Skolkovo Institute of Science and Technology

COMPARATIVE ANALYSIS OF THE ACTION OF EUBACTERIAL CLASS 1 CRISPR-CAS SYSTEMS

Doctoral Thesis

by

DARIA N. ARTAMONOVA

DOCTORAL PROGRAM IN LIFE SCIENCES

Supervisors:
Professor Konstantin Severinov
Dr. Ekaterina Semenova

Moscow – 2017

Abstract

CRISPR-Cas systems are systems of adaptive immunity in prokaryotes. They protect bacteria and archaea from the invasion of viruses and mobile genetic elements (MGEs). The systems are divided into Classes 1 and 2, depending on whether several or only one protein enters the effector CRISPR ribonucleoprotein complex (crRNP complex) (Makarova *et al*., 2015). Many practical applications of Class 2 CRISPR-Cas systems have been developed. They are applied as programmable nucleases for genome editing (Jinek *et al*., 2012; Kim *et al*., 2017; Song *et al*., 2017). Variants with an inactive nuclease site are utilized as regulators of gene expression (Kearns *et al*., 2014). There are also applications in fluorescent labeling (Deng *et al.*, 2015) and in the technique of modified chromatin immunoprecipitation (Fujita *et al*, 2013). Class 1 CRISPR-Cas systems containing multi-subunit crRNP complexes can also be used for practical purposes, notably for genome editing in prokaryotic organisms, but only if an active CRISPR-Cas system is initially present in the cell genome (Li *et al.*, 2016) so that a step of delivery of complex CRISPR machinery to the cell can be bypassed (Kim *et al.*, 2017; Song *et al.*, 2017). Thus, industrial producer strains and new laboratory strains can be obtained. Both classes of CRISPR-Cas systems are related and share some common components (Shmakov *et al.*, 2015). So fundamental discoveries related to Class 1 systems can facilitate the detection of similar mechanisms in Class 2 systems. For example, anti-CRISPR proteins encoded in the genomes of some phages and capable of hampering the action against a particular type of CRISPR-Cas systems were first discovered for Class 1 systems (Bondy-Denomy *et al.*, 2013) and later the analogists were found for Class 2 systems (Pawluk *et al.*, 2016). This discovery can be applied to regulate CRISPR-Cas during genome editing applications (Rauch *et al.*, 2017). In addition, studies of the mechanisms of action of various CRISPR-Cas systems, including Class 1 systems, is of interest for the understanding of evolutionary relations and the "arms race" between prokaryotic organisms and their viruses.

The first part of the present work is devoted to the study of primed and naïve adaptation by the I-F subtype CRISPR-Cas system of *Escherichia coli* and the *Pseudomonas aeruginosa* UCBPP-PA14 I-F subtype CRISPR-Cas system overexpressed in the heterologous *E. coli* host. It is shown that in the *P. aeruginosa* Type I-F system, in contrast with what occurs in subtype I-E, all components of CRISPR-Cas machinery are needed not only for primed, but also for naïve adaptation. The second part of the work describes the study of III-A and III-B subtypes CRISPR-Cas systems of the bacterium *Thermus thermophilus* Hb27. Naïve adaptation was not detected in this strain. The regions of the protospacer necessary for its recognition by CRISPR-Cas machinery (the seed sequence) were investigated in an attempt to identify conditions for priming. It was found that both III-A and III-B systems are resistant to a surprisingly large number of spacer-protospacer mismatches at both the 3'- and 5'-end of the protospacer. The III-B subtype system is more sensitive to mismatches than the III-A subtype. The dependence of interference on mutations in the middle of the protospacer does not have an obvious logic, so if a seed region does exist in these systems it has a complicated geometry and may be distributive. Protospacer mutations that weakened or abolished interference by type III systems were also tested for primed adaptation, but the acquisition of new spacers was not observed.

Publications

ARTICLES

1. **Vorontsova* D**, Datsenko KA, Medvedeva S, Bondy-Denomy J, Savitskaya EE, Pougach K, Logacheva M, Wiedenheft B, Davidson AR, Severinov K, Semenova E. Foreign DNA acquisition by the I-F CRISPR-Cas system requires all components of the interference machinery. *Nucleic Acids Res.* 2015 Dec 15;43(22):10848-60. doi: 10.1093/nar/gkv1261.

2. Semenova E, Savitskaya E, Musharova O, Strotskaya A, **Vorontsova D**, Datsenko KA, Logacheva MD, Severinov K. Highly efficient primed spacer acquisition from targets destroyed by the *Escherichia coli* type I-E CRISPR-Cas interfering complex. *Proc Natl Acad Sci U S A*. 2016 Jul 5;113(27):7626-31. doi: 10.1073/pnas.1602639113

CONFERENCES

1. Savitskaya E, Semenova E, **Vorontsova D**, Medvedeva S, Logacheva M, Severinov K, Autoimmune primed CRISPR adaptation in I-E and I-F systems: comparative analysis of new spacer selection mechanisms. MCCMB-2015, July, 16-19, Moscow, Russia http://mccmb.genebee.msu.ru/2015/proceedings/abstracts/84.pdf

2. **Vorontsova D**, Medvedeva S, Savitskaya E, Logacheva M, Semenova E, Severinov K. Foreign DNA acquisition by the I-F CRISPR-Cas system. ITAS-2016, September, 25-30, Repino, St. Petersburg, Russia http://itas2016.iitp.ru/en/search.html#V

3. **Vorontsova D**, Lopatina A, Minakhin L, Semenova E, Severinov K. DNA targeting during interference of III type CRISPR-Cas systems in *T. thermophiles*. CRISPR-2016, May 23-26 2016, Rehovot, Israel, p. 46

*** - former surname**

Acknowledgments

First of all, I would like to thank my supervisor Professor Konstantin Severinov for giving the opportunity to work on extremely topical issues, for opportunity for growth, for trusting me to solve complicated problems by myself, for patience and for letting me make mistakes sometimes and learn from them, for discussions and his wise advice. No less I would like to thank my co-supervisor Dr. Ekaterina Semenova. Thanks her for the kindness, willingness to help on any issue, for giving the experience and knowledge, for setting an example of an outstanding experimenter. I would like to commend Dr. Leonid Minakhin and Dr. Anna Lopatina for help in work with *T.thermophilus*. I wish to thank Dr. Ekaterina Savitskaya and Sofia Medvedeva for teaching me to analyze High-throughput sequencing data and further consultations. As well I would like to thank Dr. Mikhail Khodorkovskiy for support and giving me life advice. I would like to thank Skoltech for amazing opportunities of its PhD program, for its courses and professors. Also, I would like to express gratitude to all my colleagues at laboratories of professor Severinov, especially Matvey Kolesnik, Varvara Verkhova and Aleksandra Strotskaya. Finally, I would like to thank my parents and my husband for endless patience, understanding and help, and my friends for moral support and faith in me.

Table of Contents

Acronyms

A - adenine

bp - base pair

C - cytosine

CRISPR - Clustered Regularly Interspaced Short Palindromic Repeats

crRNA – CRISPR RNA

crRNP complex - CRISPR ribonucleoprotein complex

cryo-EM - cryo-electron microscopy

DNA - deoxyribonucleic acid

dNTPs - deoxynucleotides

DSB - double-strand break

dsDNA – double-stranded DNA

dsRNA – double-stranded RNA

EOT - efficiency of transformation

E. coli – bacteria *Escherichia coli*

FRET – fluorescence/Förster resonance energy transfer

G – guanine

HTS - High-throughput sequencing

H. hispanica - archaea *Haloarcula hispanica*

IHF – integration host factor

IPTG - isopropyl β-D-1-thiogalactopyranoside

kb – kilobase – 1000 bases of DNA or RNA

kbp – kilobase pairs – 1000 base pairs of DNA

kDa - kiloDalton

MGE - mobile genetic elements

M. mediterranea - bacteria *Marinomonas mediterranea*

PAM – protospacer adjacent motif

PCR - polymerase chain reaction

pre-crRNA - precursor crRNA

P. atrosepticum - bacteria *Pectobacterium atrosepticum*

P. aeruginosa – bacteria *Pseudomonas aeruginosa*

P. furiosus - archaea *Pyrococcus furiosus*

RNA - ribonucleic acid

RNase III - ribonuclease III

rpm – revolutions per minute

RRM - RNA recognition motif

RT - reverse transcriptase

ssDNA – single-stranded DNA

ssRNA - single-stranded RNA

S. islandicus - archaea *Sulfolobus islandicus*

S. solfataricus - archaea *Sulfolobus solfataricus*

tracrRNA - trans-activating crRNA

T - thymine

T. onnurineus - archaea *Thermococcus onnurineus*

T. thermophilus – bacteria *Thermus thermophilus*

U – uracil

w/v – weight/volume

List of figures

List of tables

Chapter 1: Review of literature

1. Protective function of CRISPR-Cas systems

1.1. The discovery of CRISPR-Cas systems

CRISPR-Cas (Clustered Regularly Interspaced Short Palindromic Repeats and CRISPR associated genes) systems (Jansen *et al.*, 2002) are adaptive immune systems encoded in the genomes of prokaryotes (Mojica *et al.*, 2005; Barrangou *et al.*, 2007; Brouns *et al.*, 2008). Attention was first called to their existence in 1987 by Japanese scientists working on the *iap* gene responsible for alkaline phosphatase isozyme conversion in *E. coli*. Their work had no connection with the protective systems of bacteria, but, next to the *iap* gene, they observed an unusual set of 29-nucleotide palindromic repeats, separated by unique sequences of 32 nucleotides in length (Ishino *et al.*, 1987). The function of this striking structure was unclear. Similar structures were found in the genomes of other bacteria and also archaea (Groenen *et al.*, 1993; Mojica *et al.*, 1995; Mojica *et al.*, 2000). A-T rich leader regions, several hundred basepairs long, flanking CRISPR loci on one side, were found during the analysis of some complete genome sequences of bacteria and archaea (Mojica *et al.*, 1993; Bult *et al.*, 1996; Klenk *et al.*, 1997; Smith *et al.*, 1997). It was shown in further *in silico* studies that the orientation of the leader sequences with respect to repeats was always the same (Jansen *et al.*, 2002). CRISPR associated genes were first discovered in the work by Jansen *et al.*, in 2002, but not until 2005, almost 20 years after the detection of CRISPR loci, it was found that some of the unique sequences of CRISPR loci in bacteria correspond to regions of bacteriophage or plasmids DNA (Bolotin *et al.*, 2005; Mojica *et al.*, 2005; Pourcel *et al.*, 2005). Moreover, it was demonstrated that mobile genetic elements (MGEs) are unable to invade a cell if its CRISPR loci contain sequences corresponding to MGEs DNA fragments. It was therefore proposed that CRISPR loci play a role in providing immunity against foreign DNA (Bolotin *et al.*, 2005; Mojica *et al.*, 2005). The RNA-interference-based mechanism of CRISPR-Cas systems (analogous to eukaryotic RNA interference) and a classification of CRISPR-Cas

systems based on amino acid sequences of CRISPR-associated proteins was first proposed in Koonin's laboratory in 2006, based on the computational genomic analysis of CRISPR loci and associated protein-coding genes (Makarova *et al.*, 2006). Finally, in 2007 the adaptive immunity function of CRISPR-Cas systems was demonstrated by Barrangou *et al.* (Barrangou *et al.*, 2007). Thus, when a bacterium is attacked by an MGE, in particular, bacteriophage, a CRISPR locus could acquire a new palindromic repeat and a unique fragment corresponding to a region of the phage genome. The acquisition could provide immunity to this phage in the future. Proteins coding by CRISPR-associated genes participate in the acquisition of immunity and protective function (Barrangou *et al.*, 2007).

1.2. General mechanisms of CRISPR protective action

CRISPR-Cas systems, along with such protective mechanisms as restriction-modification systems, receptor masking, bacteriophage exclusion (BREX) and abortive infection (Tock *et al.*, 2005; Samson *et al.*, 2013; Goldfarb *et al.*, 2015) protect prokaryotic cells from MGEs, such as viruses, plasmids and transposons. According to the CRISPRdb database (http://crispr.i2bc.paris-saclay.fr/), various types and subtypes of the CRISPR-Cas system have been found in most of known archaea and about half of bacterial genomes (Grissa *et al.*, 2007). Despite the structural and functional diversity of CRISPR-Cas systems, all of them share some common features (Barrangou & Marraffini, 2014).

CRISPR-Cas systems comprise CRISPR arrays and *cas* genes associated with them (Jansen *et al.*, 2002). CRISPR arrays consist of a leader sequence, fragments with the same sequence, called "repeats" separated by fragments with unique sequences, called "spacers" (Figure 1). The leader sequence is 5'-adjacent 100-500 nucleotide regulatory AT-rich region containing a

promoter. Repeats and spacers have a length of 21-47 and 21-72 bp, respectively, and alternate with each other (Mojica *et al.*, 1993; Jansen *et al.*, 2002; Grissa *et al.*, 2007).

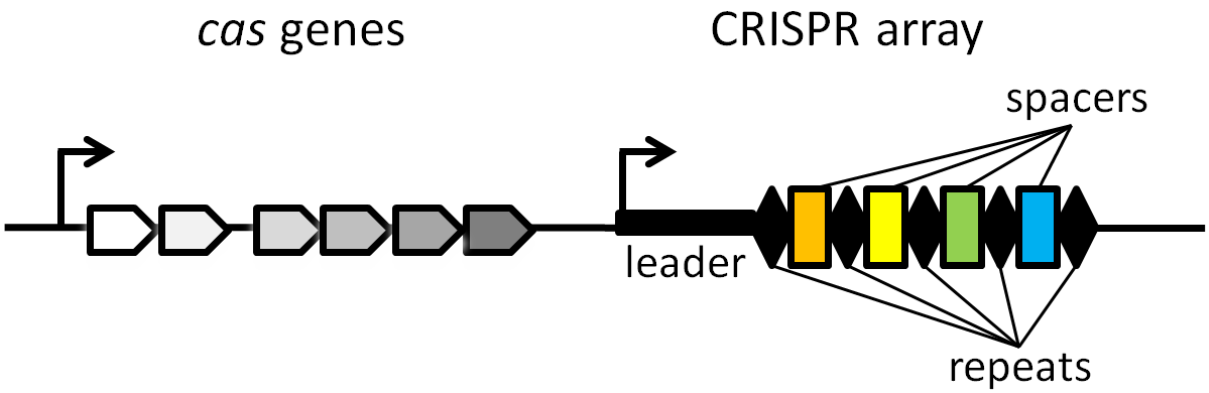


Figure 1. Organization of CRISPR-Cas systems.

CRISPR-Cas systems consist of *cas* genes and CRISPR arrays. CRISPR arrays consist of a leader sequence, unique spacer sequences (spacers) separated by repeated sequences (repeats).

CRISPR arrays can be extended as a result of the insertion of new spacers and duplication of repeats. The process of spacer acquisition is called "adaptation". Long precursor crRNA (pre-crRNA) is transcribed from the CRISPR array. The pre-crRNA is then processed by cleavage within the repeat sequences to produce mature short crRNAs, which contain a sequence of only one spacer flanked by fragments of repeats (Lillestøl *et al.*, 2006; Brouns *et al.*, 2008). In the next stage, an effector crRNP complex assembles, consisting of a single crRNA molecule and one or more Cas proteins (Brouns *et al.*, 2008). A common key feature of the protective action of CRISPR-Cas systems is the ability of the crRNP effector complex to recognize foreign DNA or RNA that is complementary (as defined by Watson–Crick pairing) to the spacer-derived region of the guide crRNAs, and cause target destruction by Cas proteins with nuclease activity (Bolotin *et al.*, 2005; Mojica *et al.*, 2005; Makarova *et al.*, 2006; Brouns *et al.*, 2008; Hale *et al.*, 2009). The process of target recognition and destruction is called

"interference" (Makarova *et al.*, 2006), and a target sequence complementary to a CRISPR spacer is called a "protospacer" (Mojica *et al.*, 2009). Thus, three common stages can be identified in the functioning of all types of CRISPR-Cas systems: 1) adaptation (the acquisition of new spacers), 2) the expression and processing of crRNA, 3) interference (recognition and destruction of the target) (Figure 2).

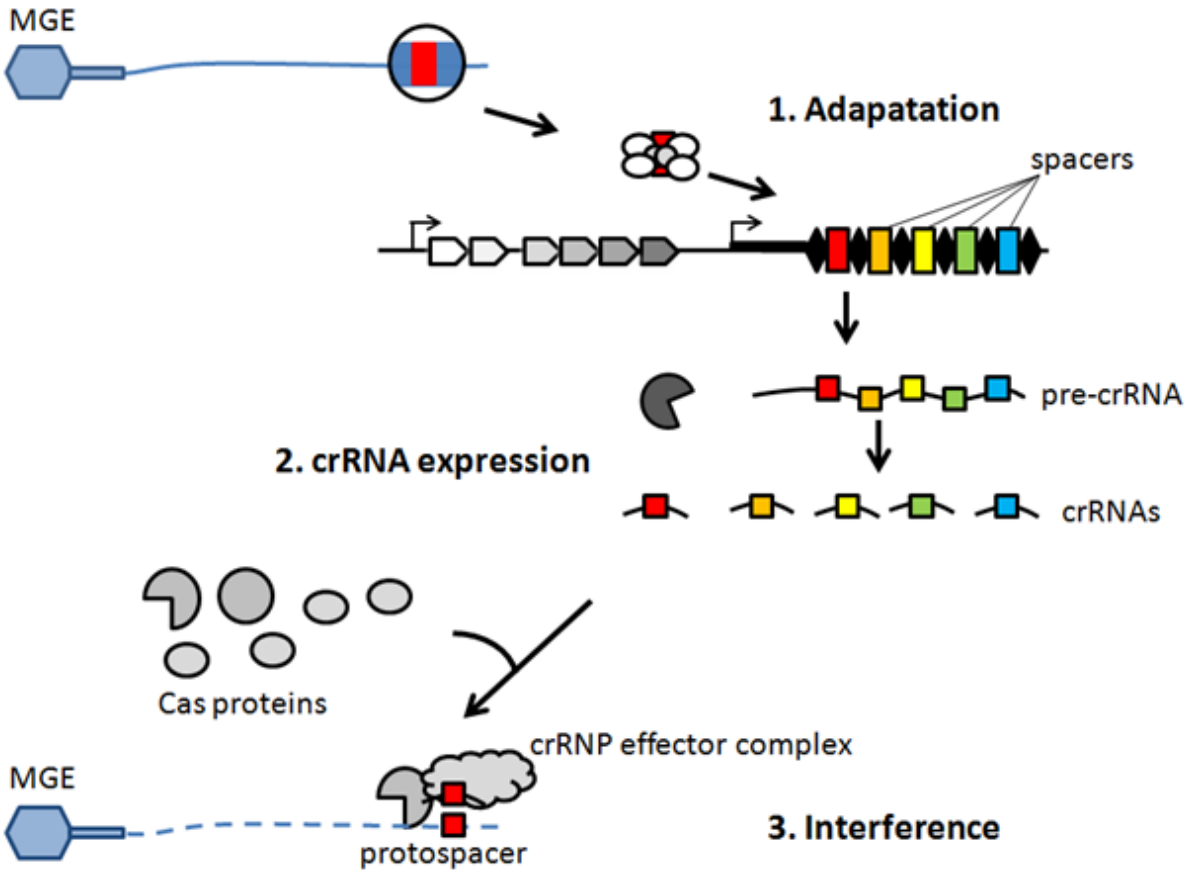


Figure 2. Mechanism of CRISPR-Cas systems action.

Three stages of immunity – adaptation, crRNA expression, and interference - are shown. At the adaptation stage Cas proteins provide acquisition of a new spacer from nucleic acid of an invader. At the next stage precursor crRNA is transcribed from the CRISPR array and processed into mature crRNAs. At the last interference stage, crRNA assembles with Cas proteins and forms a crRNP effector complex. When the MGE re-infects the cell, the crRNP complex binds with complementary to crRNA region of foreign DNA or RNA, which is called protospacer, finally it triggers the destruction of invader's nucleic acid by CRISPR-Cas machinery.

1.3. Mechanisms for avoiding autoimmune response

Obviously, each crRNA should primarily target the corresponding spacer in the CRISPR array from which it was transcribed. Therefore, the action of the CRISPR-Cas system should also be directed at its own genome. In order to avoid such autoimmune response, it is extremely important for a prokaryotic cell to distinguish a protospacer in foreign DNA from a spacer in its own genome. In most types of CRISPR-Cas systems, this problem is solved by the presence of a protospacer adjacent motif (PAM) in the foreign nucleic acid and its absence in the sequence of the CRISPR array. PAM is short motif two-eight nucleotides in length, which is located beside a protospacer. It has the individual sequence for each of subtype of CRISPR-Cas systems (Horvath *et al.*, 2008; Deveau *et al.*, 2008; Mojica *et al.*, 2009) (Figure 3).

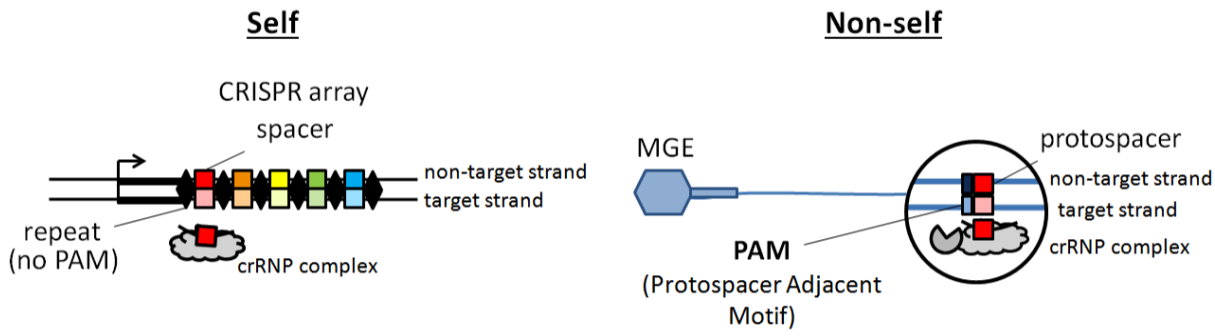


Figure 3. PAM serves to distinguish the host DNA from the foreign DNA.

PAM (protospacer adjacent motif) is a short motif beside a protospacer. It is necessary for recognition and following destruction of the target in most of CRISPR-Cas system types. Cell's genome doesn't contain PAM in its CRISPR array and it protects the cell from self-targeting.

In Type I and Type V CRISPR-Cas systems, PAM is located beside the 3' end of the protospacer's target strand, while it is located on the other side of protospacer in Type II systems (Figure 4). For Type IV systems, PAM and its location have not been determined yet. (Leenay and Beisel, 2017) (Figure 4). It is interesting that an 8-nucleotide "handle" (a residue

of a CRISPR repeat) at the 5 ' end of the crRNA but not PAM plays a major role in a process of discrimination foreign nucleic acid and the cell's own nucleic acid in Type III CRISPR-Cas systems. Type III systems are transcriptional-depended and crRNA binds with protospacer on transcript, not with DNA. Destruction of the target will occur only if there is a lack of complementarity between the specific positions of the handle and the target RNA (Marraffini and Sontheimer, 2010). However, in the recent investigation, it was suggested that uncoupling of the 5'-handle is insufficient for identification of an RNA target as the foreign one. It was also found that the RNA PAM (rPAM) is located on target RNA beside the 3' end of the protospacer (Elmore *et al.*, 2016) (Figure 4). The target for Type VI CRISPR-Cas systems is RNA and so-called the protospacer-flanking sequence (PFS) is located beside 3' end of protospacer (Abudayyeh *et al*., 2016) (Figure 4). Thus, the presence of a PAM, or non-complementarity of the repeat residue in the crRNA molecule and the corresponding sequence adjacent to target protospacer, or both, are requested for recognition and destruction of the target in addition to the complementarity between a protospacer and a crRNA.

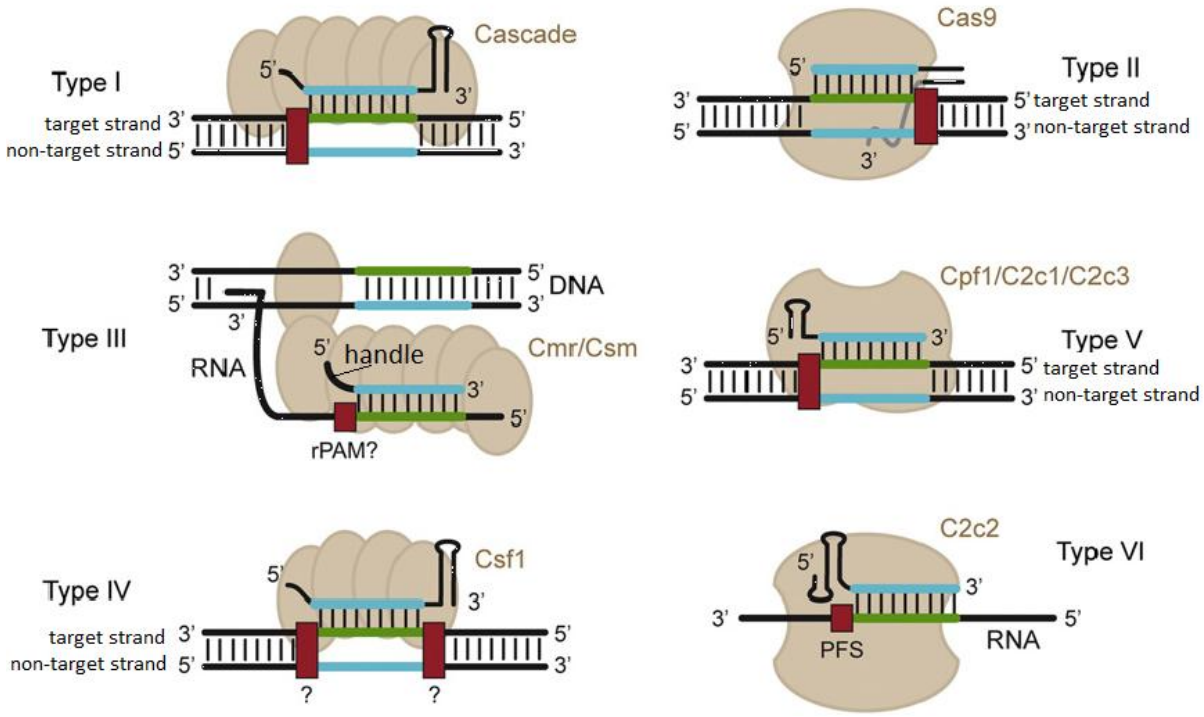


Figure 4. PAM orientation and targets for different types of CRISPR–Cas systems.

The target-bound crRNP complexes and their names for different CRISPR-Cas system types are shown. PAMs are marked in scarlet. PAM is located beside the 3'-end of the protospacer's target strand in Type I and Type V CRISPR-Cas systems, while it is located beside the other side of protospacer in Type II systems. For Type IV systems, the location of PAM has not been determined yet. A so-called protospacer-flanking sequence (PFS) is present beside the 3' of protospacer in Type VI system in which the target is RNA. In Type III transcriptional-depended systems, the "handle" (5'-end of the crRNA) serves for discrimination of invader. The lack of basepairing between it and the target RNA license the destruction, however the additional presence of RNA PAM (rPAM) was proposed in the recent study. (Adapted from Leenay and Beisel, 2017 with permission)

It remains unclear how the cell distinguishes its own DNA from foreign nucleic acid at the adaptation stage. It has been shown for several systems that new spacers are preferentially selected from actively replicating DNA (Levy *et al.*, 2015). In other cases, apparently, the cells do not distinguish between their own DNA and foreign DNA, therefore many cells die due to autoimmunity. However, in this case, acquisition of spacers from a foreign DNA can occur immediately after its penetration into the cell and the resistant cells become dominant in the population (Wei *et al.*, 2015).

2. Diversity and evolution of CRISPR-Cas systems

2.1. Classification of CRISPR-Cas systems.

According to modern classification, CRISPR-Cas systems are divided into two classes, differing in the complexity and composition of their effector crRNP complexes. Class 1 includes systems with multi-subunit crRNP complexes, and Class 2 includes systems that have only one large protein in their crRNP complex. Each of the classes currently includes three types of CRISPR-Cas systems. Each of the types is divided into subtypes, which are denoted with the letters of the Latin alphabet based on the composition and amino acid sequences of the Cas proteins entering into the crRNP complexes. Class 1, on which current dissertation is focused on, includes types I, III and functionally uncharacterized Type IV systems. Class 2 includes Type II, V and VI (Makarova *et al.*, 2015; Shmakov *et al.*, 2015). Architecture of types and subtypes is shown on Figure 5.

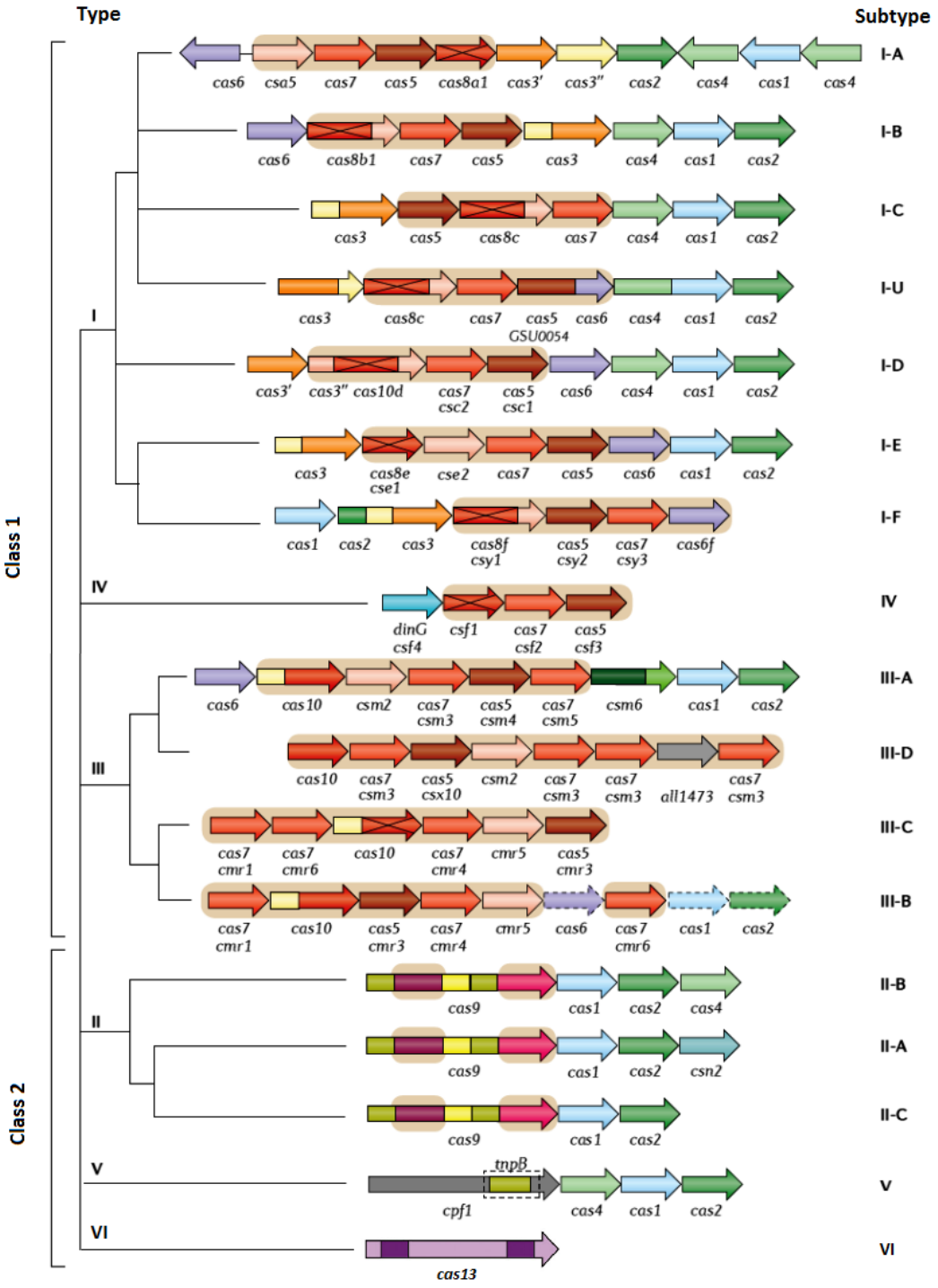


Figure 5. Classification of CRISPR-Cas systems and the architecture of different subtypes' loci.

All CRISPR-Cas systems are divided into two classes. Class 1 includes systems with multi-subunit crRNP complexes, Class 2 – with one large protein in their crRNP complexes. Each of Classes contains three types named with Roman numerals and most of types divided into subtypes named with Latin alphabet letters. Typical operon organization is shown for each CRISPR–Cas system subtype with color-coded arrows. Genes and gene regions encoding components of the crRNP complexes are highlighted with a beige background. (Adapted from Makarova *et al.*, 2015 and Koonin *et al.*, 2017 with permission)

Type I is subdivided into subtypes A through F and, in addition, subtype U which means "uncharacterized". For the archaea, it was suggested that the subtype I-G, which is very close to I-B (Vestergaard *et al.*, 2014), should be additionally identified. Type III is subdivided into subtypes A through D. Type II - into subtypes A through D. Found last types V and VI of Class 2 CRISPR-Cas systems have also been divided into several subtypes recently (for simplicity, it is not reflected on the Figure 5) (Koonin et al., 2017).

All subtypes belonging to the same Type are characterized by the presence of one signature protein. For example, for Type I this protein is Cas3, for Type III it is Cas10, and for Type II it is famous for its practical applications Cas9 (Makarova *et al.*, 2011a; Makarova *et al.*, 2011b; Makarova *et al.*, 2015).

2.2. Modular organization of CRISPR-Cas systems

Despite the diversity, there are some general principles of organization and function for all subtypes of CRISPR-Cas systems. The analysis of Cas proteins identified 4 common functional protein modules: 1) a module responsible for the expression and processing of crRNA, 2) an interference module responsible for binding crRNA and targeting invader nucleic acid, and for the destruction of the target, 3) an adaptation module responsible for the acquisition of new spacers, and 4) a regulatory non-mandatory ancillary module (Makarova *et al.*, 2015) (Figure 6).

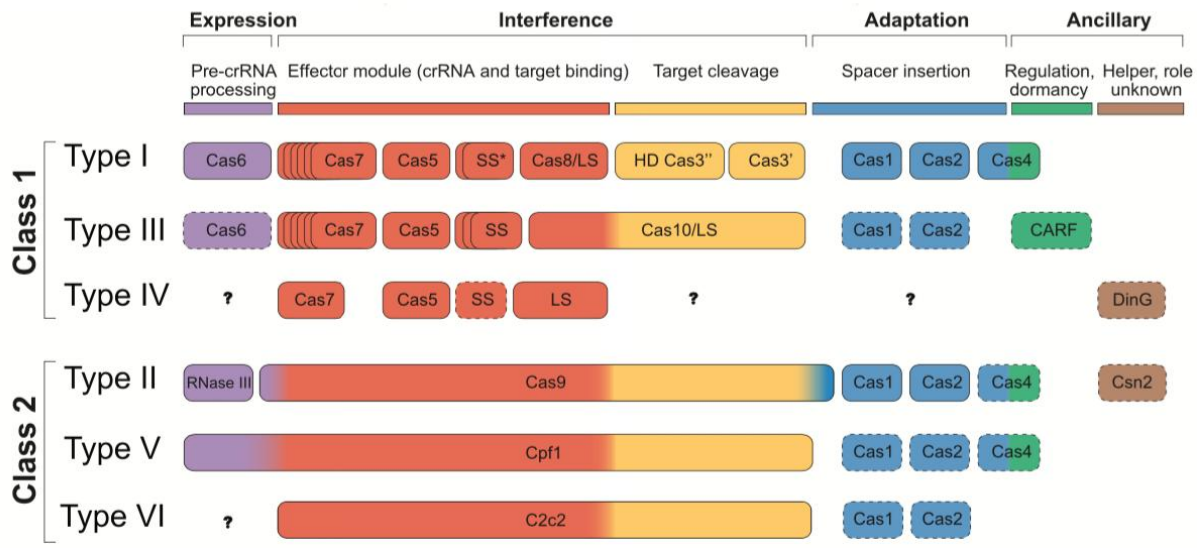


Figure 6. Modular organization of the CRISPR-Cas systems.

Cas proteins or protein regions of each of CRISPR-Cas system Type can be separated into four common functional protein modules: expression, interference, adaptation and ancillary. Small subunit and large subunit are marked on the scheme as SS and LS, respectively. (Adapted from Makarova *et al*., 2015 with permission).

The composition of the adaptation module is what diverges least across all systems. In both Class 1 and Class 2 systems, it includes the most conserved proteins Cas1 and Cas2. Cas4 protein may also be present in several subtypes. Cas6-like proteins are in most cases responsible for the processing of crRNA in Class 1 systems, while cellular RNases III, which are not directly related to CRISPR-Cas machinery, play the major role in this process in Type II systems (the most studied Class 2 systems). An interference module includes an effector module and proteins/domains which are directly responsible for target degradation. The proteins of the effector module together with crRNA form crRNP complexes. The interference module diverges most across the systems. In all subtypes of Class 1 systems, the effector modules include several common families of proteins (Makarova *et al.*, 2015). The core of crRNP complexes consists of several subunits of Cas7-family proteins, which are assembled along the crRNA. The complex also includes a protein of the Cas5 family and a so-

called “large subunit” (a protein of the Cas8 family and Cas 10 family is the large subunit, respectively, in Type I and Type III, while the Csf1 protein has been bioinformatically predicted for Type IV). So-called “small subunits” may be included in the crRNP complex in some subtypes. In addition, in the subtypes I-C, I-E and I-F, endoribonucleases, responsible for the formation of crRNAs, remain bound to mature crRNAs and are also part of the effector complex. The Cas3 protein, which has helicase and nuclease activities, but is not directly a part of the crRNP complex, is responsible for destruction of the target in the Type I system. In Type III systems, the domain with nuclease activity is present in the large subunit (Cas10 protein). Interestingly, for Class 2 single-subunit systems, a single protein – for example, the signature protein of Type II systems, Cas9, – can function not only as an interference module protein but can also participate in the processing of crRNAs and the acquisition of new spacers (Makarova *et al.*, 2015).

2.3. Evolution of CRISPR-Cas systems

It seems that CRISPR-Cas systems first appeared in thermophilic archaea and subsequently spread by horizontal gene transfer (Makarova *et al.*, 2011b). It has been suggested that the ancestral system belonged to Class 1 systems (Makarova *et al.*, 2015; Shmakov *et al.*, 2015). In general, CRISPR-Cas systems evolve rapidly, driven by the constant “arms race” between prokaryotic cells and viruses. Cas1 is the most conserved protein, both from the point of view of its presence in CRISPR-Cas loci and conservation of the amino acid sequence (Takeuchi *et al.*, 2012; Makarova, *et al.*, 2011a). To a large extent, the adaptation module and the effector module of CRISPR-Cas systems evolved independently of each other. In general, there is a very weak correspondence between CRISPR-Cas systems phylogeny and phylogeny of species they are present in (Makarova *et al*, 2015).

2.4. Prevalence of CRISPR-Cas systems subtypes

According to the CRISPRdb database, CRISPR-Cas systems are found in most known archaea and in about half of bacteria (Grissa *et al.*, 2007). The most common systems in the genomes of both archaea and bacteria are of Type I. They represent slightly over half of all the CRISPR-Cas systems that have been identified. Another quarter of all CRISPR-Cas systems are Type III. Type II systems are relatively common in bacterial genomes that have been sequenced to date, while they are not found in the genomes of archaea (Makarova *et al.*, 2015). Almost 25% and about 15% of systems are incomplete or ambiguous in archaea and bacteria, respectively.

The most abundant subtypes among bacteria are I-B, I-C, I-E (20%, 13%, 12%, respectively, from all uniquely classified and complete CRISPR-Cas systems), followed by the III-A and III-B subtypes. I-B and I-A are the most abundant subtypes for archaeon genomes (30%, 18%), followed by III-A and III-B (Makarova *et al.*, 2015). I-F subtype systems are not found in archaea (Vestergaard *et al.*, 2014). In general, different phyla of bacteria are characterized by different trends in the distribution of CRISPR-Cas systems, but variations are also possible (Makarova *et al.*, 2015).

3. Adaptation in Class 1 CRISPR-Cas systems

3.1. Main features

Adaptation is the process of the acquisition of new spacers by the CRISPR array. Adaptation makes CRISPR-Cas immune systems specific, adaptive and heritable. After inserting a spacer into the CRISPR array a prokaryotic cell will transmit this information to daughter cells, so subsequent generations will have immunity to mobile genetic elements with that specific

sequence (Mohanraju *et al.*, 2016). The integration of new spacers occurs mainly in the region bordering the leader sequence of the CRISPR array. This fact makes it possible to establish the order of acquisition of new spacers, and therefore the chronology of interactions with bacteriophages and other MGEs (Shipman *et al.*, 2016).

The process of new spacer acquisition was first detected in the Class 2 CRISPR-Cas system (Barrangou *et al.*, 2007). In Class 1 systems it was only shown in 2012 (Yosef *et al.*, 2012). It has since been demonstrated that adaptation can occur by two fundamentally different mechanisms: naïve (non-primed) adaptation and primed adaptation. Naïve adaptation occurs when the prokaryotic cell interacts with the MGE for the first time and its CRISPR array did not previously contain spacers against this foreign agent (Yosef *et al.*, 2012). Primed adaptation may occur if the CRISPR array already contained a spacer against this foreign agent, but the protospacer or PAM in the target has mutated, disabling the ability to be recognized (Datsenko *et al.*, 2012).

3.2. Proteins involved in the adaptation process

Cas1 and Cas2 proteins play a key role in adaptation. These proteins alone are sufficient for naïve adaptation in the I-E subtype (Yosef *et al.*, 2012). However, primed adaptation also requires the presence of Cas3 nuclease and the crRNP complex, which is called "Cascade" (CRISPR-associated complex for antiviral defense) for I-E CRISPR-Cas systems and sometimes for all subtypes of Type I systems (Datsenko *et al.*, 2012) (see Figure 6).

The *cas4* gene was detected in I-A, B, C, D, U subtypes of Class 1 systems (Makarova *et al.*, 2015) (See Figure 5). In the I-A subtype system of archaeon *Thermoproteus tenax* Cas1, Cas2, Cas4, and Csa1 (Cas4 homolog in archaea) form a complex *in vitro*, suggesting that these proteins participate in the process of acquiring new spacers in this subtype (Plagens *et*

al., 2012). It has also been shown that Cas4 of an archaeon *Sulfolobus solfataricus* exhibits 5'-3' ssDNA exonuclease activity *in vitro*. It was proposed that Cas4 forms single-strand 3'-overhangs of adaptation intermediates (Zhang *et al.*, 2012).

The presence of *cas1* and *cas2* genes is typical (though not necessary) only for the III-A subtype of the Type III CRISPR-Cas systems. *cas1* and *cas2* are rarely located next to the genes of the interference module of the III-B subtype, and *cas1* and *cas2* are not present in loci encoding the genes of III-C and III-D subtypes (see Figure 5). Different subtypes of Type III CRISPR-Cas systems can share Cas1 and Cas2 proteins (Makarova *et al.*, 2015).

Type IV system was predicted by means of bioinformatics, and the *cas1* gene was not found in it (see Figure 5). To date, only one experimental paper has been published, where the crystal structure of Cas2 from the Type IV system of archaeon *Thermococcus onnurineus* was described (Jung *et al.*, 2016). However, there is no data on either adaptation or interference activity for the systems of this Type.

3.3. Naïve adaptation

Only a few cases of naïve adaptation of Class 1 CRISPR-Cas systems have been detected to date (Sternberg *et al.*, 2016). Naïve adaptation has been observed: 1) in archaeon *S. solfataricus*, which has I-A and III-B system subtypes (Erdmann *et al.*, 2012; Erdmann *et al.*, 2013); 2) in *E. coli* I-E subtype under conditions of Cas1 and Cas2 overexpression (Yosef *et al.*, 2012); 3) in an *E. coli*-based strain overexpressing the *P. aeruginosa* I-F subtype CRISPR-Cas system (Vorontsova *et al.*, 2015); and 4) in *Marinomonas mediterranea* bacterium III-B subtype (Silas *et al.*, 2016). The mechanism of generation and integration of new spacers is not fully understood yet.

Subtype I-E of *E. coli* is the most studied of the Class 1 systems. The adaptation complex in *E. coli* cells is assembled from six subunits: a central Cas2 dimer binds to two Cas1 subunits and each of these Cas1 subunits binds to another one Cas1 (Nunez *et al.*, 2014). Cas1 has metal-dependent endonuclease activity against both ssDNA and dsDNA (Wiedenheft *et al.*, 2009). Cas2 has endonuclease activity towards ssRNA (Beloglazova *et al.*, 2008), but the catalytic activity of Cas2 is not required for the adaptation process (Nunez *et al.*, 2014). However, it was shown that the C-end of Cas2 is extremely important. It is required for complex formation with Cas1, which is a prerequisite for integration of new spacers (Nunez *et al.*, 2014). Once formed, the complex can bind to PAM-containing DNA fragments and produce protospacers of the specific length. It also has integrase activity and facilitates embedding of spacers in the CRISPR array in the final stage of adaptation.

The origin of intermediates for adaptation remains unclear. It was suggested in a study by Levy *et al*. that fragments of degraded DNA formed as a result of double-strand break repair could serve as intermediates utilized by Cas1-Cas2 complexes (Levy *et al.*, 2015). When a double-strand break (DSB) occurs in the *E. coli* genome, the RecBCD complex recognizes it, untwists it and performs DNA degradation up to the nearest Chi-site (Dillingham & Kowalczykowski, 2008). It has been shown that "hot spots", i.e., the positions from which new spacers are selected with increased efficiency, are often located between the sites of the replication fork delay (where DSBs occur most frequently) and the Chi-sites closest to them. Further, the artificial introduction of DSBs into a particular position of the *E. coli* genome led to increased acquisition of spacers from the regions surrounding the site of the break, up to the nearest Chi sites (Levy *et al.*, 2015).

ssDNA fragments with lengths between tens and thousands of nucleotides are formed as a result of the nuclease activity of RecBCD (Dillingham & Kowalczykowski, 2008). However, it has been shown *in vitro* that it is possible to insert only dsDNA (Nunez *et al.*, 2015b).

Therefore, if the products of DNA degradation by RecBCD complex are actually used as intermediates, there should also be an additional stage of transition from single-stranded products to double-stranded fragments ready to be embedded in the CRISPR array. This is perhaps accomplished by reannealing of single-stranded complementary degradation products before binding them to the Cas1-Cas2 complex, or the Cas1-Cas2 complex initially binds to ssDNA, and then an enzyme with DNA polymerase activity synthesizes the second strain (Amitai and Sorek, 2016).

The crystal structure of the Cas1-Cas2 complex bound to a DNA intermediate was resolved in 2015 (Figure 7). The DNA intermediate had unwound ends and a 23-nucleotide duplex in the middle. The length of the duplex is determined by the distance between Tyr22 of two Cas1 subunits (marked in Figure 7A as Cas1a and Cas1a'). These tyrosines play the role of wedges, limiting the duplex on both sides and contributing to the unwinding of DNA behind them. The central segment of the duplex is bound to the Cas2 dimer surface. The overhanging 3'-ends of DNA are important: they pass through arginine-rich channel of Cas1a-Cas1b or Cas1a'-Cas1b' to a pocket with the catalytic center (Nunez *et al.*, 2015a; Wang, *et al.*, 2015). This pocket interacts in sequence-specific manner with the 3'-TTC-5' sequence complementary to the PAM (Figure 7B). In the case of the correct PAM, the cleavage of the single-stranded overhang of intermediate is provided by two Cas1 (Cas1a and Cas1a') subunits in such a way that 5 unpaired bases ending in cytosine (from the PAM complementary sequence) remain at one 3'-end and 5 unpaired bases remain at the other 3'-end of adaptation intermediate. This mechanism determines the exact length – 33 pairs of nucleotides – of the future spacer in I-E CRISPR-Cas systems (Wang, *et al.*, 2015; Nunez *et al.*, 2015a).

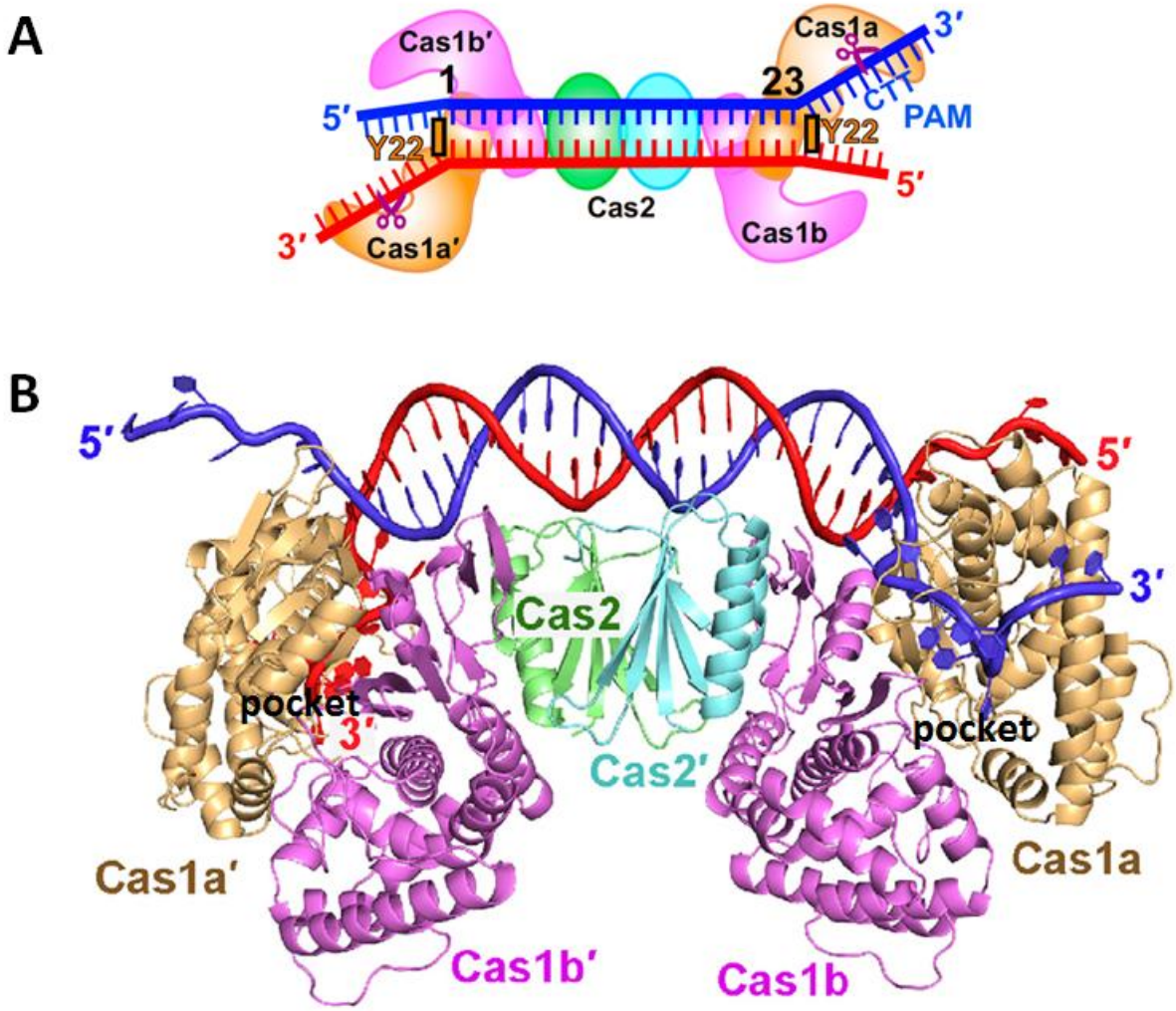


Figure 7. A scheme and a crystal structure (2.6 A) of *E. coli* DNA-bound Cas1-Cas2 complex.

A. A model of Cas1-Cas2 complex bound with double-forked DNA intermediate. In complex, Cas2 dimer (green and cyan) interacts with Cas1b and Cas1b'(magenta) and Cas1b and Cas1b' interact with Cas1a and Cas1a'(orange), respectively. The DNA intermediate has unwound ends and a 23-nucleotide duplex in the middle (positions 1 and 23 are shown). The length of the duplex is determined by the distance between tyrosine 22 residues (marked in the picture as Y22) of subunits Cas1a and Cas1a'. **B.** The crystal structure. The central part of DNA duplex is bound to surface of Cas2-Cas2' dimer. The overhanging 3'-ends of DNA are important: they pass through arginine-rich channel. The PAM-complementary sequence 3'-TTC-5' in the 3' overhang is recognized in a base-specific manner by the Cas1a or Cas1a' catalytic subunits which are formed the pocket together with Cas1b or Cas1b'. Subsequent cleavage at positions 5 nt from the duplex boundary is provided by Cas1a or Cas1a' subunit. (Adapted from Wang *et al*., 2015 with permission).

The new spacer is incorporated into the CRISPR array as a result of two successive nucleophilic attacks aimed at both ends of the repeat closest to the leader (Figure 8).

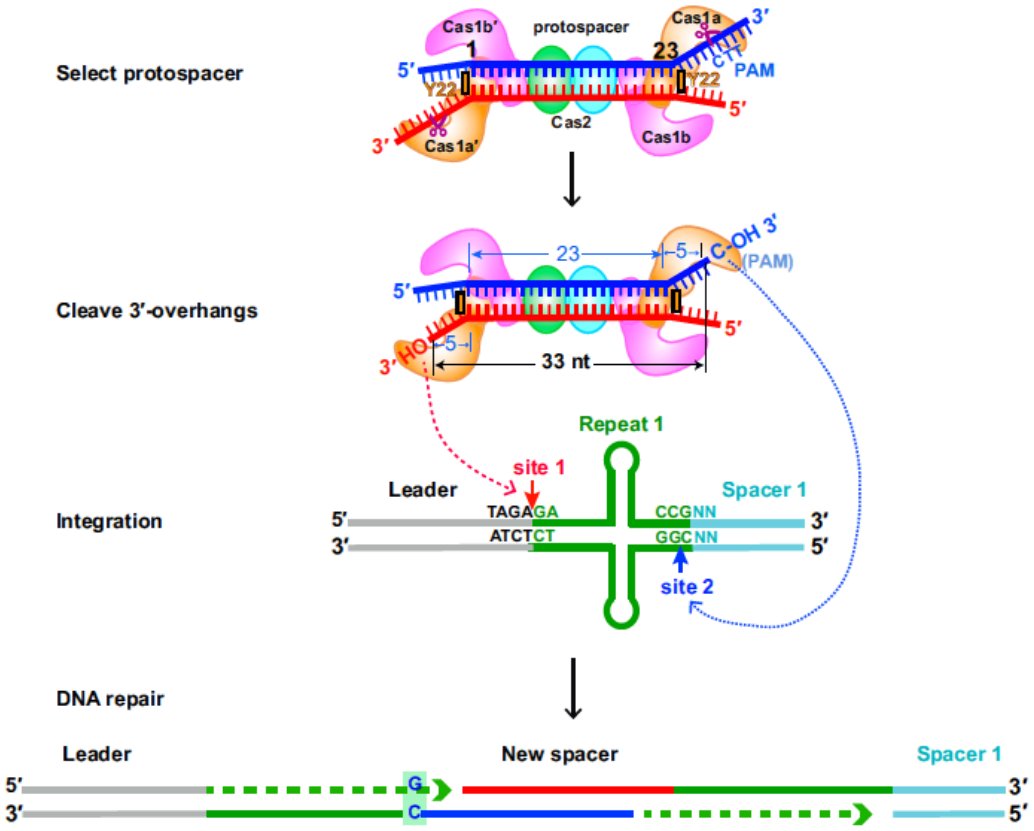


Figure 8. Model of new spacer incorporation into the host I-E CRISPR array.

At the first stage, Cas1-Cas2 complex binds to PAM-containing DNA fragment. Next, Cas1 subunits make cleavages in ssDNA ends and produce intermediates of the specific length, which is determined by Cas1-Cas2 complex architecture. After that, Cas1-Cas2 complex positions the 3'-OH groups of intermediate for nucleophilic attack and the new spacer is incorporated into the CRISPR array as a result of two successive attacks aimed at both ends of the closest to the leader repeat. At the final stage, the gapped duplex is repaired by the host unidentified enzymes. (Taken from Wang *et al.*, 2015 with permission).

It has been shown *in vitro* that the supercoiled form of DNA is required for incorporation (Nunez *et al.*, 2015b). In the first stage, the bound to the future spacer Cas1-Cas2 complex positions the 3'-OH group for nucleophilic attack aimed, in most cases, at the end of the first repeat's non-template strand, which is proximal the leader. The nucleophilic attack of the

second 3'-terminal OH-group is, correspondingly, directed to the leader-distant end of the first repeat's template strand. Thus, the new spacer is fully integrated into the array but is surrounded by single-stranded repeats. There are a template strand of the repeat on the leader-proximal side, and a non-template strand on the leader-distant side. DNA polymerase and DNA ligase are required to fill gaps and to eliminate nicks, respectively, at the final stage of adaptation. It has not yet been determined exactly which enzymes perform these functions (Nunez *et al.*, 2016; Rollie *et al.*, 2015). However, Ivančić-Baće *et al*., have demonstrated that DNA polymerase I is essential for adaptation and it was proposed that it is involved in the filling of gaps (Ivančić-Baće *et al.*, 2015). It should be noted that the attack by cytosine's 3'-OH group is preferable, and this may determine the orientation of the spacer during insertion into the array. It was shown that the precursors of spacers with or without cytosines at both 3'-ends were inserted in both orientations (Nunez *et al.*, 2015b). In addition, the leader sequence (Diez-Villasenor *et al.*, 2013; Nunez *et al.*, 2015b) and repeats (Nunez *et al.*, 2015b) are likely to take part in determining the proper orientation. Also it was shown that integration host factor (IHF) is required for spacer acquisition *in vivo* in I-E subtype *E. coli* system. The IHF binds to the leader and specifies leader-proximal spacer integration (Nunez *et al.*, 2016).

New spacer acquisition occurs in such a way that the new spacer is, in most cases, located closest to the leader (Yosef *et al.*, 2012; Nunez *et al.*, 2015b). However, it was detected that spacers were inserted into various positions of the array with location preference before the 4^{th} spacer in one of the CRISPR arrays of *S. solfataricus* (Erdmann *et al.*, 2012).

In 2016, Shipman *et al.* demonstrated the possibility of *in vivo* integration of dsDNA intermediates into the I-E CRISPR array of *E. coli*. It was shown that the presence of the correct PAM increases the frequency of incorporation and determines the direction of incorporation. Oligonucleotides not containing PAM were incorporated with approximately the same frequency in direct and reverse orientation, while oligonucleotides containing

consensus PAM were incorporated mainly in direct orientation. Mutagenesis of Cas1-Cas2 changed PAM preferences (Shipman *et al.*, 2016).

Type III systems are transcription-dependent. A crRNP complex binds to RNA, complementary to crRNA, and then introduces several cuts into RNA and a cut into the DNA from which the target RNA is transcribed (Tamulaitis *et al.*, 2017) It was found that the chimeric RT-Cas1 protein, in which Cas1 is fused with reverse transcriptase (RT), is typical instead of Cas1 for some Type III CRISPR-Cas systems (Toro *et al*., 2014; Silas *et al*., 2016; Silas *et al.*, 2017). *M. mediterranea* has the III-B subtype system and a gene coding such an RT-Cas1. Silas *et al.* demonstrated that RT-Cas1 of *M. mediterranea in vivo* allows the acquisition of new spacers from both RNA and DNA. Under conditions of overexpression of *RT-cas1*, *cas2* and *marme_0670*, adjusted to them, from a plasmid in *M. mediterranea*, the insertion of spacers originating from both DNA and RNA was detected. The deletion of the RT domain or the mutagenesis of the active site abolished the acquisition of spacers from RNA. RT-Cas1 and Cas2 proteins of *M. mediterranea* catalyze *in vitro* the ligation of RNA spacers with the CRISPR array and the subsequent reverse transcription reaction in presence of deoxynucleotides (dNTPs). In addition, dsDNA and ssDNA spacers are inserted, and the deletion of the RT domain does not inhibit the incorporation of DNA oligonucleotides (Silas *et al.*, 2016).

3.4. Primed adaptation

The phenomenon of priming was discovered in 2012 in the I-E subtype system of *E. coli*. It was shown that a target with an escape mutation in a protospacer could trigger the intensive acquisition of new spacers. Primed adaptation is at least ten times more efficient than naïve adaptation (Datsenko *et al.*, 2012; Savitskaya *et al.*, 2013). It has been suggested that primed

adaptation appeared during the struggle between bacteria and phages as a mechanism, which prevents phages from avoiding CRISPR-Cas action. Accumulation of mutations in a protospacer leads to escape from recognition and destruction, but the acquisition of new spacers triggered by partial recognition of the protospacer restores the immunity of the prokaryotic cell (Datsenko *et al.*, 2012). Mapping of the new spacers acquired during primed adaptation showed that the choice of spacers is carried out in a strand-specific manner. The new spacers are co-directed with the priming spacer so that they will target the same strand of the invader (Datsenko *et al.*, 2012; Swarts *et al.*, 2012). Interestingly, primed adaptation requires not only Cas1 and Cas2 proteins but also all the other components of the CRISPR-Cas machinery: the crRNP Cascade complex and the Cas3 nuclease (Datsenko *et al.*, 2012). Primed adaptation is thus apparently connected with interference (Datsenko *et al.*, 2012; Swarts *et al.*, 2012).

Later it was shown that new spacers come from the area adjacent to the site of priming. The frequency of new spacer selection decreases with distance from the priming site. In the I-E system, the gradient of spacer selection efficiency is asymmetric and strand-specific. Most of the spacers are selected from the upstream region relative to the non-target strain of the priming protospacer. Significantly fewer new spacers are selected from the downstream region and these spacers have an opposite that of the priming spacer (Strotskaya *et al.*, 2017). Two hypotheses have been proposed to explain the strand-specific choice of new spacers (Sternberg *et al.*, 2016). The first hypothesis suggested that the interference complex could slide along the DNA after binding with the priming protospacer until it reaches the PAM sequence, initiating the acquisition of a new spacer. According to this hypothesis, the maximum number of spacers should be selected from the region near to the priming spacer and selection should decrease with distance from this spacer, as indeed observed in the experiments. However, Savitskaya *et al.* showed that the introduction of additional AAG

PAM sequences in the target DNA did not in any way influence the selection of new spacers by the I-E subtype CRISPR-Cas system (Savitskaya *et al.*, 2013). The second hypothesis was that the formed by the action of the Cascade effector complex and the Cas3 helicase-nuclease (Beloglazova *et al.*, 2011; Sinkunas *et al.*, 2011; Westra *et al.*, 2012; Sinkunas *et al.*, 2013) products of target DNA degradation are used as donors for primed adaptation. However, there was a paradox: more efficient primed adaptation was observed in the presence of mutations in the priming protospacer, although the recognition and binding of the crRNP complex with such target is much weaker (Semenova *et al.*, 2011). This paradox was resolved in the work of Semenova *et al.*, which showed that, in fact, the instantaneous rate of primed adaptation is much higher when there are no mutations in the protospacer, but the overall yield is low because of the very high rate of degradation of such targets (Semenova *et al.*, 2016).

Künne *et al.* demonstrated the close relationship between primed adaptation and interference. It was shown that Cas3 forms 30-100 nucleotide fragments *in vitro*, which then bind to complementary fragments and are processed by the Cas1-Cas2 complex into intermediates for adaptation (Künne *et al*., 2016). Primed adaptation in the I-E subtype system of *E. coli* was studied in Musharova *et al.* (Musharova *et al*., 2017). It was shown that Cas1 interacts with DNA intermediates that are distinct from standard double-stranded DNA and have a length typical for spacers of this subtype. It was suggested that these intermediates are formed as a result of cleavage of the non-targeted chain by Cas3 nuclease (Musharova *et al*., 2017). The inconsistency with previously described data on the interaction of Cas1 with dsDNA (Nunez *et al.* 2015; Wang *et al*., 2015) can be explained by observations of adaptation at various stages: Musharova *et al.* observed the early stage, while Nunez and Wang were looking at the later stage.

Primed adaptation has also been demonstrated for subtypes I-B, I-C, I-F of Class 1 CRISPR-Cas systems, but each subtype has its own preferences for choosing new spacers relative to the priming point (Li *et al*., 2014; Rao *et al*., 2016; Richter *et al*., 2014).

4. Biogenesis of crRNAs

Pre-CRISPR RNA is formed in Class 1 CRISPR-Cas systems by a CRISPR array transcription from a promoter located in the leader region. The pre-crRNA is then processed. In most subtypes of Class 1 CRISPR-Cas systems, proteins of the Cas6 family play the key role in the processing of pre-crRNA (Brouns *et al*., 2008; Carte *et al*., 2008). Cas6 family proteins are metal-independent endoribonucleases with two RNA-binding RRM (RNA recognition motif) domains. They introduce a single break into the phosphodiester bond in the repeat, so that short single-spacer crRNAs are formed (Lillestøl *et al*., 2006; Brouns *et al*., 2008). An 8-nucleotide fragment of the repeat with a hydroxyl group is located at the 5'-end of the crRNA molecule and the remaining repeat part ending with 3'-phosphate or 2' 3'-cyclophosphate is located at the 3’-end (Wiedenheft *et al*., 2011b; Haurwitz *et al*., 2010). The Cas6 family proteins show significant structural differences, which can affect the specificity of cleavage. The mechanisms of crRNA formation have specific features in different types of Class 1 CRISPR-Cas systems (Charpentier *et al*., 2015).

Interestingly, the Cas6 protein is completely absent in the subtype I-C. Its function there is performed by the Cas5 protein, which also has two RRM domains. The Cas5 protein introduces a break in the repeat sequence of RNA in such a way that 11 but not 8 nucleotides are located at the 5' end of the crRNA (Garside *et al*., 2012). The crRNA of the I-C, I-D, I-E and I-F subtypes has a stable hairpin necessary for the processing of pre-crRNA. This hairpin is further involved in the formation of a strong interaction between the crRNA and crRNP

complex. After inserting a break in the pre-crRNA the proteins remain connected with a repeat hairpin at the 3'-end of crRNA in the I-C, I-E, and I-F subtypes and become part of the crRNP complex (Charpentier *et al*., 2015). There is no additional processing of crRNA in I-C, I-E and I-F subtype systems (Brouns et al., 2008; Garside *et al*., 2012; Nam *et al*., 2012; Jore *et al*., 2011; Wiedenheft *et al*., 2011). By contrast, the crRNA of I-A, I-B, I-D subtype systems and Type III systems undergo additional processing (Plagens *et al*., 2014; Richter *et al*., 2012; Scholz *et al*., 2013). It is also interesting that two Cas6 subunits participate in the pre-crRNA cut in the I-B subtype (Shao *et al*., 2016). For the III-A subtype system it was shown *in vivo* that not only the Cas6 protein, but also the Cas10 and Csm4 proteins are required for the primary processing of pre- crRNA (Hatoum-Aslan *et al*., 2011). In the III-A and III-B systems, additional processing of crRNA with intracellular exonuclease from 3'-end occurs, but this mechanism has not been studied sufficiently (Carte *et al*., 2008; Walker *et al*., 2017). As a result, a set of mature crRNAs of different lengths is formed in Type III systems. Most of the crRNAs were 37 or 43 nucleotides in length in the III-A subtype CRISPR-Cas system of *Staphylococcus epidermidis.* It is known that at least Csm2, Csm3, and Csm5 proteins (Hatoum-Aslan *et al*., 2011) are required for the complete maturation of crRNA systems of this subtype. The final maturation apparently occurs simultaneously with the assembly of the Csm complex. Six subunits of Csm3 are assembled on immature crRNA, and the unknown nuclease then degrades the 3'-end of the crRNA, thereby ensuring its final maturation. For the III-B CRISPR-Cas subtype of *Pyrococcus furiosus*, it was shown that most of the crRNAs included in the CMR crRNP complex were 45 or 39 nucleotides in length (Hale *et al.*, 2012). Some cases are known when CRISPR-Cas systems coexisting in the genome can "share" the *cas6* gene product between several subtypes of systems. For example, this is the case for systems of subtypes III-A and III-B in *T. thermophilus* Hb27 (Staals *et al.*, 2014).

The processing of crRNA in CRISPR-Cas systems belonging to Class 2 is also worth mentioning. Their mechanisms for cRNAs maturation are significantly diverse even within the same Type and are strikingly different from the mechanism described above for Class 1 systems. For example, in the II-A and II-B subtypes, the effector protein Cas9, ribonuclease III (RNase III) and trans-activating crRNA (tracrRNA) are involved. The tracrRNA binds to the repeat sequence in pre-crRNA on the complementarity principle, Cas9 stabilizes the structure, and RNase III (the enzyme specifically binding and cutting dsRNA) creates breaks. The 5'-end of crRNA is then degraded by the unknown RNase, not belonging to the CRISPR-Cas system machinery (Deltcheva *et al.*, 2011). In the II-C subtype system of *Neisseria meningitidis*, crRNA is transcribed from multiple promoters located inside the repeats of a CRISPR array, but RNase III and tracrRNA are still required for processing of the 3'-end of crRNA, although interference can also occur with non-processed 3'-ends (Zhang *et al.*, 2013). Expression of tracrRNA has also been shown for V-B subtype systems. This may indirectly indicate that the mechanism of crRNA maturation in this recently discovered subtype is similar to the Type II mechanism. The tracrRNA genes were not detected in the V-A subtype and Type VI systems but the pre-crRNA is transcribed and processed (Shmakov *et al.*, 2015). Since these Class 2 CRISPR-Cas systems lack tracrRNA, they may, for example, rely on processing by other cellular RNases or use their own effector proteins. It has recently been shown that the V-A subtype protein Cpf1 itself has RNase activity, which allows it to process pre-crRNA. Cpf1 forms crRNAs with a 5'-terminal hairpin (Fonfara *et al.*, 2016).

5. Interference in Class 1 CRISPR-Cas systems

5.1. Composition and assembly of crRNP complexes

CRISPR interference in Class 1 systems was first shown for the I-E subtype system of *E. coli*. The action of this system against lambda bacteriophage was demonstrated. In addition, the crRNP complex Cascade and crRNAs were isolated and purified for the first time (Brouns *et al.*, 2008). Structural analysis was performed using mass spectrometry, cryo-electron microscopy (cryo-EM) and X-ray diffraction analysis of the Cascade complex of *E. coli* (Jore *et al.*, 2011; Wiedenheft *et al.*, 2011b; Wiedenheft *et al.*, 2011a). Cascade of the I-E subtype systems is a complex of about 400 kDa with "sea-horse" shape and $Cse1_1Cse2_2Cas7_6Cas5_1Cas6_1$ stoichiometry (according to the old nomenclature, these proteins were called CasA, CasB, CasC, CasD and CasE, respectively), having crRNA that is 61 nucleotides in length (Jore *et al.*, 2011) (Figure 9).

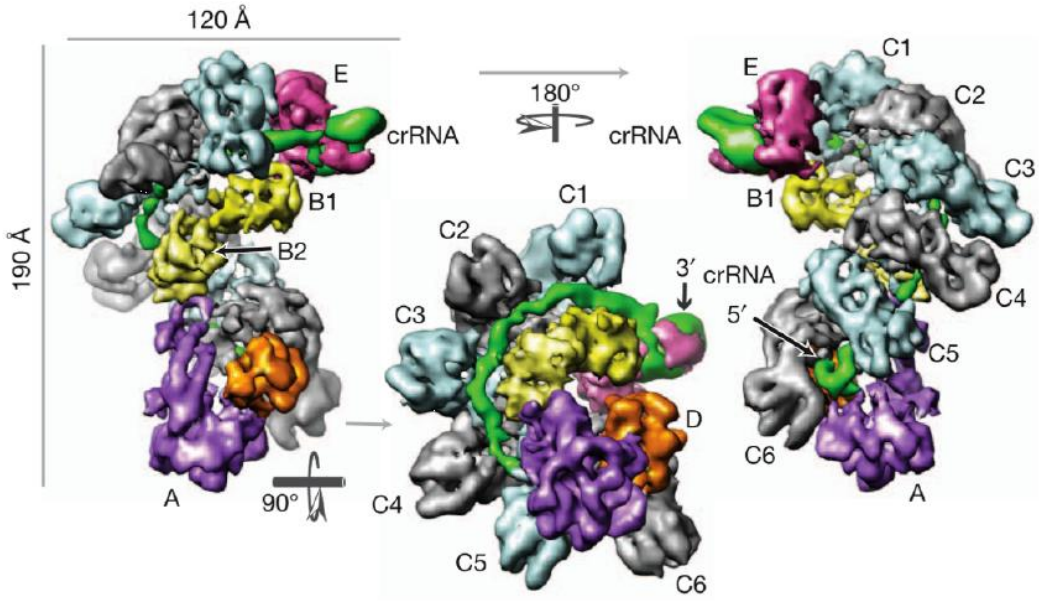


Figure 9. Cryo-EM structure of the Cascade complex from *E. coli*.

Cascade assembles into a sea-horse-shaped architecture where the crRNA (green) is positioned along a helical arrangement of six Cas7 subunits (marked as C1–6). The helical spine is capped at its ends by Cas6 (E, magenta) - the "head" and Cse1 (A, purple) – the "tail" of sea-horse. Cas 5 is marked as D (orange) and two Cse2 – as B (yellow). (Taken from Wiedenheft *et al*., 2011b with permission)

The assembly of Cascade proceeds as follows. First, Cas5 (CasD) is attached to the 5-'end repeat region of the crRNA associated with Cas6 (CasE), then 6 subunits of Cas7 (CasC) are bound to the spacer sequence, then a large Cse1 (CasA) subunit and a dimer of small Cse2 (CasB) subunits are attached. Both large and small subunits take part in binding with DNA. The large subunit is responsible for recognizing the target. After the complex has been bound with the protospacer, the Cas3 protein, which has an N-terminal nuclease HD domain and a C-terminal domain of the helicase belonging to the superfamily 2 (DExD/H), is attracted to the DNA target. The foreign DNA is then destroyed by Cas3 (Makarova *et al.*, 2006; Brouns *et al.*, 2008; Jore *et al.*, 2011).

The modular organization of CRISPR-Cas systems described in section 2.2 above indicates that some common features of systems are observed within the same Type. However, each subtype has its own special qualities, mostly in composition of the crRNP complexes (Plagens *et al.*, 2015). The helicase and nuclease domains are present on the individual Cas3' and Cas3'' proteins in the I-A subtype, and these proteins are a part of the crRNA complex (Plagens *et al.*, 2014). In subtypes I-B and I-C, there is no small subunit, but bioinformatics analysis showed that its fragment is fused to the large subunit Cas8 (Makarova *et al.*, 2015). In subtype I-D, as well as in I-A, the helicase and nuclease domains of Cas3 are separated into different protein sequences. Cas10 protein, which is typical of the Type III system, is partly present in the I-D subtype. This may indicate that the I-D subtype is the Type I system that is closest to Type III systems. The Cas3 nuclease domain Cas3" is fused with a fragment of Cas10. Cas3"-Cas10 protein is a part of the crRNP complex (Makarova *et al.*, 2015). In the I-F subtype, the Cas3 protein is fused with the Cas2 fragment into one Cas2-3 protein (Makarova *et al.*, 2011a; Makarova *et al.*, 2015).

crRNP complexes in III-A and III-D subtype CRISPR-Cas systems are called Csm-complexes, and in III-B and III-C subtypes they are called Cmr-complexes (Haft *et al.*, 2005; Rouillon *et al.*, 2013; Spilman *et al.*, 2013; Staals, *et al.*, 2013; Makarova *et al.*, 2015). The *cas10* gene is typical for all subtypes of Type III systems. It encodes a multidomain protein - a large subunit of complexes. Sometimes *cas10* is called *csm1* and *cmr2* for III-A and III-B subtypes, respectively (Tamulaitis *et al.*, 2017). The Cas10 protein has an N-terminal nuclease HD-domain and a Palm-domain with a GGDD motif (Makarova *et al.*, 2015). Small subunits, Csm2 and Cmr5, are present in both crRNP complexes in three copies (Rouillon *et al.*, 2013; Staals, *et al.*, 2013). The proteins of the Cas5 family are represented by Csm4. The proteins of the Cas7 family are represented by five identical Csm3 subunits and one Csm5 subunit in the III-A subtype complexes, and by four Cmr4, one Cmr6 and one Cmr1 subunits in the III-B subtype complexes. Csm3 and Cmr4 proteins are endoribonucleases. Interestingly, the locus of III-A systems usually includes *cas1*, *cas2* and *cas6* genes, while most of III-B do not have them and use proteins encoded by the corresponding genes of other systems (Makarova *et al.*, 2015).

5.2. Mechanisms of interference in Type I CRISPR-Cas systems

Two conditions are required for interference by Type I CRISPR-Cas systems: the complementary binding of a protospacer with crRNA located in the crRNP complex and the presence of a consensus PAM. It has been shown that certain positions in the protospacer are extremely important for its recognition by the crRNP complex. These positions form the so-called “seed” region. In Type I systems seed is located at the PAM-proximal region of the protospacer. It was shown that even a single mutation in PAM or seed sequence makes impossible the CRISPR-Cas-mediated destruction of the target DNA (Semenova *et al.*, 2011).

Similar mechanisms of target recognition and destruction are typical for all systems of Type I (Mohanraju *et al.*, 2016). The crRNP complex recognizes the correct PAM, then the target DNA is partially melted and is paired within the 7-8 nucleotide seed sequence with the crRNA, and formation of the R-loop begins (Jore *et al.*, 2011). If complementary with the seed is complete, the formation of extended R-loop occurs, and ATP-bound Cas3 helicase-nuclease is recruited (Mulepati *et al.*, 2013). Cas3 introduces a break into the DNA non-target strand and performs the unwinding and degradation of the target (Szczelkun *et al.*, 2014). Thus, PAM and seed sequence are critical for interference (Rutkauskas *et al.*, 2015). If there is an incomplete correspondence between crRNA and seed, the formation of the R-loop stops when the uncoupling occurs, the Cas3 cannot be recruited and interference does not take place (Blosser *et al.*, 2015).

5.3. The connection between interference and primed adaptation

The differences between mechanisms leading to interference and to primed adaptation in Type I systems have been described using biophysical single-molecule methods. Fluorescence resonance energy transfer (FRET) technology showed that some binding can occur without the presence of the correct PAM and without fully-complementary seed. This non-canonical binding is provided by pairing with any part of the protospacer. The binding with a protospacer with incomplete matching is short-lived (Blosser *et al.*, 2015). In this case, Cascade cannot switch into the conformation necessary for Cas3 docking. Instead, primed adaptation can be stimulated (Redding *et al.*, 2015). So the complex shows two different binding models with fully complementary or partially complementary targets, which lead to interference or primed adaptation, respectively.

Redding *et al.* showed that the recruitment of Cas3 occurs by two fundamentally different ways in case of a target with full complementarity and in case of mutated targets (Redding *et al.*, 2015). When there are mutations in the PAM or seed region of the protospacer, favoring primed adaptation rather than the destruction of the target, recruitment of Cas3 is possible only in the presence of a Cas1-Cas2 complex. The results indicate that binding to the mutant target occurs less frequently, but the average lifetime in the bound state is comparable: 57 and 40 minutes for fully matching and mutant targets, respectively. In the latter case, Cas3 is not recruited directly, but by means of Cas1-Cas2. In the case of a mutant protospacer, Cas3 is proposed to move along DNA in any direction, but there was no evidence of the introduction of single- or double-strand breaks into the target by Cas3 (Redding *et al.*, 2015). Xue *et al.*, using FRET technology, discovered that the conformation of the Cse1 (CasA) subunit of the Cascade complex determines whether the target DNA will be destroyed or whether primed adaptation will be initiated. Thus, in the absence of mutations in the PAM and seed region of the protospacer, Cse1 goes into a "closed" state, which leads to direct recruitment of Cas3 followed by interference, while the presence of mutations blocks the Cse1 in the "open" state (Xue *et al.*, 2016).

This dual role of Cascade enables the CRISPR-Cas system to efficiently destroy fully-matched targets and to initiate the acquisition of new spacers from mutated targets, thereby updating CRISPR “memory”.

5.4. Mechanisms of interference in Type III CRISPR-Cas systems

Although the shapes of crRNP in Type III systems and Type I systems are similar (Rouillon *et al.*, 2013; Spilman *et al.*, 2013; Jackson *et al.*, 2014), the mechanisms of their action are

fundamentally different. Both Csm and Cmr complexes function as ssRNases (Hale *et al.*, 2009; Tamulaitis *et al.*, 2014; Staals *et al.*, 2014) and RNA-activated ssDNases, which can cut DNA that is transcribed (Deng *et al.*, 2013; Goldberg *et al.*, 2014; Samai *et al.*, 2015). The binding of crRNP complexes to complementary RNA activates the sequence-non-specific deoxyribonuclease activity of the Cas10 subunit. Each of the Csm3/Cmr4 subunits demonstrate endoribonuclease activity (Tamulaitis *et al.*, 2014; Benda *et al.*, 2014; Ramia *et al.*, 2014), each of them cut RNA, and this action deactivates the Cas10 deoxyribonuclease activity (Elmore *et al.*, 2016; Estrella *et al.*, 2016; Kazlauskiene *et al.*, 2016). Thus, the action of Type III CRISPR-Cas systems is directed to transcriptionally-active MGE.

It is interesting to note that ability of the Csm-complex of the archaeon *T. onnurineus* to perform RNA-independent ssDNA activity directed to the targeted strand of protospacer was demonstrated in a recent *in vitro* study (Park *et al.*, 2017).

Another distinctive feature of Type III CRISPR-Cas systems is that they act by a PAM-independent mechanism. The 8-nucleotide 5'-handle of crRNA derived from the repeat serves to prevent autoimmunity. Thus, if there is a complementary pairing between the 5'-handle and the target, the system recognizes it as the CRISPR array transcript, and interference does not occur (Marraffini *et al.*, 2010; Deng *et al.*, 2013). However, it was later shown for the III-A subtype system that interactions only in the –2, -3, -4 and -5 positions of the 5’-handle are spatially possible (Kazlauskiene *et al.*, 2016). It also seems to be the case in III-B that not all nucleotides of the 5’-handle can interact with the target RNA (Osawa *et al.*, 2015). Binding in three positions only (-2 to -4) was experimentally shown to be sufficient to prevent interference (Kazlauskiene *et al.*, 2016). This is consistent with earlier data that the introduction of mutations in either the -2 and -3 or the -3 and -4 positions of the crRNA is sufficient to cancel the protection from autoimmunity (Marraffini *et al.*, 2010).

Interestingly, complete absence of the 3’-end of target RNA corresponding to the 5'-handle of crRNA did not lead to activation of Cas10 in the III-A system *in vitro* (Kazlauskiene *et al.*, 2016), but activated Cas10 in III-B systems (Elmore *et al.*, 2016; Estrella *et al.*, 2016).

Study of the III-B subtype system of the archaea *Pyrococcus furiosus* showed that uncoupling of the 5'-handle was insufficient for identification of an RNA target as foreign. It was also found that the PAM is located on targeted RNA at the 3'- end of the protospacer. This is quite surprising given that Cas10 can be activated even when there is a complete lack of RNA flanking the protospacer at the 3'-end (Elmore *et al.*, 2016).

6. Anti-CRISPR

MGEs can avoid the action of CRISPR-Cas systems by acquiring mutations in their protospacer or PAM region. Another powerful avoidance strategy was discovered in 2013. Genes encoding anti-CRISPR proteins were found in the genomes of some bacteriophages. These are the small 50-150 amino-acid residue proteins (Bondy-Denomy *et al.*, 2013; Bondy-Denomy *et al.*, 2015). Proteins that can counteract the CRISPR-Cas systems of subtypes I-E, I-F and Class2 subtype II-A systems have now also been discovered. These discoveries may indicate that this mechanism is widespread (Pawluk *et al.*, 2014; Bondy-Denomy *et al.*, 2013; Rauch *et al.*, 2017).

Anti-CRISPR genes were first discovered in the genomes of bacteriophages infecting *P. aeruginosa*. The products of these genes were able to inhibit the action of the I-F subtype CRISPR-Cas systems present in various *P. aeruginosa* strains, but did not affect the functioning of the related I-E subtype. It was found later that anti-CRISPR proteins, which inhibit the action of the I-E subtype system of *P. aeruginosa* SMC4386 and SMC4389, are

present in bacteriophages infecting *P. aeruginosa*. However, they did not inhibit the I-E subtype systems of *E. coli* or the I-F system of *P. aeruginosa*.

The mechanisms of action of these proteins differ (Bondy-Denomy *et al.,* 2015). Studies have been carried out on the mechanisms of inhibition of four anti-CRISPR proteins (AcrF1, AcrF2, AcrF3, AcrF4), which prevent the action of the I-F subtype CRISPR-Cas systems. AcrF3, in homodimeric form, interacts with the non-active ADP-bound form of nuclease-helicase Cas3. Mutagenesis of the amino-acid residues of AcrF3, which participate in the interaction with Cas3, abolishes the inhibitory effect of AcrF3 and breaks off the interaction. Apparently, these amino-acid residues are what determines the specificity of *P. aeruginosa* Cas3 recognition by AcrF3, since the amino-acid alignment of Cas3 belonging to the I-F and I-E subtypes shows that all but one of the amino-acid residues involved in Cas3-AcrF3 interactions are different (Huo *et al.,* 2014). AcrF1, AcrF2 and AcrF4 proteins interact with different subunits of the Csy complex, thereby blocking its DNA-binding activity and preventing the action of the CRISPR-Cas system at the target recognition stage. In the Csy crRNP complex, the heterodimer Csy1-Csy2 is bound with the 5'-end of crRNA, the Csy4 monomer is linked to the 3'-end, and between them, there are six Csy3 subunits covering the spacer part of the crRNA. The AcrF2 protein binds to a Csy1-Csy2 dimer. AcrF1 proteins bind to the Csy3 core along its entire length in a 1: 2 ratio and block the Csy-complex in a conformation that is unable to bind with target DNA (Maxwell *et al.,* 2016).

Several different families of anti-CRISPR proteins encoded in the genomes of various bacteriophages infecting *Pseudomonas* and *Pectobacterium atrosepticum* were subsequently described. All of them inhibit the action of I-F or I-E subtype CRISPR-Cas systems. A universal AcrF6 inhibiting I-F and I-E subtype systems of *P. aeruginosa* was also discovered. However, different amino-acid residues of AcrF6 are important for action against various

CRISPR-Cas system subtypes. Also, AcrF6 homologs were found in prophages, conjugative elements and even in regions not annotated as MGEs of different proteobacteria. Some of them inhibit the I-F systems of *P. aeruginosa*, but none of them acts against the I-E subtype (Pawluk *et al.*, 2016a).

Anti-CRISPR proteins have different sequences (Bondy-Denomy *et al.*, 2013), structures (Maxwell *et al.*, 2016; Wang *et al.*, 2016), and mechanisms of action (Bondy-Denomy *et al.*, 2015). These facts support the hypothesis of the independent evolution of inhibitors of CRISPR-Cas systems and suggest the existence of many more unknown anti-CRISPR proteins. Recently, three families of anti-CRISPR proteins inhibiting the II-C and II-A subtype of Class 2 CRISPR-Cas systems were identified. They bind with the Cas9 protein and prevent its binding with DNA and subsequent DNA cleavage. This discovery of anti-CRISPR action against Class 2 CRISPR-Cas systems confirms the hypothesis of the widespread distribution of anti-CRISPR proteins (Pawluk *et al.*, 2016b).

Chapter 2: Study of adaptation in I-F subtype CRISPR-Cas systems

1. Introduction

I-F subtype CRISPR-Cas systems are evolutionarily the closest to I-E subtype (Makarova *et al.*, 2015). The I-F crRNP complex, called the "Csy complex", has a mass of about 350 kDa. In most cases it consists of only four different proteins: 1) a large subunit Csy1; 2) Csy2 (a protein of the Cas5 family); 3) Csy3 (a protein of the Cas7 family, which is present in the complex in six copies); and 4) the Cas6-like Csy4 protein. There is no small subunit in the I-F crRNP complex. Csy1 does not belong to the Cas8 family, but is an analogue of Cas8 (Makarova *et al.*, 2011a; Wiedenheft *et al.*, 2011a; Makarova *et al.*, 2011b). Further, it was found that, in the genome of the bacterium *Shewanella putrefaciens* CN-32, for which activity of the CRISPR-Cas system has been demonstrated *in vivo*, there is no *csy1* gene encoding the large subunit and its crRNP complex includes only three types of proteins (Dwarakanath *et al.*, 2015). However, the most unique feature of I-F subtype is that it is the only subtype CRISPR-Cas system where the Cas2 protein, which plays a critical role in the adaptation stage, is fused with the Cas3 protein responsible for the interference stage.

The question arises of the influence of this fusion on the functional activity of Cas2-3 and particularly on the relationship between the adaptation and interference stages of immunity. Data on adaptation in I-F subtype CRISPR-Cas systems appeared for the first time in 2012. In the work of Cady *et al.*, single cases of incorporation of new spacers into the CRISPR array were described during the interaction of the bacterium *P. aeruginosa* PA14 with a lytic bacteriophage (Cady *et al.*, 2012). In 2014, Richter *et al.* demonstrated capacity of I-F subtype CRISPR-Cas system for primed adaptation in the bacterium *P. atrosepticum*. About 350 acquired spacers were analyzed (Richter *et al.*, 2014). The analysis showed that new spacers target both strands in contrast to I-E subtype, where most of the new spacers target the same strand as was targeted by the priming protospacer (Datsenko *et al.*, 2012). In I-F subtype, the strand bias is also observed around the priming protospacer, but, unlike I-E subtype

(Strotskaya *et al.*, 2017), an approximately equal number of new spacers originate from the upstream and downstream regions (Richter *et al.*, 2014). Moreover, in I-F subtype, most of the acquired spacers downstream from the priming point are co-directed with the priming spacer, and spacers upstream from the priming point have the opposite orientation (Richter *et al.*, 2014). This situation is reciprocal to that in I-E subtype (Strotskaya *et al.*, 2017) and I-B subtype systems (Li *et al.*, 2014). In 2015, it was demonstrated that, in conditions of phage infection, the *P. aeruginosa* PA14 I-F subtype CRISPR-Cas system is also capable of efficient primed adaptation (Westra *et al.*, 2015). The requirements for adaptation in I-F subtype remained unclear.

The present work establishes the requirements for naïve and primed adaptation by I-F CRISPR-Cas system. As an experimental system, overexpression of the I-F CRISPR-Cas system from *P. aeruginosa* PA14 in a heterologous *E. coli* strain BL21-AI background is used. The analysis of acquired spacers was carried out with high-throughput sequencing with deep coverage (1M of reads), which revealed a much larger pool of new spacers in comparison with previous studies. Adaptation by the I-F CRISPR-Cas system naturally present in *E. coli* strain ED1a (Diez-Villasenor *et al.*, 2010) was also analyzed.

2. Materials and Methods

2.1. Bacterial strains and plasmids

E. coli strains used are listed in Table 1.

Strains	Description	Reference
NovaBlue	endA1 hsdR17 (r_{K12}^{-} m_{K12}^{+}) supE44 thi-1 recA1 gyrA96 relA1 lac F′[$proA^{+}B^{+}$ $lacI^{q}ZΔM15$::Tn10] (Tet^{R})	Novagene
KD604	BL21-AI_ΔI-E_CRISPR carrying *Pseudomonas aeruginosa* CRISPR array with a single spacer: ACGCAGTTGCTGAGTGTGATCGATGCCATCAG	This study
KD606	BL21-AI_ΔI-E_CRISPR carrying *Pseudomonas aeruginosa* CRISPR array with a single spacer: ACCGGACCTTCAATCGGCCCTTCGCTGATGGC	This study
KD628	BL21-AI_ΔCRISPR carrying *Pseudomonas aeruginosa* CRISPR array with a single repeat	This study
KD675	The same as KD604 but containing a protospacer with a mismatch at position +1 (**T**CGCAGTTGCTGAGTGTGATCGATGCCATCAG) preceded by a functional GG PAM introduced in its genome in the *ompL/yihN* intergenic region corresponding to the positions 4372171-4372261 of NC_012947	This study
ED1a	*E. coli* ED1a strain with a native I-F CRISPR-Cas system	Touchon, M., and Rocha, E. P. C., 2010

Table 1. Strains used for study of adaptation in I-F subtype CRISPR-Cas systems.

KD604, KD606, KD628, KD675 were engineered from the BL21-AI strain by Dr. Kirill Datsenko and kindly provided. KD604, KD606 and KD675 contain a minimized I-F subtype *P. aeruginosa* UCBPP-PA14 CRISPR array (two repeats and one spacer) and a 134 bp-long

upstream leader region under the control of the T7 RNA polymerase promoter. KD628 contains only a leader (134 bp) and a single repeat. The sequences of the KD604 and KD606 spacers are, respectively, 5'-ACGCAGTTGCTGAGTGTGATCGATGCCATCAG-3' and 5'-ACCGGACCTTCAATCGGCCCTTCGCTGATGGC-3'. KD675 is the same as KD604 but also contains a protospacer with a mismatch at position +1 preceded by a functional GG PAM introduced in its genome.

E. coli ED1, a strain with a native I-F CRISPR–Cas system, is described elsewhere (Touchon and Rocha, 2010) and was kindly donated by Dr. Erick Denamur.

Plasmids pCas (expressing *cas1* and *cas2–3* under the control of the T7 RNA polymerase promoter) and pCsy (expressing *csy1, csy2, csy3* and *csy4* under the control of the T7 RNA polymerase promoter) were kindly donated by Dr. Blake Wiedenheft. pCsy is described in (Wiedenheft *et al.*, 2011a), pCas was constructed by cloning *cas1* and *cas2-3* genes from *P. aeruginosa* UCBPP-PA14 into the pET empty polycistronic destination vector (2Z) (AddGene ID 29776).

The pACYC-Duet-1 (Novagen) and pT7Blue (Novagen) vectors were also used in this study. The plasmids pSPA and pSPAmut were generated by cloning double-stranded oligonucleotides containing perfectly matching or A1T mutant protospacer sequences (the latter harbors an A to T substitution at the first position of the seed) and a consensus GG PAM into the EcoN I and Kpn I restriction sites of the pACYCDuet-1 vector. The pSED and pSEDmut plasmids were generated by cloning double-stranded oligonucleotides containing perfectly matching or T1A mutant protospacer sequences and consensus GG PAM in the EcoRV site of the pT7blue vector.

pHERD30T vector and pHERD30T-based plasmids, carrying genes coding for anti-CRISPR proteins under control of arabinose inducible promoter were kindly donated by Dr. Joe Bondy-Denomy and described in (Bondy-Denomy *et al.*, 2013).

2.2. Media

Bacterial cultivation was carried out using LB media (1% w/v tryptone, 1% w/v NaCl, 0.5% w/v yeast extract) or SOC (2% w/v tryptone, 0.5% w/v yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$, 20 mM glucose). Bacteria were grown at 37 °C on an orbital shaker (180 rpm), unless stated otherwise. Agar was added to LB medium up to 1.5% w/v concentration for cultivation on plates. If required, the antibiotics ampicillin (100 μg/ml), spectinomycin (50 μg/ml), chloramphenicol (34 μg/ml), gentamicin (30 μg/ml) were used for maintaining pCsy/pT7Blue, pCas, pACYC-Duet-1 and pHERD30T plasmid, respectively. LB media was supplemented with 1 mM arabinose and 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) to induce T7 polymerase expression in the BL21-AI-based strains.

2.3. Preparation and transformation of electrocompetent cells

An overnight culture of cells of the required *E. coli* strain was diluted (1:100) and grown in the LB medium up to OD_{600} 0.6–0.8. The cells were then processed to prepare electrocompetent cells using a standard protocol (Miller and Nickoloff, 1995), after which 50 μl aliquots were transformed with 10-50 ng of one or two different plasmids at the same time with a BioRad MicroPulser using the standard *E. coli* protocol, provided by the manufacturer.

After 1-h outgrowth in 1 ml of SOC medium, the cells were deposited onto an LB agar plate containing required antibiotics. The plate was incubated overnight.

2.4. Molecular cloning

Complementary oligonucleotides (synthesized by Integrated DNA Technologies) were phosphorylated with T4 Polynucleotide Kinase (New England Biolabs), as described in the protocol provided by the manufacturer. The reaction mixture was then heated to 95 °C for 10 minutes in a water bath and annealed by slowly cooling to room temperature. The plasmids were restricted with required restriction endonucleases (New England Biolabs), as described in the protocol provided by the manufacturer. Calf Intestinal Alkaline Phosphatase (New England Biolabs) was then added to the reaction. The restricted plasmids were purified with GeneJET Gel Extraction and DNA Cleanup Micro Kit (Thermo Fisher Scientific) immediately after enzymatic reaction or after 0.8% w/v agarose gel electrophoresis in 0.5x Tris/Borate/EDTA (TBE) buffer (44.5 mM Tris, 44.5 mM boric acid and 1 mM EDTA). The plasmid was ligated with double-stranded oligonucleotides taken in large excess using T4 DNA ligase (New England Biolabs) in the conditions described in the manufacturer's protocol. NovaBlue ultracompetent cells were transformed with ligation mixture according to the manufacturer's instructions and plated on agar media with antibiotics. All insert sequences were confirmed by sequencing performed by Integrated DNA Technology

2.5. Plasmid extraction

Plasmids were extracted using a GeneJET Plasmid Miniprep Kit (Thermo Fisher Scientific) from 5 ml of single colony-derived overnight culture in LB medium supplemented with antibiotics.

2.6. Site-specific mutagenesis

Mutations in selected *cas* or *csy* genes were introduced by polymerase chain reaction (PCR) with PfuUltra II Fusion HS DNA Polymerase (Agilent Technologies) using oligonucleotides containing desired mutations. The following PCR parameters were used: 95 __ – 30 seconds; [95 °C – 30 seconds, 55 °C – 1 minute, 68 °C – 1 minute/kb]*12 cycles. Following PCR, the reactions were cooled to 37 °C and proceed with DpnI (New England Biolabs) digestion and transformation of NovaBlue cells. All insert sequences were confirmed by sequencing performed by Integrated DNA Technology.

2.7. CRISPR interference assay

Electrocompetent *E. coli* ED1a cells were prepared as described above. The BL21-AI-based *E. coli* cells containing pCas and/or pCsy plasmids were grown in LB medium in the presence of inducers and antibiotics required for maintenance of the plasmids and were processed to prepare electrocompetent cells. Cells were transformed with 10 ng of target protospacer plasmids or control vector. After 1-h outgrowth in 1 ml of LB medium, 10 µl aliquots of serial dilutions of transformation mixtures were deposited onto LB agar plates containing all required antibiotics. The plates were incubated overnight and growth results recorded. Each experiment was conducted at least in triplicate.

2.8. CRISPR adaptation assay

To induce adaptation in BL21-AI-based strains, clones of cells transformed with selected plasmids were grown in LB supplemented with arabinose and IPTG for a total of 72 hours. Every 24 h aliquots of cultures were diluted (1:500) into fresh medium. The adaptation was detected by PCR with a pair of primers, one (forward) annealing in the leader sequence and another (reverse) at the CRISPR array spacer. For KD628, the reverse primer annealed downstream of the single CRISPR repeat present in this strain. Cell adaptation in KD675 was detected after overnight cultivation. After amplification, reaction products were analyzed by 1.5% w/v agarose gel electrophoresis in 0.5x TBE buffer. The procedures for ED1a cells were the same except that no inducers were added to the media and a different primer set was used for amplification.

2.9. High-throughput sequence data analysis

DNA bands corresponding to expanded arrays were cut and extracted from the gel with GeneJET Gel Extraction and DNA Cleanup Micro Kit. High-throughput sequencing of the DNA was performed by Dr. Maria Logacheva using the MiSeq Illumina system.

The received data were preprocessed and analyzed using ShortRead (Morgan *et al.*, 2009) and BioStrings (Pages *et al.*, 2012) Bioconductor packages. Sequences located between two CRISPR repeats were considered to be spacers. They were mapped onto the genome and all plasmids presented in the sample, with no mismatches allowed. The following procedure was used to assign non-unique spacers from common regions of pCas and pCsy plasmids. First, spacers originating from unique regions of each plasmid were counted. The ratio of the number of such unique spacers, normalized for the length of each plasmid, was taken as a measure of mean spacer acquisition efficiency from each plasmid. The non-unique spacers

that could have originated from either plasmid were next assigned to either pCas or pCsy based on this measure. R scripts were used for statistical analysis and Circos (Krzywinski, 2009) was used for graphical representation of the data.

3. Results

3.1. Experimental model

The *E. coli* KD604 strain containing, under the T7 RNA polymerase promoter control, a minimized *P. aeruginosa* CRISPR array (two repeats, one spacer, and a leader sequence) inserted in its chromosome, was transformed with two compatible plasmids, pCas and pCsy, expressing all *P. aeruginosa* subtype I-F CRISPR-associated genes (Figure 10).

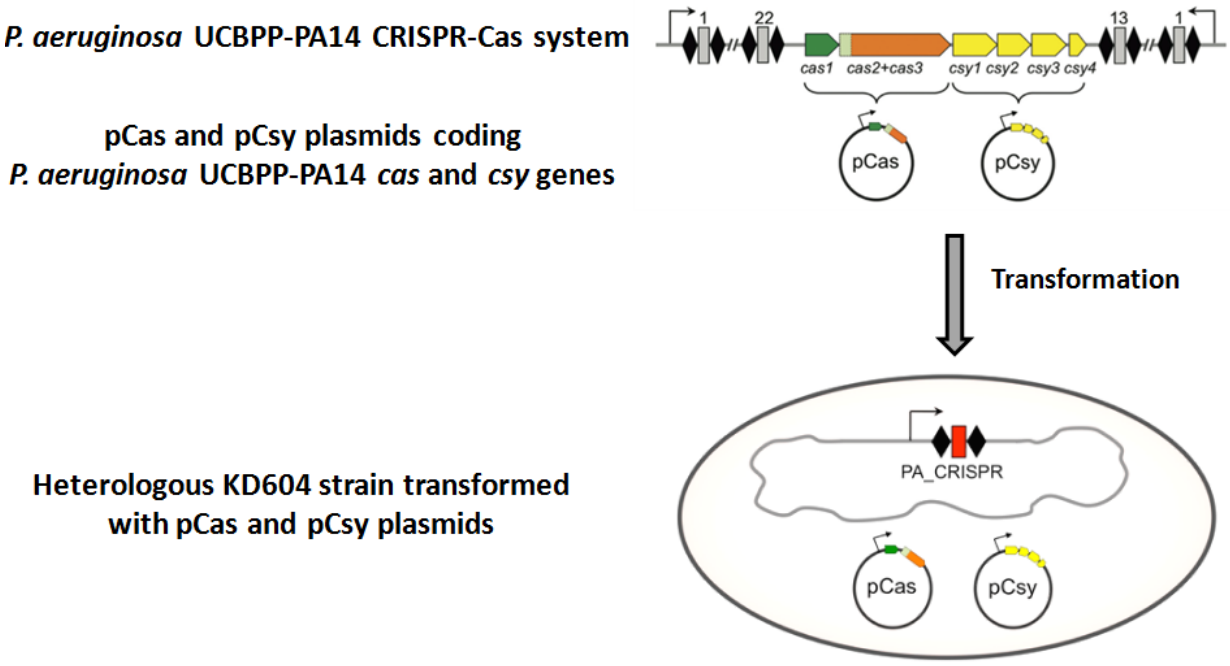


Figure 10. Creation of the experimental model for study of the *P. aeruginosa* I-F CRISPR-Cas system in heterologous *E. coli* background.

A minimized I-F CRISPR array was inserted into *E. coli* genome under control of T7 promoter and all genes of *P. aeruginosa* I-F CRISPR-Cas system were overexpressed from plasmid vectors.

To determine whether the transplanted subtype I-F immune system from *P. aeruginosa* is capable of interference in *E. coli*, a plasmid containing a protospacer flanked by a GG consensus PAM (pSPA) and a plasmid with a mutant version of this protospacer with a single mismatch in the first position of the seed region (pSPAmut) were constructed (Figure 11).

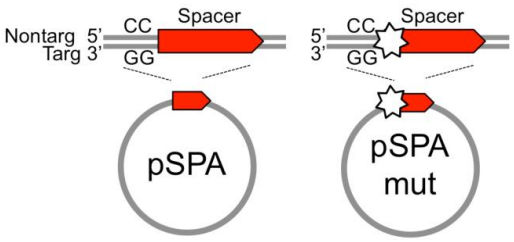


Figure 11. pACYC-based protospacer plasmids pSPA and pSPAmut.

The plasmids carry a protospacer matching the spacer present in the KD604 genomic CRISPR array bordered by a consensus GG PAM. In pSPAmut the first position of the protospacer carries a mutation introducing a mismatch with the KD604 spacer.

Pre-induced KD604 cells harboring pCas and pCsy were transformed with pSPA, pSPAmut or a control pACYC plasmid with no protospacer. The efficiency of transformation (EOT) of pSPA was ~15% of EOT of the pACYC vector. For pSPAmut, EOT decreased less, to ~70% of that of the pACYC vector (Figure 12).

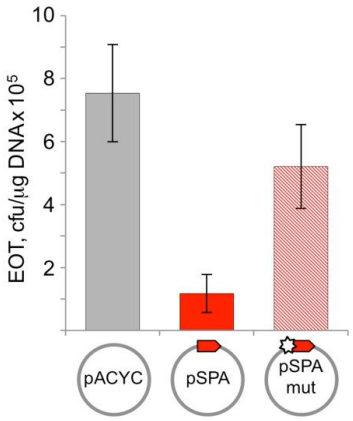


Figure 12. Efficiency of transformation of pSPA, pSPAmut, and a control vector with no protospacer.

Induced KD604 cells co-expressing *P. aeruginosa cas* and *csy* genes from pCas and pCsy, respectively, were transformed with pSPA, or pSPAmut, or pACYC plasmids. Bars represent mean EOT values obtained in three independent experiments with standard deviations.

Thus, although the effects are modest, the subtype I-F immune system from *P. aeruginosa* is capable of interfering with plasmid transformation in the *E. coli*-based KD604 strain and this

immune response relies on complementary base pairing between the crRNA-spacer sequence and the protospacer. The weak level of interference may be caused by a heterologous background (suggesting that an additional factor is required) or may simply be due to the particular spacer-protospacer pair chosen.

To determine whether the transplanted immune system from *P. aeruginosa* was capable of incorporating new spacer sequences into the minimal CRISPR locus, *E. coli*-based KD604 cells were transformed with pSPA, pSPAmut or pACYC and cultured in LB-media supplemented with arabinose and IPTG to induce expression of the *cas* and *csy* genes. Aliquots of cell cultures were subjected to PCR with a pair of primers amplifying the CRISPR array and the proximal leader sequence. Acquisition of new spacers is reflected by the appearance of PCR products longer than the 176 bp fragment amplified from the genomic DNA of the starting cells. Robust spacer acquisition was detected in cells transformed with pSPA, pSPAmut or pACYC (Figure 13).

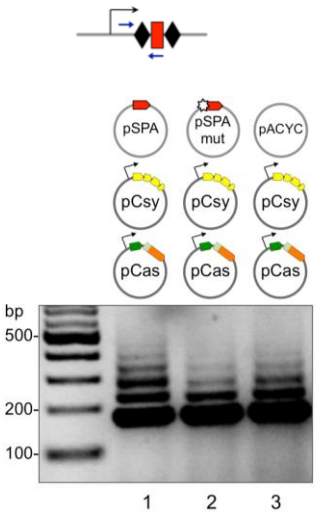


Figure 13. Spacer acquisition by the *P. aeruginosa* CRISPR-Cas system in a heterologous background in conditions of priming.

Results of a CRISPR adaptation experiment in KD604 cells transformed with pCas, pCsy, and indicated protospacer plasmids or control vector are shown. The leader-proximal end of the CRISPR array was amplified using a primer pair (schematically shown); amplification products were separated by agarose gel electrophoresis and visualized by ethidium bromide staining.

3.2. Genetic requirements for spacer acquisition by the I-F CRISPR–Cas system from *P. aeruginosa*

In fact, pSPA, pSPAmut and pACYC plasmids were not necessary for adaptation, since spacer acquisition was detectable in cells containing only pCas and pCsy plasmids (Figure 14, lane 1). However, no new spacer acquisition occurred in cells that contained individual pCas or pCsy (Figure 14, lanes 2 and 3).

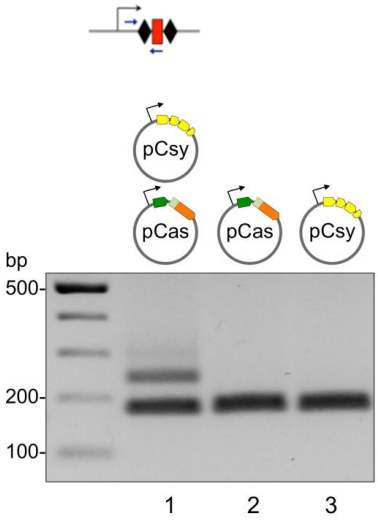


Figure 14. Spacer acquisition by the *P. aeruginosa* CRISPR-Cas system in a heterologous background in non-primed conditions.

The experiment was performed as in Figure 13 in the absence of a protospacer plasmid and in the presence of only pCas, or only pCsy, or both plasmids, as indicated.

The effects of mutations in *P. aeruginosa cas* and *csy* genes on spacer acquisition were established next. Introducing an alanine at position 268 of Cas1 (Cas1D268A) abolished adaptation (Figure 15, lane 5). This is an expected result since D268 is a conserved metal coordinating residue and substitution of the corresponding residue in *E. coli* Cas1 also abolishes adaptation (Arslan *et al*., 2014). A specific feature of subype I-F systems is a fusion of *cas2* and *cas3* homologs, which are encoded on separate genes in other CRISPR–Cas systems (Makarova *et al*., 2011a). A D124Oc mutation in the *cas2-cas3* gene, which

introduced an ochre stop codon instead of aspartate codon at position 124, after the *cas2* portion of the fused gene, abolished spacer acquisition (Figure 15, lane 2). Point mutations introducing single amino acid substitutions in the endonuclease (D124A) and helicase (D576N) domains of the Cas3 protein also prevented spacer acquisition (Figure 15, lanes 3 and 4). Deleting the *csy3* gene (Figure 15, lane 6), or mutating the catalytic residue His29 of Csy4 nuclease (Figure 15, lane 7), which is needed for generation of mature crRNA, also abolished spacer acquisition (Haurwitz *et al*., 2010).

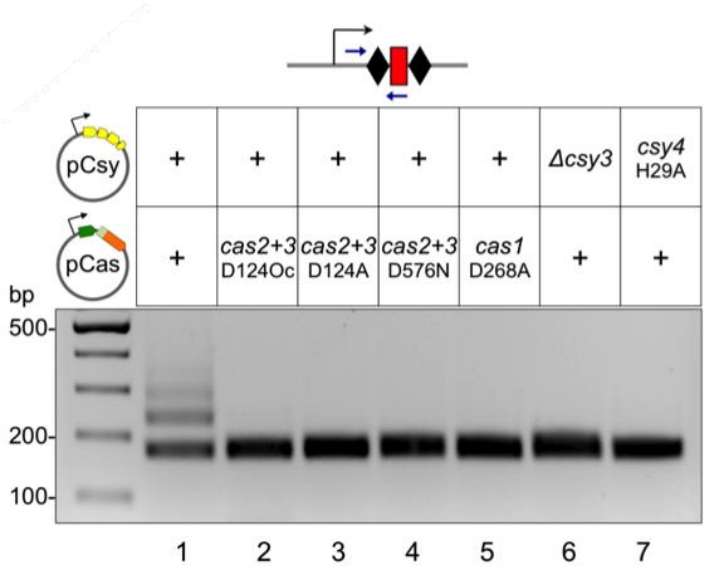


Figure 15. Active site mutations or gene deletions disrupt new spacer acquisition by the subtype I-F CRISPR-Cas system in non-primed conditions.

Results of a CRISPR adaptation experiment with KD604 cells transformed with pCas and pCsy expressing wild-type *cas/csy* genes (lane 1) or various indicated mutant versions of these genes.

We therefore concluded that spacer acquisition by the transplanted *P. aeruginosa* CRISPR–Cas system requires Cas1, Cas2, and the Csy proteins.

To determine whether adaptation depends on specific crRNA, an additional *E. coli* strain (KD606) containing a different spacer in the engineered *P. aeruginosa* CRISPR locus was used. Neither KD604 nor KD606 spacers have detectable similarity to pCas or pCsy sequences or to the *E. coli* genome (at least 13 mismatches, longest stretch of

complementarity 8 nucleotides). Just as in the case of the KD604 strain, robust spacer acquisition was observed when both pCas and pCsy were introduced into *E. coli* (KD606), but no adaptation was detected when only one plasmid was present.

When *E. coli* (KD628) cells harboring only a leader and a single CRISPR repeat were transformed with pCas and pCsy plasmids, very weak adaptation (compared to adaptation observed in cells containing two repeats and one spacer) was observed (Figure 16), indicating that removal of a spacer or one of the two repeats, both of which should affect crRNA production, inhibits adaptation. The residual adaptation was abolished when combined with the H29A mutation in *csy4*. Thus, the *P. aeruginosa* CRISPR–Cas system is capable of robust adaptation in the apparent absence of preexisting matches between the crRNA spacer and the target. *In E. coli*, such naïve adaptation requires only the Cas1 and Cas2 proteins (Yosef *et al*., 2012). However, in the *P. aeruginosa* CRISPR–Cas system, proteins that constitute the Csy effector complex are also required. The presence of crRNA strongly stimulates spacer acquisition.

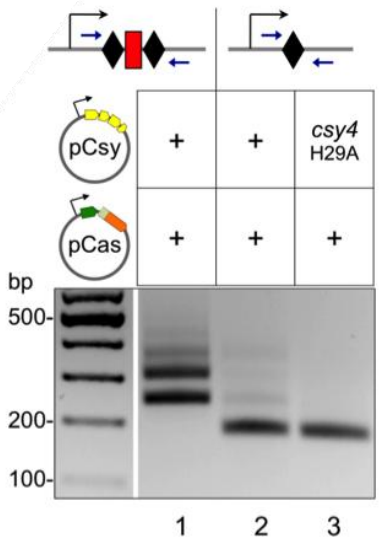


Figure 16. crRNA plays an important role in adaptation in non-primed conditions.

Results of a CRISPR adaptation experiment with KD604 cells transformed with pCas and pCsy expressing wild-type *cas*/*csy* genes (lane 1) or with KD628 cells harboring a single CRISPR repeat transformed with wild-type pCas and pCsy or indicated mutant plasmid combination.

3.3. Anti-CRISPR proteins prevent spacer acquisition by *P. aeruginosa* I-F CRISPR–Cas system

Anti-CRISPR proteins target different components of the Csy complex or the Cas3 protein (Bondy-Denomy *et al*., 2015). To determine if anti-CRISPR proteins also affect spacer acquisition, *E. coli* (KD604) cells harboring pCas and pCsy plasmids were transformed with compatible plasmids expressing distinct anti-CRISPR proteins and spacer acquisition was monitored (17). Spacer acquisition was inhibited in the presence of each of the different anti-CRISPR proteins tested (acrF1–5, Figure 17). A plasmid bearing a frameshift mutation in the beginning of the *acrF1* anti-CRISPR gene that inactivates its anti-interference function (Bondy-Denomy *et al*., 2015) also inactivated the anti-acquisition function.

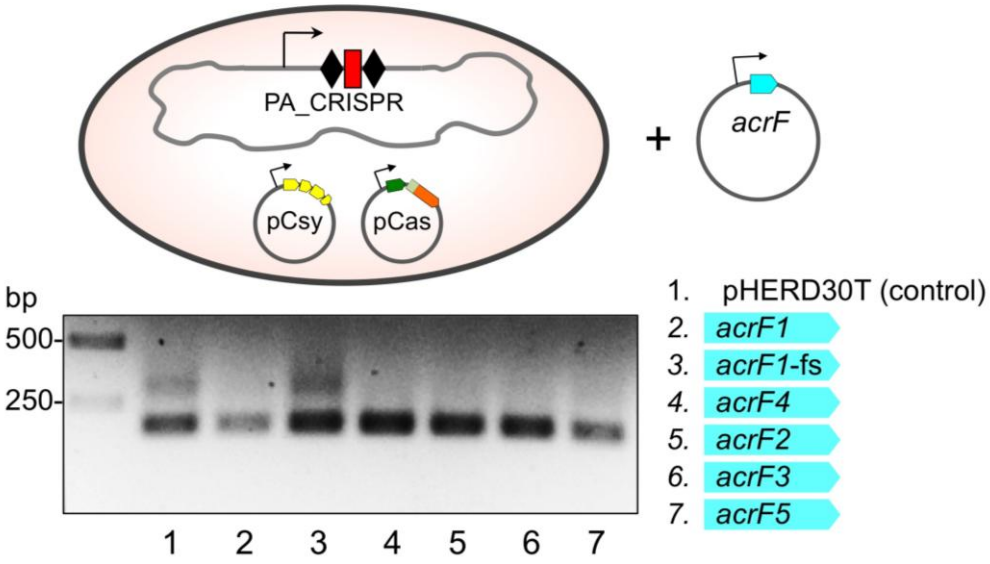


Figure 17. Diverse anti-CRISPR proteins encoded by *P. aeruginosa* bacteriophages inhibit spacer acquisition.

KD604 cells carrying pCas and pCsy expressing wild-type *cas/csy* genes were transformed with compatible pHERD30T-based plasmids expressing various anti-CRISPR proteins (AcrFs) listed on the right (*acrF1*-fs carries a frame-shift mutation). Expression of plasmid-borne gene was induced and spacer acquisition was monitored by PCR amplification.

We conclude that anti-CRISPR proteins that inhibit CRISPR interference in *P. aeruginosa* by targeting either the Csy complex or Cas3, also inhibit spacer acquisition. This supports the results of our genetic analysis that demonstrate the requirement of these proteins for adaptation in the subtype I-F CRISPR–Cas system.

3.4. The origin and distribution of spacers acquired by the I-F *P. aeruginosa* CRISPR–Cas system

To determine the origin of spacers acquired by *E. coli* (KD604) cells containing pCas and pCsy (with or without pSPA, pSPAmut or pACYC), PCR fragments corresponding to expanded CRISPR arrays were subjected to high throughput sequencing. Since most spacers mapped to plasmids rather than to the KD604 genome (<1% of all spacers), spacers from the genome were excluded from further analysis.

In cells harboring only pCas and pCsy, the ratio of unique spacers originating from pCas and pCsy was ~4 to 1 (Figure 18). The two plasmids are of similar size and have the same origin of replication. The nature of this bias, which is highly reproducible, is unknown and requires further investigation. In the presence of either pSPA or pACYC, <1% of all spacers were acquired from these additional plasmids (Figure 18). By contrast, in cells harboring pSPAmut ~30% of acquired spacers were derived from this plasmid (Figure 18). Thus, an imperfect spacer match with the target strongly stimulates and allows spacer selection from DNA sequences in *cis*, which is a hallmark of primed adaptation. Increased efficiency of spacer selection from pSPAmut did not affect the ratio of spacers acquired from pCas and pCsy (Figure 18).

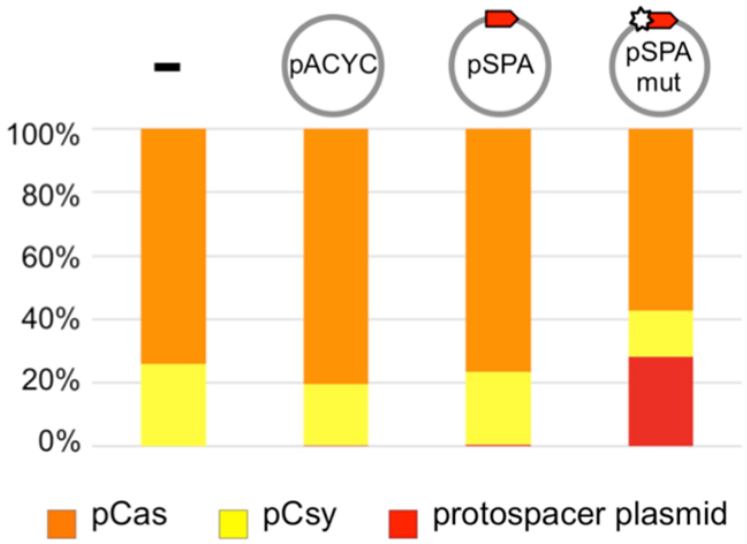


Figure 18. The origin of spacers acquired by the *P. aeruginosa* CRISPR-Cas system.

Bar graph showing the origin of plasmid-derived spacers. Spacers originating from pCas are shown in orange, those originating from pCsy are in yellow, and those from pSPA, pSPAmut, or pACYC are in red.

Next the distribution of donor protospacers and efficiency of their use were investigated (determined by the number of corresponding spacer reads) for each plasmid (Figure 19 and Figure 20). Both the distribution and the efficiency of use of donor protospacers in pCas and pCsy were found to be highly reproducible and independent of the presence of pSPA, pSPAmut or pACYC (Figure 19, correlation coefficients for spacer distributions between different samples or between biological replicates of the same sample are 0.82 or higher). Spacers derived from pCas originated from both strands and from every part of the plasmid. The overall efficiency of spacer selection was considerably higher for the *ori* region and the adjacent *rop* gene (Figure 19). 27% of all spacers originated from *ori*. A strong strand bias for selection of these spacers was observed, with 95.5% of *ori* spacers being selected from the 'inner' strand of the plasmid, as shown in Figure 19. The remaining 73% of spacers were selected with equal efficiency from both strands of the plasmid (53% from the 'inner' strand). A similar pattern was observed for pCsy: equal efficiency of spacer acquisition from both

strands throughout the plasmid backbone with strong strand bias in the *ori*. It should be noted that the *ori* sequences of pCas and pCsy are highly similar, so most spacers derived from these regions are not unique. The same distribution of acquired spacers from pCas and pCsy was observed in KD606 cells (a correlation coefficient with KD604 of 0.85 for pCas and 0.89 for pCsy) indicating that spacer acquisition preferences do not depend on the spacer of pre-existing crRNA.

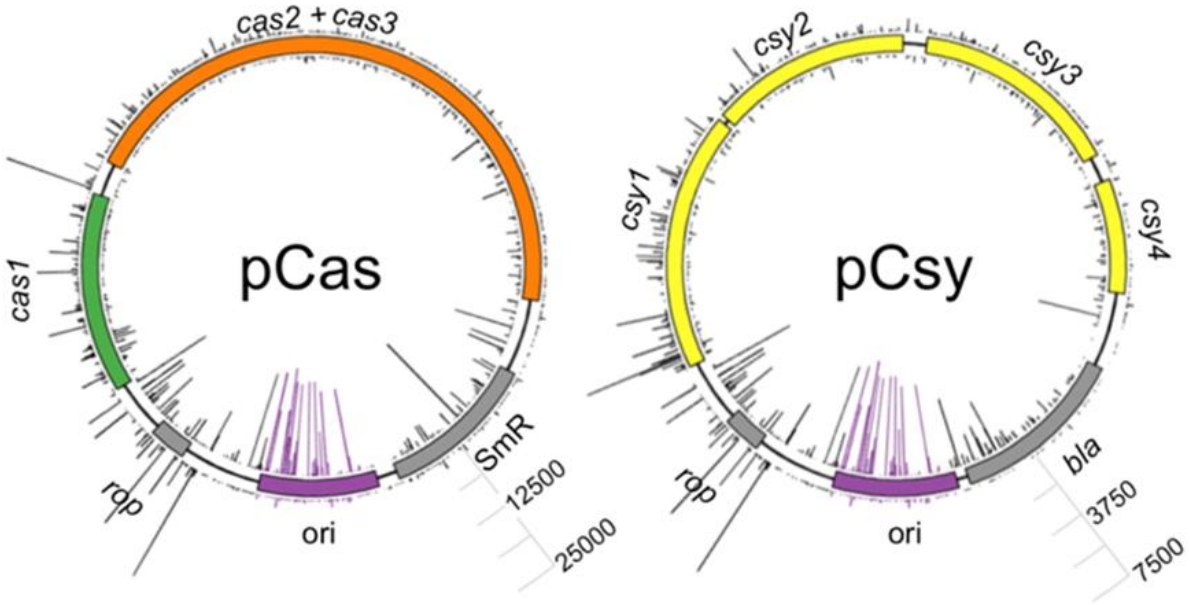


Figure 19. The distribution of spacers acquired by the *P. aeruginosa* CRISPR-Cas system from plasmids coding *cas* and *csy* genes.

Mapping of spacers acquired by cells containing pCas and pCsy on donor plasmids. The *cas* genes are shown in green and orange, *csy* genes in yellow, antibiotic resistance and repressor of primer (*rop*) genes in grey, and replication origins in purple. The height of the grey and purple bars indicates efficiency (the number of times) of spacer acquisition from this position. Bars protruding inside and outside of plasmid circles represent spacers derived from different strands of DNA. The height of bars corresponding to the most frequently acquired protospacers in both plasmids is made the same for easier comparisons. The scale bars make it possible to access spacer acquisition efficiencies for each plasmid (number of reads). Purple bars indicate spacers originating from *ori* regions. Grey bars indicate spacers originating from the rest of each plasmid.

This, and the absence of strand bias for spacers acquired from most of the pCas and pCsy sequence, is consistent with naïve, non-primed adaptation. The only exception is the *ori* region, where strong strand bias and increased overall level of spacer selection efficiency is observed that is not typical for naïve adaptation.

The pattern of spacer selection from pACYC, pSPA, and pSPAmut is shown in Figure 20. Spacers from pACYC were acquired from both strands of the plasmid, with the most actively used protospacers located at or close to the *ori* region. The distribution was similar in the case of the pSPA plasmid, with a modest bias for selection of protospacers in the vicinity of the protospacer matching KD604 crRNA. In the case of pSPAmut most spacers originated from one strand of the plasmid and there was a strong preference for selection of protospacers downstream from the protospacer matching the KD604 spacer (Figure 20).

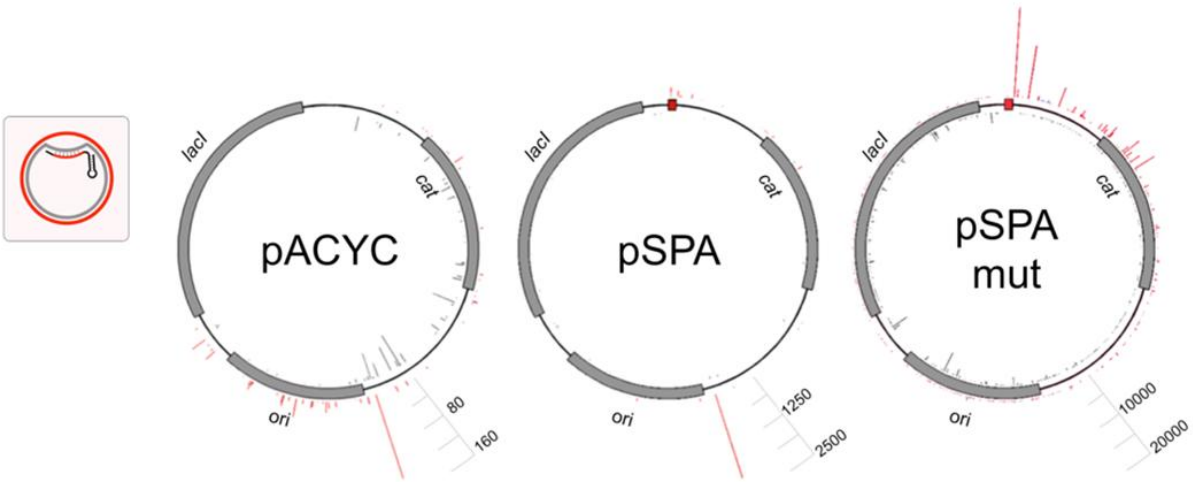


Figure 20. The distribution of spacers acquired by the *P. aeruginosa* CRISPR-Cas system from plasmids, carrying the protospacer.

Mapping of spacers acquired from pSPA, pSPAmut, or pACYC vector control in cells expressing the *cas* and *csy* genes. Where present, a protospacer matching crRNA is shown as a small red box. The inset schematically shows the structure of the R-loop formed by crRNA. The target strand is shown in grey and non-target in red. Bars showing spacers originating from each of these strands are colored accordingly. Scale bars indicate spacer acquisition efficiency for each plasmid (number of reads).

A strong (96%) bias for spacer acquisition from the non-target strand in this region was observed (Figure 20). Protospacers located closer to the priming spacer appeared to be used more efficiently than those located further away, revealing a gradient in protospacer selection efficiency as a function of distance from the priming site. Protospacers in the area located upstream from the priming spacer were used much less efficiently. Spacers from this area were more efficiently selected from the strand targeted by the KD604 crRNA spacer (97% bias).

97.76% of more than 4000 unique plasmid-derived spacers matched protospacers containing consensus GG PAM. This value was the same for pCas and pCsy-derived spacers, which were acquired in the absence of priming, and for pSPAmut spacers acquired in the course of primed adaptation (Table 2). Most spacers that appeared to originate from protospacers with non-consensus PAMs were 'derived' from regions containing a consensus PAM by 1–2 nucleotide upstream or downstream shifts (we called them "shifters") or by insertion into a CRISPR array in the opposite orientation (we called them "flippers").

	pCas	pCsy	pACYC	pSPA	pSPAmut
Plasmid-derived spacers	474318	166866	1018	3756	116066
Strand bias, %	35.56	38.67	51.57	90.07	66.06
GG PAM frequency, %	97.58	97.90	94.30	99.41	95.46
Derived spacers (shifters), %	1.20	1.10	1.77	0.24	3.46
Derived spacers (flippers), %	0.63	0.46	3.54	0.16	1.00
Strand bias corrected, %	35.48	38.67	48.62	90.23	66.23
GG PAM frequency corrected, %	99.41	99.46	99.61	99.81	99.92

Table 2. Statistics for reads corresponding to spacers derived from various plasmids in the course of adaptation by the *P. aeruginosa* CRISPR-Cas system.

Similar aberrant spacers were previously observed in the *E. coli* subtype I-E system (Shmakov *et al*., 2014). When such 'derived' spacers were considered as originating from

parental protospacers with consensus PAM, the preference for GG PAM increased to more than 99% for both pCas and pCsy and for pSPAmut. In the *E. coli* subtype I-E system, a strong bias toward protospacers with consensus, interference-proficient PAMs is indicative of primed adaptation (Datsenko *et al*., 2012).

3.5. Spacer acquisition during targeting of the *E. coli* genome

To obtain a more detailed view of donor protospacer selection upstream and downstream of the priming site an *E. coli* KD675 strain, a derivative of KD604 was used. KD675 contains a partially matching protospacer (the same as in the pSPAmut plasmid, above) with consensus PAM in its genome (Figure 21A). Primed adaption was expected to be initiated from the genomic protospacer under conditions of *cas* and *csy* genes expression. KD675 cells were transformed with pCas and pCsy plasmids, induced, and grown in the presence of antibiotics to reduce a fraction of the cells with plasmid-derived spacers in the culture. Despite self-targeting, induced KD675 cultures continued to grow normally, consistent with low levels of interference observed during plasmid transformation. As a result, while only a modest level of spacer acquisition was observed (Figure 21B), analysis of the high-throughput sequence showed that 99.7% of all reads carried just one new spacer. 30% of spacers originated from pCas and pCsy, and the remainder were from the KD675 genome. Spacers acquired from pCas and pCsy had a distribution similar to that shown in Figure 19 (correlation coefficients of 0.85 or more). Spacers from the KD675 genome originated from either side of the priming site. The efficiency of spacer selection decreased as the distance from the priming spacer increased, but remained above background levels at distances as large as 5000 base pairs (bp) (Figure 21C).

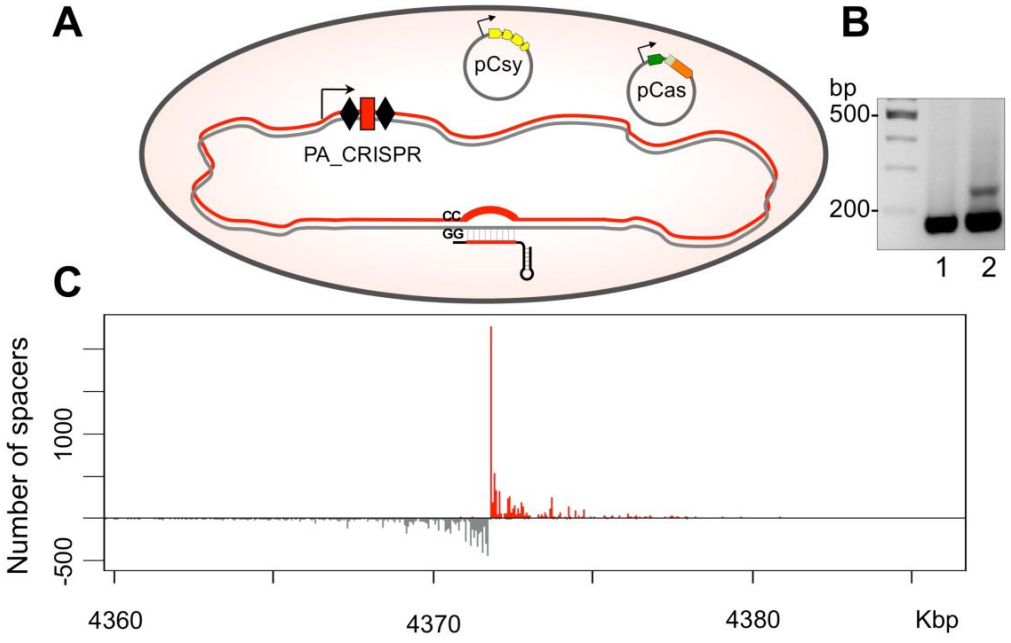


Figure 21. New spacers are acquired from regions of the genome that are distributed around the priming site.

A. An *E. coli* KD675 cell transformed with pCas and pCsy plasmids is schematically shown. KD675 does not have a CRISPR-Cas system of its own but carries a CRISPR array containing two *P. aeruginosa* subtype I-F repeats, a single spacer, and an upstream leader sequence with an upstream promoter. The spacer targets a protospacer in KD675's own genome. The structure of the R-loop formed upon recognition of the protospacer by self-targeting crRNA is shown, with the target strand shown in grey and the non-target in red. **B**. Results of a CRISPR adaptation experiment with KD675 cells transformed with pCas and pCsy plasmids. No spacer acquisition was observed in the absence of inducers (lane 1), the cells acquired new spacers after induction of *cas* and *csy* genes expression (lane 2). **C.** Mapping of acquired spacers on a region of KD675 genome within ~25 kbp of the priming site. Spacers acquired from the non-target strand are shown in red. Spacers acquired from the target strand are shown in grey.

In the downstream direction, in agreement with data obtained with the pSPAmut plasmid, spacers were selected predominantly from the non-target strand where the protospacer matching the crRNA spacer is located. In the upstream direction, spacers were selected from the opposite, target strand. Though on each side of the priming site the strand bias was very strong, when all spacers were considered together there was no overall strand bias since

spacers upstream and downstream from the priming site are selected from different strands (Table 3).

KD675 genome-derived spacers	37616
Strand bias, %	50.57
GG PAM frequency, %	91.44
Derived spacers (shifters), %	5.01
Derived spacers (flippers), %	0.93
Strand bias corrected, %	50.12
GG PAM frequency corrected, %	97.38

Table 3. Statistics for reads corresponding to spacers derived from the genome of *E. coli* KD675 cells carrying a self-targeting spacer.

3.6. Spacer acquisition by subtype I-F CRISPR–Cas system of *E. coli*

While most strains of *E. coli* contain a subtype I-E CRISPR–Cas system, the ED1a strain of *E. coli* contains a subtype I-F system (Figure 22) (Touchon *et al*., 2010; Almendros *et al*., 2012).

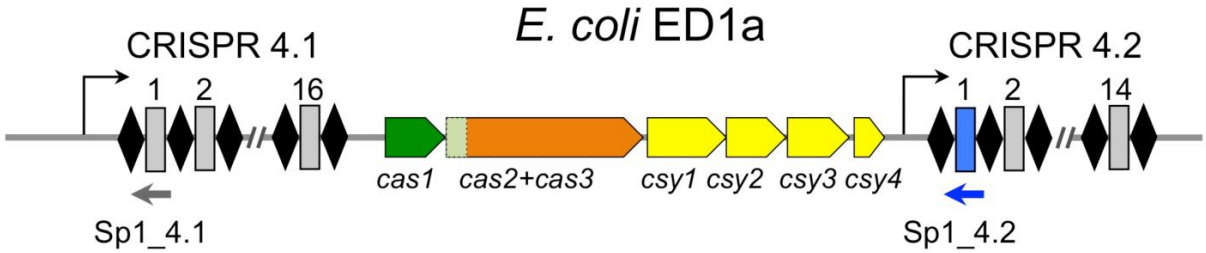


Figure 22. Organization of the I-F subtype CRISPR-Cas system from the *E. coli* ED1a strain.

Two I-F subtype CRISPR arrays (4.1 and 4.2) and all gene of I-F subtype system are present in the genome of *E. coli* ED1a.

To monitor the function of the ED1a CRISPR–Cas system, the cells were transformed with plasmids containing protospacers complementary to the first spacer in the endogenous CRISPR 4.2 array. No difference in transformation efficiency was detected for a pSED plasmid containing a protospacer with a consensus GG PAM as compared to a pSEDmut plasmid containing a protospacer with a mismatch in the seed region or pT7blue vector control. However, upon prolonged cultivation in the absence of antibiotics the pSED plasmid was lost from the culture, while pSEDmut and the pT7blue vector were retained. When the transformed cultures were cultivated and analyzed for CRISPR array expansion, PCR fragments corresponding to expanded CRISPR arrays were observed in cultures of cells containing both pSED and pSEDmut, but not in cultures harboring pT7blue (Figure 23A).

Therefore, the ED1a cells apparently undergo primed adaption, but it proceeds with the same efficiency from targets with completely or partially complementary protospacers. High-throughput sequencing of acquired spacers was performed and spacer sequences were mapped. An identical result was obtained when spacers acquired in cultures harboring either pSED or pSEDmut were analyzed. In both cases, 98% of spacers were plasmid-derived. The remaining spacers originated from the bacterial genome. The distribution of donor protospacers and the efficiency of spacer selection from the pSEDmut plasmid is shown in Figure 23B. As can be seen from Figure 23B, a gradient of spacer acquisition was also revealed in this case with strand biases upstream and downstream from the priming site matching those observed for the subtype I-F *P. aeruginosa* system.

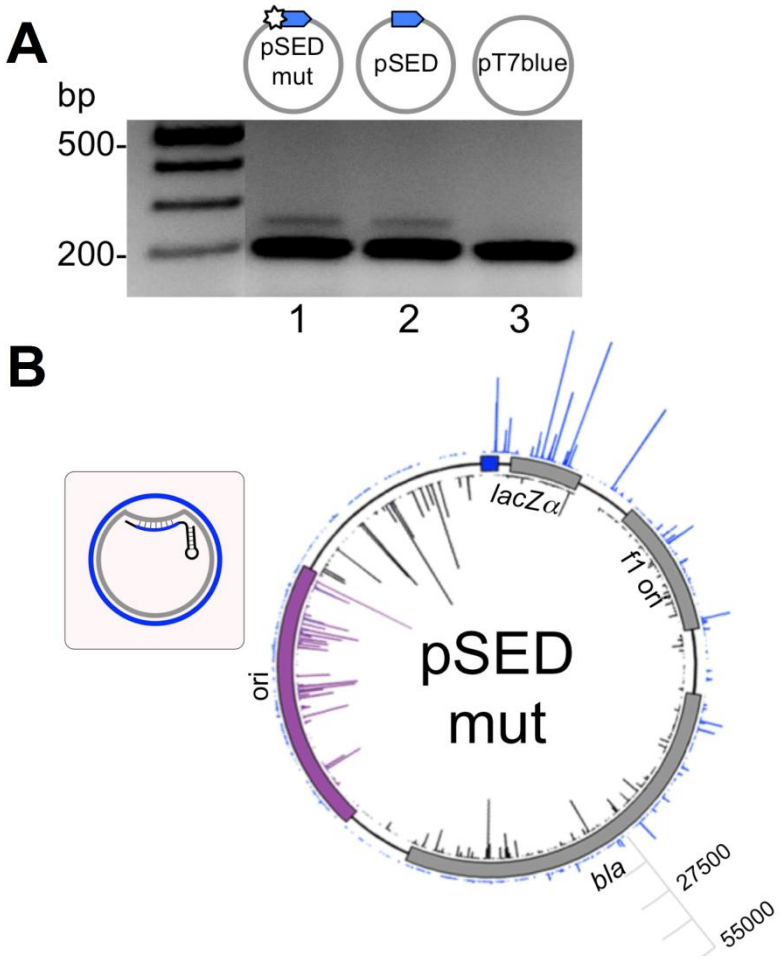


Figure 23. Spacer acquisition by the I-F subtype CRISPR-Cas system from the *E.coli* ED1a strain.

A. *E. coli* ED1a cells transformed with plasmid pSED (carrying a fully matching protospacer and a GG PAM), plasmid pSEDmut (containing single spacer-protospacer mismatch at position +1), or the pT7blue vector control (no protospacer). PCR products from the leader end of CRISPR locus 4.2 were separated by electrophoresis and visualized by ethidium bromide staining. B. Mapping of spacers acquired by *E. coli* ED1a cells transformed with pSPAmut. The priming protospacer is shown as a blue box. The inset schematically shows the structure of the R-loop formed by crRNA. The target strand is shown in grey and the non-target in blue. Bars showing spacers originating from each of these strands are colored accordingly. Bars showing spacers originating from the *ori* are colored purple. The scale bar indicates spacer acquisition efficiency for each plasmid (number of reads).

4. Discussion

The principal finding of this part of the work is that both naïve adaptation and primed adaptation are operational in the *P. aeruginosa* subtype I-F system. However, in marked contrast to the situation in the subtype I-E system in *E. coli*, the *cas1* and *cas2* genes are not sufficient for naïve adaptation. In fact, in the subtype I-F system from *P. aeruginosa*, both modes of adaptation require an intact Csy crRNP complex with a crRNA. By definition, primed adaptation in *E. coli* also requires a crRNA that contains mismatches with a protospacer in the target DNA (Datsenko *et al*., 2012). These mismatches have been shown to reduce, but not abolish binding by the Cascade. Residual interaction with the mutated protospacer recruits Cas1 and Cas2 to target DNA, resulting in primed adaptation. A very similar process must be occurring during primed adaptation by the *P. aeruginosa* I-F system, when crRNA with a spacer partially matching a protospacer guides the Csy complex (a functional ortholog of Cascade), and Cas2–3 and Cas1 to the target, leading to efficient spacer acquisition (Richter *et al*., 2014). Consistent with previous results, our analysis reveals that spacers are acquired from both sides of the priming site. Apparently, the adaptation machinery proceeds bi-directionally from the priming site with efficiency of spacer selection decreasing as the distance from the priming site increases. Experiments with an *E. coli* strain containing a spacer complementary to the *E. coli* chromosome suggest that the *P. aeruginosa* I-F system adaptation machinery is highly processive and able to acquire spacers from protospacers located several thousands of base pairs away from the priming site. Strand bias is a hallmark of primed adaptation in *E. coli*: spacers from the strand opposite to the one targeted by the Cascade are selected about ten times more frequently than from the target strand (Datsenko *et al*., 2012; Savitskaya *et al*., 2013). Overall, primed adaptation by the *P. aeruginosa* I-F system does not show such bias. However, when spacers selected from both sides of the priming site are considered separately, a very strong strand bias is observed. Upstream from

the priming site, spacers are selected from the strand targeted by the priming crRNA. In the downstream direction the strand bias is reversed and spacers are predominantly selected from non-target strand protospacers. A similar strand bias was earlier detected for spacers acquired during primed adaptation by the I-F subtype system in *P. atrosepticum* (Richter *et al*., 2014; Staals *et al*., 2016) and is also found during primed adaptation by the *E. coli* I-F subtype system studied in the present work and in recently published work of Staals *et al*., where adaptation by the I-F subtype system in *P. atrosepticum* was studied (Staals *et al*., 2016). A strong strand bias that changes its direction at the priming site is most consistent with the existence of single-stranded intermediates formed at both sides (and on different strands) of the priming site. New spacers could be selected from these intermediates upon the recognition of PAMs by the adaptation machinery. If such intermediates were double stranded, PAM sequences in both strands would have been recognized, abolishing the strand bias. This hypothesis is supported by data obtained by Musharova *et al.* in a study of I-E subtype adaptation (Musharova *et al.*, 2017)

It was suggested that in the case of the subtype I-B system from *Haloarcula hispanica*, where adaptation strictly requires priming, the Cas3 protein activity is responsible for generation of single-stranded intermediates for spacer selection (Li *et al*., 2014). Our data are consistent with this model. However, the direction of strand bias in *H. hispanica* is opposite to the one observed in I-F systems. One can speculate that this difference might be caused by opposing directionality of Cas3 helicase/nuclease action in these systems.

Also only primed adaptation was observed in the I-C subtype system of *Legionella pneumophila* (Rao *et al*., 2016). Most of the acquired spacers target the same strand as the priming spacer, but only 23 new spacers were analyzed so it is hard to make strong conclusions.

The observation of primed adaptation in transplanted *P. aeruginosa* I-F system was capable only when CRISPR interference was inactivated by a mutation that introduced a mismatch between the crRNA spacer and the priming protospacer. Interestingly, primed adaptation by the *E. coli* I-F system was observed with equal efficiency, whether there was a mismatch or full match between the crRNA spacer and the priming site. However, it should be noted that CRISPR interference by the *E. coli* I-F system was very weak, even in the conditions of a full match between the crRNA spacer and the target protospacer. It is possible that the outcome of crRNP complex interaction with a protospacer, i.e. interference or primed adaptation, may be determined not so much by the presence of mismatches between the spacer and protospacer but by the overall stability and/or life time of the complex, which could be affected by the sequence of the spacer-protospacer heteroduplex.

The *P. aeruginosa* subtype I-F CRISPR–Cas system is capable of robust adaptation in the absence of pre-existing matches between a crRNA and the target. In the *E. coli* subtype I-E system, naïve adaptation requires just the Cas1 and Cas2 proteins (Yosef *et al*., 2012). However, in the case of the *P. aeruginosa* CRISPR–Cas system, all Cas and Csy proteins are required for this process. Naïve adaptation is also strongly stimulated by the crRNA. However, the sequence of crRNA spacers appears to be unimportant, since the KD604 and KD606 strains contain different crRNAs and yet exhibit the same spacer acquisition preferences. This suggests that the role of crRNA may be limited to stabilization of the Csy complex and that the Csy complex is involved in protospacer selection in a process that is independent of crRNA-guided base paring to the target.

The observation that phage-encoded anti-CRISPR proteins inhibit both interference and acquisition supports the genetic data presented here and provides further mechanistic links between these two processes. It has recently been shown that distinct anti-CRISPR proteins operate through diverse mechanisms, by preventing DNA-binding by the Csy complex or

blocking recruitment of the Cas2–3 protein (Bondy-Denomy *et al*., 2015). The ability of anti-CRISPR proteins that inhibit CRISPR interference in *P. aeruginosa* by targeting either the Csy complex or Cas3, to also inhibit spacer acquisition supports the results of genetic analysis, which showed that Csy and Cas3 proteins are required for adaptation in the subtype I-F CRISPR–Cas system.

In *E. coli* (subtype I-E), overexpression of Cas1 and Cas2 results in more than 60% of spacers derived from protospacers with non-consensus PAM (Yosef *et al*., 2012; Shmakov *et al*., 2014). Spacers derived from protospacers not flanked by a PAM are unable to elicit an interference response, but could induce primed adaptation (Datsenko *et al*., 2012, Fineran *et al*., 2014). By contrast, both primed and naïve adaptation by the *P. aeruginosa* system result in almost absolute selectivity for protospacers with consensus PAM, therefore leading to crRNAs that are capable of interference. Presumably, the higher level of selectivity towards consensus PAM is due to involvement of the Csy complex in PAM recognition (Richter *et al*., 2014). Interestingly, recent analysis of spacer acquisition by Type II CRISPR–Cas systems showed that, in addition to the 'professional' adaptation proteins, Cas1 and Cas2, the Cas9 protein, which is an crRNA-guided endonuclease involved in target DNA cleavage, is also required for adaptation and determines selection of protospacers with correct PAMs (Wei *et al*., 2015, Heler *et al*., 2015). Involvement of PAM-recognizing proteins from the interference pathway in selection of donor protospacers during CRISPR adaptation appears to be a common strategy, which ensures that newly acquired spacers will result in crRNAs capable of eliciting direct degradation of the foreign target.

Naïve and primed adaptation by the I-F subtype system was recently studied in the bacteria *P. atrosepticum (*Staals *et al.,* 2016). That work proposed an experiment-based model of primed adaptation, according to which the first spacer in primed adaptation is acquired in most cases from the target strand and the orientation of the new spacer is opposite to the primed spacer.

However, that model is unable to describe our results obtained with the *P. aeruginosa* system. Analysis of spacer acquisition during targeting of the *E. coli* genome in priming condition showed that the first acquired spacers were derived from both strands with the same frequency.

A structural model of $Cas1_4$-Cas2-3_2 complex from *P. atrosepticum* with bound protospacer substrate DNA was described in Fagerlund *et al*. study (Fagerlund *et al*., 2017). It was shown *in vitro* that Cas1-Cas2-3 complex is able to capture and integrate spacers into CRISPR array and enzymatic activity of Cas3 is not required. This result is differing from our result. So we propose that the action of Cas3 domain and Csy complex are essential for formation of spacers' precursors and are essential at the earliest stage of adaptation.

The *P. aeruginosa* subtype I-F CRISPR–Cas system clearly prefers to acquire spacers from the ColEI origin of replication during non-primed spacer adaptation. Spacer acquisition from ori is highly biased to one strand, suggesting that it is driven by a specific structural feature of the origin itself. An extended RNA-DNA duplex and an R-loop are formed on ColEI when RNA polymerase transcribes RNA II, a transcript used to prime plasmid replication. A second RNA, RNA I, is transcribed in the opposite direction and is used to control the number of RNA II molecules (Kornberg and Baker, 2005). The structure of the ColEI origin is schematically shown in Figure 24, along with the distribution of GG PAMs and spacers acquired from this region of the pCas plasmid. Very similar distributions were observed in experiments involving KD604 and KD606 cells that contain unrelated crRNA spacers. In fact, the distribution of spacers acquired from the ColE1 origin of pSED plasmids by the *E. coli* ED1a subtype I-F CRISPR–Cas system is also very similar (the correlation coefficient is 0.8). Thus, spacers are acquired from the strand complementary to RNA II. The mechanistic reasons for this bias have yet to be clarified.

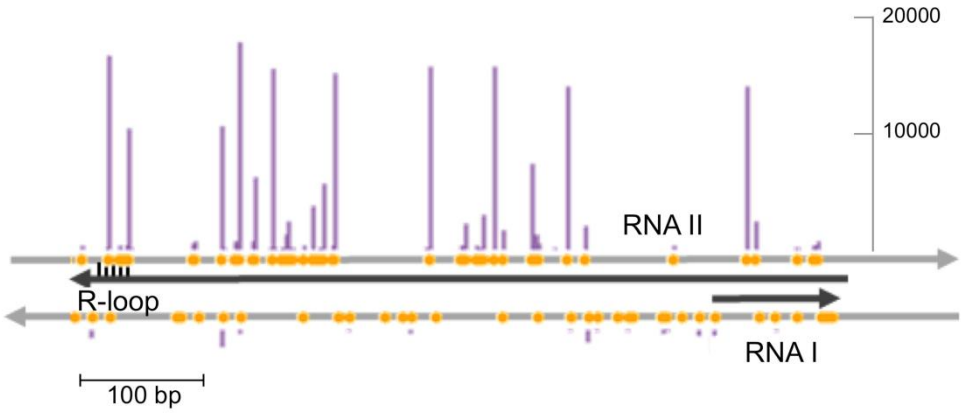


Figure 24. The distribution of spacers acquired by the *P. aeruginosa* CRISPR-Cas system at the ColE1 origin of replication.

The two strands to the ColE1 origin are shown as grey arrows. Black arrows show origin transcripts. The position of the R-loop at the 3'-terminal part of RNA II is indicated. Yellow circles indicate GG sequences (PAMs) in both strands. The purple vertical bars show the number of spacers acquired. The distribution of acquired spacers for the pCas plasmid from KD604 (see Figure 19) is shown. Very similar distributions were obtained with KD606 cells and with ED1a *E. coli* that acquired spacers from the ColE1 plasmid pSED (Figure 23B). The scale bar indicates spacer acquisition efficiency (number of reads).

The *E. coli* subtype I-E Cas3 protein is involved in copy number control of ColE1 plasmids (Ivančić-Baće *et al*., 2013), while naïve spacer acquisition by this system is targeted to the genomic replication termination region (Levy *et al*., 2015), which, in case of unidirectional ColE1 replication, coincides with *ori*. These observations suggest that there may be a deep link between naïve CRISPR adaptation and replication, which is not yet understood (Levy *et al*., 2015). Such a link could initially target CRISPR adaptation machinery to actively replicating foreign DNA, while priming would allow additional protective spacers to be specifically acquired at a later point.

Chapter 3: Study of III-A and III-B subtype CRISPR-Cas system

1. Introduction

To date, only one paper describing data on adaptation in Type III CRISPR-Cas systems in the laboratory conditions has been published (Silas *et al.*, 2016). Some of CRISPR-Cas systems carry *cas1* gene fused with reverse transcriptase domain (RT). The vast majority of these systems belongs to Type III (Toro *et al.*, 2014; Silas *et al.*, 2017). Naïve adaptation was shown for a III-B subtype system carrying such *RT-cas1* gene. However, adaptation was detected only under conditions of overexpression of *RT-cas1*, *cas2* and another adjacent gene but not in the native strain. It is interesting that spacer acquisition was observed not only in the presence of wild-type RT-Cas1 but also in the presence of RT-Cas1 with mutations at its RT active site (YADD → YAAA) or complete deletion of RT domain. It was shown that the presence of wild-type RT-Cas1 protein allows insertion of new spacers directly from RNA while in the presence of mutated RT-Cas1 spacers are acquired from DNA (Silas *et al.*, 2016).

Bacterium *T. thermophilus* Hb27 contains genes coding for proteins of III-A, III-B, I-C and another one incomplete subtype CRISPR-Cas systems and ten CRISPR arrays (Swarts *et al*., 2015). *T. thermophilus* Hb27 is a widely used laboratory strain. It carries a chromosome (1 894 877 bp) and pTT27 megaplasmid (232 605 bp) (Henne *et al*., 2004). Some genomic manipulations, such as gene replacements, could be easily done with this strain. The high growth rate is also typical for this extremely thermophilic bacterium (Cava *et al*., 2009; Carr *et al*., 2015). Thus, it can be a good model for investigation of adaptation in Type III systems.

For I-E subtype systems, primed adaptation is much more efficient than naïve adaptation (Savitskaya *et al*., 2013). Primed adaptation is caused by mutations in the PAM or seed region of the target protospacer in systems of Type I (Datsenko *et al*., 2012). It is believed that there is no PAM in Type III systems (Marraffini *et al*., 2010). The seed region of protospacers was also not defined (Staals *et al*., 2014; Peng *et al*., 2015). We considered definition of the seed

location as a pre-requisite for investigating adaptation in primed conditions by Type III systems. The subtype III-A Csm RNP complex of *T. thermophilus* Hb8 demonstrated periodic endoribonuclease activity *in vitro,* cleaving target RNA between 5th and 6th positions starting from 3'-end and further downstream with 6-nucleotide periodicity. It was also demonstrated that the introduction of mutations in positions 1-5, or 7-11, or subsequent five-nucleotide mutations abolishes cleavage at sites nearest to mutations (Staals *et al*., 2014). Also, single mismatches in positions 1 to 7 were tested. In contrast with Type I systems (Semenova *et al*., 2011), only a mismatch in the 5th position inhibited cleavage at the downstream (between 5th and 6th positions) site (Staals *et al*., 2014). Peng *et al*. have demonstrated *in vivo* that in the III-B system of archaeon *Sulfolobus islandicus* three-nucleotide mismatches (positions 28-30) between protospacer and the 3'-region of crRNA abolished interference during cell transformation with protospacer containing plasmids (Peng *et al*., 2015).

2. Materials and methods

2.1. Bacterial strains, bacteriophages, and plasmids

T. thermophilus Hb27 was used as a model for subtype III-A and III-B studies. *E. coli* DH5α (F^- Φ80*lacZ*ΔM15 Δ(*lac*ZYA-*arg*F) U169 *rec*A1 *end*A1 *hsd*R17(r_k^-, m_k^+) *pho*A *sup*E44 *thi*-1 *gyr*A96 *rel*A1 λ^-) was used for molecular cloning. pT7blue (Novagen), pWUR112 (a kind gift of Dr. John van der Oost and described in Brouns *et al*., 2005) and pMH184 (a kind gift of Dr. Jose Berenguer and described in Cava *et al*., 2007) plasmid vectors were used for mutagenesis of *T. thermophilus* Hb27. pMK18 shuttle vector (de Grado *et al*., 1999), which is able to replicate in *E. coli* and *T. thermophilus*, was a kind gift of Dr. Jose Berenguer and was used for adaptation study and for cloning of protospacers and subsequent research of seed. Bacteriophage phiKo, which is able to infect *T. thermophilus* Hb27, was kindly provided by Dr. Anna Lopatina.

2.2. Media

Modified TB medium TBM (0.8% w/v tryptone, 0.4% w/v NaCl, 0.2% w/v yeast extract in Vittel mineral water) was used for *T. thermophilus* cultivation. For cultivation on plates, agar was added to TBM medium up to 2% w/v concentration. LB medium was used for *E.coli* cultivation (see "Materials and Methods" in Chapter 2). For adaptation assay TBM-2 medium (0.25% w/v tryptone, 0.4% w/v NaCl in Vittel mineral water) was also used.

Antibiotics ampicillin (100 μg/ml), kanamycin (50 μg/ml for *E.coli* and 30 μg/ml for *T. thermophilus*), hygromycin (40 μg/ml) and bleomycin (15 μg/ml) were used to maintain pT7Blue, pMK18, pMH184 and pWUR112 plasmids, respectively, and for the corresponding selection of *T. thermophilus* mutant strains.

T. thermophilus and *E. coli* were cultivated at 70 °C and 37 °C, respectively, with aeration (150 rpm) in liquid media, unless otherwise stated.

2.3. DNA purification

Genomic DNA was purified from 1 ml of single colony-derived bacterial overnight culture using a Genomic DNA Purification Kit (Thermo Fisher Scientific) and plasmid DNA was purified from 5 ml of bacterial overnight culture using a GeneJET Plasmid Miniprep Kit (Thermo Fisher Scientific) from 5 ml of overnight culture in LB medium supplemented with antibiotics), following the protocol provided by the manufacturer.

2.4. Preparation of *E. coli* DH5α chemically competent cells

Competent cells were prepared using the standard protocol (Inoue *et al.*, 1990).

2.5 Assembling of constructions for *T. thermophilus* mutagenesis

Genes of antibiotic resistance, flanking areas of *T. thermophilus* genes and pT7blue plasmid vector were amplified by PCR. pWUR112/pMH184, and genomic DNA or pT7blue were used as a template. Phusion DNA Polymerase (New England Biolabs) was used for the PCR reaction. The PCR products were purified using GeneJET Gel Extraction and DNA Cleanup Micro Kit (Thermo Fisher Scientific) after separation by 0.8% w/v agarose gel electrophoresis in 0.5x TBE buffer (see "Materials and Methods" in Chapter 2). Finally, all fragments were assembled with Gibson Assembly Master Mix (New England Biolabs), following the protocol

provided by the manufacturer. *E. coli* DH5α competent cells were transformed with the mix and plated on agar LB medium with ampicillin.

2.6. Transformation of *T. thermophilus*

T. thermophilus was transformed with plasmids as described in de Grado *et al.* (de Grado *et al.*, 1999).

2.7. Mutagenesis of *T. thermophilus*

T. thermophilus cells were transformed with assembled T7blue-based plasmid constructions carrying an antibiotic resistance gene, were placed on agarized TBm medium supplied with correspondent antibiotic and were incubated for 18-24 hours at 70 °C. Several colonies were then restreaked on a fresh plate supplied with the antibiotic. This procedure was repeated several times. Replacement of the desired *T. thermophilus* gene with the antibiotic resistance gene was checked with PCR and Sanger sequencing of purified PCR product. All Sanger sequencing was performed by Evrogen.

2.8. Creation of plasmids with protospacers

All oligonucleotides were synthesized by Evrogen. Pairs of complementary oligonucleotides were annealed as described in chapter 2 and ligated with purified shuttle pMK18 plasmid vector processed by EcoRI and HindIII restriction endonucleases. This vector can be replicated in both *E. coli* and *T. thermophilus.* Chemically competent *E. coli* cells were transformed with ligation mix and plated on a medium supplied with kanamycin.

2.9. CRISPR adaptation assay

- with plasmids:

Clones of cells transformed with selected plasmids were grown for a total of 3 days. Every 24h aliquots of cultures were diluted (1:500) into fresh medium. Then aliquots of *T. thermophilus* cultures were heated for 15 minutes at 98 °C and adaptation was checked by PCR with a pair of primers, one (forward) annealing in the leader sequence and another (reverse) at CRISPR arrays spacers.

- with bacteriophage:

Cells were grown in TBM medium and were infected at OD_{600} 0.4 with *T. thermophilus* bacteriophage phiKo with multiplicity of infection ranging from 0.1 and 0.01. Aliquots of cell cultures were taken after overnight and 72-h incubation. Adaptation was checked by PCR as described above.

- with synthetic deoxyriboonucleotides:

450 µl of *T. thermophilus* culture at OD_{600} 0.4 were transformed with 50 µl of 20 µM single-stranded or annealed double-stranded synthetic deoxyribooligonucleotides (synthesized by Evrogen). Cells were incubated for 2 hours at 70 °C without aeration. Aliquots for analysis were taken. Then 3 ml of TBM medium were added and samples were grown overnight at 70 °C with aeration. Adaptation was checked by PCR with primers annealing to the leader and to the middle of the sequence of deoxyribooligonucleotides used for transformation (in direct and reverse orientations) after 20-h and overnight incubation.

PCR reaction products were always analyzed by 1.5% w/v agarose gel electrophoresis in 0.5x TBE buffer.

2.10. CRISPR interference assay

T. thermophilus cells were transformed with pMK18 vector (control) or with a plasmid carrying a protospacer that was fully-matched to the 1st spacer of CRISPR-8 array (1.8) or carrying mutated 1.8 protospacer. 1.5 hours after transformation, several 10-fold dilutions were made and 10 µl of each dilution were placed on a plate with kanamycin supplied medium. The cells were incubated for 18 hours at 70 °C.

3. Results

3.1. Looking for naïve adaptation in *T. thermophilus* Hb27

T. thermophilus Hb27 cells were transformed with pMK18 plasmid and incubated for 3 days with or without aeration in TBM or TBM-2 media at 60 °C, or 65 °C, or 70 °C. PCR with oligonucleotides annealing to the leader and to the first spacer for each of *T. thermophilus* Hb27 CRISPR arrays was used for monitoring acquisition of new spacers. The length of PCR products from unexpanded CRISPR arrays ranged from 114 bp to 143 bp. No spacer acquisition was detected in any of the arrays at all conditions tested (Figure 25).

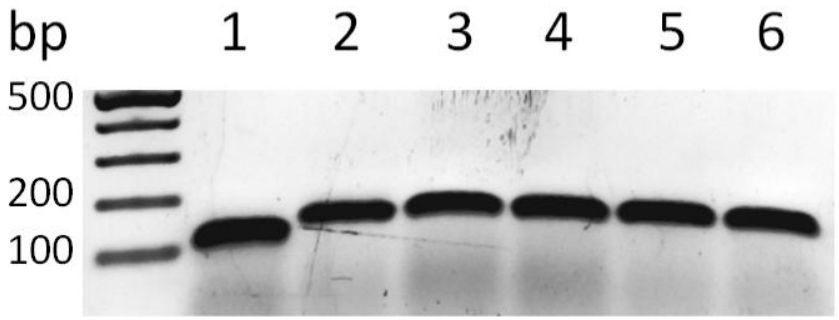


Figure 25. Absence of naïve adaptation in *T. thermophilus* Type III CRISPR arrays.

The leader-proximal ends of CRISPR arrays were amplified using primers annealing to leader regions and first spacer of each Type III array. Amplification products were separated by agarose gel electrophoresis. A typical result obtained with six *T. thermophilus* Hb27 Type III CRISPR arrays (lanes 1 to 6) is shown.

Next, cells were infected with *T. thermophilus* bacteriophage phiKo. At various times culture aliquots were monitored as described above, but no adaptation was detected.

In yet another approach described in Shipman *et al.* for detection of spacer acquisition by I-E CRISPR-Cas system of *E. coli* (Shipman *et al.*, 2016), *T. thermophilus* Hb27 cells were

transformed with spacer-sized single- or double-stranded deoxyribooligonucleotides whose sequence (5'-ATGTGGCAGTCCAACGACGACGGCTTCGGCATGTTCTTTCG-3') matched a spacer from Type III CRISPR array of *T. thermophilus* Hb8 strain. Primers annealing to the leader and to the middle of this sequence (in direct and reverse orientations) were used for PCR but no acquisition was observed using this sensitive procedure.

To summarize, all our attempts to detect naïve adaptation by Type III *T. thermophilus* systems by methods that have been successfully used with various Type I and Type II systems have failed. The question, therefore, arouses if these systems are active at laboratory conditions (note that in *E. coli*, the subtype I-E CRISPR-Cas system is inactive and genetic manipulations such an insertion of promoters upstream the genes and an overexpression of proteins are required to reveal its interference and adaptation potential).

3.2. Creation of *T. thermophilus* strains lacking III-A, III-B, or both Type III systems

The *T. thermophilus* Hb27 strain contains genes coding for proteins of both III-A and III-B subtype CRISPR-Cas system and 10 CRISPR arrays. Six of ten CRISPR arrays of *T. thermophilus* Hb27 (CRISPR-1, 2, 5, 8, 9, and 10) are shared between the III-A and III-B subtypes (Swarts *et al*., 2015) (Figure 26).

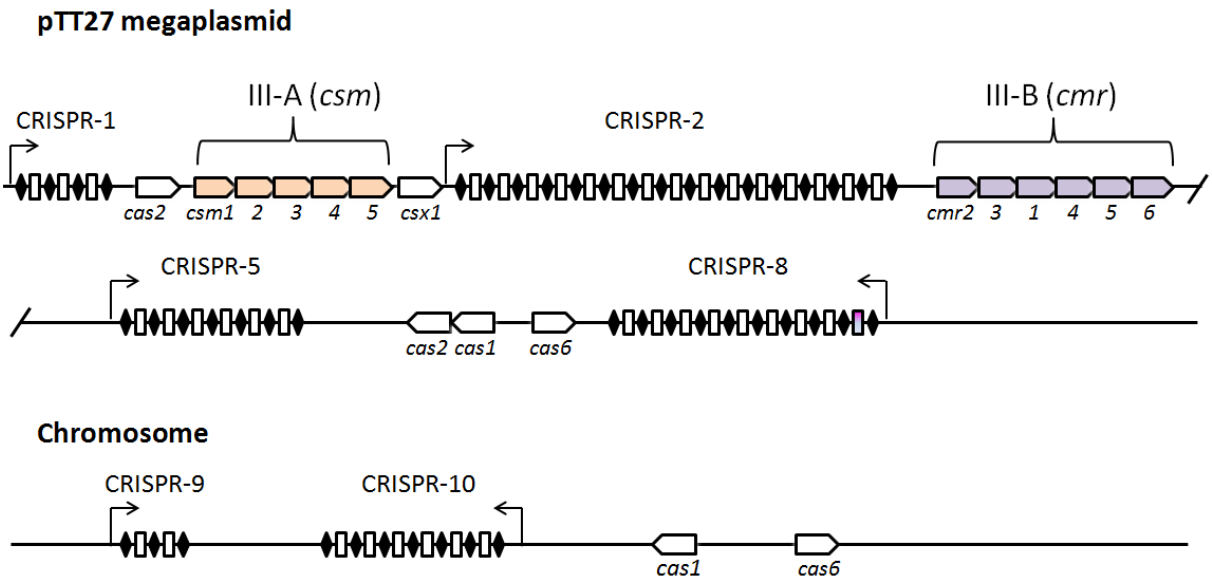


Figure 26. Organization of *T. thermophilus* Hb27 Type III CRISPR-Cas systems.

CRISPR repeats are indicated as black rhombi, spacers – as white rectangles. Promoters from which CRISPR arrays are transcribed are marked with arrows. The III-A subtype *csm* genes are indicated as orange pentagons, the III-B subtype *cmr* genes – as purple pentagons. Other Type III CRISPR related genes are shown as white pentagons.

Mutant strains were constructed in order to differentiate the action of the III-A and III-B systems. In one strain, the *cmr4* gene was replaced with a gene providing hygromycin resistance. Four copies of the Cmr4 protein with ribonuclease activity compose the core of the Cmr-complex. So the III-B subtype system must be deactivated in this mutant, which was called the "III-A strain", to indicate that only this Type III subtype is active in it. In another strain, the III-A subtype homolog of *cmr4*, the *csm3* gene, was replaced with hygromycin resistance gene to create the "III-B strain" with deactivated III-A subtype system. A control "double mutant" (DM) strain was created by replacing the *cmr4* gene by a gene providing bleomycin resistance in the III-A strain.

3.3. Both *T. thermophilus* Hb27 III-A and III-B subtype systems are capable of interference

Both III-A and III-B subtype systems show transcriptional-dependent deoxyribonuclease activity (Deng *et al.*, 2013; Goldberg *et al.*, 2014). To test the activity of the III-A and III-B strains, plasmids carrying a protospacer corresponding to the 1st spacer of the 8th CRISPR array (spacer 1.8: 5'- CTCTTTCAGGATCCACGCAAACTCCCTTCCTTGGGGCTTA-3') were created. Plasmid pMK18-dir has this protospacer in "direct" orientation with respect to transcription from the PslpA promoter from which the plasmid-borne antibiotic resistance gene is transcribed and plasmid pMK18-rev has it in the opposite orientation (Figure 27). Plac promoter, which is not expected to be active in *Thermus,* is located on the other side (relative to PslpA promoter) of the multiple cloning site where protospacer was cloned (Galindo, 2007).

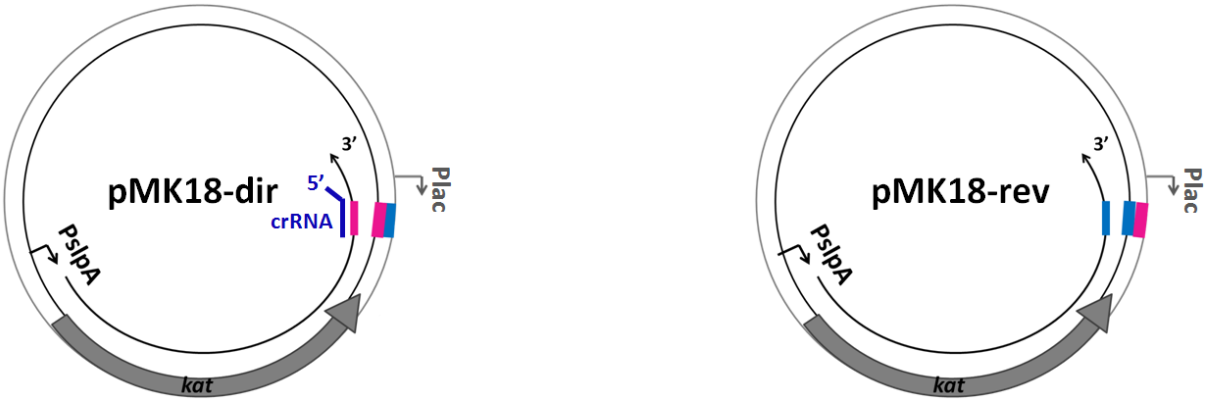


Figure 27. The scheme of pMK18 plasmids, in which the protospacer is cloned in the direct and reverse orientations.

The strand of protospacer complementary to the 1.8 crRNA, is shown in pink, the strand with the matching sequence is shown in blue. Transcript synthesized from PslpA promoter is shown as a solid black arrow.

The 1.8 crRNA spacer has the same sequence as the RNA protospacer transcribed from the PslpA promoter in the case of the pMK18-rev plasmid, and so this plasmid is not expected to

be subject to interference. “Direct” orientation means that crRNP Csm and Cmr complexes containing 1.8 crRNA could bind with the transcript. This binding should activate deoxyribonucleic activity of the Cas10 subunit of these complexes and should lead to interference.

Interference was tested for each of the strains. Cells were transformed with the control empty pMK18 vector and plasmids carrying protospacer in both orientations (Figure 28). Serial 10-fold dilutions of transformed cultures were deposited on plates with selective medium. Colony formation indicates the lack of interference.

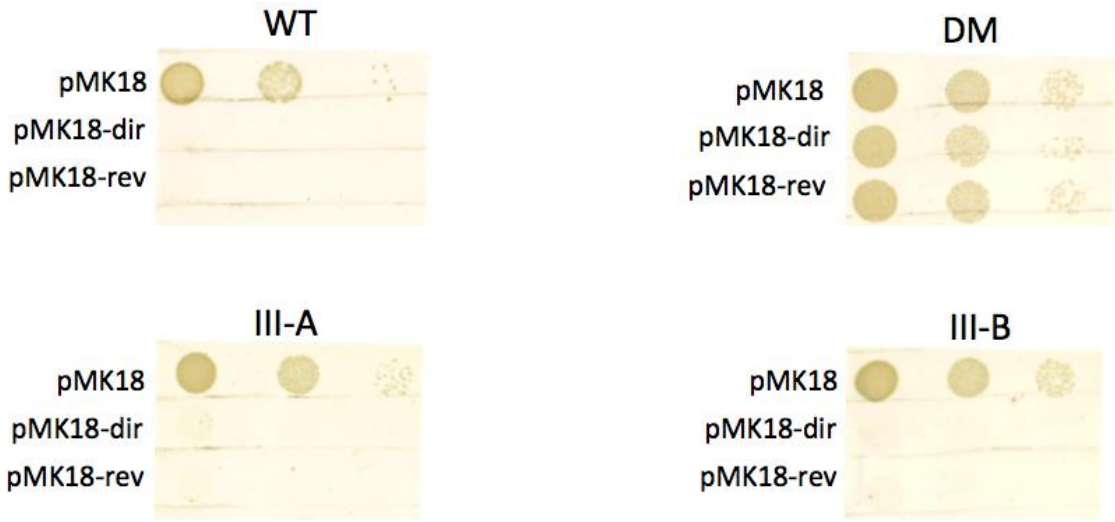


Figure 28. Type III systems of *T. thermophilus* Hb27 interfere with plasmid transformation.

The results of transformation of *T. thermophilus* Hb27 WT, DM, III-A and III-B strains with protospacer-containing plasmids or pMK18 vector are shown. Serial 10-fold dilutions of transformed cultures were deposited on plates with selective medium. Colony formation indicates the lack of interference and the absence of colonies corresponds to high level of interference.

Surprisingly, the efficiency of transformation was equally low (compared to pMK18 EOT) not only for the pMK18-dir but also for the pMK18-rev plasmid. The EOT in DM strain was equally high for all three plasmids. This indicates, firstly, that both III-A and III-B systems

are active in *T. thermophilus* Hb27. Secondly, the results suggest that the region with cloned protospacer may be transcribed from both strands (Figure 29). Indeed, RNA-Seq data obtained in our laboratory showed the presence of plasmid-borne transcripts generated from a strand opposite to the one transcribed from the PslpA promoter. The abundance of these transcripts at the site of multiple cloning site was ~50-fold lower than complementary transcripts initiated from PslpA, but apparently was sufficient to induce Type III interference.

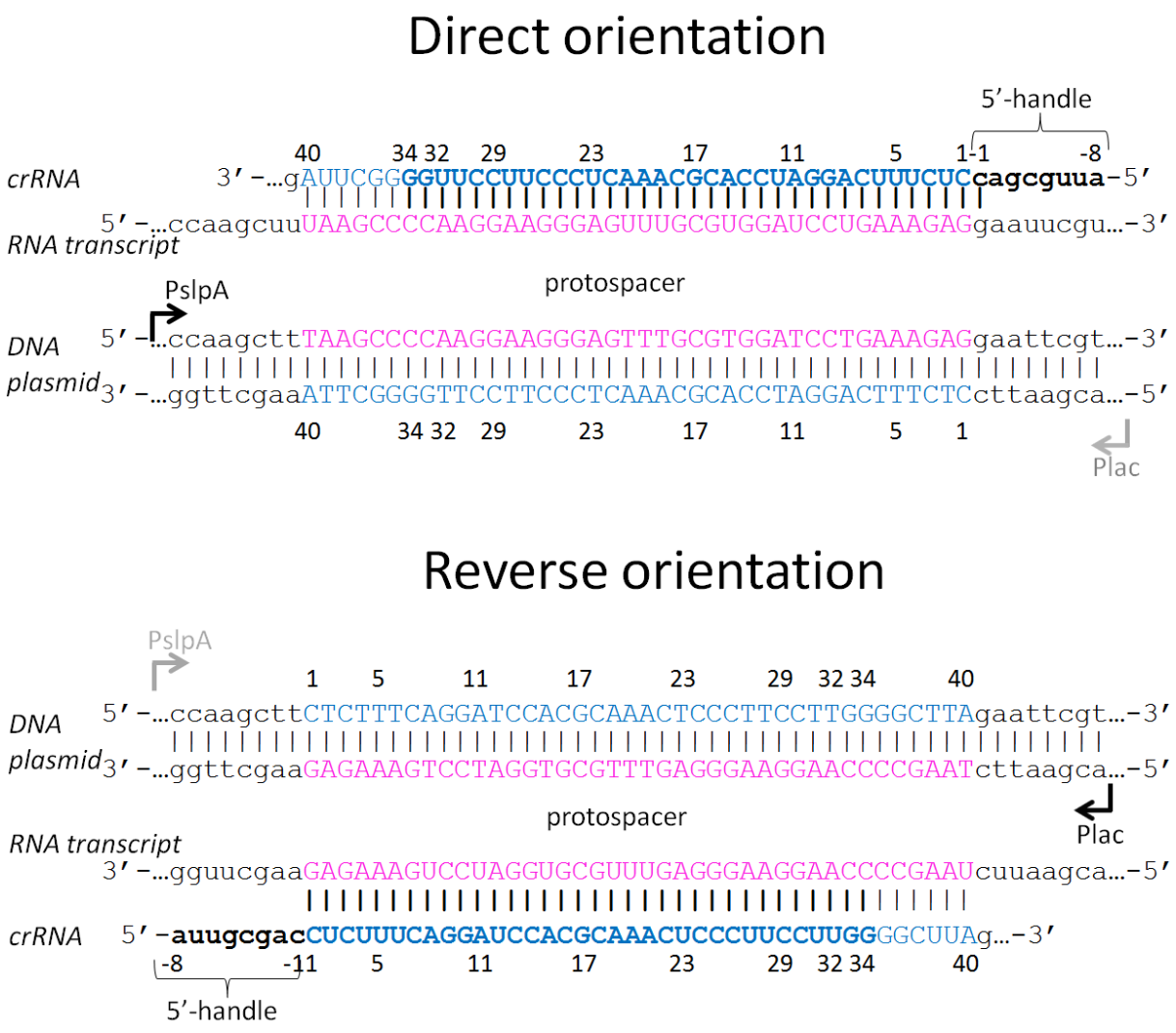


Figure 29. A scheme of Type III targeting for the direct and reverse orientation of cloned into pMK18 protospacer.

The immunity is provided by recognition of a transcript from the PslpA promoter in case of the direct orientation of the protospacer and by recognition of transcript from the opposite strand in case of the reverse orientation of the protospacer.

3.4. Dependence of interference on orientation and mutations in the protospacer

We were interested to detect a Type III seed region, i.e., locate positions of protospacer recognized by Type III effectors that are particularly important for its recognition and destruction. We reasoned that if identified, protospacer plasmids with mutations in such positions could be used to detect primed adaptation. For this purpose, a set of plasmids was constructed with mutations at different positions of protospacer introducing mismatches with 1.8 crRNA. The A<->T and G<->C substitutions in protospacer were introduced. It is known that most important protospacer region for PAM-dependent CRISPR-Cas systems is located at the PAM-proximal end. In Type III systems the 5'-handle of crRNA takes part in autoimmunity avoidance playing a role analogous to PAM in Type I and Type II systems. So a mismatch was first introduced in the +1 position from the 3'-end of the protospacer. The mutation had no effect on interference by either III-A and III-B subtypes. The number of mismatches at the 3'-end of the protospacer was therefore increased. In the III-B strain, interference was abolished after changing 5 nucleotides, while in the III-A and WT strain interference was eliminated only after 8 mismatches had been induced.

The same logic was applied to probe the importance of matches at the 5'-end of the protospacer. In this case, interference was inhibited in the III-A and III-B strains by the introduction of 18 and 15 mismatches, respectively. This is a surprising result considering that the length of the 1.8 protospacer is 40 bp and thus very large portions of it are apparently not needed for target recognition by Type III effectors. Type III systems interference was more sensitive to the simultaneous introduction of mismatches at both ends of the protospacer: mismatches at positions 1, 2, and 27-40 (a total of 16 mismatches) abolished interference in III-A, III-B, and the wild-type strains. However, mismatches in positions 1 and 27-40 or 2 and 27-40 had no effect on interference.

Point mutations that introduced mismatches in positions immediately upstream of the sites of transcript cleavage previously determined *in vitro* (Staals *et al.*, 2013; Samai *et al*., 2015) did not affect interference. However, when all six mutations introducing mismatches were present together, interference was abolished. For the III-A strain, five of these mutations (mismatches at positions 5, 11, 17, 23 and 29 or 11, 17, 23, 29 and 35) were also sufficient to abolish interference but mutations introducing mismatches in any of the four sites did not have any effect.

The III-B strain was more sensitive to mismatch-introducing point mutations, in agreement with data on probing of 3’ and 5’ ends of the protospacer with multiple mismatches. Single point mutations did not affect interference but mutations introducing just two mismatches in positions 5 and 11 inhibited interference. However, other tested combinations of double mismatches (at positions 5 and 17; 5 and 23; 5 and 29; or 11 and 17) had no effect.

The III-A strain was tolerant to mutations introducing five consecutive mismatches, but combination of these mutations with some others (i.e., a combination of 1-5 and 7-11 mismatches or 1-5 and 11) abolished interference. The III-B strain was tolerant only to two tested variants of such type of mutations, with mismatches at positions 19-23 and 31-35. The 19-23 variant was extended by an additional mismatch at either side - 18-23 and 19-24. These variants also did not inhibit interference.

In some combinations, mismatched area shift by just one nucleotide had a dramatic effect on interference without changing the overall number of mismatches between spacer and protospacer: mutations introducing mismatches in positions 1-5, 9 and 10 inhibited interference in the III-A strain, while plasmid with protospacer mismatched at positions 1-5, 8, and 9 did not form transformants.

All mutations analyzed in this work are listed in Table 4. Review of the data suggests that the general features of interference by the III-A and III-B subtype CRISPR-Cas systems of *T. thermophilus* Hb27 are similar but that the III-B subtype system is more sensitive to mismatches than III-A.

Positions of mutations in direct oriented protospacer	III-A strain	III-B strain
No protospacer	-	-
WT	+	+
Mutations in the both ends of protospacer		
27-40	+	+
26-40	+	-
24-40	+	-
23-40	-	-
1-4	+	+
1-5	+	-
1-7	+	-
1-8	-	-
1-2,27-40	-	-
1,27-40	+	-
2,27-40	+	-
Point mutations and combinations of point mutations		
5	+	+
11	+	+
17	+	+
23	+	+
29	+	+
35	+	+
5,11,17,23,29,35	-	-
5,11	+	-
5,17	+	+
5,23	+	+

5,29	+	+
11,17	+	+
11,23	+	+
5,11,17	+	-
23,29,35	+	+
5,11,17,23	+	-
5,11,17,23,29	-	-
11,17,23,29,35	-	-
Mutations of five or more consecutive positions		
1-5	+	-
7-11	+	-
13-17	+	-
19-23	+	+
25-29	+	-
31-35	+	+
18-23	+	+
19-24	+	+
1-5,7-11	-	-
1-5,13-17	-	-
1-5,25-29	-	-
1-5,7,8	-	-
1-5,11	-	-
Shifting of mutations		
1-5,8,9	+	-
1-5,9,10	-	-
7-11,13,14	+	-
7-11,14,15	-	-

Table 4. Recognition of mutant protospacers in direct orientation by III-A and III-B subtype CRISPR-Cas systems.

"+" means low EOT with the pMK18 vector carrying the protospacer with indicated mutations and therefore, active interference. "-" means high EOT and, therefore, absence of interference.

In paragraph 3.2 of this chapter, it was demonstrated that interference can also occur in case of the reverse orientation of the protospacer. Transcription of protospacer from the opposite strand was proposed as an explanation for this observation. A plasmid with protospacer in reverse orientation but harboring a "borderline" number of mismatches at the 5'-end as defined by analysis of protospacer in direct orientation (17 and 14 for III-A and III-B, respectively) was used to find out whether the direction of the protospacer influences sensitivity to mutations. The idea behind this experiment was that lower level of transcription from the strand opposite to the template strand of the PslpA promoter could make interference more sensitive to mismatches. Indeed, we observed that interference with protospacers in reverse orientation was more sensitive to mismatches in III-A and III-B strains. For the 5' end of the protospacer, interference with plasmids harboring reverse-oriented protospacers was eliminated when there were mismatches starting at positions 26 and 32 for the III-A and III-B strains, respectively (15 and 9 mismatches (Table 5). These numbers were 18 and 15 for direct orientation of protospacer, for III-A and III-B strains, respectively).

Positions of mutations in reverse oriented protospacer	III-A strain	III-B strain
No protospacer	-	-
WT	+	+
Mutations in the 5'-end of protospacer		
24-40	-	-
26-40	-	-
27-40	+	-
32-40	+	-
33-40	+	+

Table 5. Recognition of reverse-oriented mutant protospacers by III-A and III-B subtype CRISPR-Cas systems.

"+" means low EOT with the pMK18 vector carrying cloned protospacer with indicated mutations and, therefore, active interference. "-" means high EOT and, therefore, absence of interference.

3.5. Primed adaptation in *T. thermophilus* Hb27

Establishing the positions of mismatches that abolished interference allowed us to test the ability of *T. thermophilus* Type III systems to perform primed adaptation. Each strain was transformed with the plasmid which contained a protospacer variant with the minimal number of mutations that abolished interference: introducing mismatches in positions 5, 11, 17, 23 and 29 for III-A and WT strains and mismatches in positions 5 and 11 for III-B strain. Transformed cells were cultivated in TBM medium for 3 days at 70 °C. Expansion of each of six Type III CRISPR arrays was tested by PCR as described above for naïve adaptation. Unfortunately, no acquisition of new spacers was detected (Figure 30 A-D, lines 1-6).

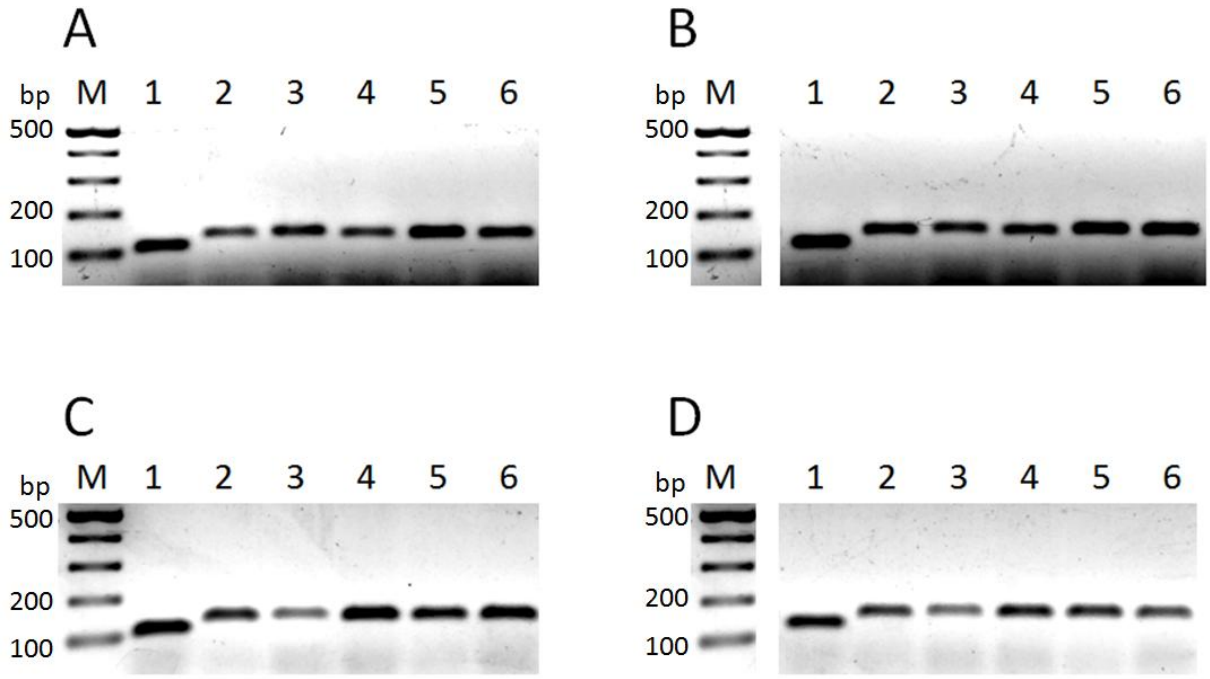


Figure 30. Absence of primed adaptation of Type III CRISPR arrays in A – III-A strain, B – III-B strain, C – WT strain and D – DM strain transformed with protospacer plasmids harboring mismatches that abolish interference.

Cells were transformed with plasmids containing mutant protospacers with mismatches in positions 5, 11, 17, 23 and 29 for III-A and WT strains and mismatches in positions 5 and 11 for III-B and DM strains. Adaptation was checked by PCR in the same way as it was done in the case of naïve adaptation. Results for six Type III CRISPR arrays of *T. thermophilus* Hb27 are shown in lanes 1-6.

Primed adaptation was also tested in the presence of protospacer plasmid with threshold number of mismatches at the 5' end (18 mismatches, from 23^{rd} to 40^{th} positions of the protospacer in the case of III-A and WT strains and 15 mismatches, from 26^{th} to 40^{th} positions of protospacer in case of III-B strain and control DM strain). Again, no array expansion was observed.

4. Discussion

Several subtypes of Type I CRISPR-Cas systems, e.g., I-B and I-C, demonstrate only the primed mode of adaptation in laboratory conditions (Li *et al*., 2014; Rao *et al*., 2016). This may be a specific feature of these systems or, more likely, the range of conditions favorable for naïve adaptation is very narrow, and they have not been reproduced in the laboratory. Perhaps such mode of adaptation is possible only in Type I systems due to some specific features of Type I interference machinery action. In contrast, for type II systems, only naïve adaptation has been described (Barrangou *et al*., 2007; Lopez-Sanchez *et al*., 2012; Heler *et al*., 2015; Sternberg *et al*., 2016). To date, for Type III systems, only naïve adaptation was observed and only for a III-B subtype system, which contains a RT-Cas1 fusion protein. Moreover, adaptation was detected only under conditions of overexpression of RT-Cas1 and Cas2 proteins but not in a natural strain (Silas *et al.*, 2016). In this part of work, *T. thermophilus* strains that are suitable for separate study of III-A and III-B subtype CRISPR-Cas systems were constructed and conditions for naïve or primed adaptation were sought. Naïve adaptation was not detected either with plasmids, phage infection, or oligo transformation assays. Primed adaptation in Type I systems is observed in the presence of mutant protospacers containing mismatches with crRNA spacer in the seed region or non-consensus PAM (Datsenko *et al*., 2012). For the study of primed adaptation by Type III systems, it was necessary to find out which mismatches between a protospacer and a crRNA abolish interference. Since the seed in Type III CRISPR-Cas systems was not known, we have attempted not only to find mutations which inhibit interference and, hence, will be suitable for detection of primed adaptation but to establish the III-A and III-B subtype systems seed sequences. Results showed that, in general, III-A and III-B subtype CRISPR-Cas systems react to the introduction of mismatches between crRNA and protospacer in a similar way

(Figure 31), but the III-B subtype is more sensitive to the number of mismatches than the III-A subtype.

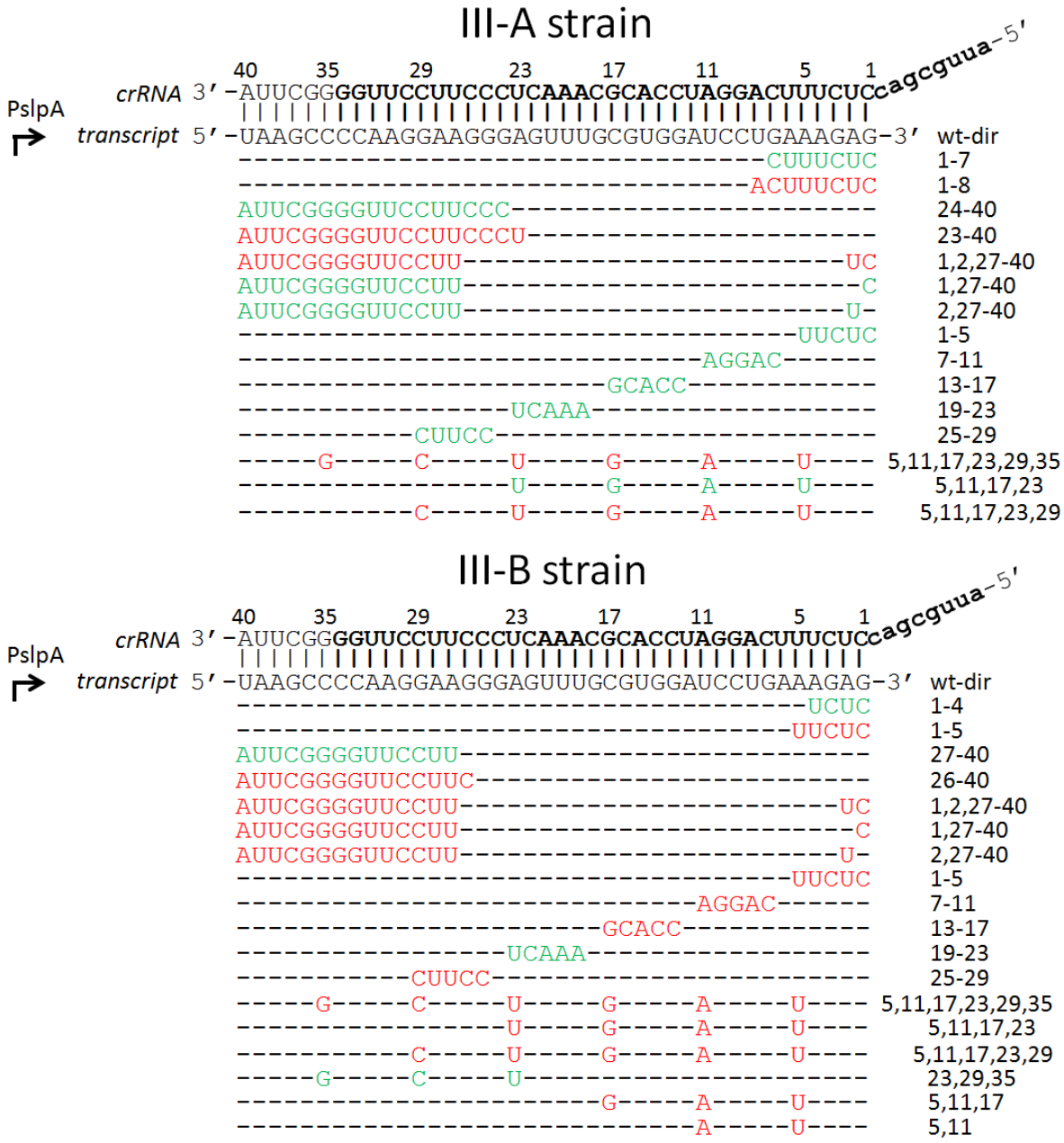


Figure 31. Summary of the data on interference by *T. thermophilus* III-A and III-B systems in with plasmids harboring protospacers with mismatches to crRNA spacer.

The sequence of crRNA is on the top, the spacer part is 40-nucleotides long. The complementary sequence of the protospacer is shown below. Nucleotide numbering starts from the 3' end of the protospacer. Only mutated nucleotides are shown in the scheme, dashes show positions complementary to crRNA spacer. Green color indicates mutations that had no effect on interference; red color shows mutations that abolished interference, leading to high efficiency of transformation of mutant plasmids.

The reason of that is unclear, but we propose that it could be linked with the level of Cmr proteins expression in the cells by analogy with the dependence of tolerance to the number of mutations on different levels of transcription of protospacer cloned in the direct and reverse orientation into pMK18 plasmid. In addition, based on the results, it can be concluded that the Type III seed sequence is not located at the flanks of the protospacer, as is the case in Type I or II systems. However, the 5'-end of a protospacer is more tolerant to mutation than the 3'-end in both III-A and III-B strains. A surprisingly large number of permissible mismatches at the 5'-end of the protospacer can be explained with the mechanism of crRNA maturation in Type III systems: after primary processing of pre-crRNA by Cas6, crRNAs are additionally processed from the 5'-end. During this process, crRNA could be partly degraded by the unknown nuclease. There is some data demonstrating participation of Csm3 and Cmr4 in the process of final maturation (Hatoum-Aslan *et al*, 2011; Hale *et al.*, 2012). Several fractions of crRNA with the same length can be formed. Unpublished RNA high-throughput sequencing data from our laboratory show that the most abundant lengths of crRNA with 1.8 spacer sequence are 42 nucleotides for the III-A strain and 40 nucleotides for III-B. In both cases, 8 nucleotides of the crRNA are taken by the 5'-handle. Thus, spacer parts of 1.8 crRNAs are only 34 and 32 ribonucleotides in the III-A and III-B strains, respectively. Thus, full matching with the 5'-end of the protospacer does not occur even in the presence of non-mutated protospacer, because the spacer part of mature crRNA in the crRNP effector complex is shorter than the protospacer.

It is interesting, that recently Pyenson *et al.* hypothesized that primed adaptation is not possible in Type III because of their tolerance to a large number of mutations in protospacer and impossibility of escapes appearance by acquiring of mutations in a protospacer (Pyenson *et al*., 2017). In Type I systems, primed adaptation helps cells to restore protection against MGE which escaped by a mutation in the seed region of protospacer or in PAM.

However it very hard for an invader to avoid action of Type III CRISPR-Cas systems because of broad targeting specificity and high-level tolerance to mutations provided by this type of systems (Pyenson *et al*., 2017).

Conclusions

Systems of Class 1 have a common predecessor (Shmakov *et al*., 2015), but, as it has been shown in this study, differences in the mechanisms of their actions are observed not only among different types and subtypes, but even among systems belonging to the same subtype, but encoded in different bacteria.

The requirements for naïve and primed adaptation by I-F CRISPR-Cas system was established *in vivo* for the first time in this work. The high-throughput sequencing of acquired spacers has been never carried out before. This analysis revealed a much larger pool of new spacers in comparison with previous studies (Richter *et al.*, 2014). It allowed establishing the shape of the distribution of acquired spacers from the region surrounding the priming point in more detail.

Several works on seed search and investigations of mutations in protospacer were undertaken before for Type III systems (Staals *et al*., 2014; Peng *et al*., 2015) but detailed analysis of impact of different mismatches between crRNA and a protospacer and combinations of these mismatches has been never carried out previously. Unfortunately, neither naïve nor primed adaptation was detected in both III-A and III-B subtypes of CRISPR-Cas systems. It should be noticed once again that hypothesis of absence of primed adaptation has been proposed by Pyenson *et al.* (Pyenson *et al*., 2017) and the results obtained in our work confirm that hypothesis.

Thus, the main conclusions of our work are as follows:

1. In contrast to I-E, all components of CRISPR-Cas machinery are required for both naïve and primed adaptation in I-F subtype system.

2. The distribution of acquired spacers from the region surrounding the priming point in I-F system is opposite to the distribution observed in I-E system.

3. Interference but no adaptation was observed by *T. thermophilus* III-A and III-B CRISPR-Cas systems.

4. Multiple mismatches at the 3'- and 5'- ends of crRNA and the target have no effect on interference by III-A and III-B systems.

5. If the seed exists in III-A and III-B subtype systems, it has a complicated geometry and may be non-continuous.

The summary of the investigation and comparison of the results with the known data on the I-E subtype CRISPR-Cas system (the most studied subtype among all Class 1 systems) are presented in Table 6.

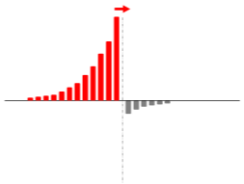
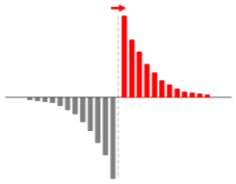
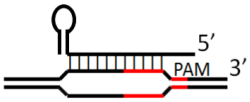
	I-E	I-F	III-A	III-B
Naïve adaptation	Cas1-Cas2 complex (Yosef *et al.*, 2012)	All components of CRISPR machinery	Not detected	Requirements are not defined (Silas *et al.*, 2016)
Primed adaptation	The shape of gradient around of priming protospacer  (Strotskaya *et al.*, 2017)	The shape of gradient around of priming protospacer 	Not detected	Not detected
Position of seed	3'-{1-5,7,8}-5'  (Semenova *et al.*, 2011)	3'-{1-8}-5'  (Wiedenheft *et al.*, 2011a)	? 	? 

Table 6. The comparative table describing some features of the action the under study subtypes I-F, III-A and III-B and the "reference" I-E subtype CRISPR-Cas systems.

Finally, it should be noticed that the investigation of action of Class 1 CRISPR-Cas system and their molecular mechanisms are important not only for fundamental science but also for practical purposes. Generally, Class 2 systems are used in industry because of their relative mechanistic simplicity, however, Class 1 systems can also be used for practical approaches, but only for manipulations with prokaryotic cells with active CRISPR-Cas systems initially located in their genomes. It is possible to seek induced and spontaneous mutations in the

certain regions of the prokaryotic cell genome and to use these systems for the screening after gene deletion and insertion by homologous recombination-based methods, for example. Cells could be transformed with the special plasmid that carries an artificial CRISPR array with a spacer against the region of interest. This way, only mutated cells will stay alive and cells with the initial sequence will be killed by their own CRISPR-Cas system. (Li *et al.*, 2016) Thus, the influence of mismatches between crRNA and the protospacer on the target's recognition and following destruction investigated for III-A and III-B CRISPR-Cas systems in this work can be applied for modification of the mutation screening method, which is described above.

This work was supported by the Foundation for Assistance to Small Innovative Enterprises program “UMNIK” on the topic “Development of an antibiotic-independent screening gene deletions method for the inactivation of genes in thermophilic bacterium *Thermus thermophilus*” under the contract N 9188ГУ/2015 (0019877) of 12/24/2015.

Bibliography

Abudayyeh OO, Gootenberg JS, Konermann S, Joung J, Slaymaker IM, Cox DB, Shmakov S, Makarova KS, Semenova E, Minakhin L, Severinov K, Regev A, Lander ES, Koonin EV, Zhang F. (2016) C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science*, 353(6299), aaf5573.

Almendros C, Guzmán NM, Díez-Villaseñor C, García-Martínez J, Mojica FJ. (2012) Target motifs affecting natural immunity by a constitutive CRISPR–Cas system in *Escherichia coli*. *PLoS One*, 7(11), e50797.

Amitai G, Sorek R. (2016) CRISPR-Cas adaptation: insights into the mechanism of action. *Nat. Rev. Microbiol*., 2, 67-76.

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

Barrangou R, Marraffini LA. (2014) CRISPR-Cas systems: Prokaryotes upgrade to adaptive immunity. *Mol. Cell*, 54(2), 234-244.

Beloglazova N, Brown G, Zimmerman MD, Proudfoot M, Makarova KS, Kudritska M, Kochinyan S, Wang S, Chruszcz M, Minor W, Koonin EV, Edwards AM, Savchenko A, Yakunin AF. (2008) A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. *J. Biol. Chem.*, 283(29), 20361-20371.

Beloglazova N, Petit P, Flick R, Brown G, Savchenko A, and Yakunin AF. (2011) Structure and activity of the Cas3 HD nuclease MJ0384, an effector enzyme of the CRISPR interference. *EMBO J.*, 30(22), 4616–4627.

Benda C, Ebert J, Scheltema RA, Schiller HB, Baumgärtner M, Bonneau F, Mann M, Conti E. (2014) Structural model of a CRISPR RNA-silencing complex reveals the RNA-target cleavage activity in Cmr4. *Mol. Cell*, 56(1), 43–54.

Blosser TR, Loeff L, Westra ER, Vlot M, Künne T, Sobota M, Dekker C, Brouns SJ, Joo C. (2015) Two distinct DNA binding modes guide dual roles of a CRISPR-Cas protein complex. *Mol. Cell*, 58 (1), 60–70.

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

Bondy-Denomy J, Pawluk A, Maxwell KL, Davidson AR. (2013) Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature*, 493(7432), 429-432.

Bondy-Denomy J, Garcia B, Strum S, Du M, Rollins MF, Hidalgo-Reyes Y, Wiedenheft B, Maxwell KL, Davidson AR. (2015) Multiple mechanisms for CRISPR-Cas inhibition by anti-CRISPR proteins. *Nature*, 526(7571), 136-139.

Brouns SJ, Wu H, Akerboom J, Turnbull AP, de Vos WM, van der Oost J. (2005) Engineering a selectable marker for hyperthermophiles. *J Biol Chem.*, 280(12), 11422-11431.

Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, Snijders AP, Dickman MJ, Makarova KS, Koonin EV, van der Oost J. (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science*, 321(5891), 960-964.

Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD, Kerlavage AR, Dougherty BA, Tomb JF, Adams MD, Reich CI, Overbeek R, Kirkness EF, Weinstock KG, Merrick JM, Glodek A, Scott JL, Geoghagen NS, Venter JC. (1996) Complete genome sequence of the methanogenic archaeon *Methanococcus jannaschii*. *Science*, 273(5278), 1058–1073.

Cady KC, Bondy-Denomy J, Heussler GE, Davidson AR, and O'Toole GA. (2012) The CRISPR/Cas adaptive immune system of *Pseudomonas aeruginosa* mediates resistance to naturally occurring and engineered phages. *J. Bacteriol*., 194(21), 5728–5738.

Carr JF, Danziger ME, Huang AL, Dahlberg AE, Gregory ST. (2015) Engineering the genome of *Thermus thermophilus* using a counterselectable marker. *J Bacteriol.*, 197(6), 1135-1144.

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

Cava F, Laptenko O, Borukhov S, Chahlafi Z, Blas-Galindo E, Gómez-Puertas P, Berenguer J. (2007) Control of the respiratory metabolism of *Thermus thermophilus* by the nitrate respiration conjugative element NCE. *Mol Microbiol*., 64(3), 630-646.

Cava F, Hidalgo A, Berenguer J. (2009) *Thermus thermophilus* as biological model. *Extremophiles*, 13(2), 213-231.

Charpentier E, Richter H, van der Oost J, & White MF. (2015) Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. *FEMS Microbiol. Rev*., 39(3), 428–441.

Datsenko KA, Pougach K, Tikhonov A, Wanner BL, Severinov K, & Semenova E. (2012) Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat. Commun*., 3, 945.

de Grado M, Castán P, Berenguer J. (1999) A high-transformation-efficiency cloning vector for *Thermus thermophilus*. *Plasmid*, 42(3), 241-245.

Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, Eckert MR, Vogel J, Charpentier E. (2011) CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*, 471(7340), 602-607.

Deng L, Garrett RA, Shah SA, Peng X, She Q. (2013) A novel interference mechanism by a type III-B CRISPR-Cmr module in *Sulfolobus*. *Mol. Microbiol*., 87(5), 1088-1099.

Deng W, Shi X, Tjian R, Lionnet T, Singer RH. (2015) CASFISH: CRISPR/Cas9-mediated in situ labeling of genomic loci in fixed cells. *Proc. Natl. Acad. Sci. U S A*., 112(38), 11870-11875.

Deveau H, Barrangou R, Garneau JE, Labonté J, Fremaux C, Boyaval P, Romero DA, Horvath P, Moineau S. (2008) Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J. Bacteriol*., 190(4), 1390-1400.

Díez-Villaseñor C, Almendros C, García-Martínez J, Mojica FJ. (2010) Diversity of CRISPR loci in *Escherichia coli*. *Microbiology*, 156(5), 1351-1361.

Díez-Villaseñor C, Guzmán NM, Almendros C, García-Martínez J, Mojica FJ. (2013) CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of *Escherichia coli*. *RNA Biol*., 10(5), 792-802.

Dillingham MS & Kowalczykowski SC. (2008) RecBCD enzyme and the repair of double-stranded DNA breaks. *Microbiol. Mol. Biol. Rev.*, 72(4), 642–671.

Dwarakanath S, Brenzinger S, Gleditzsch D, Plagens A, Klingl A, Thormann K, Randau L. (2015) Interference activity of a minimal Type I CRISPR-Cas system from *Shewanella putrefaciens*. *Nucleic Acids Res.*, 43(18), 8913-8923.

Elmore JR, Sheppard NF, Ramia N, Deighan T, Li H, Terns RM, Terns MP. (2016) Bipartite recognition of target RNAs activates DNA cleavage by the Type III-B CRISPR-Cas system. *Genes Dev.*, 30(4), 447–459.

Erdmann S, and Garrett RA. (2012) Selective and hyperactive uptake of foreign DNA by adaptive immune systems of an archaeon via two distinct mechanisms. *Mol. Microbiol.*, 85(6), 1044–1056.

Erdmann S, Shah SA, and Garrett RA. (2013) SMV1 virus-induced CRISPR spacer acquisition from the conjugative plasmid pMGB1 in *Sulfolobus solfataricus* P2. *Biochem. Soc. Trans.*, 41(6), 1449–1458.

Estrella MA, Kuo FT, Bailey S. (2016) RNA-activated DNA cleavage by the Type III-B CRISPR-Cas effector complex. *Genes Dev.*, 30(4), 460–470.

Fineran PC, Charpentier E. (2012) Memory of viral infections by CRISPR-Cas adaptive immune systems: Acquisition of new information. *Virology*, 434(2), 202–209.

Fineran PC, Gerritzen MJ, Suárez-Diez M, Künne T, Boekhorst J, van Hijum SA, Staals RH, Brouns SJ. (2014) Degenerate target sites mediate rapid primed CRISPR adaptation. *Proc. Natl. Acad. Sci. U S A.*, 111(16), 1629-1638.

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

Fujita T, Fujii H. (2013) Efficient isolation of specific genomic regions and identification of associated proteins by engineered DNA-binding molecule-mediated chromatin immunoprecipitation (enChIP) using CRISPR. *Biochem. Biophys. Res. Commun.*, 439(1), 132-136.

Fagerlund RD, Wilkinson ME, Klykov O, Barendregt A, Pearce FG, Kieper SN, Maxwell HWR, Capolupo A, Heck AJR, Krause KL, Bostina M, Scheltema RA, Staals RHJ, Fineran PC. (2017) Spacer capture and integration by a type I-F Cas1-Cas2-3 CRISPR adaptation complex. *Proc Natl Acad Sci U S A*, 114(26), E5122-E5128.

Garside EL, Schellenberg MJ, Gesner EM, Bonanno JB, Sauder JM, Burley SK, Almo SC, Mehta G, MacMillan AM. (2012) Cas5d processes pre-crRNA and is a member of a larger family of CRISPR RNA endonucleases. *RNA*, 18(11), 2020-2028.

Goldberg GW, Jiang W, Bikard D, Marraffini LA. (2014) Conditional tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting. *Nature*, 514(7524), 633–637.

Goldfarb T, Sberro H, Weinstock E, Cohen O, Doron S, Charpak-Amikam Y, Afik S, Ofir G, Sorek R. (2015) BREX is a novel phage resistance system widespread in microbial genomes. *EMBO J.*, 34(2), 169–183.

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

Groenen PM, Bunschoten AE, van Soolingen D, van Embden JD. (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(5), 1057–1065.

Haft DH, Selengut J, Mongodin EF, Nelson KE. (2005) A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput Biol*., 1(6), e60.

Hale CR, Majumdar S, Elmore J, Pfister N, Compton M, Olson S, Resch AM, Glover CV 3rd, Graveley BR, Terns RM, Terns MP. (2012) Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. *Mol. Cell*, 45(3), 292–302.

Hale CR, Zhao P, Olson S, Duff MO, Graveley BR, Wells L, Terns RM, Terns MP. (2009) RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. Cell, 139(5), 945-956.

Hatoum-Aslan A, Maniv I, Marraffini LA. (2011) Mature clustered, regularly interspaced, short palindromic repeats RNA (crRNA) length is measured by a ruler mechanism anchored at the precursor processing site. *Proc. Natl. Acad. Sci. U S A.*, 108(52), 21218–21222.

Haurwitz RE, Jinek M, Wiedenheft B, Zhou K, Doudna JA. (2010) Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science*, 329(5997), 1355–1358.

Hein S, Scholz I, Voß B, Hess WR. (2013) Adaptation and modification of three CRISPR loci in two closely related cyanobacteria. *RNA Biol*., 10(5), 852–864.

Heler R, Samai P, Modell JW, Weiner C, Goldberg GW, Bikard D, Marraffini LA. (2015) Cas9 specifies functional viral targets during CRISPR–Cas adaptation. *Nature*, 519(7542), 199–202.

Henne A, Brüggemann H, Raasch C, Wiezer A, Hartsch T, Liesegang H, Johann A, Lienard T, Gohl O, Martinez-Arias R, Jacobi C, Starkuviene V, Schlenczeck S, Dencker S, Huber R, Klenk H-P, Kramer W, Merkl R, Gottschalk G, Fritz H-J. (2004) The genome sequence of the extreme thermophile *Thermus thermophilus*. *Nat Biotechnol.*, 22, 547–553.

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

Huo Y, Nam KH, Ding F, Lee H, Wu L, Xiao Y, Farchione MD Jr, Zhou S, Rajashankar K, Kurinov I, Zhang R, Ke A. (2014) Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. *Nat. Struct. Mol. Biol.*, 21(9), 771-777.

Inoue H, Nojima H, Okayama H. (1990) High efficiency transformation of *Escherichia coli* with plasmids. *Gene*, 96(1), 23-28.

Ivančić-Baće I, Radovčić M, Bočkor L, Howard JL, Bolt EL. (2013) Cas3 stimulates runaway replication of a ColE1 plasmid in *Escherichia coli* and antagonises RNaseHI. *RNA Biol.*, 10(5), 770–778.

Ivančić-Baće I, Cass SD, Wearne SJ, Bolt EL. (2015) Different genome stability proteins underpin primed and naïve adaptation in *E. coli* CRISPR-Cas immunity. *Nucleic Acids Res.*, 43(22), 10821–10830.

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

Jansen R, Embden JD, Gaastra W, Schouls LM. (2002) Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol.*, 43(6), 1565-1575.

Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, 337(6096), 816-21.

Jore MM, Lundgren M, van Duijn E, Bultema JB, Westra ER, Waghmare SP, Wiedenheft B, Pul U, Wurm R, Wagner R, Beijer MR, Barendregt A, Zhou K, Snijders AP, Dickman MJ, Doudna JA, Boekema EJ, Heck AJ, van der Oost J, Brouns SJ. (2011) Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nat. Struct. Mol. Biol*., 18(5), 529-36.

Jung TY, Park KH, An Y, Schulga A, Deyev S, Jung JH, Woo EJ. (2016) Structural features of Cas2 from *Thermococcus onnurineus* in CRISPR-Cas system type IV. *Protein Sci*., 25(10), 1890-1897.

Kazlauskiene M, Tamulaitis G, Kostiuk G, Venclovas Č, Siksnys V. (2016) Spatiotemporal control of Type III-A CRISPR-Cas immunity: coupling DNA degradation with the target RNA recognition. *Mol. Cell*, 62(2), 295–306.

Kearns NA, Genga RM, Enuameh MS, Garber M, Wolfe SA, Maehr R. (2014) Cas9 effector-mediated regulation of transcription and differentiation in human pluripotent stem cells. *Development*, 141(1), 219-223.

Kim EJ, Kang KH, Ju JH. (2017). CRISPR-Cas9: a promising tool for gene editing on induced pluripotent stem cells. *The Korean Journal of Internal Medicine*, *32*(1), 42–61.

Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD, Richardson DL, Kerlavage AR, Graham DE, Kyrpides NC, Fleischmann RD, Quackenbush J, Lee NH, Sutton GG, Gill S, Kirkness EF, Dougherty BA, McKenney K, Adams MD, Loftus B, Peterson S, Reich CI, McNeil LK, Badger JH, Glodek A, Zhou L, Overbeek R, Gocayne JD, Weidman JF, McDonald L, Utterback T, Cotton MD,

Spriggs T, Artiach P, Kaine BP, Sykes SM, Sadow PW, D'Andrea KP, Bowman C, Fujii C, Garland SA, Mason TM, Olsen GJ, Fraser CM, Smith HO, Woese CR, Venter JC. (1997) The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature,* 390(6658), 364–370.

Koonin EV, Krupovic M. (2015) Evolution of adaptive immunity from transposable elements combined with innate immune systems. *Nat. Rev. Genet*., 16(3), 184–192.

Koonin EV, Makarova KS, & Zhang F. (2017) Diversity, classification and evolution of CRISPR-Cas systems. *Current Opinion in Microbiology*, *37*, 67-78.

Kornberg A, and Baker TA. (2005) DNA Replication. *University Science Books,* 2nd edn.

Künne T, Kieper SN, Bannenberg JW, Vogel AI, Miellet WR, Klein M, Depken M, Suarez-Diez M, Brouns SJ. (2016) Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation. *Mol. Cell*, 63(5), 852-864.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res*., 19(9), 1639–1645.

Leenay RT, Beisel CL. (2017) Deciphering, Communicating, and Engineering the CRISPR PAM. *J. Mol. Biol*., 429(2), 177-191.

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

Li Y, Pan S, Zhang Y, Ren M, Feng M, Peng N, Chen L, Liang YX, She Q. (2016) Harnessing Type I and Type III CRISPR-Cas systems for genome editing. *Nucleic Acids Res.*, 44(4), e34.

Li M, Wang R, Zhao D, Xiang H. (2014) Adaptation of the *Haloarcula hispanica* CRISPR-Cas system to a purified virus strictly requires a priming process. *Nucleic Acids Res.*, 42(4), 2483–2492.

Lillestøl RK, Redder P, Garrett RA, Brügger K. (2006) A putative viral defence mechanism in archaeal cells. *Archaea*, 2(1), 59-72.

Lopez-Sanchez MJ, Sauvage E, Da Cunha V, Clermont D, Ratsima Hariniaina E, Gonzalez-Zorn B, Poyart C, Rosinski-Chupin I, Glaser P. (2012) The highly dynamic CRISPR1 system of *Streptococcus agalactiae* controls the diversity of its mobilome. *Mol Microbiol.*, 85(6), 1057-1071.

Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV. (2006) A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct.*, 1, 7.

Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, Horvath P, Moineau S, Mojica FJ, Wolf YI, Yakunin AF, van der Oost J, Koonin EV. (2011a) Evolution and classification of the CRISPR-Cas systems. *Nat. Rev. Microbiol.*, 9(6), 467–477.

Makarova KS, Aravind L, Wolf YI, Koonin EV. (2011b) Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biology Direct.*, 6, 38.

Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, Barrangou R, Brouns SJ, Charpentier E, Haft DH, Horvath P, Moineau S, Mojica FJ, Terns RM, Terns MP, White MF, Yakunin AF, Garrett RA, van der Oost J, Backofen R, Koonin EV. (2015) An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol.*, 13(11), 722–736.

Marraffini LA, and Sontheimer EJ. (2010). Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature*, 463(7280), 568–571.

Maxwell KL, Garcia B, Bondy-Denomy J, Bona D, Hidalgo-Reyes Y, Davidson AR. (2016) The solution structure of an anti-CRISPR protein. *Nat Commun.*, 7, 13134.

Miller EM, and Nickoloff JA. (1995) *Escherichia coli* electrotransformation. *Methods Mol. Biol*., 47, 105-113.

Mojica FJ, Juez G, Rodríguez-Valera F. (1993) Transcription at different salinities of *Haloferax mediterranei* sequences adjacent to partially modified PstI sites. *Mol Microbiol.*, 9(3), 613-621.

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

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

Mojica FJ, Díez-Villaseñor C, García-Martínez J, Soria E. (2005) Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol*., 60(2), 174-182.

Mojica FJ, Díez-Villaseñor C, García-Martínez J, Almendros C. (2009) Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology*, 155(3), 733–740.

Mohanraju P, Makarova KS, Zetsche B, Zhang F, Koonin EV, van der Oost J. (2016) Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. *Science*, 353(6299), aad5147.

Morgan M, Anders S, Lawrence M, Aboyoun P, Pagès H, Gentleman R. (2009) ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics*, 25(19), 2607–2608.

Mulepati S, Bailey S. (2013) *In vitro* reconstitution of an *Escherichia coli* RNA-guided immune system reveals unidirectional, ATP-dependent degradation of DNA target. *J. Biol. Chem*., 288(31), 22184-22192.

Musharova O, Klimuk E, Datsenko KA, Metlitskaya A, Logacheva M, Semenova E, Severinov K, Savitskaya E. (2017) Spacer-length DNA intermediates are associated with Cas1 in cells undergoing primed CRISPR adaptation. *Nucleic Acids Res*., 45(6), 3297-3307.

Nam KH, Haitjema C, Liu X, Ding F, Wang H, DeLisa MP, Ke A. (2012) Cas5d protein processes pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvulg CRISPR-Cas system. *Structure*, 20(9), 1574-1584.

Nuñez JK, Kranzusch PJ, Noeske J, Wright AV, Davies CW, Doudna JA. (2014). Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nat. Struct. Mol. Biol*., 21(6), 528–534.

Nuñez JK, Harrington LB, Kranzusch PJ, Engelman AN, Doudna JA. (2015a). Foreign DNA capture during CRISPR–Cas adaptive immunity. *Nature*, 527(7579), 535-538.

Nuñez JK, Lee AS, Engelman A, Doudna JA. (2015b). Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature*, 519(7542), 193–198.

Nuñez JK, Bai L, Harrington LB, Hinder TL, Doudna JA. (2016) CRISPR Immunological Memory Requires a Host Factor for Specificity. *Mol Cell*, 62(6), 824-833.

Osawa T, Inanaga H, Sato C, Numata T. (2015) Crystal structure of the CRISPR-Cas RNA silencing Cmr complex bound to a target analog. *Mol. Cell*, 58(3), 418–430.

Pages H, Aboyoun P, Gentleman R and DebRoy S. (2012) Biostrings: String objects representing biological sequences, and matching algorithms. R package version 2.24.1.

Park KH, An Y, Jung TY, Baek IY, Noh H, Ahn WC, Hebert H, Song JJ, Kim JH, Oh BH, Woo EJ. (2017) RNA activation-independent DNA targeting of the Type III CRISPR-Cas system by a Csm complex. *EMBO Rep*., 18(5), 826-840.

Pawluk A, Bondy-Denomy J, Cheung VH, Maxwell KL, Davidson AR. (2014) A new group of phage anti-CRISPR genes inhibits the type I-E CRISPR-Cas system of *Pseudomonas aeruginosa*. *MBio*, 5(2), e00896.

Pawluk A, Amrani N, Zhang Y, Garcia B, Hidalgo-Reyes Y, Lee J, Edraki A, Shah M, Sontheimer EJ, Maxwell KL, Davidson AR. (2016a) Naturally Occurring Off-Switches for CRISPR-Cas9. *Cell*, 167(7), 1829-1838.

Pawluk A, Staals RH, Taylor C, Watson BN, Saha S, Fineran PC, Maxwell KL, Davidson AR. (2016b) Inactivation of CRISPR-Cas systems by anti-CRISPR proteins in diverse bacterial species. *Nat. Microbiol*., 1(8), 16085.

Peng W, Feng M, Feng X, Liang YX, 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(1), 406-417.

Plagens A, Tjaden B, Hagemann A, Randau L, Hensel R. (2012) Characterization of the CRISPR/Cas subtype I-A system of the hyperthermophilic crenarchaeon *Thermoproteus tenax*. *J Bacteriol*., 194(10), 2491-2500.

Plagens A, Tripp V, Daume M, Sharma K, Klingl A, Hrle A, Conti E, Urlaub H, Randau L. (2014) *In vitro* assembly and activity of an archaeal CRISPR-Cas type I-A Cascade interference complex. *Nucleic Acids Res*., 42(8), 5125-5138.

Plagens A, Richter H, Charpentier E, Randau L. (2015) DNA and RNA interference mechanisms by CRISPR-Cas surveillance complexes. *FEMS Microbiol. Rev*., 39(3), 442-463.

Pourcel C, Salvignol G, 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(3), 653-663.

Pyenson NC, Gayvert K, Varble A, Elemento O, Marraffini LA. (2017) Broad Targeting Specificity during Bacterial Type III CRISPR-Cas Immunity Constrains Viral Escape. *Cell Host Microbe*, 22(3), 343-353.e3.

Ramia NF, Spilman M, Tang L, Shao Y, Elmore J, Hale C, Cocozaki A, Bhattacharya N, Terns RM, Terns MP, Li H, Stagg SM. (2014) Essential structural and functional roles of the

Cmr4 subunit in RNA cleavage by the Cmr CRISPR-Cas complex. *Cell Rep*., 9(5), 1610–1617.

Rao C, Guyard C, Pelaz C, Wasserscheid J, Bondy-Denomy J, Dewar K, Ensminger AW. (2016) Active and adaptive *Legionella* CRISPR-Cas reveals a recurrent challenge to the pathogen. *Cell Microbiol.*, 18(10), 1319-1338.

Rauch BJ, Silvis MR, Hultquist JF, Waters CS, McGregor MJ, Krogan NJ, Bondy-Denomy J. (2017) Inhibition of CRISPR-Cas9 with Bacteriophage Proteins. *Cell*, 168(1-2), 150-158.e10.

Redding S, Sternberg SH, Marshall M, Gibb B, Bhat P, Guegler CK, Wiedenheft B, Doudna JA, Greene EC. (2015) Surveillance and processing of foreign DNA by the *Escherichia coli* CRISPR-Cas system. *Cell*, 163(4), 854–865.

Richter C, Dy RL, McKenzie RE, Watson BN, Taylor C, Chang JT, McNeil MB, Staals RH, Fineran PC. (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(13), 8516–8526.

Richter H, Zoephel J, Schermuly J, Maticzka D, Backofen R, Randau L. (2012) Characterization of CRISPR RNA processing in *Clostridium thermocellum* and *Methanococcus maripaludis*. *Nucleic Acids Res*., 40(19), 9887-9896.

Rollie C, Schneider S, Brinkmann AS, Bolt EL, White MF. (2015) Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition. *Elife*, 4, e08716.

Rouillon C, Zhou M, Zhang J, Politis A, Beilsten-Edmands V, Cannone G, Graham S, Robinson CV, Spagnolo L, White MF. (2013) Structure of the CRISPR interference complex CSM reveals key similarities with cascade. *Mol. Cell*, 52(1), 124–134.

Rutkauskas M, Sinkunas T, Songailiene I, Tikhomirova MS, Siksnys V, Seidel R. (2015) Directional R-loop formation by the CRISPR-Cas surveillance complex cascade provides efficient off-target site rejection. *Cell Rep.*, 10(9), 1534–1543.

Samai P, Pyenson N, Jiang W, Goldberg GW, Hatoum-Aslan A, Marraffini LA. (2015) Co-transcriptional DNA and RNA Cleavage during Type III CRISPR-Cas Immunity. *Cell*, 161(5), 1164–1174.

Samson JE, Magadán AH, Sabri M, Moineau S. (2013) Revenge of the phages: Defeating bacterial defences. *Nat. Rev. Microbiol*., 11(10), 675–687.

Savitskaya E, Semenova E, Dedkov V, Metlitskaya A, Severinov K. (2013) High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in *E. coli*. *RNA Biol*., 10(5), 716-725.

Scholz I, Lange SJ, Hein S, Hess WR, 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(2), e56470.

Semenova E, Jore MM, Datsenko KA, Semenova A, Westra ER, Wanner B, van der Oost J, Brouns SJ, Severinov K. (2011) Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc. Natl. Acad. Sci. U S* A, 108(25), 10098-10103.

Semenova E, Savitskaya E, Musharova O, Strotskaya A, Vorontsova D, Datsenko KA, Logacheva MD, Severinov K. (2016) Highly efficient primed spacer acquisition from targets destroyed by the *Escherichia coli* type I-E CRISPR-Cas interfering complex. *Proc. Natl. Acad. Sci. U S A*, 113(27), 7626-7631.

Shao Y, Richter H, Sun S, Sharma K, Urlaub H, Randau L, Li H. (2016). A non-stem-loop CRISPR RNA is processed by dual binding Cas6. *Structure*, 24(4), 547-554.

Shipman SL, Nivala J, Macklis JD, Church GM. (2016) Molecular recordings by directed CRISPR spacer acquisition. *Science*, 353(6298), aaf1175.

Shmakov S, Savitskaya E, Semenova E, Logacheva MD, Datsenko KA, Severinov K. (2014) Pervasive generation of oppositely-oriented spacers during CRISPR adaptation. *Nucleic Acids Res*., 42(9), 5907–5916.

Shmakov S, Abudayyeh OO, Makarova KS, Wolf YI, Gootenberg JS, Semenova E, Minakhin L, Joung J, Konermann S, Severinov K, Zhang F, Koonin EV. (2015) Discovery and functional characterization of diverse class 2 CRISPR-Cas systems. *Mol. Cell*, 60(3), 385–397.

Silas S, Mohr G, Sidote DJ, Markham LM, Sanchez-Amat A, Bhaya D, Lambowitz AM, Fire AZ. (2016) Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. *Science*, 351(6276), aad4234.

Silas S, Makarova KS, Shmakov S, Páez-Espino D, Mohr G, Liu Y, Davison M, Roux S, Krishnamurthy SR, Fu BXH, Hansen LL, Wang D, Sullivan MB, Millard A, Clokie MR, Bhaya D, Lambowitz AM, Kyrpides NC, Koonin EV, Fire AZ. (2017) On the Origin of Reverse Transcriptase-Using CRISPR-Cas Systems and Their Hyperdiverse, Enigmatic Spacer Repertoires. *MBio.*, 8(4), e00897-17.

Sinkunas T, Gasiunas G, Fremaux C, Barrangou R, Horvath P, Siksnys V. (2011) Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *EMBO J*., 30(7), 1335–1342.

Sinkunas T, Gasiunas G, Waghmare SP, Dickman MJ, Barrangou R, Horvath P, Siksnys V. (2013). *In vitro* reconstitution of Cascade-mediated CRISPR immunity in *Streptococcus thermophilus*. *EMBO J*., 32(3), 385–394.

Smith DR, Doucette-Stamm LA, Deloughery C, Lee H, Dubois J, Aldredge T, Bashirzadeh R, Blakely D, Cook R, Gilbert K, Harrison D, Hoang L, Keagle P, Lumm W, Pothier B, Qiu D, Spadafora R, Vicaire R, Wang Y, Wierzbowski J, Gibson R, Jiwani N, Caruso A, Bush D, Reeve JN, et al. (1997) Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics. *J Bacteriol.*, 179(22), 7135–7155.

Song M. (2017) The CRISPR/Cas9 system: their delivery, *in vivo* and *ex vivo* applications and clinical development by startups. *Biotechnol Prog.*, Epub ahead of print.

Spilman M, Cocozaki A, Hale C, Shao Y, Ramia N, Terns R, Terns M, Li H, Stagg S. (2013) Structure of an RNA silencing complex of the CRISPR-Cas immune system. *Mol. Cell*, 52(1), 146–152.

Staals RH, Agari Y, Maki-Yonekura S, Zhu Y, Taylor DW, van Duijn E, Barendregt A, Vlot M, Koehorst JJ, Sakamoto K, Masuda A, Dohmae N, Schaap PJ, Doudna JA, Heck AJ, Yonekura K, van der Oost J, Shinkai A. (2013) Structure and activity of the RNA-targeting Type III-B CRISPR-Cas complex of *Thermus thermophilus*. *Mol. Cell*, 52(1), 135–145.

Staals RH, Zhu Y, Taylor DW, Kornfeld JE, Sharma K, Barendregt A, Koehorst JJ, Vlot M, Neupane N, Varossieau K, Sakamoto K, Suzuki T, Dohmae N, Yokoyama S, Schaap PJ, Urlaub H, Heck AJ, Nogales E, Doudna JA, Shinkai A, van der Oost J. (2014) RNA targeting by the type III-A CRISPR-Cas Csm complex of *Thermus thermophilus. Mol. Cell*, 56(4), 518–530.

Sternberg SH, Richter H, Charpentier E, Qimron U. (2016) Adaptation in CRISPR-Cas Systems. *Mol Cell*, 61(6), 797-808.

Strotskaya A, Savitskaya E, Metlitskaya A, Morozova N, Datsenko KA, Semenova E, Severinov K. (2017) The action of *Escherichia coli* CRISPR-Cas system on lytic bacteriophages with different lifestyles and development strategies. *Nucleic Acids Res*., 45(4), 1946-1957.

Swarts DC, Mosterd C, van Passel MW, Brouns SJ. (2012) CRISPR interference directs strand specific spacer acquisition. *PLoS One*, 7(4), e35888.

Swarts DC, Koehorst JJ, Westra ER, Schaap PJ, van der Oost J. (2015) Effects of Argonaute on Gene Expression in *Thermus thermophilus*. *PLoS One*, 10(4), e0124880.

Szczelkun MD, Tikhomirova MS, Sinkunas T, Gasiunas G, Karvelis T, Pschera P, Siksnys V, Seidel R. (2014) Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. *Proc. Natl. Acad. Sci. U S A*, 111(27), 9798-803.

Takeuchi N, Wolf YI, Makarova KS, Koonin EV. (2012) Nature and intensity of selection pressure on CRISPR-associated genes. *J. Bacteriol*., 194(5), 1216–1225.

Tamulaitis G, Kazlauskiene M, Manakova E, Venclovas Č, Nwokeoji AO, Dickman MJ, Horvath P, Siksnys V. (2014) Programmable RNA Shredding by the Type III-A CRISPR-Cas System of *Streptococcus thermophilus*. *Mol Cell*., 56(4), 506-517.

Tamulaitis G, Venclovas Č, Siksnys V. (2017) Type III CRISPR-Cas Immunity: Major Differences Brushed Aside. *Trends Microbiol*., 25(1), 49-61.

Tock MR, & Dryden DT. (2005) The biology of restriction and antirestriction. *Curr. Opin. Microbiol*., 8(4), 466–472.

Toro N, Nisa-Martínez R. (2014) Comprehensive phylogenetic analysis of bacterial reverse transcriptases. *PLoS One*, 9(11), e114083.

Touchon M, and Rocha EPC. (2010) The small, slow and specialized CRISPR and anti-CRISPR of *Escherichia* and *Salmonella*. *PLoS ONE*, 5(6), e11126.

van der Oost J, Jore MM, Westra ER, Lundgren M, Brouns SJ. (2009) CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends Biochem Sci*., 34(8), 401-407.

Vestergaard G, Garrett RA, Shah SA. (2014) CRISPR adaptive immune systems of Archaea. *RNA Biol*., 11(2), 156-167.

Vorontsova D, Datsenko KA, Medvedeva S, Bondy-Denomy J, Savitskaya EE, Pougach K, Logacheva M, Wiedenheft B, Davidson AR, Severinov K, Semenova E. (2015) Foreign DNA acquisition by the I-F CRISPR-Cas system requires all components of the interference machinery. *Nucleic Acids Res*., 43(22), 10848–10860.

Walker FC, Chou-Zheng L, Dunkle JA, Hatoum-Aslan A. (2017) Molecular determinants for CRISPR RNA maturation in the Cas10–Csm complex and roles for non-Cas nucleases. *Nucleic Acids Res.*, 45 (4), 2112-2123.

Wang J, Li J, Zhao H, Sheng G, Wang M, Yin M, Wang Y. (2015) Structural and mechanistic basis of PAM-dependent spacer acquisition in CRISPR-Cas systems. *Cell*, 163(4), 840–853.

Wang X, Yao D, Xu JG, Li AR, Xu J, Fu P, Zhou Y, Zhu Y. (2016) Structural basis of Cas3 inhibition by the bacteriophage protein AcrF3. *Nat. Struct. Mol. Biol*., 23(9), 868-870.

Wei Y, Terns RM, Terns MP. (2015) Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. *Genes Dev*., 29(4), 356–361.

Westra ER, van Erp PB, Kunne T, Wong SP, Staals RH, Seegers CL, Bollen S, Jore MM, Semenova E, Severinov K, de Vos WM, Dame RT, de Vries R, Brouns SJ, van der Oost J.

(2012) CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. *Mol. Cell*, 46(5), 595–605.

Westra ER, van Houte S, Oyesiku-Blakemore S, Makin B, Broniewski JM, Best A, Bondy-Denomy J, Davidson A, Boots M, Buckling A. (2015). Parasite exposure drives selective evolution of constitutive versus inducible defense. *Current Biology*, 25(8), 1043-1049.

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

Wiedenheft B, van Duijn E, Bultema JB, Waghmare SP, Zhou K, Barendregt A, Westphal W, Heck AJ, Boekema EJ, Dickman MJ, Doudna JA. (2011a) RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proc. Natl. Acad. Sci. U S A*, 108(25), 10092-10097.

Wiedenheft B, Lander GC, Zhou K, Jore MM, Brouns SJ, van der Oost J, Doudna JA, Nogales E. (2011b). Structures of the RNA-guided surveillance complex from a bacterial immune system. *Nature*, *477*(7365), 486-489.

Xue C, Whitis NR, Sashital DG. (2016) Conformational Control of Cascade Interference and Priming Activities in CRISPR Immunity. *Mol. Cell*, 64(4), 826-834.

Yosef I, Goren MG, Qimron U. (2012) Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli. Nucleic Acids Res*., 40(12), 5569–5576.

Zhang J, Kasciukovic T, White MF. (2012) The CRISPR associated protein Cas4 is a 5' to 3' DNA exonuclease with an iron-sulfur cluster. *PLoS One*, 7(10), e47232.

Zhang Y, Heidrich N, Ampattu BJ, Gunderson CW, Seifert HS, Schoen C, Vogel J, Sontheimer EJ. (2013) Processing-independent CRISPR RNAs limit natural transformation in *Neisseria meningitidis*. *Mol. Cell*, 50(4), 488–503.

Galindo EB. (2007) Desarrollo de nuevas herramientas y protocolos de selección de proteínas termoestables en *Thermus thermophilus.* Thesis Doctoral, 73.