells: the function of Piwi proteins and the piRNA pathway in diverse organisms. Annual review of genetics *45*, 447-469.

Kawamata, T., Seitz, H., and Tomari, Y. (2009). Structural determinants of miRNAs for RISC loading and slicer-independent unwinding. Nature structural & molecular biology *16*, 953-960.

Kawamoto, H., Koide, Y., Morita, T., and Aiba, H. (2006). Base-pairing requirement for RNA silencing by a bacterial small RNA and acceleration of duplex formation by Hfq. Molecular Microbiology *61*, 1013-1022.

Kearns, N.A., Pham, H., Tabak, B., Genga, R.M., Silverstein, N.J., Garber, M., and Maehr, R. (2015). Functional annotation of native enhancers with a Cas9-histone demethylase fusion. Nature methods *12*, 401-403.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C*., et al.* (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics *28*, 1647-1649.

Ketting, R.F. (2011). The many faces of RNAi. Developmental cell *20*, 148-161.

Khatodia, S., Bhatotia, K., Passricha, N., Khurana, S.M.P., and Tuteja, N. (2016). The CRISPR/Cas Genome-Editing Tool: Application in Improvement of Crops. Frontiers in Plant Science *7*.

Khvorova, A., Reynolds, A., and Jayasena, S.D. (2003). Functional siRNAs and miRNAs exhibit strand bias. Cell *115*, 209-216.

Kim, D., Kim, J., Hur, J.K., Been, K.W., Yoon, S.H., and Kim, J.S. (2016). Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells. Nat Biotechnol *34*, 863-868.

Kim, H.J., Lee, H.J., Kim, H., Cho, S.W., and Kim, J.-S. (2009). Targeted genome editing in human cells with zinc finger nucleases constructed via modular assembly. Genome Research *19*, 1279-1288.

Kim, J.-S. (2016). Genome editing comes of age. *11*, 1573-1578.

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

Kim, Y., Kweon, J., Kim, A., Chon, J.K., Yoo, J.Y., Kim, H.J., Kim, S., Lee, C., Jeong, E., Chung, E., *et al.* (2013). A library of TAL effector nucleases spanning the human genome. Nat Biotechnol *31*, 251-258.

Kim, Y.G., Cha, J., and Chandrasegaran, S. (1996). Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. Proceedings of the National Academy of Sciences of the United States of America *93*, 1156-1160.

Kleinstiver, B.P., Pattanayak, V., Prew, M.S., Tsai, S.Q., Nguyen, N.T., Zheng, Z., and Joung, J.K. (2016). High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. Nature *529*, 490-495.

Kleinstiver, B.P., Prew, M.S., Topkar, V.V., Tsai, S.Q., and Joung, J.K. (2015). 58. Engineered Cas9

Variants with Novel PAM Specificities Expand the Targeting Range of CRISPR/Cas Nucleases. Molecular Therapy *23, Supplement 1*, S26.

Konermann, S., Brigham, M.D., Trevino, A.E., Hsu, P.D., Heidenreich, M., Cong, L., Platt, R.J., Scott, D.A., Church, G.M., and Zhang, F. (2013). Optical control of mammalian endogenous transcription and epigenetic states. Nature *500*, 472-476.

Koonin, E.V. (2016). Viruses and mobile elements as drivers of evolutionary transitions. Philosophical Transactions of the Royal Society B: Biological Sciences *371*.

Koonin, E.V., and Dolja, V.V. (2013). A virocentric perspective on the evolution of life. Current opinion in virology *3*, 546-557.

Koonin, E.V., Makarova, K.S., and Zhang, F. (2017). Diversity, classification and evolution of CRISPR-Cas systems. Curr Opin Microbiol *37*, 67-78.

Künne, T., Kieper, S.N., Bannenberg, J.W., Vogel, A.I., Miellet, W.R., Klein, M., Depken, M., Suarez-Diez, M., and Brouns, S.J. (2016). Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation. Mol Cell *63*, 852-864.

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

Künne, T., Westra, E.R., and Brouns, S.J. (2015). Electrophoretic Mobility Shift Assay of DNA and CRISPR-Cas Ribonucleoprotein Complexes. Methods in molecular biology *1311*, 171-184.

Kupczok, A., and Bollback, J.P. (2014). Motif depletion in bacteriophages infecting hosts with CRISPR systems. BMC genomics *15*, 663.

Kurtz, J., and Franz, K. (2003). Innate defence: evidence for memory in invertebrate immunity. Nature *425*, 37-38.

Kuscu, C., Arslan, S., Singh, R., Thorpe, J., and Adli, M. (2014). Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. Nat Biotechnol *32*, 677-683.

Kuznedelov, K., Mekler, V., Lemak, S., Tokmina-Lukaszewska, M., Datsenko, K.A., Jain, I., Savitskaya, E., Mallon, J., Shmakov, S., Bothner, B.*, et al.* (2016). Altered stoichiometry Escherichia coli Cascade complexes with shortened CRISPR RNA spacers are capable of interference and primed adaptation. Nucleic Acids Res *44*, 10849-10861.

Labrie, S.J., Samson, J.E., and Moineau, S. (2010). Bacteriophage resistance mechanisms. Nature reviews. Microbiology *8*, 317-327.

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

Lee, T.S., Krupa, R.A., Zhang, F., Hajimorad, M., Holtz, W.J., Prasad, N., Lee, S.K., and Keasling, J.D. (2011). BglBrick vectors and datasheets: A synthetic biology platform for gene expression. Journal of biological engineering 5, 12.

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

Lei, C., Li, S.Y., Liu, J.K., Zheng, X., Zhao, G.P., and Wang, J. (2017). The CCTL (Cpf1-assisted Cutting and Taq DNA ligase-assisted Ligation) method for efficient editing of large DNA constructs in vitro. Nucleic Acids Res.

Lemak, S., Beloglazova, N., Nocek, B., Skarina, T., Flick, R., Brown, G., Popovic, A., Joachimiak, A., Savchenko, A., and Yakunin, A.F. (2013). Toroidal structure and DNA cleavage by the CRISPR-associated [4Fe-4S] cluster containing Cas4 nuclease SSO0001 from Sulfolobus solfataricus. Journal of the American Chemical Society *135*, 17476-17487.

Lemak, S., Nocek, B., Beloglazova, N., Skarina, T., Flick, R., Brown, G., Joachimiak, A., Savchenko, A., and Yakunin, A.F. (2014). The CRISPR-associated Cas4 protein Pcal_0546 from Pyrobaculum calidifontis contains a [2Fe-2S] cluster: crystal structure and nuclease activity. Nucleic Acids Res *42*, 11144-11155.

Lenski, R.E. (1988). Experimental Studies of Pleiotropy and Epistasis in Escherichia coli. I. Variation in Competitive Fitness Among Mutants Resistant to Virus T4. Evolution *42*, 425-432.

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

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*, 15-20.

Lewis, B.P., Shih, I.h., Jones-Rhoades, M.W., Bartel, D.P., and Burge, C.B. (2003). Prediction of

Mammalian MicroRNA Targets. Cell *115*, 787-798.

Lewis, K.M., and Ke, A. (2017). Building the Class 2 CRISPR-Cas Arsenal. Mol Cell *65*, 377-379.

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

Li, S.Y., Zhao, G.P., and Wang, J. (2016). C-Brick: A New Standard for Assembly of Biological Parts Using Cpf1. ACS synthetic biology *5*, 1383-1388.

Li, Y., Pan, S., Zhang, Y., Ren, M., Feng, M., Peng, N., Chen, L., Liang, Y.X., and She, Q. (2015). Harnessing Type I and Type III CRISPR-Cas systems for genome editing. Nucleic Acids Research.

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*, 259-272.

Lin, S., Staahl, B.T., Alla, R.K., and Doudna, J.A. (2014). Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. eLife *3*, e04766.

Liu, J., Gaj, T., Yang, Y., Wang, N., Shui, S., Kim, S., Kanchiswamy, C.N., Kim, J.S., and Barbas, C.F., 3rd (2015). Efficient delivery of nuclease proteins for genome editing in human stem cells and primary cells. Nature protocols *10*, 1842-1859.

Liu, L., Chen, P., Wang, M., Li, X., Wang, J., Yin, M., and Wang, Y. (2017a). C2c1-sgRNA Complex Structure Reveals RNA-Guided DNA Cleavage Mechanism. Mol Cell *65*, 310-322.

Liu, L., Li, X., Wang, J., Wang, M., Chen, P., Yin, M., Li, J., Sheng, G., and Wang, Y. (2017b). Two Distant Catalytic Sites Are Responsible for C2c2 RNase Activities. Cell *168*, 121-134 e112.

Louwen, R., Horst-Kreft, D., de Boer, A.G., van der Graaf, L., de Knegt, G., Hamersma, M., Heikema, A.P., Timms, A.R., Jacobs, B.C., Wagenaar, J.A*., et al.* (2013). A novel link between Campylobacter jejuni bacteriophage defence, virulence and Guillain-Barre syndrome. European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology *32*, 207-226.

Louwen, R., Staals, R.H., Endtz, H.P., van Baarlen, P., and van der Oost, J. (2014). The role of CRISPR-Cas systems in virulence of pathogenic bacteria. Microbiology and molecular biology reviews : MMBR *78*, 74-88.

Luo, M.L., Jackson, R.N., Denny, S.R., Tokmina-Lukaszewska, M., Maksimchuk, K.R., Lin, W., Bothner, B., Wiedenheft, B., and Beisel, C.L. (2016). The CRISPR RNA-guided surveillance complex in Escherichia coli accommodates extended RNA spacers. Nucleic Acids Res *44*, 7385-7394.

Luo, M.L., Mullis, A.S., Leenay, R.T., and Beisel, C.L. (2015). Repurposing endogenous type I CRISPR-Cas systems for programmable gene repression. Nucleic Acids Res *43*, 674-681.

Lutze, W., and Ewing, R. (1990). Infection of phytoplankton by viruses and reduction of primary productivity. Nature *347*, 4.

Ma, X., Zhu, Q., Chen, Y., and Liu, Y.-G. (2016). CRISPR/Cas9 Platforms for Genome Editing in Plants: Developments and Applications. Molecular Plant *9*, 961-974.

Magadan, A.H., Dupuis, M.-E., Villion, M., and Moineau, S. (2012). Cleavage of Phage DNA by the Streptococcus thermophilus CRISPR3-Cas System. Plos One *7*.

Maier, L.-K., Lange, S., Stoll, B., Haas, K., Fischer, S., Fischer, E., Duchardt-Ferner, E., Wöhnert, J., Backofen, R., and Marchfelder, A. (2013). Essential requirements for the detection and degradation of invaders by the Haloferax volcanii CRISPR/Cas system IB. RNA biology *10*, 0--1.

Majsec, K., Bolt, E.L., and Ivancic-Bace, I. (2016). Cas3 is a limiting factor for CRISPR-Cas immunity in Escherichia coli cells lacking H-NS. BMC microbiology *16*, 28.

Majumdar, S., Zhao, P., Pfister, N.T., Compton, M., Olson, S., Glover, C.V., 3rd, Wells, L., Graveley, B.R., Terns, R.M., and Terns, M.P. (2015). Three CRISPR-Cas immune effector complexes coexist in Pyrococcus furiosus. RNA *21*, 1147-1158.

Makarova, K., Grishin, N., Shabalina, S., Wolf, Y., and Koonin, E. (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. Biology Direct *1*, 7.

Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J.M., Wolf, Y.I., Yakunin, A.F*., et al.* (2011). Evolution and classification of the CRISPR–Cas systems. *9*, 467-477.

Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J., Barrangou, R., Brouns, S.J., Charpentier, E., Haft, D.H*., et al.* (2015). An updated evolutionary classification of CRISPR-Cas systems. Nature reviews. Microbiology *13*, 722-736.

Makarova, K.S., Zhang, F., and Koonin, E.V. (2017a). SnapShot: Class 1 CRISPR-Cas Systems. Cell

168, 946-946 e941.

Makarova, K.S., Zhang, F., and Koonin, E.V. (2017b). SnapShot: Class 2 CRISPR-Cas Systems. Cell *168*, 328-328 e321.

Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., and Church, G.M. (2013). RNA-guided human genome engineering via Cas9. Science *339*, 823-826.

Manica, A., Zebec, Z., Steinkellner, J., and Schleper, C. (2013). Unexpectedly broad target recognition of the CRISPR-mediated virus defence system in the archaeon Sulfolobus solfataricus. Nucleic Acids Research *41*, 10509-10517.

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

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

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

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

Masepohl, B., Gorlitz, K., and Bohme, H. (1996). Long tandemly repeated repetitive (LTRR) sequences in the filamentous cyanobacterium Anabaena sp. PCC 7120. Biochimica et biophysica acta *1307*, 26-30.

Millen, A.M., Horvath, P., Boyaval, P., and Romero, D.A. (2012). Mobile CRISPR/Cas-Mediated Bacteriophage Resistance in Lactococcus lactis. PloS one *7*, e51663.

Miller, J.C., Tan, S., Qiao, G., Barlow, K.A., Wang, J., Xia, D.F., Meng, X., Paschon, D.E., Leung, E., Hinkley, S.J.*, et al.* (2011). A TALE nuclease architecture for efficient genome editing. Nat Biotechnol *29*, 143-148.

Moch, C., Fromant, M., Blanquet, S., and Plateau, P. (2016). DNA binding specificities of Escherichia coli Cas1-Cas2 integrase drive its recruitment at the CRISPR locus. Nucleic Acids Res.

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

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

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*, 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*, 85-93.

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

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

Moll, I., Leitsch, D., Steinhauser, T., and Bläsi, U. (2003). RNA chaperone activity of the Sm-like Hfq protein. EMBO reports *4*, 284-289.

Møller, T., Franch, T., Højrup, P., Keene, D.R., Bächinger, H.P., Brennan, R.G., and Valentin-Hansen, P. (2002). Hfq: a bacterial Sm-like protein that mediates RNA-RNA interaction. Molecular Cell *9*, 23-30.

Monico, C., Capitanio, M., Belcastro, G., Vanzi, F., and Pavone, F.S. (2013). Optical Methods to Study Protein-DNA Interactions in Vitro and in Living Cells at the Single-Molecule Level. International journal of molecular sciences *14*, 3961-3992.

Moon, K., and Gottesman, S. (2011). Competition among Hfq-binding small RNAs in Escherichia coli. Molecular Microbiology *82*, 1545-1562.

Mulepati, S., and Bailey, S. (2011). Structural and Biochemical Analysis of Nuclease Domain of Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR)-associated Protein 3 (Cas3). Journal of Biological Chemistry *286*, 31896-31903.

Mulepati, S., and Bailey, S. (2013). In Vitro Reconstitution of an Escherichia coli RNA-guided Immune System Reveals Unidirectional, ATP-dependent Degradation of DNA Target. The Journal of biological chemistry *288*, 22184-22192.

Mulepati, S., Heroux, A., and Bailey, S. (2014). Structural biology. Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. Science *345*, 1479-1484.

Musharova, O., Klimuk, E., Datsenko, K.A., Metlitskaya, A., Logacheva, M., Semenova, E., Severinov, K., and Savitskaya, E. (2017). Spacer-length DNA intermediates are associated with Cas1 in cells undergoing primed CRISPR adaptation. Nucleic Acids Res *45*, 3297-3307.

Nakanishi, K., Weinberg, D.E., Bartel, D.P., and Patel, D.J. (2012). Structure of yeast Argonaute with guide RNA. Nature *486*, 368-374.

Nakata, A., Amemura, M., and Makino, K. (1989). Unusual nucleotide arrangement with repeated sequences in the Escherichia coli K-12 chromosome. J Bacteriol *171*, 3553-3556.

Nam, K.H., Haitjema, C., Liu, X., Ding, F., Wang, H., DeLisa, M.P., and Ke, A. (2012a). Cas5d protein processes pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvulg CRISPR-Cas system. Structure *20*, 1574-1584.

Nam, K.H., Huang, Q., and Ke, A. (2012b). Nucleic acid binding surface and dimer interface revealed by CRISPR-associated CasB protein structures. FEBS Lett *586*, 3956-3961.

Nishimasu, H., Cong, L., Yan, W.X., Ran, F.A., Zetsche, B., Li, Y., Kurabayashi, A., Ishitani, R., Zhang, F., and Nureki, O. (2015). Crystal Structure of Staphylococcus aureus Cas9. Cell *162*, 1113-1126.

Nishimasu, H., Ran, F.A., Hsu, P.D., Konermann, S., Shehata, S.I., Dohmae, N., Ishitani, R., Zhang, F., and Nureki, O. (2014). Crystal structure of Cas9 in complex with guide RNA and target DNA. Cell *156*, 935-949.

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

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

Nunez, J.K., Kranzusch, P.J., Noeske, J., Wright, A.V., Davies, C.W., and Doudna, J.A. (2014). Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. Nat Struct Mol Biol *21*, 528-534.

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

O'Geen, H., Henry, I.M., Bhakta, M.S., Meckler, J.F., and Segal, D.J. (2015). A genome-wide analysis of Cas9 binding specificity using ChIP-seq and targeted sequence capture. Nucleic Acids Res *43*, 3389-3404.

Osawa, T., Inanaga, H., Sato, C., and Numata, T. (2015). Crystal structure of the CRISPR-Cas RNA silencing Cmr complex bound to a target analog. Mol Cell *58*, 418-430.

Paez-Espino, D., Morovic, W., Sun, C.L., Thomas, B.C., Ueda, K., Stahl, B., Barrangou, R., and Banfield, J.F. (2013). Strong bias in the bacterial CRISPR elements that confer immunity to phage. Nat Commun *4*, 1430.

Paez-Espino, D., Sharon, I., Morovic, W., Stahl, B., Thomas, B.C., Barrangou, R., and Banfield, J.F. (2015). CRISPR immunity drives rapid phage genome evolution in Streptococcus thermophilus. mBio *6*.

Papenfort, K., Bouvier, M., Mika, F., Sharma, C.M., and Vogel, J. (2010). Evidence for an autonomous 5' target recognition domain in an Hfq-associated small RNA. Proceedings of the National Academy of Sciences *107*, 20435-20440.

Papenfort, K., Pfeiffer, V., Lucchini, S., Sonawane, A., Hinton, J.C., and Vogel, J. (2008). Systematic deletion of Salmonella small RNA genes identifies CyaR, a conserved CRP-dependent riboregulator of OmpX synthesis. Molecular Microbiology *68*, 890-906.

Papenfort, K., Podkaminski, D., Hinton, J.C., and Vogel, J. (2012). The ancestral SgrS RNA discriminates horizontally acquired Salmonella mRNAs through a single GU wobble pair. Proceedings of the National Academy of Sciences *109*, E757-E764.

Papenfort, K., Sun, Y., Miyakoshi, M., Vanderpool, Carin K., and Vogel, J. (2013). Small RNA-Mediated Activation of Sugar Phosphatase mRNA Regulates Glucose Homeostasis. Cell *153*, 426-437.

Papenfort, K., and Vogel, J. (2009). Multiple target regulation by small noncoding RNAs rewires gene expression at the post-transcriptional level. Research in microbiology *160*, 278-287.

Pattanayak, V., Lin, S., Guilinger, J.P., Ma, E., Doudna, J.A., and Liu, D.R. (2013). High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. Nat Biotechnol *31*, 839-843.

Patterson, A.G., Jackson, S.A., Taylor, C., Evans, G.B., Salmond, G.P., Przybilski, R., Staals, R.H., and Fineran, P.C. (2016). Quorum Sensing Controls Adaptive Immunity through the Regulation of Multiple

CRISPR-Cas Systems. Mol Cell *64*, 1102-1108.

Patterson, A.G., Yevstigneyeva, M.S., and Fineran, P.C. (2017). Regulation of CRISPR-Cas adaptive immune systems. Curr Opin Microbiol *37*, 1-7.

Paul, J.W., and Qi, Y. (2016). CRISPR/Cas9 for plant genome editing: accomplishments, problems and prospects. Plant Cell Reports *35*, 1417-1427.

Pawluk, A., Bondy-Denomy, J., Cheung, V.H., Maxwell, K.L., and Davidson, A.R. (2014). A new group of phage anti-CRISPR genes inhibits the type I-E CRISPR-Cas system of Pseudomonas aeruginosa. mBio *5*, e00896.

Pawluk, A., Staals, R.H., Taylor, C., Watson, B.N., Saha, S., Fineran, P.C., Maxwell, K.L., and Davidson, A.R. (2016). Inactivation of CRISPR-Cas systems by anti-CRISPR proteins in diverse bacterial species. Nature microbiology *1*, 16085.

Peer, A., and Margalit, H. (2011). Accessibility and evolutionary conservation mark bacterial small-RNA target-binding regions. Journal of bacteriology *193*, 1690-1701.

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

Perez-Rodriguez, R., Haitjema, C., Huang, Q., Nam, K.H., Bernardis, S., Ke, A., and DeLisa, M.P. (2011). Envelope stress is a trigger of CRISPR RNA-mediated DNA silencing in Escherichia coli. Molecular Microbiology *79*, 584-599.

Pfeiffer, V., Papenfort, K., Lucchini, S., Hinton, J.C., and Vogel, J. (2009). Coding sequence targeting by MicC RNA reveals bacterial mRNA silencing downstream of translational initiation. Nature structural & molecular biology *16*, 840-846.

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

Plagens, A., Tripp, V., Daume, M., Sharma, K., Klingl, A., Hrle, A., Conti, E., Urlaub, H., and Randau, L. (2014). In vitro assembly and activity of an archaeal CRISPR-Cas type I-A Cascade interference complex. Nucleic Acids Res *42*, 5125-5138.

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. Journal of virology *80*, 8081-8088.

Porteus, M. (2016). Genome Editing: A New Approach to Human Therapeutics. Annual Review of Pharmacology and Toxicology *56*, 163-190.

Pougach, K., Semenova, E., Bogdanova, E., Datsenko, K.A., Djordjevic, M., Wanner, B.L., and Severinov, K. (2010). Transcription, processing and function of CRISPR cassettes in Escherichia coli. Mol Microbiol *77*, 1367-1379.

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*, 653-663.

Proctor, L., Fuhrman, J., and Ledbetter, M. (1988). Marine bacteriophages and bacterial mortality. Eos *69*, 1111-1112.

Proctor, L.M., and Fuhrman, J.A. (1990). Viral mortality of marine bacteria and cyanobacteria.

Pul, Ü., 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. Molecular Microbiology *75*, 1495-1512.

Pusch, O., Boden, D., Silbermann, R., Lee, F., Tucker, L., and Ramratnam, B. (2003). Nucleotide sequence homology requirements of HIV-1-specific short hairpin RNA. Nucleic acids research *31*, 6444-6449.

Qi, L.S., Larson, M.H., Gilbert, L.A., Doudna, J.A., Weissman, J.S., Arkin, A.P., and Lim, W.A. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell *152*, 1173-1183.

Quax, T.E., Voet, M., Sismeiro, O., Dillies, M.A., Jagla, B., Coppee, J.Y., Sezonov, G., Forterre, P., van der Oost, J., Lavigne, R.*, et al.* (2013). Massive activation of archaeal defense genes during viral infection. Journal of virology *87*, 8419-8428.

R-Development-Core-Team (2008). R: A language and environment for statistical computing (Vienna).

Ramakrishna, S., Kwaku Dad, A.B., Beloor, J., Gopalappa, R., Lee, S.K., and Kim, H. (2014). Gene disruption by cell-penetrating peptide-mediated delivery of Cas9 protein and guide RNA. Genome Res *24*, 1020-1027.

Ramanan, V., Shlomai, A., Cox, D.B., Schwartz, R.E., Michailidis, E., Bhatta, A., Scott, D.A., Zhang, F., Rice, C.M., and Bhatia, S.N. (2015). CRISPR/Cas9 cleavage of viral DNA efficiently suppresses hepatitis B virus. Sci Rep *5*, 10833.

Ran, F.A., Cong, L., Yan, W.X., Scott, D.A., Gootenberg, J.S., Kriz, A.J., Zetsche, B., Shalem, O., Wu, X., Makarova, K.S.*, et al.* (2015). In vivo genome editing using Staphylococcus aureus Cas9. *520*, 186-191.

Ran, F.A., Hsu, Patrick D., Lin, C.-Y., Gootenberg, Jonathan S., Konermann, S., Trevino, A.E., Scott, David A., Inoue, A., Matoba, S., Zhang, Y.*, et al.* (2013a). Double Nicking by RNA-Guided CRISPR Cas9 for Enhanced Genome Editing Specificity. Cell *154*, 1380-1389.

Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., and Zhang, F. (2013b). Genome engineering using the CRISPR-Cas9 system. Nat. Protocols *8*, 2281-2308.

Rath, D., Amlinger, L., Hoekzema, M., Devulapally, P.R., and Lundgren, M. (2014). Efficient programmable gene silencing by Cascade. Nucleic Acids Research.

Rebar, E.J., and Pabo, C.O. (1994). Zinc finger phage: affinity selection of fingers with new DNA-binding specificities. Science *263*, 671-673.

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

Reeks, J., Naismith, J.H., and White, M.F. (2013). CRISPR interference: a structural perspective. The Biochemical journal *453*, 155-166.

Reyon, D., Tsai, S.Q., Khayter, C., Foden, J.A., Sander, J.D., and Joung, J.K. (2012). FLASH assembly of TALENs for high-throughput genome editing. Nat Biotechnol *30*, 460-465.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends in genetics : TIG *16*, 276-277.

Richter, C., Chang, J.T., and Fineran, P.C. (2012a). Function and Regulation of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) / CRISPR Associated (Cas) Systems. Viruses-Basel *4*, 2291-2311.

Richter, C., Dy, R.L., McKenzie, R.E., Watson, B.N., Taylor, C., Chang, J.T., McNeil, M.B., Staals, R.H., and Fineran, P.C. (2014). Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. Nucleic Acids Res *42*, 8516-8526.

Richter, H., Lange, S.J., Backofen, R., and Randau, L. (2013). Comparative analysis ofCas6b processing and CRISPR RNA stability. RNA Biol *10*, 700-707.

Richter, H., Zoephel, J., Schermuly, J., Maticzka, D., Backofen, R., and Randau, L. (2012b). Characterization of CRISPR RNA processing in Clostridium thermocellum and Methanococcus maripaludis. Nucleic Acids Res *40*, 9887-9896.

Roberts, R.W., and Crothers, D.M. (1992). STABILITY AND PROPERTIES OF DOUBLE AND TRIPLE HELICES - DRAMATIC EFFECTS OF RNA OR DNA BACKBONE COMPOSITION. Science *258*, 1463-1466.

Rohwer, F., and Thurber, R.V. (2009). Viruses manipulate the marine environment. *459*, 207-212.

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

Rollins, M.F., Chowdhury, S., Carter, J., Golden, S.M., Wilkinson, R.A., Bondy-Denomy, J., Lander, G.C., and Wiedenheft, B. (2017). Cas1 and the Csy complex are opposing regulators of Cas2/3 nuclease activity. Proceedings of the National Academy of Sciences of the United States of America.

Rose, R.E. (1988). THE NUCLEOTIDE-SEQUENCE OF PACYC184. Nucleic Acids Research *16*, 355-355.

Rouet, P., Smih, F., and Jasin, M. (1994). Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. Mol Cell Biol *14*, 8096-8106.

Rouillon, C., Zhou, M., Zhang, J., Politis, A., Beilsten-Edmands, V., Cannone, G., Graham, S., Robinson, C.V., Spagnolo, L., and White, M.F. (2013). Structure of the CRISPR interference complex CSM reveals key similarities with cascade. Mol Cell *52*, 124-134.

Rudin, N., and Haber, J.E. (1988). Efficient repair of HO-induced chromosomal breaks in Saccharomyces cerevisiae by recombination between flanking homologous sequences. Mol Cell Biol *8*, 3918-3928.

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

Samai, P., Pyenson, N., Jiang, W., Goldberg, G.W., Hatoum-Aslan, A., and Marraffini, L.A. (2015). Co-transcriptional DNA and RNA Cleavage during Type III CRISPR-Cas Immunity. Cell *161*, 1164-1174.

Sampson, T.R., Napier, B.A., Schroeder, M.R., Louwen, R., Zhao, J., Chin, C.Y., Ratner, H.K., Llewellyn, A.C., Jones, C.L., Laroui, H., *et al.* (2014). A CRISPR-Cas system enhances envelope integrity mediating antibiotic resistance and inflammasome evasion. Proceedings of the National Academy of Sciences of the United States of America *111*, 11163-11168.

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

Samson, J.E., Magadan, A.H., Sabri, M., and Moineau, S. (2013). Revenge of the phages: defeating bacterial defences. Nature reviews. Microbiology *11*, 675-687.

Sapranauskas, R., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P., and Siksnys, V. (2011). The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. Nucleic Acids Res *39*, 9275-9282.

Sashital, D.G., Jinek, M., and Doudna, J.A. (2011). An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3. Nature structural & molecular biology *18*, 680-687.

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

Sauer, E. (2013). Structure and RNA-binding properties of the bacterial LSm protein Hfq. RNA biology *10*, 12-11.

Savitskaya, E., Lopatina, A., Medvedeva, S., Kapustin, M., Shmakov, S., Tikhonov, A., Artamonova, II, Logacheva, M., and Severinov, K. (2016). Dynamics of Escherichia coli type I-E CRISPR spacers over 42 000 years. Molecular ecology.

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

Schaefer, K.A., Wu, W.-H., Colgan, D.F., Tsang, S.H., Bassuk, A.G., and Mahajan, V.B. (2017). Unexpected mutations after CRISPR-Cas9 editing in vivo. *14*, 547-548.

Schirle, N.T., and MacRae, I.J. (2012). The crystal structure of human Argonaute2. Science *336*, 1037-1040.

Schmid-Hempel, P. (2005). Evolutionary ecology of insect immune defenses. Annual review of entomology *50*, 529-551.

Scholz, I., Lange, S.J., Hein, S., Hess, W.R., and Backofen, R. (2013). CRISPR-Cas systems in the cyanobacterium Synechocystis sp. PCC6803 exhibit distinct processing pathways involving at least two Cas6 and a Cmr2 protein. PLoS One *8*, e56470.

Schwarz, D.S., Ding, H., Kennington, L., Moore, J.T., Schelter, J., Burchard, J., Linsley, P.S., Aronin, N., Xu, Z., and Zamore, P.D. (2006). Designing siRNA that distinguish between genes that differ by a single nucleotide. PLoS genetics *2*, e140.

Schwarz, D.S., Hutvágner, G., Du, T., Xu, Z., Aronin, N., and Zamore, P.D. (2003). Asymmetry in the assembly of the RNAi enzyme complex. Cell *115*, 199-208.

Seed, K.D., Lazinski, D.W., Calderwood, S.B., and Camilli, A. (2013). A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. Nature *494*, 489-491.

Semenova, E., Jore, M.M., Datsenko, K.A., Semenova, A., Westra, E.R., Wanner, B., van der Oost, J., Brouns, S.J.J., and Severinov, K. (2011). Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proceedings of the National Academy of Sciences of the United States of America *108*, 10098-10103.

Semenova, E., Kuznedelov, K., Datsenko, K.A., Boudry, P.M., Savitskaya, E.E., Medvedeva, S., Beloglazova, N., Logacheva, M., Yakunin, A.F., and Severinov, K. (2015). The Cas6e ribonuclease is not required for interference and adaptation by the E. coli type I-E CRISPR-Cas system. Nucleic Acids Res *43*, 6049-6061.

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

Severinov, K., Ispolatov, I., and Semenova, E. (2016). The Influence of Copy-Number of Targeted Extrachromosomal Genetic Elements on the Outcome of CRISPR-Cas Defense. Frontiers in molecular biosciences *3*, 45.

Shah, S.A., Erdmann, S., Mojica, F.J.M., and Garrett, R.A. (2013). Protospacer recognition motifs:

Mixed identities and functional diversity. Rna Biology *10*, 891-899.

Shao, Y., Feng, L., Rutherford, S.T., Papenfort, K., and Bassler, B.L. (2013). Functional determinants of the quorum-sensing non-coding RNAs and their roles in target regulation. The EMBO journal *32*, 2158-2171.

Shao, Y., and Wang, I.N. (2009). Effect of late promoter activity on bacteriophage lambda fitness. Genetics *181*, 1467-1475.

Shen, B., Zhang, W., Zhang, J., Zhou, J., Wang, J., Chen, L., Wang, L., Hodgkins, A., Iyer, V., Huang, X.*, et al.* (2014). Efficient genome modification by CRISPR-Cas9 nickase with minimal off-target effects. *11*, 399-402.

Shinkai, A., Kira, S., Nakagawa, N., Kashihara, A., Kuramitsu, S., and Yokoyama, S. (2007). Transcription activation mediated by a cyclic AMP receptor protein from Thermus thermophilus HB8. J Bacteriol *189*, 3891-3901.

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

Shmakov, S., Abudayyeh, O.O., Makarova, K.S., Wolf, Y.I., Gootenberg, J.S., Semenova, E., Minakhin, L., Joung, J., Konermann, S., Severinov, K.*, et al.* (2015). Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. Mol Cell *60*, 385-397.

Shmakov, S., Savitskaya, E., Semenova, E., Logacheva, M.D., Datsenko, K.A., and Severinov, K. (2014). Pervasive generation of oppositely oriented spacers during CRISPR adaptation. Nucleic Acids Res *42*, 5907-5916.

Shmakov, S., Smargon, A., Scott, D., Cox, D., Pyzocha, N., Yan, W., Abudayyeh, O.O., Gootenberg, J.S., Makarova, K.S., Wolf, Y.I.*, et al.* (2017). Diversity and evolution of class 2 CRISPR-Cas systems. Nature reviews. Microbiology *15*, 169-182.

Shubin, N., Tabin, C., and Carroll, S. (2009). Deep homology and the origins of evolutionary novelty. Nature *457*, 818-823.

Sieburth, J.M., Johnson, P.W., and Hargraves, P.E. (1988). ULTRASTRUCTURE AND ECOLOGY OF AUREOCOCCUS ANOPHAGEFERENS GEN. ET SP. NOV. (CHRYSOPHYCEAE): THE DOMINANT PICOPLANKTER DURING A BLOOM IN NARRAGANSETT BAY, RHODE ISLAND, SUMMER 19851. Journal of Phycology *24*, 416-425.

Sinkunas, T., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P., and Siksnys, V. (2011). Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *30*, 1335-1342.

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

Siomi, H., and Siomi, M.C. (2009). On the road to reading the RNA-interference code. Nature *457*, 396-404.

Slaymaker, I.M., Gao, L., Zetsche, B., Scott, D.A., Yan, W.X., and Zhang, F. (2016). Rationally engineered Cas9 nucleases with improved specificity. Science *351*, 84-88.

Soper, T.J., Doxzen, K., and Woodson, S.A. (2011). Major role for mRNA binding and restructuring in sRNA recruitment by Hfq. Rna *17*, 1544-1550.

Sorek, R., Lawrence, C.M., and Wiedenheft, B. (2013a). CRISPR-Mediated Adaptive Immune Systems in Bacteria and Archaea. In Annual Review of Biochemistry, Vol 82, R.D. Kornberg, ed., pp. 237-266.

Sorek, R., Lawrence, M., and Wiedenheft, B. (2013b). CRISPR-mediated Adaptive Immune Systems in Bacteria and Archaea. Annual Review of Biochemistry *82*.

Spilman, M., Cocozaki, A., Hale, C., Shao, Y., Ramia, N., Terns, R., Terns, M., Li, H., and Stagg, S. (2013). Structure of an RNA silencing complex of the CRISPR-Cas immune system. Mol Cell *52*, 146-152.

Staals, R.H., Jackson, S.A., Biswas, A., Brouns, S.J., Brown, C.M., and Fineran, P.C. (2016). Interference-driven spacer acquisition is dominant over naive and primed adaptation in a native CRISPR-Cas system. Nat Commun *7*, 12853.

Staals, R.H., Zhu, Y., Taylor, D.W., Kornfeld, J.E., Sharma, K., Barendregt, A., Koehorst, J.J., Vlot, M., Neupane, N., Varossieau, K.*, et al.* (2014). RNA targeting by the type III-A CRISPR-Cas Csm complex of Thermus thermophilus. Mol Cell *56*, 518-530.

Staals, R.H.J., Agari, Y., Maki-Yonekura, S., Zhu, Y., Taylor, D.W., van Duijn, E., Barendregt, A., Vlot, M., Koehorst, J.J., Sakamoto, K.*, et al.* (2013). Structure and Activity of the RNA-Targeting Type III-B

CRISPR-Cas Complex of Thermus thermophilus. Molecular Cell *52*, 135-145.

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

Stern, A., and Sorek, R. (2011). The phage-host arms race: shaping the evolution of microbes. BioEssays : news and reviews in molecular, cellular and developmental biology *33*, 43-51.

Sternberg, S.H., Redding, S., Jinek, M., Greene, E.C., and Doudna, J.A. (2014). DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature *507*, 62-67.

Sternberg, S.H., Richter, H., Charpentier, E., and Qimron, U. (2016). Adaptation in CRISPR-Cas Systems. Mol Cell *61*, 797-808.

Storz, G., Vogel, J., and Wassarman, K.M. (2011). Regulation by small RNAs in bacteria: expanding frontiers. Molecular Cell *43*, 880-891.

Sugimoto, N., Nakano, M., and Nakano, S. (2000). Thermodynamics-structure relationship of single mismatches in RNA/DNA duplexes. Biochemistry *39*, 11270-11281.

Summers, W.C. (2011). In the beginning…. Bacteriophage *1*, 50-51.

Sun, C.L., Barrangou, R., Thomas, B.C., Horvath, P., Fremaux, C., and Banfield, J.F. (2013). Phage mutations in response to CRISPR diversification in a bacterial population. Environmental Microbiology *15*, 463-470.

Suttle, C.A. (2007). Marine viruses--major players in the global ecosystem. Nature reviews. Microbiology *5*, 801-812.

Swarts, D.C., Makarova, K., Wang, Y., Nakanishi, K., Ketting, R.F., Koonin, E.V., Patel, D.J., and van der Oost, J. (2014). The evolutionary journey of Argonaute proteins. Nat Struct Mol Biol *21*, 743-753.

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

Swarts, D.C., van der Oost, J., and Jinek, M. (2017). Structural Basis for Guide RNA Processing and Seed-Dependent DNA Targeting by CRISPR-Cas12a. Mol Cell *66*, 221-233 e224.

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

Tamulaitis, G., Kazlauskiene, M., Manakova, E., Venclovas, C., Nwokeoji, A.O., Dickman, M.J., Horvath, P., and Siksnys, V. (2014). Programmable RNA shredding by the type III-A CRISPR-Cas system of Streptococcus thermophilus. Mol Cell *56*, 506-517.

Taylor, D.W., Zhu, Y., Staals, R.H., Kornfeld, J.E., Shinkai, A., van der Oost, J., Nogales, E., and Doudna, J.A. (2015). Structural biology. Structures of the CRISPR-Cmr complex reveal mode of RNA target positioning. Science *348*, 581-585.

Terns, M.P., and Terns, R.M. (2011). CRISPR-based adaptive immune systems. Current Opinion in Microbiology *14*, 321-327.

Toledo-Arana, A., Dussurget, O., Nikitas, G., Sesto, N., Guet-Revillet, H., Balestrino, D., Loh, E., Gripenland, J., Tiensuu, T., Vaitkevicius, K*., et al.* (2009). The Listeria transcriptional landscape from saprophytism to virulence. Nature *459*, 950-956.

Torrella, F., and Morita, R.Y. (1979). Evidence by electron micrographs for a high incidence of bacteriophage particles in the waters of Yaquina Bay, oregon: ecological and taxonomical implications. Appl Environ Microbiol *37*, 774-778.

Tóth, E., Weinhardt, N., Bencsura, P., Huszár, K., Kulcsár, P.I., Tálas, A., Fodor, E., and Welker, E. (2016). Cpf1 nucleases demonstrate robust activity to induce DNA modification by exploiting homology directed repair pathways in mammalian cells. Biology Direct *11*, 46.

Touchon, M., Charpentier, S., Clermont, O., Rocha, E.P., Denamur, E., and Branger, C. (2011). CRISPR distribution within the Escherichia coli species is not suggestive of immunity-associated diversifying selection. J Bacteriol *193*, 2460-2467.

Touchon, M., and Rocha, E.P.C. (2010). The Small, Slow and Specialized CRISPR and Anti-CRISPR of Escherichia and Salmonella. Plos One *5*.

Tsai, S.Q., Wyvekens, N., Khayter, C., Foden, J.A., Thapar, V., Reyon, D., Goodwin, M.J., Aryee, M.J., and Joung, J.K. (2014). Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *32*, 569-576.

Ui-Tei, K., Naito, Y., Nishi, K., Juni, A., and Saigo, K. (2008). Thermodynamic stability and Watson–Crick base pairing in the seed duplex are major determinants of the efficiency of the siRNA-based off-target effect. Nucleic acids research *36*, 7100-7109.

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*, 401-407.

van der Oost, J., Westra, E.R., Jackson, R.N., and Wiedenheft, B. (2014). Unravelling the structural and mechanistic basis of CRISPR-Cas systems. Nature reviews. Microbiology *12*, 479-492.

van Embden, J.D., van Gorkom, T., Kremer, K., Jansen, R., van Der Zeijst, B.A., and Schouls, L.M. (2000). Genetic variation and evolutionary origin of the direct repeat locus of Mycobacterium tuberculosis complex bacteria. J Bacteriol *182*, 2393-2401.

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

van Houte, S., Ekroth, A.K., Broniewski, J.M., Chabas, H., Ashby, B., Bondy-Denomy, J., Gandon, S., Boots, M., Paterson, S., Buckling, A.*, et al.* (2016). The diversity-generating benefits of a prokaryotic adaptive immune system. Nature *532*, 385-388.

Vercoe, R.B., Chang, J.T., Dy, R.L., Taylor, C., Gristwood, T., Clulow, J.S., Richter, C., Przybilski, R., Pitman, A.R., and Fineran, P.C. (2013). Cytotoxic Chromosomal Targeting by CRISPR/Cas Systems Can Reshape Bacterial Genomes and Expel or Remodel Pathogenicity Islands. Plos Genetics *9*.

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*, 3738-3750.

Vogel, J., and Luisi, B.F. (2011). Hfq and its constellation of RNA. Nature Reviews Microbiology *9*, 578-589.

Vorontsova, D., Datsenko, K.A., Medvedeva, S., Bondy-Denomy, J., Savitskaya, E.E., Pougach, K., Logacheva, M., Wiedenheft, B., Davidson, A.R., Severinov, K.*, et al.* (2015). Foreign DNA acquisition by the I-F CRISPR-Cas system requires all components of the interference machinery. Nucleic Acids Res *43*, 10848-10860.

Vourekas, A., Zheng, Q., Alexiou, P., Maragkakis, M., Kirino, Y., Gregory, B.D., and Mourelatos, Z. (2012). Mili and Miwi target RNA repertoire reveals piRNA biogenesis and function of Miwi in spermiogenesis. Nat Struct Mol Biol *19*, 773-781.

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

Wang, X., Kim, Y., Ma, Q., Hong, S.H., Pokusaeva, K., Sturino, J.M., and Wood, T.K. (2010). Cryptic prophages help bacteria cope with adverse environments. *1*, 147.

Waters, L.S., and Storz, G. (2009). Regulatory RNAs in bacteria. Cell *136*, 615.

Watkins, J.N.E., Kennelly, W.J., Tsay, M.J., Tuin, A., Swenson, L., Lee, H.-R., Morosyuk, S., Hicks, D.A., and SantaLucia, J.J. (2011). Thermodynamic contributions of single internal rA·dA, rC·dC, rG·dG and rU·dT mismatches in RNA/DNA duplexes. Nucleic Acids Research *39*, 1894-1902.

Wee, L.M., Flores-Jasso, C.F., Salomon, W.E., and Zamore, P.D. (2012). Argonaute divides its RNA guide into domains with distinct functions and RNA-binding properties. Cell *151*, 1055-1067.

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

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

Weinberger, A.D., Wolf, Y.I., Lobkovsky, A.E., Gilmore, M.S., and Koonin, E.V. (2012). Viral diversity threshold for adaptive immunity in prokaryotes. mBio *3*, e00456-00412.

Westra, E.R., Buckling, A., and Fineran, P.C. (2014). CRISPR-Cas systems: beyond adaptive immunity. Nature reviews. Microbiology *12*, 317-326.

Westra, E.R., Nilges, B., van Erp, P.B., van der Oost, J., Dame, R.T., and Brouns, S.J. (2012a). Cascade-mediated binding and bending of negatively supercoiled DNA. RNA Biol *9*, 1134-1138.

Westra, E.R., Pul, U., Heidrich, N., Jore, M.M., Lundgren, M., Stratmann, T., Wurm, R., Raine, A., Mescher, M., Van Heereveld, L.*, et al.* (2010). H-NS-mediated repression of CRISPR-based immunity in Escherichia coli K12 can be relieved by the transcription activator LeuO. Mol Microbiol *77*, 1380-1393.

Westra, E.R., Semenova, E., Datsenko, K.A., Jackson, R.N., Wiedenheft, B., Severinov, K., and Brouns, S.J. (2013). Type I-E CRISPR-Cas Systems Discriminate Target from Non-Target DNA through Base Pairing-Independent PAM Recognition. PLoS Genet *9*, e1003742.

Westra, E.R., Swarts, D.C., Staals, R.H., Jore, M.M., Brouns, S.J., and van der Oost, J. (2012b). The CRISPRs, they are a-changin': how prokaryotes generate adaptive immunity. Annu Rev Genet *46*, 311-

339.

Westra, E.R., van Erp, P.B., Künne, T., Wong, S.P., Staals, R.H., Seegers, C.L., Bollen, S., Jore, M.M., Semenova, E., Severinov, K., *et al.* (2012c). CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. Mol Cell *46*, 595-605.

Westra, E.R., van Houte, S., Oyesiku-Blakemore, S., Makin, B., Broniewski, J.M., Best, A., Bondy-Denomy, J., Davidson, A., Boots, M., and Buckling, A. (2015). Parasite Exposure Drives Selective Evolution of Constitutive versus Inducible Defense. Current biology : CB *25*, 1043-1049.

Weterings, E., and Chen, D.J. (2008). The endless tale of non-homologous end-joining. *18*, 114-124.

Wickham, H. (2009). Ggplot2 : elegant graphics for data analysis (New York: Springer).

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

Wiedenheft, B., Sternberg, S.H., and Doudna, J.A. (2012). RNA-guided genetic silencing systems in bacteria and archaea. Nature *482*, 331-338.

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

Wigley, D.B. (2013). Bacterial DNA repair: recent insights into the mechanism of RecBCD, AddAB and AdnAB. Nature reviews. Microbiology *11*, 9-13.

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

Wright, A.V., Nunez, J.K., and Doudna, J.A. (2016). Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. Cell *164*, 29-44.

Wu, P., Nakano, S., and Sugimoto, N. (2002). Temperature dependence of thermodynamic properties for DNA/DNA and RNA/DNA duplex formation. European journal of biochemistry *269*, 2821-2830.

Wu, X., Kriz, A.J., and Sharp, P.A. (2014a). Target specificity of the CRISPR-Cas9 system. Quantitative biology *2*, 59-70.

Wu, X., Scott, D.A., Kriz, A.J., Chiu, A.C., Hsu, P.D., Dadon, D.B., Cheng, A.W., Trevino, A.E., Konermann, S., Chen, S., *et al.* (2014b). Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. Nat Biotechnol *32*, 670-676.

Xue, C., Seetharam, A.S., Musharova, O., Severinov, K., SJ, J.B., Severin, A.J., and Sashital, D.G. (2015). CRISPR interference and priming varies with individual spacer sequences. Nucleic Acids Res *43*, 10831-10847.

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

Yamada, M., Watanabe, Y., Gootenberg, J.S., Hirano, H., Ran, F.A., Nakane, T., Ishitani, R., Zhang, F., Nishimasu, H., and Nureki, O. (2017). Crystal Structure of the Minimal Cas9 from Campylobacter jejuni Reveals the Molecular Diversity in the CRISPR-Cas9 Systems. Mol Cell *65*, 1109-1121 e1103.

Yamano, T., Nishimasu, H., Zetsche, B., Hirano, H., Slaymaker, I.M., Li, Y., Fedorova, I., Nakane, T., Makarova, K.S., Koonin, E.V., *et al.* (2016). Crystal Structure of Cpf1 in Complex with Guide RNA and Target DNA. Cell *165*, 949-962.

Yang, H., Gao, P., Rajashankar, K.R., and Patel, D.J. (2016). PAM-Dependent Target DNA Recognition and Cleavage by C2c1 CRISPR-Cas Endonuclease. Cell *167*, 1814-1828 e1812.

Yekta, S., Shih, I.-h., and Bartel, D.P. (2004). MicroRNA-directed cleavage of HOXB8 mRNA. Science Signaling *304*, 594.

Yoda, M., Kawamata, T., Paroo, Z., Ye, X., Iwasaki, S., Liu, Q., and Tomari, Y. (2009). ATP-dependent human RISC assembly pathways. Nature structural & molecular biology *17*, 17-23.

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

Yosef, I., Goren, M.G., Kiro, R., Edgar, R., and Qimron, U. (2011). High-temperature protein G is essential for activity of the Escherichia coli clustered regularly interspaced short palindromic repeats (CRISPR)/Cas system. Proceedings of the National Academy of Sciences *108*, 20136-20141.

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

Yosef, I., Shitrit, D., Goren, M.G., Burstein, D., Pupko, T., and Qimron, U. (2013). DNA motifs determining the efficiency of adaptation into the Escherichia coli CRISPR array. Proceedings of the National Academy of Sciences of the United States of America *110*, 14396-14401.

Young, J.C., Dill, B.D., Pan, C., Hettich, R.L., Banfield, J.F., Shah, M., Fremaux, C., Horvath, P., Barrangou, R., and Verberkmoes, N.C. (2012). Phage-induced expression of CRISPR-associated proteins is revealed by shotgun proteomics in Streptococcus thermophilus. PLoS One *7*, e38077.

Zebec, Z., Manica, A., Zhang, J., White, M.F., and Schleper, C. (2014). CRISPR-mediated targeted mRNA degradation in the archaeon Sulfolobus solfataricus. Nucleic Acids Res *42*, 5280-5288.

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*, 210-219.

Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., Essletzbichler, P., Volz, S.E., Joung, J., van der Oost, J., Regev, A*., et al.* (2015). Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell *163*, 759-771.

Zetsche, B., Heidenreich, M., Mohanraju, P., Fedorova, I., Kneppers, J., DeGennaro, E.M., Winblad, N., Choudhury, S.R., Abudayyeh, O.O., Gootenberg, J.S*., et al.* (2017). Multiplex gene editing by CRISPR-Cpf1 using a single crRNA array. Nat Biotechnol *35*, 31-34.

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

Zhao, H., Sheng, G., Wang, J., Wang, M., Bunkoczi, G., Gong, W., Wei, Z., and Wang, Y. (2014). Crystal structure of the RNA-guided immune surveillance Cascade complex in Escherichia coli. Nature *515*, 147-150.

Zhu, J., and Wartell, R.M. (1999). The effect of base sequence on the stability of RNA and DNA single base bulges. Biochemistry *38*, 15986-15993.

Zuris, J.A., Thompson, D.B., Shu, Y., Guilinger, J.P., Bessen, J.L., Hu, J.H., Maeder, M.L., Joung, J.K., Chen, Z.Y., and Liu, D.R. (2015). Cationic lipid-mediated delivery of proteins enables efficient protein-based genome editing in vitro and in vivo. Nat Biotechnol *33*, 73-80.

About the author

Tim Andreas Künne was born on February 15th 1988 in Düsseldorf, Germany. In 2007 he started to study Biotechnology at Wageningen University, the Netherlands. He obtained his BSc in 2010 and his MSc (*cum laude*) in 2012. In 2010 he performed research for his BSc thesis at the department of Molecular Biology at Wageningen University on cell wall remodelling and signalling during root nodule formation in *Medicago truncatula*. In 2011 he performed research for his MSc thesis at the department of Microbiology at Wageningen University on the CRISPR-Cas immune system in *Escherichia coli*. Later in 2012 he participated in the Erasmus exchange program for 6 months, to perform research for his MSc internship in the Roslin Institute, University of Edinburgh, Scotland. Here, he characterized a knock-out cell line of mouse embryonic stem cells.

In September 2012, after returning from Scotland, Tim started his PhD research project at Wageningen University in the laboratory of Microbiology. He continued to study the type I-E CRISPR-Cas system in *E. coli* under the supervision of Dr. Stan Brouns and Prof. John van der Oost. The work done in the following 4 -5 years focussed on the mechanisms of this fascinating immune system and is described in this thesis.

List of publications

Künne, T., Biewenga, L., Kieper, S.N., van Esveld, S., ten Buren, E.B.J., van der Oost, J., and Brouns, S.J. Development of Cascade into a programmable RNA-guided nuclease. *Manuscript in preparation*.

Künne, T., Zhu, Y., da Silva, F., Konstantinides, N., van der Oost, J., and Brouns, S.J. Opposite effects of guanine and cytosine protospacer mismatches in direct CRISPR interference and priming. *Manuscript in preparation*.

Vlot, M., Houkes, J., Lochs, S., Swarts, D.C., Zheng, P., Künne, T., Mohanraju, P., Anders, C., Jinek, M., van der Oost, J., Dickman, M.J., Brouns, S.J. Target DNA binding by Type I and II but not Type V crRNA-effector complexes is impaired by bacteriophage DNA glucosylation. *Submitted*

Künne, T., Kieper, S.N., Bannenberg, J.W., Vogel, A.I., Miellet, W.R., Klein, M., Depken, M., Suarez-Diez, M., and Brouns, S.J. (2016). Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation. Mol Cell 63, 852-864.

Künne, T., Westra, E.R., and Brouns, S.J. (2015). Electrophoretic Mobility Shift Assay of DNA and CRISPR-Cas Ribonucleoprotein Complexes. Methods in molecular biology 1311, 171-184.

Blosser, T.R., Loeff, L., Westra, E.R., Vlot, M., Künne, T., Sobota, M., Dekker, C., Brouns, S.J., and Joo, C. (2015). Two distinct DNA binding modes guide dual roles of a CRISPR-Cas protein complex. Mol Cell 58, 60-70.

Fineran, P.C., Gerritzen, M.J., Suarez-Diez, M., Künne, T., Boekhorst, J., van Hijum, S.A., Staals, R.H., and Brouns, S.J. (2014). Degenerate target sites mediate rapid primed CRISPR adaptation. Proceedings of the National Academy of Sciences of the United States of America 111, E1629-1638.

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

Westra, E.R., van Erp, P.B., Künne, T., Wong, S.P., Staals, R.H., Seegers, C.L., Bollen, S., Jore, M.M., Semenova, E., Severinov, K., et al. (2012). CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. Mol Cell 46, 595-605.

Overview of completed training activities

Discipline-specific activities

Meetings & conferences

- CRISPR meeting 2015, The Adaptive Prokaryotic Immune System CRISPR-Cas. New York (USA), 2015*
- NWO-ALW molecular genetics meeting. Lunteren (NL), 2015**
- Host microbe interactomics symposium. Wageningen (NL), 2014*
- Annual Meeting NWO study group Protein Research, Nucleic Acids and Lipids & Membranes. Veldhoven (NL), 2014*
- NWO-ALW molecular genetics meeting. Lunteren (NL), 2014*
- CRISPR meeting 2014. Berlin (DE), 2014*
- CRISPR meeting 2013. St. Andrews (UK), 2013

*poster presentation, **oral presentation

Courses

- VLAG Metabolic engineering. Wageningen (NL), 2015
- VLAG Applied biocatalysis. Wageningen (NL), 2014
- Cat-AgroFood PacBio seminar. Wageningen (NL), 2014
- VLAG Homology modelling seminar. Nijmegen (NL), 2013

General courses

- InDesign essentials. Wageningen (NL), 2016
- Essential skills in data managment. Utrecht (NL), 2016
- PhD workshop carousel. Wageningen (NL), 2015
- Scientific writing. Wageningen (NL), 2015
- Advanced scientific artwork. Wageningen (NL), 2014
- Reviewing a scientific paper. Wageningen (NL), 2014
- AFSG giving and receiving feedback. Wageningen (NL), 2013
- VLAG PhD week. Baarlo (NL), 2013

Optionals

- PhD trip MIB-SSB. California (USA), 2015
- Organising committee, Antiviral defence mechanisms symposium. Wageningen (NL), 2014
- Bacterial Genetics group meetings
- PhD meetings Laboratory of Microbiology
- Microbiology seminars

Acknowledgements

After almost 5 years of challenges, success and failure, it is time to put an end to this chapter of my life. But ends don't have to be goodbyes. It is very important to thank all the people that have supported me, contributed to my work, or made their impression in other ways. But it is also important to not take yourself too seriously, so here it goes.

First off, I want to thank **Stan**. You put your trust in me to be your first PhD student and gave me this amazing opportunity. You have been the perfect supervisor, giving me all the freedom or all the guidance when needed. You always showed personal interest and support, never making me feel restricted in my possibilities. Very importantly, you always managed to see things positive and motivate me, which can't be easy. If anything, you are too nice. I can say that now that I am finished. In that spirit, keep calm and CRISPR on!

John, you may have been the man in the background for me, but you are the coolest and most approachable professor I have met. Thank you for being my promotor. I really appreciate the family that you have created in BacGen, which makes for an amazing working environment.

Edze, you are the one to blame for making me do a PhD on CRISPR. You first inspired me as my supervisor during my MSc thesis. Later, I enjoyed being your colleague for a little while. Your only mistake? Setting the bar too high for those who come after you ;)

Spending around 20 hours a week in a small room with the same people, leaves quite the impression. I had the pleasure to serve my time in two different cells, I mean offices, with two different groups of people. I truly enjoyed the time in the Dreijen office thanks to the following people. **Marnix**, you share the same fate as me and you always helped me with my woodworking problems. **Daan**, you were always there to look down on us and share your knowledge about cleavage. **Jorrit**, or Daan2, you are equally good at looking down on people and always advised us on the newest trends in online gambling. **Becca**, although only for a short time, you brought female charm to our office and we had a lot of fun with you. **Edze**, you created an atmosphere of excitement in the office by building the world's tallest office peanut butter jar tower, which threatened to collapse at any time. **Stan**, who would imagine it could be so much fun to sit in an office with your boss? **Peter**, you confirmed every stereotype we had about New Zeeland and I want to thank you very much for Chapter 4.

After moving to Helix, I had a great new office team that supported me through the writing of my thesis. **Teunke**, you were the big sister of our office, always making sure everyone had a nice weekend, getting personally involved and providing moral support. **Nico**, the man who knows everything and can't stop talking, awesome combination. You were the most entertaining technical support I could have wished for. **Alex**, the only one who had the means, a hammer and a strike-drone, to make Nico stop talking. I enjoyed listening to your stories and fun facts about Nico. **Tijs**, as a new office member you brought balance to the force by joining the more quiet side of the room. **Hanne**, especially in combination with Nico, you strongly contributed to a lively and entertaining atmosphere in the office. **Lucia**, it was great to hear about the business side of science for a change.

A great thank you also goes to my students, who have greatly contributed to the work in this book. **Sebastian**, **Anne**, **Lieuwe**, **Willem**, **Gerdina**, **Nico**, **Jasper**, **Emiel** and **Fausia**. You have taught me an important leadership skill. Knowing that there is no greater feeling than relaxing in your chair while others do your work for you. May you enjoy the same benefits during your PhDs or other careers.

Thanks also to all the colleagues in the 'bacterial defence team' that I had the pleasure to work with. **Raymond**, the man of international mystery. You were endlessly helpful in the lab, especially during my MSc thesis. You always had a good story at hand, benefits of a sketchy social environment. I both enjoyed and disliked that you moved to the other side of the globe for a while. **Yifan**, it was a great pleasure to have you join team CRISPR, even though on the wrong system. I enjoyed our talks about China, Chinese food and whiskey. **Ioannis**, it was great fun to supervise the practical with you and learn a bit about Cas9. **Franklin**, it was great to have you and your endless surge of ideas here, even though you were sometimes hard to follow. **Wen**, your endless energy is highly contagious, great! Pew pew. **Prathna**, thanks a lot for helping me with my new project. **Sebastian**, I am very happy that you turned from my student into my colleague. **Jochem**, I enjoyed exchanging travel experiences, of which you have a lot. **Patrick**, I enjoyed your fun first slide during presentations.

To all other (former) colleagues at BacGen. **Tom**, **Elleke**, **Sjoerd**, **Tijn**, **Melvin**, **Bas**, **Tessa**, **Mark**, **Servé**, **Joyshree**, **Jeroen**, **Lione**. Thank you all for the great times and environment! Thanks to **Wim**, **Sjon**, **Phillipe**, **Merlijn**, **Steven**, **Anja**, **Carolien, Mirella** for the fun technical and administrative support. **Gosse** and **Caroline**, thank you for your great and inspiring teaching, which got me excited about microbiology in the first place. Thank you to all the other unnamed members of **MIB** and **SSB** for the good times at the Dreijen and in Helix.

Also many thanks to all the collaborators that have been involved in this work. **Misha** and **Martin** from Delft for their help with chapter 5. **Luuk** for great mutual exchanges about Cas3. A huge thank you goes to **Maria**, for handling all the bio-informatics in this thesis. You were a great help and you are amazing at what you do.

Willem, thank you for shaping such a great department and always being open to suggestions and excuses to have some beers with our colleagues.

A special thanks also to my paranymphs, **Benoit** and **Marnix**, for all the free labour you have done/are hopefully going to do for me.

A

My dear family, **Mama**, **Papa**, **Niklas**, **Johannes**, thank you so much for everything you did that has brought me to this point in my life. Be it your love and support, a good spanking, escalating sibling rivalry or any other traumatic experiences that I might have suppressed. Your influences have shaped me into the person that I am today and for that I am forever grateful.

Thank you also to **Tobi**, for becoming the first Dr. in my family. You inspired me to pursue an academic career.

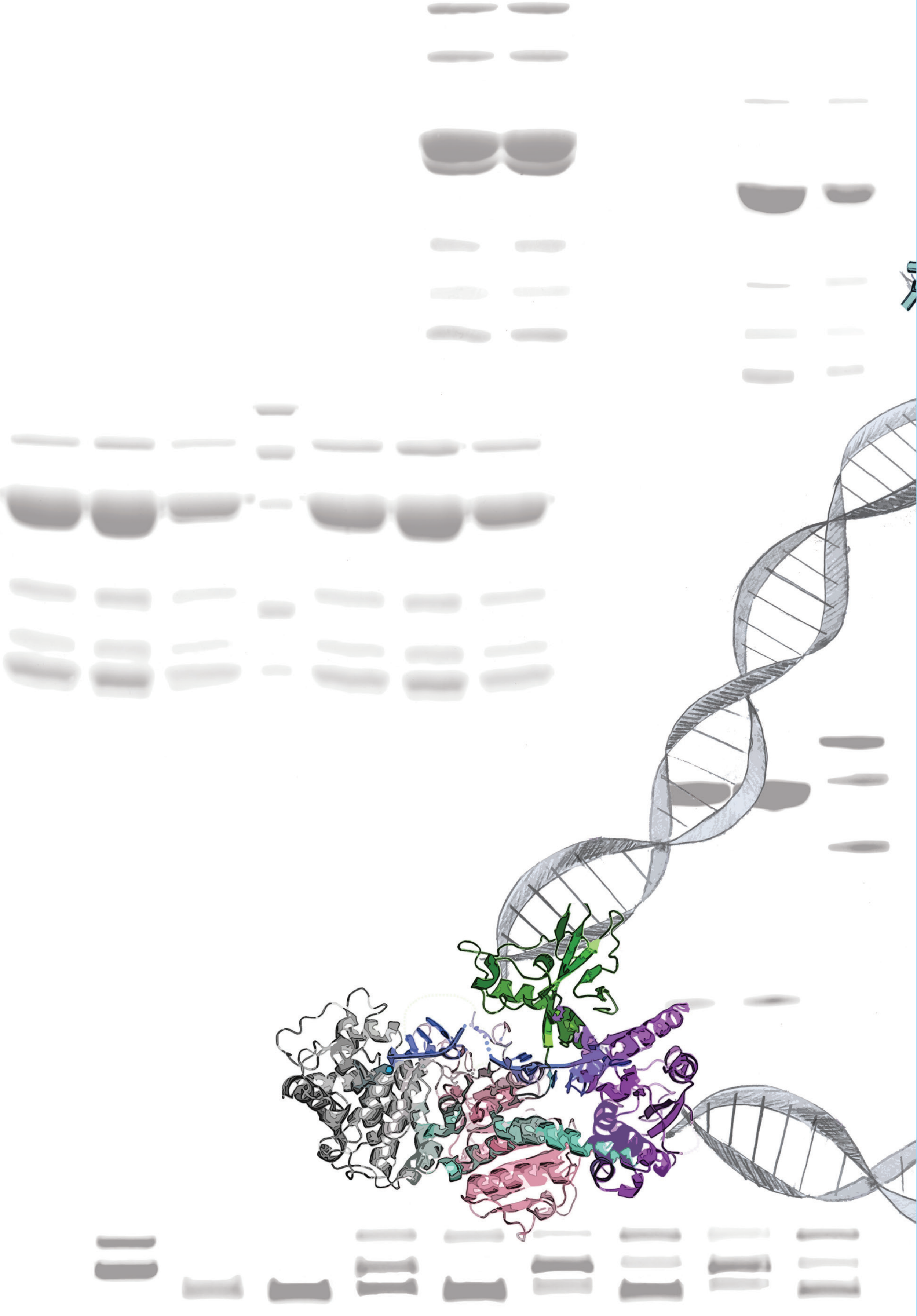
Last and most importantly, **Maomao**. At this point I am not sure if I should say, without you this book would have never been finished, or, without you this book would have been finished a year earlier. Either way, you are the best thing that ever happened to me. You showed me how to live and enjoy my life in the now, instead of waiting for it to start one day in the future. You have opened my eyes to the world beyond what I knew and thereby made me a better person. You have also turned me into a total 吃货, which you will regret once my metabolism slows down. I love you.

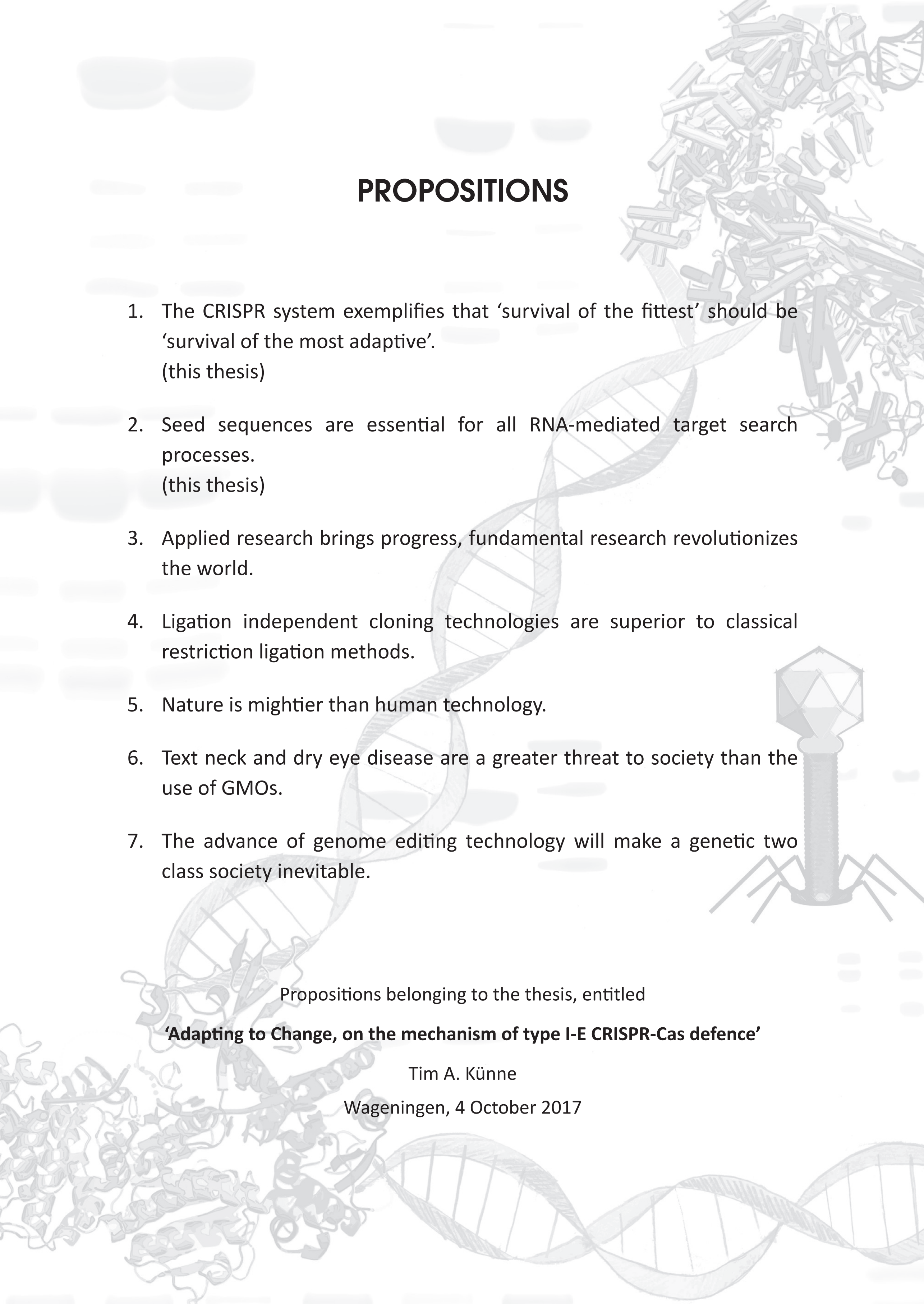
About the cover

The background of the cover is made up of photos of DNA and protein gels that have been at the centre of this thesis. The protein gels show bands representing the three protein-complexes that are studied and applied in this work, Cascade, Cas1-2 and Cas3. The DNA gels show the results of typical *in vitro* assays performed with these proteins. The large triangular illustration represents the type I-E CRISPR-Cas immune system of *E. coli*. At the edges, crystal structures of the Cas proteins are shown, Cascade (top), Cas1-2 (right) and Cas3 (left). The proteins are connected by nucleic acids. This illustrates the interactions of the Cas proteins and the DNA/RNA of the CRISPR array to form the complex CRISPR-Cas immune system. The virus in the middle is trapped by the synergistic action of the Cas proteins.

The research described in this thesis was financially supported by a European Research Council starting grant (639707) and a Netherlands Organisation for Scientific Research (NWO) Vidi grant (864.11.005) to Dr S.J.J. Brouns.

Cover design: Illiana Boshoven-Gkini | www.AgileColor.com
Lay-out: Tim Künne
Printed: GVO drukkers & vormgevers B.V. Ede, on recycled, FSC-certified paper





PROPOSITIONS

1. The CRISPR system exemplifies that ‘survival of the fittest’ should be ‘survival of the most adaptive’.
   (this thesis)

2. Seed sequences are essential for all RNA-mediated target search processes.
   (this thesis)

3. Applied research brings progress, fundamental research revolutionizes the world.

4. Ligation independent cloning technologies are superior to classical restriction ligation methods.

5. Nature is mightier than human technology.

6. Text neck and dry eye disease are a greater threat to society than the use of GMOs.

7. The advance of genome editing technology will make a genetic two class society inevitable.

Propositions belonging to the thesis, entitled

‘Adapting to Change, on the mechanism of type I-E CRISPR-Cas defence’

Tim A. Künne

Wageningen, 4 October 2017