Skolkovo Institute of Science and Technology

CHARACTERIZATION AND APPLICATION OF CRISPR-Cas ENZYMES

Doctoral Thesis

by

IANA FEDOROVA

DOCTORAL PROGRAM IN LIFE SCIENCES

Supervisor
Professor Konstantin Severinov

Moscow – 2020

I hereby declare that the work presented in this thesis was carried out by myself at Skolkovo Institute of Science and Technology, Moscow, except where due acknowledgement is made, and has not been submitted for any other degree.

Candidate (Iana Fedorova)

Supervisor (Prof. Konstantin Severinov)

TABLE OF CONTENTS

ABSTRACT

To maintain the integrity of the genome and to avoid being killed by viruses, prokaryotes developed diverse specialized defense mechanisms. Among these mechanisms are the CRISPR-Cas adaptive immunity systems, which were discovered relatively recently. Their action is based on Cas nucleases, enzymes, which in complex with guide CRISPR RNAs (crRNAs), can specifically recognize invader nucleic acid and cleave it, preventing the further steps of infection. The ability of CRISPR-Cas ribonucleoprotein complexes to target distinct DNA sites using guide crRNAs of different sequences led to development of powerful biotechnology instruments. The SpCas9 nuclease from *Streptococcus pyogenes*, the first Cas nuclease used for genome editing in human cells, remains the best characterized and the most used Cas enzyme to date. However, numerous CRSPR-Cas systems can be found in different bacterial and archaeal species.

Although all CRISPR-Cas systems rely on similar principles of nucleic acids recognition, they are incredibly diverse. Most of these systems are not biochemically characterized to date. Studying of CRISPR-Cas effectors of different types, as well as Cas orthologs from various bacterial and archaeal species, allows one to develop new biotechnology instruments and advance the understanding of prokaryotic defense mechanisms.

This thesis is devoted to studies of CRISPR-Cas systems, the biochemical characterization of different Cas effectors and, finally, their application in genome editing.

Chapters 1 and 2 describe the biochemical characterization of several Type II-C Cas9 orthologues. These enzymes originated from *Pasteurella pneumotropica*, a gram-negative bacterium isolated from multiple mammalian species, *Clostridium celluloliticum*, a promising biofuel producer isolated from compost, and *Defluviimonas sp. 20V17*, a bacterium inhabiting deep sea hydrothermal vents. We show that the effector Cas nucleases PpCas9, CcCas9 and DfCas9 have different properties, require novel PAM (protospacer adjacent motif) sequences for efficient DNA cleavage and have smaller sizes than SpCas9. This makes them attractive candidates for the development of new biotechnology tools. In particular, we show that PpCas9 efficiently cleaves genomic DNA in human cells and, hence, can be used for genome modification in mammals.

Chapter 3 and Chapter 4 are devoted to studies of Type V-A CRISPR-Cas systems effector Cas12a (former Cpf1). In Chapter 3 we show that the Cas12a structure is distinct from that of Cas9 and propose a model of Cas12a DNA cleavage mechanism. The distinct organization of Cas12a enzymes confers unique properties compared with Cas9. In Chapter 4 we show that AsCas12a from *Acidominococcus sp.* and LbCas12a from *Lachnospiraceae* can process their pre-crRNA into mature crRNAs targeting different DNA sites without the help from any other enzymes. We used this property of Cas12a to create a multiplex gene editing system which allows one to simultaneously modify several genomic sites in human cells using a single CRISPR array.

Chapter 5 is devoted to Cas12e effectors (formerly CasX), members of the Type V-E CRISPR-Cas enzymes. Despite low sequence similarity between Cas12e and Cas12a proteins, their domain organization is quite similar, which results in common mechanisms of DNA cleavage. Due to this, both nucleases generate staggered ends at the DNA cut site, in contrast to Cas9, which generates blunt ends. We used high throughput sequencing to precisely map the cut site positions of DpbCas12e from *Deltaproteobacteria* and compared its DNA cleavage pattern to that of AsCas12a. We found that these enzymes cleave DNA in a highly similar manner and that the length of DNA overhangs generated by Cas12e can be increased using guide RNAs with shorter spacer segments.

Along with the characterization of the effector nucleases, it is important to study the overall mechanism of CRISPR-Cas systems action. The mechanism of acquisition of new spacers into the CRISPR array remains underinvestigated to date. The main intriguing questions is how DNA fragments from invader nucleic acids become spacers in the CRISPR array. In Chapter 6 we used HTS-based method FragSeq to define precursors of CRISPR array spacers – DNA fragments, which are produced in the course of CRISPR adaptation – after selection from foreign DNA but before integration into the array.

ACKNOWLEDGEMENTS

First of all, I would like to express my deepest appreciation to my mentor Konstantin Severinov for the guidance and encouragement he has provided throughout my time as his student. Konstantin will always be an example for me on how to see the beauty of science, share passion and energy with others, even when you need to struggle with challenges to do good science in our country. I have been extremely lucky to have a supervisor who devoted a lot of time for me at the beginning of my PhD and gave a lot of independence at the end. One day Konstantin said that I can be as a lone sailboat in the sea of science to learn how to sail. And he gave me opportunities to learn how to write grants, papers and even organize work in the lab. And I understood how is it difficult but exciting.

I would like to extend my deepest gratitude to Mikhail Khodorkovskii, the Director of Nanobio center in Saint Petersburg Peter the Great Polytechnic University where I conducted most of my research. Without Mikhail's support and guidance, this work wouldn't be done. He helped me a lot on this long journey.

I'm deeply indebted to Feng Zhang and members of the Zhang lab for giving me an opportunity to work with them at the Broad Institute on the characterization of Cas12a proteins. They showed me all the technics necessary for CRISPR-Cas nuclease characterization and demonstrated how fast, efficient, and exciting science could be and that we should try to find an application of fundamental research where it is possible. My special thanks to Bernd Zetsche, who was teaching me how to work with eukaryotic cells and how to use CRISPR-Cas for genome editing. I am also thankful to Sourav R. Choudhury who shared his great experience of AAV purification.

I am grateful to our small Saint Petersburg research group, people with whom we side by side were working last three years on the characterization of small Cas9 orthologs: Polina Selkova, Aleksandra Vasileva, Georgii Pobegalov, and Anatolii Arseniev, as well as Olga Musharova, my friend and colleague from Moscow. Also, Tatyana Artamonova's help cannot be overestimated. Thanks to her for mass spectrometry analysis of numerous recombinant proteins which were purified during this project. Many thanks to Anna Shiriaeva for advice on data analysis. I very much appreciate conversations with Sergey Shmakov and his help with the bioinformatic search for *cas* genes. I also would like to acknowledge the help of Lidia Rakcheeva who was bravely doing a lot of paperwork for our projects. Without Lidia writing grant reports would be much difficult. I am grateful to Mary Sokolova for her friendship and support all these years.

Special thanks to the students whom I am mentoring: Max Kazalov, Marina Abramova, Nataliia Shcheglova, and Irina Francuzova, who wanted to join me and with whom I learned a lot. Thanks also to all members of Severinov's lab and Nanobio center, for their help and creating of warm atmosphere at the bench.

I am extremely grateful to my family: my parents, my husband Sasha, and our son Danya who was born during my work on this Thesis. You all are supporting me and encourage me to try to do good science, to think and to create, try to be an "artist". Dear family, your patience cannot be underestimated.

I also want to acknowledge all members of my Individual Thesis Committee for fruitful discussions during annual reviews.

And last but not least I would like to thank Skoltech and people who created Skoltech community. Freedom and good science, world standards of education – this makes Skoltech a perfect place to study. I hope that Skoltech will be developing further and more such universities will appear in Russia.

PUBLICATIONS

1. **Fedorova, I.***; Arseniev, A.*; Selkova, P.; Pobegalov, G.; Goryanin, I.; Vasileva, A.; Musharova, O.; Abramova, M.; Kazalov, M.; Zyubko, T.; Artamonova, T.; Artamonova, D.; Shmakov, S.; Khodorkovskii, M.; Severinov, K. DNA Targeting by Clostridium Cellulolyticum CRISPR-Cas9 Type II-C System. *Nucleic Acids Res.* 2020, *48* (4), 2026–2034. https://doi.org/10.1093/nar/gkz1225.

2. **Fedorova, I.***; Vasileva,A.*; Selkova,P.; Abramova,M.; Arseniev, A.; Pobegalov, G.; Kazalov, M.; Musharova, O.; Goryanin, I.; Artamonova, D.; Zyubko, T.; Shmakov, S.; Artamonova, T.; Khodorkovskii, M.; Severinov, K. PpCas9 from *Pasteurella pneumotropica* - a compact Type II-C Cas9 ortholog active in human cells. *Nucleic Acids Res.* 2020 **SUBMITTED**

3. Yamano, T.; Nishimasu, H.; Zetsche, B.; Hirano, H.; Slaymaker, I. M.; Li, Y.; **Fedorova, I.;** Nakane, T.; Makarova, K. S.; Koonin, E. V.; Ishitani, R.; Zhang, F.; Nureki, O. Crystal Structure of Cpf1 in Complex with Guide RNA and Target DNA. *Cell* 2016, *165* (4), 949–962. https://doi.org/10.1016/j.cell.2016.04.003.

4. Zetsche, B.; Heidenreich, M.; Mohanraju, P.; **Fedorova, I.**; Kneppers, J.; DeGennaro, E. M.; Winblad, N.; Choudhury, S. R.; Abudayyeh, O. O.; Gootenberg, J. S.; Wu, W. Y.; Scott, D. A.; Severinov, K.; van der Oost, J.; Zhang, F. Multiplex Gene Editing by CRISPR-Cpf1 Using a Single crRNA Array. *Nat. Biotechnol.* 2017, *35* (1), 31–34. https://doi.org/10.1038/nbt.3737.

5. Selkova,P.; Vasileva,A.; Pobegalov, G.; Musharova, O.; Arseniev, A.; Kazalov, M.; Zyubko, T.; Shcheglova, N.; Artamonova, T.; Khodorkovskii, M.; Severinov, K. and **Fedorova, I.** Position of Deltaproteobacteria Cas12e nuclease cleavage sites depends on spacer length of guide RNA. *RNA biology* 2020 https://doi.org/10.1080/15476286.2020.1777378.

6. Shiriaeva, A. A.; Savitskaya, E.; Datsenko, K. A.; Vvedenskaya, I. O.; **Fedorova, I.;** Morozova, N.; Metlitskaya, A.; Sabantsev, A.; Nickels, B. E.; Severinov, K.; Semenova, E. Detection of Spacer Precursors Formed in Vivo during Primed CRISPR Adaptation. *Nat Commun* 2019, *10* (1), 4603. https://doi.org/10.1038/s41467-019-12417-w.

*equal contribution

CONFERENCES

1. **Fedorova I**, Zetsche B, Severinov K, Zhang F. Cpf1: multiplex gene editing. Skoltech & MIT Conference: “Shaping the Future: Big Data, Biomedicine and Frontier Technologies”. Russia, Moscow. 25-26 April, 2017

2. **Fedorova Iana**. Studying of CRISPR-Cas Type II-C systems. Conference: “Biotech Club”. Russia, Moscow. 26 October, 2018

3. **Iana Fedorova**, George Pobegalov, Anatoliy Arseniev, Aleksandra Vasileva, Polina Selkova, Olga Musharova, Ignat Goryanin, Darya Artamonova, Tatyana Zubko, Sergey Shmakov, Konstantin Severinov. Cas9 diversity: characterization of Type II-C PpCas9, CcCas9, DfCas9 and DsCas9 nucleases. Conference:“CRISPR 2019”. Canada, Quebec. 17-20 June, 2019

PATENTS

1. DNA cleavage instrument based on CcCas9 protein from bacterium *Clostridium celluloliticum*, Russia, №RU2712497, 29.02.2020

2. DNA cleavage instrument based on DfCas9 protein from bacterium *Defluviimonas sp*, Russia, №RU2712492, 29.02.2020

3. DNA cleavage instrument based on PpCas9 protein from bacterium *Pasteurella pneumotropica,* Russia, application № 2019118061 11.06.2019 (patent granted but not published yet)

4. DNA cleavage instrument based on DsCas9 protein from bacterium *Demequina sediminicola,* Russia, application № 2019118066 11.06.2019 (patent granted but not published yet)

5. Using of PpCas9 protein from bacterium *Pasteurella pneumotropica* for modification of genomic DNA *in vivo,* Russia, application № 2019136264 11.11.2019 (patent granted but not published yet)

LIST OF SYMBOLS, ABBREVIATIONS

CRISPR – Clustered Regularly Interspaced Short Palindromic Repeats

Cas – CRISPR-associated

DNA – deoxyribonucleic acid

RNA – ribonucleic acid

crRNA – CRISPR RNA

dsDNA – double stranded DNA

ssDNA – single stranded DNA

nt – nucleotides

bp – base pairs

aa – amino acids

HTS – High Throughput Sequencing

PCR – polymerase chain reaction

tracrRNA – trans – activating CRISPR RNA

TSL domain – Target Strands Loading domain

BREX system – Bacteriophage Exclusion system

R-M system – Restriction- Modification system

DR – Direct Repeat

TALEN – Transcription Activator-Like Effector Nuclease

ZNF – Zinc Finger Nuclease

HDR – Homology Directed Repair

NHEJ – Non-Homologous End Joining

TS – Target Strand

NTS – Non Target Strand

INTRODUCTION

1. CRISPR-Cas systems as one of the bacterial defense mechanisms

Similarly to eukaryotes, bacterial and archaeal cells need to protect themselves from numerous mobile genetic elements with which they live side by side: plasmids, transposons and viruses of prokaryotes - bacteriophages (1). The latter in particular are widely spread and appear to be the most abundant group of organisms on Earth (2–4). Bacteriophage genetic diversity is extremely high. They rapidly evolve and acquire new properties allowing fast and efficient infection of and replication in cellular life forms. The "arms race" between mobile genetic elements and prokaryotes leads to incessant evolution of both: the infection mechanisms of invaders and the defense systems of bacteria and archaea are constantly honed and improved.

The evolution of prokaryotic immune systems takes place due to accumulation of mutations as well as through extensive gene duplication and horizontal gene transfer (5). As a consequence, protein sequences of defense genes are extremely diverse, which makes their identification by bioinformatics a non-trivial task. Nevertheless, development of bioinformatic search approaches have led to a steady stream of new predicted bacterial and archaeal systems (6–9). These searches are facilitated by the fact that genes coding for defense systems tend to cluster into so-called defense islands in prokaryotic genomes (10). The defense islands often carry several defense systems of different types, with different mechanisms of action. As a result, the cell provided with a versatile multi-pronged arsenal of protective agents.

Most of antiviral prokaryotic defense systems known to date act either through programmable cell suicide in response to the infection or through self/nonself-discrimination, when the invader's genome is degraded or modified to prevent its replication, while the host genomic DNA remains intact.

The first group includes toxin-antitoxin systems and abortive infection systems. In brief, their work is based on the balance of an intercellular toxic element (typically, a protein) and its neutralizer (a protein or an RNA)(11–13). Stress, such as bacteriophage infection, imbalances the system and unleashes the toxin, leading to cell lysis or dormancy.

The second group includes restriction-modification systems (R-M), BREX, prokaryotic Argonaute systems and CRISPR-Cas systems, the latter being in the focus of this thesis (13–15). All of these systems rely on recognition of invader's genome and distinguishing it from the host DNA. While the BREX mechanism of protection remains unclear, the rest of these defense systems rely on nuclease cleavage of the invader's genome for its subsequent degradation or gene silencing through degradation of foreign RNA transcripts (13–15). The mechanism of invader recognition and self/nonself-discrimination is one of the most intriguing questions when studying an immune

system. The specific recognition of invader genome can be accomplished through the binding of a protein or a nucleic acid. The latter usually relies on complementarity between nucleic acids, which plays a central role in the transfer of information in living organisms, and thus, seems to be a natural and logical principle for specific recognition of foreign genomes (7). Strikingly, bacterial defense mechanisms recognizing invaders through this principles, the CRISPR-Cas and prokaryotic Argounate systems, were found only recently, while R-M systems, which identify parasitic DNA through protein binding, have been known for about seventy years (16, 17).

CRISPR-Cas systems initially were noticed by Yoshizumi Ishino and colleagues in 1987 when they were cloning the *iap* gene of *E. coli* and unintentionally captured a part of what we now know to be a Type I-E CRISPR-Cas system array adjacent to the target gene (18). They observed several repeats interspaced by unique spacer sequences, which formed a cassette. The direct repeats (DR) were mistaken for regulatory elements of the *iap* gene. Although other scientific groups continued to detect similar arrays in bacterial and archaeal genomes (19, 20), only in 2005 it was shown that some of the spacer sequences match to fragments of mobile genetic elements, that infect the host organism (21–23). The fact that spacer sequences take their origin from mobile genetic elements gave raise to the idea of their involvement in adaptive prokaryotic immunity. At the same time the mechanism of CRISPR-Cas immunity based on nucleic acids complementary pairing was proposed (22–24) and experimentally confirmed in different prokaryotic organisms (25–27).

In the next decade CRISPR-Cas systems were intensively studied. The increased interest in this defense mechanism was due to the demonstrated application of CRISPR-Cas systems in genome editing (28). As a result, to date, the basic principles of CRISPR-Cas systems function are well known, although many interesting details remain to be unveiled.

2. CRISPR-Cas mechanism of action

The CRISPR-Cas immune response is based on the action of RNA-guided Cas nucleases (CRISPR-associated nucleases). The RNA-component of defense system is encoded by the CRISPR array and the nuclease component(s) is encoded by the *cas* genes. CRISPR arrays are generally adjacent to the *cas* genes clusters and together form CRISPR-Cas loci (Figure 1).

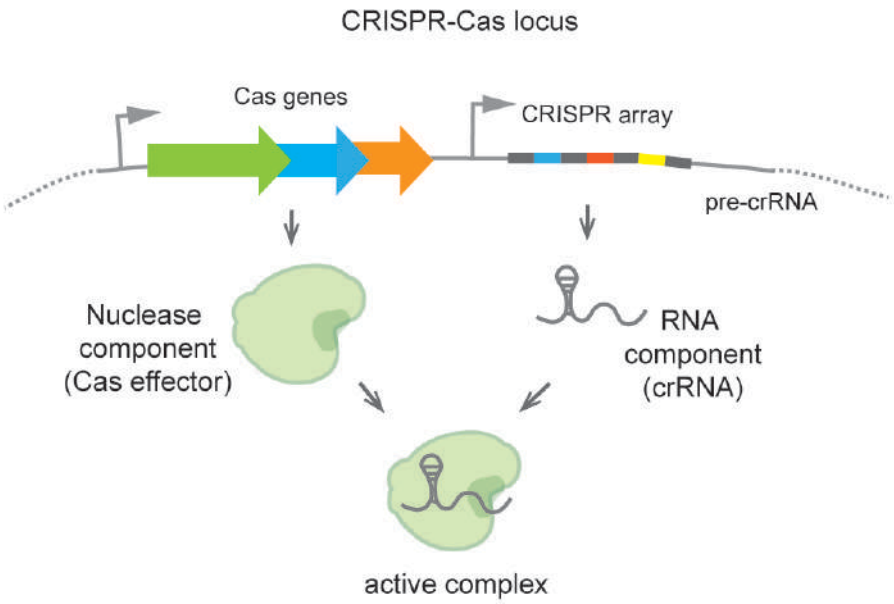


Figure1. CRISPR-Cas loci organization.
The nuclease component of the defense system is encoded by one or several *cas* genes and the RNA component is encoded by the CRISPR array. Binding of the nuclease and RNA components results in the formation of RNA-guided nuclease complex (effector), which defends the prokaryotic cell from mobile genetic elements.

The CRISPR-Cas immune response takes place in three stages: adaptation, expression (crRNA biogenesis) and interference.

2.1. Adaptation

Among prokaryotic defense systems, CRISPR-Cas are the only ones that function via a *bona fide* adaptive immunity mechanism. Living side by side with numerous infectious agents necessitates not only fast and efficient, but also selective and targeted protection from mobile genetic elements, to prevent wasting of cellular resources on threats that are nor relevant here and now. This is achieved through an elegant and sophisticated storage of information about the invaders in CRISPR arrays, the hallmarks of CRISPR-Cas systems.

CRISPR arrays are genomic regions consisting of direct repeats (DR) of 25-35 bp interspaced by unique sequences, spacers, of the similar length. Often DRs are palindromic and fold into hairpins when transcribed in RNA (29, 30). As was noted earlier, many spacers are derived from the genomes of invaders infecting the cell. Thus, CRISPR arrays constitute a memory of past and/or current acts of genetic aggressions.

This “recording tape” in the CRISPR array is continuously updated through the addition of new spacers (25, 31, 32). The process of acquisition of new spacers into CRISPR array is called

"CRISPR adaptation". This process is carried out predominantly by two enzymes, Cas1 and Cas2, with, depending on the CRISPR-Cas system type, the help from other Cas proteins (33–35). Following the invasion by a mobile genetic element of the cell, its genome can occasionally can be shredded into small fragments by endogenous nucleases (36). These small fragments, pre-spacers, are processed further and bind by the heterohexameric Cas1-Cas2 complex. The details of this process are not well understood to date. Chapter 6 of this work describes determination of pre-spacers intermediates produced in *Escherichia coli* during the adaptation process.

The Cas1-Cas2 complex incorporates pre-spacers into CRISPR array (37) (Figure 2).

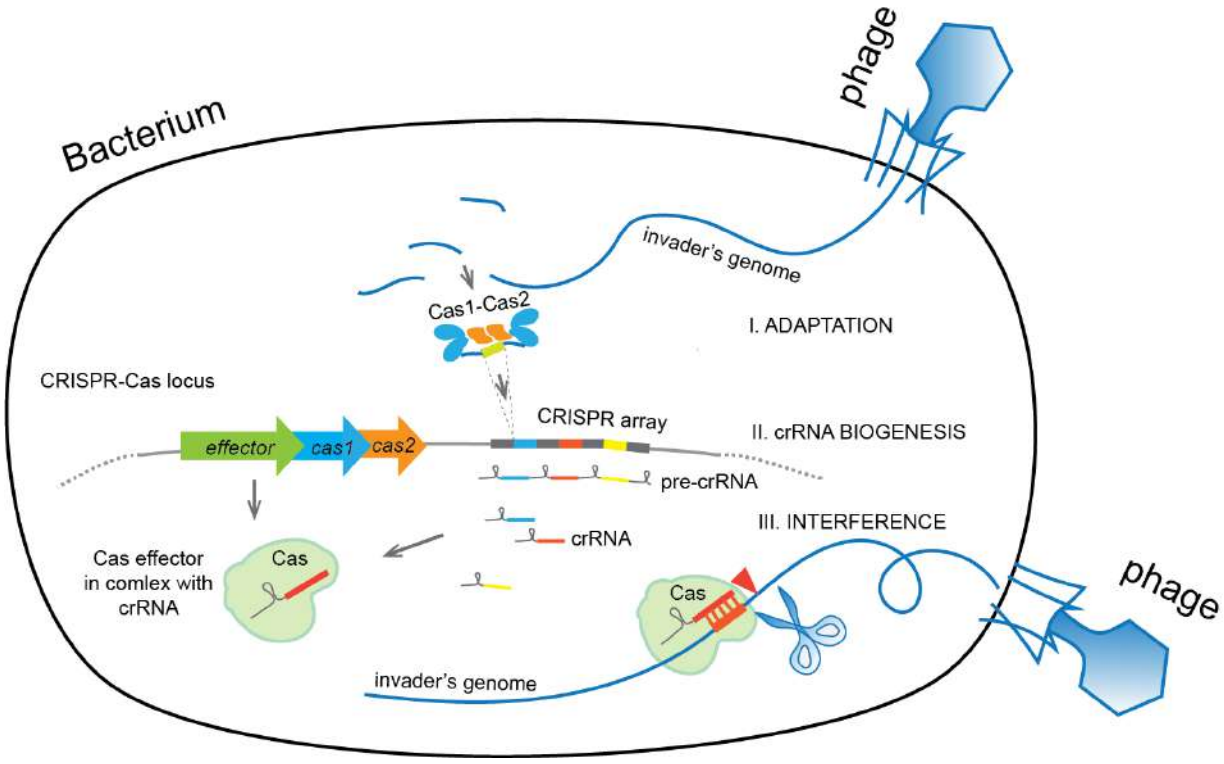


Figure 2. A scheme illustrating the CRISPR-Cas defense process.

Spacer incorporation occurs at the beginning of CRISPR array, near the so-called "leader sequence", an AT-rich region, which is in most types of CRISPR-Cas system is necessary for transcription of the array into a long pre-crRNA (38). The leader-proximal direct repeat is duplicated during new spacer incorporation (33). Some RNA-targeting CRISPR-Cas systems acquire spacers from invader transcripts through reverse transcription by a reverse-transcriptase encoded in the CRISPR-Cas locus and often fused to Cas1 (39). As a result of spacer incorporation, the beginning of the CRISPR array contains spacers corresponding bacteriophages or other invaders, which were encountered recently and which are likely to be encountered again. Thus, CRISPR array is a "recording tape", storing the information about the previous cell infections, and, moreover, about the order of infectious events in a certain time span.

2.2. Expression of CRISPR array and crRNA biogenesis

The CRISPR array is typically transcribed into a long precursor (pre-crRNA). The transcription starts from a promoter located within the leader sequence. Next, pre-crRNA is processed into short mature crRNAs, the main agents providing recognition of specific sites in invader's genome through complementary base pairing (40, 41) (Figure 2). In different CRISPR-Cas variants the pre-cRNA processing is conducted by *i*) endogenous host nucleases with a help from additional transacting RNA, *ii*) multisubunit Cas effector, or *iii*) single-subunit Cas effector. The last case is one of the topics of this thesis and in described in Chapter 4.

Mature crRNAs consist of conserved parts derived from flanking DRs and a variable guide part, derived from spacers. The spacer part of crRNA is 17-35 nt long and can be complementary paired with a protospacer in the invader's genome. In different types of CRISPR-Cas systems the DR part of crRNA can be located at 5'-, 3'-end or both ends of the spacer part. The conserved DR motifs, in particular their secondary structure, allow crRNA to bind to Cas proteins to form a ribonucleoprotein complex (42). This complex may include only one protein, the Cas effector nuclease, or consist of several Cas proteins each of which performs a certain function. The crRNA-Cas effector complex performs the last step of immune response – degradation of mobile genetic element - in a process, called "CRISPR-Cas interference".

2.3. CRISPR-Cas interference

The crRNA-Cas effector complex is searching for DNA targets by three-dimensional diffusion (43–45) and binds to protospacers through complementary base pairing between with crRNA spacer segment. This binding leads to subsequent cleavage of the target (Figure 2).
For efficient recognition and nucleic acid cleavage DNA targeting crRNA-Cas effector complexes require the presence of a 1-8 nucleotide long protospacer adjacent motif (PAM), flanking the target protospacer (46). This sequence serves for distinguishing self from non-self DNA: the absence of PAM near spacers in CRISPR array prevents autoimmunity and cleavage of the host genome.

PAM recognition is the first step in protospacer cleavage by the CRISPR-Cas machinery. Specific Cas proteins or domains (in case of a single-subunit effectors) interact with the PAM sequence of the invader genome. Next, Cas proteins unwind the dsDNA fragment adjacent to PAM, allowing complementary pairing between crRNA and the target (47, 48). This hybridization produces an R-loop structure - crRNA-target DNA strand hybrid with a displaced non target DNA strand (49). The fulfillment of all these conditions leads to interference – the cleavage of dsDNA by the effector complex.

The cleaved invader's nucleic acid is undergoing subsequent degradation by endogenous nucleases and, in addition, can serve as a source of new spacers (50). The diversity of spacers targeting different sites of a certain mobile genetic element ensures efficient protection of the host. Thus, this Cas proteins-mediated degradation of the invader not only prevents an ongoing infection, but also provides immunity against future invasions of similar infecting agents through the interference-adaptation coupling.

The understanding of the principles of CRISPR-Cas interference gave rise to an idea of applying Cas enzymes for genome editing in eukaryotic organisms: using guide crRNAs of different sequences allows programmable targeting of Cas effectors to any desirable site in the genome and introduction of DNA double stranded breaks, which can lead to genome modification (28, 51, 52). The application of Cas enzymes in genome editing will be discussed further.

3. CRISPR-Cas systems diversity and classification

CRISPR-Cas systems are remarkably diverse and widely distributed in bacteria and archaea. In 2019 K. Makarova and colleagues analyses 13,116 complete bacterial and archaeal genomes and concluded that CRISPR-Cas systems are present in 85.5 % of the archaea and in 40 % of bacteria (53). These systems are extremely variable in terms of genomic loci architecture, *cas* genes composition and their sequences, functions of Cas proteins and even their origin. Although a significant number of CRISPR-Cas systems was identified through bioinformatics searches and some of them were biochemically characterized, every year the analysis of genomic and metagenomic data reveals new members of CRISPR-Cas family (8, 9, 54, 55).

High diversity of CRISPR-Cas systems as well as the need to keep a track of new variants requires their rational classification and *cas* genes nomenclature (56). Since there are no universal *cas* genes and the frequent shuffling of the adaptation and effector protein modules during evolution of CRISPR-Cas systems takes place, the current classification is based on combined information on the presence of signature *cas* genes specific for a certain type of CRISPR-Cas system, organization of the loci, and phylogeny of Cas1, the most conserved Cas proteins (56).

The recent classification, which was updated in 2019, divides all CRISPR-Cas systems into two major classes (53). Class 1 comprises systems, where interference is mediated by a large multisubunit complex of Cas proteins, with each subunit performing a certain function. Class 2, includes CRISPR-Cas systems, where all activities of target recognition and cleavage are confined to a single effector Cas protein of complex multidomain organization (Figure 3). The major part of this thesis is dedicated to Class 2 Cas nucleases, and they will be discussed in more detail.

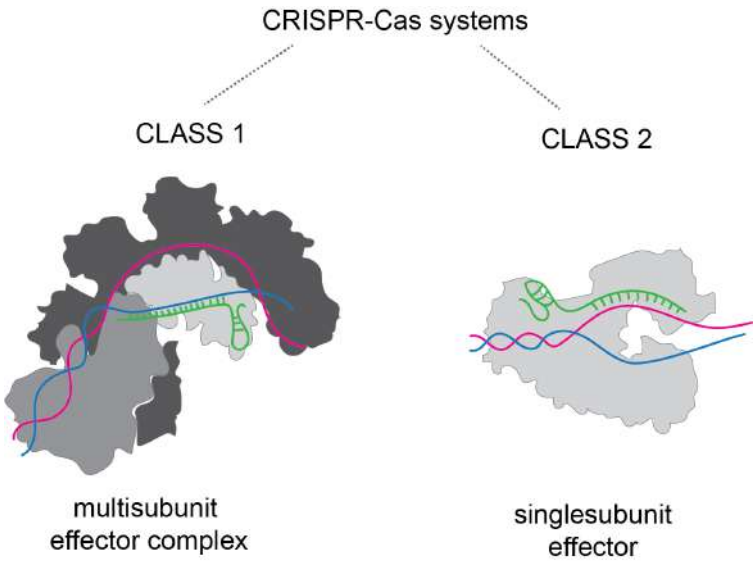


Figure 3. Two Classes of CRISPR-Cas systems.
Class 1 includes systems, where interference is mediated by a large effector complex consisting of several Cas proteins; Class 2 includes systems where the role of effector is played by a single protein.

CRISPR-Cas systems of each Class are additionally divided into three types: Class 1 includes types I, III, and IV; Class 2 – types II, V, and VI (Figure 4). Systems from Class 1 can be further subdivided into 16 subtypes and from Class 2 - into 17 subtypes.

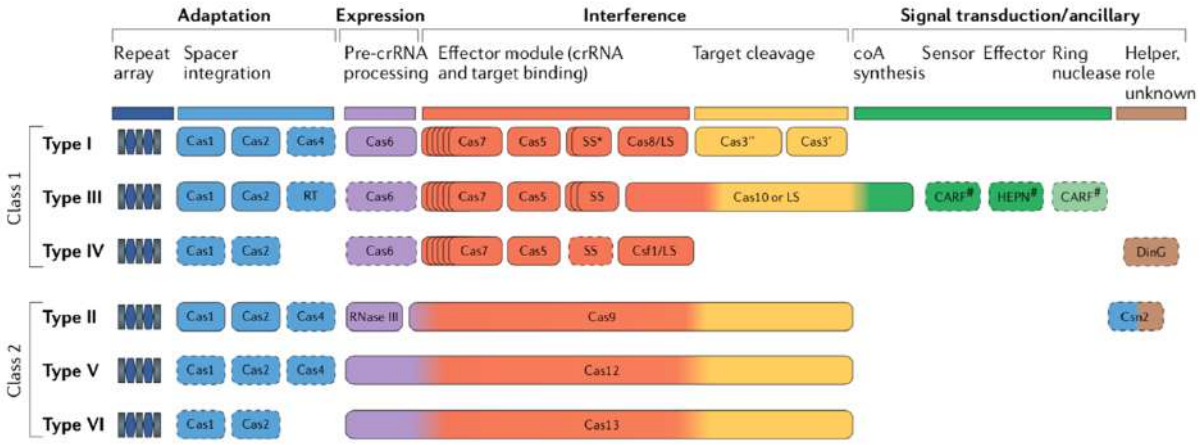


Figure 4. Modular organization of CRISPR-Cas systems and their classification. *Reproduced with permission from (53)*

Cas proteins belonging to any CRISPR-Cas system can be grouped into modules based of their functions (53, 56). Three major modules are: the adaptation module, the expression module, and the interference, or effector module (Figure 4). Although this division into modules is an approximation, because some Cas proteins, especially of Class 2, participate in all stages of defense, it provides an insight into the architecture of the systems of different subtypes (53).

The organization of the systems of Class 2 is simpler than that of Class 1: the effector module, which responsible for DNA or RNA cleavage, consists of only one polypeptide. The effector nucleases related to Class 2 (Cas9, Cas12, and Cas13 families of proteins) integrate all of the functions required for the interference, such as PAM recognition (or PFS, PAM analogue for Cas13), target nucleic acid unwinding, R-loop formation and, finally, nuclease activity.

The integration of all these functions into a single large protein made the Cas enzymes of Class 2 the perfect instruments for programmable nucleic acids cleavage and modification in eukaryotic cells. These systems, with interference activity located in one protein, are easy to reconstruct in human cells: they are more likely to form a functioning recombinant complex compared with the complicated and bulky Class 1 effector modules. Properties of the Cas effectors related to different Types of Class 2 CRISPR-Cas systems, as well as their application in genome engineering will be discussed further.

4. CRISPR-Cas systems of Class 2 and their applications in biotechnology

4.1. Type II CRISPR-Cas systems

Class 2 Type II CRISPR-Cas systems protect prokaryotic hosts from DNA invaders. Their effector module is represented by DNA-cleaving Cas9 proteins. For DNA recognition Cas9 effectors use two RNAs: a crRNA which mediates the complementary pairing with the DNA target site and tracrRNA, an axillary RNA necessary for crRNA processing and proper folding of the RNA-protein DNA cleavage complex (51, 52, 57) (Figure 5).

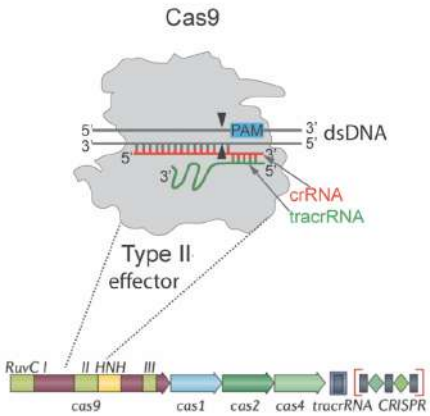


Figure 5. A scheme of a Type II CRISPR-Cas locus and the Cas9 effector complex with target DNA.

Type II loci consist of the effector module, represented by *cas9* gene, the adaptation module, represented by *cas1, cas2* and sometimes *cas4* or *csn2* genes, a CRISPR array and a sequence

coding for tracrRNA. The active ribonucleoprotein effector complex consisting of Cas9, tracrRNA and crRNA is shown above. *Reproduced with permission from (53) with modifications.*

TracrRNA is partially hybridized with conserved 3'- part of crRNA derived from DR, forming a heteroduplex. The DR fragment of crRNA paired with tracrRNA plays a role of a scaffold necessary for forming of an active DNA cleavage complex by the effector (51, 58) (Figure 5).

The 20-25nt spacer segment of crRNA pairs with target DNA and allows Cas9-crRNA-tracrRNA complex to recognize specific sites in invader's genome. Target recognition requires, along with complementary pairing of crRNA with the target, a PAM sequence at the 3'-side of the protospacer. Cas9 orthologues from different species often require different PAM sequences (46, 59, 60). PAM specificity is determined by PAM interacting domain of Cas9 and may be mutated to change PAM preference of the enzyme (51, 61).

For DNA cleavage Cas9 uses two nuclease domains: RuvC and HNH (51, 52, 58, 62). These domains cleavage different strand of target DNA: HNH cleaves the strand which is paired with crRNA (further – a target strand), while RuvC cleaves the other strand (further – a non-target strand) (51, 52). Mutations in RuvC or HNH domains active sites allow to generate Cas9 nickases – enzymes which cleave only one strand, target or non-target, depending on which domain is deactivated (51, 63). Mutations in the active sites of both domains produce a dead version of Cas9, an enzyme which in complex with crRNA and tracrRNA can specifically bind DNA but is unable to cleave it.

For DNA cleavage, both RuvC and HNH domains require the presence of divalent metal ions, typically Mg^{2+}(51, 52), although the preference to certain ions can depend on the habitat of the bacterial host. Cas9 nucleases typically cleave DNA three nucleotides upstream of PAM, producing predominantly blunt ends at the cleavage site (51, 52).

Besides interference, Cas9 proteins typically also participate in acquisition of new spacers in the CRISPR array, along with the major proteins of the adaptation module, Cas1 and Cas2 (37, 64). The process of crRNA maturation also depends on Cas9. In complex with tracrRNA, Cas9 binds to DR-derived segments of pre-crRNA through hybridization between two RNAs. Next, the non-Cas nuclease of the host, Rnase III introduces breaks in pre-crRNA-tracrRNA hybrids producing mature crRNA (57).

In 2012 Jinek et al. and Gasiunas et al. showed that a minimal Type II CRISPR-Cas9 DNA cleavage complex can be reconstructed *in vitro* (51, 52): Cas9 in complex with tracrRNA and crRNA was targeted to a certain DNA site by using of spacer segment of crRNA complementary to the target site. This finding demonstrated that CRISPR-Cas nucleases, and particularly Cas9 enzymes, can be programmed to cleave a specific DNA site. This gave rise an idea that CRISPR-

Cas effectors can be used to cleave genomic DNA in non-host cells, allowing genome modification in different organisms, substituting the widely used at the time TALEN (Transcription Activator-Like Effector Nucleases) and ZFN (Zinc Finger Nucleases) genomic editors. In contrast to TALENs and ZFNs, which rely on protein domains for DNA target sites recognition, Cas nucleases are guided by crRNAs. Optimization of DNA-binding motifs of ZFN and TALEN is time consuming and expensive compared to synthesis of crRNAs with a certain spacer segment sequence (65, 66). Moreover, Jinek et al. fused tracrRNA with crRNA into a single molecule, so-called single guide RNA (further sgRNA). This reduced the number of Cas9 minimal complex components to two and simplified the Cas9 DNA-targeting system even further.

In 2013 Cong et al. and Mali et al. showed that Cas9-guideRNA complex can be reconstructed in human cells by expressing the Cas9 gene and sequences coding for sgRNAs from eukaryotic promoters (28). These discoveries opened up the field of CRISPR-Cas genome editing. Since then, the growing number of CRISPR-Cas applications as well as a constant pull for new Class 2 enzymes suited for various properties accelerated the pace of Cas enzymes search and characterization. The focus of bioinformatics searches was on CRISPR-Cas systems of new types as well as on new Cas orthologues, particularly on Cas9s. Studies performed by E. Koonin and colleagues showed that genes coding for Cas9 proteins are unevenly distributed between archaea and bacteria: only several of archaea encode Type II CRISPR-Cas systems while numerous bacteria contain them (9, 53, 67). The proposed reason for this was the requirement of RNAse III, an enzyme absent from archaea, for pre-crRNA processing (53). Thus, most of Cas9 orthologues known to date originate from bacterial species, although recently several archaeal Cas9s were identified (9).

Although all Cas9 effectors use similar mechanisms to cleave DNA, through the use of HNH and RuvC domains, they demonstrate variability in PAM preferences, their sizes, temperature requirements and optimal conditions for function depending on their bacterial host. In addition, Type II CRISPR-Cas systems demonstrate slight differences in organization of their loci. Together, the phylogeny of *cas* genes and the presence of axillary adaptation genes *csn1* and *cas4* in the locus allow to divide Type II systems into three different subtypes (Type II-A, Type II-B and Type II-C) (53, 68).

4.1.1. SpCas9 and its applications

The first CRISPR-Cas nuclease which was shown by Cong et al. to be active in human cells was SpCas9 from *Streptococcus pyogenes* (28). According to the current classification, SpCas9 is a Type II-A CRISPR-Cas nuclease (53). This enzyme remains the most popular Cas

nuclease and the best characterized to date. Although, the first *in vitro* reconstructions of Cas9 DNA cleavage complex were done using enzymes from both *S. pyogenes* and *S. thermophilus* CRISPR-Cas Type II systems (51, 52), Cong et al. tested SpCas9 in human cells (28). Since then numerous biotechnology tools and instruments of genome modification were developed based on this nuclease (63, 69–74).

As is typical for Type II-A nucleases, SpCas9 is a large enzyme with a size of 1368 amino acids. For efficient DNA binding and cleavage, it requires the presence of relatively short two nucleotide PAM sequence 5'-NGG-3' (51). Though this sequence is frequently encountered in DNA, the strict requirement for PAM restricts the range of possible SpCas9 targets. Since SpCas9 source organism *S. pyogenes* is a human pathogen, SpCas9 is active at 30 – 45°C, which is close to the temperature of human body (75). SpCas9 is highly efficient in cleaving genomic DNA in mammalian cells as well as DNA of prokaryotes and other eukaryotes living at similar temperatures.

The basic approach of CRISPR-Cas eukaryotic genome editing and modification relies on the introduction of double-stranded DNA breaks in a desirable site of genomic DNA using Cas9 proteins charged with sgRNA of a certain sequence. This double stranded break is repaired by endogenous reparation systems – NHEJ (Non-homologous end joining) or HDR (homology directed repair). Most of double-stranded breaks are repaired through NHEJ – direct double stranded break ends ligation, which does not require any reparation template and produces insertions or deletions in the repaired DNA molecule, called "indels". Two-thirds of such repair events lead to frameshift mutation in genes and can be used for generation of gene knockouts (Figure 6). The HDR pathway requires a single stranded or double stranded DNA template for reparation. HDR happens less frequently than NHEJ but allows to incorporate new sequences into eukaryotic genome to modify or edit its sequence.

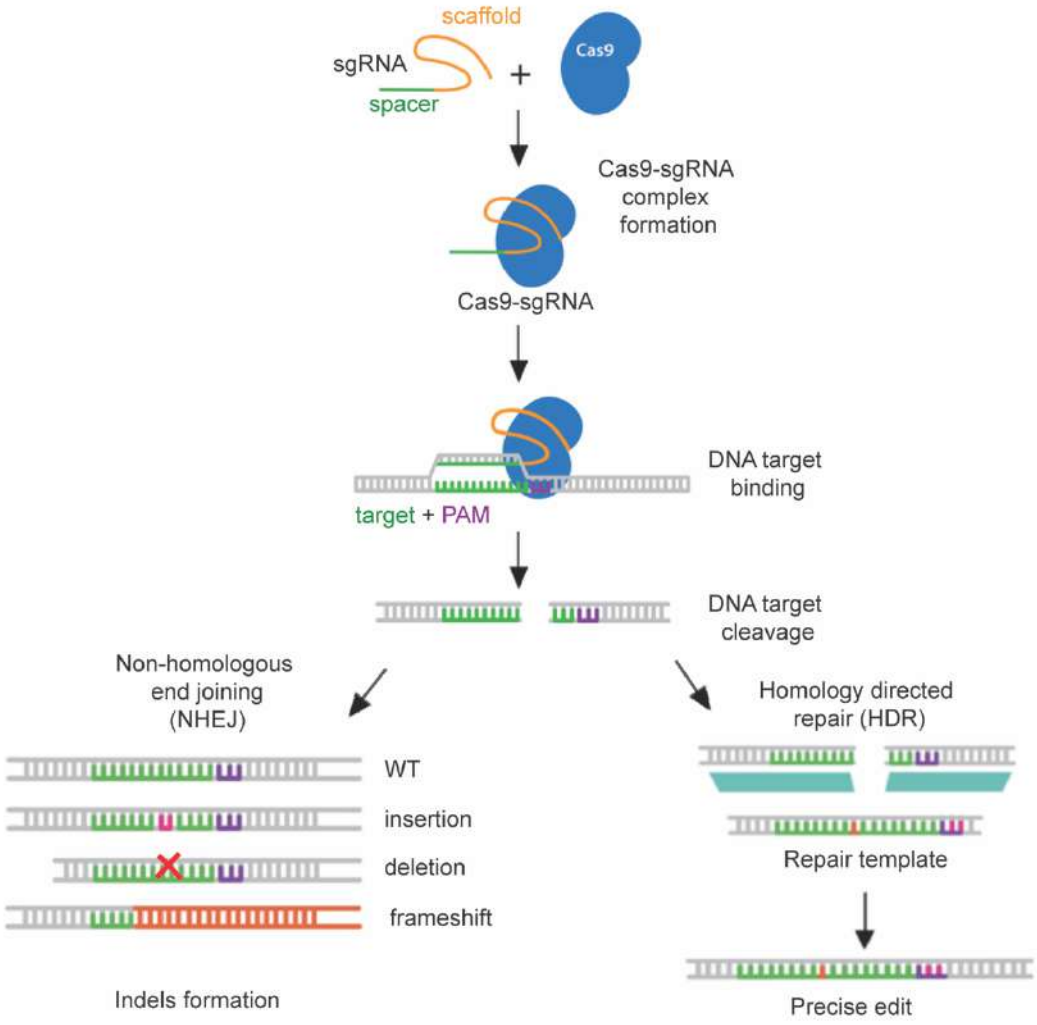


Figure 6. CRISPR-Cas genome modification through non-homologous end joining (NHEJ) or homology directed repair (HDR). *Adapted from Addgene web site https://www.addgene.org/guides/crispr/ with modifications.*

SpCas9 demonstrates relatively high efficiency in genome modification, through both NHEJ and HDR, in plant, insect, fish, and mammalian, including human, cell lines (76–79). To introduce a double-stranded break in genomic DNA the CRISPR-SpCas9 system should be delivered into cells. This can be done by delivery of SpCas9-sgRNA ribonucleoprotein complexes; through delivering of DNA coding for SpCas9 gene and sgRNA positioned under control of eukaryotic promoters; or through delivery of SpCas9 mRNA and sgRNA molecules (60, 77). SpCas9-sgRNA ribonucleoprotein complexes currently are used for injection in zygotes, as well as for transformation of cell lines, including plant protoplasts (77, 80). Delivery of CRISPR-Cas systems in the DNA form can be done by plasmid transfection into a cell line or by viral delivery into an organism. The use of adeno associated viruses (AAV) provides fast and efficient way of

delivery of genetic material into the dividing as well as non-dividing cells. Furthermore, AAV are safe for human, which was demonstrated in several clinical trials (81, 82). They are divided into several serotypes specific to a certain organ or tissue, allowing targeted delivery of a transgene (81, 83, 84). The small size of the capsid allows AAV particles to efficiently spread in the tissue from the site of injection, and thus allows for efficient modification in a particular organ.

Unfortunately, AAV particles packaged with DNA longer than 4.8 kb have reduced infectivity (85). Due to this, the SpCas9 DNA cleavage system, including the effector gene (4,104 bp), sgRNA coding sequence as well as promoters and terminators, cannot be packaged into a single AAV. This is why CRISPR-SpCas9 system is typically delivered using dual AAV system (86, 87). The need of simultaneous transduction of a cell by two kinds of AAV particles reduces the efficiency of CRISPR-SpCas9 system delivery. Multiplex gene editing (simultaneous targeting of several genes) with SpCas9 requires even higher viral capacity to package several sequences coding for sgRNAs as well as several promoters, driving each guide RNA expression.

Another drawback of SpCas9, besides of its large size, is it's not perfect specificity, although it can be improved by modification through rational SpCas9 mutagenesis and directed evolution. (88–90).

Bioinformatics searches for CRISPR-Cas systems of other types, as well as Cas9 orthologs, allow one to find Cas nucleases with properties different from SpCas9: of smaller size, with different PAM requirements and even different mechanisms of nucleic acids cleavage. While Cas proteins with novel PAM and guide RNA requirements can broaden the range of possible CRISPR-Cas targets and provide "orthogonal" instruments for simultaneous regulation and editing of multiple genes (91), Cas nucleases with different mode of DNA cleavage and other unique properties can be used for developing of completely new tools (92).

CRISPR-Cas9 systems are also used for mutagenesis of prokaryotic cells. To date, most approaches of bacterial genome mutagenesis also rely on SpCas9 (93, 94). But this strategy is often challenging due to the need of heterogeneous expression of SpCas9: it requires knowledge on functional promoters to drive the expression of *cas9* gene and introduction of DNA carrying the CRISPR-SpCas9 system into sometimes difficult-to-transform organisms. The use of endogenous CRISPR-Cas systems should facilitate the modification of bacteria of biotechnological and industrial importance.

4.1.2. SpCas9 orthologs

Bioinformatics searches for SpCas9 orthologs in genomes of different bacteria reveal numerous enzymes, which are not biochemically characterized to date. These enzymes can have novel PAM

requirements and unique temperature and reaction condition requirements. Indeed, the recent *in vitro* screening study performed on 79 Cas9 orthologs by Gasiunas et al. showed that Cas9 nucleases may require completely different PAM sequences, which potentially can broaden the range of possible targets. Analysis of Cas9-coding genes from known genomes reveals that the sizes of these nucleases range from 800 to 1600 amino acids (63) (Figure 7).

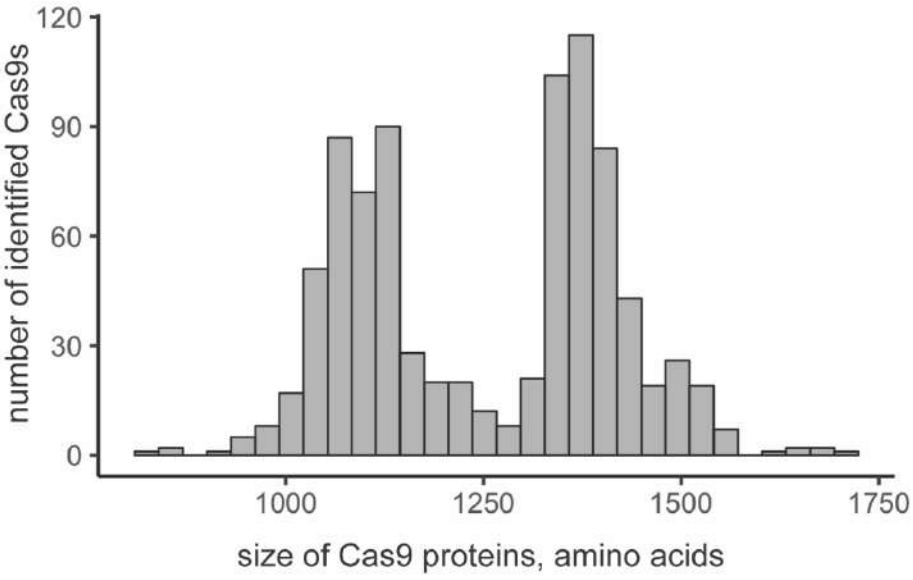


Figure 7. Length of Cas9 orthologs.

The histogram was plotted using results of bioinformatic search for Cas9 orthologs kindly provided by S. Shmakov.

The bioinformatic analysis of the CRISPR-Cas loci allows one to predict tracrRNA and crRNA sequences, active sites of Cas9 nucleases and sometimes to approximately determine PAM. Nevertheless, the biochemical characterization of predicted Cas9 nucleases is necessary to show that the system is active and to precisely determine the PAM requirements (95). To date several Cas9 orthologs from different bacterial and archaeal species were shown to be active *in vitro* and some of them appeared to be active in human cells (Table 1). These nucleases belong to Type II-A (SpCas9, St1Cas9, St3Cas9, SaCas9, ScCas9), Type II-B (FnCas9), or Type II-C (GeoCas9, CdCas9, Nme1Cas9) CRISPR-Cas systems.

Table 1. Cas9 orthologs active in human cells.

Cas9 ortholog	Bacterial host	PAM N – A, T, C or G H – A,T or C R – A or G Y – C or T W – A or T	Gene size, kb	Demonstration of genome modification activity in eukaryotes	Reference
SpCas9	*Streptococcus pyogenes*	NGG	4.1	January 2013	(28)
FnCas9	*Francisella novicida*	NGG	4.9	October 2019	(96)
ScCas9	*Streptococcus canis*	NNG	4.1	October 2018	(97)
St3Cas9	*Streptococcus thermophilus*	NGGNG	4.2	January 2015	(80)
SauriCas9	*Staphylococcus Auricularis*	NNGG	3.3	March 2020	(98)
CjCas9	*Campylobacter jejuni*	NNNNRYAC	3.0	February 2017	(59)
SaCas9	*Staphylococcus aureus*	NNGRRT	3.2	April 2015	(60)
Nme1Cas9	*Neisseria meningitidis* strain 8013	NNNNGNTT	3.2	November 2013	(91)
Nme2Cas9	*Neisseria meningitidis* strain De11444	NNNNCC	3.2	February 2019	(99)
CdCas9	*Corynebacterium diphtheriae*	NNRHHHY	3.2	April 2019	(100)
GeoCas9	*Geobacillus stearothermophilus*	NNNNCRAA	3.2	November 2017	(101)
St1Cas9	*Streptococcus thermophilus*	NNAGAAW	3.4	November 2013	(91)

As was mentioned earlier, Cas9 nucleases, in particular listed in Table 1, are very diverse. The Type II-B FnCas9 from *Francisella novicida* requires the same PAM as SpCas9, 5'-NGG-3' (102), is significantly larger than SpCas9 (with 4.9 kb of gene size), demonstrates higher specificity and generates 5'-overhangs at the cut site in contrast to SpCas9, which produces blunt-ends (96). The recently characterized *Staphylococcus auricularis* Cas9, SauriCas9, also recognizes the 5'-NGG-3' PAM but it is much smaller than SpCas9 (and FnCas9), which makes this nuclease advantageous for viral delivery (98). GeoCas9 derived from thermophilic bacterium *Geobacillus stearothermophilus* is resistant to degradation in human plasma and active at temperatures up to 70 °C, compared with 45 °C for SpCas9 (101).

Cas9 nucleases from closely related strains and even from different CRISPR-Cas loci coexisting in the same bacterium can be very diverse (91, 99, 103). In 2013 Esvelt et al. showed that Nme1Cas9 is active in human cells. This nuclease requires a long three-nucleotide PAM 5'-

NNNNGNTT-3'(91). Nme1Cas9 was derived from *Neisseria meningitidis,* human pathogen, many strains of which were sequenced. In 2019 Edraki et al. analyzed all available *Neisseria meningitidis* strains and found NmeCas9 orthologs (99). These proteins were highly similar to NmeCas9 but divergent in PAM interacting domain sequences. Characterization of these nucleases allowed to find Nme2Cas9, recognizing a shorter PAM 5'-NNNNCC-3'(99).

Another remarkable example of Cas9 ortholog diversity are St1Cas9 and St3Cas9 nucleases from *S. thermophilus* (80, 91, 103). These effectors are encoded by two different Type II-A CRISPR-Cas loci encoded in the *S. thermophilus* genome. St1Cas9 and St3Cas9 coexist in the *S. thermophilus* cell and protect their host from DNA invaders recognizing targets with 5'-NNAGAAW-3' and 5'-NGGNG-3' PAMs, respectively (80, 91). St1Cas9 and St3Cas9 have different sizes: their genes are about 3.4 and 4.2 kb, respectively.

Clearly, characterization of new Cas9 orthologs with different properties expands the range of CRISPR-Cas9 applications, providing enzymes with different properties. A significant number of Cas9 nucleases from Table 1 are smaller than SpCas9. This is due to the targeted search of Cas9 orthologs of small size conducted in last several years. The small Cas9 nucleases active in human cells, which recognize short and distinct PAMs, are of high interest for their potential application in biomedicine: they allow packaging in single AAV particles have a wide range of accessible genome targets.

Unfortunately, it appeared that many small Cas9 enzymes require long complicated PAM sites: GeoCas9 requires 5'-NNNNCRAA-3', SaCas9 - 5'-NNGRRT-3', Nme1Cas9 - 5'-NNNNGNTT-3', CjCas9 - 5'-NNNNRYAC-3', which restricts the choice of potential targets. The recently found SauriCas9 and Nme2Cas9 afford a partial solution to this problem – though small, these effectors recognize 5'-NNGG-3' and 5'-NNNNCC-3' PAMs, and expand the range of possible all-in-one AAV targeting to G and C-rich sequences (Figure 8).

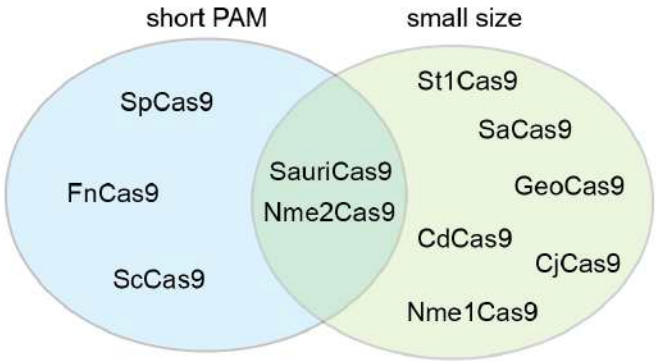


Figure 8. Cas9 effectors active in human cells.

Only several Cas9 orthologs are small enough to be delivered as all-in-one AAV particle and have a relatively short PAM.

Nevertheless, the need for additional small-sized Cas9 orthologs remains. For example, SauriCas9 and Nme2Cas9 cannot be used for modification of "A" or "T"-rich genomic regions. CdCas9 nuclease from *Corynebacterium diphtheriae*, although requiring a promiscuous PAM, demonstrated moderate activity in human cell when injected as RNP in mice zygotes and failed to introduce indels when delivered on a plasmid (100).

In this work, we biochemically characterized several Type II-C Cas9 orthologues of small size. The characterization of CcCas9 effector from bacterium *Clostridium cellulolyticum*, a promising biofuel producer, is described in Chapter 1 of the thesis. Chapter 2 describes the characterization of two CRISPR-Cas9 effectors derived from bacteria inhabiting different environment: DfCas9 from *Defluviimonas sp.20V17* isolated from hydrothermal vents and PpCas9 from *Pasteurella pneumotropica*, isolated from various mammalian species. CcCas9, DfCas9, and PpCas9 are Cas9 effectors of small size, require novel, relatively short PAM sequences: 5'-NNNNGNA-3' (CcCas9), 5'-NNRNAY-3' (DfCas9), 5'-NNNNRT-3' (PpCas9) and actively cleave DNA *in vitro*. Moreover, PpCas9 introduces indels in human cells genome with efficiency comparable to that of SpCas9.

4.2. Type V CRISPR-Cas systems

The successful application of SpCas9 for gene editing in human cells promoted bioinformatics searches for new CRISPR-Cas effectors: both Cas9 orthologs and effectors from completely different families. This led to discovery of a number of new types of CRISPR-Cas effectors and expanded the list of CRISPR-Cas defense systems. Among the first biochemically characterized Class 2 nucleases, were Cas12a enzymes from Type V CRISPR-Cas systems. In 2012 Schunder and colleagues analyzed *Francisella tularensis* CRISPR-Cas loci and along with a Type II system found another, previously unknown Type of CRISPR-Cas loci in substrain *novicidia U112*. The system contained a CRISPR array, the adaptation module (*cas1, cas2, cas4*) and the effector *cas* gene, of unknown kind (102). This protein was characterized by Zetsche et al. in 2015 and appeared to be a member of new type of Class 2 effectors, known today as Type V (104). Since 2015 a number of Type V CRISPR-Cas systems were found and according current classification they divided into 10 subtypes, based predominantly on the effector gene phylogeny (53).

The adaptation module in Type V systems typically consists of *cas1, cas2* and often *cas4* genes. Type V Cas effectors, known as Cas12, presumably originated from TnpB (transposase B, a component of a transposon) (105). Cas12a nucleases predominantly cleave DNA invaders and

rely on a single RuvC domain, in contrast to Cas9, which uses RuvC and HNH domains for DNA cleavage. The Cas12 DNA cleavage mechanism will be discussed further in the text.

4.2.1. Cas12a nucleases and their application

The first Cas12 enzyme characterized by Zetsche et al., in 2015 was FnCas12a (formerly FnCpf1, "CRISPR from *Prevotella* and *Francisella 1*") from *Francisella novicidia U112* (104). The CRISPR-FnCas12a locus contains *cas1, cas2* and *cas4* genes. The locus does not encode any axillary RNAs, such as tracrRNA (Figure 9). The effector ribonucleoprotein complex consists of only Cas12a protein and crRNA. The conserved nucleotides of DR-derived segment of crRNA form a hairpin (104). This hairpin adopts a pseudoknot structure coordinated by a magnesium ion and serves as a scaffold for proper crRNA-Cas protein complex formation (106). Similarly to other CRISPR-Cas effectors, Cas12a binds to DNA through complementary pairing between crRNA and the target DNA strand and requires a PAM sequence flanking the 5'-end of the protospacer. In contrast to Cas9, Cas12a cleaves DNA at PAM-distal part of the protospacer, producing staggered DNA ends.

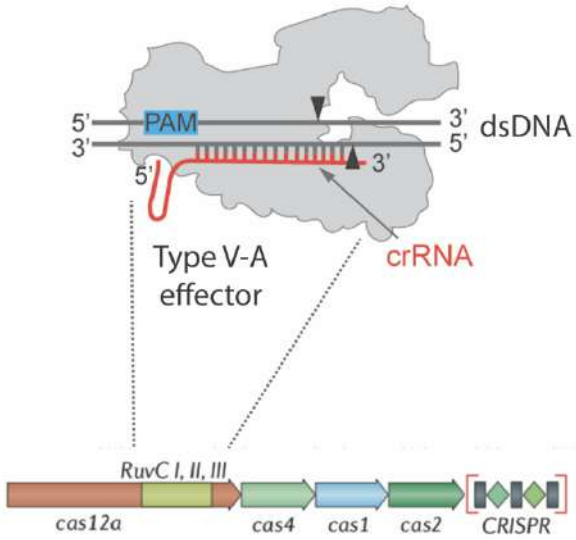


Figure 9. A scheme of a Type V-A CRISPR-Cas locus and the Cas12a effector complex with target DNA.

Type V-A loci consist of the effector module, represented by the *cas12a* gene, the adaptation module, represented by *cas1, cas2*, *cas4*, and a CRISPR array. The active ribonucleoprotein effector complex consisting of Cas12a and crRNA is shown above. *Reproduced with permission from (53) with modifications.*

Along with FnCas9, Zetsche et al. characterized AsCas12a from *Acidaminococcus* and LbCas1 from *Lachnospiraceae* (104). It was shown that these effectors are active in eukaryotes

and can be used as genome editing instruments in human cells, as well as in animal, plant, and fish cells (107, 108). Later, it is appeared that FnCas9 also is able to produce indels in eukaryotic genome (109), although it failed to modify DNA in human cells in Zetsche at al. studies (104). Moreover, it was shown that orthologous proteins Mb3Cas12a from *Moraxella bovoculi AAX11_00205* and EeCas12a from *Eubacterium eligens* efficiently introduce indels in eukaryotic genome too (110, 111).

FnCas12a, AsCas12a, LbCas12a, Mb3Cas12a, and EeCas12a are about 1300 amino acids long and, as most of Cas12a nucleases, recognize thymidine-rich PAMs (5'-TTN-3' – the optimal PAM for FnCas12a and 5'-TTTN-3' - for the others) (104), which expands the range of possible DNA targets.

Cas12a proteins as well as the other Type V systems effectors have very different domain organization compared to that of Cas9. The crystal structures of AsCas12a and LbCas12a were solved in 2016 and revealed that Cas12a nucleases adopt a bilobed architecture: the so-called REC and NUC lobes are separated by a positively charged channel, to which crRNA-target DNA heteroduplex is bound upon target recognition (106, 112). The NUC lobe consists of RuvC domain and three domains, not related to any domains of Cas9 nucleases. Similarly to WED and PI domains in Cas9, two of these domains interact with the crRNA scaffold and PAM (112). Mutations in the third unique domain led to generation of a AsCas12a nickase and due to this, this domain (Nuc) was proposed to be the second nuclease responsible for DNA cleavage along with the RuvC domain (112). Chapter 3 of this thesis describes the study of AsCas12a structure and determination of the role of Nuc domain.

The structures of other Type V nucleases (AacCas12b from Type V-B1 and DpbCas12e from Type V-E (113)) (114), as well as FnCas12a (115) revealed that the Nuc domain of Cas12a is not a DNA cleavage domain but rather the target strand loading domain (TSL), which helps to exchange the DNA strands in the RuvC catalytic pocket to allow cleavage of both DNA strands by a single active site (Figure 10). The strand exchange is mediated by sharp target strand bending by the TSL domain, which helps to replace the cleaved non-target DNA strand in the RuvC domain. Although composed of different amino acid sequences, the TSL domains are thought to be present in all Type V effectors and play a similar role (113, 114), (Figure 10).

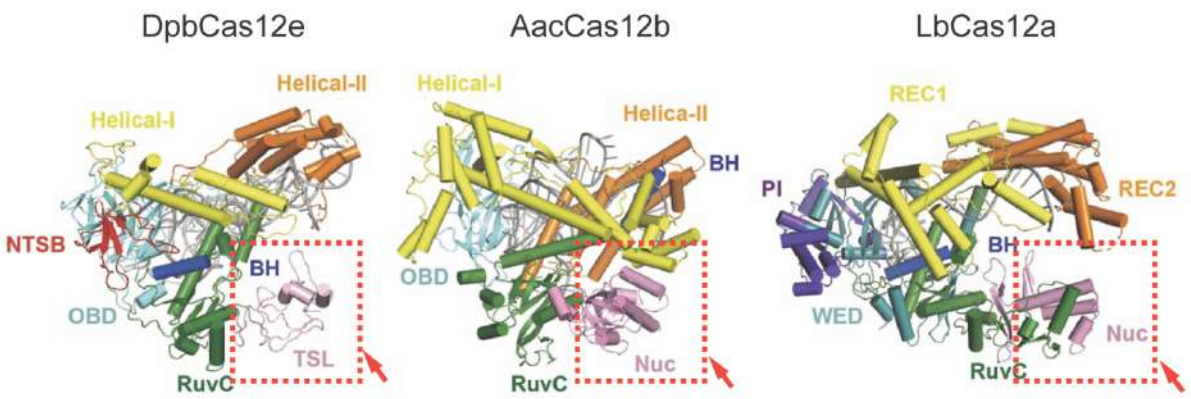


Figure 10. Target strand loading (TSL or Nuc) domain position in different members of Type V effectors: Cas12e, Cas12b, and Cas12a.

TSL or Nuc are colored in pink and indicated by dotted red frames. *Reproduced and modified with permission from (113)*.

The bending of the target DNA strand needed to make it accessible by the RuvC domain, causes a shift of the target strand cleavage position; as a result, Type V effectors produce several nucleotide 5'-overhangs at the DNA cleavage site. Chapter 5 of this work, beside other results, describes precise determination of cut site positions of Type V nucleases.

Together, these findings resulted in a model of sequential DNA cleavage by a single active site through strand displacement, which is common for the majority of Type V CRISPR-Cas systems effectors (Figure 11) (113–115).

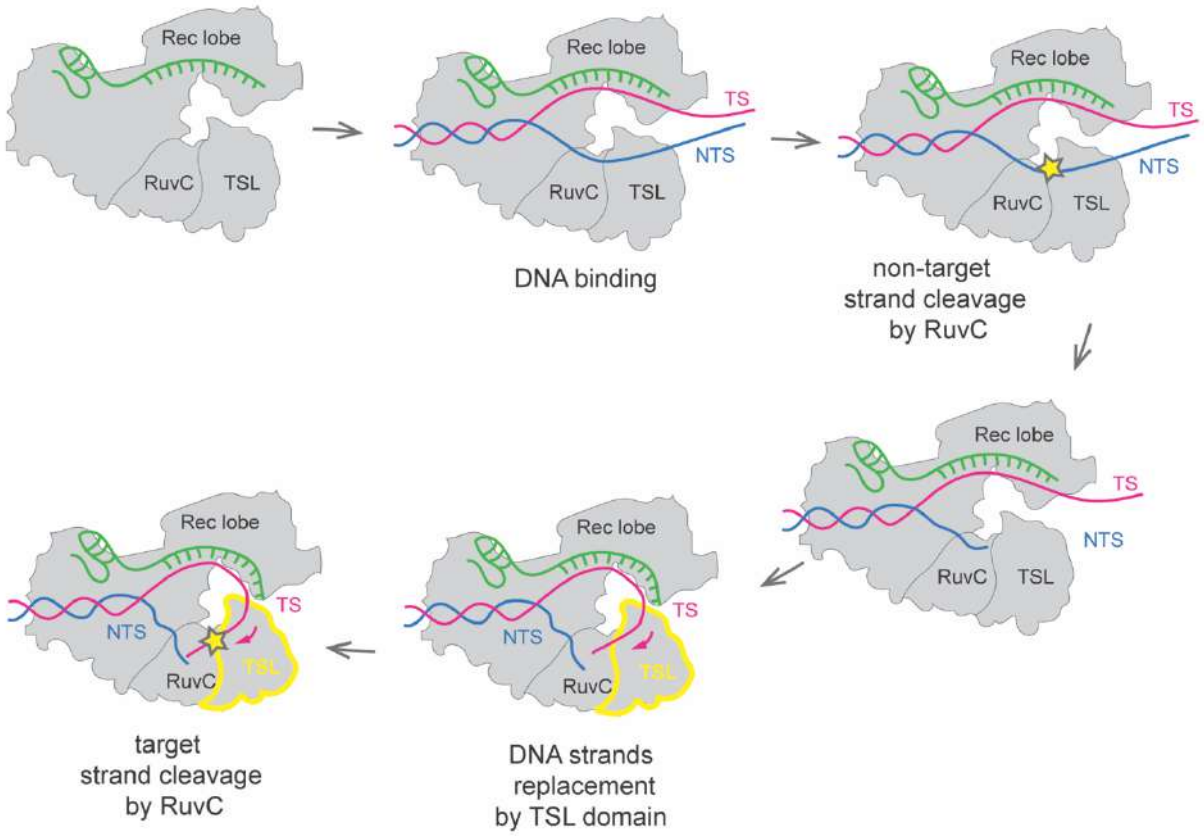


Figure 11. A scheme DNA cleavage by Cas12a.

After binding to the DNA target, the RuvC domain of Cas12a cleaves the non-target strand of DNA (NTS). Further, target strand loading domain (TSL) replaces, through a conformational rearrangement, the non-target strand with the target strand (TS), placing it into the RuvC catalytic pocket. The target strand cleavage by the RuvC domain results in double-stranded break in DNA. TS is indicated in pink, NTS - in blue, crRNA is shown in green, the activation of TSL domain is shown by yellow color, the DNA cleavage steps are shown by asterisks. The scheme illustrates a model proposed in (113–115)

After the cleavage of both strands of the double-stranded DNA target, the PAM-distal product is released, while the PAM-proximal DNA fragment remains bound to Cas12a (48). This locks Cas12a in an activated state, in which the RuvC active site is exposed to the solution. The activated post-DNA cleavage Cas12a is able to cleave non-target ssDNA as part of "collateral damage" (116). This target-activated nonspecific ssDNA cleavage activity was successfully employed for detection of nucleic acids (116).

As it turned out, Cas12a differs from Cas9 not only in the mode of DNA cleavage but also in maturation of crRNA. While Cas9 effector plays a minor role in pre-crRNA processing, protecting the crRNA-tracrRNA duplex and allowing RNAse III to produce the mature crRNAs of a particular size (57), Cas12a is a main player in the processing of the primary CRISPR array transcript. Chapter 4 of the thesis demonstrates that Cas12a can process pre-crRNA on its own and

describes the application of this Cas12a property for multiplex gene editing (simultaneous modification of several genes).

4.2.2. Cas12e nucleases and their application

Besides the Cas12a enzymes, effectors related to other subtypes of Type V CRISPR-Cas systems were applied for genome editing. In 2017 Burstein et al. bioinformatically identified and partially characterized by experiments in model bacteria Class 2 CRISPR-Cas systems of Type V-E (9). The Type V-E CRISPR-Cas loci, besides the effector gene, encode a CRISPR array, an adaptation module (*cas1, cas2* and *cas4* genes), as well as tracrRNA (Figure 12). The Cas12e enzyme, together with tracrRNA and crRNA, forms an effector complex which recognizes double-stranded DNA targets flanked at the 3'-end by a PAM sequence, and introduce a break generating staggered DNA ends similar to those produced by Cas12a (113). Liu et al. reported that Cas12e generates 5'-overhangs much longer than Cas12a – about 10 nucleotides long. In Chapter 5 of the thesis we precisely mapped the DpbCas12e and AsCas12a DNA cleavage sites to compare the cleavage patterns of these two nucleases.

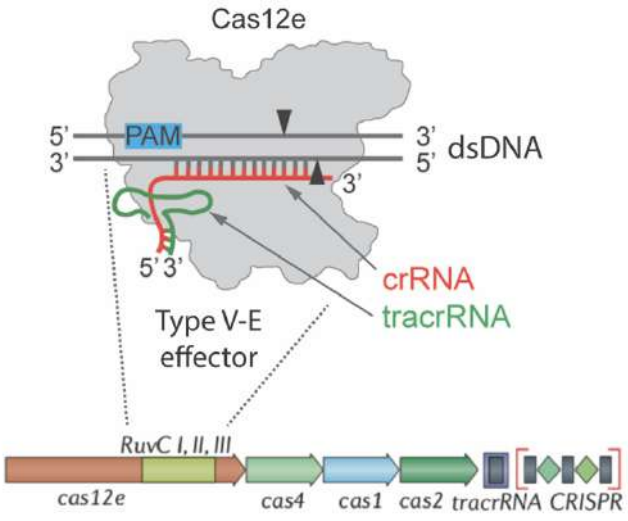


Figure 12. A scheme of Type V-E CRISPR-Cas locus.

Type V-E loci consist of the effector module, represented by the *cas12e* gene (formerly *casX*), the adaptation module, represented by *cas1, cas2* and *cas4*, a sequence coding for tracrRNA and a CRISPR array. The active effector complex consisting of Cas12e, tracrRNA and crRNA is shown above. *Reproduced with permission from (53) with modifications.*

Although the detailed biochemical characterization was done using DpbCas12a, a nuclease found in metagenomic DNA and related to *Deltaproteobaceria* species, further studies showed that an orthologous protein PlmCas12a from *Planctomycetes* demonstrates higher activity in eukaryotes (113). Thus, due to its small size PlmCas12e can be considered as a promising genome editing instrument, which potentially can be delivered by all-in-one AAV particles.

4.2.3. Other Type V CRISPR-Cas systems effectors

The family of Type V CRISPR-Cas systems is very diverse. Besides the discussed above Type V effectors, a number of nucleases of other subfamilies of Type V, were biochemically characterized, and some of them demonstrated activity in eukaryotic cells.

The Type V-B CRISPR-Cas systems were found by Shmakov et. al in 2015 (8). The Type V-B locus from *Alicyclobacillus acidoterrestris* was shown to be active in bacteria. It consists of the effector gene, the adaptation module, a CRISPR array and tracrRNA gene. The effector protein AaCas12b (formerly AaC2C1) has a size similar to that of SpCas9 and in complex with crRNA and tracrRNA introduces double-stranded breaks in targets flanked by 5'-TTN-3' PAM, generating staggered DNA ends. In 2018 Teng et al. repurposed AaCas12b for genome editing and regulation of transcription (117).

Definitely worth mentioning are Type V-F CRISPR-Cas systems. The effector proteins of this subtype are remarkably small: the member of the V-F1 family Un1Cas12f1 found in uncultivated archaeon (initially named Cas14a) is only about 500 amino acids long. Uni1Cas12f1 in complex with tracrRNA and crRNA cleaves dsDNA targets flanked by T-rich PAM or ssDNA without any PAM sequence (118, 119). Similarly to V-A nucleases, Un1Cas12f1 demonstrates target-activated nonspecific ssDNA cleavage activity (116, 119).

Other V-F1 family nucleases, Mi2Cas12f1, PtCas12f1, AsCas12f1, and CnCas12f1, also require two-nucleotide PAMs for dsDNA cleavage (118). Although their activity in eukaryotes remains to be tested, they appear promising for further genome editing applications due to their small (400-600 amino acids) size and short PAMs (118).

Surprisingly, Yan and colleagues demonstrated that in contrast to other Type V nucleases Type V-G effectors cleave not DNA, but RNA molecules (120). Due to their small size (about 800 amino acids) and no PAM requirements, the V-G effectors were proposed for *in vivo* transcriptome engineering applications, although the activity in human cells has not been demonstrated yet for any effector of this type (120). Nucleases of V-I and V-H subtypes are also small enzymes holding promising for further application in biotechnology, although not enough studied to date (120).

4.3. Type VI CRISPR-Cas systems

Although the thesis is predominantly dedicated to DNA cleaving Cas nucleases, it is necessary to discuss Type VI CRISPR-Cas systems, recognizing RNA molecules. They were discovered in 2015 (8), and subfamily VI-A was biochemically characterized in 2016 (121). The Type VI-A locus of *Leptotrichia shahii* consists of the adaptation module, a CRISPR array, and the effector nuclease gene *LshCas13a* (formerly LshC2C2). The effector recognizes RNA molecules through complementary pairing between the target and crRNA and cleaves them using two HEPN domains. For efficient RNA cleavage Cas13a requires PFS (protospacer flanking sequence) and prefers unstructured RNA target sites. It was shown that Cas13a cleaves collateral RNA in addition to crRNA-targeted ssRNA. This property of Cas13a was used by Gootenberg et al. in 2017 for creation of nucleic acid detection tool SHERLOCK (Specific High-Sensitivity Enzymatic Reporter UnLOCKing)(122)

AIM OF THE STUDY

The aim of my PhD thesis project was to explore the diversity of CRISPR-Cas effectors, reveal their unique properties, and apply them, whenever possible, for genome modification of eukaryotes.

To pursue this, the following specific goals were set:

1. Biochemically characterize several Cas9 orthologs of small size from different bacterial species (Chapters 1 and 2)
2. Apply small-sized Cas9 orthologs for genome modification of human cells (Chapter 2)
3. Investigate Cas12a domain organization (Chapter 3)
4. Study Cas12a crRNA biogenesis and apply it for multiplex gene editing (Chapter 4)
5. Compare DNA cleavage pattern of Cas12a and Cas12e nucleases (Chapter 5)
6. Determine how CRISPR interference is linked with the adaptation process in bacteria (Chapter 6)

RESULTS

Chapter I

DNA targeting by *Clostridium cellulolyticum* CRISPR-Cas9 Type II-C system

Introduction

In this chapter, we characterized CRISPR-Cas Type II-C system from *Clostridium cellulolyticum,* a bacterium considered to be a promising biofuel producer. Using small RNA sequencing and plasmid interference screening we show that this defense system is active in model bacterium *E. coli* and can protect it from DNA invaders. To study the effector protein, CcCas9, we purified its recombinant version and performed DNA cleavage experiments *in vitro*. As a result, we show that CcCas9 is a relatively small nuclease, which efficiently introduces double-stranded breaks in DNA targets flanked by relatively short, two nucleotide PAM sequence 5'-NNNNGNA-3'.

Contribution

I conceived the study. I designed and participated in all experiments, in particular: designed the plasmids used in the work; performed RNA-Seq to determine the sequences of guiding RNAs and analyzed its results; performed a significant part of *in vitro* experiments to determine CcCas9 PAM sequence. Anatolii Arseniev, an equally contributing author, conducted the purification of recombinant CcCas9, performed biochemical assays, including experiments on determination of CcCas9 temperature requirements.

I prepared the Figures and wrote the manuscript with the help from other authors. I would like to thank all authors for their contributions.

2026–2034 *Nucleic Acids Research, 2020, Vol. 48, No. 4* *Published online 16 January 2020*
doi: 10.1093/nar/gkz1225

DNA targeting by *Clostridium cellulolyticum* CRISPR–Cas9 Type II-C system

Iana Fedorova[1,*,†], Anatolii Arseniev[2,†], Polina Selkova[1], Georgii Pobegalov[2], Ignatiy Goryanin[1], Aleksandra Vasileva[1], Olga Musharova[1], Marina Abramova[2], Maksim Kazalov[2], Tatyana Zyubko[1], Tatyana Artamonova[2], Daria Artamonova[1], Sergey Shmakov[1,3], Mikhail Khodorkovskii[2] and Konstantin Severinov[1,4,*]

[1]Skolkovo Institute of Science and Technology, Center of life sciences, Skolkovo, Russia, [2]Peter the Great St. Petersburg Polytechnic University, Saint Petersburg, Russia, [3]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, USA and [4]Center for Precision Genome Editing and Genetic Technologies for Biomedicine, Institute of Gene Biology, Russian Academy of Sciences, Moscow, Russia

Received July 20, 2019; Revised December 16, 2019; Editorial Decision December 17, 2019; Accepted December 20, 2019

ABSTRACT

Type II CRISPR–Cas9 RNA-guided nucleases are widely used for genome engineering. Type II-A SpCas9 protein from *Streptococcus pyogenes* is the most investigated and highly used enzyme of its class. Nevertheless, it has some drawbacks, including a relatively big size, imperfect specificity and restriction to DNA targets flanked by an NGG PAM sequence. Cas9 orthologs from other bacterial species may provide a rich and largely untapped source of biochemical diversity, which can help to overcome the limitations of SpCas9. Here, we characterize CcCas9, a Type II-C CRISPR nuclease from *Clostridium cellulolyticum* H10. We show that CcCas9 is an active endonuclease of comparatively small size that recognizes a novel two-nucleotide PAM sequence. The CcCas9 can potentially broaden the existing scope of biotechnological applications of Cas9 nucleases and may be particularly advantageous for genome editing of *C. cellulolyticum* H10, a bacterium considered to be a promising biofuel producer.

INTRODUCTION

CRISPR–Cas systems are bacterial and archaeal immune systems that protect their hosts from invaders such as plasmids or bacteriophages. The immune mechanism is based on the function of Cas ribonucleoprotein effector complexes composed of Cas nucleases and CRISPR RNAs (crRNAs). crRNAs are encoded in CRISPR arrays consisting of repeats and intervening unique spacers. Some spacers are derived from invader's DNA and are introduced into CRISPR arrays during the infection. The CRISPR array is transcribed into a pre-crRNA, which is processed further to short mature crRNAs containing a single spacer and flanking repeat sequences. Complementary pairing between crRNA spacer segment and the invader genome allows Cas nucleases to specifically recognize foreign targets and degrade them, thus preventing the spread of the infection.

The crRNAs with investigator defined spacer sequences allow one to guide Cas nucleases to virtually any desirable target. Because of their relative simplicity, single-subunit Cas nucleases of Type II CRISPR–Cas systems form the basis of multiple genome editing applications. Since 2013 Type II CRISPR-based instruments are used for genome modification and transcription regulation in eukaryotic, including human, cells (1). Alongside with eukaryotic genome editing, there is a large demand for genome engineering of microorganisms useful in biotechnology and several efficient CRISPR-based methods of bacterial genome editing have been developed (2–4). Most of these genome editing approaches rely on the use of the SpCas9 protein, the most investigated to date effector nuclease from *Streptococcus pyogenes* Type II-A CRISPR–Cas system (5). Despite high DNA cleavage efficiency, SpCas9 has several limitations due to its large size, a strict requirement for an NGG PAM (protospacer adjacent motif essential for target DNA recognition) and imperfect specificity.

Bioinformatic searches for Cas9 orthologs and their subsequent biochemical characterization reveal nucleases with different properties, which can broaden Cas9 proteins application. Thus, SaCas9 from *Staphylococcus aureus* and CjCas9 from *Campylobacter jejuni*, two small size Cas9 orthologs with PAM requirements 5′- NNGRRT-3′ and 5′-NNNNRYAC-3′, respectively, were shown to be active in human cells (6,7). In 2014, Fonfara *et al.* using bioinfor-

*To whom correspondence should be addressed. Tel: +79062713200; Email: femtokot@gmail.com
Correspondence may also be addressed to Konstantin Severinov. Tel: +79854570284; Email: K.Severinov@skoltech.ru
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

matics approaches detected a Type II-C system CRISPR–Cas in *Clostridium cellulolyticum* genome but no functional characterization of this system was performed (8). The mesophilic cellulolytic bacterium *C. cellulolyticum* is considered to be a promising biofuel producer since it can directly convert plant biomass to lactate, acetate, ethanol and hydrogen (9). Fast and efficient approaches of *C. cellulolyticum* genome engineering will be required to improve the fermentation properties of this microorganism. To date, several CRISPR–Cas-based strategies were applied to change the *C. cellulolyticum* genome, all of them relying on SpCas9 due to the lack of any information about the host CRISPR–Cas system (PAM requirements, guide crRNAs sequences, protospacer length etc.) (10–12). Studying of *C. cellulolyticum* Type II-C CRISPR–Cas system and, in particular, its effector Cas9 nuclease, could facilitate genome modification of this bacterium and provide an additional small-size Cas9 effector for biotechnology or biomedicine. Here, we demonstrate that *C. cellulolyticum* H10 CcCas9 protein is an active RNA-guided nuclease, which efficiently introduces double-stranded breaks in DNA targets flanked by two-nucleotide 5′-NNNNGNA-3′ PAM. To facilitate further application of CcCas9 in biotechnology, we determined the main features of this CRISPR–Cas system, such as crRNA and tracrRNA sequences, the range of temperatures required for *in vitro* activity and created a nickase version of CcCas9, which could be suitable for *C. cellulolyticum* H10 genome editing by a single-nick-assisted homologous recombination (11).

MATERIALS AND METHODS

Plasmids cloning

The entire predicted CRISPR–Cas Type II-C system locus of *C. cellulolyticum* including flanking regions (100 nt upstream of putative tracrRNA coding sequence and 180 nt downstream of the last DR) was PCR amplified with primers locus_F and locus_R using *C. cellulolyticum* H10 genomic DNA (DSMZ 5812) as a template. The resulting fragment was inserted into XbaI and HindIII digested pACYC184 vector using NEBuilder HiFi DNA Assembly Cloning Kit (NEB, E5520). To obtain pET21a_CcCas9 plasmid, CcCas9 coding sequence was PCR amplified with CcCas9_F and CcCas9_R primers using *C. cellulolyticum* H10 genomic DNA as a template. The resulting fragment was inserted into XhoI and NheI digested pET21a vector by NEBuilder HiFi DNA Assembly Cloning Kit (NEB, E5520). The vectors maps are presented in the Supplementary Table S1.

Plasmid transformation interference screening

Randomized 7N plasmid libraries carried a protospacer sequence flanked by seven randomized nucleotides (Supplementary Table S1). To create the library the ssDNA oligo Library_f containing randomized nucleotides was double-stranded through single stage PCR with Library_r primer (Evrogen). This fragment was assembled with PUC19 fragment synthesized through PCR using primers PUC19_F and PUC19_R by NEBuilder HiFi DNA Assembly Cloning Kit (NEB, E5520). The mix was transformed to *Escherichia coli* DH5alpha strain and plated to media supplemented with 100 μg/ml ampicillin. The plates were incubated at 37°C. Eighteen hours after transformation >50 000 colonies were washed off the plates, and the plasmid library was extracted by Qiagen Plasmid Maxi kit (Qiagen 12162). HTS analysis of the library showed representation of 15716 PAM variants. The library plasmid map is presented in the Supplementary Table S1. Competent *E. coli* Star cells carrying pACYC184_CcCas9_locus or an empty pACYC184 vector were transformed with 7N PAM plasmid libraries and plated to 100 μg/ml ampicillin and 25 μg/ml chloramphenicol containing agar plates. After 16 h, cells were harvested and DNA was extracted using Qiagen Plasmid Maxi kit (Qiagen 12162). PAM-containing sequences were PCR amplified using M13_f and M13_r primers and sequenced using Illumina platform with pair-end 150 cycles (75 + 75).

Bacterial RNA sequencing

E. coli DH5alpha carrying pACYC184_CcCas9_locus were grown 16 h at 37°C in LB (Luria Bertani) medium supplemented with 25 μg/ml chloramphenicol. Bacteria were resuspended in TRIzol (ThermoFisher, 15596026). Total RNA was purified using Direct-Zol RNA kit (Zymo research, R2051). RNA was DNase I (Zymo research) treated and 3′ dephosphorylated with T4 PNK (NEB, M0201). Ribo-Zero rRNA Removal Kit (Gram-Negative Bacteria) kit (Illumina, 15066012) was used to remove ribosomal RNA. HTS samples were prepared using NEBNext Multiplex Small RNA Library Prep Set for Illumina (NEB, E7300). The library was sequenced using Illumina platform with pair-end 150 cycles (75 + 75).

RNA sequencing analysis

HTS results of RNA sequencing were aligned to the reference plasmid pACYC184_CcCas9_locus using BWA aligner (13). Determined coordinates of 5′ and 3′ RNA ends were used to reconstruct the full-length RNA sequences. The resulting fragments were analyzed using Geneious 11.1.2. Filtered 40–130 nt-length sequences were used to generate the alignment.

In vitro DNA cleavage assays

DNA cleavage reactions were performed using the recombinant CcCas9 protein and linear dsDNA targets. The reaction conditions were: 1× CutSmart (NEB, B7204) buffer, 1 mM DTT, 30 nM DNA, 400 nM CcCas9, 2 μM crRNA, 2 μM tracrRNA. Samples were incubated at an appropriate temperature for 20 min (unless otherwise stated). Further, 4× loading dye containing 10 mM Tris–HCl, pH 7.8, 40% glycerol, 40 mM EDTA, 0.01% bromphenol blue, 0.01% xylene cyanol was added to stop the reaction. Reaction products were analyzed by electrophoresis in 1.5% agarose gels or, where indicated, in 1× TBE polyacrylamide gels. Pre-staining with ethidium bromide or post-staining with SYBR gold stain (ThermoFisher, 11494) was used for visualization of bands on agarose or polyacrylamide gels, correspondingly.

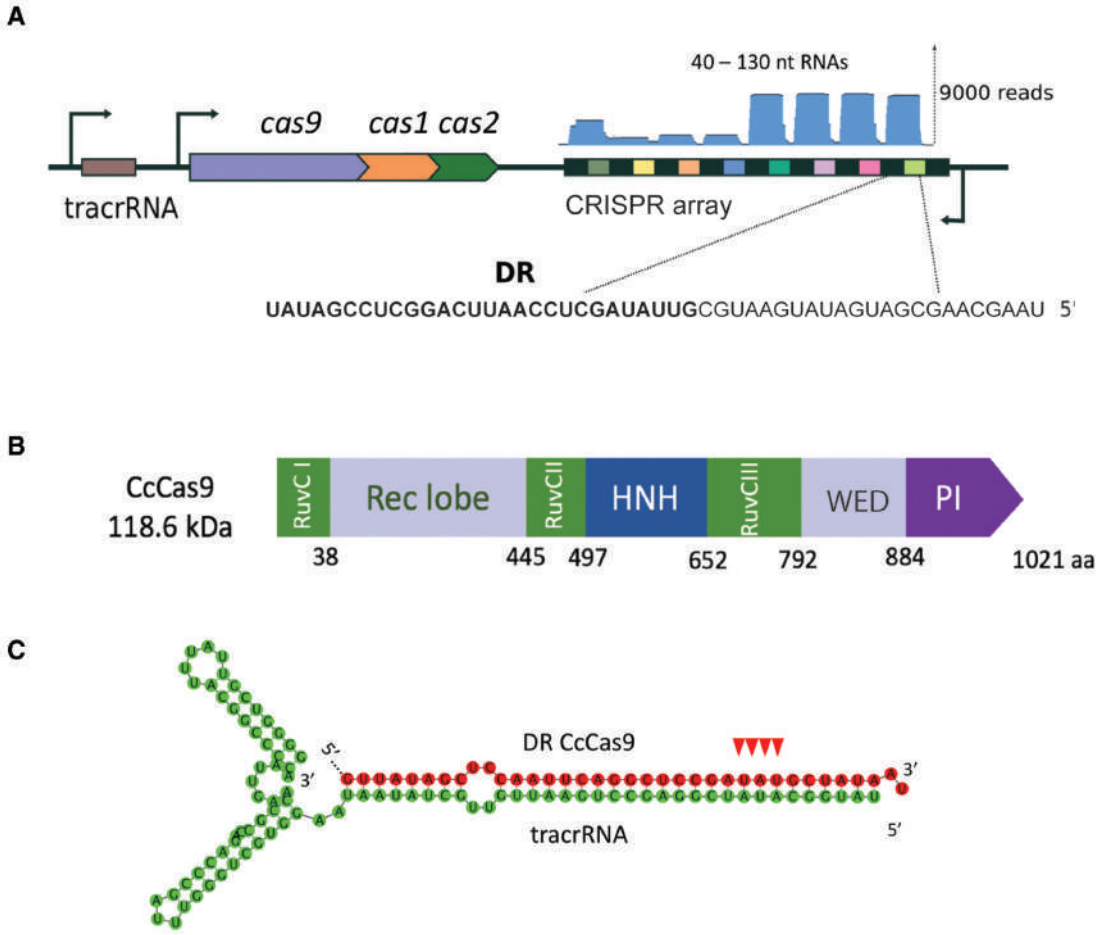


Figure 1. Organization of *Clostridium cellulolyticum* H10 CRISPR–Cas Type II-C locus. (**A**) A scheme of the *C. cellulolyticum* H10 CRISPR–Cas locus. DRs (direct repeats) are shown as black rectangles, spacers are indicated by rectangles of different colors. The tracrRNA coding sequence is shown as a brown rectangle. The *cas* genes are labeled. The direction of transcription is indicated with black arrows. Mapping of small RNAs reads revealed by RNA-seq is shown at the top of CRISPR array in blue. A sequence of typical mature crRNA is expanded below with the DR part shown in bold typeface. (**B**) Domain organization of the CcCas9 protein. (**C**) *In silico* co-folding of *C. cellulolyticum* H10 CRISPR–Cas Type II-C system DR and putative tracrRNA. The DR sequence is colored in red, the tracrRNA sequence is colored in green. The cleavage sites introduced during crRNA maturation are indicated with red arrows. Co-folding was performed using Geneious software, free energy of structure shown is −80.50 kcal/mol.

All *in vitro* DNA cleavage reactions were performed at 45°C unless otherwise stated. For testing the activity of CcCas9 at different temperatures a mix of CcCas9 protein with *in vitro* transcribed crRNA–tracrRNA in the cleavage buffer, and the DNA substrates, also in the cleavage buffer, were first incubated separately at the chosen temperature for 10 min, combined, and incubated for additional 10 min at same temperature.

For *in vitro* PAM screens, 100 nM linear DNA 7N PAM library was incubated with 400 nM CcCas9, 1 μM crRNA and 1 μM tracrRNA. Reactions without crRNA were used as negative controls. The reactions were performed at 45°C for 20 min. Reaction products were separated by electrophoresis in agarose gels. Uncleaved DNA fragments were extracted from the gel using Zymo Clean Gel Recovery kit (Zymo research, D4007). HTS libraries were prepared using Ultra II DNA library prep kit (NEB, E7646). Samples were sequenced using MiniSeq Illumina with single-end 150 cycles. All RNAs used in this study are listed in Supplementary Table S2.

Computational sequence analysis

For PAM screens results analysis, Illumina reads were filtered by requiring an average Phred quality (Q score) of at least 20. Resulting reads were mapped against the corresponding reference sequence using BWA (13). All unmapped reads were discarded from the analysis. The degenerate 7-nucleotide region was extracted from the sequences.

For interference PAM screens analysis, depleted PAM sequences were determined by comparing the number of each PAM counts for CRISPR CcCas9 sample and control. The representation of unique PAM in both samples, as well as PAM representation of initial 7N library was >15 000 PAM variants. WebLogo was used to generate a logo based on 887 of statistically significantly (one-sided Pearson chi-square test with a P-value $< 10^{-12}$) depleted PAM sequences (listed in Supplementary File S2). In case of *in vitro* PAM determination screens 16 364 and 16 363 unique PAM sequences were found, respectively, for the depleted and control samples. Depletion values of PAM sequence positions

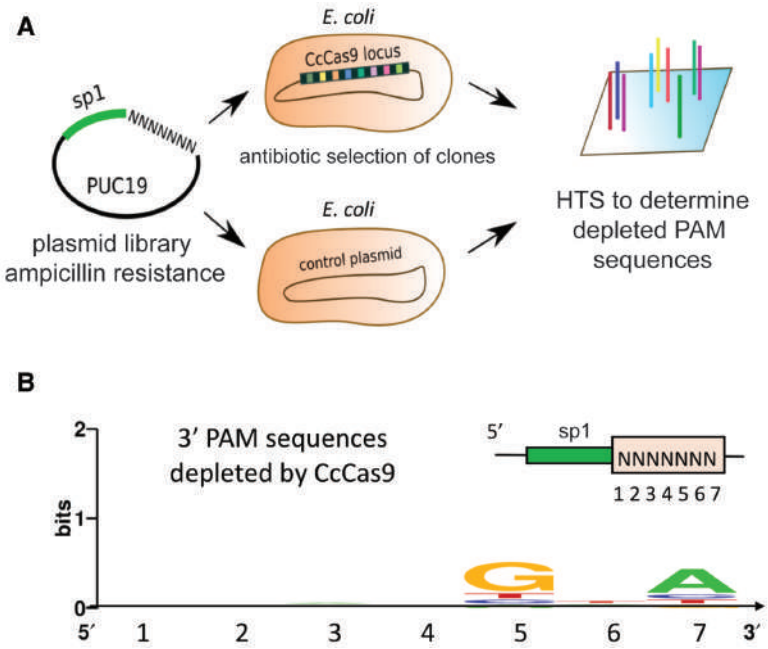


Figure 2. Determination of CcCas9 PAM sequence using plasmid transformation interference screening. (**A**) A scheme of the bacterial interference screen experiment. *Escherichia coli* cells carrying the CcCas9 locus were transformed by PUC19-based library carrying the protospacer sequence flanked by seven randomized nucleotides and plated on ampicillin containing plates. The presence of an interference-proficient PAM decreases the frequency of plasmids with this PAM among ampicillin-resistant colonies. Comparison of PAM representation in CcCas9 locus carrying cells and in control cells without the CcCas9 locus reveals depleted PAM sequences and allows one to deduce the PAM consensus. (**B**) *Clostridium cellulolyticum* CRISPR–Cas Type II-C system PAM sequence logo determined by plasmid transformation interference screening.

were counted according to (14). The frequencies of each PAM variants in depleted and control samples were processed by R script. The frequencies of PAM variants were also used for PAM wheel construction.

Recombinant protein purification

For recombinant CcCas9 purification competent *E. coli* Rosetta cells were transformed with pET21a_CcCas9 plasmid and grown till $OD_{600} = 0.6$ in 500 ml LB media supplemented with 100 μg/ml ampicillin. The target protein synthesis was induced by the addition of 1 mM IPTG. After 18 h of growth at 22°C, cells were centrifuged at 4000g, the pellet was resuspended in lysis buffer containing 50 mM Tris–HCl pH 8.0 (4°C), 500 mM NaCl, 1 mM β-mercaptoethanol and 10 mM imidazole supplemented with 1 mg/ml lysozyme (Sigma) and cells were lysed by sonication. The cell lysate was centrifuged at 16 000g (4°C) and filtered through 0.45 μm filters. The lysate was applied to 1 ml HisTrap HP column (GE Healthcare) and CcCas9 was eluted by imidazole gradient in the same buffer without lysozyme. After affinity chromatography, fractions containing CcCas9 were applied on a Superose 6 Increase 10/300 GL (GE Healthcare) column equilibrated with a buffer containing 50 mM Tris–HCl pH 8.0 (4°C), 500 mM NaCl, 1 mM DTT. Fractions containing CcCas9 monomer were pooled and concentrated using 30 kDa Amicon Ultra-4 centrifugal unit (Merc Millipore, UFC803008). Glycerol was added to final concentration of 10% and samples were flash-frozen in liquid nitrogen and stored at −80°C. Purity of CcCas9 was assessed by denaturing 8% PAGE and the integrity of recombinant protein was confirmed by mass spectrometry.

RESULTS AND DISCUSSION

Clostridium cellulolyticum H10 CRISPR–Cas II-C system: locus organization

The *C. cellulolyticum* H10 type II-C CRISPR–Cas locus was bioinformatically found by Fonfara *et al.* in 2014 but up to date there is no information about the activity of this system. The CRISPRFinder tool (https://crispr.i2bc.paris-saclay.fr/Server/) revealed an array composed of nine 36-bp DRs (direct repeats) interspaced by 31-bp spacers in the proximity of the *cas* genes operon (Figure 1A). A Blast search using spacer sequences as queries revealed no matches to sequences from publicly accessible databases. The *C. cellulolyticum* H10 *cas* genes comprise the CcCas9 effector nuclease gene and the adaptation module composed of *cas1* and *cas2* genes. Being a II-C type Cas nuclease, CcCas9 has a relatively small size (1021 amino acids or 118 kDa) compared to the widely used SpCas9 (1368 amino acids/158 kDa). Alignment of the CcCas9 amino acid sequence with the previously characterized small-size Type II-A SaCas9 protein from *S. aureus* shows the presence of all domains necessary for nuclease activity (Figure 1B, Supplementary Figure S1). Upstream of *cas* genes, we identified a putative tracrRNA-encoding sequence with an anti-repeat partially complementary to DRs. *In silico* co-folding of part of DR with the putative tracrRNA predicts a stable secondary structure (Figure 1C).

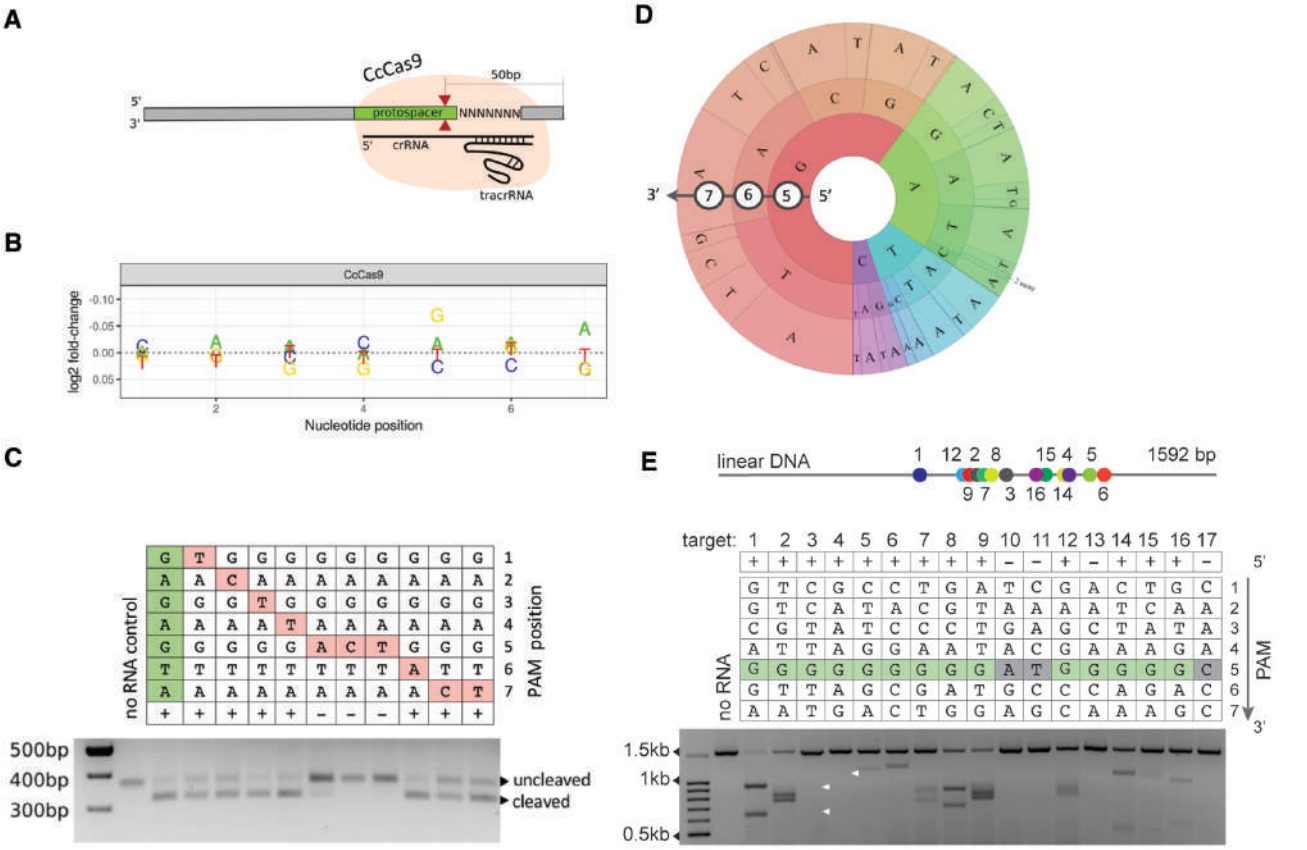


Figure 3. *In vitro* cleavage of DNA targets by CcCas9. (**A**) Scheme of the DNA library used for *in vitro* PAM screening experiment. Cleavage of 7N DNA library with CcCas9 generates DNA products shortened by 50 bp. The location of the cleavage site is shown with red arrows. (**B**) Analysis of depletion of PAM library sequences after *in vitro* cleavage. (**C**) Single-nucleotide substitutions in the 5th position of PAM prevent DNA cleavage by CcCas9. An agarose gel showing the results of electrophoretic separation of cleavage products of targets with PAM sequences shown at the top is presented. The +/– signs signify, correspondingly, whether cleavage was or was not observed. Bands corresponding to cleaved and uncleaved DNA fragments are indicated. (**D**) Wheel representation of *in vitro* PAM screen results for fifth, sixth and seventh nucleotide positions of PAM. Nucleotide positions from the inner to outer circle match the PAM positions moving away from the protospacer. For a given sequence, the area of the sector in the PAM wheel is proportional to the relative depletion in the library. (**E**) *In vitro* cleavage of different 20-bp target sites on a linear DNA fragment by CcCas9. The PAM sequences corresponding to each target are shown in the table. The +/– signs signify, correspondingly, whether cleavage was or was not observed. The conserved G at the fifth position is indicated by green color. Below, a gel showing results of *in vitro* cleavage of targets with indicated PAMs is presented. White arrows indicate positions of poorly visible bands. Above, a scheme showing the relative positions of the targeted DNA sites on the linear DNA fragment is presented.

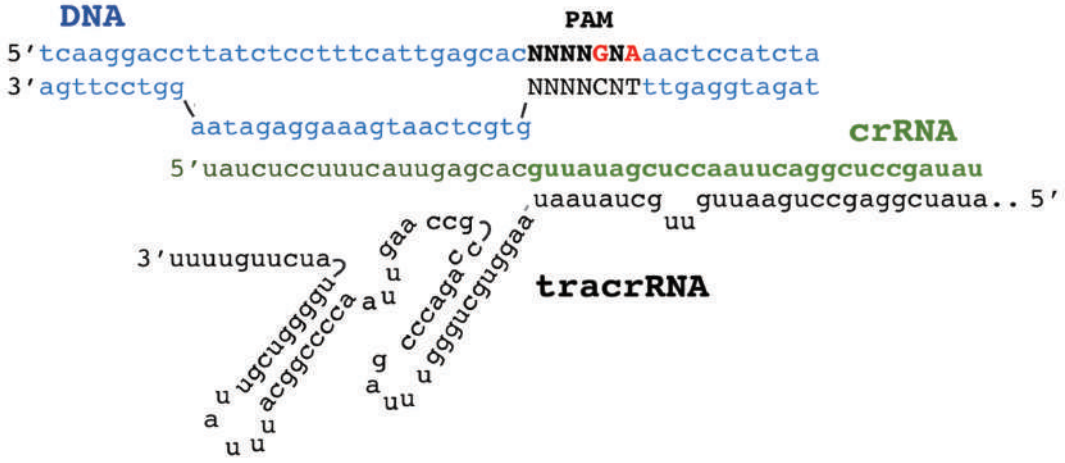


Figure 4. A scheme of the CcCas9 DNA–cleavage complex. DNA is shown in blue, crRNA in green and tracrRNA in black.

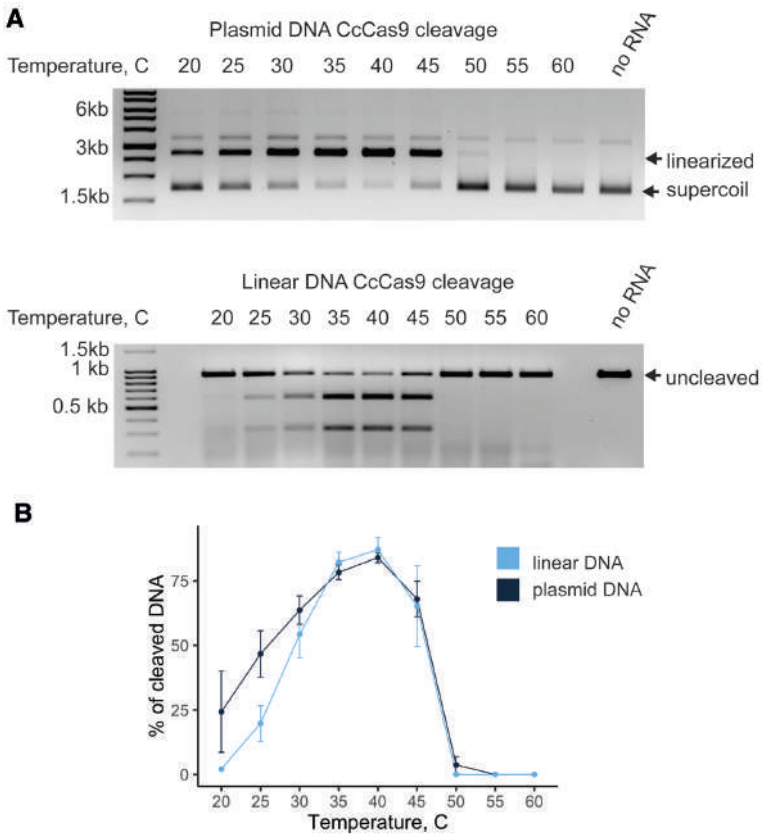


Figure 5. Activity of CcCas9 at different temperatures. (**A**) CcCas9 was incubated with tracrRNA, crRNA and a 2.7 kb plasmid DNA (above) or a 921 bp linear DNA fragment (below) containing a target sequence at indicated temperatures for 10 min. Products were separated by agarose gel electrophoresis. (**B**) CcCas9 was incubated with tracrRNA, crRNA and plasmid or linear DNA as in panel A. Cleavage efficiency (in per cent) was calculated as a ratio of intensity of staining of cleaved bands to the combined intensity of cleaved and uncleaved bands. Mean values and standard deviations obtained from three independent experiments are shown.

The entire CRISPR–Cas locus of *C. cellulolyticum* H10 with adjacent non-coding sequences likely containing promoters was cloned into *E. coli* pACYC184 plasmid vector for heterologous expression. Although *E. coli* cells carry a CRISPR–Cas system of their own, it belongs to a different class (type I-E), relies on different kinds of crRNAs, and is inactive at least at laboratory conditions (15). Thus, no influence of resident CRISPR–Cas on the function of *C. cellulolyticum* H10 CRISPR–Cas is expected. To determine the polarity of *C. cellulolyticum* H10 CRISPR array transcription and confirm the tracrRNA sequence, small RNAs present in *E. coli* heterologously expressing *C. cellulolyticum* H10 CRISPR–Cas locus were sequenced. We found that the CRISPR array is actively transcribed in the orientation opposite to the *cas* genes transcription and mature crRNAs corresponding to every spacer in the array could be detected (Figure 1A). This could be due to efficient processing of pre-crRNA or, alternatively, due to transcription from internal promoters embedded into the repeat sequence, as has been observed in some Type II-C systems (16). Indeed, we noted that the terminal nine nucleotides of *C. cellulolyticum* H10 DRs have a sequence similar to bacterial extended −10 promoter consensus element, as is also the case for *Neisseria meningitidis* CRISPR–Cas II-C system, where transcription initiation within each repeat has been shown experimentally (16). Each *C. cellulolyticum* H10 crRNA contains 23–26 nt of spacer sequence and 24–28 nt of DR. The tracrRNA coding sequence is also expressed, generating variably sized, 70–107 nt, products. In the natural host, the length of mature crRNAs and tracrRNA could be slightly different from those obtained during heterologous expression in *E. coli*.

Determination of CcCas9 PAM by DNA interference screening

Given robust expression of *C. cellulolyticum* crRNAs in *E. coli*, we performed a bacterial interference screen to determine the CcCas9 protospacer adjacent motif (PAM) sequence (Figure 2A). Based on the knowledge about organization of known Cas9-guide RNAs–target DNA complexes and the direction of *C. cellulolyticum* CRISPR array transcription, we designed a plasmid-based PAM library carrying a 30-bp protospacer sequence matching the first spacer in the *C. cellulolyticum* CRISPR array flanked at one side with seven randomized nucleotides (Figure 2A). *E. coli* cells carrying a compatible plasmid with the CcCas9 locus or an empty vector were transformed with the library and plated

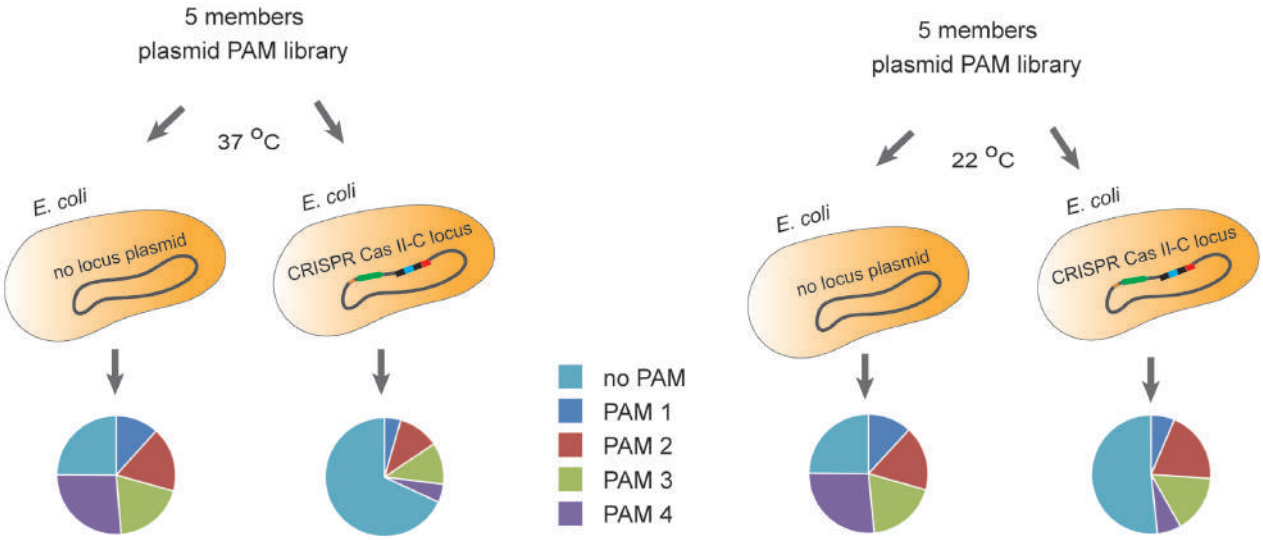


Figure 6. The influence of temperature on *Clostridium cellulolyticum H10* CRISPR–Cas II-C locus interference. A plasmid library composed of five members carrying protospacer matching the first spacer in the CRISPR array and flanked by 5′-ACAGGTA-3′(PAM 1), 5′-CGGTGTA-3′ (PAM 2), 5′-TGAAGAA-3′ (PAM 3), 5′-ATTGGAA-3′ (PAM 4), and 5′-TTCATAT-3′ (no PAM) sequence was transformed in *E. coli* cells carrying a plasmid with the *C. cellulolyticum H10* CRISPR–Cas II-C locus or a control plasmid. Cells were plated on LB media supplemented with ampicillin and chloramphenicol and grown for 18 h at 37°C (left panel) or 22°C (right panel). The plasmid DNA was extracted from grown colonies and HTS was used to estimate the representation of each library member. The pie charts showing PAM representation in colonies formed are shown below. Each colored sector represents a fraction of corresponding PAM sequence.

on a medium that only allowed the growth of cells carrying both plasmids. High-throughput sequencing of the targeted protospacer region amplified from plasmids extracted from pooled transformant colonies revealed depletion of 887 out of 16 384 library members in cells carrying the CcCas9 locus compared to control cells (Supplementary File S2). Most of depleted variants had a 5′-NNNNGNA-3′ sequence, indicating that CcCas9 prefers purines at positions 5 and 7 of the non-target DNA strand downstream of the protospacer (Figure 2B).

In vitro cleavage of DNA by CcCas9

Based on the interference screening experiments results we proceeded to reconstitute CcCas9 DNA cleavage *in vitro*. A recombinant CcCas9 was purified (Supplementary Figure S2) and tested for its ability to cleave linear DNA PAM libraries containing a target site flanked with seven randomized nucleotides (Figure 3A). Since *C. cellulolyticum* H10 was isolated from decayed grass in a compost pile (17), we first performed DNA cleavage reactions at 33°C, the reported optimal growth temperature (17), but did not detect any cleavage. The change of reaction temperature to 45°C led to observable library DNA cleavage. Uncleaved DNA fragments as well as a negative control (original DNA PAM library incubated with DNA cleavage reaction components in the absence of crRNA) were sequenced using the Illumina platform. Comparison of PAM variants representation in experimental and control samples allowed us to determine PAM sequences depleted in the presence of the CcCas9 effector complex. The analysis revealed that recombinant CcCas9 in complex with *in vitro* synthesized tracrRNA and crRNA was able to cleave DNA targets with 'NNNNGNA' PAM at the 3′-flank, in agreement with results obtained during *in vivo* interference screening (Figure 3B), although A at the 7th position was less conserved comparing to G at the 5th position.

To validate CcCas9 PAM sequence preferences, single-nucleotide substitutions in the deduced consensus PAM sequence were introduced and individually tested for cleavage efficiency (Figure 3C). The results confirmed the importance of a G at the fifth position and a less strict preference for an A at the seventh position (Figure 3C). To further investigate CcCas9 PAM sequence preferences, in particular, to identify individual sequences representing functional PAMs and the relative activity of each sequence, we used the PAM wheel approach developed by Leenay *et al.* (18) for results visualization. The PAM wheel confirmed the 5′-NNNNGNA-3′ motif with a moderate preference for an A at the seventh position but also revealed a slight bias for an A in addition to G in the fifth position (Figure 3D).

We next tested CcCas9 DNA cleavage activity on different targets flanked by the 5′-NNNNGNA-3′ consensus PAM as well as 5′-NNNNGNN-3′ PAM sequences (Figure 3E). Several 20-bp target sites with CcCas9 PAM in a 1592-bp PCR fragment of human grin2b gene were selected, the corresponding crRNAs synthesized, and *in vitro* cleavage reactions were performed with recombinant CcCas9 charged with these crRNAs. Control crRNAs recognizing sequences flanked by PAMs with no G at the fifth position were also tested. As can be seen from Figure 3E, CcCas9 did not recognize targets flanked by control sequences with substitutions of G at the fifth position. On the other hand, the CcCas9 nuclease recognized and cleaved not only targets with 5′-NNNNGNA-3′ consensus PAM, but also targets flanked by 5′-NNNNGNN-3′ sequences, confirm-

ing that 5′-NNNNGNN-3′ PAMs are functional. Similar results were obtained when in *in vitro* DNA cleavage by CcCas9 was performed using a supercoiled plasmid carrying the cloned grin2b gene fragment (Supplementary Figure S3). The cleavage efficiency of CcCas9 on different DNA targets varied significantly, which is likely a combination of contributions by protospacer sequences and by identity of 'N' nucleotides in the PAM. Overall, based on plasmid transformation interference screening results and *in vitro* DNA cleavage data, we conclude that CcCas9 recognizes a two-nucleotide 5′-NNNNGNA-3′ PAM, with requirement for an A in seventh position being not very stringent. To the best of our knowledge, this PAM is distinct from PAM sequences of known Cas9 nucleases.

Experiments described above were conducted using the tripartite system composed of CcCas9, crRNA and tracrRNA. To simplify the CcCas9 DNA-cleavage process, we sought to design sgRNA, a single guide RNA where crRNA is fused to tracrRNA. Several sgRNA variants were tested, but none were active *in vitro* (Supplementary Figure S4). Thus, the CcCas9 DNA minimal cleavage system to date consists of three components: CcCas9 nuclease, tracrRNA and crRNA (Figure 4). Additional studies might reveal the requirements for a functional sgRNA sequence in this system.

One of the possible applications of CcCas9 is genome modification of its host, *C. cellulolyticum.* To facilitate further use of CcCas9 for editing of *C. cellulolyticum* via the single-nick-assisted HR strategy proposed by Xu *et al.* (11), we generated a CcCas9 nickase version by mutating the aspartic acid D8 to alanine in the active site of CcCas9 RuvC nuclease domain. The incubation of D8A CcCas9 mutant with a double-stranded DNA target in the presence of crRNA and tracrRNA led to cleavage of only one DNA strand, as expected (Supplementary Figure S5).

Activity of CcCas9 at different temperatures

Based on the initial observations showing that DNA cleavage by CcCas9 is temperature-dependent, we decided to determine the dependence of its nuclease activity on temperature. Incubation of CcCas9, crRNA, tracrRNA and plasmid carrying a protospacer flanked by consensus PAM sequence 5′-ACAGGTA-3′ at different temperatures led to efficient cleavage of the target in a temperature range of 25–45°C with maximal cleavage at 40°C (Figure 5A and B). CcCas9 cleavage of a linear DNA fragment carrying the same target site showed similar temperature activity profile.

Given the observed differences in CcCas9 *in vitro* DNA cleavage efficiency at room temperature and at 37°C, we compared the CcCas9 CRISPR–Cas II-C system interference activity at 22°C and 37°C. To this end, we used an equimolar mixture of five PUC19-based plasmids carrying a protospacer matching the first spacer in the CRISPR array and flanked by 5′-ACAGGTA-3′, 5′-CGGTGTA-3′, 5′-TGAAGAA-3′ and 5′-ATTGGAA-3′ CcCas9 PAM variants and a 5′-TTCATAT-3′ sequence as a 'no PAM' control. This 5-members PAM library was transformed into competent *E. coli* cells carrying pACYC184_CcCas9_locus plasmid or pACYC184 vector as a control. Cells were plated on LB medium supplemented with ampicillin and chloramphenicol and grown for 18 h at either 22 or 37°C. Plasmid DNA was purified from colonies formed at each temperature and HTS of PAM-containing regions was performed to determine the changes in representation of library members (Figure 6, Supplementary Table S3). Analysis of HTS results showed the decrease in the frequency of 5′-NNNNGNA-3′ PAM-containing plasmids in cells carrying the CcCas9 locus due to interference and corresponding increase of the 'no PAM' plasmid representation at 37°C as well as at 22°C compared to control (Supplementary File S1, Supplementary Figure S6). The observed effect was stronger in colonies formed at 37°C than at 22°C. Plasmids with different 5′-NNNNGNA-3′ PAM sequences showed different depletion levels. Thus, the temperature dependence of *C. cellulolyticum* CRISPR–Cas II-C system can be observed in bacteria as well as *in vitro*.

CONCLUSION

Despite the extensive use of Cas9 nucleases for genome engineering, to date, only several Cas9 orthologs can be considered as well-characterized. Given the diversity of Type II CRISPR–Cas systems, Cas9 orthologs can show significant variations in PAM requirements, specificity and other biochemical properties. In this work, we functionally characterized CRISPR–Cas system from *Clostridium cellulolyticum* H10. When introduced in *E. coli*, the *C. cellulolyticum* CRISPR–Cas system shows high levels of crRNA expression, as well as interference against plasmid transformation. The *C. cellulolyticum* Cas9 effector, CcCas9, is a Type II-C endonuclease and thus has a relatively small (compared to other Type II effector proteins) molecular weight. This nuclease in complex with tracrRNA and crRNA actively cleaves DNA targets flanked by two-nucleotide PAM sequence 5′-NNNNGNA-3′. Most other small Type II-C Cas9 effectors have more complex PAM requirements, i.e. NmeCas9, CjeCas9 and GeoCas9 require, 5′-NNNNGNTT-3′, 5′-NNNNRYAC-3′ and 5′-NNNNCNAA-3′, respectively (7,19–20). The simple, two-nucleotide PAM of CcCas9 may thus be considered as an advantage for future biotechnology applications. Whereas further studies are needed to check the ability of CcCas9 to edit eukaryotic genomes, we envision that the CRISPR–Cas system characterized here potentially can be conveniently used as an instrument for *C. cellulolyticum* H10 genome engineering.

DATA AVAILABILITY

Raw sequencing data have been deposited with the National Center for Biotechnology Information Sequence Read Archive under BioProject ID PRJNA554628.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the Center for Precision Genome Editing and Genetic Technologies for Biomedicine, IGB RAS, for the equipment

FUNDING

Ministry of Science and Higher Education of the Russian Federation Subsidy Agreement 14.606.21.0006 [RFMEFI60617X0006]. Funding for open access charge: grant 075-15-2019-1661 from the Ministry of Science and Higher Education of the Russian Federation.
Conflict of interest statement. None declared.

REFERENCES

1. Cong,L., Ran,F.A., Cox,D., Lin,S., Barretto,R., Habib,N., Hsu,P.D., Wu,X., Jiang,W., Marraffini,L.A. *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science (New York, N.Y.)*, **339**, 819–823.
2. Jiang,W., Bikard,D., Cox,D., Zhang,F. and Marraffini,L.A. (2013) RNA-guided editing of bacterial genomes using CRISPR–Cas systems. *Nat. Biotechnol.*, **31**, 233–239.
3. Ronda,C., Pedersen,L.E., Sommer,M.O. and Nielsen,A.T. (2016) CRMAGE: CRISPR optimized MAGE recombineering. *Sci. Rep.*, **6**, 19452.
4. Jiang,Y., Qian,F., Yang,J., Liu,Y., Dong,F., Xu,C., Sun,B., Chen,B., Xu,X., Li,Y. *et al.* (2017) CRISPR-Cpf1 assisted genome editing of Corynebacterium glutamicum. *Nat. Commun.*, **8**, 15179.
5. Jinek,M., Chylinski,K., Fonfara,I., Hauer,M., Doudna,J.A. and Charpentier,E. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (New York, N.Y.)*, **337**, 816–821.
6. Ran,F.A., Cong,L., Yan,W.X., Scott,D.A., Gootenberg,J.S., Kriz,A.J., Zetsche,B., Shalem,O., Wu,X., Makarova,K.S. *et al.* (2015) In vivo genome editing using Staphylococcus aureus Cas9. *Nature*, **520**, 186–191.
7. Kim,E., Koo,T., Park,S.W., Kim,D., Kim,K., Cho,H.Y., Song,D.W., Lee,K.J., Jung,M.H., Kim,S. *et al.* (2017) In vivo genome editing with a small Cas9 orthologue derived from Campylobacter jejuni. *Nat. Commun.*, **8**, 14500.
8. Fonfara,I., Le Rhun,A., Chylinski,K., Makarova,K.S., Lecrivain,A.L., Bzdrenga,J., Koonin,E.V. and Charpentier,E. (2014) Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR–Cas systems. *Nucleic Acids Res.*, **42**, 2577–2590.
9. Desvaux,M. (2005) Clostridium cellulolyticum: model organism of mesophilic cellulolytic clostridia. *FEMS Microbiol. Rev.*, **29**, 741–764.
10. Xu,T., Li,Y., Shi,Z., Hemme,C.L., Li,Y., Zhu,Y., Van Nostrand,J.D., He,Z. and Zhou,J. (2015) Efficient genome editing in Clostridium cellulolyticum via CRISPR–Cas9 nickase. *Appl. Environ. Microbiol.*, **81**, 4423–4431.
11. Xu,T., Li,Y., He,Z., Van Nostrand,J.D. and Zhou,J. (2017) Cas9 nickase-assisted RNA repression enables stable and efficient manipulation of essential metabolic genes in Clostridium cellulolyticum. *Front. Microbiol.*, **8**, 1744.
12. Pyne,M.E., Bruder,M.R., Moo-Young,M., Chung,D.A. and Chou,C.P. (2016) Harnessing heterologous and endogenous CRISPR–Cas machineries for efficient markerless genome editing in Clostridium. *Sci. Rep.*, **6**, 25666.
13. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
14. Maxwell,C.S., Jacobsen,T., Marshall,R., Noireaux,V. and Beisel,C.L. (2018) A detailed cell-free transcription-translation-based assay to decipher CRISPR protospacer-adjacent motifs. *Methods*, **143**, 48–57.
15. Pougach,K., Semenova,E., Bogdanova,E., Datsenko,K.A., Djordjevic,M., Wanner,B.L. and Severinov,K. (2010) Transcription, processing and function of CRISPR cassettes in Escherichia coli. *Mol. Microbiol.*, **77**, 1367–1379.
16. Zhang,Y., Heidrich,N., Ampattu,B.J., Gunderson,C.W., Seifert,H.S., Schoen,C., Vogel,J. and Sontheimer,E.J. (2013) Processing-independent CRISPR RNAs limit natural transformation in Neisseria meningitidis. *Mol. Cell*, **50**, 488–503.
17. Petitdemange,E., Caillet,F., Giallo,J. and Gaudin,C. (1984) Clostridium cellulolyticum sp. nov., a cellulolytic, mesophilic: species from decayed grass. *Int. J. Syst. Evol. Microbiol.*, **34**, 155–159.
18. Leenay,R.T., Maksimchuk,K.R., Slotkowski,R.A., Agrawal,R.N., Gomaa,A.A., Briner,A.E., Barrangou,R. and Beisel,C.L. (2016) Identifying and visualizing functional PAM diversity across CRISPR–Cas systems. *Mol. Cell*, **62**, 137–147.
19. Ciaran,M.L., Thomas,J.C. and Gang,Bao (2016) The Neisseria meningitidis CRISPR–Cas9 system enables specific genome editing in mammalian cells. *Mol. Ther.*, **24**, 645–654.
20. Harrington,L.B., Paez-Espino,D., Staahl,B.T., Chen,J.S., Ma,E., Kyrpides,N.C. and Doudna,J.A. (2017) A thermostable Cas9 with increased lifetime in human plasma. *Nat. Commun.*, **8**, 1424.

Supplementary File S1

Figure S1. CcCas9 and SaCas9 protein alignment. The putative protein domains are shown by colored blocks.

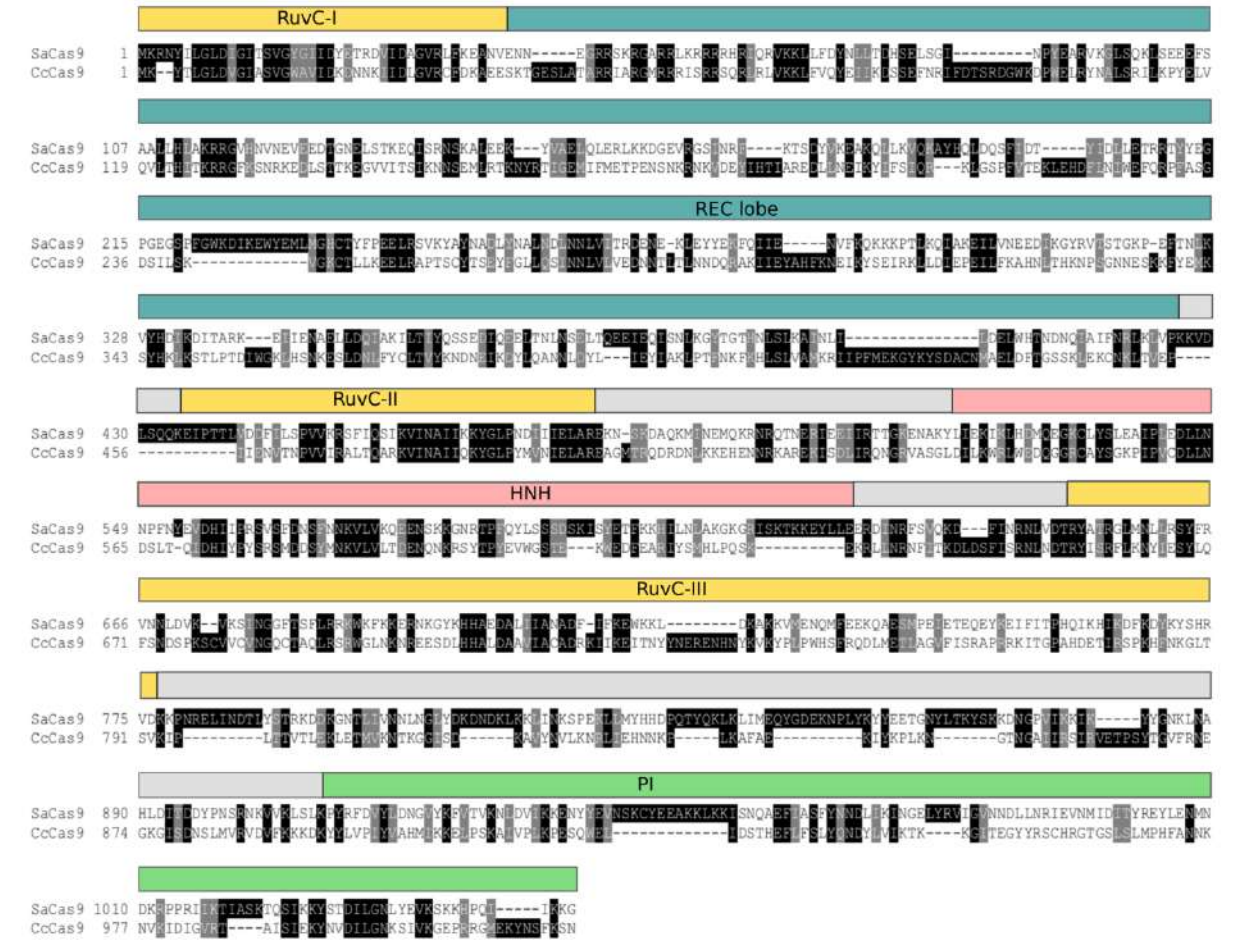


Table S1. DNA sequences used in this study.

<table>
<tr><td rowspan="2">PUC19 7N fragment</td><td>ctatgaccatgattacgccaagctNNNNNNNGCATTCATATCATCGCTTGCTTATTTTTTA
cccgggtaccgagctcga</td></tr>
<tr><td>https://benchling.com/s/seq-9AUgIqLVZMC8P2fi0fHy</td></tr>
<tr><td>Pacyc184_CcCas9_locus plasmid</td><td>https://benchling.com/s/seq-JtQiuWbwPZSgoYDPVQdo</td></tr>
<tr><td>pET21a_CcCas9</td><td>https://benchling.com/s/seq-vDHDkNAp45fykKGHOVCk</td></tr>
<tr><td>Plasmids with different CcCas9 target sites (PUC19 with human grin2b gene fragment)</td><td>https://benchling.com/s/seq-UBYcoVMlXuLO5PUYrmMc</td></tr>
<tr><td rowspan="2">DNA fragment with different CcCas9 target sites (human grin2b gene fragment)</td><td>https://benchling.com/s/seq-FvBzwSlYhkYoQthrlbZw</td></tr>
<tr><td>GAGAGAGATGGCCAAGGCTTATATTCTATAGAGCATTATGTCCTTAGTTTGATGCATAGAA
TAAGATTTAGGGTCATATGTGGAAGTAAAAAGGAAGGAGTTCTTTGTAGGTAAAAGGTGGC
AAATTATATGAAAATACGGTATCAGTCATTTTAGGGAAGTCACGACTATAGGATGGCATCA
GGAAAAAAAAAGGAACATTTTTCAAATGTGGCTCTAACATTACTTCAGCTGCTAATGGTAT
TTGTTTAAGTTTCTGTATTTTGGTGTATAAATAGATTGGAGTAATATGTGTTCCTTATAAT
AATTGGTTATATGAGAGGCAGTTCCACGTAGTGTAATAGAACACATATTGGAATACAAAAG
TCAGAAGATCTGGGTTCAGGTTTGTTTCACTTAATTGATTGGTTCATGGTCTTAGAACACT
TAGCTTCTCTGAGCCTTGGCGTCAACATTTATAAAAATGGTGATAATAATGTTTTTCTTAT
TTTATTCCCTACTGGGTCATTGTAAGGATCAATTGAGGCAATGTTTTAAAACTACTAGTCA
TGTATCAGTTGTTCTTGTAGTTTAATATTAAGAGCCAGATACTAACAAGGTTACTAAAGAA
TTTTCTGGCTGTTGTCCTCATTGAGGCAAACATAAGGTGAAGGCAGCAAGAATGCAGGGCT
TGTGTACTTATAGCCCCCCACATCCAGTTTATCCAGCCCATGTTCTGTTGCTCACCTCTGC
TGAGCACGTTTTTCTGCTCACTTTGTCTGGCCTTGCTTTCCTTCAGCCCAAGAACAGTACA
AGGGTGGGCTGTAACAGGAGGGCCAGGAGATTTGTGTATGCATACTCGCATGGCTACCTGG
ACCACTCACAACCTCTTTTCCTCCTTTGTCTCTGCCTGTAGCTGCCAATGACTATAGCAAT
AGCACCTTTTATTGCCTTGTTCAAGGATTTCTGAGGCTTTTGAAAGTTTCATTTTCTCTCA
TTCTGCAGAGCAAATACCAGAGATAAGAGAGTAGGCTGGTAGATGGAGTTGGGTTTGGTGC
TCAATGAAAGGAGATAAGGTCCTTGAATTGCAGTATCTAGCCTCTTCTAAGACAGGTTACG
TGATGTAGATCCTATTTTAACATGCTCTTTCTTTGTGTTTGCAGGGAGTCGACGAGTTGAA
GATGAAGCCCAGAGCGGAGTGCTGTTCTCCCAAGTTCTGGTTGGTGTTGGCCGTCCTGGCC
GTGTCAGGCAGCAGAGCTCGTTCTCAGAAGAGCCCCCCCAGCATTGGCATTGCTGTCATCC
TCGTGGGCACTTCCGACGAGGTGGCCATCAAGGATGCCCACGAGAAAGATGATTTCCACCA
TCTCTCCGTGGTACCCCGGGTGGAACTGGTAGCCATGAATGAGACCGACCCAAAGAGCATC
ATCACCCGCATCTGTGATCTCATGTCTGACCGGAAGATCCAGGGGGTGGTGTTTGCTGATG
ACACAGACCAGGAAGCCATCGCCCAGATCCTCGATTTCATTTCAGCACAGACTCTCACCCC
CATCCTGGGCATCCACGGGGGCTCCTCTATGATAATGGCAGATAAGGTAAAAAGGGGCTGC
AGGGAG

target site PAM DNA cleavage products

target1 TGAGGCAAACATAAGGTGAA GGCAGCA 648bp + 944bp
target2 TAACAGGAGGGCCAGGAGAT TTGTGTA 821bp + 771bp
target3 AGCAATAGCACCTTTTATTG CCTTGTT 926bp + 666bp
target4 CGACTCCCTGCAAACACAAA GAAAGAG 1132bp + 460bp
target5 ACGGCCAACACCAACCAGAA CTTGGGA 1196bp + 396bp
target6 GAACGAGCTCTGCTGCCTGA CACGGCC 1226bp + 366bp
target7 GGAAAAGAGGTTGTGAGTGG TCCAGGT 858bp + 734bp
target8 TATAGTCATTGGCAGCTACA GGCAGAG 893bp + 699bp
target9 TGTAACAGGAGGGCCAGGAG ATTTGTG 819bp + 773bp
target10 CTACATCACGTAACCTGTCT TAGAAGA -
target11 TCCGCTCTGGGCTTCATCTT CAACTCG -
target12 ACAAGGGTGGGCTGTAACAG GAGGGCC 807bp + 785bp
target13 CACCAACCAGAACTTGGGAG AACAGCA -
target14 ATCTACATCACGTAACCTGT CTTAGAA 1091bp + 501bp
target15 AAGAGGCTAGATACTGCAAT TCAAGGA 1066bp + 526bp
target16 GATAAGAGAGTAGGCTGGTA GATGGAG 1014bp + 578bp</td></tr>
</table>

	target17 TATCTCCTTTCATTGAGCAC CAAACCC -

Locus_F	caagaagatcatcttattaatcagataaaatatttctagaTATGGTAGCAAATATGAATGTAAAGTG
Locus_R	tagcaatttaactgtgataaactaccgcattaaagcttGTACTATTTGAGGGTCGTAGTTTGTGGATA TAATTTC
CcCas9_F	aaggagatatacatatggctagcATGAAATATACATTAGGTCTTGATGTTG
CcCas9_R	tggtggtggtggtgctcgagGTTGGATTTGAAACTATTATATTTCTCCATCCCACG
PUC19_F	cccgggtaccgagctcga
PUC19_R	agcttggcgtaatcatggtcatag
Library_f	ctatgaccatgattacgccaagctNNNNNNNGCATTCATATCATCGCTTGCTTATTTTTTAcccgggt accgagctcga
Library_r	tcgagctcggtacccgggt
M13_f	GTTGTAAAACGACGGCCAGTG
M13_r	AGCGGATAACAATTTCACACAGGA

Table S2. RNA sequences used in this study.

	Sequence:
CcCas9 crRNA	**GGG**uaucuccuuucauugagcacGUUAUAGCUCCAAUUCAGGCUCCGAUAUGCUAUAAU
CcCas9 tracrRNA	**GGG**AUUAUGGCAUAUCGGAGCCUGAAUUGUUGCUAUAAUAAGGUGCUGGGUUUAGCCC AGACCGCCAAGUUAACCCCGGCAUUUAUUGCUGGGGUAUCUUGUUUU

SpCas9 crRNA	**GGG**uaucuccuuucauugagcacGUUUUAGAGCUAUGCUGUUUUGAAUGGUCCCAAAAC
SpCas9 tracrRNA	**GGG**AACCAUUCAAAACAGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUUUUU
CcCas9 sgRNA 1	uaucuccuuucauugagcacGUUAUAGCUCCAAUUCAGGCUCCGAUAUGCUAUAAUGAAAAUUAUGGCAUAUCGGAGCCUGAAUUGUUGCUAUAAUAAGGUGCUGGGUUUAGCCCAGACCGCCAAGUUAACCCCGGCAUUUAUUGCUGGGG
CcCas9 sgRNA 2	uaucuccuuucauugagcacGUUAUAGCUCCAAUUCAGGCUCCGAUAUGAAAAUAUCGGAGCCUGAAUUGUUGCUAUAAUAAGGUGCUGGGUUUAGCCCAGACCGCCAAGUUAACCCCGGCAUUUAUUGCUGGGG
CcCas9 sgRNA 3	uaucuccuuucauugagcacGUUAUAGCUCCAAUUCAGGCUCCGAAAGGAGCCUGAAUUGUUGCUAUAAUAAGGUGCUGGGUUUAGCCCAGACCGCCAAGUUAACCCCGGCAUUUAUUGCUGGGG
CcCas9 sgRNA 4	uaucuccuuucauugagcacGUUAUAGCUCCAAUUCAGGAAACUGAAUUGUUGCUAUAAUAAGGUGCUGGGUUUAGCCCAGACCGCCAAGUUAACCCCGGCAUUUAUUGCUGGGG
CcCas9 sgRNA 5	uaucuccuuucauugagcacGUUAUAGCUCCAAUUCAGGCUCCGAAAGGAGCCUGAAUUGUUGCUAUAAUAAGGUGCUGGGUUUAGCCCAGACCGCCAAGUUAA

DR sequences colored in red. GGG sequences – consequence of T7 RNA polymerase RNA synthesis are in bold

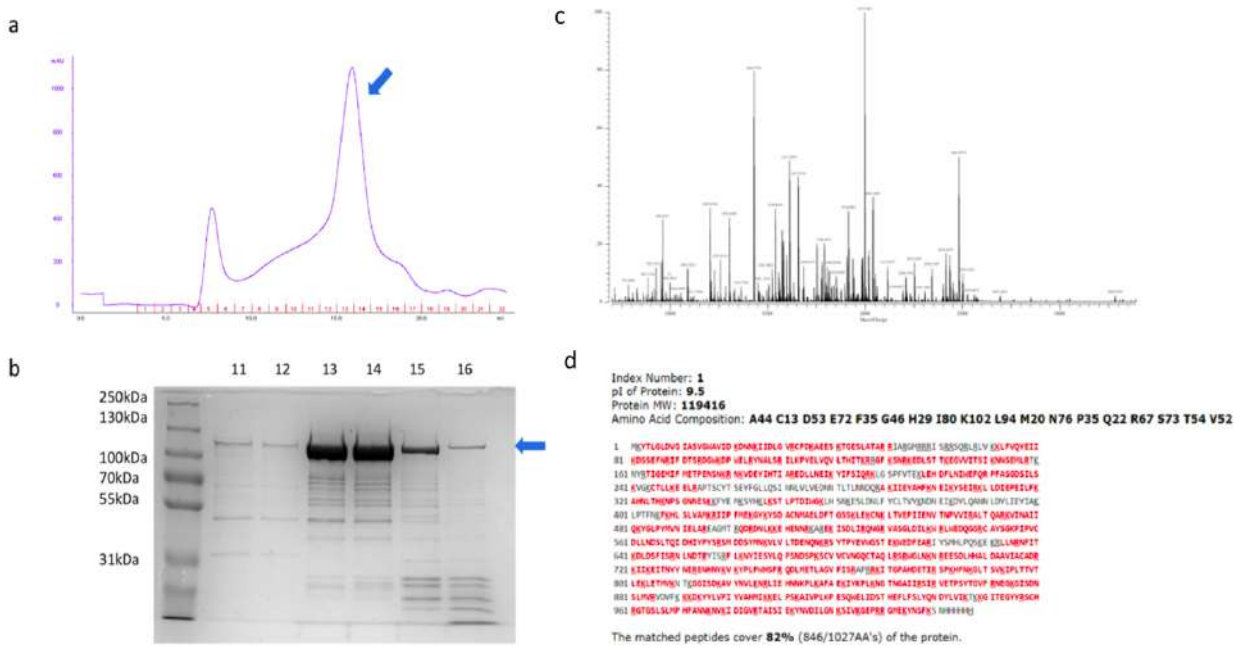


Figure S2. CcCas9 recombinant protein purification.

a. Size exclusion chromatography elution of CcCas9 protein. Monomer fraction is marked with blue arrow. The fractions numbers are written along x-axis in red.
b. SDS PAGE gel electrophoresis of size exclusion chromatography fractions 11-16. CcCas9 bands position is showed with the blue arrow.
c. Mass spectrumof tryptic hydrolyzate of gel strip corresponding to CcCas9 protein recorded by FT ICR MS (Varian) in positive MALDI mode
d. The protein sequence coverage determined by http://prospector.ucsf.edu/prospector/mshome.htm

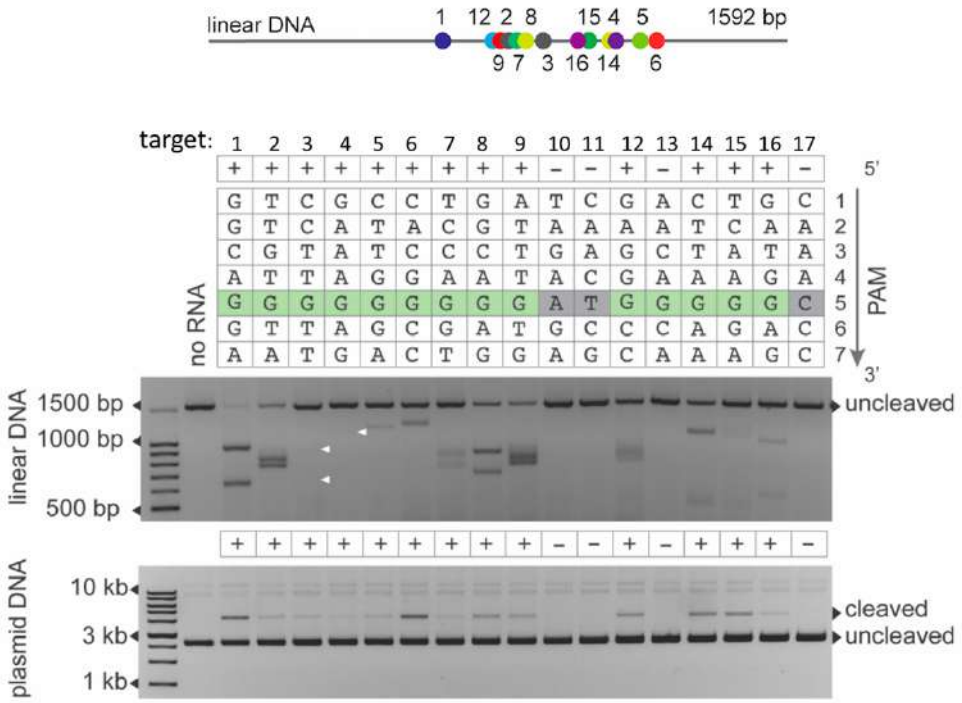


Figure S3. *In vitro* DNA cleavage of different 20-bp sites by CcCas9.
The PAM sequences corresponding to the targets are shown in the table. The +/- signs signify, correspondingly, whether cleavage was or was not observed. The conservative G in the 5th position is indicated by green color. Below two gels showing results of *in vitro* cleavage of targets with indicated PAMs on linear DNA and on plasmid DNA are presented. The white arrows indicate the positions of poorly visible bands. Above, a scheme showing the relative positions of the targeted DNA sites on the linear DNA fragment is presented.

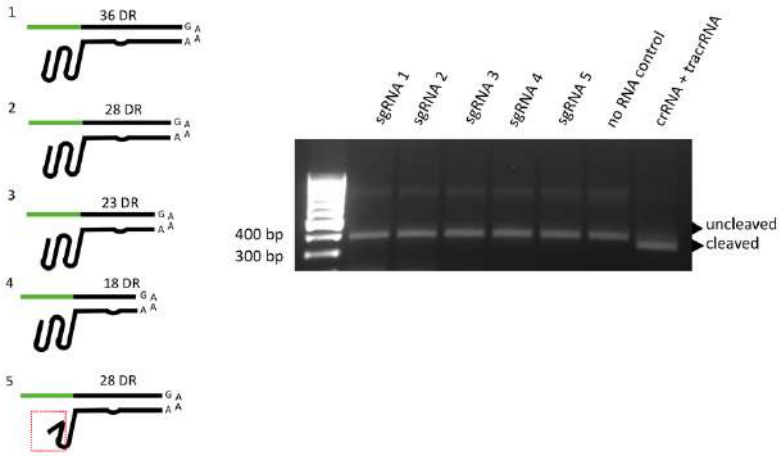


Figure S4. CcCas9 sgRNA design.
Five variants of sgRNA were used for CcCas9 DNA cleavage reactions *in vitro* (left). To estimate DNA cleavage efficiency reactions products were loaded to 1.5% agarose gel. In opposite to crRNA-tracrRNA-CcCas9 complex sgRNA-CcCas9 complexes were not able to cleave DNA targets *in vitro*. The experiment was conducted three times with three independently synthetized sgRNAs sets, which length was tested by denaturing PAGE electrophoresis.

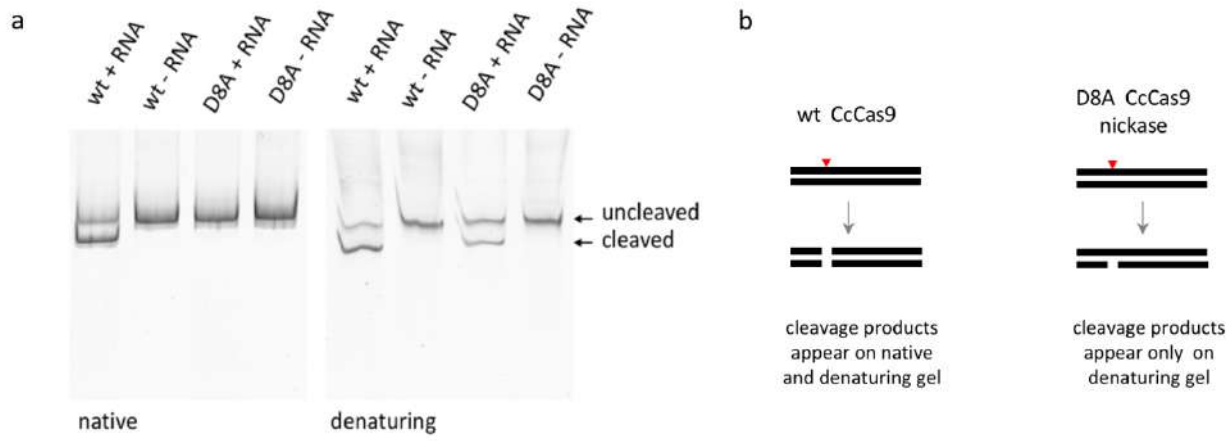


Figure S5. CcCas9 D8A mutant is a nickase.

a. wt CcCas9 protein and D8A CcCas9 DNA cleavage reactions results. Linear DNA cleavage products were loaded to native (left) or denaturing (right) PAGE gel.
b. Illustration of expectable DNA cleavage products in case of wt CcCas9 protein and a D8A CcCas9 nickase version.

Table S3. The influence of temperature on *Clostridium cellulolyticum H10* CRISPR Cas II-C locus interference.

Fractions of plasmid library members extracted from cells carrying pACYC184 (control) or pACYC184_CcCas9_locus grown at 37°C or at 22°C.

		pACYC184 carrying cells				
		37C rep1	37C rep2	37C rep3	37C mean	STD
PAM 1	ACAGGTA	0,117	0,115	0,119	0,117	0,001
PAM 2	CGGTGTA	0,178	0,176	0,175	0,177	0,001
PAM 3	TGAAGAA	0,197	0,193	0,190	0,193	0,002
PAM 4	ATTGGAA	0,263	0,264	0,265	0,264	0,001
no PAM	TTCATAT	0,246	0,252	0,250	0,249	0,002
		pACYC184_CcCas9_locus carrying cells				
		37C rep1	37C rep2	37C rep3	37C mean	STD
PAM 1	ACAGGTA	0,033	0,049	0,052	0,045	0,007
PAM 2	CGGTGTA	0,082	0,125	0,127	0,111	0,018
PAM 3	TGAAGAA	0,079	0,130	0,132	0,114	0,021
PAM 4	ATTGGAA	0,052	0,050	0,048	0,050	0,001
no PAM	TTCATAT	0,754	0,646	0,641	0,681	0,045

		pACYC184 carrying cells				
		22C rep1	22C rep2	22C rep3	22C mean	STD
PAM 1	ACAGGTA	0,118	0,117	0,120	0,118	0,001
PAM 2	CGGTGTA	0,176	0,175	0,176	0,176	0,000
PAM 3	TGAAGAA	0,189	0,191	0,193	0,191	0,001
PAM 4	ATTGGAA	0,268	0,268	0,264	0,266	0,002
no PAM	TTCATAT	0,249	0,249	0,248	0,249	0,000
		pACYC184_CcCas9_locus carrying cells				
		22C rep1	22C rep2	22C rep3	22C mean	STD
PAM 1	ACAGGTA	0,071	0,059	0,061	0,064	0,005
PAM 2	CGGTGTA	0,199	0,192	0,200	0,197	0,003
PAM 3	TGAAGAA	0,165	0,150	0,157	0,158	0,005
PAM 4	ATTGGAA	0,071	0,062	0,062	0,065	0,004
no PAM	TTCATAT	0,494	0,537	0,520	0,517	0,016

Bacterial interference experiments conducted at different temperatures: statistical analysis of the results.

ANOVA (analysis of variants) results for “no PAM” plasmid fraction:

	Df	SumSq	MeanSq	F value	Pr(>F)
temperature	1	0.0202	0.0202	17.71	0.00296 **
presence of CRISPR CcCas9 locus	1	0.3668	0.3668	322.11	9.52e-08 ***
temperature: presence of CRISPR CcCas9 locus	1	0.0198	0.0198	17.43	0.00310 **
Residuals	8	0.0091	0.0011		

The difference of “no PAM” plasmid fraction in cells carrying CRISPR CcCas9 locus comparing to control cells is significant: $F(1,8) = 322.11$ p-value = 9.52e-08 (< 0.05)

The difference between “no PAM” plasmid fraction in cells grown at different temperature is significant: $F(1,8) = 17.71$ p-value = 0.00296 (< 0.05)

ANOVA also indicates the interaction between “temperature” factor and “presence of CRISPR CcCas9 locus” factor. The effect is illustrated on graph below. The increasing of “no PAM” plasmid fraction, which illustrates the overall interference level, significantly higher in cells carrying CRISPR Cas locus grown at 37°C than at 22°C.

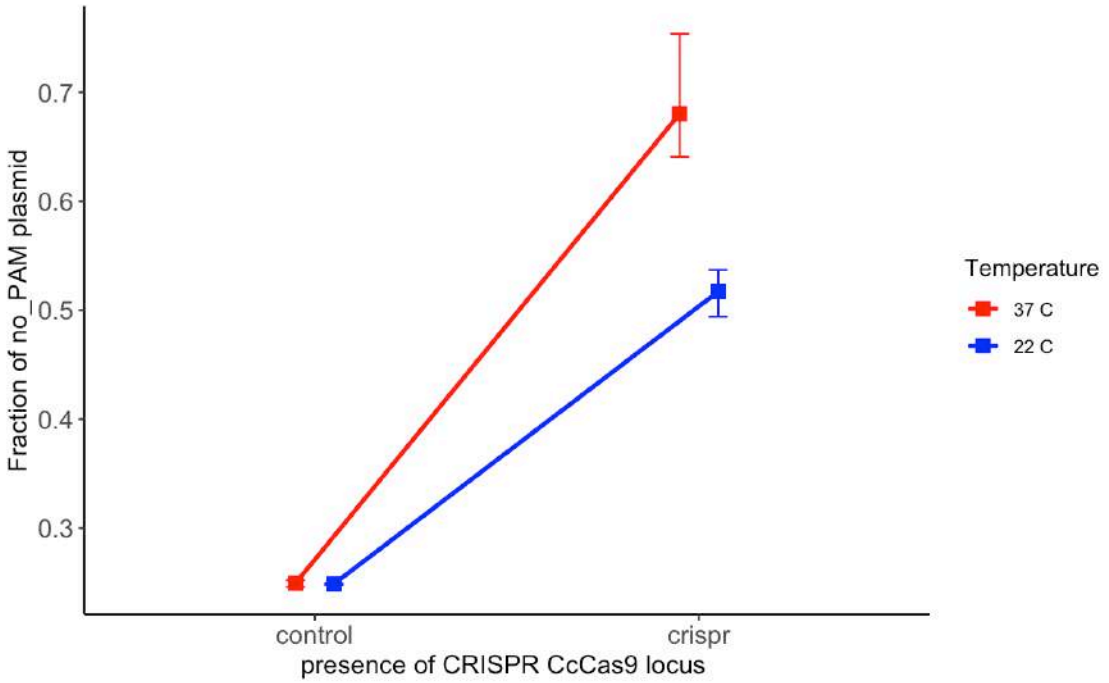


Figure S6. Fraction of "no PAM" plasmid in *E.coli* cells carrying *Clostridium cellulolyticum H10* CRISPR Cas II-C system grown at different temperatures.

A plasmid library composed of five members carrying a protospacer matching the first spacer in the CRISPR array and flanked by 5'-ACAGGTA-3'(PAM 1), 5'-CGGTGTA-3' (PAM 2), 5'-TGAAGAA-3' (PAM 3), 5'-ATTGGAA-3' (PAM 4), and 5'-TTCATAT-3' (no PAM) sequence was transformed in *E. coli* cells carrying a plasmid with the *C. cellulolyticum H10* CRISPR-Cas II-C locus or a control plasmid. Cells were plated on LB media supplemented with antibiotics, and grown for 18 h at 37 °C or 22 °C. The plasmid DNA was extracted from grown colonies and HTS was used to estimate the representation of each library member. "no PAM" plasmid fraction illustrates the overall interference level in cells. Cells carrying *Clostridium cellulolyticum H10* CRISPR Cas II-C system indicated as "crispr" and cells carrying an empty pACYC184 as "control". The results, obtained on cells grown at 37°C and 22°C showed in red and blue color correspondingly. Mean values and standard deviations obtained from three independent experiments are shown.

Tukey multiple comparisons of means:

	diff	lwr	upr	**p adj**
22C:control-37C:control	-0.0006666667	-0.08890101	0.08756767	0.9999945
37C:crispr-37C:control	0.4310000000	0.34276566	0.51923434	**0.0000013**
22C:crispr-37C:control	0.2676666667	0.17943233	0.35590101	0.0000490
37C:crispr-22C:control	0.4316666667	0.34343233	0.51990101	0.0000013
22C:crispr-22C:control	0.2683333333	0.18009899	0.35656767	**0.0000481**
22C:crispr-37C:crispr	-0.1633333333	-0.25156767	-0.07509899	**0.0015743**

The difference of "no PAM" plasmid fraction in cells carrying the CRISPR CcCas9 locus compared to control is significant at 37 °C (p-value= **0.0000013**), as well as at 22 °C (p-value = **0.0000481**)
There is the significant difference between "no PAM" plasmid fraction in cells carrying CRISPR CcCas9 locus at 37 °C and at 22 °C (p-value = **0.0015743**).

Chapter II

PpCas9 from *Pasteurella pneumotropica* - a compact Type II-C Cas9 ortholog active in human cells

Introduction

This chapter continues the characterization of small-sized Cas9 orthologs described in Chapter 1. Here were characterized two CRISPR-Cas Type II-C systems, from *Defluviimonas sp.20V17* – a bacterium inhabiting deep sea hydrothermal vents - and from *Pasteurella pneumotropica* - a gram-negative bacterium isolated from multiple mammalian species. Using experiments in *E. coli* heterologously expressing these defense systems genes as well as *in vitro* we showed that DfCas9 and PpCas9 are active nucleases of small size with novel, 5'- NNRNAY-3' and 5'-NNNNRT-3' PAMs. To test the activity of DfCas9 and PpCas9 in human cells their genes as well as sequences coding of corresponding guide RNAs were cloned into plasmids under regulation of eukaryotic promoters. The analysis of genomic DNA of human cells transfected by plasmids carrying PpCas9 CRISPR system showed that it actively introduces indels in DNA sites flanked with 5'-NNNNRTT-3' PAM. The high throughput sequencing analysis of possible off-target sites showed that PpCas9 is specific enough to be considered as a promising candidate for further use in genome editing.

Contribution

I conceived the study. I designed and participated in all experiments, in particular: designed the plasmids used in the work; performed RNA-Seq to determine DfCas9 guiding RNAs sequences and analyzed its results; performed a significant part of biochemical assays and tested the activity of the nucleases in human cells with the help of co-authors. Aleksandra Vasileva, an equally contributing author, performed all bioinformatics analysis (except of RNA-Seq analysis), performed significant part of *in vitro* experiments, and conducted the study of PpCas9 specificity in human cells.

I and Aleksandra Vasileva prepared all Figures. I drafted the manuscript with contributions and insights from all authors. I would like to thank all authors for their help.

PpCas9 from *Pasteurella pneumotropica* - a compact Type II-C Cas9 ortholog active in human cells

Iana Fedorova[1]*†, Aleksandra Vasileva[1]†, Polina Selkova[1], Marina Abramova[2], Anatolii Arseniev[2], Georgii Pobegalov[2], Maksim Kazalov[2], Olga Musharova[1,3], Ignatiy Goryanin[1], Daria Artamonova[1], Tatyana Zyubko[2], Sergey Shmakov[4], Tatyana Artamonova[2], Mikhail Khodorkovskii[2], and Konstantin Severinov[1,5,6]*

[1]Skolkovo Institute of Science and Technology, Center of Life Sciences, Moscow, Russia
[2]Peter the Great St. Petersburg Polytechnic University, Saint Petersburg, Russia
[3]Institute of Molecular Genetics, Russian Academy of Sciences, Moscow, Russia
[4]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, USA
[5]Center for Precision Genome Editing and Genetic Technologies for Biomedicine, Institute of Gene Biology, Russian Academy of Sciences, Moscow, Russia
[6]Waksman Institute for Microbiology, Rutgers, The State University of New Jersey, Piscataway, New Jersey, USA

† Joint first authors
* To whom correspondence should be addressed:
Tel: +79062713200; Email: femtokot@gmail.com (I.F.)
Tel: +79854570284; Email: severik@waksman.rutgers.edu (K.S.)

Abstract

CRISPR-Cas defence systems opened up the field of genome editing due to the ease with which effector Cas nucleases can be programmed with guide RNAs to access desirable genomic sites. Type II-A SpCas9 from *Streptococcus pyogenes* was the first Cas9 nuclease used for genome editing and it remains the most popular enzyme of its class. Nevertheless, SpCas9 has some drawbacks including big size and restriction to targets flanked by an "NGG" PAM sequence. The more compact Type II-C Cas9 orthologs can help to overcome the size limitation of SpCas9. Yet, only a few Type II-C nucleases were fully characterized to date. Here, we characterized two Cas9 II-C orthologs, DfCas9 from *Defluviimonas sp.20V17* and PpCas9 from *Pasteurella pneumotropica*. Both DfCas9 and PpCas9 cleave DNA *in vitro* and have novel PAM requirements. We show that PpCas9 nuclease is active in human cells. This small nuclease requires an "NNNNRTT" PAM orthogonal to that of SpCas9 and thus can broaden the range of Cas9 applications in biomedicine and biotechnology.

Introduction

CRISPR-Cas are bacterial and archaeal immune systems that degrade invaders genomes using RNA-guided Cas nucleases. The CRISPR-Cas loci consist of CRISPR arrays and CRISPR-associated *cas* genes. CRISPR arrays are composed of repeats separated by intervening unique spacers. Some spacers are derived from invaders DNA and are acquired by CRISPR arrays from genetic mobile elements such as plasmids or bacteriophages (1–3). The CRISPR array is transcribed into a long precursor crRNA and further processed to mature short CRISPR RNAs (crRNAs), each containing a part of repeat and a single spacer sequence (4). Mature crRNAs bind to Cas effector proteins and guide them to regions of invader genomes complementary to crRNA spacer segment (5). The specific recognition of nucleic acid targets leads to activation of Cas effector nucleases and subsequent degradation of invader's genome (6–8).

The ability to guide Cas nucleases to DNA targets of choice using crRNAs of different sequences led to development of efficient and easy-to-use genome engineering tools (9–11). CRISPR-Cas systems vary in terms of effector complexes architecture and mechanisms of action. In CRISPR-Cas class II systems' effectors Cas9, the crRNA binding and DNA cleavage functions are combined in a single, albeit large proteins and this simplicity led to their extensive use (12).

The Cas9 effector nuclease of Type II-A CRISPR-Cas system from *Streptococcus pyogenes* was the first Cas nuclease to be successfully harnessed for genome engineering in human cells (11, 13). Despite biochemical characterization of several other Type II-A Cas9 orthologs, the SpCas9 still remains the most investigated and highly used enzyme of its class due to its high efficiency and requirements for a relatively short PAM (protospacer adjacent motif) - several nucleotides flanking the target site that are essential for efficient DNA recognition and cleavage.

The Cas9 effector proteins of Type II-C CRISPR-Cas systems have generally a smaller size than Type II-A counterparts, which allows the simultaneous delivery of Type II-C Cas9 gene and sequences coding for guide RNAs in a single adeno-associated viral (AAV) particle (14–16). Type II-C effectors from *Neisseria meningitidis* strain 8013 (NmeCas9, (17)) and strain De11444 (Nme2Cas9, (16)), *Campylobacter jejuni* (CjCas9, (14)), *Corynebacterium diphtheriae* (CdCas9, (18)), *Geobacillus stearothermophilus* (GeoCas9, (19)), and *Staphylococcus auricularis* (SauriCas9, (15)) were characterized and shown to mediate genome editing in human cells. Together with a small-sized Type II-A Cas9 from *Staphylococcus aureus* (SaCas9, (20)), these nucleases comprise a group of small Cas9 enzymes, whose use may be advantageous during the development of AAV-based genome editing platforms.

Despite the advantages of their size, small Cas9 nucleases characterized to date tend to require long PAMs - 5′-NNGRRT-3′ for SaCas9; 5′-NNNVRYAC-3′ for CjCas9; 5′-

NNNNGNTT-3′ for NmeCas9; 5′-NNRHHHY-3′ for CdCas9; 5′-NNNNCRAA-3′ for GeoCas9 -- which narrows the choice of targets available for editing. Only small orthologs Nme2Cas9 and recently found SauriCas9 require short PAM sequences 5'-NNNNCC-3'and 5'-NNGG-3' correspondingly.

Characterization of new small size nucleases with shorter PAMs will expand the tool kit of genomic editors and increase the number of editable sites. Here, we functionally characterized two small-sized Type II-C Cas9 orthologs from *Defluviimonas sp.20V17*, a bacterium inhabiting deep-sea hydrothermal vents (21, 22) and *Pasteurella pneumotropica* (*Rodentibacter pneumotropicus*), a gram-negative bacterium isolated from multiple mammalian species (23). Using *in vitro* studies and/or experiments in bacteria we show that *Defluviimonas sp.20V17* and *Pasteurella pneumotropica* CRISPR-Cas Type II-C systems encode active Cas9 nucleases that efficiently cleave target DNA with novel, 5'-NNRNAY-3' (DfCas9) and 5'-NNNNRT-3' (PpCas9) PAMs. In addition to *in vitro* DNA cleavage activity, the PpCas9 nuclease exhibits activity in human cells, efficiently introducing indels in HEK293T genome targets flanked by a 5'-NNNNRTT-3' PAM.

Material and methods

Plasmids cloning

The predicted CRISPR-Cas Type II-C system locus of ***Defluviimonas sp.20V17*** including three spacers in the CRISPR array was PCR amplified with primers locus_DfCas9_F and locus_DfCas9_R using ***Defluviimonas sp.20V17*** genome DNA (DSMZ 24802) as a template. The resulting fragment was inserted into XbaI and HindIII digested pACYC184 vector using NEBuilder HiFi DNA Assembly Cloning Kit (NEB, E5520).

To obtain pET21a_DfCas9 plasmid, DfCas9 coding sequence was PCR amplified with DfCas9_F and DfCas9_R primers using bacterial genome DNA as a template. To obtain pET21a_PpCas9 plasmid, PpCas9 coding sequence was synthetized as g-block (IDT). The DfCas9 or PpCas9 coding fragments were inserted into XhoI and NheI digested pET21a vector by NEBuilder HiFi DNA Assembly Cloning Kit (NEB, E5520). The vectors maps and primers are presented in the Supplementary Table S1.

For expression in human cells PpCas9 gene was codon-optimized and inserted into plasmid under regulation of CMV promoter. SgRNA expression was driven by U6 promoter. The vector map is presented in the Supplementary Table S1.

Plasmid transformation interference screening

To determine DfCas9 PAM sequence a randomized 7N plasmid library carried a protospacer sequence flanked by seven randomized nucleotides was used (Supplementary Table S1). To create the library the ssDNA oligo Library_F containing randomized nucleotides was double-stranded through single stage PCR with Library_R primer (Evrogen). This fragment was assembled with PUC19 fragment synthesized through PCR using primers PUC19_F and PUC19_R by NEBuilder HiFi DNA Assembly Cloning Kit (NEB, E5520). The mix was transformed to *E. coli* DH5alpha strain and plated to media supplemented with 100μg/ml ampicillin. The plates were incubated at 37 °C. 18 hours after transformation more than 50 000 colonies were washed off the plates, and the plasmid library was extracted by Qiagen Plasmid Maxi kit (Qiagen 12162). The library plasmid map is presented in the Supplementary Table S1. Competent *E. coli* Star cells carrying pACYC184_DfCas9_locus or an empty pACYC184 vector were transformed with 7N PAM plasmid libraries and plated to 100μg/ml ampicillin and 25μg/ml chloramphenicol containing agar plates. After 16 hours cells were harvested and DNA was extracted using Qiagen Plasmid Maxi kit (Qiagen 12162). PAM-coding sequences were PCR amplified using M13_f and M13_r primers and sequenced using Illumina platform with pair-end 150 cycles (75+75).

Bacterial RNA sequencing

E. coli DH5alpha carrying pACYC184_DfCas9_locus plasmid were grown for 16 hours at 37 °C in LB (Luria Bertani) medium supplemented with 25 μg/ml chloramphenicol. Bacteria were resuspended in TRIzol (Thermo Fisher Scientific, 15596026). Total RNA was purified using Direct-Zol RNA kit (Zymo research, R2051). RNA was DNase I (Zymo research) treated and 3' dephosphorylated with T4 PNK (NEB, M0201). Ribo-Zero rRNA Removal Kit (Gram-Negative Bacteria) kit (Illumina, 15066012) was used to remove ribosomal RNA. HTS samples were prepared using NEBNext Multiplex Small RNA Library Prep Set for Illumina (NEB, E7300). The library was sequenced using Illumina platform with pair-end 150 cycles (75+75).

RNA sequencing analysis

HTS results of RNA sequencing were aligned to the reference plasmid pACYC184_DfCas9_locus using BWA aligner (24). Determined coordinates of 5' and 3' RNA ends were used to reconstruct the full-length RNA sequences. The resulting fragments were analyzed using Geneious 11.1.2. Filtered 40-130 nt-length sequences were used to generate the alignment.

In vitro DNA cleavage assays

DNA cleavage reactions were performed using the recombinant DfCas9 or PpCas9 proteins and linear dsDNA targets. The reaction conditions were: 1×CutSmart (NEB, B7204) buffer, 0.5 mM DTT, 20 nM DNA, 400 nM recombinant protein, 2 μM crRNA, 2 μM tracrRNA.
Samples were incubated at 37°C (DfCas9) or 42°C (PpCas9) for 30 min (unless otherwise stated). Further, 4X loading dye containing 10 mM Tris-HCl, pH 7.8, 40% glycerol, 40 mM EDTA, 0.01% bromophenol blue, 0.01% xylene cyanol was added to stop the reaction. Reaction products were analyzed by electrophoresis in 1.5% agarose gels. Pre-staining with ethidium bromide was used for visualization of bands on agarose.

For *in vitro* PpCas9 PAM screening, 100 nM linear DNA 7N PAM library was incubated with 400 nM recombinant protein, 5 μM crRNA, 5 μM tracrRNA. Reactions without crRNA were used as negative controls. The reaction was performed at 42°C for 30 min. Reaction products were separated by electrophoresis in agarose gels. Uncleaved DNA fragments were extracted from the gel using Zymo Clean Gel Recovery kit (Zymo research, D4007). HTS libraries were prepared using Ultra II DNA library prep kit (NEB, E7646). Samples were sequenced using MiniSeq Illumina with pair-end 300 cycles.

For testing the activity of DfCas9 or PpCas9 at different temperatures a mix of the corresponding protein with *in vitro* transcribed crRNA-tracrRNA in the cleavage buffer, and the DNA substrates, also in the cleavage buffer, were first incubated separately at the chosen temperature for 2 min, combined, and incubated for additional 10 minutes at same temperature. The following concentrations were used: 12nM DNA, 240nM PpCas9 or DfCas9, 1,2 μM crRNA, 1,2 μM tracrRNA – for *in vitro* cleavage of a linear DNA fragment; 4nM DNA, 80nM PpCas9 or DfCas9, 400nM crRNA, 400nM tracrRNA - for *in vitro* cleavage of a plasmid DNA.
All RNAs used in this study are listed in Supplementary Table S2.

Computational sequence analysis

For analysis of PpCas9 *in vitro* PAM screening results as well as DfCas9 plasmid interference screening in bacteria, Illumina reads were filtered by requiring an average Phred quality (Q score) of at least 20. Resulting reads were mapped against the corresponding reference sequence using BWA (24). All unmapped reads were discarded from the analysis. The degenerate 7-nucleotide region was extracted from the sequences. 16301 unique PAM sequences were found both for the depleted and control samples for DfCas9 PAM screening, and 16384 unique PAM sequences were found for PpCas9 PAM screening analysis. WebLogo was used to generate logo based on statistically significantly (one-sided Pearson chi-square test with a p-value less than 10^{-}

[12]) depleted PAM sequences (122 and 79 PAMs for DfCas9 and PpCas9, respectively; Supplementary Files S2, S3). For PAM wheel construction the depletion values of PAM sequence positions were counted according to Maxwell et al., 2018 (25).

Recombinant protein purification

For recombinant DfCas9 and PpCas9 purification competent *E. coli* Rosetta cells were transformed with pET21a_DfCas9 or pET21a_PpCas9 plasmid and grown till OD_{600} = 0.6 in 500 ml LB media supplemented with 100 μg/ml ampicillin. The target protein synthesis was induced by the addition of 1 mM IPTG. After 5 hours of growth at 25°C, cells were centrifugated at 4000g, the pellet was resuspended in lysis buffer containing 50 mM Tris-HCl pH 8.0 (4°C), 500 mM NaCl, 1 mM beta-mercaptoethanol and 10 mM imidazole supplemented with 1 mg/ml lysozyme (Sigma) and cells were lysed by sonication. The cell lysate was centrifuged at 16000g (4°C) and filtered through 0.45 μm filters. The lysate was applied to 1 ml HisTrap HP column (GE Healthcare) and DfCas9 or PpCas9 was eluted by 300 mM imidazole in the same buffer without lysozyme. After affinity chromatography fractions containing the nuclease were applied on a Superdex200 Increase 10/300 GL (GE Healthcare) column equilibrated with a buffer containing 50 mM Tris-HCl pH 8.0 (4°C), 500 mM NaCl, 1 mM DTT. Fractions containing DfCas9 or PpCas9 monomer were pooled and concentrated using 30 kDa Amicon Ultra-4 centrifugal unit (Merc Millipore, UFC803008). Glycerol was added to final concentration of 10% and samples were flash-frozen in liquid nitrogen and stored at -80 °C. Purity of the nucleases was assessed by denaturing 8% PAGE and the integrity of recombinant protein was confirmed by mass spectrometry.

Cell culture and transfection

HEK293RT cells were maintained in Dulbecco's modified Eagle's Medium (DMEM) supplemented with 10% fetal bovine serum at 37°C with 5% CO_2 incubation. Cells were seeded into 24-well plates (Eppendorf) one day prior to transfection. Cells were transfected using Lipofectamine 2000 (Thermo Fisher Scientific) following the manufacturer's recommended protocol. For each well of a 24-well plate a total of 500 ng plasmids was used. Three days after transfection cells were harvested and genomic DNA was extracted using QuickExtract solution (Lucigen, QE0950).

Indel frequency analysis

The genomic DNA from transfected cells was obtained as described above. Genomic region surrounding the CRISPR target site was amplified using two-step PCR. At the first step primers combining target-specific sequences and Illumina adapter overhangs were used (Supplementary Table S5).

First-Round PCR Forward Primer:

5' **CTCTTTCCCTACACGACGCTCTTCCGATCT**NNNN [target-specific sequence] 3'

First-Round PCR Reverse Primer:

5' **TCAGACGTGTGCTCTTCCGATCT** [target-specific sequence] 3'

The result amplicons were used as template in the second step PCR. This step introduced 8N barcode and flow cell linker adaptors using primers containing a sequence that anneals to the Illumina primer sequence introduced in first step.

Second-Round PCR Forward Primer:

5'AATGATACGGCGACCACCGAGATCTACA**CTCTTTCCCTACACGACGCTCTTCCGATCT** 3'

Second-Round PCR Reverse Primer:

5'CAAGCAGAAGACGGCATACGAGAT**NNNNNNNN**GTGACTGGAGT**TCAGACGTGTGCTCTTCCGATCT** 3'

The second round PCR products were loaded to agarose electrophoresis. The PCR fragments were gel-extracted using Cleanup Standard kit (Evrogen, BC022) and sequenced using Illumina (pair-end 150+150 or 75+75 cycles). Illumina reads were checked for the number of substitutions in regions covering sequences of primers used in PCR. These regions by experiment design don't include any indels and shouldn't contain a lot of errors. Thus, this step allows to filter out erroneous reads. As a threshold we used an average error rate 0.24 ± 0.06% per base proposed by Pfeiffer et al., 2018. Filtered reads were merged using custom script. Indel frequencies were estimated using CRISPResso2 analysis package (26). The window of 20 bp around the gRNA site and quantification window center corresponding to 3 nucleotides from the 3' end of the guide were provided to detect possible mutations. The resulting indel percentage was calculated as [Indels % in transfected cells] – [Indels % in untransfected cells] for a certain region of genomic DNA.

For T7 endonuclease I indel detection assay genomic region surrounding the CRISPR target site was PCR amplified using primers listed in Supplementary Table S1. The PCR fragments were gel-extracted using Cleanup Standart kit (Evrogen, BC022). Next, the indel detection assay was performed using T7 endonuclease I (NEB, M0302) according to manufacturer's recommended protocol. In brief, after incubation with T7 endonuclease I PCR products were loaded to agarose native gel and stained with ethidium bromide for 10 minutes.

Results

Defluviimonas sp.20V17 and *P. pneumotropica* CRISPR Cas II-C systems: organization of the loci

Bioinformatics searches of small-sized Cas9 proteins reveal multiple orthologues belonging to Type II-C type CRISPR-Cas systems (27–29). Most of these proteins are not biochemically characterized. CRISPR-Cas systems from *Defluviimonas sp.20V17* and *Pasteurella pneumotropica* carry intact sequences of *cas* genes and were chosen for further characterization. The *P. pneumotropica* CRISPR-Cas Type II-C locus contains an array composed of four 36-bp DRs (direct repeats) interspaced by 30-bp spacers in the proximity of the *cas* genes operon (Figure 1A). The available *Defluviimonas sp.20V17* genome assembly is fragmented and comprises 236 discrete contigs (22). The *Defluviimonas sp.20V17* Type II-C CRISPR-Cas system is located at the end of one contig and the leader-proximal part of the array is missing. Based on available information, the *Defluviimonas sp.20V17* Type II-C CRISPR array contains at least 30 DRs interspaced by 30-bp spacers (Figure 1A). A BLAST search using spacer sequences from arrays of both systems as queries revealed no matches to sequences from public databases. The adaptation modules of *P. pneumotropica* and *Defluviimonas sp.20V1* Type II-C CRISPR-Cas loci *P. pneumotropica* and *Defluviimonas sp.20V1* include *cas1* and *cas2* genes. Both loci contain *cas9* genes encoding relatively small Type II-C effectors: DfCas9 is 1079 amino acids long while PpCas9 is 1055 amino acids long. Upstream of *cas* genes in both systems we identified a putative tracrRNA-encoding sequences partially complementary to DRs. In both cases *in silico* co-folding of part of DR with the putative tracrRNA predicts stable secondary structures (Figure 1B).

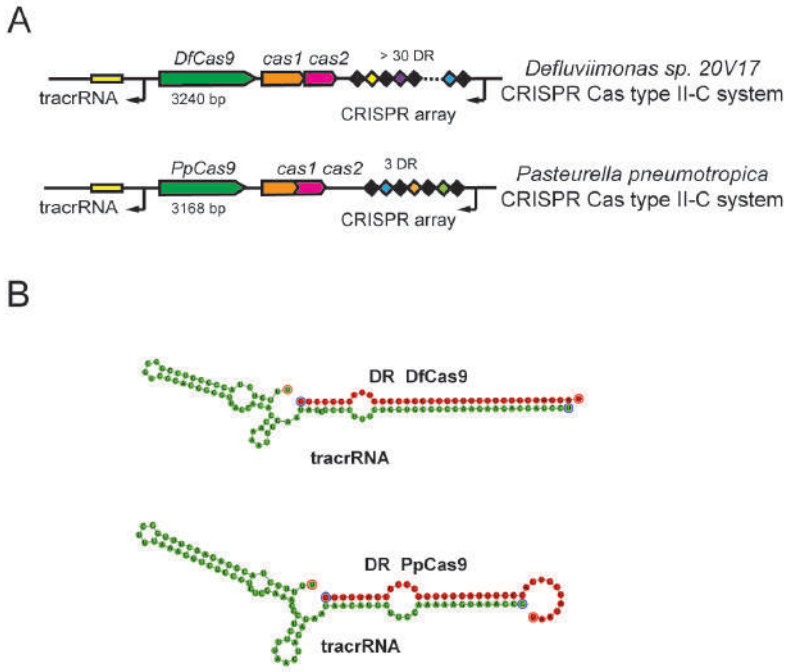


Figure 1. *Defluviimonas sp. 20V17 and P. pneumotropica* CRISPR-Cas Type II-C locuses.

A. Organization of *Defluviimonas sp. 20V17* and *P. pneumotropica* CRISPR-Cas Type II-C loci. DRs are shown as black rectangles, spacers are indicated by rectangles of different colors. The tracrRNA coding sequences are shown as yellow rectangles. The *cas* genes are labeled. Direction of transcription is indicated with black arrows.

B. *In silico* co-folding of *Defluviimonas sp. 20V17 and P. pneumotropica* DRs and putative tracrRNAs. The DR sequences are colored in red, the tracrRNA sequences are colored in green.

Characterization of *Defluviimonas sp.20V17* DfCas9 nuclease

To study *Defluviimonas sp.20V17* CRISPR-Cas Type II-C system we cloned the corresponding locus into the pACYC184 plasmid for heterologous expression. Only a fragment of CRISPR array containing four DRs adjacent to *cas* genes was used for cloning. To check the efficiency of transcription of RNA components of the cloned CRISPR-Cas system, small RNAs present in *E. coli* harboring plasmid-borne *Defluviimonas sp.20V17* locus were sequenced. We found that the shortened *Defluviimonas sp.20V17* CRISPR array was actively transcribed in an orientation opposite to that of *cas* genes transcription and processed crRNAs were detected (Figure 2A). The tracrRNAs coding sequence was also expressed generating 72–83 nt products.

Given high expression levels of *Defluviimonas sp. 20V17* crRNA and tracrRNA in *E. coli*, we performed plasmid transformation interference screening in the heterologous host to determine the DfCas9 protospacer adjacent motif (PAM) sequence (Figure 2B). The plasmid transformation interference screening is based on transformation of *E. coli* cells carrying a plasmid with a

CRISPR-Cas locus or an empty vector with a library of compatible plasmids bearing a protospacer sequence matching one of the spacers in the CRISPR array and flanked by seven randomized nucleotides. Transformed cells are plated on a medium that selects for cells carrying both plasmids. Since successful recognition of targets with interference-proficient PAM by Cas9 nuclease leads to plasmid destruction, underrepresentation of interference-proficient PAM library members is expected in transformants carrying the CRISPR-Cas locus comparing to control cells.

One of the spacers in *Defluviimonas sp. 20V17* CRISPR array was used as a protospacer for plasmid PAM library construction. Plasmid transformation interference screening in *E. coli* and subsequent high-throughput sequencing of the targeted protospacer region amplified from plasmids extracted from pooled transformant colonies revealed depletion of library members with 5'-NNRNAYN-3' sequences adjacent to protospacer in cells carrying *Defluviimonas sp. 20V17* CRISPR-Cas Type II-C locus compared to control cells.

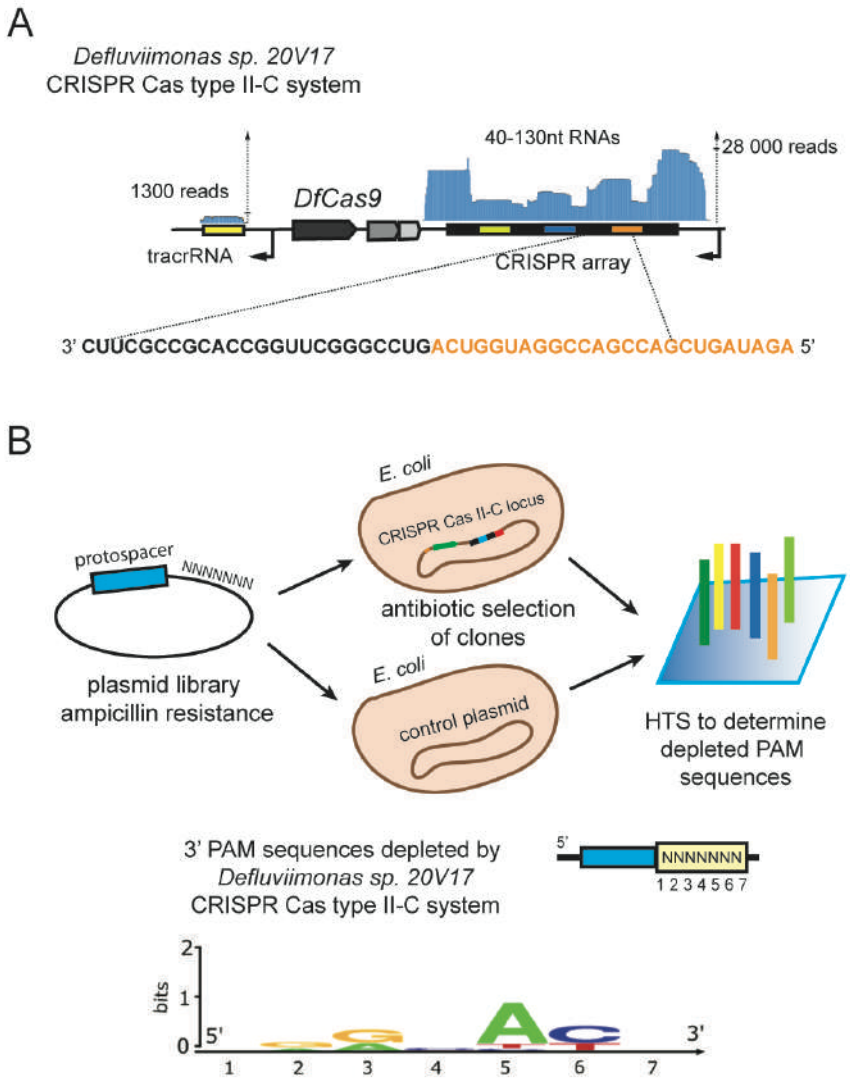


Figure 2. Studying of *Defluviimonas sp. 20V17* CRISPR-Cas Type II-C system in bacteria.

A. Identification of *Defluviimonas sp. 20V17* crRNAs. Reads (blue) are mapped at the top of the CRISPR array. A sequence of a typical mature crRNA sequence is expanded at the bottom with spacer part shown in orange. The direction of transcription is indicated with black arrows.

B. Determination of DfCas9 PAM sequences using plasmid transformation interference screening. Above - a scheme of the interference screen experiment. Below - DfCas9 PAM sequence logo determined from the PAM screening. PAM position numbers correspond to nucleotides immediately following the protospacer in the 5'-3' direction.

To reconstruct the DfCas9 DNA cleavage reaction *in vitro*, recombinant DfCas9 was purified and crRNA and tracrRNA molecules were synthesized by T7 RNA polymerase (Supplementary File S1, Supplementary Figure S1, Supplementary Table S3). As a DNA target we used a linear DNA fragment carrying a protospacer sequence flanked by 5'-AAAAACG-3' PAM selected based on plasmid transformation interference screening results. The incubation of DfCas9-crRNA-tracrRNA ribonucleoprotein complex with the DNA target in a buffer supplemented with Mg^{2+} at 37 °C for 30 minutes led to DNA cleavage (Figure 3A). To further examine the DfCas9 PAM preferences, single-nucleotide substitutions in the deduced consensus PAM sequence were introduced and individually tested for cleavage efficiency (Figure 3A). The replacement of purines to pyrimidines in the 3rd position, as well as substitutions in the 5th and 6th positions of the 5'-AAAAACG-3' PAM prevented DNA cleavage. These results confirmed the PAM consensus determined by plasmid transformation interference screening and allowed us to conclude that DfCas9 nuclease requires a 5'-NNRNAY-3' PAM.

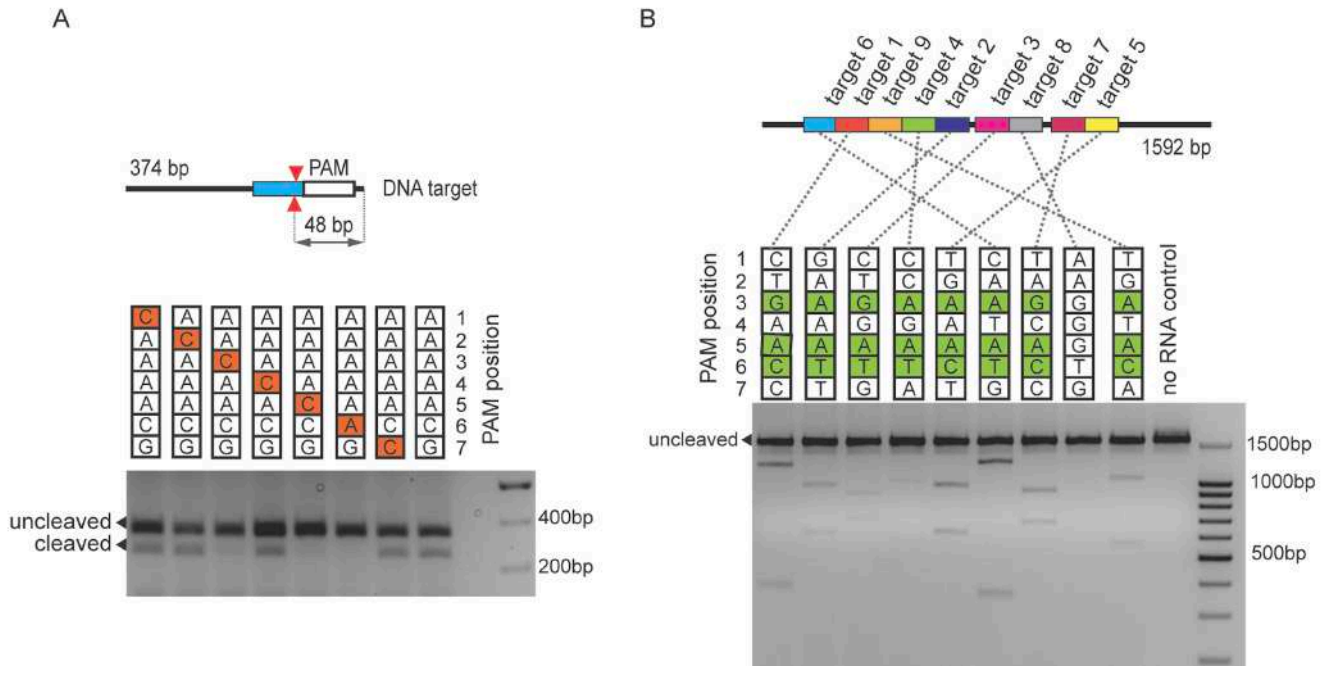


Figure 3. *In vitro* cleavage of DNA targets by DfCas9.

A. Single-nucleotide substitutions in the 3th, 4th, and 6th positions of PAM prevent DNA cleavage by DfCas9. An agarose gel showing the results of electrophoretic separation of cleavage products of targets with PAM sequences shown at the top is presented. Bands corresponding to cleaved and uncleaved DNA fragments are indicated. The scheme above shows the position of expected DNA cleavage site.

B. DfCas9 efficiently cleaves different DNA targets with 5'-NNRNAYN-3' PAM consensus *in vitro*. The scheme above shows the positions of different target sites in the grin2b gene fragment. Below, a gel showing results of *in vitro* cleavage of targets with indicated PAMs is presented.

Next, we tested DfCas9 DNA cleavage activity on different targets. Several 20-bp targets flanked by 5'-NNRNAYN-3' consensus PAM sequence were chosen on a 1592 bp linear DNA fragment of the GRIN2b gene. DNA cleavage reactions were performed using crRNAs DfCas9 charged with crRNAs corresponding to different target sites (Figure 3B, Supplementary Table S3). As can be seen, DfCas9 nuclease successfully introduced double-stranded breaks in all DNA targets, and did not cleave a site with an 5'-AAGGGTG-3' located at the place of PAM, which was used as a negative control.

Overall, we conclude that DfCas9 nuclease specifically cleaves DNA targets flanked with 5'-NNRNAY-3' PAM sequence at the 3' side of protospacers.

Characterization of PpCas9 nuclease from *Pasteurella pneumotropica*

Due to the lack of *P. pneumotropica* genomic DNA at our disposal, all experiments with the PpCas9 effector nuclease were performed *in vitro.* A recombinant PpCas9 was purified from *E. coli* Rosetta cells, the bioinformatically predicted crRNAs and tracrRNA were synthesized *in vitro*. To assess the PpCas9 nuclease activity we tested its ability to cleave linear DNA PAM libraries containing a target site flanked with seven randomized nucleotides at the 3' end (Figure 4A). PpCas9 in complex with crRNA and tracrRNA was incubated with PAM library at 42 °C for 30 minutes, uncleaved molecules were purified after agarose gel electrophoresis and sequenced (along with negative control - original PAM library incubated with PpCas9-tracrRNA in the absence of crRNA) using an Illumina platform. Comparison of PAM variants representation in the depleted sample and the control allowed us to determine the PpCas9 PAM logo. The results showed that PpCas9 prefers targets flanked by a 5'-NNNNATT- 3' PAM (Figure 4B), where T in the 7th position is less conserved than nucleotides in the 5th and the 6th positions.

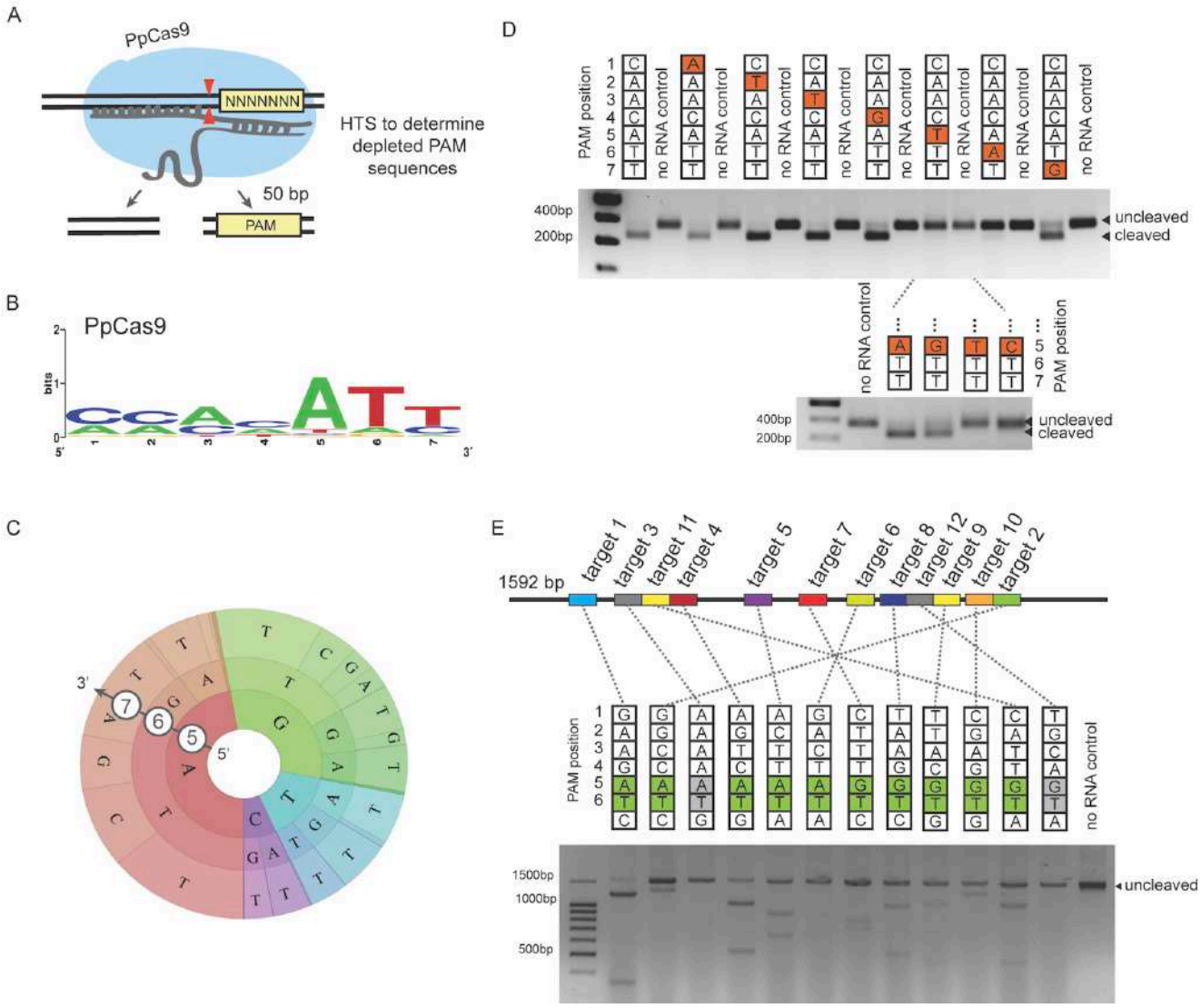


Figure 4. *In vitro* cleavage of DNA targets by PpCas9.

A. A scheme of the *in vitro* PAM screening experiment.
A linear DNA PAM library containing a target site flanked with seven randomized nucleotides at the 3' end is incubated with PpCas9 charged with appropriate crRNA and tracrRNA. This leads to the cleavage of library members carrying functional PAM sequences and generates DNA products shortened by 50 bp. Uncleaved PAM library molecules are recovered, which depletes the library. The uncleaved molecules, as well as negative control (original PAM library incubated with PpCas9-tracrRNA in the absence of crRNA) are sequenced. Comparison of PAM variants representation in the depleted sample and the control allows to determine the PpCas9 PAM logo.

B. Weblogo of PpCas9 PAM sequences depleted after *in vitro* PAM screening.

C. Single-nucleotide substitutions in the 5th and 6th positions of PAM prevent DNA cleavage by PpCas9. An agarose gel showing the results of electrophoretic separation of targets with PAM sequences shown at the top after incubation with the PpCas9 effector complex is presented. Bands corresponding to cleaved and uncleaved DNA fragments are indicated.

D. Wheel representation of *in vitro* PAM screen results for 5th, 6th, and 7th nucleotide positions of PAM. Nucleotide positions from the inner to the outer circle match PAM positions moving away from the protospacer. For a given sequence, the area of the sector in the PAM wheel is proportional to the relative depletion in the library.

E. PpCas9 efficiently cleaves different DNA targets with 5'-NNNNRTN-3' PAM consensus *in vitro*. The scheme above shows the positions of different target sites in a 1592 bp GRIN2b gene fragment. Below, a gel showing results of *in vitro* cleavage of targets with indicated PAMs is presented.

To further investigate PpCas9 PAM sequence preferences, identify individual sequences representing functional PAMs, and determine the relative activity of each PAM sequence, we used the PAM wheel approach developed by Leenay et al. for results visualization (30). The PAM wheel confirmed the importance of T in the 6th position and variability of the 7th PAM nucleotide with a slight bias for a T. In addition, this approach revealed a preference for purines in the 5th position (Figure 4C).

Next, we made single-nucleotide substitutions in consensus 5'-CAACATT-3' PAM and tested these targets individually for PpCas9 cleavage efficiency (Figure 4D). The experiment results confirmed the preference for both G and A in the 5th position, the importance of T in the 6th position, and tolerance for all four possible nucleotides in the 7th position of PAM. Overall, the results allowed us to disregard the small preference for a T in the 5th position and conclude that for efficient *in vitro* DNA cleavage PpCas9 requires a 5'-NNNNRTN-3' consensus PAM. To validate the proposed PAM consensus we tested whether PpCas9 is able to cleave different DNA targets flanked with this consensus. A 1592 bp linear DNA fragment was used as a target bearing several 20-nt PpCas9 target sites flanked by 5'-NNNNRTN-3' PAM (Figure 4E). PpCas9 successfully cleaved most targets, confirming the deduced PAM consensus.

DfCas9 and PpCas9 temperature preferences

The range of optimal working temperatures is one of the factors which determine Cas nuclease application. Temperature dependence of DfCas9 and PpCas9 nuclease activities was determined using targets flanked by corresponding consensus PAMs on either a linear DNA fragment or on a plasmid (Figure 5). DfCas9 efficiently cleaved the plasmid DNA in a temperature range of 20-37 °C with maximal cleavage at 35 °C. PpCas9 demonstrated lower activity at 20 °C and efficiently cleaved its targets between 25 and 47 °C with maximal activity at 40 °C. In contrast to PpCas9, DfCas9 demonstrated different efficiencies of linear and supercoiled DNA cleavage.

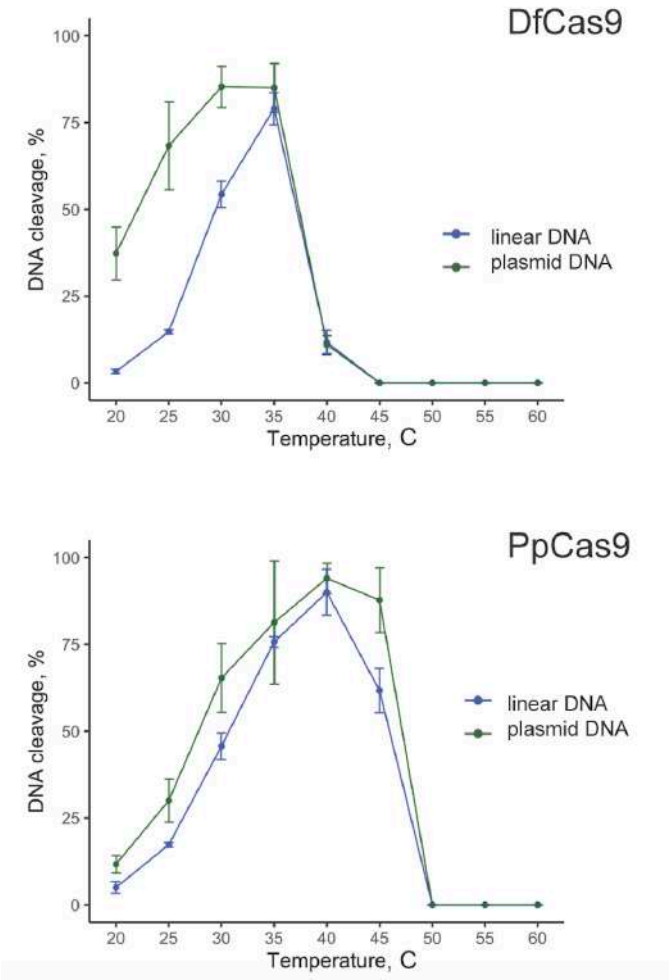


Figure 5. Cleavage activity of DfCas9 and PpCas9 at different temperatures.

DfCas9 or PpCas9 was incubated with cognate tracrRNA and crRNA and a 2.7 kb plasmid DNA or a 921 bp linear DNA fragment containing target sequences at indicated temperatures for 10 min. Products were separated by agarose gel electrophoresis. Cleavage efficiency (in percent) was calculated as a ratio of intensity of cleaved bands to the combined intensity of cleaved and uncleaved bands. Mean values and standard deviations obtained from three independent experiments are shown.

DfCas9 and PpCas9 sgRNA design

To facilitate the use of DfCas9 and PpCas9 as programmable nucleases we sought to design single guide RNA (sgRNAs) where crRNA and tracrRNA are fused. Several PpCas9 and DfCas9 sgRNAs variants were tested in *in vitro* DNA cleavage experiments (Supplementary File S1, Supplementary Figure S2). As a result, we determined DfCas9 and PpCas9 sgRNA forms, which supported efficient DNA cleavage *in vitro* (Figure 6).

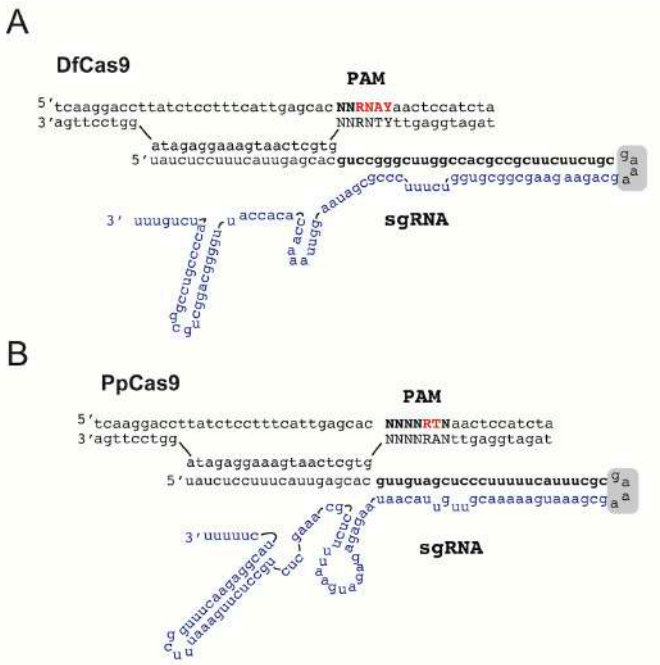


Figure 6. The DfCas9 and PpCas9 minimal *in vitro* DNA cleavage systems.

A. A scheme of recognition by the DfCas9-sgRNA complex of a DNA target flanked by 5'-NNRNAY-3' PAM (Y stands for pyrimidines, R stands for purines). The crRNA-tracrRNA linker is indicated with a grey box. The part of sgRNA that originated from tracrRNA is shown in blue.

B. A scheme of recognition by the PpCas9-sgRNA complex of a DNA target flanked by 5'-NNNNRTN-3 PAM (R stands for purines). The crRNA-tracrRNA linker is indicated by a grey box. The part of sgRNA that originated from tracrRNA is shown in blue.

PpCas9 nuclease is active in human cells

Next, we tested DfCas9 and PpCas9 activity in human cells. Codon optimized DfCas9 and PpCas9 genes, as well as the SpCas9 gene, which was used as a positive control, were cloned into plasmid vectors under regulation of constitutive CMV promoter. A GFP coding sequence was fused to in frame with nuclease open reading frames through a sequence encoding P2A self-cleaving peptide. Appropriate sgRNA coding sequences were introduced into plasmids upstream of nuclease genes under the control of the U6 promoter (Figure 7A). PpCas9 and DfCas9 were targeted to two human genes, EMX1 and GRIN2b, with two targeting sites tried in each gene (Supplementary Table S4). SpCas9 was targeted to the GRIN2b gene only.

Extension of complementary region between a target sequence and the guide RNA spacer part increased the genome editing activity for several Type II-C nucleases (14, 18). Given this, for each site we used sgRNAs with spacer sequences of two lengths: 20 and 24 nt. In the case of SpCas9 only sgRNAs with optimal 20 nt spacer sequences were used. Plasmids were transfected

into human HEK293T cells and production of recombinant Cas9 proteins was confirmed by Western blot analysis. The efficiency of transfection was about 30%. Two days after transfection genomic DNA was extracted from a heterogeneous population of modified and unmodified cells and indel formation (nucleotides insertion or deletions) was assessed using the T7 endonuclease I detection assay.

No genome modification was detected in cells transfected with DfCas9 (data not shown). In contrast, PpCas9 introduced indels in EMX1.1 and GRIN2b.1 sites (Figure 7B) but failed to modify the EMX1.2 and GRIN2b.2 sites. Where cleavages were observed, sgRNAs with spacer sequences of 24 nt were more effective, in agreement with data for other Type II-C effectors (14, 18).

Next, the activity of PpCas9 at additional sites in GRIN2b or EMX1 genes flanked by 5'-NNNNRTN-3' was tested. PpCas9 modified more targets in GRIN2b compared to EMX1, possibly due to DNA methylation or other factors which can impede the binding of the nuclease to genomic DNA. Although the *in vitro* DNA cleavage experiments demonstrated only a slight preference for T in the 7th PAM position, in human cells PpCas9 efficiently cleaved most of the targets flanked by the 5'-NNNNRTT-3' PAM and failed to cleave targets flanked by PAMs with no T in the 7th position. This suggests that T in the 7th PAM position, although non-essential for *in vitro* DNA cleavage, is highly important for DNA recognition in human cells.

The length requirements of PpCas9 sgRNA spacer needed for optimal genome editing was investigated in further detail. HEK293T cells were transfected by PpCas9 carrying plasmids analogous to those described above but coding for sgRNAs of different spacer length targeting two sites in the human genome. The assessment of DNA modification efficiency was performed using HTS sequencing of targeted sites. The results showed that PpCas9 efficiently introduces double-stranded breaks in genomic DNA with sgRNAs of 21-24 nt spacer length (Figure 7C). The highest levels of genome modification were achieved when sgRNAs with 22-23 nt spacers were used.

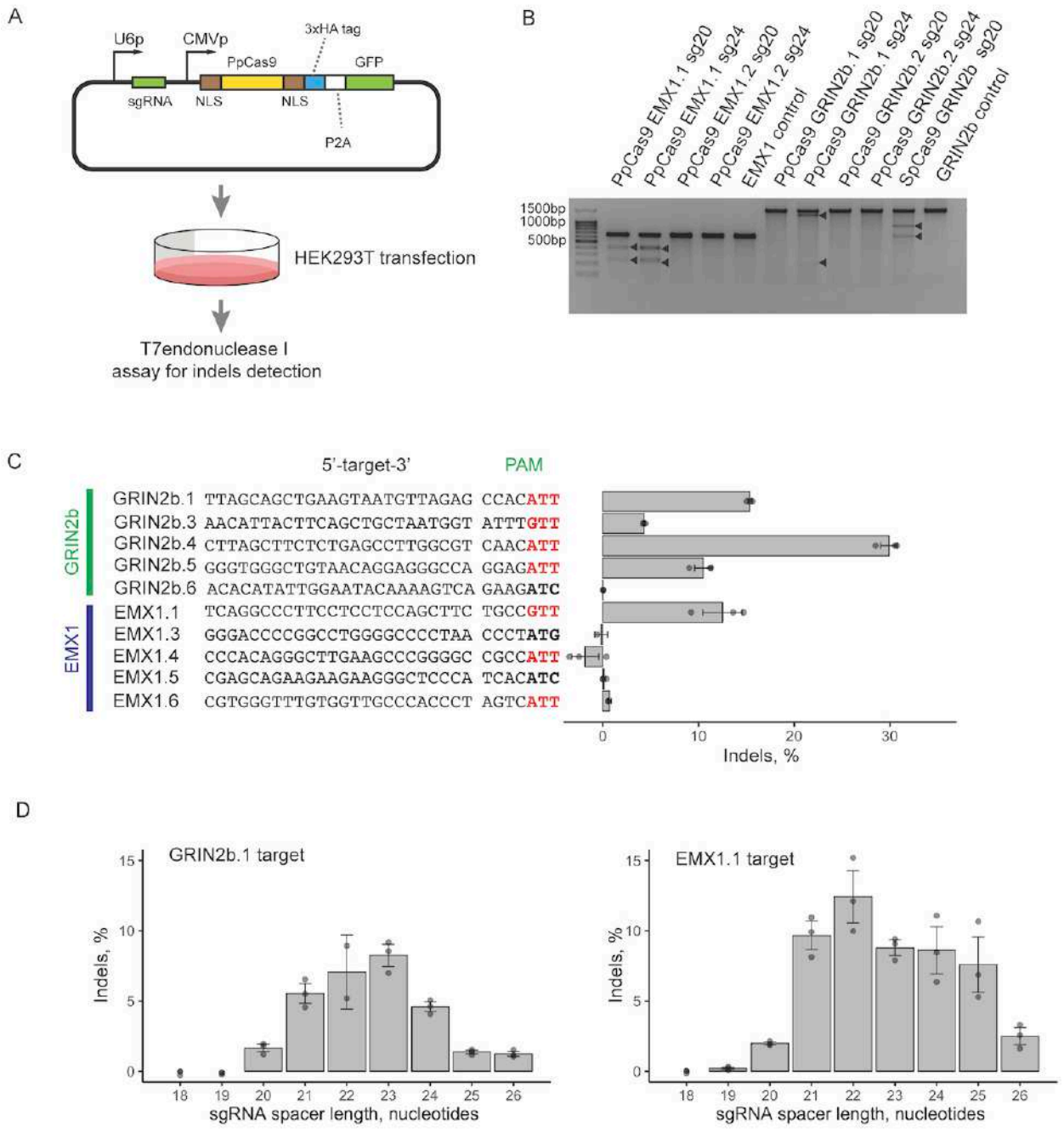


Figure 7. PpCas9 nuclease activity in human HEK293T cells.

A. Scheme of the PpCas9 nuclease activity assessment experiment. Above - a scheme showing design of a plasmid used for PpCas9 gene and sgRNAs expression. The PpCas9 gene is shown as a yellow rectangle, NLS (nuclear localization signals) as brown rectangles, GFP gene as a green rectangle. CMV promoter and U6 promoters are indicated with black arrows. The sgRNA coding sequence is shown as a green rectangle. The plasmid was transfected into HEK293T cells and genomic DNA was extracted from a heterogeneous population of modified and unmodified cells for indel frequency assessment through HTS of the targeted region or *in vitro* assay with T7 endonuclease I.

B. Results of T7 endonuclease I indel detection assay showing PpCas9-mediated cleavage of EMX1 and GRIN2b genes in HEK293T genome.

C. PpCas9 indel formation efficiency at different genomic sites. Left – genomic DNA target sites with corresponding PAM sequences. 5'-NNNNRTT-3' PAM are shown in red. Right - indel frequency estimated by HTS analysis. Mean values and standard deviations obtained from three biological replicas are shown.

D. The influence of sgRNA spacer length on PpCas9-mediated indel formation efficiency in EMX1 and GRIN2b genes. HEK293T cells were transfected with PpCas9_sgRNA plasmids (as in panel A) coding for sgRNAs with different lengths of spacer segments. Left - results for the GRIN2b.1 target, right – for the EMX1.1 target. Mean values and standard deviations obtained from three biological replicas are shown.

The specificity of PpCas9

Use of Cas nucleases in biotechnology requires sufficient specificity of target recognition. We conducted a preliminary test of PpCas9 genome cleavage specificity. As on-targets we chose two DNA sites in EMX1 and GRIN2b genes, which were efficiently cleaved by PpCas9 in previous experiments (EMX1.1 and GRIN2b.1). HEK293T cells were transfected with plasmids carrying the PpCas9 genome-editing system targeting these sites. SgRNAs with spacer length of 21 nt were used to direct PpCas9 to on-target sequences. Three days post-transfection genomic DNA was extracted and indel frequency at the on-target as well as at likely off-target sites (sequences differing by up to 3 nucleotides from on-target sites) was assessed by targeted amplicon high-throughput sequencing. The computational analysis using CRISPResso2 detected some modifications at off-target site 6 and off-target site 2 for GRIN2b and EMX1, respectively (Figure 8), although all of the identified indels were single nucleotide substitutions, which may result from sequencing errors rather than off-target activity. We did not observe either insertions or deletions at these off-target sites, in contrast to on-target regions, where PpCas9 generated indels of different lengths. These preliminary results indicate that PpCas9 is quite specific, though additional studies using the more accurate methods are needed to fully estimate its specificity.

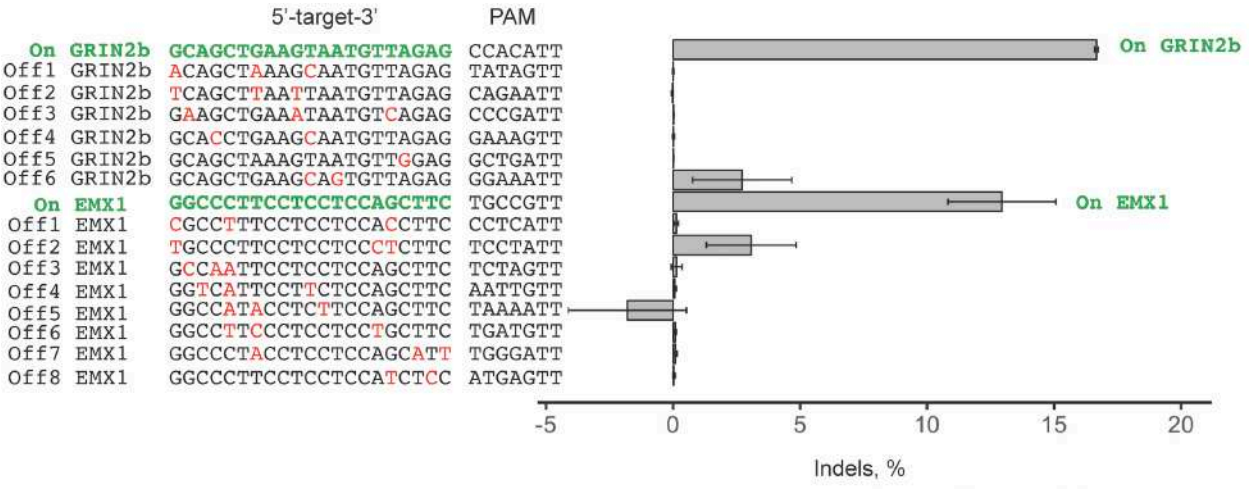


Figure 8. The specificity of genomic DNA cleavage by PpCas9.

The indel frequency at two on-target as well as at corresponding off-target sequences was assessed by targeted amplicon sequencing of genomic DNA of HEK293T cells transfected with plasmids carrying the PpCas9 genome editing system. Left - sequences of on-target sites (in green) and off-target sites are shown. Mismatches in off-target sequences are shown in red. Right – frequencies of indel formation in each site.

Discussion

The genome editing applications require efficient delivery of CRISPR-Cas systems into targeted organs or tissues. Due to their small size, AAV particles are spreading within a tissue around the site of injection. The use of AAV particles as delivery vehicles allows high levels of Cas nuclease gene expression in recipient cells and efficient indel formation in targeted genome sites (14, 29, 31). As a result, the use AAV for CRISPR editors delivery allows to assess the mutagenesis effects in 1-2 months after injection in adult animals. This provides considerable saving of time compared to conventional strategies of the generation of animal models with targeted mutations.

The length of the SpCas9 gene, coding for most popular Cas effector, does not allow to deliver it simultaneously with sequences coding for sgRNAs and promoters in a single AAV particle. The alternative delivery in two mating AAVs is less effective due to the lower frequency of co-transduction. Shorter Type II-C Cas9 orthologs provide a source of “small-size” nucleases suitable for all-in-one particle delivery. Most of Type II-C Cas9 nucleases with demonstrated activity in human cells have long complicated PAM sequences (NmeCas9, CjeCas9, GeoCas9, CdCas9 require PAM sequences 5’-NNNNGNTT-3’, 5’-NNNNRYAC-3’, 5’-NNNNCRAA-3’, and 5’-NNRHHHY-3’ (H stands for A, T or C), correspondingly) and/or lower target cleavage efficiency compared to SpCas9. Exception to this are Nme2Cas9 and SauriCas9 which require 5’-NNNNCC-3’and 5’-NNGG-3’ PAMs, correspondingly, and are highly efficient in human cells.

The PpCas9 nuclease from *P. pneumotropica* characterized in this study requires a novel short PAM sequence 5'-NNNNRTT-3' (5'-NNNNRT-3' for *in vitro* DNA cleavage) and demonstrates activity in human cells. At 1055 amino acids, PpCas9 is similar to small Cas9 orthologs SaCas9, CjCas9, Nme2Cas9 and SauriCas9 (1053, 984, 1082 and 1061 amino acids, correspondingly), and thus can be delivered with sgRNA sequences via a single size-restricted vector such as AAV. We therefore envision that PpCas9 potentially could be used as a genome-editing instrument, although additional studies of its efficiency should be conducted. Indeed, PpCas9 cleaved DNA targets in GRIN2b more efficiently than in EMX1, which may reflect its preference for certain genome methylation patterns, folding of DNA and/or other factors. While PpCas9 specificity also should be studied in more detail, preliminary data obtained in this study demonstrate that this nuclease does not possess high off-targeting activity.

The *Defluviimonas sp. 20V17* nuclease DfCas9, also characterized in this study, did not show observable genome-editing activity in human cells. Yet, this small size nuclease with distinct PAM requirements is active at least in *E. coli* and thus may also be biotechnological application of the CRISPR-Cas9 technology.

Data availability

Raw sequencing data have been deposited with the National Center for Biotechnology Information Sequence Read Archive under BioProject ID PRJNA629762 and PRJNA629763.

Funding

This work was supported by the Ministry of Education and Science of the Russian Federation Subsidy Agreement 14.606.21.0006 (Project identifier RFMEFI60617X0006)

References

1. Pourcel,C., Salvignol,G. and Vergnaud,G. (2005) CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology (Reading, Engl.)*, **151**, 653–663.

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

3. Bolotin,A., Quinquis,B., Sorokin,A. and Ehrlich,S.D. (2005) Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology (Reading, Engl.)*, **151**, 2551–2561.

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

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

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

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

8. Garneau,J.E., Dupuis,M.-È., Villion,M., Romero,D.A., Barrangou,R., Boyaval,P., Fremaux,C., Horvath,P., Magadán,A.H. and Moineau,S. (2010) The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature*, **468**, 67–71.

9. Ronda,C., Pedersen,L.E., Sommer,M.O.A. and Nielsen,A.T. (2016) CRMAGE: CRISPR Optimized MAGE Recombineering. *Sci Rep*, **6**, 19452.

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

11. Cong,L., Ran,F.A., Cox,D., Lin,S., Barretto,R., Habib,N., Hsu,P.D., Wu,X., Jiang,W., Marraffini,L.A., *et al.* (2013) Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science*, **339**, 819–823.

12. Makarova,K.S., Wolf,Y.I., Iranzo,J., Shmakov,S.A., Alkhnbashi,O.S., Brouns,S.J.J., Charpentier,E., Cheng,D., Haft,D.H., Horvath,P., *et al.* (2019) Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.*, 10.1038/s41579-019-0299-x.

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

14. Kim,E., Koo,T., Park,S.W., Kim,D., Kim,K., Cho,H.-Y., Song,D.W., Lee,K.J., Jung,M.H., Kim,S., *et al.* (2017) In vivo genome editing with a small Cas9 orthologue derived from Campylobacter jejuni. *Nat Commun*, **8**, 14500.

15. Hu,Z., Wang,S., Zhang,C., Gao,N., Li,M., Wang,D., Wang,D., Liu,D., Liu,H., Ong,S.-G., *et al.* (2020) A compact Cas9 ortholog from Staphylococcus Auricularis (SauriCas9) expands the DNA targeting scope. *PLoS Biol.*, **18**, e3000686.

16. Edraki,A., Mir,A., Ibraheim,R., Gainetdinov,I., Yoon,Y., Song,C.-Q., Cao,Y., Gallant,J., Xue,W., Rivera-Pérez,J.A., *et al.* (2019) A Compact, High-Accuracy Cas9 with a Dinucleotide PAM for In Vivo Genome Editing. *Mol. Cell*, **73**, 714-726.e4.

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

18. Hirano,S., Abudayyeh,O.O., Gootenberg,J.S., Horii,T., Ishitani,R., Hatada,I., Zhang,F., Nishimasu,H. and Nureki,O. (2019) Structural basis for the promiscuous PAM recognition by Corynebacterium diphtheriae Cas9. *Nat Commun*, **10**, 1968.

19. Harrington,L.B., Paez-Espino,D., Staahl,B.T., Chen,J.S., Ma,E., Kyrpides,N.C. and Doudna,J.A. (2017) A thermostable Cas9 with increased lifetime in human plasma. *Nat Commun*, **8**, 1424.

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

21. Jiang,L., Xu,H., Shao,Z. and Long,M. (2014) Defluviimonas indica sp. nov., a marine bacterium isolated from a deep-sea hydrothermal vent environment. *INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY*, **64**, 2084–2088.

22. Jiang,L., Long,M. and Shao,Z. (2014) Draft Genome Sequence of Defluviimonas indica Strain 20V17T, Isolated from a Deep-Sea Hydrothermal Vent Environment in the Southwest Indian Ocean. *Genome Announcements*, **2**, e00479-14, 2/3/e00479-14.

23. Jawetz,E. (1950) A Pneumotropic Pasteurella of Laboratory Animals. I. Bacteriological and Serological Characteristics of the Organism. *Journal of Infectious Diseases*, **86**, 172–183.

24. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

25. Maxwell,C.S., Jacobsen,T., Marshall,R., Noireaux,V. and Beisel,C.L. (2018) A detailed cell-free transcription-translation-based assay to decipher CRISPR protospacer-adjacent motifs. *Methods*, **143**, 48–57.

26. Clement,K., Rees,H., Canver,M.C., Gehrke,J.M., Farouni,R., Hsu,J.Y., Cole,M.A., Liu,D.R., Joung,J.K., Bauer,D.E., *et al.* (2019) CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat. Biotechnol.*, **37**, 224–226.

27. Mir,A., Edraki,A., Lee,J. and Sontheimer,E.J. (2018) Type II-C CRISPR-Cas9 Biology, Mechanism, and Application. *ACS Chem. Biol.*, **13**, 357–365.

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

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

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

31. Lau,C.-H. and Suh,Y. (2017) In vivo genome editing in animals using AAV-CRISPR system: applications to translational research of human disease. *F1000Res*, **6**, 2153.

Figures legends

Figure 1. *Defluviimonas sp. 20V17 and P. pneumotropica* CRISPR-Cas Type II-C locuses.

A. Organization of *Defluviimonas sp. 20V17* and *P. pneumotropica* CRISPR-Cas Type II-C loci. DRs are shown as black rectangles, spacers are indicated by rectangles of different colors. The tracrRNA coding sequences are shown as yellow rectangles. The *cas* genes are labeled. Direction of transcription is indicated with black arrows.
B. *In silico* co-folding of *Defluviimonas sp. 20V17 and P. pneumotropica* DRs and putative tracrRNAs. The DR sequences are colored in red, the tracrRNA sequences are colored in green.

Figure 2. Studying of *Defluviimonas sp. 20V17* CRISPR-Cas Type II-C system in bacteria.

A. Identification of *Defluviimonas sp. 20V17* crRNAs. Reads (blue) are mapped at the top of the CRISPR array. A sequence of a typical mature crRNA sequence is expanded at the bottom with spacer part shown in orange. The direction of transcription is indicated with black arrows.

B. Determination of DfCas9 PAM sequences using plasmid transformation interference screening. Above - a scheme of the interference screen experiment. Below - DfCas9 PAM sequence logo determined from the PAM screening. PAM position numbers correspond to nucleotides immediately following the protospacer in the 5'-3' direction.

Figure 3. *In vitro* cleavage of DNA targets by DfCas9.

A. Single-nucleotide substitutions in the 3^{th}, 4^{th}, and 6^{th} positions of PAM prevent DNA cleavage by DfCas9. An agarose gel showing the results of electrophoretic separation of cleavage products of targets with PAM sequences shown at the top is presented. Bands corresponding to cleaved and uncleaved DNA fragments are indicated. The scheme above shows the position of expected DNA cleavage site.

B. DfCas9 efficiently cleaves different DNA targets with 5'-NNRNAYN-3' PAM consensus *in vitro*. The scheme above shows the positions of different target sites in the grin2b gene fragment. Below, a gel showing results of *in vitro* cleavage of targets with indicated PAMs is presented.

Figure 4. *In vitro* cleavage of DNA targets by PpCas9.

A. A scheme of the *in vitro* PAM screening experiment.
A linear DNA PAM library containing a target site flanked with seven randomized nucleotides at the 3' end is incubated with PpCas9 charged with appropriate crRNA and tracrRNA. This leads to the cleavage of library members carrying functional PAM sequences and generates DNA products shortened by 50 bp. Uncleaved PAM library

molecules are recovered, which depletes the library. The uncleaved molecules, as well as negative control (original PAM library incubated with PpCas9-tracrRNA in the absence of crRNA) are sequenced. Comparison of PAM variants representation in the depleted sample and the control allows to determine the PpCas9 PAM logo.

B. Weblogo of PpCas9 PAM sequences depleted after *in vitro* PAM screening.

C. Single-nucleotide substitutions in the 5th and 6th positions of PAM prevent DNA cleavage by PpCas9. An agarose gel showing the results of electrophoretic separation of targets with PAM sequences shown at the top after incubation with the PpCas9 effector complex is presented. Bands corresponding to cleaved and uncleaved DNA fragments are indicated.

D. Wheel representation of *in vitro* PAM screen results for 5th, 6th, and 7th nucleotide positions of PAM. Nucleotide positions from the inner to the outer circle match PAM positions moving away from the protospacer. For a given sequence, the area of the sector in the PAM wheel is proportional to the relative depletion in the library.

E. PpCas9 efficiently cleaves different DNA targets with 5'-NNNNRTN-3' PAM consensus *in vitro*. The scheme above shows the positions of different target sites in a 1592 bp GRIN2b gene fragment. Below, a gel showing results of *in vitro* cleavage of targets with indicated PAMs is presented.

Figure 5. Cleavage activity of DfCas9 and PpCas9 at different temperatures.

DfCas9 or PpCas9 was incubated with cognate tracrRNA and crRNA and a 2.7 kb plasmid DNA or a 921 bp linear DNA fragment containing target sequences at indicated temperatures for 10 min. Products were separated by agarose gel electrophoresis. Cleavage efficiency (in percent) was calculated as a ratio of intensity of cleaved bands to the combined intensity of cleaved and uncleaved bands. Mean values and standard deviations obtained from three independent experiments are shown.

Figure 6. The DfCas9 and PpCas9 minimal *in vitro* DNA cleavage systems.

A. A scheme of recognition by the DfCas9-sgRNA complex of a DNA target flanked by 5'-NNRNAY-3' PAM (Y stands for pyrimidines, R stands for purines). The crRNA-tracrRNA linker is indicated with a grey box. The part of sgRNA that originated from tracrRNA is shown in blue.

B. A scheme of recognition by the PpCas9-sgRNA complex of a DNA target flanked by 5'-NNNNRTN-3 PAM (R stands for purines). The crRNA-tracrRNA linker is indicated by a grey box. The part of sgRNA that originated from tracrRNA is shown in blue.

Figure 7. PpCas9 nuclease activity in human HEK293T cells.

A. Scheme of the PpCas9 nuclease activity assessment experiment. Above - a scheme showing design of a plasmid used for PpCas9 gene and sgRNAs expression. The PpCas9 gene is shown as a yellow rectangle, NLS (nuclear localization signals) as brown rectangles, GFP gene as a green rectangle. CMV promoter and U6 promoters are indicated with black arrows. The sgRNA coding sequence is shown as a green rectangle. The plasmid was transfected into HEK293T cells and genomic DNA was extracted from a heterogeneous population of modified and unmodified cells for indel frequency assessment through HTS of the targeted region or *in vitro* assay with T7 endonuclease I.

B. Results of T7 endonuclease I indel detection assay showing PpCas9-mediated cleavage of EMX1 and GRIN2b genes in HEK293T genome.

C. PpCas9 indel formation efficiency at different genomic sites. Left – genomic DNA target sites with corresponding PAM sequences. 5'-NNNNRTT-3' PAM are shown in red. Right - indel frequency estimated by HTS analysis. Mean values and standard deviations obtained from three biological replicas are shown.

D. The influence of sgRNA spacer length on PpCas9-mediated indel formation efficiency in EMX1 and GRIN2b genes. HEK293T cells were transfected with PpCas9_sgRNA plasmids (as in panel A) coding for sgRNAs with different lengths of spacer segments. Left - results for the GRIN2b.1 target, right – for the EMX1.1 target. Mean values and standard deviations obtained from three biological replicas are shown.

Figure 8. The specificity of genomic DNA cleavage by PpCas9.

The indel frequency at two on-target as well as at corresponding off-target sequences was assessed by targeted amplicon sequencing of genomic DNA of HEK293T cells transfected with plasmids carrying the PpCas9 genome editing system. Left - sequences of on-target sites (in green) and off-target sites are shown. Mismatches in off-target sequences are shown in red. Right – frequencies of indel formation in each site.

Supplementary Table S1. DNA sequences used in this study

locus_DfCas9_F primer	atctcaagaagatcatcttattaatcagataaaatatttctagaTGGA CGGAACACAGGGGCCGGACGTG
locus_DfCas9_R primer	caatttaactgtgataaactaccgcattaaagcttATCACTGAA CATGTCCCTGATGCTGATCGAG
pACYC184_DfCas9_locus plasmid	https://benchling.com/s/seq-aePyMryY8SWwdax2uXmp
CRISPR-PpCas9 locus	https://benchling.com/s/seq-2rHKjZaY71sBPevLcrXM
DfCas9_R primer	gagtgcggccgcaagcttAACTGTCCCATGCGGGATC GTGTGAACCCGCCCCAATTCGTCGATCCGCAC
DfCas9_F primer	gaaggagatatacatatgATGTATCGTTTCGCTTTCGA CCTCGGAACCAAC
pET21a_DfCas9 plasmid	https://benchling.com/s/seq-Rcl0JPmGbUo5J2ef8tQq
pET21a_PpCas9 plasmid	https://benchling.com/s/seq-RSWQPFNAnE05Xg3X8gkV
7N_PUC19_DfCas9_library	https://benchling.com/s/seq-bt807FquL2f5Zdbb9kgy
PUC19_R primer	agcttggcgtaatcatggtcatag
PUC19_F primer	cccgggtaccgagctcga
Library_F primer	ctatgaccatgattacgccaagctgatcatgatcgacatgatcc cgaaggtctaNNNNNNNcccgggtaccgagctcga
Library_R primer	tcgagctcggtacccggg
M13_f primer	GTTGTAAAACGACGGCCAGTG
M13_r primer	AGCGGATAACAATTTCACACAGGA
1592 bp DNA fragment used for *in vitro* reactions	https://benchling.com/s/seq-atpZZkfQvY1aiGMgr25K
U6_sgRNA_BsmBI_CMV_PpCas9_P2A_GFP	https://benchling.com/s/seq-zFrQDDaVHSsxwMuyNuf8
U6_sgRNA_BsmBI_CMV_SpCas9_P2A_GFP	https://benchling.com/s/seq-g4bqSa99fdbF7JPn7eJp
GRIN2b_T7_endoI_fragment	https://benchling.com/s/seq-FCZIYuNF0rPKw4wXKBvD
EMX1_T7_endoI_fragment	https://benchling.com/s/seq-BuPF1uTcGA3WMTyOGtf0

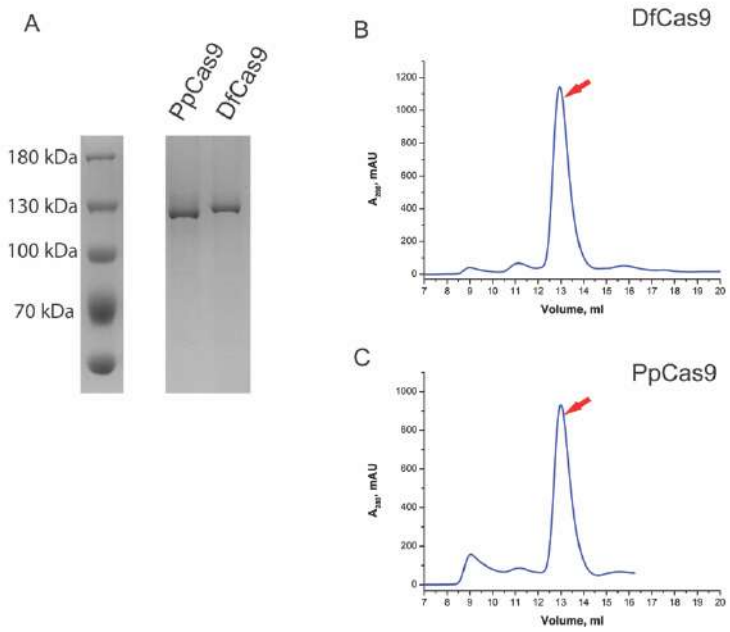
Supplementary Table S2. RNA sequences used in this study

PpCas9 crRNA	**GGG**tatctcctttcattgagcacGTTGTAGCTCCCTTTTTCATTTCGC
PpCas9 tracrRNA	**GG**GCGAAATGAAAAACGTTGTTACAATAAGAGATGAATTTCTCGCAAAGCT CTGCCTCTTGAAATTTCGGTTTCAAGAGGCATCTTTTT
DfCas9 crRNA	**GGG**tatctcctttcattgagcacGTCCGGGCTTGGCCACGCCGCTTCTTCTGCTAGGAT
DfCas9 tracrRNA	**GGG**TCCTAGCAGAAGAAGCGGCGTGGTCTTTCCCGCGATAAGGTTAAAACCACACCAT TGGGGCAGGCTGCGGCCTGCCCCATCTGTTT
DfCas9 sgRNA 1	**GGG**CTTGGCCACGCCGCTTCGAAAGAAGCGGCGTGGTC TTTCCCGCGATAAGGTTAAAACCACACCATTGGGGCAGGCTGCGGCCTGCCCCATCTGTTT
DfCas9 sgRNA 2	**GGG**CTTGGCCACGCGAAAGCGTGGTCTTTCCCGCGATAA GGTTAAAACCACACCATTGGGGCAGGCTGCGGCCTGCCCCATCTGTTT
DfCas9 sgRNA 3	**GGG**TATCTCCTTTCATTGAGCACGTCCGGGCTTGGCCACGCCGCTTCTTCTGCGAAAGC AGAAGAAGCGGCGTGGTCTTTCCCGCGATAAGGTTAAAACCACACCATTGGGGCAGGC TGCGGCCTGCCCCATCTGTTT
PpCas9 sgRNA1	**GGG**TATCTCCTTTCATTGAGCACGTTGTAGCTCCCTTTTTCATTTCGCGAAAGCGAAATG AAAAACGTTGTTACAATAAGAGATGAATTTCTCGCAAAGCTCTGCCTCTTGAAATTTCGG

	TTTCAAGAGGCATCTTTTT
PpCas9 sgRNA2	**GGG**TATCTCCTTTCATTGAGCACGTTGTAGCTCCCTTTTTCATTTCGCAGTGCTATAATGAA AATTATAGCACTGCGAAATGAAAAACGTTGTTACAATAAGAGATGAATTTCTCGCAAAGC TCTGCCTCTTGAAATTTCGGTTTCAAGAGGCATCTTTTT

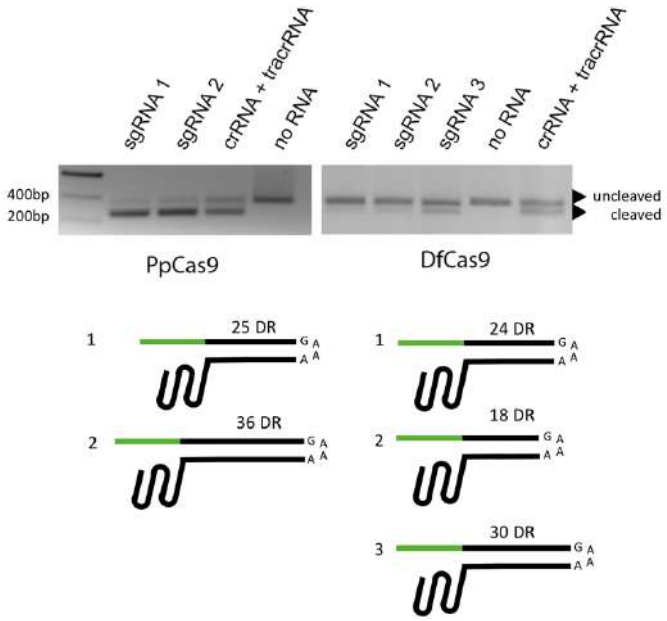
Supplementary Table S3. DNA targets in 1592 bp fragment used in *in vitro* DNA cleavage experiments.

PpCas9 grin2b targets	protospacer	PAM
target 1	ATATTGGAATACAAAAGTCA	GAAGATC
target 2	GTGGGCACTTCCGACGAGGT	GGCCATC
target 3	GCCTTGGCGTCAACATTTAT	AAAAATG
target 4	GGCAATGTTTTAAAACTACT	AGTCATG
target 5	CAAGAATGCAGGGCTTGTGT	ACTTATA
target 6	CTCTGCCTGTAGCTGCCAAT	GACTATA
target 7	TGAGCAGAAAAACGTGCTCA	CTTTGTC
target 8	TGGTGCTCAATGAAAGGAGA	TAAGGTC
target 9	CTAGCCTCTTCTAAGACAGG	TTACGTG
target 10	ATCCTCGTGGGCACTTCCGA	CGAGGTG
target 11	TATTTTATTCCCTACTGGGT	CATTGTA
target 12	AGGAGATAAGGTCCTTGAAT	TGCAGTA
DfCas9 grin2b targets	protospacer	PAM
target 1	TTGTAGTTTAATATTAAGAG	CCAGATA
target 2	CTCAATGAGGACAACAGCCA	GAAAATT
target 3	CAGAACATGGGCTGGATAAA	CTGGATG
target 4	ATCAATTAAGTGAAACAAAC	CTGAACC
target 5	CTCTGCAGAATGAGAGAAAA	TGAAACT
target 6	ATCTTCTGACTTTTGTATTC	CAATATG
target 7	GCTGCCAATGACTATAGCAA	TAGCACC
target 8	CTTCAGCCCAAGAACAGTAC	AAGGGTG
target 9	TATTAAACTACAAGAACAAC	TGATACA



Supplementary Figure S1. DfCas9 and PpCas9 purification.

A. SDS PAAG gel electrophoresis of purified PpCas9 and DfCas9 recombinant proteins.
B. Size exclusion chromatography elution of DfCas9 protein. Monomer fraction is marked with red arrow.
C. Size exclusion chromatography elution of PpCas9 protein. Monomer fraction is marked with red arrow.



Supplementary Figure S2. PpCas9 and DfCas9 sgRNA design.

Several variants of sgRNA were used for PpCas9 (left panel) or DfCas9 (right panel) DNA cleavage reactions *in vitro*. To estimate DNA cleavage efficiency reactions products were loaded to 1.5% agarose gel.

Supplementary Table S4. DNA targets in human cells genome used in the study.

PpCas9 grin2b targets	Protospacer (target site)	PAM
PpCas9 GRIN2b 1 sg20	CAGCTGAAGTAATGTTAGAG	CCACATT
PpCas9 GRIN2b 1 sg24	TTAGCAGCTGAAGTAATGTTAGAG	CCACATT
PpCas9 GRIN2b 2 sg20	AATAAGAAAAACATTATTAT	CACCATT
PpCas9 GRIN2b 2 sg24	ATAAAATAAGAAAAACATTATTAT	CACCATT
PpCas9 EMX1 1 sg20	GCCCTTCCTCCTCCAGCTTC	TGCCGTT
PpCas9 EMX1 1 sg24	TCAGGCCCTTCCTCCTCCAGCTTC	TGCCGTT
PpCas9 EMX1 2 sg20	GGAGGTGACATCGATGTCCT	CCCCATT
PpCas9 EMX1 2 sg24	CATTGGAGGTGACATCGATGTCCT	CCCCATT
SpCas9 GRIN2b sg20	ACCTTTTATTGCCTTGTTCA	AGG

Supplementary Table S5. First-Round PCR primers used in In-Del frequency analysis.

GRIN2b_1_target_different_ spacer_lengths_F	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNN**TAC GGTATCAGTCATTTTAGGGAAGTCACG**
GRIN2b_1_target_different_ spacer_lengths_R	TCAGACGTGTGCTCTTCCGATCT**ATGTGTTCTATTA CACTACGTGGAACTGCC**
EMX_1_target_different_spacer _lengths_F	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNN**CCT CCTGAGTTTCTCATCTGTGCCCCTCC**
EMX_1_target_different_spacer _lenghts_R	TCAGACGTGTGCTCTTCCGATCT**GGAGGTGACATC GATGTCCTCCCCATTGG**
GRIN2b_1_target_different_ targets_F	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNN**GG GAAGTCACGACTATAGGATGGCATCAGG**
GRIN2b_1_target_different_ targets_R	TCAGACGTGTGCTCTTCCGATCT**GAACACATATTAC TCCAATCTATTTATACACC**
GRIN2b_3_target_F	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNN**GG GAAGTCACGACTATAGGATGGCATCAGG**
GRIN2b_3_target_R	TCAGACGTGTGCTCTTCCGATCT**GAACACATATTAC TCCAATCTATTTATACACC**
GRIN2b_6_target_F	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNN**GG TTTGTTTCACTTAATTGATTGGTTCATGG**
GRIN2b_6_target_R	TCAGACGTGTGCTCTTCCGATCT**ATTGATCCTTACA ATGACCCAGTAGGG**
GRIN2b_7_target_F	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNN**CAC TTTGTCTGGCCTTGCTTTCCTTCAGC**
GRIN2b_7_target_R	TCAGACGTGTGCTCTTCCGATCT**TTGTGAGTGGTCC AGGTAGCCATGCG**
GRIN2b_8_target_F	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNN**ATA ATTGGTTATATGAGAGGCAGTTCCACG**

GRIN2b_8_target_R	TCAGACGTGTGCTCTTCCGATCT**GCTAAGTGTTCTAAGACCATGAACCAAT**
EMX_1_target_different_targets_F	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNN**CCCAGGTGAAGGTGTGGTTCCAGAACC**
EMX_1_target_different_targets_R	TCAGACGTGTGCTCTTCCGATCT**CAATGCGCCACCGGTTGATGTGATGG**
EMX_3_target_F	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNN**CTGTGAATGTTAGACCCATGGGAGCAGC**
EMX_3_target_R	TCAGACGTGTGCTCTTCCGATCT**TCAGGCTGAGCTGAGAGCCTGATGGG**
EMX_4_target_F	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNN**AGCTGGACTCTGGCCACTCCCTGG**
EMX_4_target_R	TCAGACGTGTGCTCTTCCGATCT**GAGAAGGCCAAGTGGTCCCAGGCC**
EMX_5_target_F	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNN**AAACGGCAGAAGCTGGAGGAGGAAGGG**
EMX_5_target_R	TCAGACGTGTGCTCTTCCGATCT**GGAGGTGACATCGATGTCCTCCCCATTGG**
EMX_14_target_F	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNN**GAAGCAGGCCAATGGGGAGGACATCG**
EMX_14_target_R	TCAGACGTGTGCTCTTCCGATCT**GGAGTGGCCAGAGTCCAGCTTGGG**
GRIN2b_1_target_1_off-target_F	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNN**ATGACAACAAAGATACAGAATACCAGAAGC**
GRIN2b_1_target_1_off-target_R	TCAGACGTGTGCTCTTCCGATCT**GGTGTGATGCTAGGTCTTTCTAACTTTTCC**
GRIN2b_1_target_2_off-target_F	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNN**ACAGAGAGACCCTTTAATTGAAGCCAGG**
GRIN2b_1_target_2_off-target_R	TCAGACGTGTGCTCTTCCGATCT**GAGTGGTGGAAAAGGGGATAGAGTGG**
GRIN2b_1_target_3_off-target_F	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNN**ACACCTCCCATTGTACACACTTGGGAG**
GRIN2b_1_target_3_off-target_R	TCAGACGTGTGCTCTTCCGATCT**GTGCTTTTAACAGGATGAAGTGGATTGGG**
GRIN2b_1_target_4_off-target_F	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNN**TGAAAATAGAGATACCATCTACCAAAATCG**
GRIN2b_1_target_4_off-target_R	TCAGACGTGTGCTCTTCCGATCT**TAGCGATCTTTCTAACTTCTTAATGAAGG**
GRIN2b_1_target_5_off-target_F	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNN**GAACTGAATGAAAATAACAACACAACATAC**
GRIN2b_1_target_5_off-target_R	TCAGACGTGTGCTCTTCCGATCT**GCTTAGGTTATTGATTTGAGACTTTTCTCC**
GRIN2b_1_target_6_off-target_F	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNN**TTTGAAACTAATAAGAACAGACAACATACC**
GRIN2b_1_target_6_off-target_R	TCAGACGTGTGCTCTTCCGATCT**TGAGATCTTTCTAGCTTTCTGATGTGGG**
EMX_1_target_1_off-target_F	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNN**AATCAAAATTTTTCCATGAGGGAGAACAGC**
EMX_1_target_1_off-target_R	TCAGACGTGTGCTCTTCCGATCT**TTGAATTGAGTTCAGGGTGGTGGAAAGG**

EMX_1_target_2_off-target_F	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNN**AGTCAAGATCTGCCCTCAGCCATGTGG**
EMX_1_target_2_off-target_R	TCAGACGTGTGCTCTTCCGATCT**GGATGTCCCAGCTGAAGCACAGAGAGC**
EMX_1_target_3_off-target_F	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNN**ACTGGGGTTCACTTCCTCCTTGTTGCC**
EMX_1_target_3_off-target_R	TCAGACGTGTGCTCTTCCGATCT**GACCTGTGGGTTTTGAAGAAGATGTGGG**
EMX_1_target_4_off-target_F	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNN**TCTGGCCCTGTGTGGGCTTTGATATTGC**
EMX_1_target_4_off-target_R	TCAGACGTGTGCTCTTCCGATCT**TTCAGCTGAGTACTGGTCAGCACACCTG**
EMX_1_target_5_off-target_F	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNN**GCTGGTGAGAGCTTACCTCCACTCAGG**
EMX_1_target_5_off-target_R	TCAGACGTGTGCTCTTCCGATCT**GGATTCTCAGAATGGACTGTCTGAGCTTCC**
EMX_1_target_6_off-target_F	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNN**AGAGGGAAAGAAGGCTGTGTCGGAGCC**
EMX_1_target_6_off-target_R	TCAGACGTGTGCTCTTCCGATCT**CCCCCATCCCACCCCAAGGATGTTCC**
EMX_1_target_7_off-target_F	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNN**GTTATTTATCTCCAAAGAGAAGAGAAAGGG**
EMX_1_target_7_off-target_R	TCAGACGTGTGCTCTTCCGATCT**CCTAGTCTGCCATATATGCTTAAAATGG**
EMX_1_target_8_off-target_F	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNN**ATTACCCAGTCTCTGGTAGTTCTTTATAGC**
EMX_1_target_8_off-target_R	TCAGACGTGTGCTCTTCCGATCT**ATATTTGTCCCTGCCCAAAACTCATGC**
EMX_1_target_9_off-target_F	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNN**CACATTTTAATTTCCGTCTTTTACCTTTCC**
EMX_1_target_9_off-target_R	TCAGACGTGTGCTCTTCCGATCT**GAGGTGTCATTTGATACAGGAATTGACC**

Chapter III

Crystal structure of Cpf1 in complex with guide RNA and target DNA

Introduction

This chapter is dedicated to the study of Cas12a, an effector enzyme of Type V-A CRISPR-Cas systems. In this work, the crystal structure of AsCas12a (former AsCpf1) was solved and the mechanism of Cas12a DNA cleavage by two nuclease domains was proposed.

The study shows that structure of Cas12a is different from that of Cas9: it has an additional domain adjacent to the RuvC domain that we called Nuc. Mutational analysis of AsCas12a showed that mutations in the RuvC domain active center abolish the double strand DNA cleavage, while mutations in the Nuc domain abolish target strand cleavage. As Arginine 1226 of the Nuc domain interacts with the RuvC active site, it was proposed that Cas12a first cleaves the non-target strand by the RuvC domain, which activates the Nuc domain through Arginine 1226 along with other amino acids and leads to Nuc-mediated target strand DNA cleavage.

Further studies of other Type-V Cas effectors performed by other groups showed that Nuc domain is likely responsible for interchange of DNA strands in the RuvC domain needed to achieve double-strand DNA cleavage. These corrections notwithstanding, the obtained here results shed a light on Cas12a proteins architecture and unique structure properties.

Contribution

Work described in this chapter was done during my internship at the Broad Institute, at the Feng Zhang laboratory. I was working with Bernd Zetsche, who recently characterized then new type of CRISPR-Cas nucleases, Cas12a.

When I joined the work, the first author, Takashi Yamano with colleagues solved the crystal structure of AsCas12a. It appeared that in addition to the RuvC domain there is another domain, the Nuc domain, which was hypothesized to cleave the second DNA strand. Further mutagenesis analysis was needed to prove the hypothesis.

My contribution to this work was functional analysis of AsCas12a Nuc domain through performing of *in vitro* DNA cleavage reactions with nuclease mutant forms (Figure 6F). I incubated the lysate of HEK293 cells expressing different mutant forms of AsCas12a with a crRNAs and corresponding DNA targets to find which of the mutants nicks DNA. Use of 5'- or 3'-end fluorescent labeled linear DNA targets allowed me to show that mutation in 1226 Arginine to Alanine abolishes AsCas12a target DNA strand cleavage.

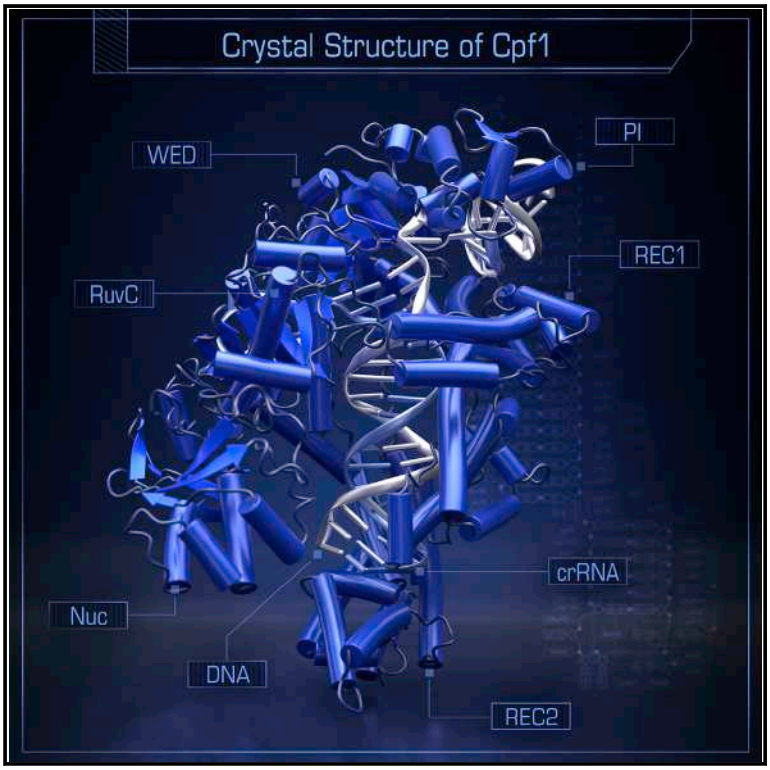
The first author of the paper and the corresponding authors wrote the manuscript.

Article

Cell

Crystal Structure of Cpf1 in Complex with Guide RNA and Target DNA

Graphical Abstract



Authors

Takashi Yamano, Hiroshi Nishimasu, Bernd Zetsche, ..., Ryuichiro Ishitani, Feng Zhang, Osamu Nureki

Correspondence

zhang@broadinstitute.org (F.Z.), nureki@bs.s.u-tokyo.ac.jp (O.N.)

In Brief

The structure of Cpf1, a type V CRISPR-Cas effector nuclease, in complex with crRNA and its target DNA provides mechanistic insights into RNA-guided DNA cleavage by Cpf1 and establishes a framework for rational engineering of the CRISPR-Cpf1 toolbox.

Highlights

- Crystal structure of *Acidaminococcus sp.* Cpf1 in complex with crRNA and target DNA
- Mechanistic insights into Cpf1-induced, staggered DNA double-strand breaks
- Recognition of the 5′-TTTN-3′ PAM via base and shape readout mechanisms
- Striking similarity and major differences between the structures of Cpf1 and Cas9

Accession Numbers

5B43

Yamano et al., 2016, Cell *165*, 949–962
May 5, 2016 ©2016 Elsevier Inc.
http://dx.doi.org/10.1016/j.cell.2016.04.003

CellPress

Cell

Crystal Structure of Cpf1 in Complex with Guide RNA and Target DNA

Takashi Yamano,[1,11] Hiroshi Nishimasu,[1,2,11] Bernd Zetsche,[3,4,5,6,7] Hisato Hirano,[1] Ian M. Slaymaker,[3,4,5,6] Yinqing Li,[3,4,5,6] Iana Fedorova,[3,4,5,6,8,9] Takanori Nakane,[1] Kira S. Makarova,[10] Eugene V. Koonin,[10] Ryuichiro Ishitani,[1] Feng Zhang,[3,4,5,6,*] and Osamu Nureki[1,*]

[1]Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo 113-0032, Japan
[2]JST, PRESTO, Tokyo 113-0032, Japan
[3]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[4]McGovern Institute for Brain Research
[5]Department of Brain and Cognitive Sciences
[6]Department of Biological Engineering
Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[7]Department of Developmental Pathology, Institute of Pathology, Bonn Medical School, 53127 Bonn, Germany
[8]Peter the Great St. Petersburg Polytechnic University, St. Petersburg, 195251, Russia
[9]Skolkovo Institute of Science and Technology, Skolkovo, 143026, Russia
[10]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA
[11]Co-first author
*Correspondence: zhang@broadinstitute.org (F.Z.), nureki@bs.s.u-tokyo.ac.jp (O.N.)
http://dx.doi.org/10.1016/j.cell.2016.04.003

SUMMARY

Cpf1 is an RNA-guided endonuclease of a type V CRISPR-Cas system that has been recently harnessed for genome editing. Here, we report the crystal structure of *Acidaminococcus sp.* Cpf1 (AsCpf1) in complex with the guide RNA and its target DNA at 2.8 Å resolution. AsCpf1 adopts a bilobed architecture, with the RNA-DNA heteroduplex bound inside the central channel. The structural comparison of AsCpf1 with Cas9, a type II CRISPR-Cas nuclease, reveals both striking similarity and major differences, thereby explaining their distinct functionalities. AsCpf1 contains the RuvC domain and a putative novel nuclease domain, which are responsible for cleaving the non-target and target strands, respectively, and for jointly generating staggered DNA double-strand breaks. AsCpf1 recognizes the 5′-TTTN-3′ protospacer adjacent motif by base and shape readout mechanisms. Our findings provide mechanistic insights into RNA-guided DNA cleavage by Cpf1 and establish a framework for rational engineering of the CRISPR-Cpf1 toolbox.

INTRODUCTION

The microbial adaptive immune system CRISPR-Cas (clustered regularly interspaced short palindromic repeats and CRISPR-associated proteins) helps bacteria and archaea defend themselves against the invasion of foreign nucleic acids (Marraffini, 2015; Wright et al., 2016). The CRISPR-Cas systems encompass arrays of direct repeats that are separated by unique spacers derived from foreign DNA. The repeat arrays are transcribed into long transcripts (precursors of CRISPR RNAs), which are then processed to yield small CRISPR RNAs (crRNAs), consisting of a spacer and a portion of the adjacent direct repeat. The crRNAs form a complex with Cas endonucleases, and in some cases with accessory Cas proteins as well, and serve as guides to target and cleave the cognate foreign nucleic acid, thus achieving interference. DNA recognition by Cas-crRNA complexes requires the presence of a protospacer adjacent motif (PAM) near the target site, which contributes to self versus non-self discrimination (Westra et al., 2013). The diverse spectrum of the CRISPR-Cas systems is broadly divided into two classes, depending on the architecture of the interference module (Makarova et al., 2015): class 1 systems use a complex of several Cas proteins, as exemplified by Cascade (Brouns et al., 2008; Redding et al., 2015), and class 2 systems use a single enzyme, such as Cas9 (Jinek et al., 2012; Gasiunas et al., 2012). Cas9 is a dual RNA-guided endonuclease that recognizes, binds, and cleaves target DNA, and it has been harnessed to create precision genome engineering tools (Cong et al., 2013; Mali et al., 2013).

Following the initial demonstration of the feasibility of using Cas9 to edit mammalian genomes, there was a burst of efforts to further adapt this endonuclease for a range of applications, from high-throughput gain-of-function screening (Gilbert et al., 2014; Konermann et al., 2015) to targeted modulation of histone marks (Hilton et al., 2015; Kearns et al., 2015). The development of these applications has been furthered, in part, through rational engineering, made possible by extensive biochemical and biophysical studies and the availability of several crystal structures of Cas9 (Nishimasu et al., 2014; Jinek et al., 2014; Anders et al., 2014; Nishimasu et al., 2015; Jiang et al., 2015, 2016; Hirano et al., 2016). Structure-guided engineering and direct evolution approaches have led to variants of Cas9 with enhanced target specificity (Slaymaker et al., 2016; Kleinstiver et al., 2016) or altered PAM requirements (Kleinstiver et al., 2015a, 2015b).

Recently, a second class 2 (type V) effector protein, Cpf1, has been harnessed for genome editing (Zetsche et al., 2015). Similar to Cas9, Cpf1 can be reprogrammed to target DNA sites of interest through complementarity to a guide RNA. However, Cpf1 possesses several unique features that distinguish it from Cas9 and could provide for a substantial expansion of the genome editing toolbox. First, Cpf1 is guided by a single crRNA, whereas Cas9 uses a crRNA and a second small RNA species, a *trans*-activating crRNA (tracrRNA) (Deltcheva et al., 2011). Second, Cpf1 recognizes a T-rich PAM, in contrast to the G-rich PAM favored by Cas9 (Fonfara et al., 2014; Karvelis et al., 2015). Third, Cpf1 generates staggered ends in its PAM-distal target site (Zetsche et al., 2015), whereas Cas9 creates blunt ends within the PAM-proximal target site (Garneau et al., 2010). Fourth, Cpf1 contains the RuvC domain but lacks a detectable second endonuclease domain (Zetsche et al., 2015), whereas Cas9 uses the HNH and RuvC endonuclease domains to cleave the target and non-target DNA strands, respectively (Jinek et al., 2012; Gasiunas et al., 2012). Together, these observations imply major differences in the target DNA recognition and cleavage mechanisms between Cas9 and Cpf1.

To clarify how Cpf1 recognizes and cleaves DNA targets, we determined the crystal structure of *Acidaminococcus sp.* Cpf1 (AsCpf1) in complex with the crRNA and its double-stranded DNA target containing the 5′-TTTN-3′ PAM. AsCpf1 adopts a bilobed architecture that accommodates the crRNA-target DNA heteroduplex in the central channel. AsCpf1 recognizes the crRNA scaffold and the 5′-TTTN-3′ PAM in structure- and sequence-dependent manners. AsCpf1 contains a RuvC endonuclease domain and a putative novel nuclease domain, which are located at positions suitable to induce staggered DNA double-strand breaks. The structural comparison of AsCpf1 with Cas9 reveals both striking structural similarity and substantial differences between the two class 2 effector proteins, thus explaining their distinct functionalities and suggesting their functional convergence.

RESULTS

Overall Structure of the AsCpf1-crRNA-Target DNA Complex

We solved the 2.8-Å resolution crystal structure of the full-length AsCpf1 (residues 1–1307) in complex with a 43-nt crRNA, a 34-nt target DNA strand, and a 10-nt non-target DNA strand containing a 5′-TTTN-3′ PAM, by the single-wavelength anomalous diffraction (SAD) method (Figures 1 and S1 and Table S1). The structure revealed that AsCpf1 adopts a bilobed architecture consisting of an α-helical recognition (REC) lobe and a nuclease (NUC) lobe, with the crRNA-target DNA heteroduplex bound to the positively charged, central channel between the two lobes (Figures 1C, 1D, and S2). The REC lobe consists of the REC1 and REC2 domains, whereas the NUC lobe consists of the RuvC domain and three additional domains, denoted A, B, and C (Figure 1C).

A Dali search (Holm and Rosenström, 2010) detected no structural similarity between the REC1, REC2, and the A, B, and C domains and any of the available protein structures. Sequence database searches using PSI-BLAST (Altschul et al., 1997) and HHPred (Söding et al., 2005) also failed to detect significant similarity between these domains and any protein sequences in the current databases. Thus, these Cpf1 domains have no detectable homologs outside the Cpf1 protein family and appear to adopt novel structural folds (Figures 1C and S3). The REC1 domain comprises 13 α helices, and the REC2 domain comprises ten α helices and two β strands that form a small antiparallel sheet (Figures S3A and S3B). Domains A and B play functional roles similar to those of the WED (Wedge) and PI (PAM-interacting) domains of Cas9 (Anders et al., 2014; Nishimasu et al., 2015; Hirano et al., 2016), respectively, although the two domains of AsCpf1 are structurally unrelated to the WED and PI domains (described below). Domain C is involved in DNA cleavage (described below). Thus, domains A, B, and C are referred to as the WED, PI, and Nuc domains, respectively. The WED domain is assembled from three separate regions (WED-I-III) in the Cpf1 sequence (Figures 1A, S3A, and S3C). The WED domain can be divided into a core subdomain comprising a nine-stranded, distorted antiparallel β sheet (β1-β8 and β11) flanked by seven α helices (α1-α6 and α9) and a subdomain comprising two β strands (β9 and β10) and two α helices (α7 and α8) (Figures S3A and S3C). Examination of the Cpf1 sequence alignment revealed that helices α7 and α8 are not conserved among Cpf1 homologs (Zetsche et al., 2015) (Figure S4). The PI domain comprises seven α helices (α1-α7) and a β hairpin (β1 and β2) and is inserted between the WED-II and WED-III regions, whereas the REC lobe is inserted between the WED-I and WED-II regions (Figures 1A, S3A, and S3B). As discussed previously (Zetsche et al., 2015), the RuvC domain contains the three motifs (RuvC-I-III) that form the endonuclease active center. A characteristic helix (referred to as the bridge helix) is located between the RuvC-I and RuvC-II motifs and connects the REC and NUC lobes (described below) (Figures 1A, 1C, and 1D). The Nuc domain is inserted between the RuvC-II and RuvC-III motifs.

Structure of the crRNA and Target DNA

The crRNA consists of the 24-nt guide segment (G1–C24) and the 19-nt scaffold (A(−19)–U(−1)) (referred to as the 5′ handle) (Figures 2A and 2B). The nucleotides G1–C20 in the crRNA and dC1–dG20 in the target DNA strand form the 20-bp RNA-DNA heteroduplex (Figures 2A and 2B). The nucleotide A21 in the crRNA is flipped out and adopts a single-stranded conformation. No electron density was observed for the nucleotides A22–C24 in the crRNA and dT21–dG24 in the target DNA strand, suggesting that these regions are flexible and disordered in the crystal structure. The nucleotides dG(−10)–dT(−1) in the target DNA strand and dC(−10*)–dA(−1*) in the non-target DNA strand form a duplex structure (referred to as the PAM duplex) (Figures 2A and 2B).

The crystal structure revealed that the crRNA 5′ handle adopts a pseudoknot structure, rather than a simple stem-loop structure predicted from its nucleotide sequence (Zetsche et al., 2015) (Figures 2A and 2C). Specifically, the G(−6)–A(−2) and U(−15)–C(−11) in the 5′ handle form a stem structure, via five Watson-Crick base pairs (G(−6):C(−11)–A(−2):U(−15)), whereas C(−9)–U(−7) in the 5′ handle adopt a loop structure. U(−1) and U(−16) form a non-canonical U•U base pair (Figure 2D). U(−10) and A(−18) form a reverse Hoogsteen A•U

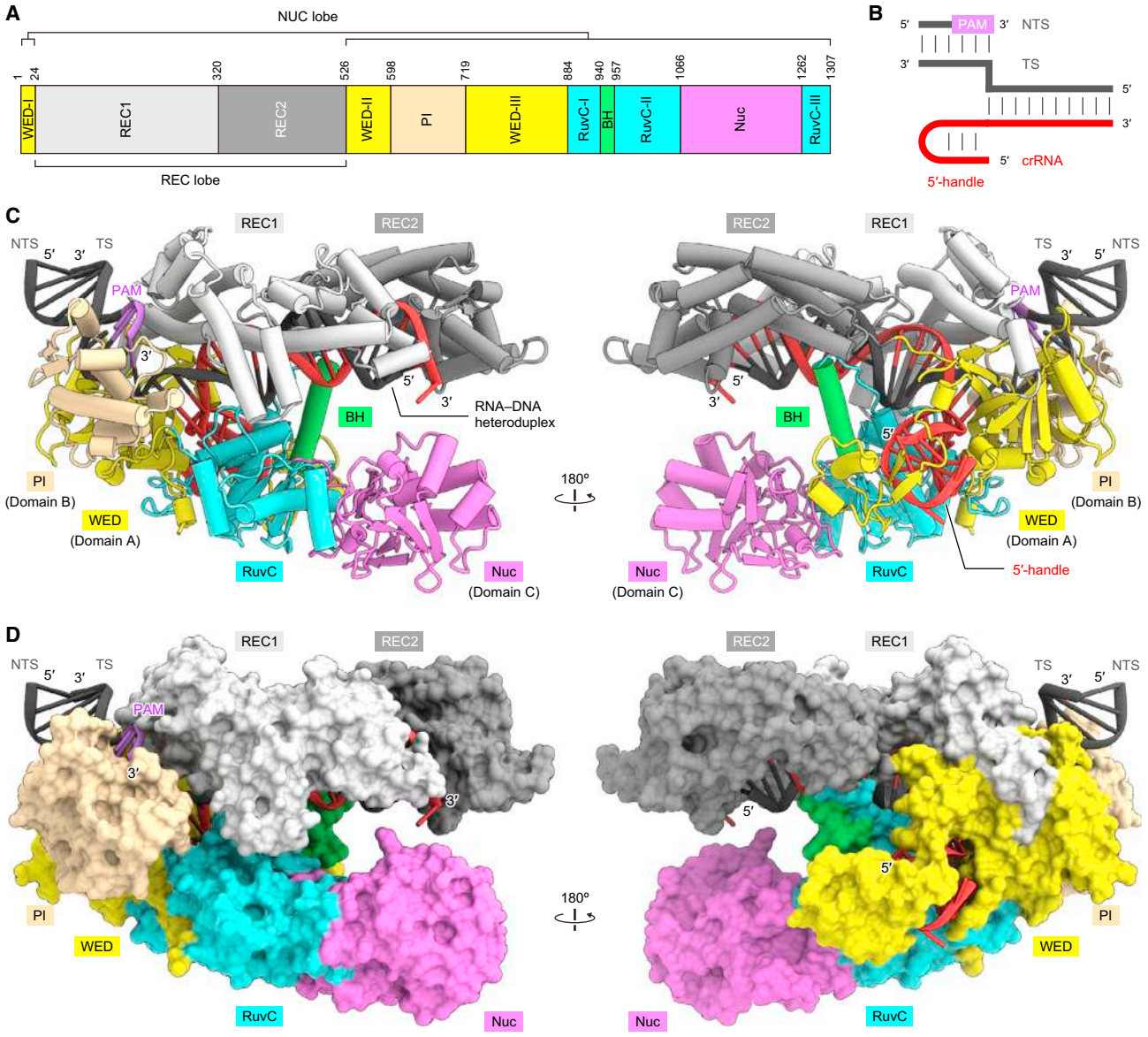


Figure 1. Overall Structure of the AsCpf1-crRNA-Target DNA Complex

(A) Domain organization of AsCpf1. BH, bridge helix.

(B) Schematic representation of the crRNA and target DNA. TS, target DNA strand; NTS, non-target DNA strand.

(C and D) Cartoon (C) and surface (D) representations of the AsCpf1-crRNA-DNA complex. Molecular graphic images were prepared using CueMol (http://www.cuemol.org).

See also Figures S1, S2, and S3 and Table S1.

base pair, and participate in pseudoknot formation (Figure 2E). The O4 and the 2′-OH of U(−10) hydrogen bond with the 2′-OH and the N1 of A(−19), respectively (Figure 2E). In addition, the N3 and the O4 of U(−17) hydrogen bond with the O4 of U(−13) and the N6 of A(−12), respectively, thereby stabilizing the pseudoknot structure (Figure 2F). Importantly, U(−1), U(−10), U(−16) and A(−18) in the crRNA are conserved among the CRISPR-Cpf1 systems (Zetsche et al., 2015), indicating that Cpf1 crRNAs form similar pseudoknot structures.

Recognition of the 5′ Handle of the crRNA

The 5′ handle of the crRNA is bound at the groove between the WED and RuvC domains (Figure 2G). The U(−1)•U(−16) base pair in the 5′ handle is recognized by the WED domain in a base-specific manner. U(−1) and U(−16) hydrogen bond with His761 and Arg18/Asn759, respectively, while U(−1) stacks on His761 (Figure 2H). These interactions explain the previous finding that the U•U base pair at this position is critical for the Cpf1-mediated DNA cleavage (Zetsche et al., 2015). The N6 of

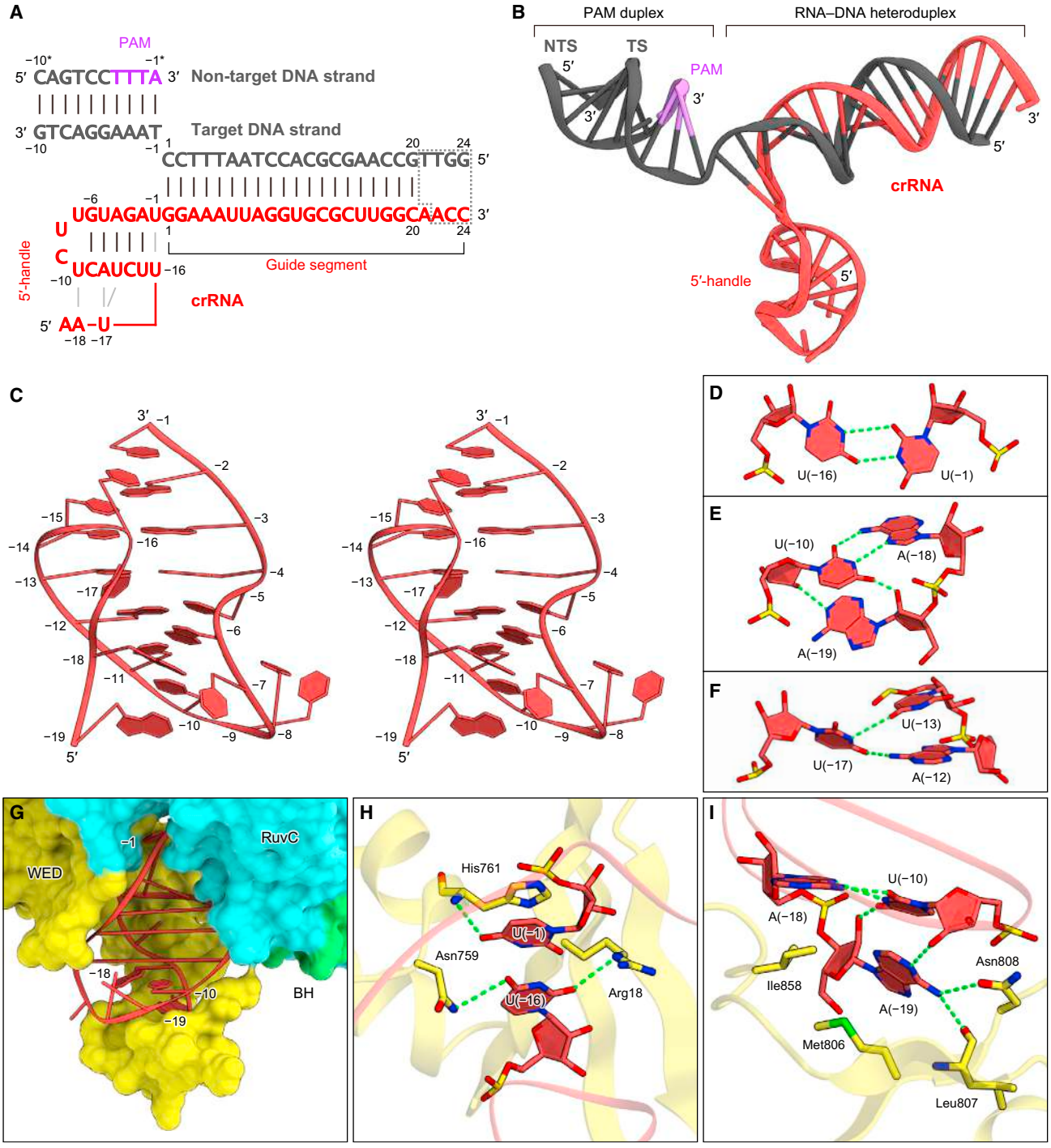


Figure 2. Structure of the crRNA and Target DNA

(A) Schematic representation of the AsCpf1 crRNA and the target DNA. The disordered region is surrounded by dashed lines.

(B) Structure of the AsCpf1 crRNA and the target DNA.

(C) Structure of the crRNA 5′ handle (stereo view).

(D–F) Close-up view of the U(−1)•U(−16) base pair (D), the reverse Hoogsteen U(−10)•A(−18) base pair (E), and the U(−13)-U(−17)-U(−12) base triple (F). Hydrogen bonds are shown as dashed lines.

(G) Binding of the crRNA 5′ handle to the groove between the WED and RuvC domains.

(H and I) Recognition of the 3′ end (H) and the 5′ end (I) of the crRNA 5′ handle. Hydrogen bonds are shown as dashed lines.

A(−19) hydrogen bonds with Leu807 and Asn808, while the base moieties of A(−18) and A(−19) form stacking interactions with Ile858 and Met806, respectively (Figure 2I). Moreover, the phosphodiester backbone of the 5′ handle forms an extensive network of interactions with the WED and RuvC domains (Figure 3). The residues involved in the crRNA 5′ handle recognition are largely conserved in the Cpf1 protein family (Zetsche et al., 2015) (Figure S4), highlighting the functional relevance of the observed interactions between AsCpf1 and the crRNA.

Recognition of the crRNA-Target DNA Heteroduplex

The crRNA-target DNA heteroduplex is accommodated within the positively charged, central channel formed by the REC1, REC2, and RuvC domains and is recognized by the protein in a sequence-independent manner (Figures 3, 4A, 4B, and S2). The PAM-distal and PAM-proximal regions of the heteroduplex are recognized by the REC1-REC2 domains and the WED-REC1-RuvC domains, respectively (Figures 3 and 4A–4C). Arg951 and Arg955 in the bridge helix, which interact with the sugar-phosphate backbone of the target DNA strand (Figure 4B), are conserved among the Cpf1 family members (Zetsche et al., 2015) (Figure S4). Notably, the sugar-phosphate backbone of the nucleotides G1–A8 in the crRNA forms multiple contacts with the WED and REC1 domains (Figures 3 and 4C), and the base pairing within the 5-bp PAM-proximal "seed" region is important for Cpf1-mediated DNA cleavage (Zetsche et al., 2015). These observations suggest that, in the Cpf1-crRNA complex, the seed of the crRNA guide is preordered in a nearly A-form conformation and serves as the nucleation site for pairing with the target DNA strand, as observed in the Cas9-sgRNA complex (Jiang et al., 2015). In addition, the backbone phosphate group between dT(−1) and dC1 of the target DNA strand (referred to as the +1 phosphate) is recognized by the side chain of Lys780 and the main-chain amide group of Gly783 (Figure 4C). This interaction results in the rotation of the +1 phosphate group, thereby facilitating base pairing between dC1 in the target DNA strand and G1 in the crRNA, as also observed in the Cas9-sgRNA-target DNA complexes (Anders et al., 2014; Nishimasu et al., 2015). The residues involved in the heteroduplex recognition are conserved in most members of the Cpf1 family (Zetsche et al., 2015) (Figure S4), and the R176A, R192A, G783P, and R951A mutants exhibited reduced activities (Figure 4D), confirming their functional relevance. Together, these observations reveal the RNA-guided DNA recognition mechanism of Cpf1.

Unexpectedly, the present structure revealed that the 24-nt crRNA guide and the target DNA strand form a 20-bp, rather than 24-bp, RNA-DNA heteroduplex (Figure 4A). The side chain of Trp382 in the REC2 domain forms a stacking interaction with the C20:dG20 base pair in the heteroduplex and thus prevents base pairing between A21 and dT21 (Figure 4E). Indeed, the W382A mutant showed reduced activity (Figure 4D), highlighting its functional importance. Trp382 is conserved in some members of the Cpf1 family, whereas others contain aromatic residues in this position (Zetsche et al., 2015) (Figure S4). These observations indicate that Cpf1 recognizes the 20-bp RNA-DNA heteroduplex and can explain the previous finding that the *Francisella novicida* Cpf1 (FnCpf1) cleaved the target DNA in a similar manner, using either the 20- or 24-nt guide-containing crRNA (Zetsche et al., 2015).

Recognition of the 5′-TTTN-3′ PAM

The PAM duplex adopts a distorted conformation with a narrow minor groove, as often observed in AT-rich DNA (Rohs et al., 2009), and is bound to the groove formed by the WED, REC1, and PI domains (Figures 5A and S5A). The PAM duplex is recognized by the WED-REC1 and PI domains from the major and minor groove sides, respectively (Figure 5B). The dT(−1):dA(−1*) base pair in the PAM duplex does not form base-specific contacts with the protein (Figure 5B), consistent with the lack of specificity in the fourth position of the 5′-TTTN-3′ PAM. Lys607 in the PI domain is inserted into the narrow minor groove and plays critical roles in the PAM recognition (Figure 5B). The O2 of dT(−2*) forms a hydrogen bond with the side chain of Lys607, whereas the nucleobase and deoxyribose moieties of dA(−2) form van der Waals interactions with the side chains of Lys607 and Pro599/Met604, respectively (Figure 5C). Modeling of the dG(−2):dC(−2*) base pair indicated that a steric clash exists between the N2 of dG(−2) and the side chain of Lys607 (Figure S5B), suggesting that dA(−2):dT(−2*), but not dG(−2):dC(−2*), is accepted at this position. These structural observations can explain the requirement of the third T in the 5′-TTTN-3′ PAM. The 5-methyl group of dT(−3*) forms a van der Waals interaction with the side-chain methyl group of Thr167, whereas the N3 and N7 of dA(−3) form hydrogen bonds with Lys607 and Lys548, respectively (Figure 5D). Modeling of the dG(−3):dC(−3*) base pair indicated that a steric clash exists between the N2 of dG(−3) and the side chain of Lys607 (Figure S5C). These observations are consistent with the requirement of the second T in the PAM. The 5-methyl group of dT(−4*) is surrounded by the side-chain methyl groups of Thr167 and Thr539, whereas the O4′ of dA(−4) forms a hydrogen bond with the side chain of Lys607 (Figure 5E). Notably, the N3 and O4 of dT(−4*) form hydrogen bonds with the N1 of dA(−4) and the N6 of dA(−3), respectively (Figure 5E). Modeling indicated that dA(−3) would sterically clash with the modeled base pairs dT(−4):dA(−4*), dG(−4):dC(−4*), and dC(−4):dG(−4*) (Figure S5D). These structural observations are consistent with the requirement of the first T in the PAM. The K548A and M604A mutants exhibited reduced activities (Figure 5F), confirming that Lys548 and Met604 participate in the PAM recognition. More importantly, the K607A mutant showed almost no activity (Figure 5F), indicating that Lys607 is critical for the PAM recognition. Together, these results demonstrate that AsCpf1 recognizes the 5′-TTTN-3′ PAM via a combination of base and shape readout mechanisms. Thr167 and Lys607 are conserved throughout the Cpf1 family, and Lys548, Pro599, and Met604 are partially conserved (Zetsche et al., 2015) (Figure S4). These observations indicate that the Cpf1 homologs from diverse bacteria recognize their T-rich PAMs in similar manners, although the fine details of the interaction could vary.

The RuvC-like Endonuclease and a Putative Second Nuclease Domain

The RuvC domain comprises a typical RNase H fold, consisting of a five-stranded mixed β sheet (β1–β5) flanked by three

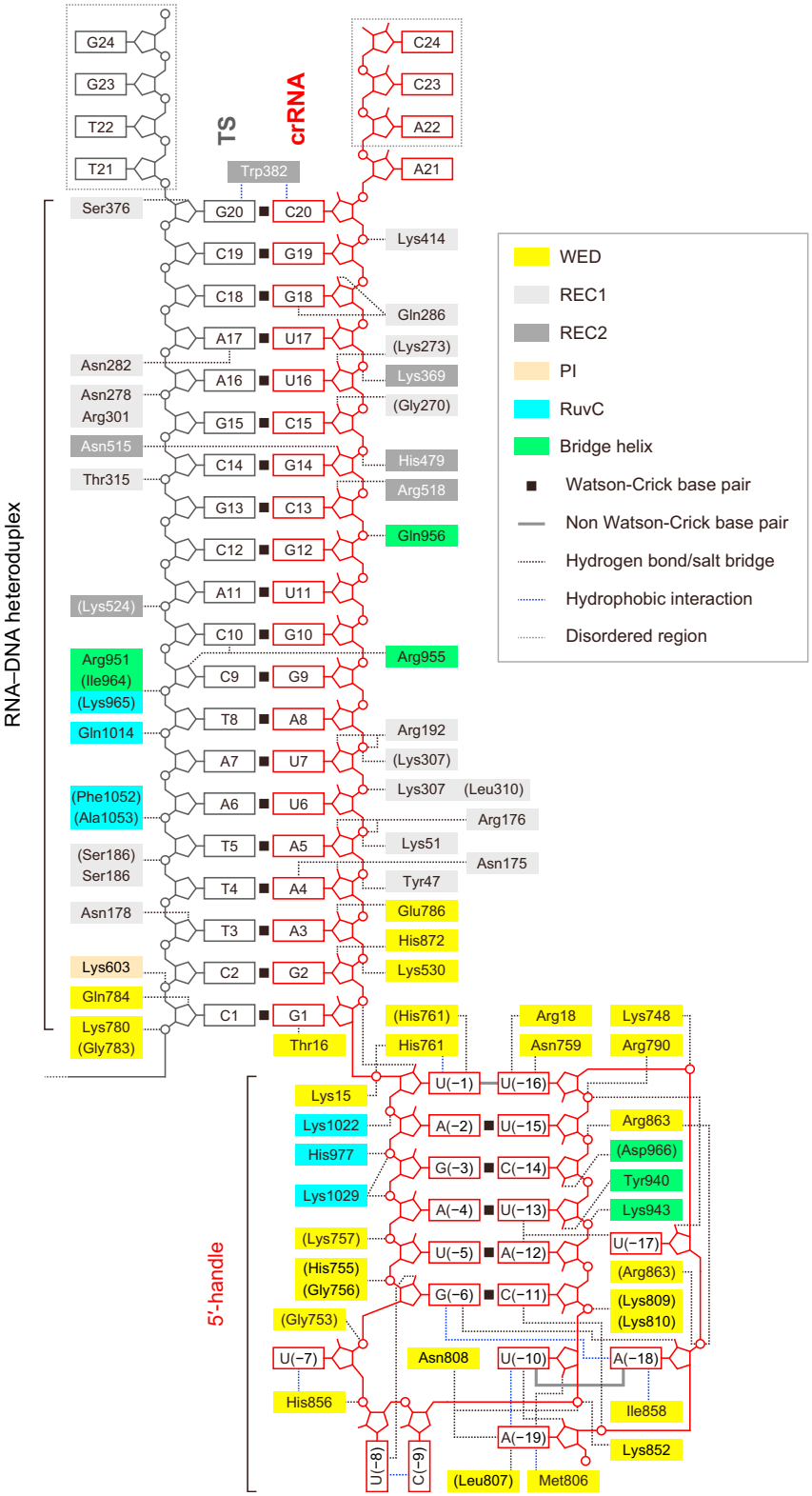


Figure 3. Schematic of Nucleic Acid Recognition by Cpf1

AsCpf1 residues that interact with the crRNA and the target DNA via their main chain are shown in parentheses. Water-mediated hydrogen-bonding interactions are omitted for clarity.
See also Figure S4.

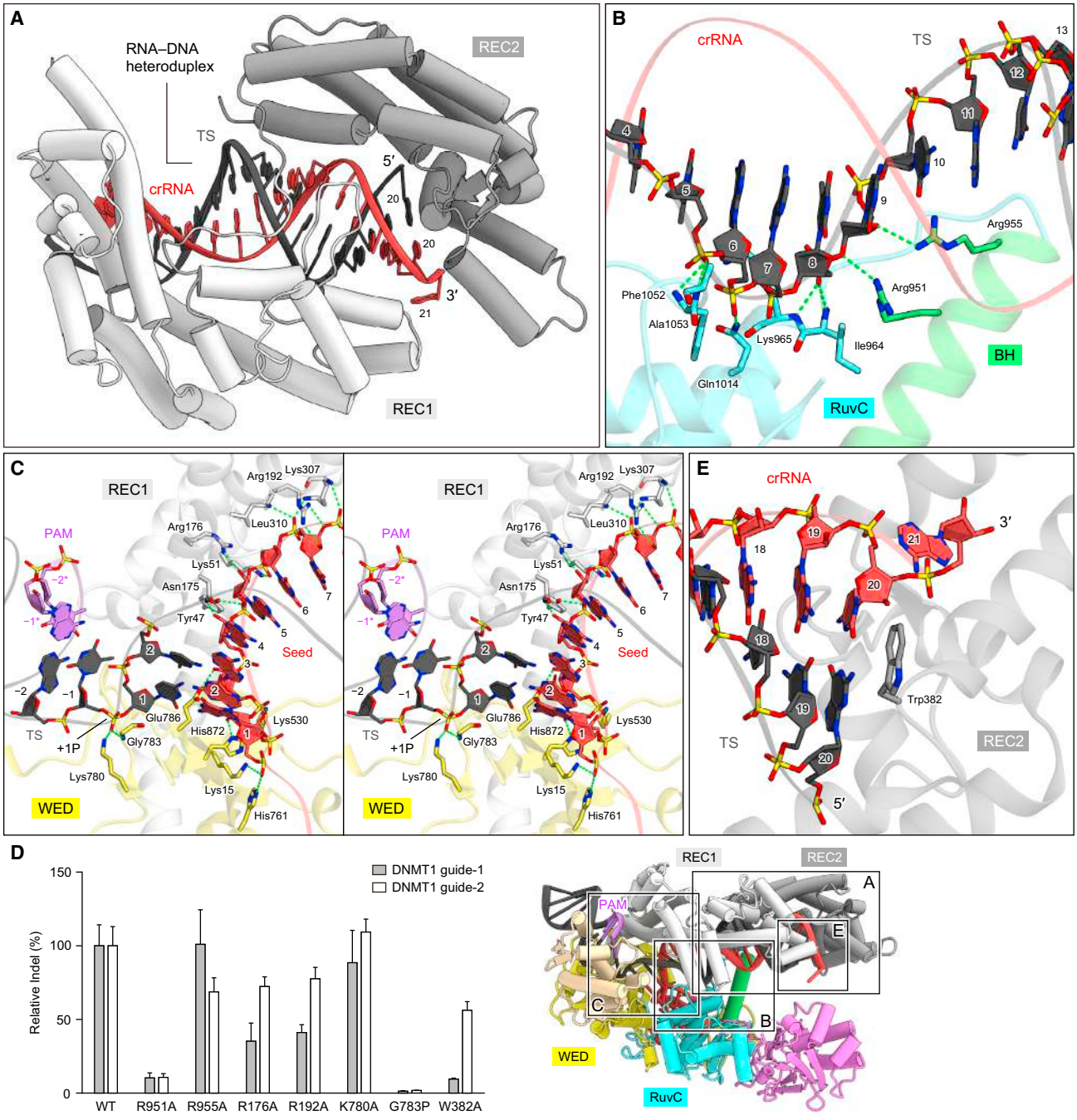


Figure 4. Recognition of the crRNA-Target DNA Heteroduplex

(A) Recognition of the crRNA-target DNA heteroduplex by the REC1 and REC2 domains.

(B) Recognition of the target DNA strand by the bridge helix and the RuvC domain. Hydrogen bonds are shown as dashed lines.

(C) Recognition of the crRNA seed region and the +1 phosphate group (+1P) (stereo view). Hydrogen bonds are shown as dashed lines.

(D) Mutational analysis of the nucleic-acid-binding residues. Effects of mutations on the ability to induce indels at two *DNMT1* targets were examined (n = 3, error bars show mean ± SEM).

(E) Stacking interaction between the 20th base pair in the heteroduplex and Trp382 of the REC2 domain.

α helices (α1–α3), and two additional α helices and three β strands (Figure 6A). The conserved, negatively charged residues Asp908, Glu993, and Asp1263 form an active site similar to that of the Cas9 RuvC domain (Nishimasu et al., 2014; Anders et al., 2014) (Figure 6B). As observed in FnCpf1 (Zetsche et al., 2015), the D908A and E993A mutants had almost no activity, whereas

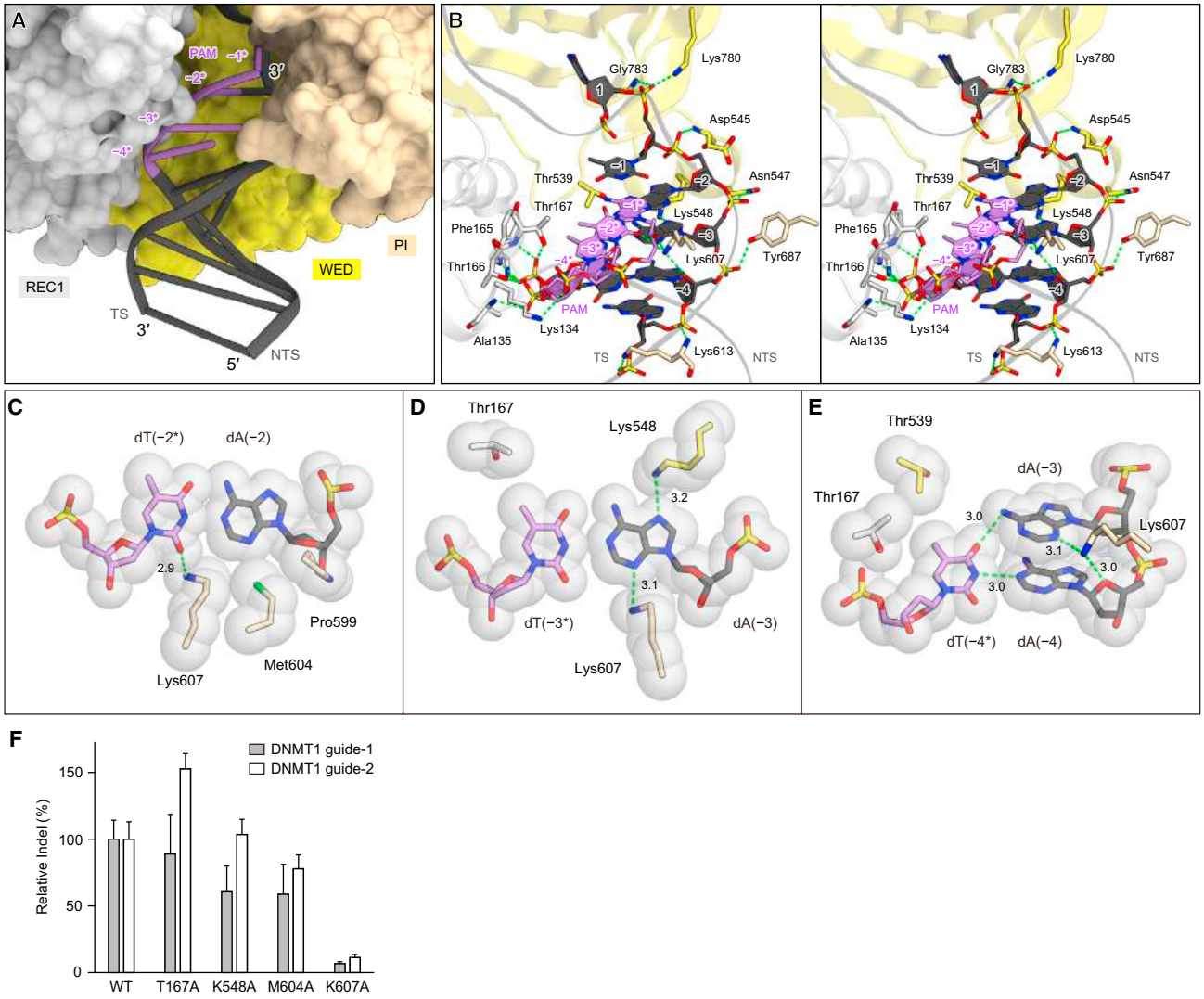


Figure 5. Recognition of the 5′-TTTN-3′ PAM

(A) Binding of the PAM duplex to the groove between the WED, REC1, and PI domains.

(B) Recognition of the 5′-TTTN-3′ PAM (stereo view). Hydrogen bonds are shown as dashed lines.

(C–E) Recognition of the dA(−2):dT(−2*) (C), dA(−3):dT(−3*) (D), and dA(−4):dT(−4*) (E) base pairs.

(F) Mutational analysis of the PAM-interacting residues. Effects of mutations on the ability to induce indels at two *DNMT1* targets were examined (n = 3, error bars show mean ± SEM).

See also Figure S5.

the D1263A mutant exhibited significantly reduced activity (Figure 6C), confirming the roles of Asp908, Glu993, and Asp1263 in DNA cleavage. Notably, the bridge helix is inserted between strand β3 and helix α1 in the RNase H fold and interacts with the REC2 domain (Figures 6A and 6D). The main-chain carbonyl group of Gln956 in the bridge helix forms a hydrogen bond with the side chain of Lys468 in the REC2 domain (Figure 6E). In addition, Trp958 in the RuvC domain is accommodated in the hydrophobic pocket formed by Leu467, Leu471, Tyr514, Arg518, Ala521, and Thr522 in the REC2 domain (Figure 6E). These residues, with the exceptions of Leu467 and Ala521, are highly conserved among the Cpf1 family members (Zetsche et al., 2015) (Figure S4), and the W958A mutant exhibited reduced activity (Figure 6C). These observations highlight the functional importance of the bridge helix-mediated interaction between the REC and NUC lobes.

The crystal structure revealed the presence of the Nuc domain, which is inserted between the RuvC-II (strand β5) and RuvC-III (helix α3) motifs in the RuvC domain. The Nuc domain is connected to the RuvC domain via two linker loops (referred to as L1 and L2) (Figure 6A). The Nuc domain comprises five α helices and nine β strands and lacks detectable structural or sequence similarity to any known nucleases or proteins. Notably, the conserved polar residues Arg1226 and Asp1235 and the

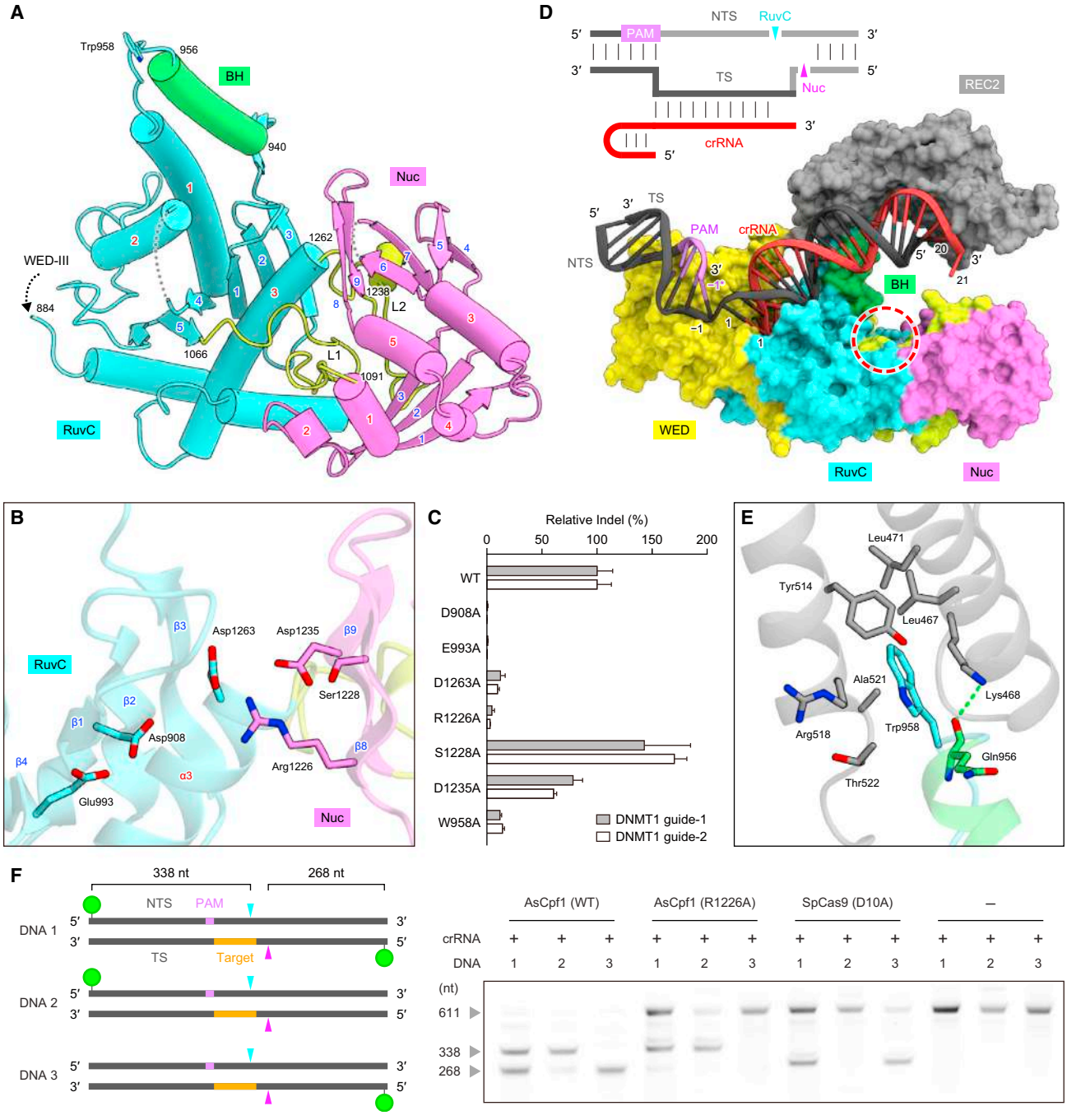


Figure 6. RuvC and Nuc Nuclease Domains

(A) Structures of the RuvC and Nuc domains. The α helices (red) and β strands (blue) in the RuvC (RNase H fold) and Nuc domains are numbered. Disordered regions are shown as dashed lines.

(B) Active site of the RuvC domain.

(C) Mutational analysis of key residues in the RuvC and Nuc domains. Effects of mutations on the ability to induce indels at two *DNMT1* targets were examined (n = 3, error bars show mean ± SEM). Indel values are normalized against wild-type AsCpf1.

(D) Spatial arrangement of the nuclease domains relative to the potential cleavage sites of the target DNA. The catalytic center of the RuvC domain is indicated by a red circle. The REC1 and PI domains are omitted for clarity. A schematic of the crRNA and target DNA is shown above the structure. The DNA strands not contained in the crystal structure are represented in light gray.

(E) Interaction between Trp958 and the hydrophobic pocket in the REC2 domain.

(legend continued on next page)

partially conserved Ser1228 are clustered in the proximity of the active site of the RuvC domain (Zetsche et al., 2015) (Figures 6B and S4). The S1228A mutant showed DNA cleavage activity comparable to that of wild-type AsCpf1 (Figure 6C). In contrast, the D1235A mutant exhibited reduced activity, whereas the R1226A mutant showed almost no activity (Figure 6C), indicating that Arg1226 is critical for DNA cleavage. Further characterization revealed that the R1226A mutant acts as a nickase that cleaves the non-target DNA strand, but not the target strand (Figure 6F), indicating that the Nuc and RuvC domains cleave the target and non-target DNA strands, respectively (Figure 6D). As in FnCpf1 (Zetsche et al., 2015), the mutations of the catalytic residues in the AsCpf1 RuvC domain abolished the cleavage of both DNA strands (Figure S6), suggesting that the cleavage of the non-target strand by the RuvC domain is a prerequisite for the target strand cleavage by the Nuc domain, presumably via a conformational change in the complex. However, further functional and structural studies are required to fully characterize the RNA-guided DNA cleavage mechanism of Cpf1.

DISCUSSION

The present structure of the AsCpf1-crRNA-target DNA complex provides mechanistic insights into RNA-guided DNA cleavage by Cpf1. The structural comparison between Cpf1 and Cas9, the only available structures of class 2 (single protein) effectors, illuminated the considerable similarity in their overall architectures, which was unanticipated given the lack of sequence similarity outside the RuvC domain (Figures 7A–7D). Both effector proteins are roughly the same size and adopt distinct bilobed structures, in which the two lobes are connected by the characteristic bridge helix and the crRNA-target DNA heteroduplex is accommodated in the central channel between the two lobes (Figures 7A and 7B). However, despite this overall similarity, only the RuvC nuclease domains of Cas9 and Cpf1 are homologous, whereas the rest of the proteins share neither sequence nor structural similarity.

One of the striking features of the Cas9 structure is the nested arrangement of the two unrelated HNH and RuvC nuclease domains, which cleave the target and non-target DNA strands, respectively (Figures 7A and 7C). In Cas9, the HNH domain is inserted between strand β4 and helix α2 of the RNase H fold in the RuvC domain (Nishimasu et al., 2014; Anders et al., 2014) (Figure 7E). In contrast, Cpf1 lacks the HNH domain and instead contains the Nuc domain, which is inserted at a different position (albeit also between the RuvC-II and RuvC-III motifs), i.e., between strand β5 and helix α3 of the RNase H fold (Figure 7F). Our mutational analysis suggested that the Nuc domain is a bona fide nuclease responsible for the target DNA strand cleavage, although the domain is relatively poorly conserved within the Cpf1 family and lacks sequence or structural similarity to any characterized nuclease (or any other protein outside the Cpf1 family). Notably, the Nuc domain of Cpf1 is located at a suitable position to cleave the single-stranded region of the target DNA strand outside the heteroduplex (Figures 7B and 7D), whereas the HNH domain of Cas9 cleaves the target DNA strand within the heteroduplex (Jinek et al., 2012; Gasiunas et al., 2012) (Figure 7C). These structural differences can explain why Cpf1 induces a staggered DNA double-strand break in the PAM-distal site, whereas Cas9 creates a blunt end in the PAM-proximal site (Zetsche et al., 2015). Unlike *Streptococcus pyogenes* Cas9 (SpCas9), in which inactivation of the RuvC nuclease turns the enzyme into a nickase that cleaves the target strand, an active RuvC domain is required for the cleavage of both strands by AsCpf1 (Figure 6F), suggesting that in Cpf1 the non-target strand cleavage by the RuvC domain is a prerequisite of the target strand cleavage by the Nuc domain. Together, these findings indicate that, despite the overall structural similarity and the apparent analogous roles of the two nuclease domains, there are substantial mechanistic differences between SpCas9 and AsCpf1. Further biochemical and structural studies with different members of the Cas9 and Cpf1 families are required to determine the generality of these distinctions between the two effector proteins and to completely elucidate the catalytic mechanism of Cpf1.

The structural comparison between Cpf1 and Cas9 revealed a striking degree of apparent structural and functional convergence between Cpf1 and Cas9, which is compatible with the previously proposed scenario of independent evolution of the effectors in the different types and subtypes of class 2 (Shmakov et al., 2015). Intriguingly, Cpf1 and Cas9 employ distinct structural features and recognize the seed region in the crRNA and the +1 phosphate group in the target DNA to achieve RNA-guided DNA targeting. In Cas9, the seed region is anchored by an arginine cluster in the bridge helix between the RuvC and REC domains, whereas the +1 phosphate group is recognized by the "phosphate lock" loop between the RuvC and WED domains (Anders et al., 2014; Nishimasu et al., 2015) (Figure S7A). In contrast, in Cpf1, the seed region is anchored by the WED and REC domains, whereas the +1 phosphate group is recognized by the WED domain (Figure S7B). Structural analyses of additional class 2 effectors, as well as the transposon-encoded TnpB proteins, which appear to be the evolutionary ancestors of the RuvC domains in the type II and type V effectors (Shmakov et al., 2015), are expected to shed further light on the evolution of this remarkable class of RNA-guided endonucleases.

The AsCpf1 structure also revealed notable differences in the PAM recognition mechanism between Cpf1 and Cas9. In Cas9, the PAM nucleotides in the non-target DNA strand are primarily read out from the major groove side, via hydrogen-bonding interactions with specific residues in the PI domain. In SpCas9, the second G and third G in the 5′-NGG-3′ PAM are recognized by Arg1333 and Arg1335 in the PI domain, via bidentate hydrogen

(F) The AsCpf1 R1226A mutant is a nickase cleaving the non-target DNA strand. The wild-type or the R1226A mutant (inactivation of the Nuc domain) of AsCpf1 was incubated with crRNA and the target DNA, which was labeled at the 5′ ends of both strands (DNA 1) or at the 5′ end of either the non-target strand (DNA 2) or the target strand (DNA 3). The cleavage products were analyzed by 10% polyacrylamide TBE-Urea denaturing gel electrophoresis. The SpCas9 D10A mutant (inactivation of the RuvC domain) is a nickase cleaving the target strand and was used as a control.
See also Figure S6.

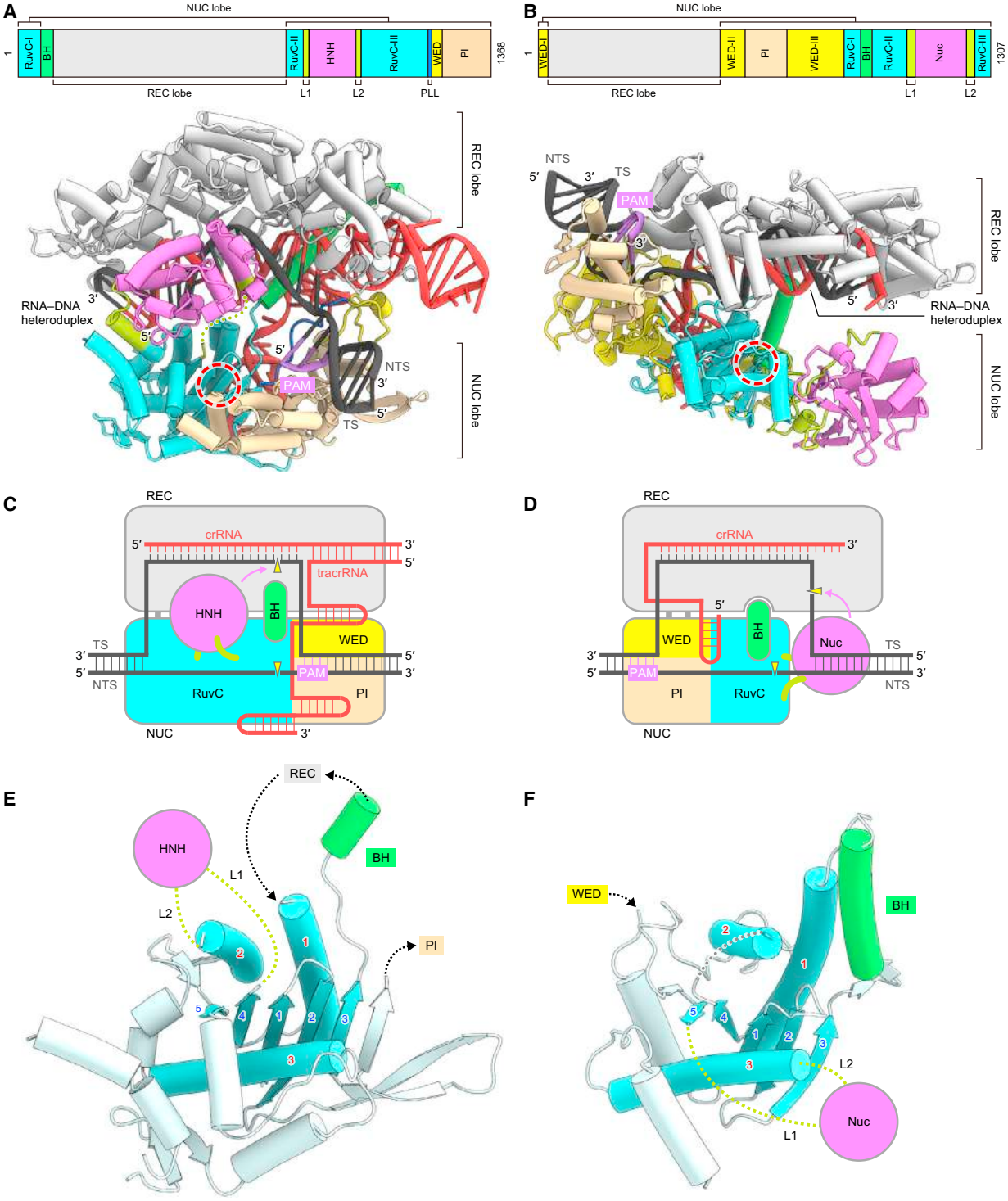


Figure 7. Comparison between Cas9 and Cpf1

(A and B) Comparison of the domain organizations and overall structures between SpCas9 (PDB: 4UN3) (A) and AsCpf1 (B). The catalytic centers of the RuvC domain are indicated by a red circle.

(C and D) Models of RNA-guided DNA cleavage by Cas9 (C) and Cpf1 (D).

(E and F) Comparison of the RuvC domains of SpCas9 (PDB: 4UN3) (E) and AsCpf1 (F). The secondary structures of the conserved RNase H fold are numbered. See also Figure S7.

bonds, respectively (Anders et al., 2014) (Figure S7A). In contrast, in AsCpf1, the PAM nucleotides in both the target and non-target DNA strands are read out by the PI domain from both the minor and major groove sides. In particular, as observed in other protein-DNA complexes (Rohs et al., 2009), the conserved lysine residue (Lys607 in AsCpf1) in the PI domain is inserted into the narrow minor groove of the PAM duplex and plays critical roles in the PAM recognition (Figure S7B). These structural observations show that, whereas Cas9 recognizes the PAM primarily via a base readout mechanism, Cpf1 combines base and shape readout to recognize the PAM. These mechanistic differences in the PAM recognition can explain why Cas9 orthologs recognize G-rich, diverse PAM sequences, whereas the widely different members of the Cpf1 family recognize similar T-rich PAMs (Zetsche et al., 2015).

In summary, the present structure of AsCpf1, combined with the mutational analysis of the two nuclease domains, provides mechanistic insights into the RNA-guided DNA recognition and cleavage by this recently discovered CRISPR-Cas effector protein and highlights the similarity and differences between the type V (Cpf1) and type II (Cas9) effectors. The structural analysis of Cas9 has enabled the design of numerous Cas9 variants with improved features and novel functions. Thus, the structural information described here will facilitate the engineering of Cpf1 and further increase the utility of the CRISPR-Cpf1 toolbox.

EXPERIMENTAL PROCEDURES

Sample Preparation

The gene encoding full-length AsCpf1 (residues 1–1307) was cloned between the *Nde*I and *Xho*I sites of the modified pE-SUMO vector (LifeSensors). The AsCpf1 protein was expressed at 20°C in *Escherichia coli* Rosetta2 (DE3) (Novagen) and was purified by chromatography on Ni-NTA Superflow (QIAGEN) and HiTrap SP HP (GE Healthcare) columns. The protein was incubated overnight at 4°C with TEV protease to remove the His_6-SUMO-tag and was then passed through the Ni-NTA column. The protein was further purified by chromatography on a HiLoad Superdex 200 16/60 column (GE Healthcare). The selenomethionine (SeMet)-labeled AsCpf1 protein was expressed in *E. coli* B834 (DE3) (Novagen) and purified using a protocol similar to that used for the native protein. The crRNA was purchased from Gene Design. The target and non-target DNA strands were purchased from Sigma-Aldrich. The purified AsCpf1 protein was mixed with the crRNA, the target DNA strand, and the non-target DNA strand (molar ratio, 1:1.5:2.3:3.4), and then the reconstituted AsCpf1-crRNA-target DNA complex was purified by gel filtration chromatography on a Superdex 200 Increase column (GE Healthcare), in buffer consisting of 10 mM Tris-HCl (pH 8.0), 150 mM NaCl, and 1 mM DTT.

Crystallography

The purified AsCpf1-crRNA-target DNA complex was crystallized at 20°C by the hanging-drop vapor diffusion method. The crystallization drops were formed by mixing 1 μl of complex solution ($A_{280\ nm}$ = 10) and 1 μl of reservoir solution (8%–10% PEG 3350, 100 mM sodium acetate [pH 4.5], and 10%–15% 1,6-hexanediol) and then were incubated against 0.5 ml of reservoir solution. The SeMet-labeled complex was crystallized by mixing 1 μl of complex solution ($A_{280\ nm}$ = 10) and 1 μl of reservoir solution (27%–30% PEG 400, 100 mM sodium acetate [pH 4.0], and 200 mM lithium sulfate). The native crystals were cryoprotected in a solution consisting of 11% PEG 3350, 100 mM sodium acetate [pH 4.5], 15% 1,6-hexanediol, and 30% ethylene glycol. The Se-Met-labeled crystals were cryoprotected in a solution consisting of 35% PEG 400, 100 mM sodium acetate (pH 4.0), 200 mM lithium sulfate, and 150 mM NaCl. X-ray diffraction data were collected at 100 K on the beamlines BL41XU at SPring-8 and PXI X06SA at the Swiss Light Source. The X-ray diffraction data were processed using DIALS (Waterman et al., 2013) and AIMLESS (Evans and Murshudov, 2013). The structure was determined by the Se-SAD method, using PHENIX AutoSol (Adams et al., 2010). The structure model was automatically built using Buccaneer (Cowtan, 2006), followed by manual model building using COOT (Emsley and Cowtan, 2004) and structural refinement using PHENIX (Adams et al., 2010).

Generation of the AsCpf1 Mutants

The human codon-optimized AsCpf1 mutants were cloned using the Golden Gate strategy (Engler et al., 2009). Briefly, wild-type AsCpf1 (pY010) was used as the template to amplify two PCR fragments, using primers containing the *Bsm*BI restriction sites. *Bsm*BI digestion results in distinct 5′ overhangs that either are compatible with the *Hind*III or *Xba*I overhangs of the recipient vector or will reconstitute the desired point mutation at the junction of the two AsCpf1 DNA pieces.

Cleavage Activity of AsCpf1 in 293FT Cells

The plasmid expressing the wild-type or mutants of AsCpf1 with N- and C-terminal nuclear localization tags (400 ng) and the plasmid expressing the crRNA (100 ng) were used to transfect human embryonic kidney 293FT cells at 75%–90% confluency in a 24-well plate, using the Lipofectamine 2000 reagent (Life Technologies). Genomic DNA was extracted using QuickExtract DNA Extraction Solution (Epicenter). Indels were analyzed by deep sequencing, as previously described (Hsu et al., 2013).

Synthesis of crRNAs

The crRNA for in vitro cleavage assay was synthesized using the HiScribe T7 High-Yield RNA Synthesis Kit (NEB). DNA oligos corresponding to the reverse complement of the target RNA sequence were synthesized from IDT and annealed to a short T7 priming sequence. T7 transcription was performed for 4 hr and then the RNA was purified using Agencourt RNAClean XP beads (Beckman Coulter).

Preparation of AsCpf1-Containing Cell Lysate

HEK293 cells, growing in six-well plates, were transfected with AsCpf1 expression plasmids (2 μg) using the Lipofectamine 2000 reagent. After 48 hr, the cells were harvested by washing with DPBS (Life Technologies) and then were resuspended in 0.25 ml of lysis buffer (20 mM HEPES [pH 7.5], 100 mM KCl, 5 mM $MgCl_2$, 1 mM DTT, 5% glycerol, 0.1% Triton X-100, and 1× cOmplete Protease Inhibitor Cocktail Tablets [Roche]). After 10-min sonication and 20-min centrifugation (20,000 × *g*), the supernatants were frozen for subsequent use in in vitro cleavage assays.

In Vitro Cleavage Assay

The in vitro cleavage assay was performed with a mammalian cell lysate containing either AsCpf1 or SpCas9 protein, at 37°C for 20 min in cleavage buffer (1× CutSmart buffer [NEB] and 5 mM DTT). The cleavage reaction used 500 ng of synthesized crRNA and 200 ng of target DNA. To prepare the substrate DNA, a 611-bp region containing the target sequence with the 5′-TTTA-3′ PAM was amplified by PCR, using the pUC19 vector as a template. To generate fluorescent-labeled substrates, PCR primers were labeled by the 5′ EndTag Nucleic Acid Labeling System (Vector Laboratories); the forward and reverse primers were labeled to generate the labeled non-target and target strands, respectively. Reactions were processed with a Zymoclean Gel DNA Recovery Kit (Zymo Research) and were run on a 10% polyacrylamide TBE-Urea gel. The gel was visualized using an Odyssey CLx Imaging System (Li-Cor). For the RuvC domain mutants, the processed reactions were run on TBE 6% polyacrylamide or TBE-Urea 6% polyacrylamide gels (Life Technologies), and the gels were then stained with SYBR Gold (Invitrogen).

ACCESSION NUMBERS

The accession number for the atomic coordinates of the AsCpf1-crRNA-target DNA complex reported in this paper has been uploaded to the Protein Data Bank: 5B43.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, and one table and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2016.04.003.

AUTHOR CONTRIBUTIONS

T.Y. crystallized the complex. T.Y. and H.N. performed the structural analysis. B.Z., I.M.S., Y.L., and I.F. performed the functional analysis. T.N. and R.I. assisted with the structural determination. H.H., K.S.M., and E.V.K. analyzed the data. T.Y., H.N., H.H., E.V.K., F.Z., and O.N. wrote the manuscript with help from all authors. H.N., F.Z., and O.N. directed and supervised all of the research.

ACKNOWLEDGMENTS

We thank Arisa Kurabayashi for assistance with vector construction. We thank the beamline scientists at PXI X06SA at the Swiss Light Source and BL41XU at SPring-8 for assistance with data collection. H.N. is supported by JST, PRESTO, JSPS KAKENHI (26291010 and 15H01463), and the Platform for Drug Discovery, Informatics, and Structural Life Science from the Ministry of Education, Culture, Sports, Science and Technology. K.S.M. and E.V.K. are supported by intramural funds of the US Department of Health and Human Services (to the National Library of Medicine). F.Z. is supported by the NIH through the NIMH (5DP1-MH100706 and 1R01-MH110049); a Waterman Award from the National Science Foundation; the New York Stem Cell, Simons, Paul G. Allen Family, and Vallee Foundations; and B. Metcalfe. F.Z. is a New York Stem Cell Foundation Robertson Investigator. F.Z. is a founder of Editas Medicine and a scientific advisor for Editas Medicine and Horizon Discovery. O.N. is supported by the Basic Science and Platform Technology Program for Innovative Biological Medicine from the Japan Agency for Medical Research and Development, AMED, the Council for Science, and the Platform for Drug Discovery, Informatics, and Structural Life Science from the Ministry of Education, Culture, Sports, Science and Technology. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the NIH.

Received: March 18, 2016
Revised: March 31, 2016
Accepted: March 31, 2016
Published: April 21, 2016

REFERENCES

Adams, P.D., Afonine, P.V., Bunkóczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.W., Kapral, G.J., Grosse-Kunstleve, R.W., et al. (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr. D Biol. Crystallogr. *66*, 213–221.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

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

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

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

Cowtan, K. (2006). The Buccaneer software for automated model building. 1. Tracing protein chains. Acta Crystallogr. D Biol. Crystallogr. *62*, 1002–1011.

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

Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallogr. D Biol. Crystallogr. *60*, 2126–2132.

Engler, C., Gruetzner, R., Kandzia, R., and Marillonnet, S. (2009). Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. PLoS ONE *4*, e5553.

Evans, P.R., and Murshudov, G.N. (2013). How good are my data and what is the resolution? Acta Crystallogr. D Biol. Crystallogr. *69*, 1204–1214.

Fonfara, I., Le Rhun, A., Chylinski, K., Makarova, K.S., Lécrivain, A.L., Bzdrenga, J., Koonin, E.V., and Charpentier, E. (2014). Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. Nucleic Acids Res. *42*, 2577–2590.

Garneau, J.E., Dupuis, M.E., Villion, M., Romero, D.A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadán, A.H., and Moineau, S. (2010). The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature *468*, 67–71.

Gasiunas, G., Barrangou, R., Horvath, P., and Siksnys, V. (2012). Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc. Natl. Acad. Sci. USA *109*, E2579–E2586.

Gilbert, L.A., Horlbeck, M.A., Adamson, B., Villalta, J.E., Chen, Y., Whitehead, E.H., Guimaraes, C., Panning, B., Ploegh, H.L., Bassik, M.C., et al. (2014). Genome-scale CRISPR-mediated control of gene repression and activation. Cell *159*, 647–661.

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

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

Holm, L., and Rosenström, P. (2010). Dali server: conservation mapping in 3D. Nucleic Acids Res. *38*, W545–W549.

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

Jiang, F., Zhou, K., Ma, L., Gressel, S., and Doudna, J.A. (2015). A Cas9-guide RNA complex preorganized for target DNA recognition. Science *348*, 1477–1481.

Jiang, F., Taylor, D.W., Chen, J.S., Kornfeld, J.E., Zhou, K., Thompson, A.J., Nogales, E., and Doudna, J.A. (2016). Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. Science *351*, 867–871.

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

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

Karvelis, T., Gasiunas, G., Young, J., Bigelyte, G., Silanskas, A., Cigan, M., and Siksnys, V. (2015). Rapid characterization of CRISPR-Cas9 protospacer adjacent motif sequence elements. Genome Biol. *16*, 253.

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

Kleinstiver, B.P., Prew, M.S., Tsai, S.Q., Nguyen, N.T., Topkar, V.V., Zheng, Z., and Joung, J.K. (2015a). Broadening the targeting range of Staphylococcus aureus CRISPR-Cas9 by modifying PAM recognition. Nat. Biotechnol. *33*, 1293–1298.

Kleinstiver, B.P., Prew, M.S., Tsai, S.Q., Topkar, V.V., Nguyen, N.T., Zheng, Z., Gonzales, A.P., Li, Z., Peterson, R.T., Yeh, J.R., et al. (2015b). Engineered CRISPR-Cas9 nucleases with altered PAM specificities. Nature *523*, 481–485.

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

Konermann, S., Brigham, M.D., Trevino, A.E., Joung, J., Abudayyeh, O.O., Barcena, C., Hsu, P.D., Habib, N., Gootenberg, J.S., Nishimasu, H., et al. (2015). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature *517*, 583–588.

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

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

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

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

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

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

Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S., and Honig, B. (2009). The role of DNA shape in protein-DNA recognition. Nature *461*, 1248–1253.

Shmakov, S., Abudayyeh, O.O., Makarova, K.S., Wolf, Y.I., Gootenberg, J.S., Semenova, E., Minakhin, L., Joung, J., Konermann, S., Severinov, K., et al. (2015). Discovery and functional characterization of diverse class 2 CRISPR-Cas systems. Mol. Cell *60*, 385–397.

Slaymaker, I.M., Gao, L., Zetsche, B., Scott, D.A., Yan, W.X., and Zhang, F. (2016). Rationally engineered Cas9 nucleases with improved specificity. Science *351*, 84–88.

Söding, J., Biegert, A., and Lupas, A.N. (2005). The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. *33*, W244–W248.

Waterman, D.G., Winter, G., Parkhurst, J.M., Fuentes-Montero, L., Hattne, J., Brewster, A., Sauter, N.K., Evans, G., and Rosenstrom, P. (2013). The DIALS framework for integration software. CCP4 Newsletter, Summer 2013. http://www.ccp4.ac.uk/newsletters/newsletter49/content.html.

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

Wright, A.V., Nuñez, J.K., and Doudna, J.A. (2016). Biology and applications of CRISPR systems: harnessing nature's toolbox for genome engineering. Cell *164*, 29–44.

Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., Essletzbichler, P., Volz, S.E., Joung, J., van der Oost, J., Regev, A., et al. (2015). Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell *163*, 759–771.

Supplemental Figures

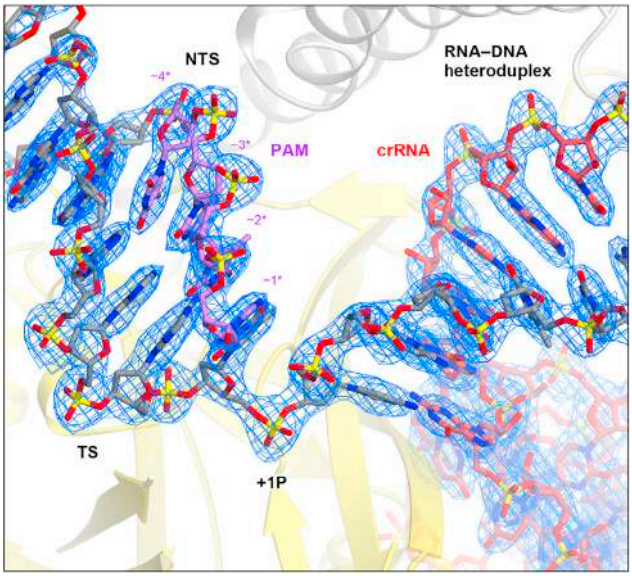


Figure S1. Electron Density Map, Related to Figure 1
The $2mF_O - DF_C$ electron density map (contoured at 2.0 σ) for the bound nucleic acids is shown as a blue mesh. +1P, +1 phosphate.

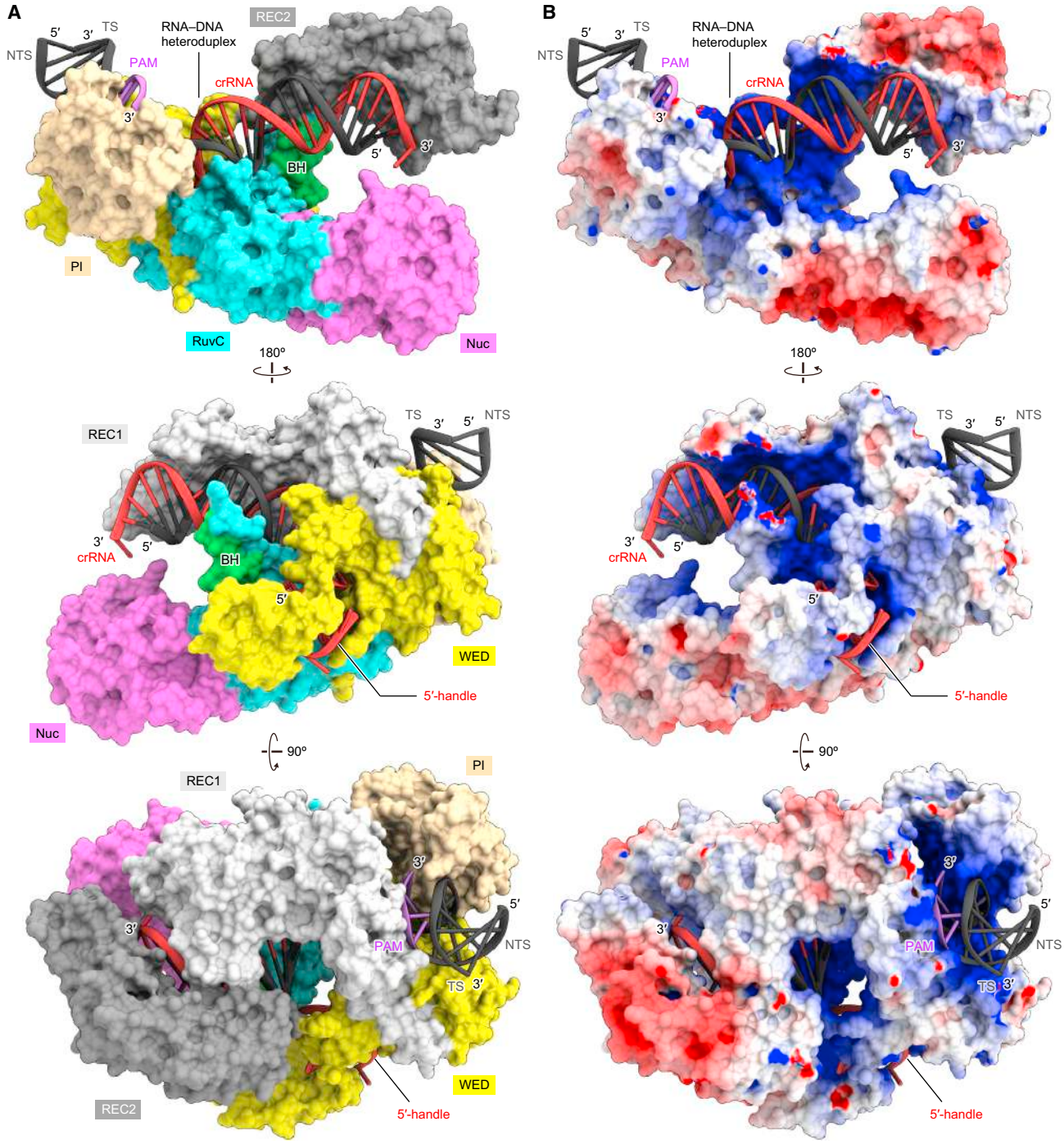


Figure S2. Molecular Surface of AsCpf1, Related to Figure 1

(A and B) Surface representations of the AsCpf1-crRNA-target DNA complex, colored according to domain (A) and electrostatic potential (B). The REC1 and REC2 domains are omitted for clarity in the top and middle panels, respectively. BH, bridge helix.

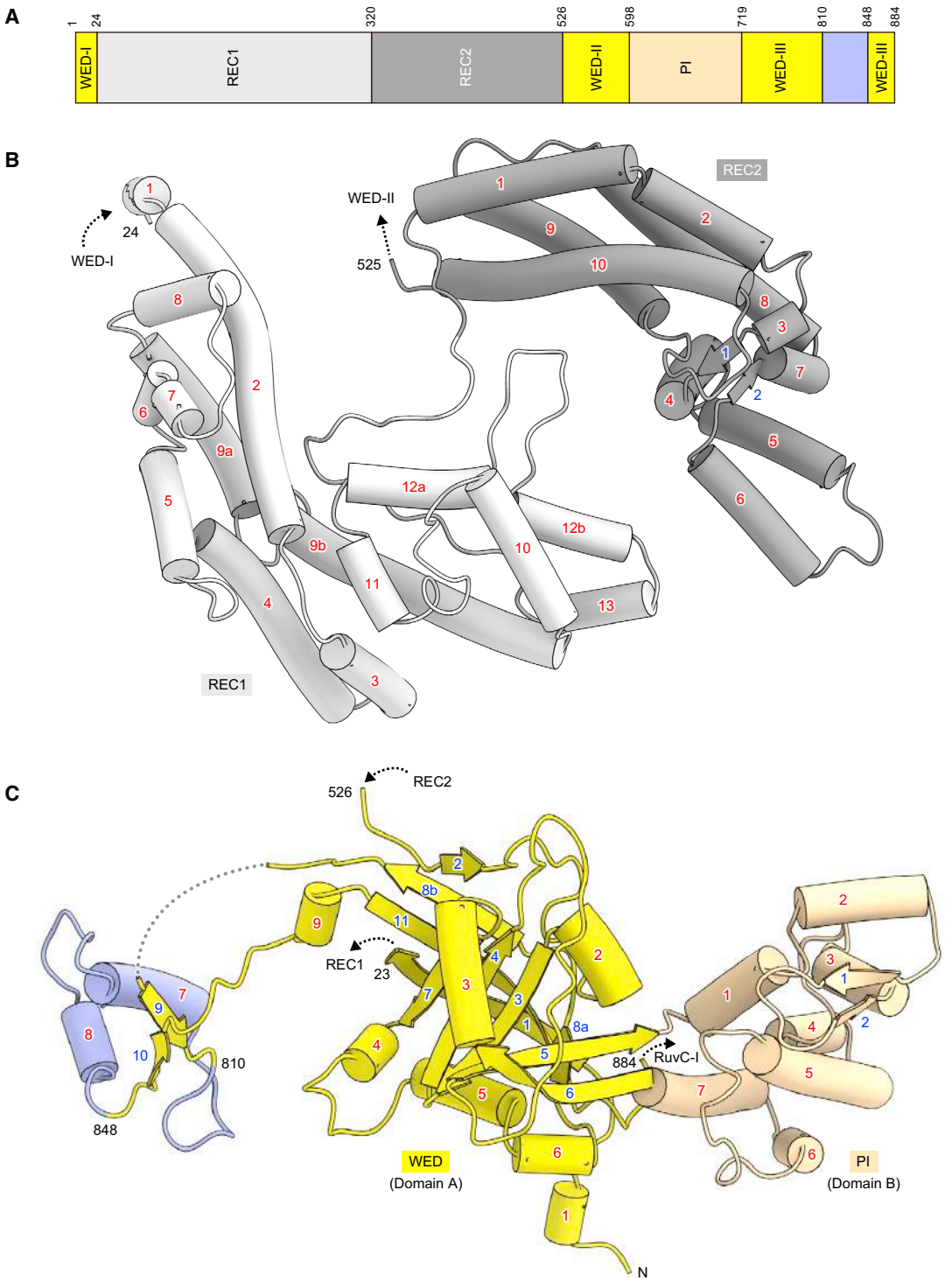


Figure S3. Structure of REC1, REC2, WED, and PI Domains, Related to Figure 1

(A) Domain organization of the REC1, REC2, WED and PI domains of AsCpf1. The less conserved region in the WED domain is colored pale blue.

(B) Structure of the REC1 and REC2 domains.

(C) Structure of the WED and PI domains. The disordered regions are shown as dashed lines.

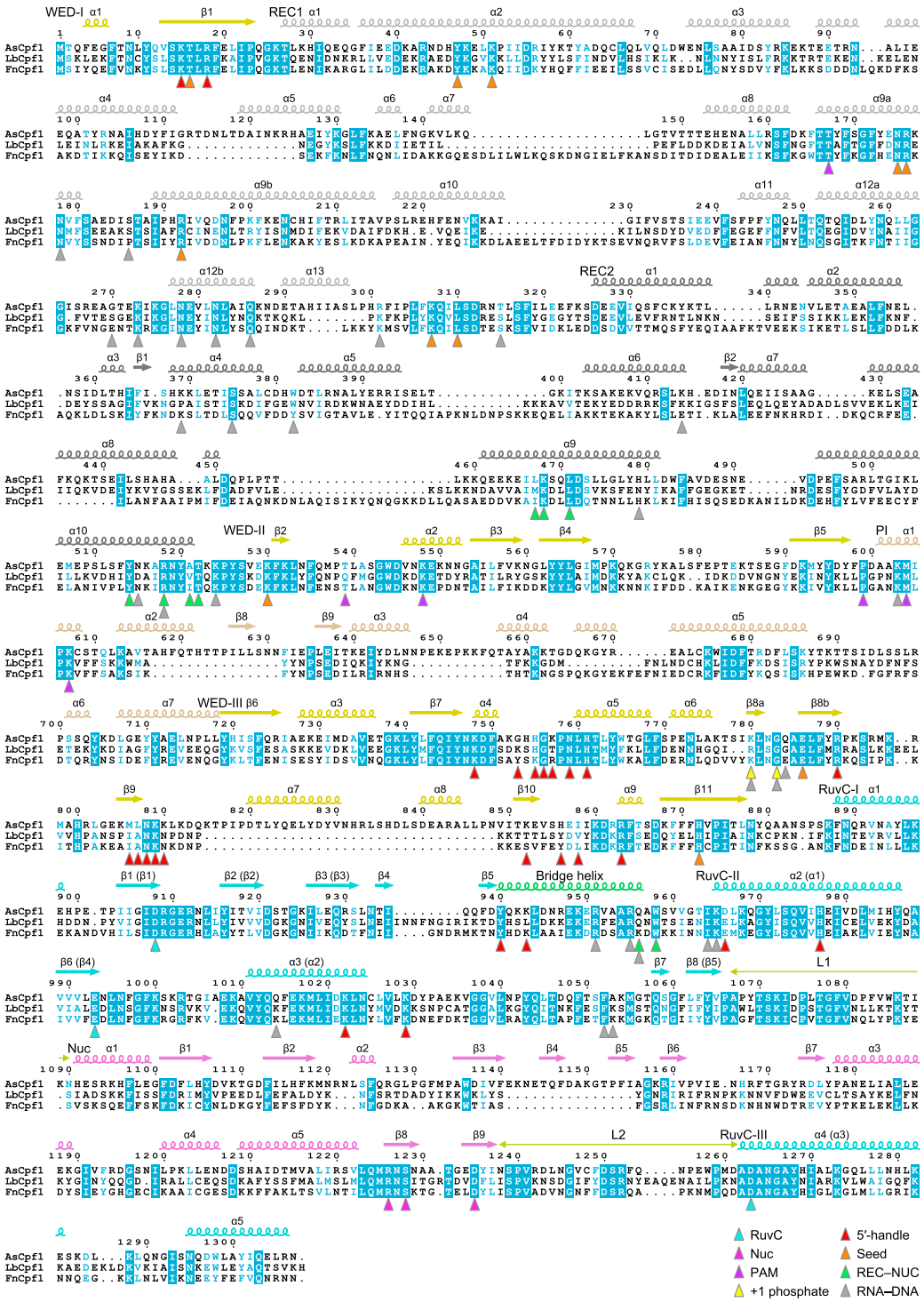


Figure S4. Multiple Sequence Alignment of Cpf1 Proteins, Related to Figure 3

The secondary structures are shown above the sequences, and the key residues are indicated by triangles. As, *Acidaminococcus sp. BV3L6*; Lb, *Lachnospiraceae bacterium ND2006*; Fn, *Francisella novicida U112*. The figure was prepared using Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo) and ESPript (http://espript.ibcp.fr).

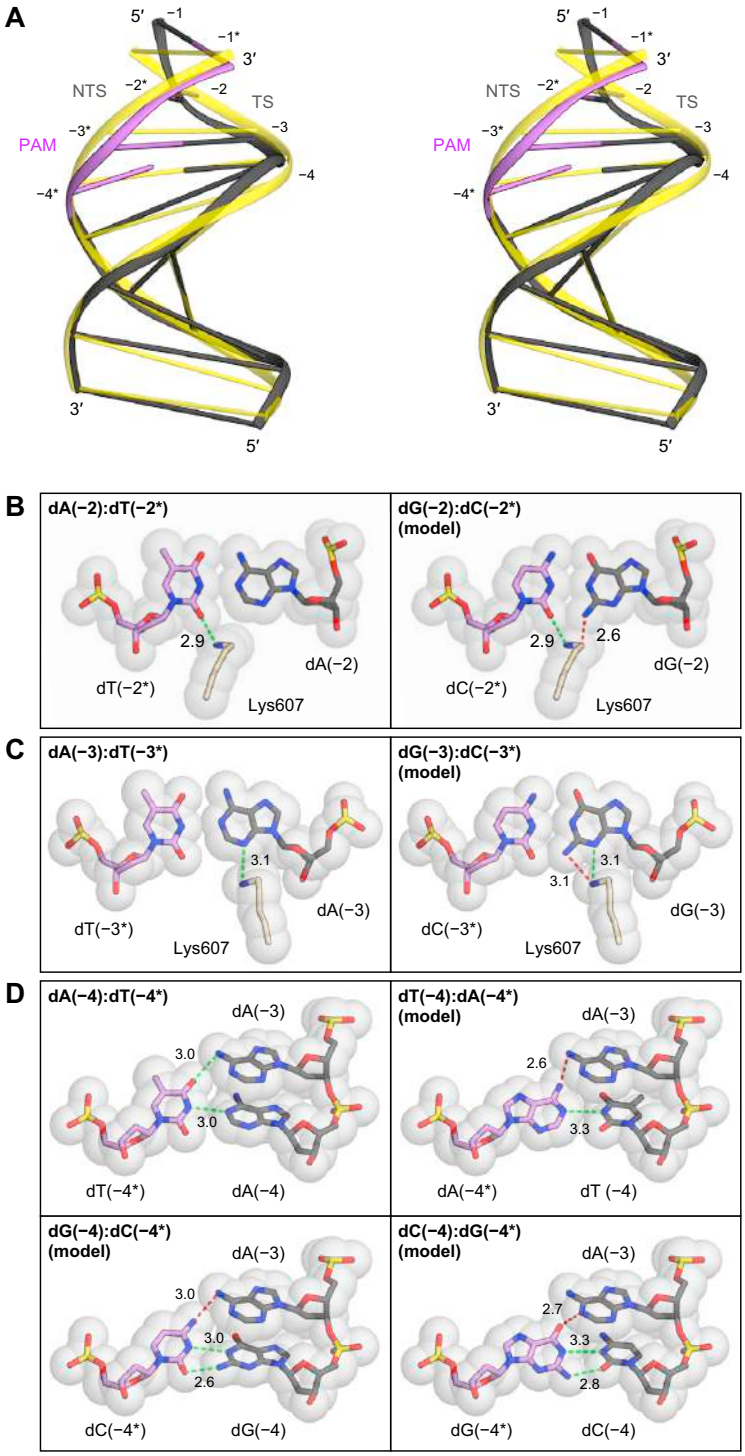


Figure S5. Structure and Recognition of the PAM Duplex, Related to Figure 5

(A) Superimposition of the PAM duplex onto a B-form DNA duplex (stereo view). The 5′-TTTN-3′ PAM is highlighted in light purple, and the B-form DNA duplex is colored yellow.

(B) Specific recognition of the dA(−2):dT(−2*) base pair. The modeled dG(−2):dC(−2*) base pair would form steric clashes with Lys607 in the PI domain.

(C) Specific recognition of the dA(−3):dT(−3*) base pair. The modeled dG(−3):dC(−3*) base pair would form steric clashes with Lys607 in the PI domain.

(D) Specific recognition of the dA(−4):dT(−4*) base pair. The modeled base pairs, dT(−4):dA(−4*), dG(−4):dC(−4*) and dC(−4):dG(−4*), would form steric clashes with dA(−3) in the target DNA strand. In (B) and (C), potential favorable and unfavorable interactions are depicted as green and red dashed lines, respectively.

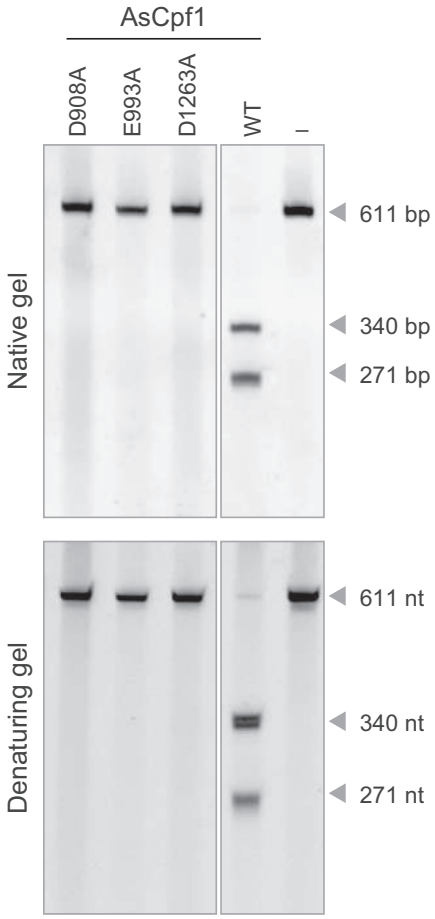


Figure S6. Mutational Analysis of the RuvC Catalytic Residues, Related to Figure 6

The wild-type or mutants of AsCpf1 were incubated with the crRNA and double-stranded DNA target, and the reaction products were resolved on native TBE and denaturing TBE-Urea polyacrylamide gels. The gels were stained with SYBR Gold (Invitrogen). The mutations of the RuvC catalytic residues (D908A, E993A and D1263A) abolished the cleavage of both the target and non-target DNA strands.

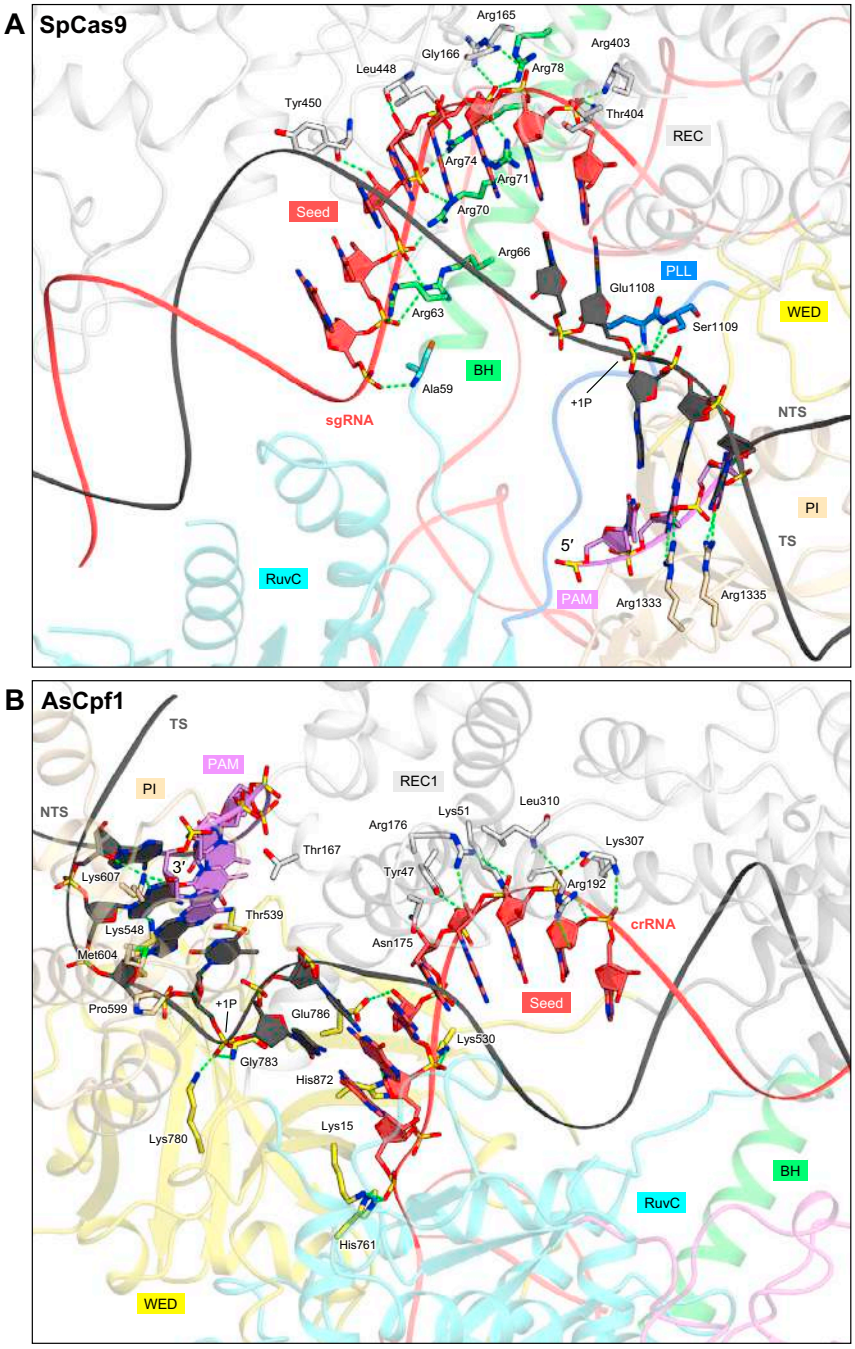


Figure S7. RNA-Guided DNA-Targeting Mechanisms of Class 2 CRISPR Effectors, Related to Figure 7

(A and B) SpCas9 (PDB: 4UN3) (A) and AsCpf1 (B). Key protein residues and nucleotides in the seed region and the PAM duplex are shown as stick models. Hydrogen bonds are shown as dashed lines. PLL, phosphate lock loop.

Cell, Volume *165*

Supplemental Information

Crystal Structure of Cpf1 in Complex with Guide RNA and Target DNA

Takashi Yamano, Hiroshi Nishimasu, Bernd Zetsche, Hisato Hirano, Ian M. Slaymaker, Yinqing Li, Iana Fedorova, Takanori Nakane, Kira S. Makarova, Eugene V. Koonin, Ryuichiro Ishitani, Feng Zhang, and Osamu Nureki

Supplemental Experimental Procedures

Substrate DNA for *in vitro* cleavage

Target (complement) / PAM

CGGGGCTGGCTTAACTATGCGGCATCAGAGCAGATTGTACTGAGAGTGCACCATATGCGGTGTG
AAATACCGCACAGATGCGTAAGGAGAAAATACCGCATCAGGCGCCATTCGCCATTCAGGCTGCG
CAACTGTTGGGAAGGGCGATCGGTGCGGGCCTCTTCGCTATTACGCCAGCTGGCGAAAGGGGGA
TGTGCTGCAAGGCGATTAAGTTGGGTAACGCCAGGGTTTTCCCAGTCACGACGTTGTAAAACGA
CGGCCAGTGAATTCGAGCTCGGTACCCGGGGATCCTTTCGAGCTCGGTACCCGGGGATCCTTTA
GAGAAGTCATTTAATAAGGCCACTGTTAAAAAGCTTGGCGTAATCATGGTCATAGCAGCTTGGC
GTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCACACAACATA
CGAGCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTCACATTAATTG
CGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGG
CCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGC

crRNA oligo for *in vitro* transcription (reverse complement)

Guide / 5′ handle / T7 promoter

GTGGCCTTATTAAATGACTTCTCATCTACAAGAGTAGAAATTACCCTATAGTGAGTCGTATTAA
TTTC

NGS primers

DNMT1-1_For	GCTTAGAGCAGGCGTGCTGCA
DNMT1-1_Rev	CTCAAACGGTCCCCAGAGGGTT
DNMT1-2_For	TGAACGTTCCCTTAGCACTCTGCC
DNMT1-2_Rev	CCTTAGCAGCTTCCTCCTCC

Table S1. Data Collection and Refinement Statistics, Related to Figure 1.		
	Native	SeMet
Data collection		
Beamline	SLS PXI X06SA	SPring-8 BL41XU
Wavelength (Å)	1.000	0.979
Space group	$P2_12_12_1$	$P4_12_12$
Cell dimensions		
a, *b*, *c* (Å)	81.5, 136.7, 196.9	191.5, 191.5, 124.2
α, *β*, *γ* (°)	90, 90, 90	90, 90, 90
Resolution (Å)*	196–2.80 (2.88–2.80)	191–2.8 (2.88–2.80)
R_{merge}	0.089 (0.32)	0.155 (2.08)
R_{pim}	0.048 (0.18)	0.030 (0.42)
$I/\sigma I$	8.6 (2.2)	22.3 (2.8)
Completeness (%)	99.0 (99.3)	100 (100)
Multiplicity	4.4 (4.5)	51.4 (48.6)
CC(1/2)	0.99 (0.73)	1.00 (0.91)
Refinement		
Resolution (Å)	56.2–2.8	
No. reflections	54,241	
R_{work} / R_{free}	0.216 / 0.255	
No. atoms		
Protein	10,168	
Nucleic acid	1,657	
Ion	1	
Solvent	37	
B-factors (Å^2)		
Protein	71.3	
Nucleic acid	70.8	
Ion	57.4	
Solvent	51.9	
R.m.s. deviations		
Bond lengths (Å)	0.002	
Bond angles (°)	0.493	
Ramachandran plot (%)		
Favored region	97.0	
Allowed region	3.0	
Outlier region	0.0	

*Values in parentheses are for the highest resolution shell.

Chapter IV

Multiplex gene editing by CRISPR-Cpf1 using a single crRNA array

Introduction

This chapter continues the studies of Cas12a enzymes. We show that in contrast to Cas9, Cas12a processes its long pre-crRNA into short mature crRNAs without the help from any other enzymes. Thus, only the effector Cas12a protein is responsible for crRNA biogenesis. We used this Cas12a property for multiplex gene editing in human cells using a single CRIPSR array. Although the multiplex editing is possible with Cas9 too, it requires several promoters to drive the expression of each guide RNA. Cas12a CRISPR array transcript, in contrast, can be cleaved by the Cas12a into mature crRNAs and, hence, requires only one promoter sequence for expression of the whole set of crRNA targeting different genome sites.

We used a dual AAV system to target three different genes in mice brain using AsCas12a multiplex gene editing system and demonstrated that this approach allows to efficiently modify several genes simultaneously in an adult organism.

Contribution

This work also was done during my internship at the Broad Institute, at Feng Zhang laboratory. I was working with Bernd Zetsche and assisted with biochemical assays. In particular, I performed the gels showing that AsCas12a and LbCas12a cleave pre-crRNA *in vitro* and prepared the products of this reaction for RNA sequencing (Figures 1b, c).

To show that multiplex editing system modifies several genes simultaneously in a single cell, I performed FACS of human cells transfected by a plasmid carrying the Cas12a system and prepared samples for HTS, although the bioinformatical analysis of sequencing results was performed by other authors (Figure 2d, e).

The manuscript was written by the first authors of the paper and the corresponding authors.

Published in final edited form as:
Nat Biotechnol. 2017 January ; 35(1): 31–34. doi:10.1038/nbt.3737.

Multiplex gene editing by CRISPR-Cpf1 through autonomous processing of a single crRNA array

Bernd Zetsche[#1,2,3,4,5], **Matthias Heidenreich**[#1,2,3,4], **Prarthana Mohanraju**[#6], **Iana Fedorova**[1,2,3,4,9,10], **Jeroen Kneppers**[1,6], **Ellen M. DeGennaro**[1,7], **Nerges Winblad**[1,2,3,4], **Sourav R. Choudhury**[1,2,3,4], **Omar O. Abudayyeh**[1,2,3,4,7], **Jonathan S. Gootenberg**[1,2,3,4,8], **Wen Y. Wu**[6], **David A. Scott**[1,2], **Konstantin Severinov**[9,11,12], **John van der Oost**[6,†], and **Feng Zhang**[1,2,3,4,†]

[1] Broad Institute of MIT and Harvard, Cambridge, MA 02142 [2] McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139 [3] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139 [4] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 [5] Department of Developmental Pathology, Institute of Pathology, Bonn Medical School, Sigmund Freud Street 25, 53127 Bonn, Germany [6] Laboratory of Microbiology, Department of Agrotechnology and Food Sciences, Wageningen University, Stippeneng 4, 6708 WE Wageningen, Netherlands [7] Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, MA 02139 [8] Department of Systems Biology, Harvard Medical School, Boston, MA 02115 [9] Skolkovo Institute of Science and Technology, Skolkovo, 143025, Russia. [10] Peter the Great St.Petersburg Polytechinc University, St. Petersburg, 195251, Russia [11] Waksman Institute for Microbiology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA [12] Institute of Molecular Genetics, Russian Academy of Sciences, Moscow, 123182, Russia

[#] These authors contributed equally to this work.

Although multiplex gene editing is possible with Cas9, it requires relatively large constructs or simultaneous delivery of multiple plasmids [7-11], both of which are problematic for multiplex screens or *in vivo* applications. In contrast, Cpf1 only requires one Pol III promoter to drive several small crRNAs (39nt per crRNA). We confirmed that Cpf1 alone is sufficient for array processing[1,2] using an artificial CRISPR pre-crRNA array consisting of four spacers separated by direct repeats (DRs) from the CRISPR locus of *Francisella*

[†] To whom correspondence should be addressed: F.Z. (zhang@broadinstitute.org) or J.v.d.O. (john.vanderoost@wur.nl). We recently harnessed Cpf1, the effector nuclease of a novel type V-A CRISPR system, for genome editing in mammalian cells. Cpf1 does not require additional factors for CRISPR RNA (crRNA) processing, providing a simple route to multiplex targeting. Here, we show that two Cpf1 orthologs are capable of multiplex gene editing in mammalian cells as well as in the mouse brain by using a customized single CRISPR array.

Author Contributions

B.Z., M.H., J.v.d.O., and F.Z. conceived this study and designed the experiments. B.Z., M.H., P.M., Y.F., J.K., E.M.D., N.W., S.C., O.O.A., and J.S.G. conducted the experiments. K.S., J.v.d.O., F.Z. supervised this project. B.Z., M.H., J.v.d.O., and F.Z. wrote the manuscript with input from all authors.

novicida (FnCpf1) and two Cpf1 orthologs with activity in mammalian cells, *Acidaminococcus* Cpf1 (AsCpf1) and *Lachnospiraceae* Cpf1 (LbCpf1) (**Figure 1b and Supplementary figure 1**). Small RNAseq showed that AsCpf1 cleavage products correlate to fragments resulting from cuts at the 5' end of DR hairpins, identical to the cleavage pattern we observed in *E.coli* heterologously expressing FnCpf1 CRISPR systems[1] (**Figure 1c**).

We further validated these results by generating AsCpf1 mutants that are unable to process arrays. Guided by the crystal structure of AsCpf1[3], we mutated five conserved amino acid residues likely to disrupt array processing (H800A, K809A, K860A, F864A, and R790A) [3]. All mutations interfered with pre-crRNA processing but not DNA cleavage activity *in vitro* (**Figure 1d and Supplementary figure 2a, b**), an effect that was also observed for FnCpf1[2]. AsCpf1 recognizes specific nucleotides at the 5' flank of the DR stem loop. Substitution of these nucleotides weakens or abolishes RNA cleavage (**Supplementary figure 3a**). Dosage tests with the five AsCpf1 mutants revealed that mutants K809A, K860A, F864A, and R790A show pre-crRNA processing when used at high concentration (**Supplementary figure 3b**) or for extended incubation times (**Supplementary figure 3c**), but H800A was inactive regardless of dose and time.

We next tested if this mutant retains DNase activity in human embryonic kidney (HEK) 293T cells using three guides. Insertion/deletion (indel) frequency at the *DNMT1* and *GRIN2b* loci were identical between wild-type and H800A AsCpf1, whereas indel frequencies at the *VEGFA* locus were higher in cells transfected with wild-type AsCpf1, demonstrating that the RNA and DNA cleavage activity can be separated in mammalian cells (**Figure 1e**).

Cpf1 mediated RNA cleavage needs to be considered when designing lenti-virus vectors for simultaneous expression of nuclease and guide (**Figure 2f**). Lenti virions carry a (+) strand RNA copy of the sequence flanked by long terminal repeats (LTR), allowing Cpf1 to bind and cleave at DR sequences. Hence, reversing the orientation of the DR is expected to result in (+) strand lenti RNAs not susceptible to Cpf1 mediated cleavage. We designed a lenti vector encoding AsCpf1 and a crRNA expression cassette. We transduced HEK293T cells with a MOI (multiplicity of infection) of <0.3 and analyzed indel frequencies in puromycin selected cells 10 days post infection. Using guides encoded on a reversed expression cassette targeting *DNMT1*, *VEGFA,* or *GRIN2b* resulted in robust indel formation for each targeted gene (**Figure 2g**).

We leveraged the simplicity of Cpf1 crRNA maturation to achieve multiplex genome editing in HEK293T cells using customized CRISPR arrays. We chose four guides targeting different genes (*DNMT1*, *EMX1*, *VEGFA*, and *GRIN2b*) and constructed three arrays with variant DR and guide lengths for expression of pre-crRNAs (**Figure 2a**). Indel events were detected at each targeted locus in cells transfected with array-1 or -2. However, the crRNA targeting *EMX1* resulted in indel frequencies of <2% when expressed from array-3. Overall, array-1 performed best, with all guides showing indel levels comparable to those mediated by single crRNAs (**Figure 2b**). Furthermore, small RNAseq confirmed that autonomous, Cpf1-mediated pre-crRNA processing occurs in mammalian cells (**Figure 2c**). Using arrays

with guides in different orders resulted in similar indel frequencies, suggesting that positioning within an array is not crucial for activity (**Supplementary figure 4a, b**).

To confirm that multiplex editing occurs within single cells, we generated AsCpf1-P2A-GFP constructs to enable fluorescence-activated cell sorting (FACS) of transduced single cells (**Figure 2d**) and clonal expansion. We used next generation deep sequencing (NGS) to compare edited loci within clonal colonies derived from cells transfected with either pooled single guides or array-1. Focusing on targeted genes edited at every locus (indels ≥95%) shows that multiplex editing occurs more frequently in colonies transfected with array-1 (6.4% all targets, 12.8% three targets, 48.7% two targets) than in pooled transfection (2.4% all targets, 3.6% three targets, 11.9% two targets).

We next tested multiplex genome editing in neurons using AsCpf1. We designed a gene-delivery system based on adeno-associated viral vectors (AAVs) for expression of AsCpf1. We generated a dual vector system in which AsCpf1 and the CRISPR-Cpf1 array were cloned separately (**Figure 2f**). We constructed a U6 promoter-driven Cpf1 array targeting the neuronal genes *Mecp2*, *Nlgn3*, and *Drd1*. This plasmid also included an green fluorescent protein (GFP) fused to the KASH nuclear transmembrane domain [4] to enable FACS of targeted cell nuclei [5].

We first transduced mouse primary cortical neurons *in vitro* and observed robust expression of AsCpf1 and GFP-KASH one week after viral delivery. SURVEYOR nuclease assay on purified neuronal DNA confirmed indel formations in all three targeted genes (**Supplementary figure 5**). Next, we tested whether AsCpf1 can be expressed in the brains of living mice for multiplex genome editing *in vivo*. We stereotactically injected our dual vector system in a 1:1 ratio into the hippocampal dentate gyrus (DG) of adult male mice. Four weeks after viral delivery we observed robust expression of AsCpf1 and GFP-KASH in the DG (**Figure 2g, h**). Consistent with previous studies [5, 6], we observed ~75% co-transduction efficiency of the dual viral vectors (**Supplementary figure 2c**). We isolated targeted DG cell nuclei by FACS (**Supplementary figure 4**) and quantified indel formation using NGS. We found indels in all three targeted loci with ~23%, ~38%, and ~51% indel formation in *Mecp2*, *Nlgn3*, and *Drd1*, respectively (**Supplementary figure 4d, e**). We quantified the effectiveness of biallelic disruption of the autosomal gene *Drd1* and found ~47% of all sorted nuclei (i.e. ~87% of all *Drd1*-edited cells) harbored biallelic modifications (**Figure 2i**). Next, we quantified the multiplex targeting efficiency in single neuronal nuclei. Our results show that ~17% of all transduced neurons were modified in all three targeted loci (**Figure 2j**). Taken together, our results demonstrate the effectiveness of AAV-mediated delivery of AsCpf1 into the mammalian brain and simultaneous multi-gene targeting *in vivo* using a single array transcript.

Taken together, these data highlight the utility of Cpf1 array processing in designing simplified systems for *in vivo* multiplex gene editing. Although multiplex gene editing is possible with Cas9, it requires relatively large constructs or simultaneous delivery of multiple plasmids [7-11], both of which are problematic for multiplex screens or *in vivo* applications. In contrast, Cpf1 only requires one Pol III promoter to drive several small crRNAs (39nt per crRNA). Hence, this system has the potential to simplify guide RNA

delivery for many genome editing applications where targeting of multiple genes is desirable.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We would like to thank F. A. Ran for helpful discussions and overall support, and Bas Cartigny and Jara van den Bogaerde for technical assistance, and the entire Zhang laboratory for support and advice. M.H was supported by the Human Frontiers Scientific Program. O.A.A. is supported by a Paul and Daisy Soros Fellowship and a Friends of the McGovern Institute Fellowship. J.S.G. is supported by a D.O.E. Computational Science Graduate Fellowship. E.D.G is supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB), of the National Institutes of Health (5T32EB1680). K.S. is supported by an NIH grant GM10407, Russian Science Foundation grant 14-14-00988, and Skoltech. J.v.d.O. is supported by Netherlands Organization for Scientific Research (NWO) through a TOP grant (714.015.001). F.Z. is supported by the NIH through NIMH (5DP1-MH100706 and 1R01-MH110049), the New York Stem Cell, Poitras, Simons, Paul G. Allen Family, and Vallee Foundations; and David R. Cheng, Tom Harriman, and B. Metcalfe. F.Z. is a New York Stem Cell Foundation Robertson Investigator. The authors plan to make the reagents widely available to the academic community through Addgene and to provide software tools via the Zhang lab website (www.genome-engineering.org).

Reference

1. Zetsche B, et al. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell. 2015; 163:759–771. [PubMed: 26422227]
2. Fonfara I, Richter H, Bratovič M, Rhun A, Charpentier E. The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. Nature. 2016
3. Yamano T, et al. Crystal Structure of Cpf1 in Complex with Guide RNA and Target DNA. Cell. 2016
4. Ostlund C, et al. Dynamics and molecular interactions of linker of nucleoskeleton and cytoskeleton (LINC) complex proteins. Journal of cell science. 2009; 122:4099–4108. [PubMed: 19843581]
5. Swiech L, et al. In vivo interrogation of gene function in the mammalian brain using CRISPR-Cas9. Nat Biotechnol. 2015; 33:102–106. [PubMed: 25326897]
6. Konermann S, et al. Optical control of mammalian endogenous transcription and epigenetic states. Nature. 2013; 500:472–476. [PubMed: 23877069]
7. Kabadi AM, Ousterout DG, Hilton IB, Gersbach CA. Multiplex CRISPR/Cas9-based genome engineering from a single lentiviral vector. Nucleic Acids Research. 2014; 42
8. Nissim L, Perli SD, Fridkin A, Perez-Pinera P, Lu TK. Multiplexed and Programmable Regulation of Gene Networks with an Integrated RNA and CRISPR/Cas Toolkit in Human Cells. Molecular Cell. 2014; 54:698–710. [PubMed: 24837679]
9. Sakuma T, Nishikawa A, Kume S, Chayama K, Yamamoto T. Multiplex genome engineering in human cells using all-in-one CRISPR/Cas9 vector system. Scientific Reports. 2014; 4:5400. [PubMed: 24954249]
10. Tsai SQ, et al. Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. Nature biotechnology. 2014; 32:569–576.
11. Xie K, Minkenberg B, Yang Y. Boosting CRISPR/Cas9 multiplex editing capability with the endogenous tRNA-processing system. Proceedings of the National Academy of Sciences. 2015; 112:3570–3575.

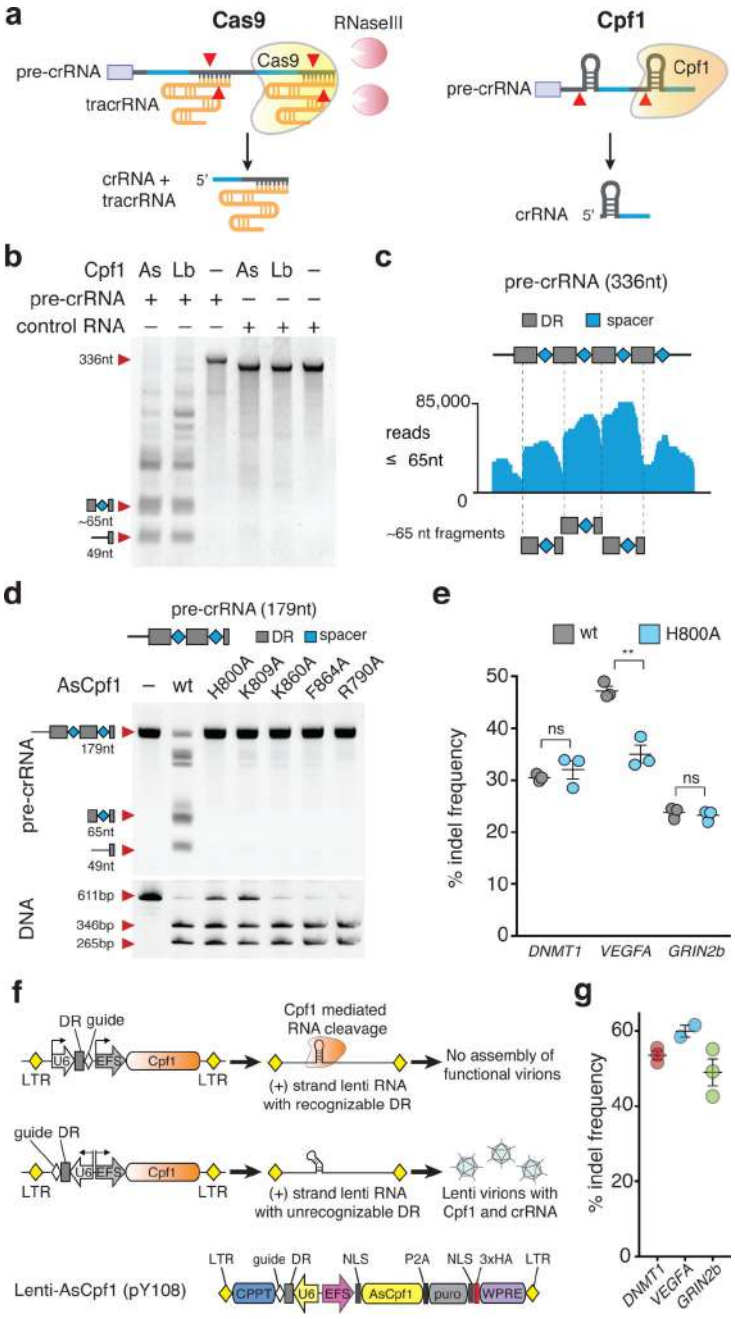


Figure 1. Cpf1 mediated processing of pre-crRNA is independent of DNA cleavage

(a) Schematic of pre-crRNA processing for Cas9 and Cpf1. Cleavage sites indicated with red triangles. **(b)** *In vitro* processing of FnCpf1 pre-crRNA transcript (80 nM) with purified AsCpf1 or LbCpf1 protein (~320 nM), cropped gel image. **(c)** RNAseq analysis of FnCpf1 pre-crRNA cleavage products, as shown in (b). A high fraction of sequence reads smaller than 65nt are cleavage products of spacers flanked by DR sequences, cropped gel images. **(d)** Pre-crRNA (top) and DNA cleavage (bottom) mediated by AsCpf1 point mutants. H800A, K809A, K860A, F864A, and R790A fail to process precrRNA but retain DNA

cleavage activity *in vitro*. 330 nM pre-crRNA was cleaved with 500 nM Cpf1 in 15 min and 25 nM DNA was cleaved with 165 nM Cpf1 in 30 min. **(e)** Indel frequencies mediated by AsCpf1H800A are comparable to wt AsCpf1, bars are mean of 3 technical replicates from one experiment, error bars are SEM. (Student *t*-test; ns = not significant; ** = p-value 0.003). **(f)** Schematic of lenti-Cpf1 construct with the U6::DR cassette in different orientations (top and middle), (+)-strand lenti RNA with recognizable DRs are susceptible to Cpf1 mediated degradation, preventing functional virion formation. Schematic of lenti-AsCpf1 (pY108) construct (bottom). **(g)** Indel frequencies analyzed by SURVEYOR nuclease assay after puromycin selection 10 days after transduction with lenti-AsCpf1 in HEK cells, bars are mean of 2 or 3 individual infections, error bars are SEM. U6, Pol III promoter; CMV, cytomegalovirus promoter; NLS, nuclear localization signal; HA, hemagglutinin tag; DR, direct repeat sequence; P2A, porcine teschovirus-1 2A self-cleaving peptide; LTR, long terminal repeat; WPRE, Woodchuck Hepatitis virus posttranscriptional regulatory element.

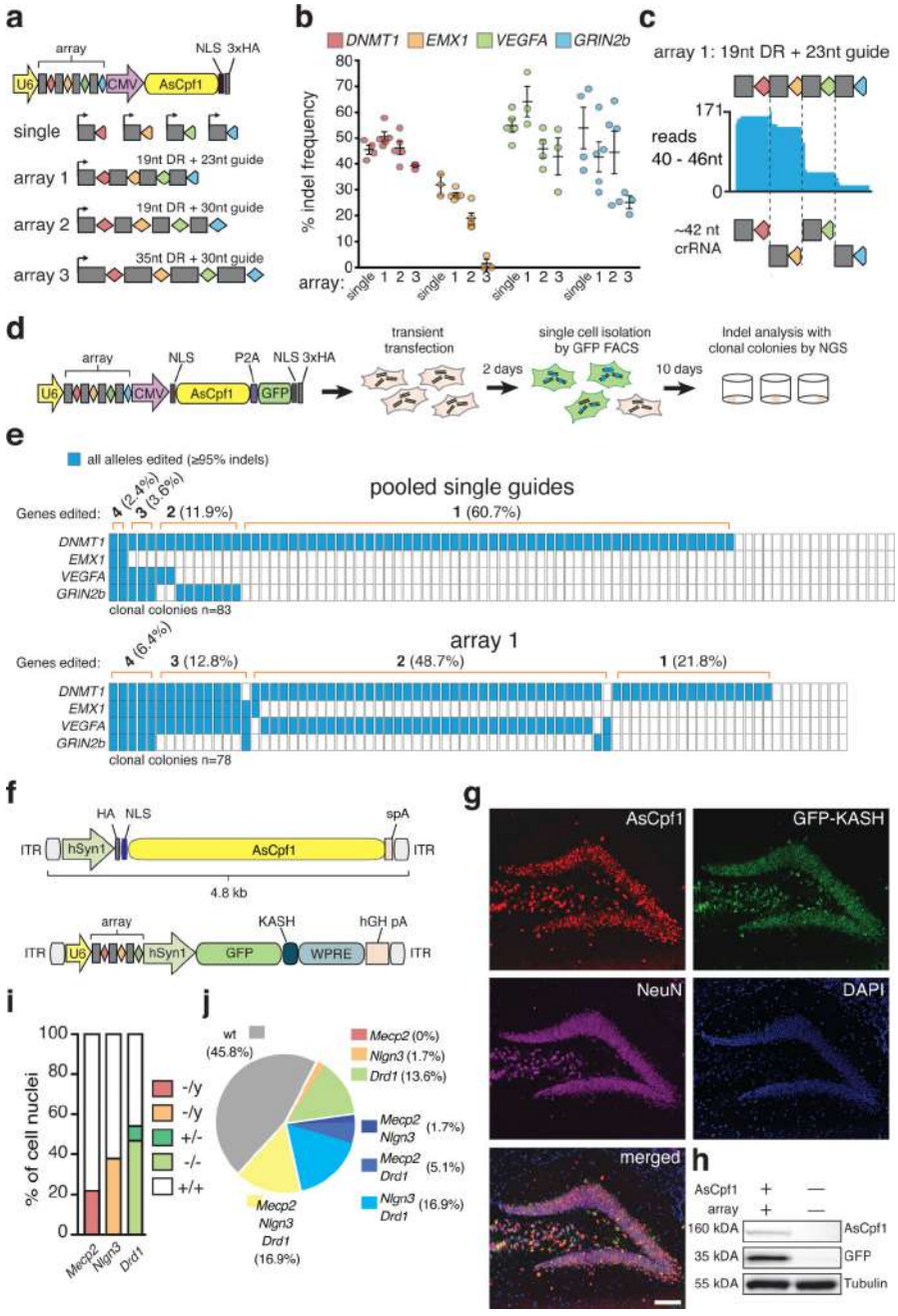


Figure 2. Cpf1-mediated multiplex gene editing in mammalian cells and mouse brain

(a) Schematic of multiplex gene editing with AsCpf1, using a single plasmid approach. **(b)** Genome editing at four different genomic loci mediated by AsCpf1 with different versions of artificial CRISPR arrays (array-1, crRNAs in their mature form (19nt DR with 23nt guide); array-2, crRNAs are in an intermediate form (19nt DR with 30nt guide); array-3 crRNAs are in their unprocessed form (35nt DR with 30nt guides)). Indels were analyzed by SURVEYOR nuclease assay 3 days post transfection; bars are mean of two individual experiments with 3 to 5 technical replicates, error bars are SEM. **(c)** Small RNAseq reads

from HEK cells transfected with AsCpf1 and array-1 show fragments corresponding to mature crRNA for each of the four guides. **(d)** Schematic for analysis of indel events in clonal colonies 48 hours after transient transfection. **(e)** Quantification of indel events measured by NGS in clonal colonies from HEK cells transiently transfected with pooled single guide plasmids or plasmid carrying array-1. Colonies were expanded for 10 days after sorting. Each column represents one clonal colony; blue rectangles indicate target genes with all alleles edited. **(f)** Schematic of AAV vector design for multiplex gene editing. Bottom: grey rectangles, direct repeat; diamonds, spacer (red: *Mecp2*, orange: *Nlgn3*, green: *Drd1*). **(g)** Immunostaining of dorsal DG 4 weeks after stereotactic AAV injection (Representative image of $n = 4$ mice). Brain sections were co-stained with anti-HA (red), anti-GFP (green) and anti-NeuN (magenta) antibodies. Nuclei were labeled with DAPI (blue). Scale bar: 100 um. **(h)** Western blot analysis of DG expressing HA-AsCpf1 and GFP-KASH (Representative blot from $n = 4$ mice). **(i)** Fraction of mono- and biallelic modifications of autosomal gene *Drd1* is shown (*Mecp2* and *Nlgn3*: x-chromosomal). **(j)** Analysis of multiplexing efficiency in individual cells. ITR, inverted terminal repeat; spA, synthetic polyadenylation signal; hSyn1, human synapsin 1 promoter; ANC1, Syne Homology nuclear transmembrane domain; hGH pA, human growth hormone polyadenylation signal;

Supplementary Information

Multiplex gene editing by CRISPR-Cpf1 through autonomous processing of a single crRNA array

Bernd Zetsche[1,2,3,4,5*], Matthias Heidenreich[1,2,3,4*], Prarthana Mohanraju[6*], Iana Fedorova[1,2,3,4,9,10], Jeroen Kneppers[1,6], Ellen M. DeGennaro[1,7], Nerges Winblad[1,2,3,4], Sourav R. Choudhury[1,2,3,4], Omar O. Abudayyeh[1,2,3,4,7], Jonathan S. Gootenberg[1,2,3,4,8], Wen Y. Wu[6], David A. Scott[1,2], Konstantin Severinov[9,11,12], John van der Oost[6†], and Feng Zhang[1,2,3,4†]

[1] Broad Institute of MIT and Harvard, Cambridge, MA 02142

[2] McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139

[3] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139

[4] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139

[5] Department of Developmental Pathology, Institute of Pathology, Bonn Medical School, Sigmund Freud Street 25, 53127 Bonn, Germany

[6] Laboratory of Microbiology, Department of Agrotechnology and Food Sciences, Wageningen University, Stippeneng 4, 6708 WE Wageningen, Netherlands

[7] Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, MA 02139

[8] Department of Systems Biology, Harvard Medical School, Boston, MA 02115

[9] Skolkovo Institute of Science and Technology, Skolkovo, 143025, Russia.

[10] Peter the Great St.Petersburg Polytechinc University, St. Petersburg, 195251, Russia

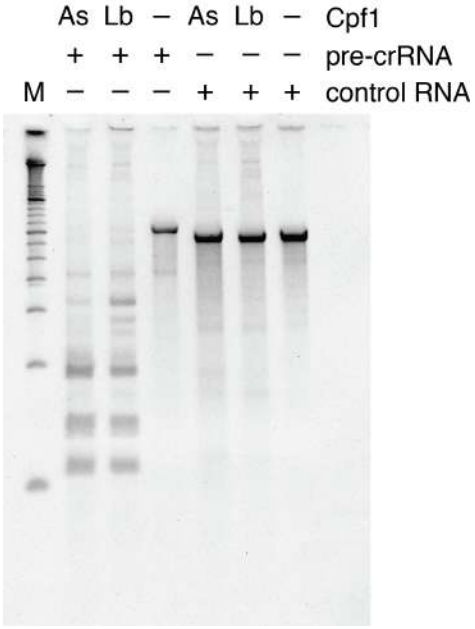
[11] Waksman Institute for Microbiology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

[12] Institute of Molecular Genetics, Russian Academy of Sciences, Moscow, 123182, Russia

*These authors contributed equally to this work.

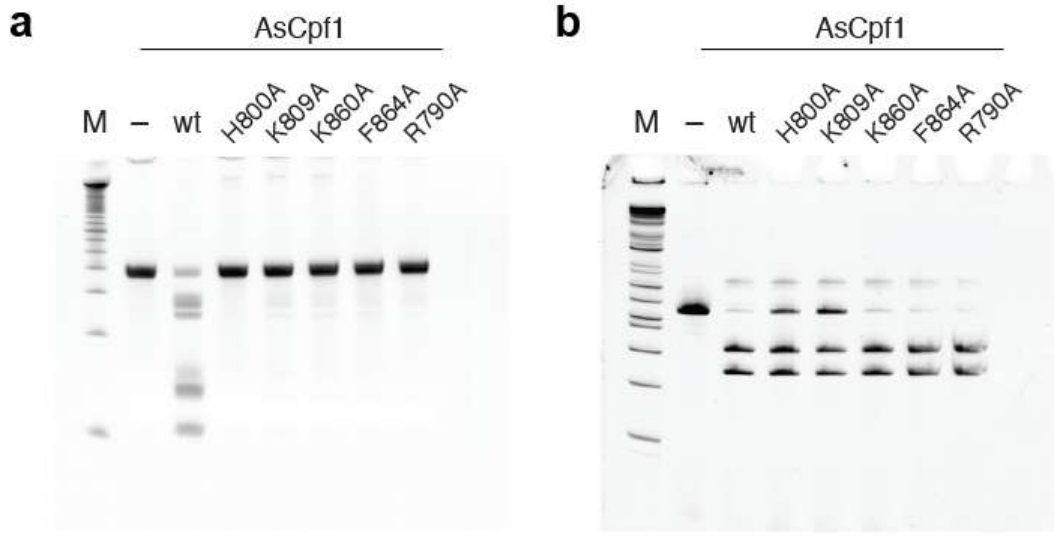
[†] To whom correspondence should be addressed: F.Z. (zhang@broadinstitute.org) or J.v.d.O. (john.vanderoost@wur.nl)

SUPPLEMENTARY FIGURES



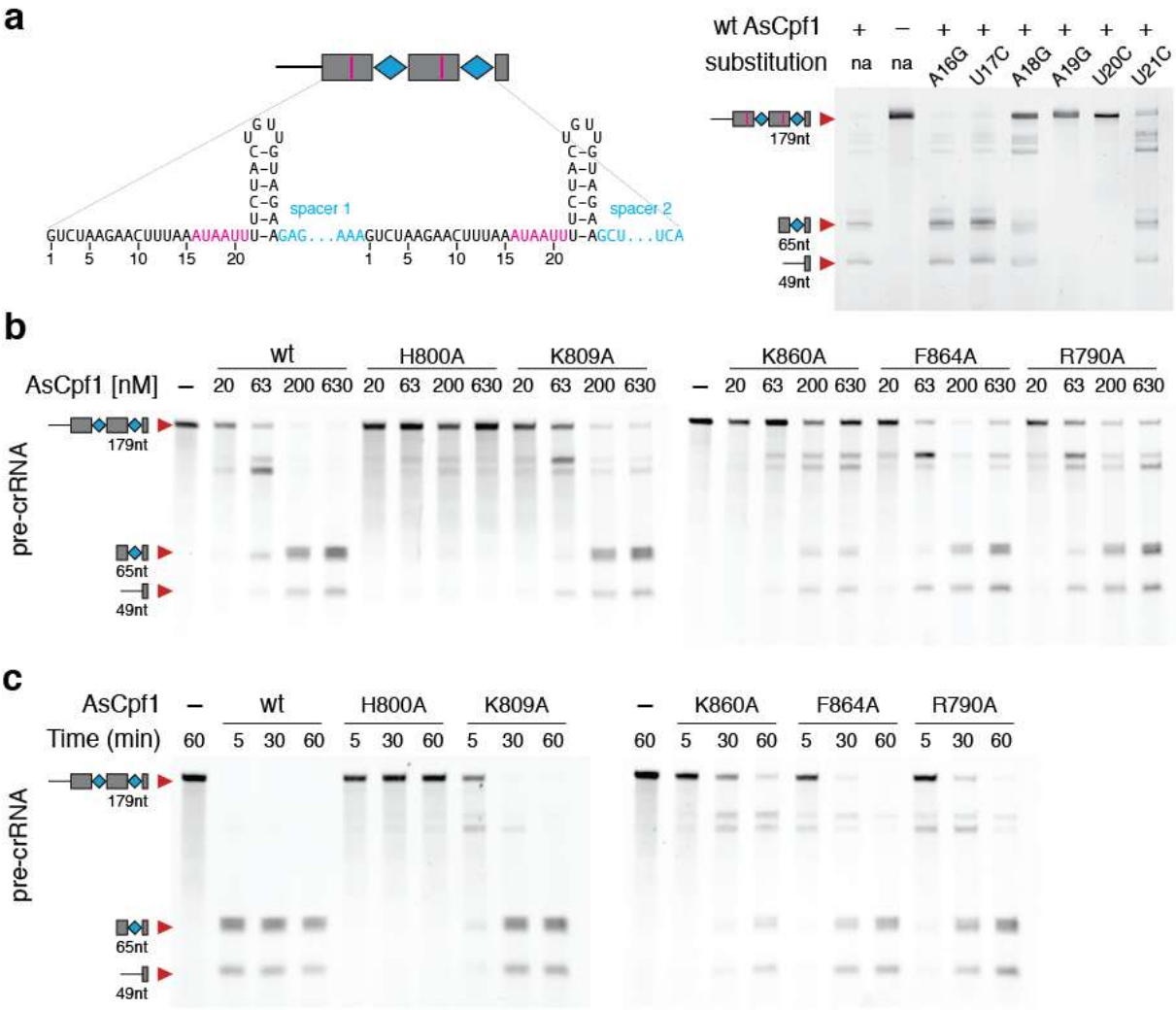
Supplementary figure 1 | Full gel image of figure 1b

Full gel image for in vitro processing of FNCpf1 pre-crRNA transcript with purified AsCpf1 or LpCpf1 protein. M = DNA standard



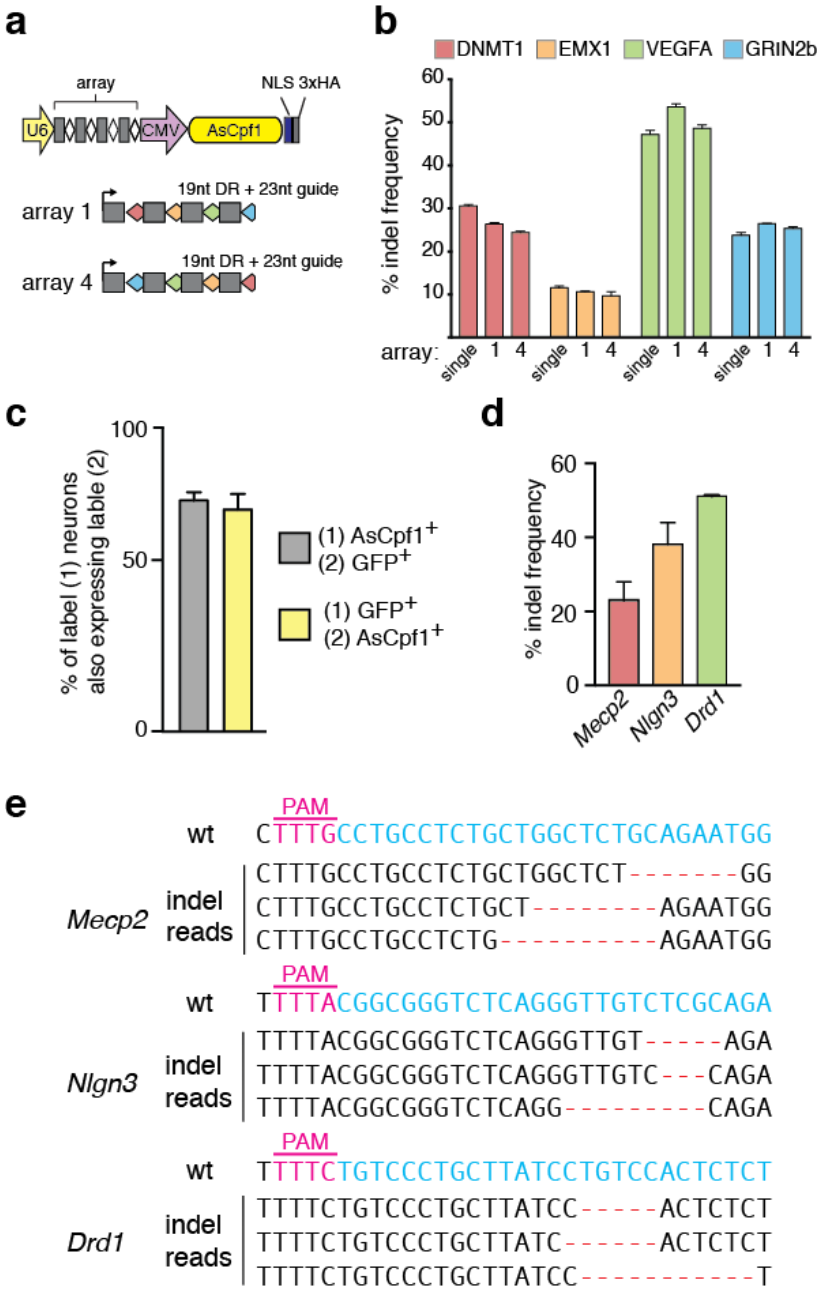
Supplementary figure 2 | Full gel images of figure 1d

(a) Full gel image for pre-crRNA cleavage. **(b)** Full gel image for DNA cleavage. M = DNA standard.



Supplementary figure 3 | Cpf1 mediated pre-crRNA cleavage is sequence and dose dependent.

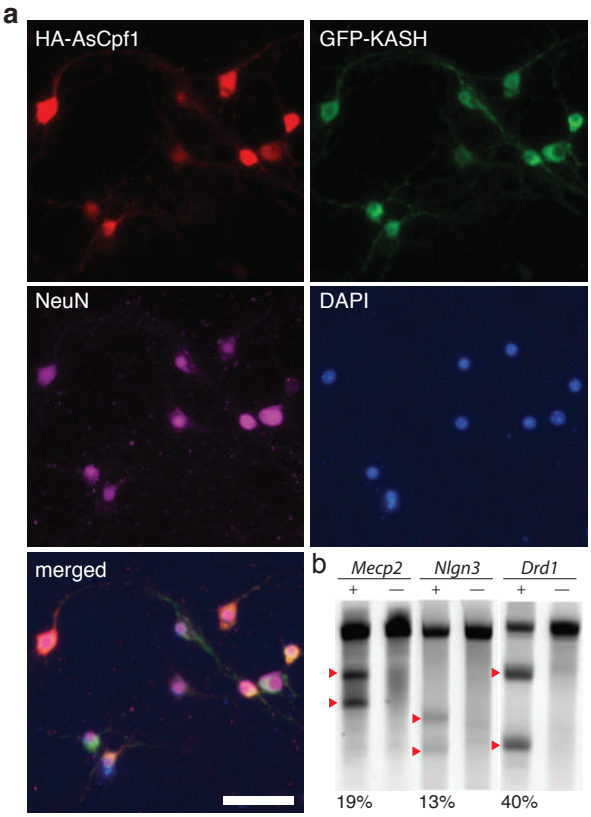
(a) Cpf1 mediated pre-crRNA processing is sequence dependent. Single nucleotide substitutions at position A19 and U20 abolish RNA cleavage *in vitro*. 200 nM pre-crRNA was cleaved with 500 nM Cpf1 in 1 hour. **(b, c)** AsCpf1 point mutants, with the exception of H800A, are active at high dose. (c) Titration of AsCpf1 mutants reveals pre-crRNA processing at high AsCpf1 protein concentration. (d) Prolonged incubation time allows pre-crRNA processing by AsCpf1 point mutants. Only H800A does not process pre-crRNA to mature crRNA at high dose. 165 nM pre-crRNA was incubated with the indicated concentration (c) or with 500 nM AsCpf1 protein (d) for 30 min.



Supplementary figure 4 | Indel levels are not influenced by guide order.

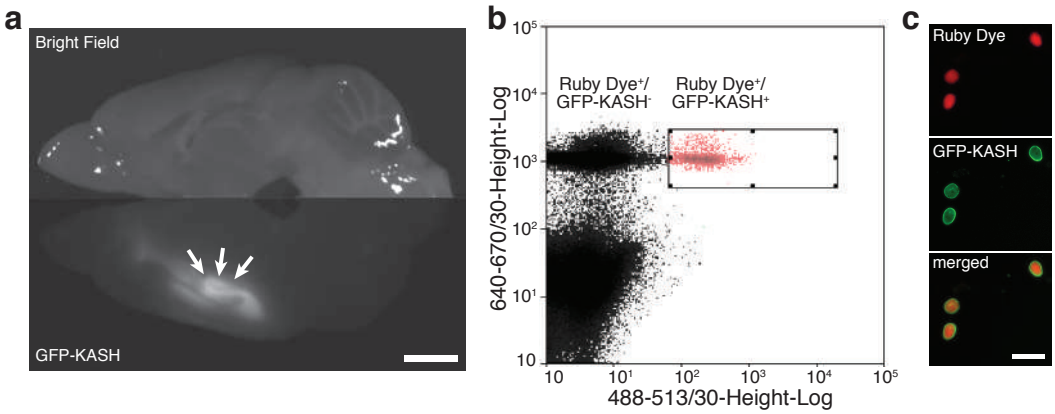
(a) Schematic of multiplex gene editing with AsCpf1, using a single plasmid approach. Two arrays with guides in reversed order are compared (array-1 and array-4). **(b)** Quantification of indel frequencies measured by Surveyor nuclease assay. Guides expressed from array-1 and array-4 result in similar indel frequencies for each targeted gene. **(c)** Quantification of neurons efficiently transduced by the dual-vector system (*n* = 581 nuclei from 3 mice). **(d)** NGS indel

analysis of modified *Mecp2*, *Nlgn3* and *Drd1* loci in single nuclei (*n* = 59 cells from 2 male mice, error bars represent mean ± SEM). **(e)** Representative mutation patterns detected by NGS. Blue, wild-type (wt) sequence; red dashes, deleted bases; PAM sequence marked in magenta.



Supplementary figure 5 | AAV delivery of AsCpf1 and multiplex gene editing in primary neurons.

(a) Immunostaining of AsCpf1 (anti-HA antibody, red) and GFP-KASH (anti-GFP antibody, green) in primary cortical neurons (anti-NeuN antibody, magenta) 7 days after viral infection with dual vector system. Nuclei were labeled with DAPI (blue). Scale bar: 25 um. **(b)** SURVEYOR nuclease assay showing indel formations (+) in all 3 targeted loci. Control neurons (-) were infected with AsCpf1 only (Bottom: Indel percentage; representative images from *n* = 3 independent experiments)



Supplementary figure 4 | *In vivo* delivery of AAV dual vector system and sorting of targeted cell nuclei from intact brain.

(a) Sagittal dissection of adult mouse brain 4 weeks after viral delivery shows infected hippocampal formation (bottom). **(b)** Representative FACS plot showing Ruby Dye^{+}/GFP-KASH^{-} and Ruby Dye^{+}/GFP-KASH^{+} nuclei populations. **(c)** Representative images of sorted Ruby Dye^{+}/GFP-KASH^{+} nuclei used for NGS indel analysis. Scale bars: 2 mm in (a), 25 um in (c).

SUPPLEMENTARY TABLES

Supplementary Table 1 | Sequences of pre-crRNA arrays used for *in vitro* cleavage reaction

4 spacer pre-crRNA	GGGGGUCUUUUUUUGCUGAUUUAGGCAAAAACGGGUCU AAGAACUUUAAAUAAUUUCUACUGUUGUAGAUGAGAAG UCAUUUAAUAAGGCCACUGUUAAAAGUCUAAGAACUUU AAAUAAUUUCUACUGUUGUAGAUGCUACUAUUCCUGUG CCUUCAGAUAAUUCAGUCUAAGAACUUUAAAUAAUUUC UACUGUUGUAGAUGUCUAGAGCCUUUUGUAUUAGUAGC CGGUCUAAGAACUUUAAAUAAUUUCUACUGUUGUAGAU UAGCGAUUUAUGAAGGUCAUUUUUUUGUCUAGCUUUAA UGCGGUAGUUUAUCACAGUUAAAUUGCUAACG
2 spacer pre-crRNA	UAGGUCUUUUUUUGCUGAUUUAGGCAAAAACGGGUCUA AGAACUUUAAAUAAUUUCUACUGUUGUAGAUGAGAAGU CAUUUAAUAAGGCCACUGUUAAAAGUCUAAGAACUUUA AAUAAUUUCUACUGUUGUAGAUGCUACUAUUCCUGUGC CUUCAGAUAAUUC
control RNA	UACGCCAGCUGGCGAAAGGGGGAUGUGCUGCAAGGCGA UUAAGUUGGGUAACGCCAGGGUUUUCCCAGUCACGACG UUGUAAAACGACGGCCAGUGAAUUCGAGCUCGGUACCC GGGNNNNNNNNGAGAAGUCAUUUAAUAAGGCCACUGUU AAAAAGCUUGGCGUAAUCAUGGUCAUAGCUGUUUCCUG UGUGAAAUUGUUAUCCGCUCACAAUUCCACACAACAUA CGAGCCGGAAGCAUAAAGUGUAAAGCCUGGGGUGCCUA AUGAGUGAGCUAACUCACAUUAAUUGCGUU

Supplementary Table 2 | Cpf1 guide sequences used for single and pre-crRNA array expression

DNMT1 23nt guide	CTGATGGTCCATGTCTGTTACTC
EMX1 23nt guide	TGGTTGCCCACCCTAGTCATTGG
VEGFA 23nt guide	CTAGGAATATTGAAGGGGGCAGG
GRIN2b 23nt guide	GTGCTCAATGAAAGGAGATAAGG
DNMT1 30nt guide	CTGATGGTCCATGTCTGTTACTCGCCTGTC
EMX1 30nt guide	TGGTTGCCCACCCTAGTCATTGGAGGTGAC
VEGFA 30nt guide	CTAGGAATATTGAAGGGGGCAGGGGAAGGC
GRIN2b 30nt guide	GTGCTCAATGAAAGGAGATAAGGTCCTTGA

Supplementary Table 3 | DNA oligonucleotides for array cloning

array 1 T1	AGATCTGATGGTCCATGTCTGTTACTCAATTTCTACTCTTGT AGATTGGTTGCCCAC
array 1 T2	CCTAGTCATTGGAATTTCTACTCTTGTAGATCTAGGAATAT TGAAGGGGGCAGGAATTTCTACTCTTGTAGATGTGCTCAAT GAAAGGAGATAAGG
array 1 B1	AAAACCTTATCTCCTTTCATTGAGCACATCTACAAGAGTAG AAATTCCTGCCCCCTT
array 1 B2	CAATATTCCTAGATCTACAAGAGTAGAAATTCCAATGACTA GGGTGGGCAACCAATCTACAAGAGTAGAAATTGAGTAACA GACATGGACCATCAG
array 2 T1	AGATCTGATGGTCCATGTCTGTTACTCGCCTGTCAATTTCTA CTCTTGTAGATTGGTTGCCCACCCTAGTC
array 2 T2	ATTGGAGGTGACAATTTCTACTCTTGTAGATCTAGGAATAT TGAAGGGGGCAGGGGAAGGCAATTTCTACTCTTGTAGATG TGCTCAATGAAAGGAGATAAGGTCCTTGA
array 2 B1	AAAATCAAGGACCTTATCTCCTTTCATTGAGCACATCTACA AGAGTAGAAATTGCCTTCCCCTGCCCCCTT
array 2 B2	CAATATTCCTAGATCTACAAGAGTAGAAATTGTCACCTCCA ATGACTAGGGTGGGCAACCAATCTACAAGAGTAGAAATTG ACAGGCGAGTAACAGACATGGACCATCAG
array 3 T1	AGATGTCAAAAGACCTTTTTAATTTCTACTCTTGTAGATCT GATGGTCCATGTCTGTTACTCGCCTGTCGTCAAAAGACCTT TTTAATTTCTACTCTTGTAGATTGGTTGCCCACCCTAGTCAT TGGAGGTGACGTCAAAAGACCTTTTTAATTTCTACTCTTGT AGATCTAGGAATATT
array 3 T2	GAAGGGGGCAGGGGAAGGCGTCAAAAGACCTTTTTAATTT CTACTCTTGTAGATGTGCTCAATGAAAGGAGATAAGGTCCT TGAGTCAAAAGACCTTTTTAATTTCTACTCTTGTAGAT
array 3 B1	AAAAATCTACAAGAGTAGAAATTAAAAAGGTCTTTTGACT CAAGGACCTTATCTCCTTTCATTGAGCACATCTACAAGAGT AGAAATTAAAAAGGTCTTTTGACGCCTTCCCCTGCCCCCTT CAATATTCCTAGATCTACAAGAGTAGAAATTAAAAAGGTC TTTTGACGTCACCTCCAA
array 3 B2	TGACTAGGGTGGGCAACCAATCTACAAGAGTAGAAATTAA AAAGGTCTTTTGACGACAGGCGAGTAACAGACATGGACCA TCAGATCTACAAGAGTAGAAATTAAAAAGGTCTTTTGAC

array 4 T1	AGATGTGCTCAATGAAAGGAGATAAGGAATTTCTACTCTTG TAGATCTAGGAATATT
array 4 T2	GAAGGGGGCAGGAATTTCTACTCTTGTAGATTGGTTGCCCA CCCTAGTCATTGGAATTTCTACTCTTGTAGATCTGATGGTCC ATGTCTGTTACTC
array 4 B1	AAAAGAGTAACAGACATGGACCATCAGATCTACAAGAGTA GAAATTCCAATGACTAG
array 4 B2	GGTGGGCAACCAATCTACAAGAGTAGAAATTCCTGCCCCC TTCAATATTCCTAGATCTACAAGAGTAGAAATTCCTTATCT CCTTTCATTGAGCAC

Supplementary Table 4 | PCR primers for amplification of DNA regions for SURVEYOR nuclease assay

DNMT1 for	CTGGGACTCAGGCGGGTCAC
DNMT1 rev	CCTCACACAACAGCTTCATGTCAGC
EMX1 for	CCATCCCCTTCTGTGAATGT
EMX1 rev	GGAGATTGGAGACACGGAGA
VEGFA for	CTCAGCTCCACAAACTTGGTGCC
VEGFA rev	AGCCCGCCGCAATGAAGG
GRIN2b for	GCATACTCGCATGGCTACCT
GRIN2b rev	CTCCCTGCAGCCCCTTTTTA
Mecp2 for	GGTCTCATGTGTGGCACTCA
Mecp2 rev	TGTCCAACCTTCAGGCAAGG
Nlgn3 for	GTAACGTCCTGGACACTGTGG
Nlgn3 rev	TTGGTCCAATAGGTCATGACG
Drd1 for	TGGCTAAGCCTGGCCAAGAACG
Drd1 rev	TCAGGATGAAGGCTGCCTTCGG

Supplementary Table 5 | PCR primers for amplification of DNA regions for next generation sequencing

NGS DNMT1 for	CCATCTCATCCCTGCGTGTCTCCTGAACGT TCCCTTAGCACTCTGCC
NGS DNMT1 rev	CCTCTCTATGGGCAGTCGGTGATGCCTTAG CAGCTTCCTCCTCC
NGS EMX1 for	CCATCTCATCCCTGCGTGTCTCCGGGCTCC CATCACATCAACCG
NGS EMX1 rev	CCTCTCTATGGGCAGTCGGTGATGCCAGAG TCCAGCTTGGGCCC
NGS VEGFA for	CCATCTCATCCCTGCGTGTCTCCCAGGGGT CACTCCAGGATTCCA
NGS VEGFA rev	CCTCTCTATGGGCAGTCGGTGATGCATTGG CGAGGAGGGAGCAG
NGS GRIN2b for	CCATCTCATCCCTGCGTGTCTCCGTTCAAG GATTTCTGAGGCTTTTGAAAG
NGS GRIN2b rev	CCTCTCTATGGGCAGTCGGTGATGGGGCTT CATCTTCAACTCGTCGAC
NGS Mecp2 for	CCATCTCATCCCTGCGTGTCTCCGGAA AAGTCAGAAGACCAGG
NGS Mecp2 rev	CCTCTCTATGGGCAGTCGGTGATGGTGGGG TCATCATACATAGG
NGS Nlgn3 for	CCATCTCATCCCTGCGTGTCTCCACCCCGA GGATGGTGTCTCG
NGS Nlgn3 rev	CCTCTCTATGGGCAGTCGGTGATGGGTAGA AGGCGTAGAAGTAGG
NGS Drd1 for	CCATCTCATCCCTGCGTGTCTCCAAGCCAC CGGAAGTGCTTTCC
NGS Drd1 rev	CCTCTCTATGGGCAGTCGGTGATGCACAGC TTTCCAGGGCATGACC

SUPPLEMENTARY METHODS

Cpf1 protein purification

Humanized Cpf1 were cloned into a bacterial expression vector (6-His-MBP-TEV-Cpf1, a pET-based vector kindly given to us by Doug Daniels). Two liters of Terrific Broth growth media with 100 μg/mL ampicillin was inoculated with 10 mL overnight culture Rosetta (DE3) pLyseS (EMD Millipore) cells containing the Cpf1 expression construct. Growth media plus inoculant was grown at 37°C until the cell density reached 0.2 OD600, then the temperature was decreased to 21°C. Growth was continued until OD600 reached 0.6 when a final concentration of 500 μM IPTG was added to induce MBP-Cpf1 expression. The culture was induced for 14-18 h before harvesting cells and freezing at -80°C until purification. Cell paste was resuspended in 200 mL of Lysis Buffer (50 mM Hepes pH 7, 2M NaCl, 5 mM $MgCl_2$, 20 mM imidazole) supplemented with protease inhibitors (Roche cOmplete, EDTA-free) and lysozyme. Once homogenized, cells were lysed by sonication (Branson Sonifier 450) then centrifuged at 10,000 x g for 1 h to clear the lysate. The lysate was filtered through 0.22 micron filters (Millipore, Stericup) and applied to a nickel column (HisTrap FF, 5 mL), washed, and then eluted with a gradient of imidazole. Fractions containing protein of the expected size were pooled, TEV protease (Sigma) was added, and the sample was dialyzed overnight into TEV buffer (500 mM NaCl, 50 mM Hepes pH 7, 5 mM MgCl, 2 mM DTT). After dialysis, TEV cleavage was confirmed by SDS-PAGE, and the sample was concentrated to 500 μL prior to loading on a gel filtration column (HiLoad 16/600 Superdex 200) via FPLC (AKTA Pure). Fractions from gel filtration were analyzed by SDS-PAGE; fractions containing Cpf1 were pooled and concentrated to 200 μL (50mM Tris-HCl pH7.5, 2mM DTT, 5% glycerol, 500mM NaCl) and either used directly for biochemical assays or frozen at -80°C for storage.

In vitro synthesis of pre-crRNA arrays

Pre-crRNA arrays were synthesized using the HiScribe™ T7 High Yield RNA Synthesis Kit (NEB). PCR fragments coding for arrays, with a short T7-priming sequence on the 5' end, were utilized as templates for *in vitro* transcription reaction (Supplementary Table 1). T7 transcription was performed for 4 h and then RNA was purified using the MEGAclear™ Transcription Clean-Up Kit (Ambion).

***In vitro* cleavage assay**

In vitro cleavage was preformed with purified recombinant proteins for AsCpf1 and LbCpf1. Cpf1 protein together with *in vitro* transcribed pre-crRNA arrays were incubated at 37°C in cleavage buffer (20mM Tris HCl, 50mM KCl supplemented with RNase Inhibitor Murine (NEB)) for 5 min to 1 h, as indicated in figure legends. Each cleavage reaction contained 20 to 630 nM of Cpf1 protein and 165 or 330 nM of synthetized pre-crRNA array, as indicated in figure legends. For DNA cleavage, 25 nM of target was cleaved with 165 nM Cpf1 and 340 nM crRNA for 30 min at 37°C. Reactions were stopped with proteinase K (Qiagen), heat denaturation and run on 10% TBE-Urea polyacrylamid gels. Gels were stained with SYBR Gold DNA stain (Life Technologies) for 10 min and imaged with a Gel Doc™ EZ gel imaging system (Bio-rad).

Pre-crRNA array design and cloning

crRNAs were designed as four oligos (IDT) consisting of direct-repeats, each one followed by a crRNA (Supplementary Table 3). The oligos favored a one-directional annealing through their sticky-end design. The oligonucleotides (final concentration 10 μM) were annealed in 10X T4 ligase buffer (final concentration 1X; NEB) and T4 PNK (5 units; NEB). Thermocycler conditions were adjusted to 37°C for 30 min, 95°C for 5 min followed by a -5°C/min ramp down to 25°C. The annealed oligonucleotides were diluted 1:10 (final concentration 1 μM) and ligated into *BsmB*I-cut pcDNA-huAsCpf1-U6 (pY26), utilizing T7 DNA ligase (Enzymatics), in room temperature for 30 min. The constructs were transformed into STBL3 bacteria and plated on ampicillin-containing (100 g/ml) agar plates. Single colonies were grown in standard LB media (Broad Facilities) for 16 h. Plasmid DNA was harvested from bacteria according to QIAquick Spin Miniprep protocol (QIAGEN). Guide sequences targeting human genes are listed in Supplementary Table 2.

Cell culture and transfection

Human embryonic kidney 293T (HEK293T) cell line (Life Technologies) were maintained in Dulbecco's modified Eagle's Medium (DMEM) + GLUTAMAX (Gibco) supplemented with 10% FBS (HyClone) at 37°C with 5% CO2 incubation. HEK293FT cells were seeded onto 24-well plates (Corning) 24 h before transfection. Cells were transfected using Lipofectamine 2000

(Life Technologies) at 70–80% confluency following the manufacturer's recommended protocol. For each well of a 24-well plate, a total of 500 ng plasmid DNA was use, each well represents one technical replicate.

Surveyor nuclease assay for genome modification

HEK293T cells were transfected with DNA as described above. Cells were incubated at 37°C for 72 h post-transfection before genomic DNA extraction. Genomic DNA was extracted using the QuickExtract DNA Extraction Solution (Epicentre) following the manufacturer's protocol. Briefly, pelleted cells were resuspended in QuickExtract solution and incubated at 65 °C for 15 min, 68 °C for 15 min, and 98 °C for 10 min. The genomic region flanking the CRISPR target site for each gene was PCR amplified (primers listed in Supplementary Table 4), and products were purified using QIAQuick PCR purification Kit (Qiagen) following the manufacturer's protocol. 200 ng total of the purified PCR products were mixed with 1 µl 10X Taq DNA Polymerase PCR buffer (Enzymatics) and ultrapure water to a final volume of 10 µl, and subjected to a re-annealing process to enable heteroduplex formation: 95 °C for 10 min, 95 °C to 85 °C ramping at −2 °C/s, 85 °C to 25 °C at −0.25 °C/s, and 25 °C hold for 1 min. After re-annealing, products were treated with Surveyor nuclease and Surveyor enhancer S (IDT) following the manufacturer's recommended protocol, and analyzed on 10% Novex TBE polyacrylamide gels (Life Technologies). Gels were stained with SYBR Gold DNA stain (Life Technologies) for 10 min and imaged with a Gel Doc gel imaging system (Bio-rad). Quantification was based on relative band intensities. Indel percentage was determined by the formula, $100 \times (1 - (1 - (b + c)/(a + b + c))1/2)$, where a is the integrated intensity of the undigested PCR product, and b and c are the integrated intensities of each cleavage product.

Small RNA extraction from cells

HEK293T cells were harvested 48 h after transfection and the total RNA was extracted from with the miRNeasy mini kit (Qiagene) according to manufactures conditions. rRNA was removed using the bacterial Ribo-Zero rRNA removal kit (Illumina).

NGS analysis of *in vitro* and *in vivo* cleavage pattern

RNA-seq libraries were prepared using a derivative of a previously described method [1]. Briefly, after PNK treatment in absence and presence of ATP (enrichment of 5'OH and 3'P respectively)

RNA cleavage products were poly-A tailed with E.coli Poly(A) Polymerase (NEB), ligated to 5' RNA adapters using T4 RNA ligase I (NEB) and reverse transcribed with AffinityScript Multiple Temperature Reverse Transcriptase (Agilent Technologies). cDNA was amplified by a fusion PCR method to attach the Illumina P5 adapters as well as unique sample-specific barcodes to the target amplicons [2]. PCR products were purified by gel-extraction using QiaQuick PCR purification Kit (Qiagen) following the manufacturer's recommended protocol. DNA samples from single nuclei were pre-amplified with SURVEYOR primers (Supplementary Table 4) and nested-PCR was performed with NGS primers (Supplementary Table 5) before Illumina barcodes were added. Finally, barcoded and purified DNA samples were quantified by Qubit 2.0 Fluorometer (Life Technologies) and pooled in an equimolar ratio. Sequencing libraries were then sequenced with the Illumina MiSeq Personal Sequencer (Life Technologies).

RNA-sequencing analysis

The prepared cDNA libraries were pooled and sequenced on a MiSeq (Illumina). Pooled sequencing reads were assigned to their respective samples on the basis of their corresponding barcodes and aligned to the proper CRISPR array template sequence using BWA 3. Interval lists were generated using the paired-end alignment coordinates and the intervals were used to extract entire transcript sequences using Galaxy tools (http://usegalaxy.org)[3] . The extracted transcript sequences were analyzed using Geneious 9.

AAV DNA constructs

The AAV hSyn1-HA-NLS-AsCpf1-spA vector was generated by PCR amplifying the AsCpf1 encoding sequence using forward PCR primer including HA-NLS (5'-aacacaggaccggtgccaccatgtacccatacgatgttccagattacgcttcgccgaagaaaaagcgcaaggtcgaagcgtccacacagt tcgagggctttaccaacctgtatcaggtgagc-3') and reverse PCR primer including a short poly A signal (spA)(5'-gcggccgcacacaaaaaaccaacacacagatctaatgaaaataaagatcttttattgaattcttagttgcgcagctcctggatgtaggccagc c-3') [4], and cloning of the resulting PCR template into AAV backbone under control of the human *Synapsin* 1 promoter (hSyn1). For the generation of AAV U6-DR(*Sap*I)-hSyn1-GFP-KASH-hGH (not shown) and U6-*Mecp2-Nlgn3-Drd1* array-hSyn1-GFP-KASH-hGH vectors, gene blocks (Integrated DNA Technologies) encoding for U6-DR(*Sap*I) and U6-*Mecp2-Nlgn3-*

Drd1 array, respectively, have been cloned into AAV hSyn-GFP-KASH-hGH backbone (Addgene PX552). All constructs were verified by Sanger sequencing.

Production of AAV vectors

AAV1 particles in DMEM culture medium were produced as described previously [5]. Briefly, HEK293FT cells were transfected with transgene plasmid, AAV1 serotype plasmid and pDF6 helper plasmid using polyethyleneimine (PEI). DMEM culture medium containing low titer AAV1 particles was collected after 48 h and sterile filtered. For high titer AAV1/2 production, HEK293FT cells were transfected with AAV1 and AAV2 serotype plasmids in equal ratios, transgene plasmid and pDF6 helper plasmid. 48 h after transfection, cells were harvested and high titer AAV1/2 virus was purified on heparin affinity column [5]. The titer of AAV vectors was determined by real-time quantitative PCR (qPCR) using probe and primers specific for the hSyn1 promoter sequence (Integrated DNA Technologies).

Primary cortical neuron culture

Mice used to obtain neurons for tissue cultures were sacrificed according to the protocol approved by the Broad's Institutional Animal Care and Use Committee (IACUC). Primary neurons were prepared from postnatal day P0.5 mouse brains and plated on laminin/PDL coated coverslips (VWR). Briefly, cortices were dissected in ice-cold HBSS (Sigma) containing 50 ug/ml penicillin/streptomycin (Thermo Fisher) and incubated for 10 min at 37°C with HBSS containing 125 Units papain (Worthington Biochemical) and 400 Units DNase I (Sigma). After enzymatic digestion, the tissues were washed twice in HBSS and gently triturated with a fire-polished Pasteur pipette. Cells were then transferred into neuronal growth medium (Neurobasal A medium, supplemented with B27, Glutamax (Life Technologies) and penicillin/streptomycin) and grown at 37°C and 5% CO_2. For inhibition of glia cell proliferation, cytosine-beta-D-arabinofuranoside (AraC, Sigma) at a final concentration of 10 μM was added to the culture medium after 48 h and replaced by fresh culture medium after 72 h. For AAV1 transduction, cultured neurons were infected with low titer AAV1 as described previously [5]. One week after transduction, neurons were harvested for isolating genomic DNA (QuickExtract DNA extraction buffer (Epicentre)), or fixed in 4% paraformaldehyde (PFA) for immunofluorescence staining.

Stereotactic injection of AAV1/2 into the mouse brain

The Broad's Institutional Animal Care and Use Committee (IACUC) approved all animal procedures described here. Craniotomy was performed on adult (12-16 weeks) male C57BL/6N mice according to approved procedures, and 1 µl of 1:1 AAV mixture (AAV hSyn1-HA-NLS-AsCpf1-spA: 2.25 x 10^{12} Vg/ml; AAV U6-*Mecp2-Nlgn3-Drd1* array-hSyn1-GFP-KASH-hGH: 9.7 x 10^{12} Vg/ml) was injected into the dorsal dentate gyrus (anterior/posterior: -1.7; mediolateral: +/-0.6; dorsal/ventral: -2.15). The pipette was held in place for 3-5 min post injection to prevent leakage. After injection, the incision was sutured and post-operative analgesics were administered according to approved IACUC protocol for three days following surgery.

Purification of cell nuclei from intact brain tissue

Cell nuclei from AAV1/2 injected hippocampal tissue were purified as described previously [4]. Briefly, dissected tissue was homogenized in ice-cold homogenization buffer (HB) (320 mM sucrose, 5 mM CaCl, 3 mM $Mg(Ac)_2$, 10 mM Tris pH7.8, 0.1 mM EDTA, 0.1 % NP40, 0.1 mM PMSF, 1 mM β-mercaptoethanol) using 2 ml Type A and B Dounce homogenizer (Sigma). For gradient centrifugation, OptiPrep™ density gradient medium (Sigma) was used. Samples were centrifuged at 10,100 x g (7,500 rpm) for 30 min at 4°C (Beckman Coulter, SW28 rotor). Cell nuclei pellets were resuspended in 65 mM β-glycerophosphate (pH 7.0), 2 mM $MgCl_2$, 25 mM KCl, 340 mM sucrose and 5% glycerol. Number and quality of purified nuclei was examined using bright field microscopy.

Fluorescent Activated Cell Sorting (FACS) of cell nuclei

Purified cell nuclei were co-labeled with Vybrant® DyeCycle™ Ruby Stain (1:500, Life Technologies) and sorted using a Beckman Coulter MoFlo Astrios EQ cell sorter (Broad Institute Flow Cytometry Core). Single and population (250-500 nuclei) GFP-KASH$^+$ and GFP-KASH$^-$ nuclei were collected in 96 well plates containing 5 µl of QuickExtract DNA

extraction buffer (Epicentre) and spun down at 2,000 x g for 2 min. Each 96 well plate included two empty wells as negative control.

Western blot analysis

AAV injected dentate gyrus tissues were lysed in 100 µl of ice-cold RIPA buffer (Cell Signaling Technologies) containing 0.1% SDS and protease inhibitors (Roche, Sigma) and sonicated in a Bioruptor sonicator (Diagenode) for 1 min. Protein concentration was determined using the BCA Protein Assay Kit (Pierce Biotechnology, Inc.). Protein samples were separated under reducing conditions on 4-15% Tris-HCl gels (Bio-Rad) and analyzed by Western blotting using primary antibodies: mouse anti-HA (Cell Signaling Technologies 1:500), mouse anti-GFP (Roche, 1:500), rabbit anti-Tubulin (Cell Signaling Technologies, 1:10,000) followed by secondary anti-mouse and anti-rabbit HRP antibodies (Sigma-Aldrich, 1:10,000). Blots were imaged with Amersham Imager 600.

Immunofluorescent staining

4 weeks after viral delivery, mice were transcardially perfused with PBS followed by PFA according to approved IACUC protocol. 30 µm free floating sections (Leica, VT1000S) were boiled for 2 min in sodium citrate buffer (10 mM tri-sodium citrate dehydrate, 0.05% Tween20, pH 6.0) and cooled down at RT for 20 min. Sections were blocked with 4% normal goat serum (NGS) in TBST (137 mM NaCl, 20 mM Tris pH 7.6, 0.2% Tween-20) for 1 h. Primary antibodies were diluted in TBST with 4% NGS and sections were incubated overnight at 4°C. After 3 washes in TBST, samples were incubated with secondary antibodies for 1 h at RT. After 3 times washing with TBST, sections were mounted using VECTASHIELD HardSet Mounting Medium including DAPI and visualized with confocal microscope (Zeiss LSM 710, Ax10 ImagerZ2, Zen 2012 Software). Following primary antibodies were used: mouse anti-NeuN (Millipore, 1:50-1:400); chicken anti-GFP (Aves Labs, 1:200-1:400); rabbit anti-HA (Cell Signaling Technologies, 1:100). Anti-HA signaling was amplified using biotinylated anti-rabbit (1:200) followed by streptavidin AlexaFluor® 568 (1:500) (Life Technologies). Anti-chicken AlexaFluor®488 and anti-mouse AlexaFluor®647 secondary antibodies (Life Technologies) were used at 1:1000.

Reference

1. Heidrich, N., Dugar, G., Vogel, J., and Sharma, C. Investigating CRISPR RNA Biogenesis and Function Using RNA-seq. *In CRISPR, M. Lundgren, E. Charpentier, and P.C. Fineran, eds. (Springer New York)*, pp. 1-21. (2015).
2. Hsu, P.D. et al. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature biotechnology* **31**, 827-832 (2013).
3. Giardine, B. et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* **15**, 1451-1455 (2005).
4. Swiech, L. et al. In vivo interrogation of gene function in the mammalian brain using CRISPR-Cas9. *Nat Biotechnol* **33**, 102-106 (2015).
5. Konermann, S. et al. Optical control of mammalian endogenous transcription and epigenetic states. *Nature* **500**, 472-476 (2013).

Chapter V

Position of *Deltaproteobacteria* Cas12e nuclease cleavage sites depends on spacer length of guide RNA

Introduction:

This chapter is dedicated to another member of Type V effectors, DpbCas12e nuclease from *Deltaproteobacteria* (formerly DpbCasX). In 2019 Liu et al. showed that DpbCas12b domain organization is quite similar to Cas12a and that this nuclease produces long, 10 nt 5'-overhangs at DNA cut site. Here, we compared *in vitro* DNA cleavage patterns of Cas12a and Cas12e using HTS on several DNA targets and showed that similarly to Cas12a, Cas12e produces 3-5 nt 5'-overhangs, in contrast to the previously published data. In addition, we show, that the 5'-overhangs, generated by DpbCas12e can be made longer by using crRNA with shorter spacer segments.

Contribution:

I conceived the study and designed the experiments. Most of the experiments were performed by the first authors of the paper: Polina Selkova performed all biochemical assays and Aleksandra Vasileva bioinformatically analyzed the HTS results. Konstantin Severinov and I wrote the manuscript with the help from the first authors. We would like to thank all authors for their help.

RNA BIOLOGY
https://doi.org/10.1080/15476286.2020.1777378

Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

OPEN ACCESS Check for updates

Position of Deltaproteobacteria Cas12e nuclease cleavage sites depends on spacer length of guide RNA

Polina Selkova[a,*], Aleksandra Vasileva[a,*], Georgii Pobegalov [b], Olga Musharova [c,d], Anatolii Arseniev[a], Maksim Kazalov[b], Tatyana Zyubko[b], Nataliia Shcheglova[b], Tatyana Artamonova[b], Mikhail Khodorkovskii[b], Konstantin Severinov[a,d], and Iana Fedorova [c]

[a]Center for Precision Genome Editing and Genetic Technologies for Biomedicine, Institute of Gene Biology, Russian Academy of Sciences, Moscow, Russia; [b]Peter the Great St. Petersburg Polytechnic University, Saint Petersburg, Russia; [c]Skolkovo Institute of Science and Technology, Center of Life Sciences, Moscow, Russia; [d]Institute of Molecular Genetics, Russian Academy of Sciences, Moscow, Russia

ABSTRACT

Cas12e proteins (formerly CasX) form a distinct subtype of Class II type V CRISPR-Cas effectors. Recently, it was shown that DpbCas12e from *Deltaproteobacteria* and PlmCas12e from *Planctomycetes* can introduce programmable double-stranded breaks in mammalian genomes. Thus, along with Cas9 and Cas12a Class II effectors, Cas12e could be harnessed for genome editing and engineering. The location of cleavage points in DNA targets is important for application of Cas nucleases in biotechnology. DpbCas12e was reported to produce extensive 5′-overhangs at cleaved targets, which can make it superior for some applications. Here, we used high throughput sequencing to precisely map the DNA cut site positions of DpbCas12e on several DNA targets. In contrast to previous observations, our results demonstrate that DNA cleavage pattern of Cas12e is very similar to that of Cas12a: DpbCas12e predominantly cleaves DNA after nucleotide position 17–19 downstream of PAM in the non-target DNA strand, and after the 22nd position of target strand, producing 3–5 nucleotide-long 5′-overhangs. We also show that reduction of spacer sgRNA sequence from 20nt to 16nt shifts Cas12e cleavage positions on the non-target DNA strand closer to the PAM, producing longer 6–8nt 5′-overhangs. Overall, these findings advance the understanding of Cas12e endonucleases and may be useful for developing of DpbCas12e-based biotechnology instruments.

ARTICLE HISTORY
Received 8 February 2020
Revised 10 May 2020
Accepted 14 May 2020

KEYWORDS
CRISPR-Cas; DpbCas12e; CasX; AsCas12a; SpCas9; sgRNA; cut site mapping; AsCpf1

Introduction

RNA-guided effector nucleases from Class II CRISPR-Cas bacterial defence systems are widely used as biotechnology instruments. These enzymes found numerous applications in targeted genome editing, regulation of transcription, and epigenetic modulation [1]. SpCas9 nuclease from *Streptococcus pyogenes* was the first Cas nuclease used for genome editing in eukaryotes [2,3]. It belongs to type II of Class II CRISPR-Cas nucleases and remains the best-characterized and most widely used Cas protein to date [1,4]. In addition to SpCas9, other type II Cas nucleases were successfully used in genome engineering [5–8].

Besides Cas9 enzymes, CRISPR-Cas effectors of other Class II types have found biotechnological applications [9,10]. Thus, Cas12a proteins belonging to type V-A CRISPR-Cas effectors possess specific features distinguishing them from the Cas9 proteins, due to their distinct domain organization. For instance, unlike Cas9, Cas12a is able to catalyse maturation of crRNA in the absence of other factors, which facilitates multiplex genome editing [11,12]. In 2017 Burstein et al. discovered new Cas protein belonging to Class II type V-E CRISPR-Cas systems – Cas12e (formerly CasX) [4,13]. Although Cas12e proteins demonstrate some similarities to Cas12a in domain organization, which will be discussed later, Cas12e and Cas12a effectors are quite distinct. For instance, in opposite to Cas12a which requires only crRNA for DNA recognition, Cas12e are dual-RNA-guided effectors (Fig. 1) [4,13].

In 2019 Jun-Jie Liu et al. using electron microscopy determined the domain composition of DpbCas12e protein from *Deltaproteobacteria* and also, demonstrated that DpbCas12e, as well as PlmCas12e from *Planctomycetes*, have genome editing activity in human cells [14]. Analysis of DpbCas12e-sgRNA-DNA complex revealed non-target-strand binding (NTSB) and target-strand loading (TSL) domains [14]. The TSL domain is located in a position analogous to that of the so-called ‘Nuc’ domain in Cas12a [14,15]. These domains perform similar functions in Cas12e and Cas12a enzymes: after non-target DNA strand cleavage by the RuvC domain, they bend sgRNA-DNA duplex. This conformational change allows the target DNA strand to be cleaved by the RuvC domain [14]. Thus, both Cas12e and Cas12a, rely on a single nuclease domain for double stranded DNA cleavage, in contrast to Cas9, which uses distinct domains, HNH and RuvC, to cleave each DNA strand

CONTACT Konstantin Severinov severik@waksman.rutgers.edu; Iana Fedorova femtokot@gmail.com Skolkovo Institute of Science and Technology, Center of Life Sciences, Moscow, Russia

*Joint Authors

Supplemental data for this article can be accessed here

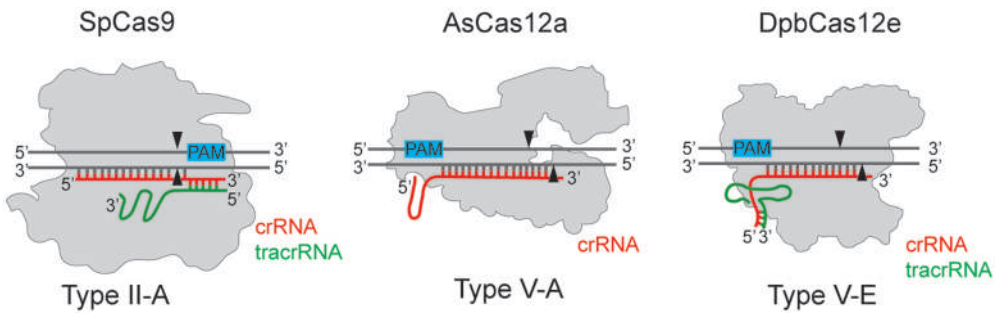


Figure 1. Cas12a and Cas12e belong to Class II Type V CRISPR-Cas effectors, subtypes V-A and V-E, correspondingly.
In contrast to Cas12a, Cas12e enzymes require tracrRNA in addition to crRNA for DNA target recognition. crRNA indicated in red, tracrRNA indicated in green. PAM sequences are shown with blue rectangles. SpCas9 – Cas9 from *Streptococcus pyogenes* (1,368 amino acids), AsCas12a – Cas12a from *Acidaminococcus sp.* (1,307 amino acids), DpbCas12e – Cas12e from *Deltaproteobacteria* (986 amino acids). The pairing between DNA and RNA molecules, as well as indicated positions of DNA cleavage sites shown are schematic.

[14,16,17]. In Cas12e and Cas12a, a large structural change alters accessibility of DNA strands for the RuvC nuclease and in this way compensates the lack of the second nuclease domain [14,17].

Possibly due to the similarity of the DNA cleavage mechanism, both Cas12a and Cas12e generate products with staggered ends. In contrast, Cas9 proteins mainly produce blunt ends [14,18,19]. Interestingly, Jun-Jie Liu et al. found that DpbCas12e produces staggered ends about 10-nucleotides long [14], which is longer than 3–5nt overhangs usually produced by Cas12a proteins. The 5′-overhangs produced by Cas12a and Cas12e, potentially can be used for *in vivo* or *in vitro* insertion of DNA fragments into genome through direct DNA ligation [20]. Determination of precise positions of Cas12e DNA cleavage is important for applications of these nucleases in biotechnology [20,21].

In most studies published to date, mapping of Cas nucleases cut sites was performed using Sanger sequencing of fragments produced in *in vitro* DNA cleavage reactions [22,23], which does not reveal the entire distribution of DNA along the target. High throughput sequencing (HTS) allows much more comprehensive, both quantitative and qualitative, characterization of cleavage sites. Although HTS has been used for both Cas9 and Cas12a-induced DSB (double-stranded DNA breaks) determination, the goal of most of these experiments was the evaluation of off-targeting effects *in vivo* or *in vitro* but not the precise determination of DNA cleavage sites positions [18,19,24,25,26]. The cell-based assays for *in vivo* mapping of DSBs miss information about the precise DNA cleavage positions due to modulation of initial DSBs by endogenous cell nucleases and reparation processes.

Here, we performed HTS mapping of DpbCas12e cut sites produced during *in vitro* DNA cleavage reactions using six different dsDNA targets. We determined DpbCas12e cut sites positions distribution and found the average length of 5′-overhangs generated by this nuclease. Our results show that Cas12e DNA cleavage pattern is very similar to that produced by the Cas12a proteins. We also show that the reducing of the length of spacer segment of sgRNA from 20nt to 16nt can significantly increase the length of 5′-overhangs generated by Cas12e nuclease. These findings can inform development of new DpbCas12e-based genome engineering tools.

Materials and methods

Plasmids

For expression of DpbCas12e, AsCas12a and SpCas9 in *E. coli* pET21a-based genetic vectors carrying the corresponding genes were cloned. The maltose binding protein (MBP) was added to DpbCas12e N-termini through TEV protease cleavage site to increase solubility of DpbCas12e nuclease. The plasmids maps are presented in the Supplementary Table S1.

Recombinant proteins purification

For recombinant SpCas9 and AsCas12a proteins purification competent *E. coli* Rosetta cells were transformed with pET21a_SpCas9, and pET21a_AsCas12a plasmid and grown till OD_{600} = 0.6 in 500 ml LB media supplemented with 100 µg/ml ampicillin. The protein synthesis was induced by adding 1 mM IPTG. After 6 hours of incubation at 22°C the cells were centrifugated at 4000 g and the pellet was lysed by sonication in lysis buffer containing 50 mM Tris-HCl pH = 8 (4°C), 500 mM NaCl, 1 mM beta-mercaptoethanol and 10 mM imidazole supplemented with 1 mg/ml lysozyme (Sigma-Aldrich L6876). The cell lysate was centrifuged at 16,000 g (4°C) and filtered through 0.22 µm filters. The lysate was applied to 1 ml HisTrap HP column (GE Healthcare) and SpCas9 or AsCas12a were eluted by 300 mM imidazole. After affinity chromatography the sample was applied on a Superdex200 Increase 10/300 GL (GE Healthcare) column equilibrated with a buffer containing 50 mM Tris-HCl pH = 8 (4°C), 500 mM NaCl, 1 mM DTT. Fractions containing SpCas9 or AsCas12a monomers were pooled and concentrated using 30 kDa Amicon Ultra-4 centrifugal unit (Merc Millipore, UFC803008). Glycerol was added to a concentration of 10% and the proteins were flash-frozen in liquid nitrogen and stored at −80°C. The purity of the protein was assessed by denaturing 10% PAGE.

For recombinant DpbCas12e purification competent *E. coli* Rosetta cells were transformed with pET21a_DpbCas12e plasmid and grown till OD 600 = 0.6 in 500 ml L of TB (Terrific broth) media supplemented with 100 g/ml ampicillin according to DpbCas12e purification protocol described earlier (Jun-Jie Liu et al.) The protein synthesis was induced by addition of 1 mM IPTG. After 18 hours of incubation at 16°C the cells were centrifugated at 4000 g and the pellet was lysed by sonication in the lysis buffer containing 50 mM Hepes-HCl pH = 7.5, 500 mM NaCl, 10% glycerol, 1 mM beta-mercaptoethanol supplemented with 1 mg/ml lysozyme (Sigma-Aldrich L6876). The cell lysate was centrifuged at 16,000 g and processed through 0.22 µm filters. The lysate was applied to the 1 ml HisTrap HP Column (GE Healthcare) and DpbCas12e was eluted by 300 mM imidazole. After affinity chromatography the sample was either digested with TEV protease for 16 hours at 4°C to cleave the MBP-tag or applied straight on the Superdex200 Increase 10/300 GL (GE Healthcare) column equilibrated with a buffer containing 50 mM Hepes-HCl pH = 7.5, 500 mM NaCl, 1 mM DTT and 10% glycerol. Fractions containing DpbCas12e monomers were pooled and concentrated using 30 kDa Amicon Ultra-4 centrifugal unit (Merc Millipore, UFC803008) to a concertation of about 2 mg/mL, aliquoted and flash-frozen in liquid nitrogen with subsequent storage at −80°C. The purity of the protein was assessed by denaturing 10% PAGE.

In vitro DNA cleavage assays and samples preparation for HTS

DNA cleavage reactions were performed using the recombinant proteins MBP_DpbCas12e (or DpbCas12e where indicated), AsCas12a or SpCas9, *in vitro* synthesized guide RNAs and linear dsDNA targets. dsDNA targets were prepared by PCR amplifications of pUC19 plasmid (Targets 1, 2, 6), and human grin2b gene (Targets 3, 4, 5) using primers listed in the Supplementary Table S2. The full sequences of DNA targets are presented in the Supplementary Table S1.

Guide RNAs were synthesized *in vitro* using HiScribe T7 High Yield RNA Synthesis Kit (NEB, E20140). The sequences of guide RNAs used in this study are presented in the Supplementary Table S3. To pre-form an active ribonucleoprotein complexes the recombinant proteins were mixed with guide RNAs in 1x reaction buffer (20 mM HEPES, pH 7.5, 10 mM $MgCl_2$, 150 mM KCl, 1% glycerol, 1 mM DTT for DpbCas12e or 1x CutSmart buffer (NEB B7204 S, 1 mM DTT) and incubated at room temperature for 10 min. All RNAs used in this study are listed in Supplementary Table S3. Further the DNA targets were added to ribonucleoprotein complexes to the final concentration of the components in the cleavage reactions: 50 nM DNA, 400 nM recombinant protein (1600 nM in case of DpbCas12e), 2 µM guide RNA and 1x reaction buffer in 20 µl final volume. The DNA cleavage reaction mix was incubated at 37°C for 30 min. The reaction was stopped by the adding of 0.5 µl proteinase K (Thermo Fisher Scientific, EO0491) and the subsequent incubation for 30 min at 37°C. The DNA cleavage reaction products were separated by 1.5% agarose gel electrophoresis. The cleaved DNA fragments were isolated from the gel and purified by GeneJET Gel Extraction Kit (Thermo Fisher Scientific, K0692) (two DNA fragments produced by the nuclease cleavage were isolated from agarose gel as one mix in case of each sample). DNA fragments were eluted with 40 µl water.

End repair was performed using modified protocol for DNA blunting used in ChIP assays described by Blecher-Gonen et al. 25 µl of end repair buffer (1x T4 DNA ligase buffer (NEB, B0202), 0.1 mg ml-1 BSA, 0.1 mM dNTPs), 1 µl T4 PNK (Thermo Fisher Scientific, EK0031) and 1 µl T4 polymerase (NEB, M0203) were added to each 40 µl sample (total reaction volume was 67 µl). The reactions were incubated in a thermal cycler using the following settings: 15 min at 15°C, 15 min at 25°C, chilling to 4°C. Agencourt RNAClean XP (Beckman Coulter) beads in 2.5x ratio were used to cleanup the DNA after the end-repair reaction. The DNA fragments were eluted in 40 µl of 10 mM Tris-HCl (pH 8.0).

For A-base addition reaction 6 µl 10xNEB buffer 2 (NEB, B7002), 0.1 µl dATP (100 mM) and 10.9 µl of nuclease-free water were added to each 40 µl DNA sample volume. This yielded to the following final concentration: 1xNEB buffer 2 (NEB, B7002) and 167 µM dATP. 3 µl of Klenow Fragment (3′→5′ exo-) (NEB, M0212) were added to the samples and the reactions were incubated at 37°C for 30 min in the thermal cycler. Agencourt RNAClean XP (Beckman Coulter) beads in 2.5x ratio were used to cleanup the DNA after the reaction. The DNA fragments were eluted in 60 µl of 10 mM Tris-HCl (pH 8.0). The further sample preparation for high throughput sequencing (HTS) was performed using NEBNext Ultra II DNA Prep Kit for Illumina (NEB, E7645) starting with the adaptor ligation step. The samples were sequenced using Illumina platform with pair-end 150 cycles (75 + 75) or 300 cycles (150 + 150).

Computational Sequence analysis

HTS of each DNA cleavage reaction products produced in average 200 000 sequencing paired-end reads. To determine the cut sites positions of the cleaved DNA molecules the following pipeline was used. Forward and reverse reads (R1 and R2) were mapped against whole DNA target molecule using BWA (Li, Durbin, 2009). All unmapped reads were discarded from the analysis. The PAM sequence, a spacer and 10 nucleotides after the spacer were chosen as the region of interest. Reads matching the reference upstream the region of interest contained the information about target DNA strand (TS) cleavage position. Reads matching the reference downstream the region of interest contained information about non-target DNA strand (NTS) cleavage position. The number of analysed reads for each target is listed in Supplementary File S2 and Supplementary File S3. These reads were compared to the reference to determine the cut site positions. The number of nucleotides after the PAM sequence (for TS cleavage position determination) or before the end of the region of interest (for NTS cleavage position determination) were counted (Supplementary File S2, Supplementary File S3). To estimate the range of possible lengths of overhangs and calculate the maximal probabilities of their generation, distances between DSBs were computed as

Distance = [cut site position on TS] – [cut site position on NTS]. Distances were calculated between all possible TS and NTS DNA cleavage positions. Relative frequencies of generated overhangs were calculated as a sum of [relative frequency of cut site positions on TS] x [relative frequency of cut site positions on NTS] for all combinations of TS and NTS producing the overhangs of certain length.

Results

Mapping of DpbCas12e and AsCas12a cut sites

For precise determination of Cas12e DNA cut site positions and comparison with Cas12a, we performed DNA cleavage reactions of several different targets *in vitro*. Recombinant versions of wild-type DpbCas12e from *Deltaproteobacteria* and AsCas12a from *Acidaminococcus sp*. were purified from *E. coli* cells (Supplementary Fig. S1). Due to a poor solubility of DpbCas12e, it was purified as an N-terminal fusion to MBP (maltose binding protein). The widely used SpCas9 nuclease from *Streptococcus pyogenes*, which predominantly produces double-strand DNA breaks with blunt endings, was used as a control. DpbCas12e, AsCas12a, and SpCas9 effector nucleases require different protospacer adjacent motifs (PAMs) for efficient target recognition and cleavage. DpbCas12e and AsCas12a require upstream PAMs with, respectively, 5′-TTCN-3′ and 5′-TTTV-3′ consensus; the SpCas9 PAM is 5′-NGG-3′ and is located downstream of the target. To compare the DpbCas12e and AsCas12a DNA cut site positions, we used DNA targets with a 'TTTCCN' sequence at the 5′ flank, which allows recognition of almost the same target by both enzymes (shifted by one nucleotide). Downstream, the targets were flanked by the SpCas9 PAM. Thus, all three Cas proteins were able to recognize almost identical DNA target sequences (shifted by several nucleotides for efficient recognition of cognate PAMs).

Six different 20-nt targets with an appropriate PAM were chosen randomly (Supplementary Table S1) and used to study cleavage by DpbCas12e or AsCas12a loaded with appropriate sgRNAs. Incubation of target-bearing linear DNA fragments with DpbCas12e-sgRNA or AsCas12a-crRNA led to efficient DNA cleavage and generation of products of expected lengths. The SpCas9-sgRNA ribonucleoprotein complex cleaved only four out of six DNA targets. All successful DNA cleavage reactions were processed according to procedure outlined in Fig. 2 to determine cut site positions using high throughput sequencing on Illumina platform. In brief, the cleavage products were separated from uncleaved DNA, blunt-ended with T4 polynucleotide kinase and T4 polymerase, which produces 5′-P and 3′-OH ends and fills-in 5′-overhangs, respectively. Next, DNA was 3′ A-tailed using the exonuclease deficient Klenow DNA polymerase fragment. This allowed to avoid unwanted processing/modification of DNA ends formed after cleavage prior to sequencing. For further steps we used components of the NEBNext Ultra II kit starting from adaptors ligation step and bypassing the DNA ends preparation step. In brief, Illumina NEBNext adaptors were ligated to DNA through TA cloning. Next, uracil residues incorporated in the adaptors were cut out by the USER enzyme, a component of NEBNext Ultra II which combines uracil DNA glycosylase and endonuclease VIII activity. Next, the samples were barcoded, according to the NEBNext Ultra II protocol, and sequenced on Illumina platform. The distribution of cleavage sites was revealed by analysing on average 100 000 reads for each cleavage reaction.

The analysis of the results showed that, as expected, unlike SpCas9, which mainly generated blunt-end cut sites, DpbCas12e and AsCas12a produced DNA cleavage products with staggered ends (Fig. 2A). The use of HTS allowed us to observe not only the most frequent cleavage sites but distributions of cut site positions. In agreement with previous studies [27,28] multiple cleavage sites by AsCas12a were observed – the nuclease predominantly cleaved DNA after nucleotide positions 17–19 downstream of PAM in the non-target DNA strand (further NTS) and after positions 21–23 in the target strand (further TS) (Fig. 3A, Supplementary Fig. S2).

Due to flexibility in cleavage sites, Cas12a cuts exhibit a wide distribution of 2–6nt overhangs within each target (Fig. 3B). Similarly, DpbCas12e predominantly cleaved DNA after nucleotides 17–19 downstream of its PAM in the NTS (Fig. 3A, Supplementary Fig. S2). In neither of the six DNA targets tested did we detect NTS cleavage after positions 12–14 position, previously reported by Jun-Jie Liu et al. In the target strand, DpbCas12e predominantly cleaved after the 22nd position.

We also performed AsCas12a and DpbCas12e DNA cleavage reaction of the same 20nt protospacer sequence used by Jun-Jie Liu et al, incorporated into a long 500nt linear DNA target (Supplementary Table S1). The cleavage of this target also resulted in producing 3–5nt overhangs for both nucleases (Supplementary Fig. S3). It is important to mention, that Jun-Jie Liu et al. used a much shorter DNA fragment, which could possibly affect the cleavage position [14]. Overall, we conclude that the DpbCas12e nuclease produces 3–5nt 5′-overhangs (Fig. 3B).

Since experiments described above were conducted with MBP fusion of DpbCas12e, we determined the DNA cleavage pattern of DpbCas12e without MBP. As a result, no significant differences between DpbCas12e without MBP and the MBP-fused protein were observed (Supplementary Fig. S4).

The influence of sgRNA spacer segment length on DpbCas12e cleavage sites

Previous studies of Cas12a showed that the spacer length of crRNA can influence the position of Cas12a DNA cleavage sites [20]. Using crRNA with a spacer length of 18nt caused a shift of cut site position in NTS to positions 13–15 instead of position 18 observed when Cas12a effector was charged with crRNAs whose spacer segments were 20 nucleotides or longer [20]. Due to DNA cleavage pattern similarity between Cas12a and Cas12e, we were interested to determine if sgRNA spacer length can also affect the position of DNA cleavage site by DpbCas12e. DpbCas12e charged with sgRNAs of different spacer length (16, 18, 20, 22 and 24nt) was used to cleave three different DNA targets. Analysis of cleavage products by HTS showed that DpbCas12e begins to cleave NTS closer to PAM when shorter than 20nt spacer sgRNAs are used: 16nt spacer sgRNAs produced 6–8nt 5′-overhangs, while sgRNAs with 20, 22, and 24nt spacer segments led to cleavage after target positions 17–19, producing 3–5nt overhangs (Fig. 4A–B, Supplementary File S3).

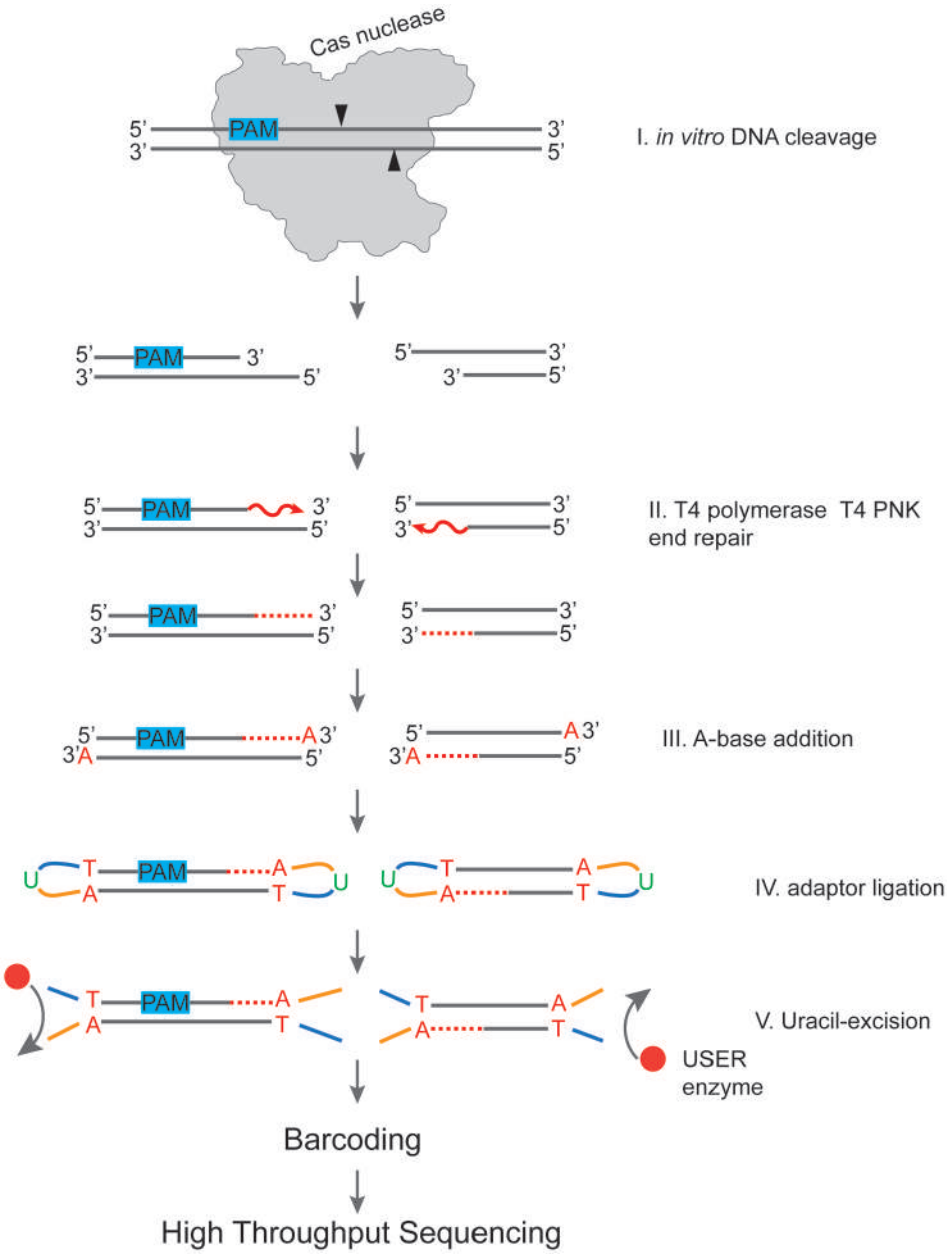


Figure 2. A workflow of sample preparation for determination of positions of DNA cleavage sites produced by Cas nucleases *in vitro*.
The cleaved DNA fragments generated by Cas nuclease during *in vitro* DNA cleavage reaction (Step I) are blunted using T4 PNK and T4 DNA polymerase (Step II). A-base is added to 3′ ends (Step III) for further ligation of Illumina NEBNext sequencing adaptors containing uridine (Step IV). Uridines are cleaved out using the NEB USER enzyme, which combines uracil DNA glycosylase and endonuclease VIII activity. Next, the samples are barcoded to produce DNA libraries ready for high throughput sequencing.

The reduction of the sgRNA spacer length may compromise the DNA cleavage efficiency of the effector. Indeed, using of sgRNAs with spacer lengths shorter than 20 nt led to lower DpbCas12e DNA cleavage efficiency, although the effect was not dramatic and the nuclease activity was sufficient for effective introduction of double-stranded breaks (Fig. 5).

Discussion

In this work we determined DNA cleavage sites produced by DpbCas12e in *in vitro* reactions. Analysis of cleaved DNA fragments by HTS allowed us to show not only the most frequent cut site positions but revealed the distribution of DNA cleavage sites along the targeted DNA molecules. In agreement with the previous results, our data demonstrates that in contrast to SpCas9, DpbCas12e and AsCas12a produce staggered ends at cut sites. The DNA cleavage positions in case of each nuclease slightly vary depending on the target sequence, though the overall cleavage pattern remains the same. We show that DpbCas12e and AsCas12a, CRISPR-Cas effectors of distinct subtypes V-E and V-A, have similar distribution of DNA cut site positions. Both enzymes introduce cuts after nucleotides 17–19 downstream of PAM in the non-target

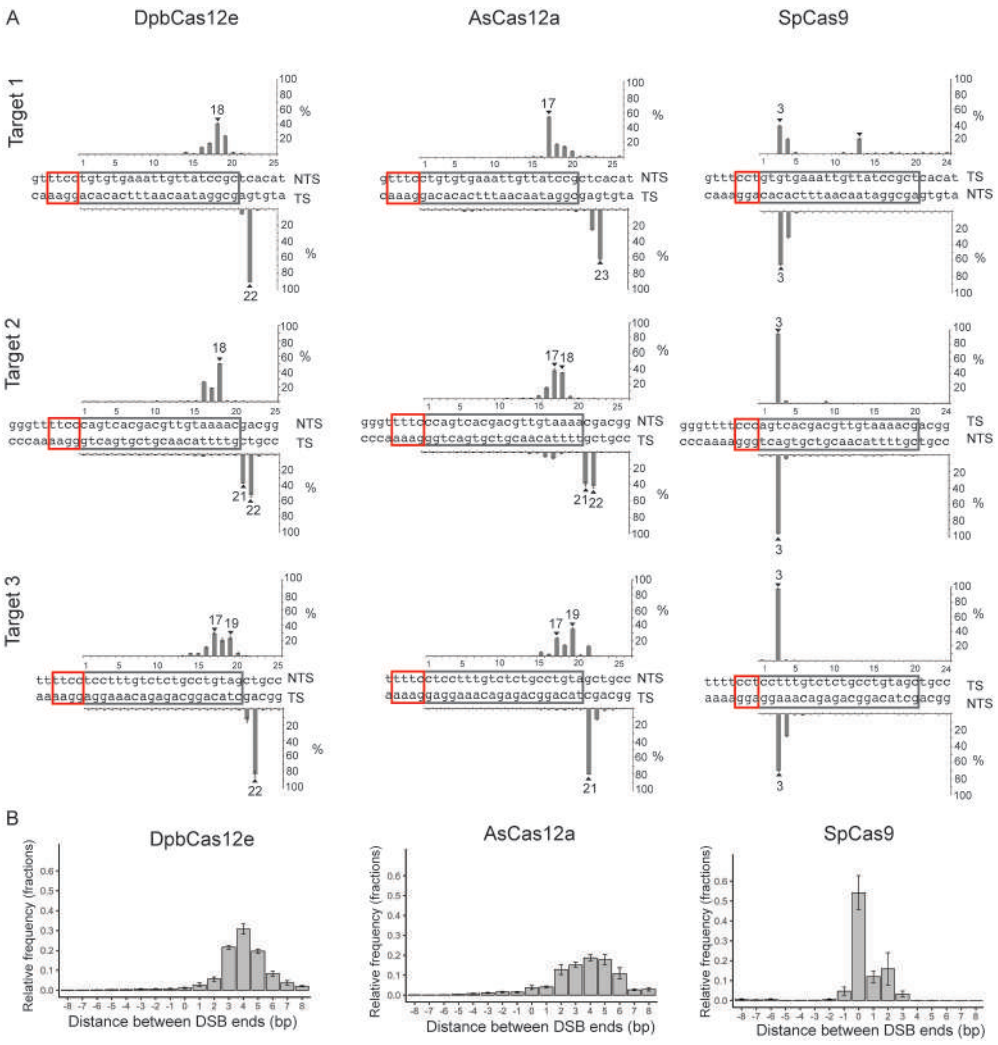


Figure 3. Determination of DpbCas12e, AsCas12a, and SpCas9 cut sites positions by high throughput sequencing of *in vitro* DNA cleavage reaction products. A) Histograms showing mapping the cut sites on the target and non-target DNA strands in case of DpbCas12e, AsCas12a, and SpCas9. Results for three different DNA targets are shown, PAMs are indicated with red rectangles. The numbering of nucleotide positions starts from the end of the PAM and is shown along the target DNA sequence. For each DNA target sequence, the histograms of cut site positions frequency in percentage for the corresponding DNA strand are shown. Each column represents the fraction of DNA cleavage events after the corresponded nucleotide. 'NTS' stands for non-target DNA strand, 'TS' – for target DNA strand. The most frequent cut site positions are shown with black triangles. Mean values obtained from three independent experiments with standard deviations are shown. B) The overhang lengths produced by DpbCas12e, AsCas12a, and SpCas9. Histograms of differences between cut site positions on DNA TS and NTS were calculated based on DNA cleavage data obtained for all six targets (panel A and Supplementary Fig. S2). Distances between DSB ends were calculated as Distance = [cut site position on TS] – [cut site position on NTS]. Distances were calculated between all possible TS and NTS DNA cleavage positions. Relative frequencies of generated overhangs were calculated as a sum of [relative frequency of cut site positions on TS] x [relative frequency of cut site positions on NTS] for all combinations of TS and NTS producing overhangs of a certain length.

DNA strand. The target DNA strand cleavage position is after nucleotides 21–23 for AsCas12a and predominantly after nucleotide 22 for DpbCas12e. This DNA cleavage pattern leads to generation of 3–5nt overhangs by both enzymes. We did not observe 10nt 5′-overhangs in DNA cleaved by DpbCas12e, that were reported earlier, and that could have been advantageous for certain biotechnological applications.

Nevertheless, we show that the length of 5′-overhangs generated by DpbCas12e can be modulated: using of sgRNAs with shorter, 16 nucleotides spacer segment increased the 5′-overhangs of cleavage products by 3nt, producing long 6–8nt staggered overhangs instead of 3–5nt overhangs, produced when 20nt spacer segment was used. Longer overhangs generated at cut sites may be potentially advantageous for *in vitro* and *in vivo* ligation of DNA fragments into Cas12e-generated breaks in double-stranded DNA. Indeed, this strategy was successfully used by Chao Lei et al. for incorporation of DNA fragments into plasmids *in vitro* using Cas12a nuclease [20].

Overall, our data on cut site position determination can be useful for potential development of DpbCas12e-based biotechnology instruments as well as for understanding of the mechanisms of Cas12e nucleases action in molecular details. The approach applied here for DNA cleavage pattern determination is general and can be used for characterization of cleavage sites by different Cas and non-Cas nucleases.

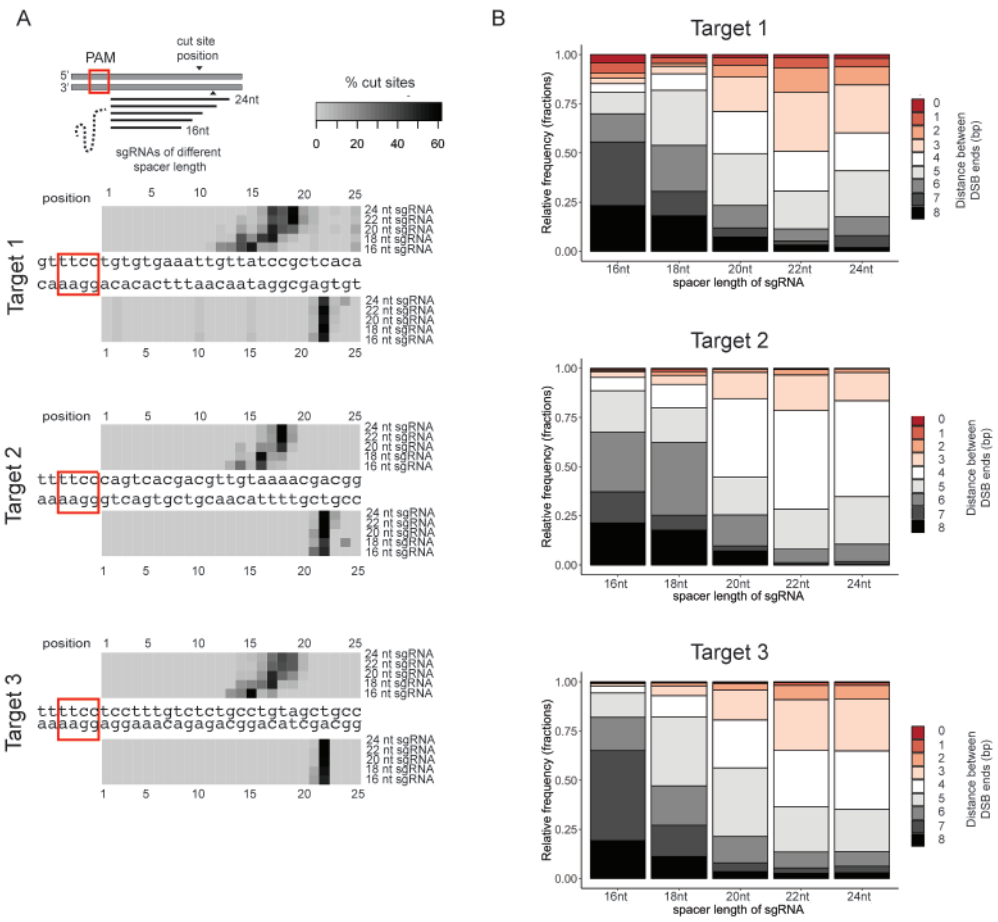


Figure 4. The influence of sgRNA spacer length on cleavage by DpbCas12e. A) Results of mapping of *in vitro* cleavage sites produced in three different DNA targets by DpbCas12e complexed with sgRNAs of different spacer lengths, PAMs are indicated with red rectangles; the numbering of nucleotide positions starts from the end of the PAM and is shown along the target DNA sequence. Heatmaps for each DNA target sequence show the positions of cut sites for the corresponding DNA strand for sgRNAs with indicated spacer lengths. Each heatmap cell intensity represents the fraction of DNA cleavage events after the corresponding nucleotide. The heatmaps are drawn based on mean values obtained from three independent experiments. B) The range of lengths of 5′-overhangs, produced by DpbCas12e in complex with sgRNAs with 16, 18, 20, 22, or 24nt spacers. The differences between cut site positions on DNA TS and NTS were calculated based on data shown in panel A. Distance between DSB ends was calculated as Distance = [cut site position on TS] – [cut site position on NTS]. Distances were calculated between all possible TS and NTS DNA cleavage positions. Relative frequencies of generated overhangs were calculated as a sum of [relative frequency of cut site positions on TS] x [relative frequency of cut site positions on NTS] for all combinations of TS and NTS producing overhangs of a certain length. The most abundant distances of 0–8nt were used to plot a stacked barchart. The length of each sector of the columns represents the fraction of overhangs of certain length.

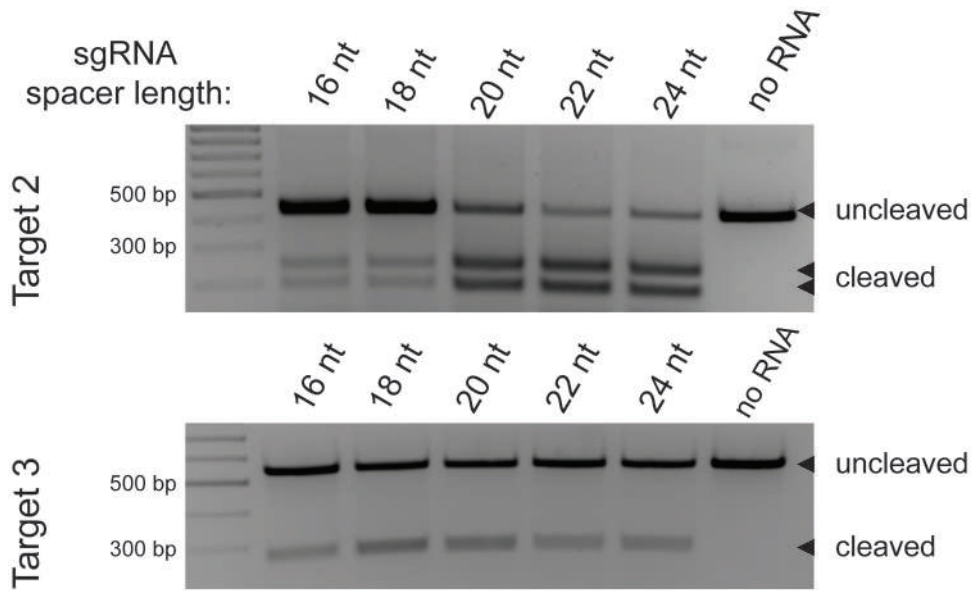


Figure 5. DpbCas12e DNA *in vitro* cleavage using sgRNAs of different spacer length. Above – a gel showing the results of *in vitro* cleavage of Target 2 using sgRNAs of 16nt, 18nt, 20nt, 22nt or 24nt spacer length. Below – similar gel for Target 3.

Acknowledgments

We thank Aleksandr Koshkin for insightful comments and suggestions. We are grateful to the Skoltech Genomics Core facility for sequencing.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Ministry of Science and Higher Education of the Russian Federation under Grant 075-15-2019-1661, as well as the Russian Science Foundation Grant 19-14-00323 to K.S.

Data availability

Raw sequencing data have been deposited with the National Center for Biotechnology Information Sequence Read Archive under BioProject ID PRJNA605170

ORCID

Georgii Pobegalov http://orcid.org/0000-0003-0836-0732
Olga Musharova http://orcid.org/0000-0003-2496-0420
Iana Fedorova http://orcid.org/0000-0001-6144-173X

References

[1] Wang H, La Russa M, Qi LS. CRISPR/Cas9 in genome editing and beyond. Annu Rev Biochem. 2016;85:227–264.
[2] Jinek M, Chylinski K, Fonfara I, et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012;337:816–821.
[3] Cong L, Ran FA, Cox D, et al. Multiplex genome engineering using CRISPR/Cas systems. Science. 2013;339:819–823.
[4] Makarova KS, Haft DH, Barrangou R, et al. Evolution and classification of the CRISPR–Cas systems. Nat Rev Microbiol. 2011;9:467–477.
[5] Ran FA, Cong L, Yan WX, et al. In vivo genome editing using Staphylococcus aureus Cas9. Nature. 2015;520:186–191.
[6] Kim E, Koo T, Park SW, et al. In vivo genome editing with a small Cas9 orthologue derived from Campylobacter jejuni. Nat Commun. 2017;8:14500.
[7] Lee CM, Cradick TJ, Bao G. The neisseria meningitidis CRISPR-Cas9 system enables specific genome editing in mammalian cells. Mol Ther. 2016;24:645–654.
[8] Harrington LB, Paez-Espino D, Staahl BT, et al. A thermostable Cas9 with increased lifetime in human plasma. Nat Commun. 2017;8:1424.
[9] Strecker J, Ladha A, Gardner Z, et al. RNA-guided DNA insertion with CRISPR-associated transposases. Science. 2019;365:48–53.
[10] Abudayyeh OO, Gootenberg JS, Essletzbichler P, et al. RNA targeting with CRISPR–Cas13. Nature. 2017;550:280–284.
[11] Fonfara I, Richter H, Bratovič M, et al. The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. Nature. 2016;532:517–521.
[12] Zetsche B, Heidenreich M, Mohanraju P, et al. Multiplex gene editing by CRISPR–Cpf1 using a single crRNA array. Nat Biotechnol. 2017;35:31–34.
[13] Burstein D, Harrington LB, Strutt SC, et al. New CRISPR–Cas systems from uncultivated microbes. Nature. 2017;542:237–241.
[14] Liu -J-J, Orlova N, Oakes BL, et al. CasX enzymes comprise a distinct family of RNA-guided genome editors. Nature. 2019;566:218–223.
[15] Yamano T, Nishimasu H, Zetsche B, et al. Crystal structure of Cpf1 in complex with guide RNA and target DNA. Cell. 2016;165:949–962.
[16] Nishimasu H, Ran FA, Hsu PD, et al. Crystal structure of Cas9 in complex with guide RNA and target DNA. Cell. 2014;156:935–949.
[17] Swarts DC, van der Oost J, Jinek M. Structural Basis for Guide RNA Processing and Seed-Dependent DNA Targeting by CRISPR-Cas12a. Molecular Cell. 2017;66:221-233.e4. doi:10.1016/j.molcel.2017.03.016
[18] Yan WX, Mirzazadeh R, Garnerone S, Scott D, Schneider MW, Kallas T, Custodio J, Wernersson E, Li Y, Gao L, et al. BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. Nat Commun. 2017;8:15058.
[19] Wienert B, Wyman SK, Richardson CD, et al. Unbiased detection of CRISPR off-targets in vivo using DISCOVER-Seq. Science. 2019;364:286–289.
[20] Lei C, Li S-Y, Liu J-K, Zheng X, Zhao G-P, Wang J. The CCTL (Cpf1-assisted Cutting and Taq DNA ligase-assisted Ligation) method for efficient editing of large DNA constructs in vitro. Nucleic Acids Res. 2017;6:e74.
[21] Anzalone AV, Randolph PB, Davis JR, et al. Search-and-replace genome editing without double-strand breaks or donor DNA. Nature. 2019;576:149–157.
[22] Zetsche B, Gootenberg JS, Abudayyeh OO, et al. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell. 2015;163:759–771.
[23] Teng F, Cui T, Feng G, et al. Repurposing CRISPR-Cas12b for mammalian genome engineering. Cell Discov. 2018;4:63.
[24] Kim D, Bae S, Park J, et al. Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. Nat Methods. 2015;12:237–243.
[25] Tsai SQ, Nguyen NT, Malagon-Lopez J, et al. CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets. Nat Methods. 2017;14:607–614.
[26] Cameron P, Fuller CK, Donohoue PD, et al. Mapping the genomic landscape of CRISPR–Cas9 cleavage. Nat Methods. 2017;14:600–606.
[27] Li S-Y, Zhao G-P, Wang J. C-Brick: a new standard for assembly of biological parts using Cpf1. ACS Synth Biol. 2016;5:1383–1388.
[28] Kim D, Kim J, Hur JK, et al. Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells. Nat Biotechnol. 2016;34:863–868.

Supplementary File S1

Supplementary Table S1

DNA sequences used in this study

pet21a_SpCas9	https://benchling.com/s/seq-IUEr7evPHwi0w8wSjLYU
pet21a_AsCas12a	https://benchling.com/s/seq-o1hUaNe9J5QrAonvF9hX
pet21a_MBP_Cas12e	https://benchling.com/s/seq-ZqqROfpn1swlsEwJYOlw
Target_1 full sequence	https://benchling.com/s/seq-9gSboMXOppMXHzGB2CEl
Target_2 full sequence	https://benchling.com/s/seq-YYU1ovv8gXzEasmzunz2
Target_3 full sequence	https://benchling.com/s/seq-8xzpBhaAWUuaMISt8RY6
Target_4 full sequence	https://benchling.com/s/seq-v2trp0ztUvueLXEgPVj2
Target_5 full sequence	https://benchling.com/s/seq-PdVLoAj9Ot7puZkbAvRZ
Target_6 full sequence	https://benchling.com/s/seq-tjpease80QL6K9r4rtkg
Target_with_protospacer from Jun-Jie Liu et al. 2019	https://benchling.com/s/seq-UafrBPV3EwW0vl8izGpZ

Supplementary Table S2

Primers used for dsDNA targets amplification

Target 1 primers	forward	CGATTCATTAATGCAGCTGGCACG
	reverse	GTAAAACGACGGCCAGT
Target 2 primers	forward	CTTTATGCTTCCGGCTCG
	reverse	GCGGCATCAGAGCAGATTGTAC
Target 3 primers	forward	TTCTGGCTGTTGTCCTCATTGAG
	reverse	CATCTTCAACTCGTCGACTCC
Target 4 primers	forward	GAGAGAGATGGCCAAGGCTT
	reverse	CTATTACACTACGTGGAACTGCC
Target 5 primers	forward	AACATGCTCTTTCTTTGTGTTTGC
	reverse	CTCCCTGCAGCCCCTTTTTAC
Target 6 primers	forward	CGCCTTTGAGTGAGCTGATACCGC
	reverse	GTAAAACGACGGCCAGT

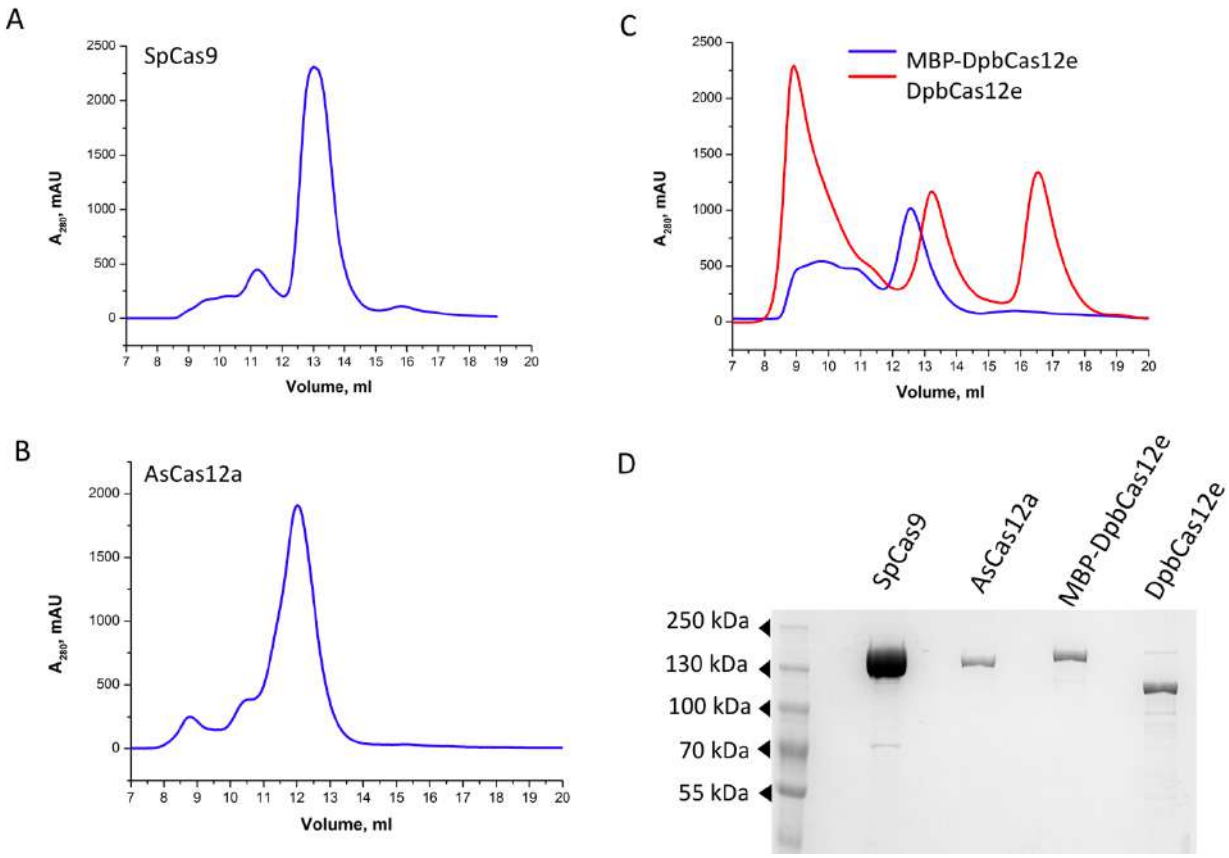
Supplementary Table S3

RNAs used in this study

Single guide RNAs DpbCas12e	
DpbCas12e sgRNA Target 1	GGGCGCGTTTATTCCATTACTTTGGAGCCAGTCCCAGCGACT ATGTCGTATGGACGAAGCGCTTATTTATCGGAGAGAAACCGA TAAGTAAAACGCATCAAAGtgtgtgaaattgttatccgc
DpbCas12e sgRNA Target 2	GGGCGCGTTTATTCCATTACTTTGGAGCCAGTCCCAGCGACT ATGTCGTATGGACGAAGCGCTTATTTATCGGAGAGAAACCGA TAAGTAAAACGCATCAAAGcagtcacgacgttgtaaaac
DpbCas12e sgRNA Target 3	GGGCGCGTTTATTCCATTACTTTGGAGCCAGTCCCAGCGACT ATGTCGTATGGACGAAGCGCTTATTTATCGGAGAGAAACCGA TAAGTAAAACGCATCAAAGtcctttgtctctgcctgtag
DpbCas12e sgRNA Target 4	GGGCGCGTTTATTCCATTACTTTGGAGCCAGTCCCAGCGACT ATGTCGTATGGACGAAGCGCTTATTTATCGGAGAGAAACCGA TAAGTAAAACGCATCAAAGtgatgccatcctatagtcgt
DpbCas12e sgRNA Target 5	GGGCGCGTTTATTCCATTACTTTGGAGCCAGTCCCAGCGACT ATGTCGTATGGACGAAGCGCTTATTTATCGGAGAGAAACCGA TAAGTAAAACGCATCAAAGaccatctctccgtggtaccc
DpbCas12e sgRNA Target 6	GGGCGCGTTTATTCCATTACTTTGGAGCCAGTCCCAGCGACT ATGTCGTATGGACGAAGCGCTTATTTATCGGAGAGAAACCGA TAAGTAAAACGCATCAAAGcgactggaaagcgggcagtg
DpbCas12e sgRNA Target 1 16nt	GGGCGCGTTTATTCCATTACTTTGGAGCCAGTCCCAGCGACT ATGTCGTATGGACGAAGCGCTTATTTATCGGAGAGAAACCGA TAAGTAAAACGCATCAAAGtgtgtgaaattgttat
DpbCas12e sgRNA Target 1 18nt	GGGCGCGTTTATTCCATTACTTTGGAGCCAGTCCCAGCGACT ATGTCGTATGGACGAAGCGCTTATTTATCGGAGAGAAACCGA TAAGTAAAACGCATCAAAGtgtgtgaaattgttatcc
DpbCas12e sgRNA Target 1 20nt	GGGCGCGTTTATTCCATTACTTTGGAGCCAGTCCCAGCGACT ATGTCGTATGGACGAAGCGCTTATTTATCGGAGAGAAACCGA TAAGTAAAACGCATCAAAGtgtgtgaaattgttatccgc
DpbCas12e sgRNA Target 1 22nt	GGGCGCGTTTATTCCATTACTTTGGAGCCAGTCCCAGCGACT ATGTCGTATGGACGAAGCGCTTATTTATCGGAGAGAAACCGA TAAGTAAAACGCATCAAAGtgtgtgaaattgttatccgctc

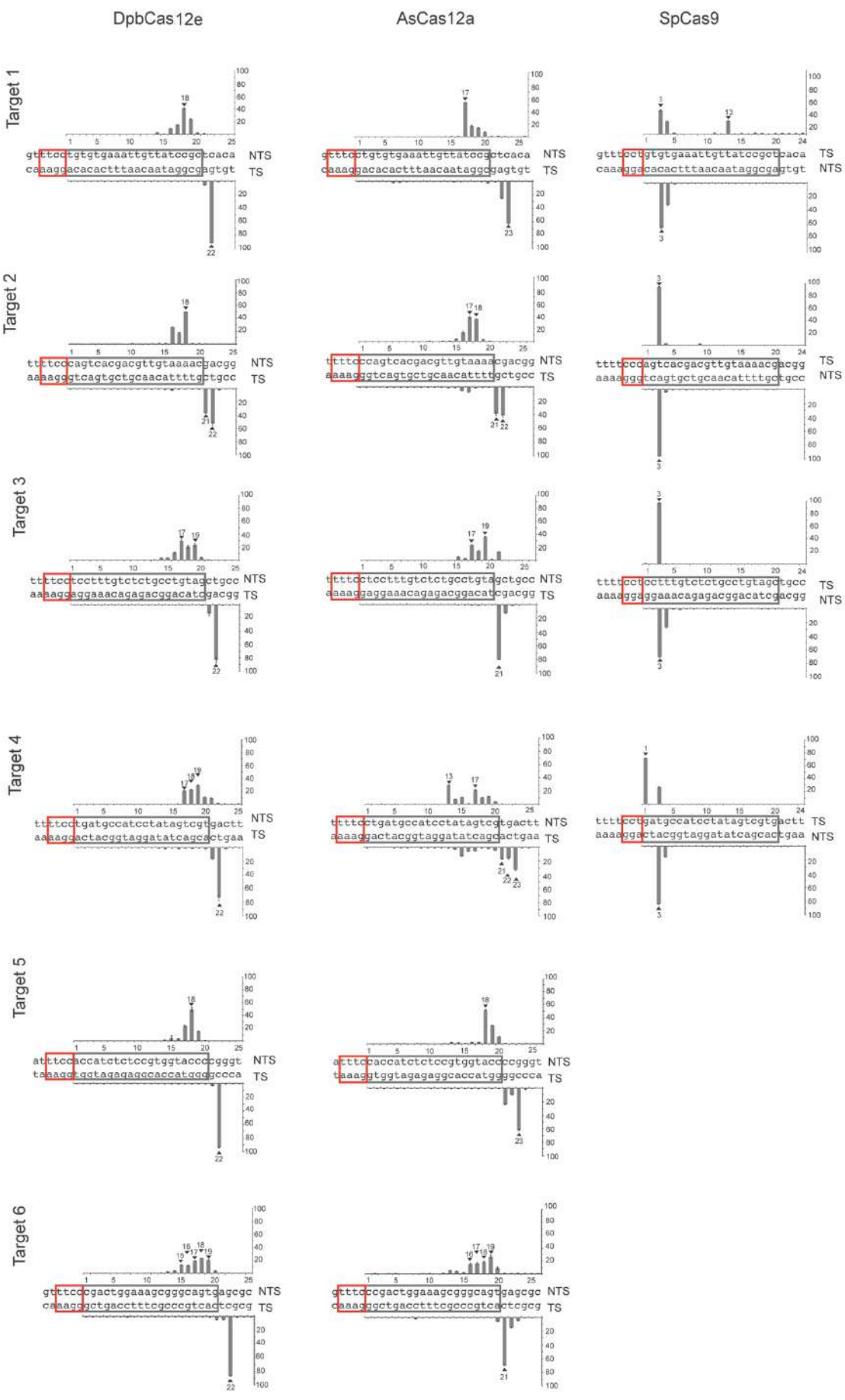
DpbCas12e sgRNA Target 1 24nt	GGGCGCGTTTATTCCATTACTTTGGAGCCAGTCCCAGCGACT ATGTCGTATGGACGAAGCGCTTATTTATCGGAGAGAAACCGA TAAGTAAAACGCATCAAAGtgtgtgaaattgttatccgctca c
DpbCas12e sgRNA Target 2 16nt	GGGCGCGTTTATTCCATTACTTTGGAGCCAGTCCCAGCGACT ATGTCGTATGGACGAAGCGCTTATTTATCGGAGAGAAACCGA TAAGTAAAACGCATCAAAGcagtcacgacgttgta
DpbCas12e sgRNA Target 2 18nt	GGGCGCGTTTATTCCATTACTTTGGAGCCAGTCCCAGCGACT ATGTCGTATGGACGAAGCGCTTATTTATCGGAGAGAAACCGA TAAGTAAAACGCATCAAAGcagtcacgacgttgtaaa
DpbCas12e sgRNA Target 2 20nt	GGGCGCGTTTATTCCATTACTTTGGAGCCAGTCCCAGCGACT ATGTCGTATGGACGAAGCGCTTATTTATCGGAGAGAAACCGA TAAGTAAAACGCATCAAAGcagtcacgacgttgtaaaac
DpbCas12e sgRNA Target 2 22nt	GGGCGCGTTTATTCCATTACTTTGGAGCCAGTCCCAGCGACT ATGTCGTATGGACGAAGCGCTTATTTATCGGAGAGAAACCGA TAAGTAAAACGCATCAAAGcagtcacgacgttgtaaaacga
DpbCas12e sgRNA Target 2 24nt	GGGCGCGTTTATTCCATTACTTTGGAGCCAGTCCCAGCGACT ATGTCGTATGGACGAAGCGCTTATTTATCGGAGAGAAACCGA TAAGTAAAACGCATCAAAGcagtcacgacgttgtaaaacgac g
DpbCas12e sgRNA Target 3 16nt	GGGCGCGTTTATTCCATTACTTTGGAGCCAGTCCCAGCGACT ATGTCGTATGGACGAAGCGCTTATTTATCGGAGAGAAACCGA TAAGTAAAACGCATCAAAGtcctttgtctctgcct
DpbCas12e sgRNA Target 3 18nt	GGGCGCGTTTATTCCATTACTTTGGAGCCAGTCCCAGCGACT ATGTCGTATGGACGAAGCGCTTATTTATCGGAGAGAAACCGA TAAGTAAAACGCATCAAAGtcctttgtctctgcctgt
DpbCas12e sgRNA Target 3 20nt	GGGCGCGTTTATTCCATTACTTTGGAGCCAGTCCCAGCGACT ATGTCGTATGGACGAAGCGCTTATTTATCGGAGAGAAACCGA TAAGTAAAACGCATCAAAGtcctttgtctctgcctgtag
DpbCas12e sgRNA Target 3 22nt	GGGCGCGTTTATTCCATTACTTTGGAGCCAGTCCCAGCGACT ATGTCGTATGGACGAAGCGCTTATTTATCGGAGAGAAACCGA TAAGTAAAACGCATCAAAGtcctttgtctctgcctgtagct
DpbCas12e sgRNA Target 3 24nt	GGGCGCGTTTATTCCATTACTTTGGAGCCAGTCCCAGCGACT ATGTCGTATGGACGAAGCGCTTATTTATCGGAGAGAAACCGA TAAGTAAAACGCATCAAAGtcctttgtctctgcctgtagctg c

crRNAs AsCas12a	
AsCas12a crRNA Target 1	GGGTAATTTCTACTCTTGTAGATctgtgtgaaattgttatccg
AsCas12a crRNA Target 2	GGGTAATTTCTACTCTTGTAGATccagtcacgacgttgtaaaa
AsCas12a crRNA Target 3	GGGTAATTTCTACTCTTGTAGATctcctttgtctctgcctgta
AsCas12a crRNA Target 4	GGGTAATTTCTACTCTTGTAGATctgatgccatcctatagtcg
AsCas12a crRNA Target 5	GGGTAATTTCTACTCTTGTAGATcaccatctctccgtggtacc
AsCas12a crRNA Target 6	GGGTAATTTCTACTCTTGTAGATccgactggaaagcgggcagt
Single guide RNAs SpCas9	
SpCas9 sgRNA Target 1	GGGagcggataacaatttcacacGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCT
SpCas9 sgRNA Target 2	GGGcgttttacaacgtcgtgactGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCT
SpCas9 sgRNA Target 3	GGGgctacaggcagagacaaaggGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCT
SpCas9 sgRNA Target 4	GGGcacgactataggatggcatcGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCT
SpCas9 sgRNA Target 5	GGGggggtaccacggagagatggGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCT
SpCas9 sgRNA Target 6	GGGtcactgcccgctttccagtcGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCT



Supplementary Figure S1. Purification of recombinant SpCas9, AsCas12a, MBP-DpbCas12e and DpbCas12e proteins.

A. Size exclusion chromatography elution of SpCas9 protein. The fractions numbers are written along x-axis.
B. Size exclusion chromatography elution of AsCas12a protein. The fractions numbers are written along x-axis.
C. Size exclusion chromatography elution of MBP-DpbCas12e (red line) and DpbCas12e (blue line) proteins. The fractions numbers are written along x-axis.
D. SDS PAAG gel electrophoresis of purified SpCas9, AsCas12a, MBP-DpbCas12e and DpbCas12e proteins.


DpbCas12e
AsCas12a
SpCas9
Target 1
Target 2
Target 3
Target 4
Target 5
Target 6
NTS
TS

Chapter VI

Detection of spacer precursors formed *in vivo* during primed CRISPR adaptation

Introduction:

In this chapter we studied adaptation – the acquisition of new spacers into CRISPR array. Although it was established that adaptation process is conducted by the Cas1-Cas2 complex, a lot of details of the process remain to be unveiled. In this work we used host-genome-targeting CRISPR-Cas Type I-E system of *E. coli* as a model of primed adaptation, acquisition of new spacers, mediated by CRISPR-Cas interference. We extracted short DNA fragments from cells undergoing the adaptation process and sequenced them using FragSeq (strand-specific, high-throughput sequencing of DNA fragments), a procedure developed by my fellow student Anna Shiriaeva. We found that fragments accumulating in a bacterium carry a PAM sequence, have a length of about 35 nt and share a common asymmetrical structure with 3'-overhang on the PAM-derived end. This suggests that these fragments are "prespacers" - intermediates of adaptation between the protospacer selection and spacer integration steps.

Contribution

In this work I assessed how CRISPR-Cas self-targeting affect the growth of *E. coli*: measured the growth rates of the cultures (Figure 1b). The first author of the paper, Anna Shiriaeva, and the corresponding authors wrote the manuscript.

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-12417-w OPEN

Detection of spacer precursors formed in vivo during primed CRISPR adaptation

Anna A. Shiriaeva[1,2,3], Ekaterina Savitskaya[1,4], Kirill A. Datsenko[3], Irina O. Vvedenskaya[5], Iana Fedorova[1,2], Natalia Morozova[1,2], Anastasia Metlitskaya[4], Anton Sabantsev[2], Bryce E. Nickels[5*], Konstantin Severinov[1,2,3,4*] & Ekaterina Semenova[3*]

Type I CRISPR-Cas loci provide prokaryotes with a nucleic-acid-based adaptive immunity against foreign DNA. Immunity involves adaptation, the integration of ~30-bp DNA fragments, termed prespacers, into the CRISPR array as spacers, and interference, the targeted degradation of DNA containing a protospacer. Interference-driven DNA degradation can be coupled with primed adaptation, in which spacers are acquired from DNA surrounding the targeted protospacer. Here we develop a method for strand-specific, high-throughput sequencing of DNA fragments, FragSeq, and apply this method to identify DNA fragments accumulated in *Escherichia coli* cells undergoing robust primed adaptation by a type I-E or type I-F CRISPR-Cas system. The detected fragments have sequences matching spacers acquired during primed adaptation and function as spacer precursors when introduced exogenously into cells by transformation. The identified prespacers contain a characteristic asymmetrical structure that we propose is a key determinant of integration into the CRISPR array in an orientation that confers immunity.

[1] Center of Life Sciences, Skolkovo Institute of Science and Technology, 1 Nobel St., Moscow 121205, Russia. [2] Peter the Great St. Petersburg Polytechnic University, 29 Polytechnicheskaya St., St. Petersburg 195251, Russia. [3] Department of Molecular Biology and Biochemistry, Waksman Institute, Rutgers University, 190 Frelinghuysen Rd., Piscataway, NJ 08854, USA. [4] Institute of Molecular Genetics, Russian Academy of Sciences, 2 Akademika Kurchatova Sq., Moscow 123182, Russia. [5] Department of Genetics, Waksman Institute, Rutgers University, 190 Frelinghuysen Rd., Piscataway, NJ 08854, USA. *email: bnickels@waksman.rutgers.edu; severik@waksman.rutgers.edu; semenova@waksman.rutgers.edu

CRISPR interference in the *Escherichia coli* type I-E system is performed by the Cascade complex, composed of a crRNA and several Cas proteins[1–3]. Initial binding of Cascade to a protospacer flanked by a 3-bp protospacer adjacent motif (PAM)[4] results in the formation of an R-loop containing an RNA–DNA heteroduplex formed between the crRNA and target strand, and extrusion of single-stranded DNA derived from the nontarget strand[2,5–10]. Cas3, a single-stranded nuclease and 3′–5′ helicase, is recruited to the Cascade–protospacer complex and cleaves the nontarget strand to initiate unwinding and degradation of the targeted DNA[6,10,11]. In vitro, Cas3 can translocate on DNA as a component of a larger complex that includes Cascade and the key proteins of CRISPR adaptation, Cas1 and Cas2 [12].

CRISPR adaptation in the *E. coli* I-E system is mediated by a Cas1–Cas2 complex that can facilitate spacer acquisition in the absence of interference, a process termed naive adaptation[13–16]. The Cas1–Cas2 complex incorporates synthetic double-stranded DNA fragments associated with consensus 5′-AAG-3′/3′-TTC-5′ PAM (PAM^{AAG}) into the CRISPR array in orientation dictated by the PAM sequence and conferring immunity[17]. However, the state of the natural prespacers captured by Cas1–Cas2 in cells and the mechanism ensuring integration of a prespacer in a specific orientation remains unknown.

In primed CRISPR adaptation, interference-driven DNA degradation initiated at a priming protospacer (PPS) is coupled with acquisition of spacers from DNA in the PPS region[18–20]. One hallmark of primed adaptation is that nearly all PPS-region sequences from which spacers are acquired contain a consensus PAM^{AAG}[18–20]. A second hallmark of primed adaptation is that spacer acquisition occurs in a bidirectional, orientation-dependent manner relative to the PAM of the PPS. In particular, the non-transcribed strand of spacers acquired from the PAM-proximal region (upstream) or PAM-distal region (downstream) is derived from the nontarget strand or target strand, respectively[21]. Available in vivo models of primed adaptation that contain a plasmid-borne PPS or phage-borne PPS are limited due to difficulties in detecting bidirectional spacer acquisition or by high rates of cell lysis[18,19,21]. In particular, analysis of spacer acquisition from circular targets, especially small plasmids, is complicated due to overlapping gradients of protospacers located both upstream and downstream of the PPS[18,19,21]. Use of long linear PPS-containing phage genomes imposes difficulties associated with phage biology such as the inability to detect adaptation for some phages or high rates of cell lysis caused by the others[21].

Here we construct a robust in vivo model for primed adaptation consisting of an *E. coli* type I-E CRISPR–Cas self-targeting locus encoding a crRNA that targets a chromosomal protospacer. We develop a strand-specific, high-throughput sequencing method for analysis of DNA fragments, FragSeq, and use this method to detect short fragments derived from the DNA surrounding the targeted protospacer. The detected fragments have sequences matching spacers acquired during primed adaptation, contain ~3- to 4-nt overhangs derived from excision of genomic DNA within a PAM, are generated in a bidirectional, orientation-dependent manner relative to the targeted protospacer, require the functional integrity of machinery for interference and adaptation to accumulate, and function as spacer precursors when introduced exogenously into cells by transformation. DNA fragments with a similar structure accumulate in cells undergoing primed adaptation in a type I-F CRISPR–Cas self-targeting system. We propose that the asymmetrical structure of the spacer precursors detected in this work is a key determinant of spacer integration into the CRISPR array in orientation conferring immunity.

Results

Type I-E self-targeting leads to robust primed adaptation. To overcome limitations of primed adaptation systems with plasmid-borne PPS or phage-borne PPS, we constructed a derivative of *E. coli* K12 with a type I-E CRISPR–Cas locus containing a spacer, Sp^{yihN}, encoding a crRNA targeting a chromosomal protospacer in the non-essential gene *yihN* (Fig. 1a; Supplementary Table 1). Induction of *cas* gene expression in self-targeting cells leads to inhibition of cell growth accompanied by an increase in cell length (Fig. 1b). Furthermore, analysis of chromosomal DNA by high-throughput sequencing shows that induction of *cas* gene expression causes a dramatic loss of ~300 kb of chromosomal DNA in the PPS region (Fig. 1c, Supplementary Fig. 1a, b, Supplementary Table 2). Loss of PPS-region DNA is also observed in cells containing a catalytically inactive Cas1 variant ($Cas1^{H208A}$) [22] but is not observed in cells containing a nuclease-deficient Cas3 variant ($Cas3^{H74A}$)[10] or cells in which Sp^{yihN} is replaced by a spacer targeting M13 phage (Sp^{M13})[9] (Supplementary Fig. 1a, Supplementary Table 3). Similar results are obtained using methods for analysis of double-stranded or single-stranded DNA (Supplementary Fig. 1b, Supplementary Table 2), indicating that interference-driven degradation of both the target and nontarget strands occurs in the self-targeting strain. The results establish that induction of *cas* gene expression results in interference-driven degradation of PPS-region DNA in the type I-E CRISPR–Cas self-targeting system.

To determine whether interference-driven degradation of PPS-region DNA is coupled with spacer acquisition from PPS-region sequences, we analyzed CRISPR arrays by PCR (Fig. 1d). Results indicate that ~20% of arrays acquire a spacer in cells in which *cas* gene expression is induced, while no spacer acquisition is detected in cells in which *cas* gene expression is not induced (Fig. 1d). Furthermore, no spacer acquisition is detected in cells in which Sp^{yihN} is replaced by Sp^{M13} (Fig. 1d), indicating that spacer acquisition requires interference-driven degradation of PPS-region DNA. High-throughput sequencing analysis of amplicons derived from arrays that have acquired a spacer indicate that the self-targeting system exhibits the defining hallmarks of primed adaptation. In particular, >95% of spacers are acquired from a PAM^{AAG}-containing protospacer in the PPS region and, furthermore, spacer acquisition occurs in a bidirectional, orientation-dependent manner characteristic of the *E. coli* I-E system[21] (Fig. 1e, Supplementary Tables 4, 5). We conclude that the type I-E CRISPR–Cas self-targeting strain provides a robust in vivo model system for primed adaptation.

FragSeq detects PPS-region-derived fragments. It has been proposed that interference-driven DNA degradation produces fragments that serve as spacer precursors in primed adaptation[19,23]. To test this model, we developed a method for strand-specific, high-throughput sequencing of DNA fragments, FragSeq. To perform FragSeq, we isolated genomic DNA fragments <700 bp in length, denatured the fragments, ligated single-stranded adapters to the 5′ and 3′ ends of the fragments, amplified the ligation products by PCR, and analyzed the sequences of the fragments by high-throughput sequencing. Because the library construction steps in FragSeq do not involve tailing—i.e., the addition of non-templated nucleotides onto fragment ends—the 5′- and 3′-end sequences of the fragments can be identified with single-nucleotide resolution. We applied FragSeq to identify products of degradation in self-targeting cells undergoing primed adaptation (Fig. 2a, Supplementary Figs. 2–4, Supplementary Tables 6–12 and Methods). Results show accumulation of fragments derived from PPS-region DNA in wild-type cells but not in cells containing inactive variants of Cas1 or

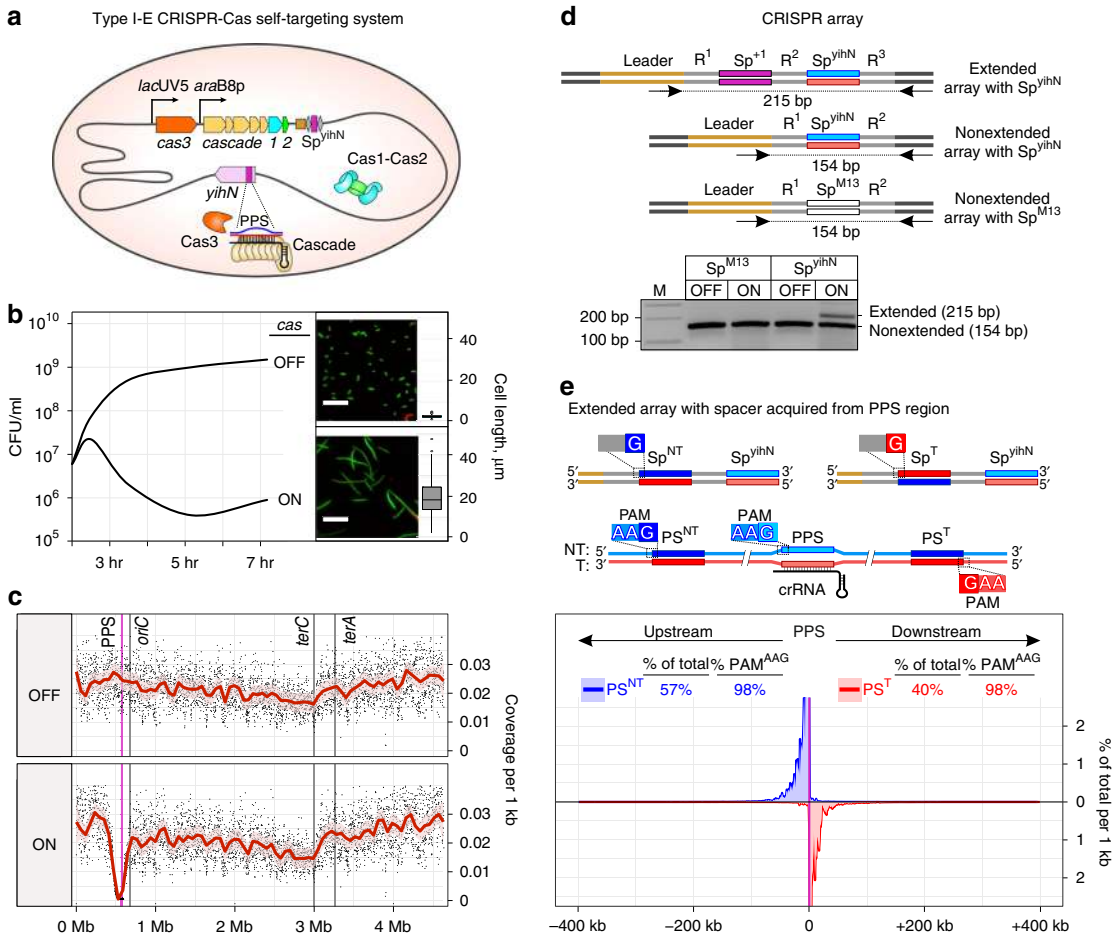


Fig. 1 Interference-driven DNA degradation coupled with spacer acquisition in a type I-E self-targeting system. **a** Type I-E self-targeting system. Shaded oval, *E. coli* cell; gray line, chromosome; orange, tan, blue, and green pentagons, *cas* genes; brown rectangle, CRISPR-array leader sequence; gray diamonds, repeats; purple rectangle, spacer targeting *yihN* (Sp^{yihN}); *lac*UV5 and *ara*B8p, promoters; mauve pentagon, *yihN*; PPS, priming protospacer within *yihN*; blue line, nontarget strand; red line, target strand; black line, crRNA. **b** Effect of self-targeting on cell growth. Growth curve for cultures in which *cas* gene expression is induced (ON) or not induced (OFF). Green, viable cells; red, non-viable cells; scale bar, 20 μm. Boxplot: the central line, median; hinges, the first and third quartiles; whiskers, 1.5 × IQR; $n = 125$. **c** Effect of self-targeting on genomic DNA content. *oriC*, site of replication origin; *terA* and *terC*, sites of replication termination; dot, coverage per 1 kb; red line, Loess smoothing; pink shading, 99% confidence interval. **d** Effect of self-targeting on spacer acquisition: PCR. Schematics depict an extended array containing Sp^{yihN} and acquired spacer Sp^{+1}, a nonextended array containing Sp^{yihN}, or an array containing a spacer targeting M13 phage (Sp^{M13}). Blue line, non-transcribed strand of Sp^{yihN}; red line, transcribed strand of Sp^{yihN} (directs synthesis of crRNA); R, repeats. Arrows below arrays represent the positions of primers used in PCR; sizes of PCR amplicons are indicated. Results show PCR analysis of cells containing an array with Sp^{yihN} or Sp^{M13}. M, double-stranded DNA marker. **e** Effect of self-targeting on spacer acquisition: high-throughput sequencing analysis. Top, extended arrays with spacers acquired from PPS-region protospacers. Bottom, results. Sp^{NT}, spacer with non-transcribed strand derived from nontarget strand (NT, blue) and transcribed strand derived from target strand (T, red); Sp^{T}, spacer with non-transcribed strand derived from target strand (T, red) and transcribed strand derived from nontarget strand (NT, blue). PS^{NT}, protospacer for Sp^{NT}; PS^{T}, protospacer for Sp^{T}. Plot shows percentage of spacers per 1 kb derived from PS^{NT} (blue) or PS^{T} (red). Mean of three biological replicates is shown. Source data are provided as a Source Data file

Cas3, or cells in which Sp^{yihN} is replaced by Sp^{M13} (Fig. 2a, Supplementary Fig. 3a, Supplementary Table 7). Thus, accumulation of PPS-region-derived fragments in cells undergoing primed adaptation requires the functional integrity of both interference and adaptation.

Analysis of length distributions of the PPS-region-derived fragments indicates that they are produced in a bidirectional, orientation-dependent manner reminiscent of spacer acquisition (Fig. 2b). The most abundant nontarget-strand fragments ($Frag^{NT}$) and target-strand fragments ($Frag^{T}$) emanating from the PAM-proximal region of the PPS (upstream) are 32- to 34-nt and 36- to 38-nt, respectively, and the most abundant $Frag^{NT}$ and $Frag^{T}$ emanating from the PAM-distal region of the PPS (downstream) are 36- to 38-nt and 32- to 34-nt, respectively (Fig. 2b). In addition, the relative abundance of complementary 32- to 34-nt and 36- to 38-nt fragments shows a positive correlation (Pearson correlation coefficient 0.48, Supplementary Table 11), suggesting that the fragments identified by FragSeq represent individual strands of double-stranded DNA products having lengths similar to that of spacers (~30 bp). Alignments of the chromosomal sequences associated with the 5′ or 3′ ends of complementary fragments reveals the presence of a consensus 5′-AAG-3′/3′-TTC-5′ PAM derived from sequences associated with the 5′ ends of 32- to 34-nt fragments and the 3′ ends of 36- to 38-nt fragments (Fig. 2c, Supplementary Tables 9, 10). Thus, the results of FragSeq suggest that cells undergoing primed adaptation accumulate 33- or 34-bp double-stranded DNA fragments containing a 3′ end, 4- or 3-nt overhang derived from excision of

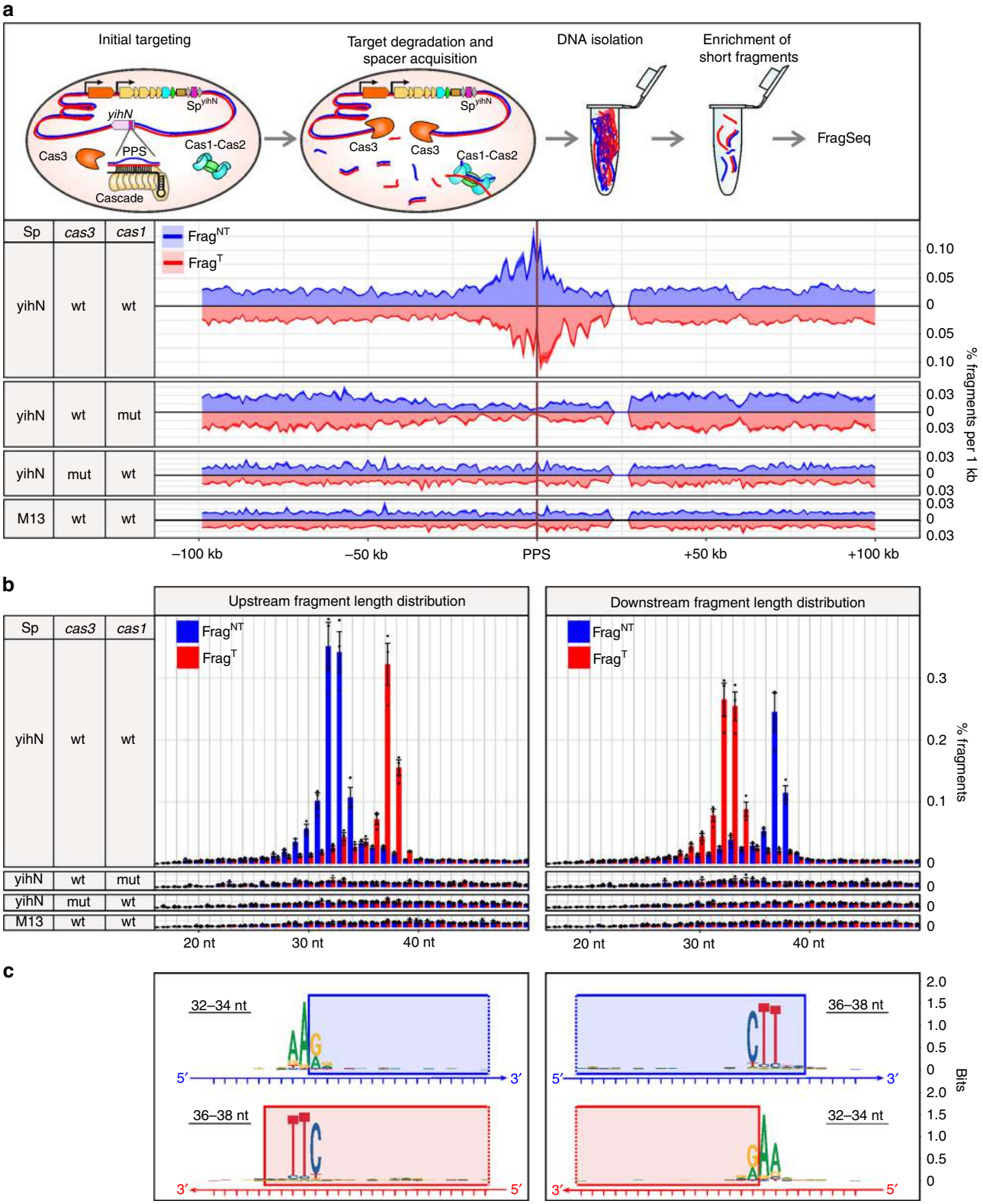


Fig. 2 FragSeq detection of DNA fragments in cells with the type I-E self-targeting system. **a** Effect of self-targeting on PPS-region DNA fragment distributions. Top, events occurring in cells upon induction of *cas* gene expression. Bottom, FragSeq results. Coverage plots show mean of three biological replicates. Blue, nontarget-strand-derived fragments ($Frag^{NT}$); red, target-strand-derived fragments ($Frag^{T}$). **b** Length distributions of PPS-region-derived fragments (mean ± SEM of three biological replicates). **c** Sequence alignments of genomic DNA from which PPS-region fragments are derived. Blue rectangles, sequences present in $Frag^{NT}$; red rectangles, sequences present in $Frag^{T}$. Source data are provided as a Source Data file

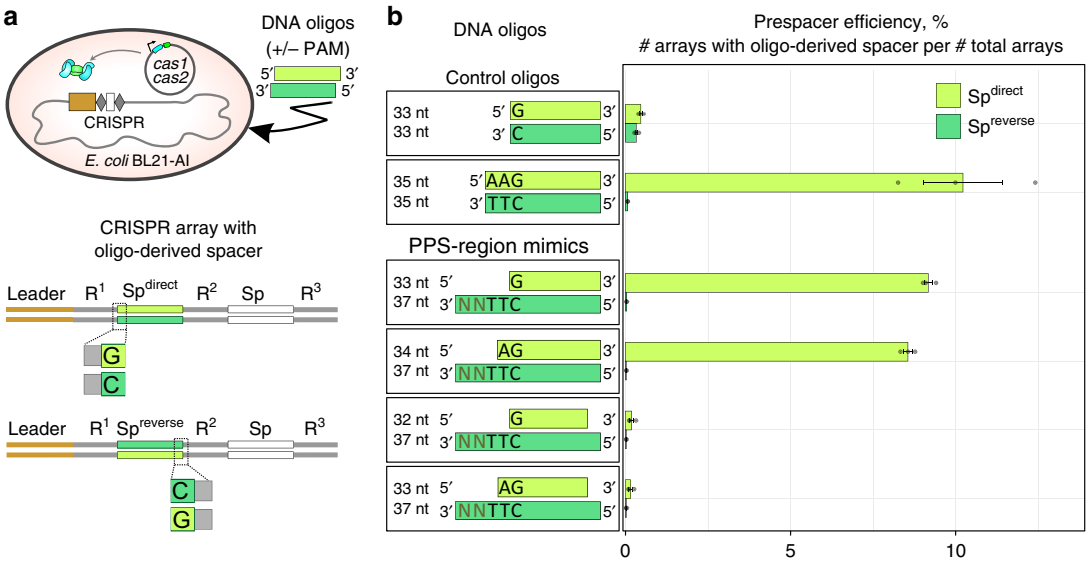


Fig. 3 Synthetic mimics of DNA fragments detected in cells undergoing primed adaptation function as prespacers. **a** Prespacer efficiency assay. Top, introduction of synthetic DNA into cells containing a CRISPR array and plasmid that directs expression of *cas1* and *cas2*. Bottom, integration of synthetic DNA into the CRISPR array occurs in either a direct (Sp^{direct}) or reverse ($Sp^{reverse}$) orientation. **b** Results. Left, oligonucleotides analyzed. Right, percentage of arrays containing oligo-derived spacers having a direct (light green) or reverse (dark green) orientation (mean ± SEM of three biological replicates). Source data are provided as a Source Data file

a PAM-containing sequence (Fig. 2c). Furthermore, the relative abundance of these fragments and spacers acquired during primed adaptation that have an identical sequence shows a positive correlation (Pearson correlation coefficient 0.5–0.6, Supplementary Table 12). Accordingly, the results strongly suggest the fragments accumulating in cells undergoing primed adaptation are products of an intermediate step between protospacer selection and spacer integration.

PPS-region-derived fragments function as prespacers. To directly test whether the PPS-region-derived fragments detected by FragSeq serve as substrates for spacer integration, we performed a prespacer efficiency assay[17] (Fig. 3a). We tested synthetic mimics corresponding to the most abundant PPS-region-derived fragments (Fig. 3b, Supplementary Tables 13–16). Results show that 33- or 34-bp synthetic mimics containing a 3′-end, 4- or 3-nt overhang on the PAM-derived end, respectively, and a blunt PAM-distal end were integrated into arrays with an efficiency similar to a control fragment containing a consensus PAM^{AAG} (~10% prespacer efficiency; Fig. 3b, Supplementary Tables 14, 15). In addition, the synthetic mimics and PAM^{AAG}-containing control fragment were integrated in a direct orientation with the G:C of the PAM positioned adjacent to the first repeat in the array (Fig. 3, Supplementary Table 15). Introduction of a 5′-end, 1-nt overhang on the PAM-distal end reduced prespacer efficiency by ~45-fold (Fig. 3b, Supplementary Table 15). The results establish that PPS-region-derived fragments containing a 3′-end overhang on the PAM-derived end and blunt PAM-distal end function as efficient spacer precursors.

Prespacers in I-E and I-F systems exhibit similar structures. In a prior work, we developed an *E. coli* strain that provides a model system for studies of self-targeting by the type I-F CRISPR–Cas system from *Pseudomonas aeruginosa*[24] (Fig. 4a). Compared with the orientation bias in spacer acquisition observed in type I-E systems, orientation bias in type I-F systems is reversed. In particular, the non-transcribed strand of spacers acquired from the PAM-proximal region of the PPS (upstream) or PAM-distal region of the PPS (downstream) are derived from the target strand or nontarget strand, respectively in type I-F. To determine whether spacer precursors could be detected in the type I-F system, we performed FragSeq analysis in cells undergoing primed adaptation (Fig. 4b, Supplementary Tables 17–21). Similar to the type I-E system, we detect accumulation of spacer-sized double-stranded DNA fragments containing a 3′-end, 5-nt overhang on the PAM-derived end (Fig. 4b). Thus, in spite of exhibiting opposite orientation bias in spacer acquisition, primed adaptation in type I-E and type I-F systems involves generation of spacer precursors with a similar structure (Fig. 4c).

Discussion

In summary, we have identified spacer precursors produced as products of an intermediate step (or steps) between protospacer selection and spacer integration for type I-E and type I-F CRISPR–Cas systems. Accumulation of spacer precursors in the type I-E system requires the functional integrity of components of interference and adaptation (Fig. 5) indicating that protospacer selection involves coordination between the interference machinery and adaptation machinery (Fig. 5a). Strikingly, spacer precursors detected during primed adaptation in both type I-E and type I-F systems share an asymmetrical structure characterized by a 3′-end overhang on the PAM-derived end. Thus, we propose that spacer precursors detected in this work are products generated during universal steps of prespacer processing in type I CRISPR–Cas systems relying on Cas1 and Cas2 and lacking auxiliary adaptation proteins. We further propose that the asymmetrical structure of the spacer precursors detected in this work is a key determinant of the sequential integration of prespacers into the CRISPR array (Fig. 5b). In addition, the FragSeq method reported in this work should be applicable, essentially without modification, to identify spacer precursors that form in vivo in any CRISPR–Cas system.

Methods

Bacterial strains and plasmids. The *E. coli* strains used in this study are listed in Supplementary Table 1. Red recombinase-mediated gene-replacement technique was used to obtain strains KD403, KD518 and KD753[25].

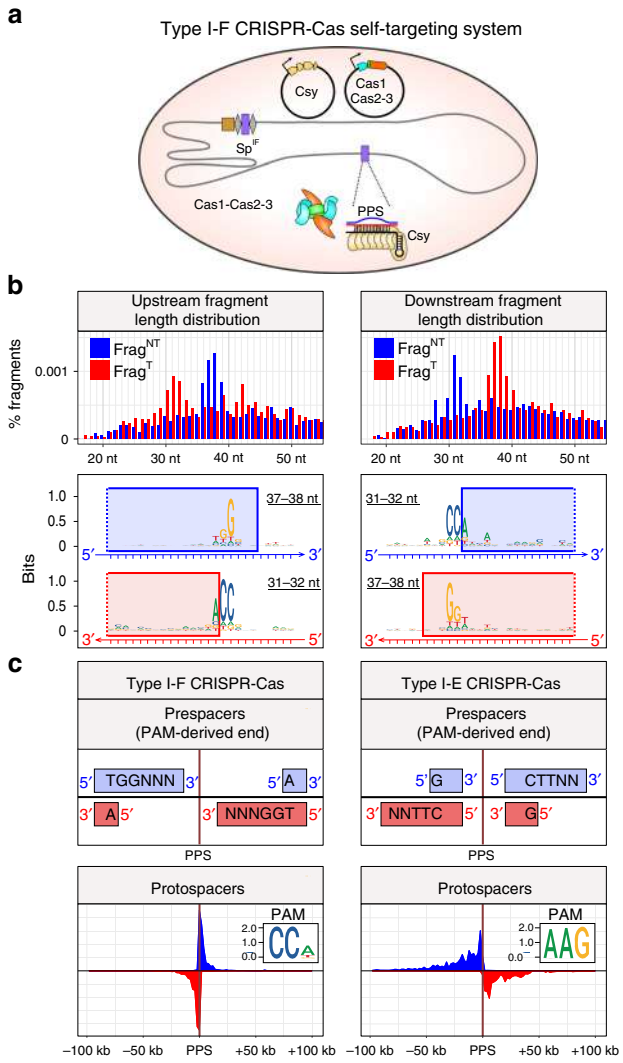


Fig. 4 Identification of spacer precursors generated in a type I-F self-targeting system by FragSeq. **a** Components of type I-F CRISPR–Cas self-targeting system. Shaded oval, *E. coli* cell; gray line, chromosomal DNA; black line, plasmid DNA; orange, tan, blue, and green pentagons, *cas* and *csy* genes; brown rectangle, array leader sequence; gray diamonds, array repeat sequences; purple rectangles, spacer and chromosomal PPS targeted by spacer-derived crRNA; Csy, type I-F effector complex; Cas1–Cas2-3, complex of Cas1 and Cas2-3 proteins. **b** FragSeq results: length distributions of fragments (top) and sequence features of PPS-region sequences from which fragments are derived (bottom). Logos for 31–32-nt fragments were generated by aligning sequences 10-nt upstream to 15-nt downstream of the fragment 5′ end. Logos for 37–38-nt fragments were generated by aligning sequences 20-nt upstream to 5-nt downstream of the fragment 3′ end. Blue rectangles, sequences present in $Frag^{NT}$; red rectangles, sequences present in $Frag^{T}$. **c** Comparison of PPS-region-derived fragments and PPS-region protospacers in type I-F and type I-E self-targeting systems. Inset, logo derived from alignment of PPS-region PAMs

Plasmid pCas1 + 2 for the expression of type I-E *cas1* and *cas2* genes as well as plasmids pCas and pCsy for expression type I-F *cas* and *csy* genes were described earlier[13,24].

Growth conditions. For analysis of CRISPR-mediated self-targeting by the type I-E system, overnight culture of KD403 strain grown at 37 °C in LB medium was diluted 100-fold into 10 ml of fresh LB and incubated at 37 °C until OD_{600} reached 0.3. The culture was divided into two portions, *cas* genes inducers, IPTG and L-(+)-arabinose were added at 1 mM concentration to one portion, and cultures with and without inducers were incubated at 37 °C for 7 h. At various time points postinduction, the cells were plated with serial dilutions on 1.5% LB agar plates for counting colony forming units (CFUs) or were monitored using fluorescent microscopy.

In assays using strains KD403, KD518, KD753 and KD263 that were followed by sequencing of total genomic DNA, short DNA fragments or newly acquired spacers, similar conditions of culture growth and *cas* genes induction were applied, except that overnight cultures were diluted 100-fold in 100 ml of LB and grown at 30 °C. Five hours postinduction, 10 ml of cells were pelleted by centrifugation at 3000×*g* for 5 min at 4 °C, washed with 10 ml of PBS, pelleted by centrifugation at 3000×*g* for 5 min at 4 °C and resuspended in 1 ml of PBS. The cells were divided into 125-µl aliquots and stored at −70 °C before they were used for DNA isolation.

For analysis of short DNA fragments generated during self-targeting by the type I-F system, cultures of strain KD675 transformed with plasmids pCas and pCsy were grown at 37 °C in LB supplemented with 100 µg/ml ampicillin and 50 µg/ml spectinomycin. Overnight cultures were diluted 200-fold into 10 ml of LB without antibiotics, grown at 37 °C until OD_{600} reached 0.3 and supplemented with 1 mM IPTG and 1mM L-(+)-arabinose. The cells were harvested 24 h postinduction and prepared for DNA isolation as described above for strains KD403, KD518, KD753 and KD263.

Fluorescence microscopy. Cultures grown with or without induction of *cas* gene expression were analyzed using a LIVE/DEAD viability kit (Thermo Scientific) at 5 h after induction. Viable cells in each culture were detected by addition of 20 µM SYTO9, green fluorescent dye that can penetrate through intact cell membranes. Non-viable cells in each culture were detected by addition of 20 µM propidium iodide dye, which cannot enter viable cells. Sample chambers were made using a microscope slide (Menzel–Gläser) with two strips on the upper and lower edges formed by double-sided sticky tape (Scotch TM). To obtain a flat substrate required for high-quality visualization of bacteria, a 1.5% agarose solution was placed between tape strips and covered with another microscopic slide. After solidification of the agarose, the upper slide was removed and several agarose pads were formed; 1 µl of each cell suspension (with and without induction) was placed on an agarose pad. The microscopic chamber was sealed using a coverslip (24 × 24 mm, Menzel–Gläser).

Fluorescence microscopy was performed using Zeiss AxioImager.Z1 upright microscope. Fluorescence signals in green (living cells) and red (dead cells) fluorescent channels were detected using Zeiss Filter Set 10 and Semrock mCherry-40LP filter set, respectively. Fluorescent images of self-targeting cells were obtained using Cascade II:1024 back-illuminated EMCCD camera (Photometrics). The microscope was controlled using AxioVision Microscopy Software (Zeiss). All image analysis was performed using ImageJ (Fiji) with ObjectJ plugin used for measurements of cell length[26].

High-throughput sequencing of total genomic DNA. Total genomic DNA was purified by GeneJET Genomic DNA Purification Kit (Thermo Fisher Scientific). Sequencing libraries were prepared either by NEBNext® Ultra™ II DNA Library Prep Kit for Illumina (NEB) or by Accel-NGS® 1S Plus DNA Library Kit (Swift Biosciences) and sequenced on a NextSeq 500 platform.

Raw reads were analyzed in R with ShortRead and Biostrings packages[27]. Reads with no more than two bases with quality <20 were mapped to the KD403 reference genome using Unipro UGENE platform[28]. Bowtie2 was used as a tool for alignment with end-to-end alignment mode and 1 mismatch allowed[29]. The BAM files were analyzed by Rsamtools package and reads with the MAPQ score equal to 42 were selected and used for downstream coverage analysis[30]. Mean coverage over non-overlapping 1 kb bins was calculated and normalized to the total coverage (the sum of means).

High-throughput sequencing of newly acquired spacers. Cell lysates were prepared by resuspending cells in water and heating at 95 °C for 5 min. Cell debris was removed from lysates by centrifugation at 16×*g* for 1 min. For the analysis of spacer acquisition in strains KD263 and KD403, lysates were used in PCR reactions containing primers LDR-F2 (ATGCTTTAAGAACAAATGTATACTTTTAG) and Ec_minR (CGAAGGCGTCTTGATGGGTTTG) (25 cycles, $T_a = 52$ °C) (Supplementary Table 22). Reaction products were separated by agarose gel electrophoresis (Fig. 1d; the uncropped image of the gel is available in the Source Data file). To obtain amplicons derived from extended CRISPR arrays in strain KD403, PCR reactions were performed using primers LDR-F2 (ATGCTTTAAGAACAAATGTATACTTTTAG) and autoSp2_R (AATAGCGAACAACAAGGTCGGTTG) (30 cycles, $T_a = 52$ °C) (Supplementary Table 22). Reaction products were separated by agarose gel electrophoresis, and the amplicon derived from the extended array was purified from the gel using a GeneJET Extraction Kit (Thermo Fisher Scientific) and sequenced on a NextSeq 500 system.

Bioinformatic analysis was performed in R using ShortRead and Biostrings packages[27]. Bases with quality <20 were substituted with N and spacer sequences were extracted from the reads containing two or more CRISPR repeats. Spacers of length 33 bp were mapped to the KD403 genome to identify 33-bp protospacer sequences with 0 mismatches. Spacers that aligned to a single position in the chromosome were used to determine protospacer distribution along the genome.

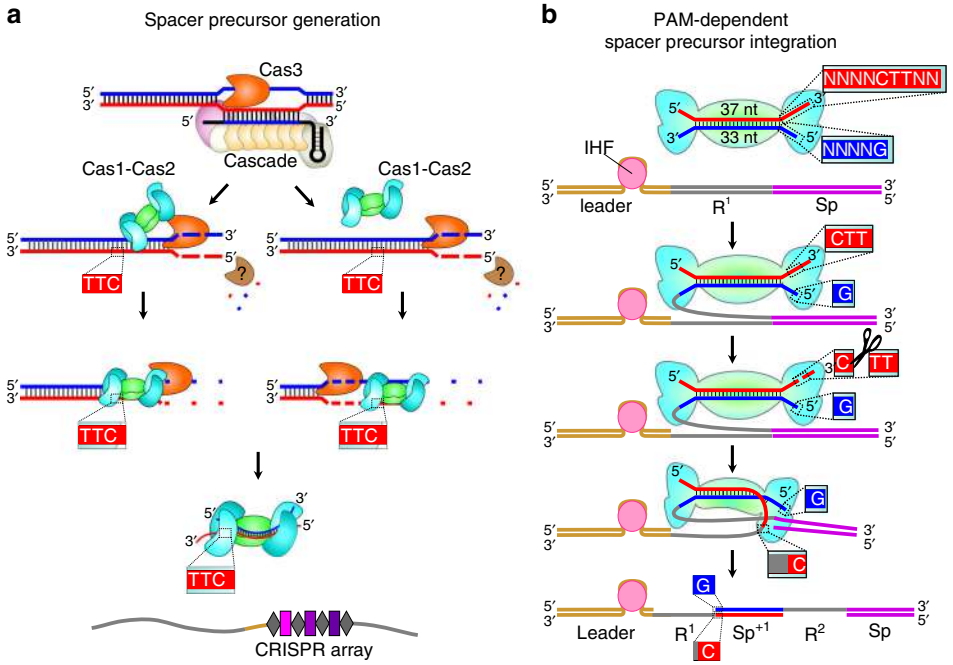


Fig. 5 Model of primed adaptation in type I-E CRISPR–Cas systems. **a** Generation of spacer precursors involves coordination between interference and adaptation. Pathway on left depicts direct coordination between interference and adaptation in which Cas1–Cas2 associates with Cas3 as it moves along DNA[12,34]. Pathway on right depicts indirect coordination between interference and adaptation in which Cas1–Cas2 captures products of Cas3-mediated DNA degradation[19,23]. Both pathways generate spacer precursors containing a 3′-end overhang on the PAM-derived end and blunt PAM-distal end. Model depicts rapid degradation of DNA not selected as a spacer precursor by Cas3 and an unknown nuclease (brown). Binding of prespacers to Cas1–Cas2 prevents prespacer degradation. **b** Sequential integration of spacer precursors. First, binding of IHF to the leader stimulates integration of the blunt PAM-distal end between the leader and first repeat sequence[35]. Second, the 3′-overhang present on the PAM-derived end is cleaved by Cas1[36] or DnaQ exonucleases[37,38] facilitating integration between the first repeat sequence and first spacer of the array. The order of events depicted results in integration of the spacer precursor in a direct orientation with respect to the PAM (see Fig. 3a)

Spacers arising from protospacers due to potential slippage or flippage were removed from analysis[31] (Supplementary Tables 4, 5).

Prespacer efficiency assay. Prespacer efficiency assay was performed according to the following protocol[17]. Overnight culture of BL21-AI cells containing a plasmid pCas1 + 2 was diluted 30-fold into 9 ml of LB supplemented with 50 μg/ml streptomycin, 13 mM L-(+)-arabinose and 1 mM IPTG and grown at 37 °C for 2 h. Cells were harvested by centrifugation at +4 °C (1 ml of cells per transformation), washed twice with cold water and resuspended in 50 μl of a solution containing 3.125 μM complementary oligonucleotides (Supplementary Table 13). Electroporation was carried out in a 1-mm gap cuvette at a voltage of 1.8 kV. 3 ml of LB supplemented with 50 μg/ml streptomycin was added to the electroporated cells and the cultures were incubated at 37 °C during 2 h. Lysates of cell cultures were prepared and used in PCR reactions containing a primer BLCRdir complementary to the leader sequence (GGTAGATTGTGACTGGCTTAAAAAATC) and a primer BLCRreverse complementary to the preexisting spacer in the array (GTTTGAGCGATGATATTTGTGCTC), respectively (Supplementary Table 22). Amplicons corresponding to extended and nonextended CRISPR arrays were isolated using GeneJET PCR Purification Kit (ThermoFisher Scientifc) and sequenced on a NextSeq 500 platform. Bioinformatic analysis was performed in R using ShortRead and Biostrings packages[27]. Reads containing the bases with Phred quality <14 were removed from analysis and reads containing at least one CRISPR repeat were further analyzed. Newly acquired spacers were extracted from the expanded reads and mapped to the genome, plasmid and transforming oligonucleotide sequence with two mismatches allowed; 33-bp oligo-derived spacers that were cut between AA and G before integration were considered as properly processed. For simplicity, only properly processed oligo-derived spacers inserted into the CRISPR array in direct (GCCCAATTTACTACTCGTTCTGGTGTTTCTCGT) or reverse (ACGAGAAACACCAGAACGAGTAGTAAATTGGGC) orientation were included into analysis.

Isolation of DNA fragments generated in vivo. Total genomic DNA was isolated from cultures of strains KD403, KD518, KD753, KD263 and KD675 by collecting 1.25 ml of cell suspensions by centrifugation, resuspending cells in 125 μl of PBS, adding 2 ml of lysis buffer (0.6% SDS, 12 μg/ml proteinase K in 1× TE buffer) and incubating at 55 °C for 1 h. Two milliliters of phenol:chloroform:isoamyl alcohol (25:24:1) (pH 8) was added to the lysate, the solution was gently mixed, and the aqueous and organic phases separated by centrifugation at 7000×*g* for 10 min at room temperature. The upper aqueous phase containing total genomic DNA was collected and the residual phenol was removed by the addition of 2 ml of chloroform:isoamyl alcohol (24:1). The solution was gently mixed, centrifuged at 7000×*g* for 10 min at room temperature. The upper DNA-containing fraction was transferred to a fresh tube; 0.2 M NaCl, 15 μg/ml of Glycoblue (Invitrogen) and two volumes of cold 100% ethanol were added, and the solution was incubated at −80 °C overnight. Precipitated DNA was recovered by centrifugation at 21,000×*g* for 30 min at 4 °C. Pellets were washed twice with 80% ethanol, resuspended in 200 μl of 1× TE buffer, and treated with 1 mg/ml RNase A at 37 °C for 30 min to remove the residual RNA. DNA was isolated by phenol:chloroform:isoamyl alcohol extraction and ethanol precipitation as described above.

DNA fragments <700 bp in length were isolated from 9 μg of total genomic DNA using a Select-a-Size DNA Clean & Concentrator kit (Zymo Research) according to manufacturer's recommendations. To ensure the binding of fragments <50 bp to the column filter, the volume of 100% ethanol added to the fraction prior to on-filter purification was increased from 290 μl to 600 μl. DNA fragments were eluted with 2 × 50 μl of elution buffer, pooled and purified by ethanol precipitation. A total of 100 μl of DNA was mixed with 10 μl of 3 M NaOAc (0.1×V), 1 μl of 10 mg/ml glycogen (0.01×V) and 330 μl of 100% ethanol, vortexed, and incubated overnight at −80 °C. DNA was recovered by centrifugation at 21,000×*g* for 30 min at 4 °C. Pellets were washed three times with 80% cold ethanol, air dried for ~5 min, and resuspended in 5 μl of nuclease-free water.

High-throughput sequencing of DNA fragments: FragSeq. The DNA oligo i116 that served as a 3′ adapter was adenylated using 5′ DNA Adenylation Kit (NEB), purified by ethanol precipitation as above and diluted to 10 μM with nuclease-free water (Supplementary Table 23).

DNA fragments <700 bp (in 5 μl of water) were heat-denatured at 95 °C for 5 min, cooled to 65 °C, and mixed with 0.5 μM adenylated oligo i116, 1× NEBuffer 1, 5 mM $MnCl_2$ and 10 pmol of thermostable 5′ App DNA/RNA ligase (NEB) in 10-μl reaction volume. The mixture was incubated at 65 °C for 1 h, heated at 90 °C for 3 min, and cooled to 4 °C on ice. Ligated products were combined with 1× T4 RNA ligase buffer, 12% PEG 8000, 10 mM DTT, 60 μg/ml BSA and 10 U of T4 RNA ligase 1 (NEB) in a 25-μl reaction volume. The reaction was incubated at 16 °C for 16 h; 25 μl of 2× loading dye was added, and the products were separated by electrophoresis on 10% 7 M urea slab gels (equilibrated and run in 1× TBE buffer). The gel was stained with SYBR Gold nucleic acid gel stain, bands were visualized

on a UV transilluminator, and products of ~40 to ~500 nt were excised from the gel and recovered as described in Vvedenskaya et al.[32]. Briefly, the excised gel slice was crushed, 400 µl of 0.3 M NaCl in 1× TE buffer was added, and the mixture incubated at 70 °C for 10 min. The eluate was collected using a Spin-X column. After the first elution step, the elution procedure was repeated, eluates were pooled, and DNA was isolated by ethanol precipitation and resuspended in 15 µl of nuclease-free water.

Next, the 3′ adapter-ligated DNA fragments were adenylated using 5′ DNA Adenylation Kit (NEB) in a 20-µl reaction following the manufacturer's recommendations. Nuclease-free water was added to 100 µl, DNA fragments were purified by ethanol precipitation and resuspended in 5 µl of nuclease-free water. The two-step ligation procedure described above was repeated using 5 µl of adenylated 3′-ligated DNA fragments, 0.5 µM of barcoded oligos i112, i113, i114 or i115 that served as 5′ adapters (barcodes were used as internal controls; Supplementary Table 23), 10 pmol of thermostable 5′ App DNA/RNA ligase at the first ligation step, and 10 U of T4 RNA ligase 1 at the second ligation step. Reactions were stopped by addition of 25 µl of 2× loading dye, and the products were separated by electrophoresis on 10% 7 M urea slab gels (equilibrated and run in 1× TBE buffer). DNA products of ~70 to ~500 nt in size were excised and eluted from the gel as described above, isolated by ethanol precipitation, and resuspended in 20 µl of nuclease-free water.

To amplify DNA, 2–8 µl of adapter-ligated DNA fragments were added to a mixture containing 1× Phusion HF reaction buffer, 0.2 mM dNTPs, 0.25 µM Illumina RP1 primer, 0.25 µM Illumina index primer and 0.02 U/µl Phusion HF polymerase in a 30-µl reaction (Supplementary Table 24). PCR was performed with an initial denaturation step of 30 s at 98 °C, amplification for 15 cycles (denaturation for 10 s at 98 °C, annealing for 20 s at 62 °C and extension for 15 s at 72 °C), and a final extension for 5 min at 72 °C. Amplicons were isolated by electrophoresis using a non-denaturing 10% slab gel (equilibrated and run in 1× TBE). The gel was stained with SYBR Gold nucleic acid gel stain and species of ~150 to ~300 bp were excised. DNA products were eluted from the gel with 600 µl of 0.3 M NaCl in 1× TE buffer at 37 °C for 3 h, purified by ethanol precipitation, and resuspended in 25 µl of nuclease-free water. Barcoded libraries were sequenced on Illumina NextSeq 500 platform in high output mode.

Bioinformatic analysis was performed in R using ShortRead and Biostrings packages[27]. Bases with quality <20 were substituted with N. After adapter trimming, all reads were compared to each other to reveal clusters of overamplified reads containing the same insert and combination of unique molecular identifiers conjugated to adapters. For each cluster, a consensus sequence was extracted and used together with non-overamplified reads for further alignment to KD403 reference genome with two mismatches allowed. Only reads with a length 16–100 nt uniquely aligned to the genome were further analyzed (Supplementary Fig. 4). Logos were generated using ggseqlogo package[33].

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

A reporting summary for this Article is available as a Supplementary Information file. Raw sequencing data obtained in this study are available in Sequence Read Archive (BioProject Accession: PRJNA552808). The source data underlying Figs. 1b, d, e, 2a, b, 3b and Supplementary Figs. 1a and 3a are provided as a Source Data file. All data are available from the corresponding author upon reasonable request.

Code availability

Custom code and information about software used in this study is available at GitHub (https://github.com/AnnaBioLogic/Shiriaeva_et_al_2019).

Received: 21 May 2019; Accepted: 8 September 2019;
Published online: 10 October 2019

References

1. Brouns, S. J. et al. Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960–964 (2008).
2. Jore, M. M. et al. Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nat. Struct. Mol. Biol.* **18**, 529–536 (2011).
3. Wiedenheft, B. et al. Structures of the RNA-guided surveillance complex from a bacterial immune system. *Nature* **477**, 486–489 (2011).
4. Mojica, F. J., Diez-Villasenor, C., Garcia-Martinez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733–740 (2009).
5. Hayes, R. P. et al. Structural basis for promiscuous PAM recognition in type I-E Cascade from E. coli. *Nature* **530**, 499–503 (2016).
6. Hochstrasser, M. L. et al. CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. *Proc. Natl Acad. Sci. USA* **111**, 6618–6623 (2014).
7. Mulepati, S., Orr, A. & Bailey, S. Crystal structure of the largest subunit of a bacterial RNA-guided immune complex and its role in DNA target binding. *J. Biol. Chem.* **287**, 22445–22449 (2012).
8. Sashital, D. G., Wiedenheft, B. & Doudna, J. A. Mechanism of foreign DNA selection in a bacterial adaptive immune system. *Mol. Cell* **46**, 606–615 (2012).
9. Semenova, E. et al. Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc. Natl Acad. Sci. USA* **108**, 10098–10103 (2011).
10. Westra, E. R. et al. CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. *Mol. Cell* **46**, 595–605 (2012).
11. Mulepati, S. & Bailey, S. In vitro reconstitution of an Escherichia coli RNA-guided immune system reveals unidirectional, ATP-dependent degradation of DNA target. *J. Biol. Chem.* **288**, 22184–22192 (2013).
12. Dillard, K. E. et al. Assembly and translocation of a CRISPR-Cas primed acquisition complex. *Cell* **175**, 934–946.e15 (2018).
13. Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic Acids Res.* **40**, 5569–5576 (2012).
14. Nunez, J. K., Harrington, L. B., Kranzusch, P. J., Engelman, A. N. & Doudna, J. A. Foreign DNA capture during CRISPR-Cas adaptive immunity. *Nature* **527**, 535–538 (2015).
15. Nunez, J. K. et al. Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nat. Struct. Mol. Biol.* **21**, 528–534 (2014).
16. Nunez, J. K., Lee, A. S., Engelman, A. & Doudna, J. A. Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature* **519**, 193–198 (2015).
17. Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. Molecular recordings by directed CRISPR spacer acquisition. *Science* **353**, aaf1175 (2016).
18. Datsenko, K. A. et al. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat. Commun.* **3**, 945 (2012).
19. Swarts, D. C., Mosterd, C., van Passel, M. W. & Brouns, S. J. CRISPR interference directs strand specific spacer acquisition. *PLoS ONE* **7**, e35888 (2012).
20. Savitskaya, E., Semenova, E., Dedkov, V., Metlitskaya, A. & Severinov, K. High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in E. coli. *RNA Biol.* **10**, 716–725 (2013).
21. Strotskaya, A. et al. The action of Escherichia coli CRISPR-Cas system on lytic bacteriophages with different lifestyles and development strategies. *Nucleic Acids Res.* **45**, 1946–1957 (2017).
22. Babu, M. et al. A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. *Mol. Microbiol.* **79**, 484–502 (2011).
23. Kunne, T. et al. Cas3-derived target DNA degradation fragments fuel primed CRISPR adaptation. *Mol. Cell* **63**, 852–864 (2016).
24. Vorontsova, D. et al. Foreign DNA acquisition by the I-F CRISPR-Cas system requires all components of the interference machinery. *Nucleic Acids Res.* **43**, 10848–10860 (2015).
25. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc. Natl Acad. Sci. USA* **97**, 6640–6645 (2000).
26. Vischer, N. O. et al. Cell age dependent concentration of Escherichia coli divisome proteins analyzed with ImageJ and ObjectJ. *Front. Microbiol.* **6**, 586 (2015).
27. Morgan, M. et al. ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* **25**, 2607–2608 (2009).
28. Okonechnikov, K., Golosova, O. & Fursov, M., team, U. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* **28**, 1166–1167 (2012).
29. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
30. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
31. Shmakov, S. et al. Pervasive generation of oppositely oriented spacers during CRISPR adaptation. *Nucleic Acids Res.* **42**, 5907–5916 (2014).
32. Vvedenskaya, I. O., Goldman, S. R. & Nickels, B. E. Preparation of cDNA libraries for high-throughput RNA sequencing analysis of RNA 5′ ends. *Methods Mol. Biol.* **1276**, 211–228 (2015).
33. Wagih, O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* **33**, 3645–3647 (2017).
34. Redding, S. et al. Surveillance and processing of foreign DNA by the Escherichia coli CRISPR-Cas system. *Cell* **163**, 854–865 (2015).
35. Nunez, J. K., Bai, L., Harrington, L. B., Hinder, T. L. & Doudna, J. A. CRISPR immunological memory requires a host factor for specificity. *Mol. Cell* **62**, 824–833 (2016).

36. Wang, J. et al. Structural and mechanistic basis of PAM-dependent spacer acquisition in CRISPR-Cas systems. *Cell* **163**, 840–853 (2015).
37. Drabavicius, G. et al. DnaQ exonuclease-like domain of Cas2 promotes spacer integration in a type I-E CRISPR-Cas system. *EMBO Rep.* **19**, e45543 (2018).
38. Kim, S., Loeff, L., Colombo, S., Brouns, S. J. J. & Joo, C. Selective prespacer processing ensures precise CRISPR-Cas adaptation. Preprint at https://www.biorxiv.org/content/10.1101/608976v1 (2019).

Acknowledgements

We thank Dr. Dibyendu Kumar and Dr. Min Tu for performing high-throughput sequencing for this project at Waksman Genomics Core Facility, Rutgers University. The microscopy experiments were carried out using scientific equipment of the Center of Shared Usage "The analytical center of nano- and biotechnologies of SPbPU". This work was supported by NIH grant GM10407 (K.S.), NIH grant GM118059 (B.E.N.) and Russian Science Foundation grant 14–14–00988 (K.S.).

Author contributions

A.A.S, I.O.V., B.E.N., K.S. and E.Se. designed the experiments. A.A.S., E.Sa., K.A.D., I.O.V., I.F., N.M., A.M., A.S. and E.Se. performed the experiments. A.A.S. and E.Sa. analyzed the high-throughput sequencing data. A.A.S., B.E.N., K.S. and E.Se. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-019-12417-w.

Correspondence and requests for materials should be addressed to B.E.N., K.S. or E.S.

Peer review information *Nature Communications* thanks Ailong Ke, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

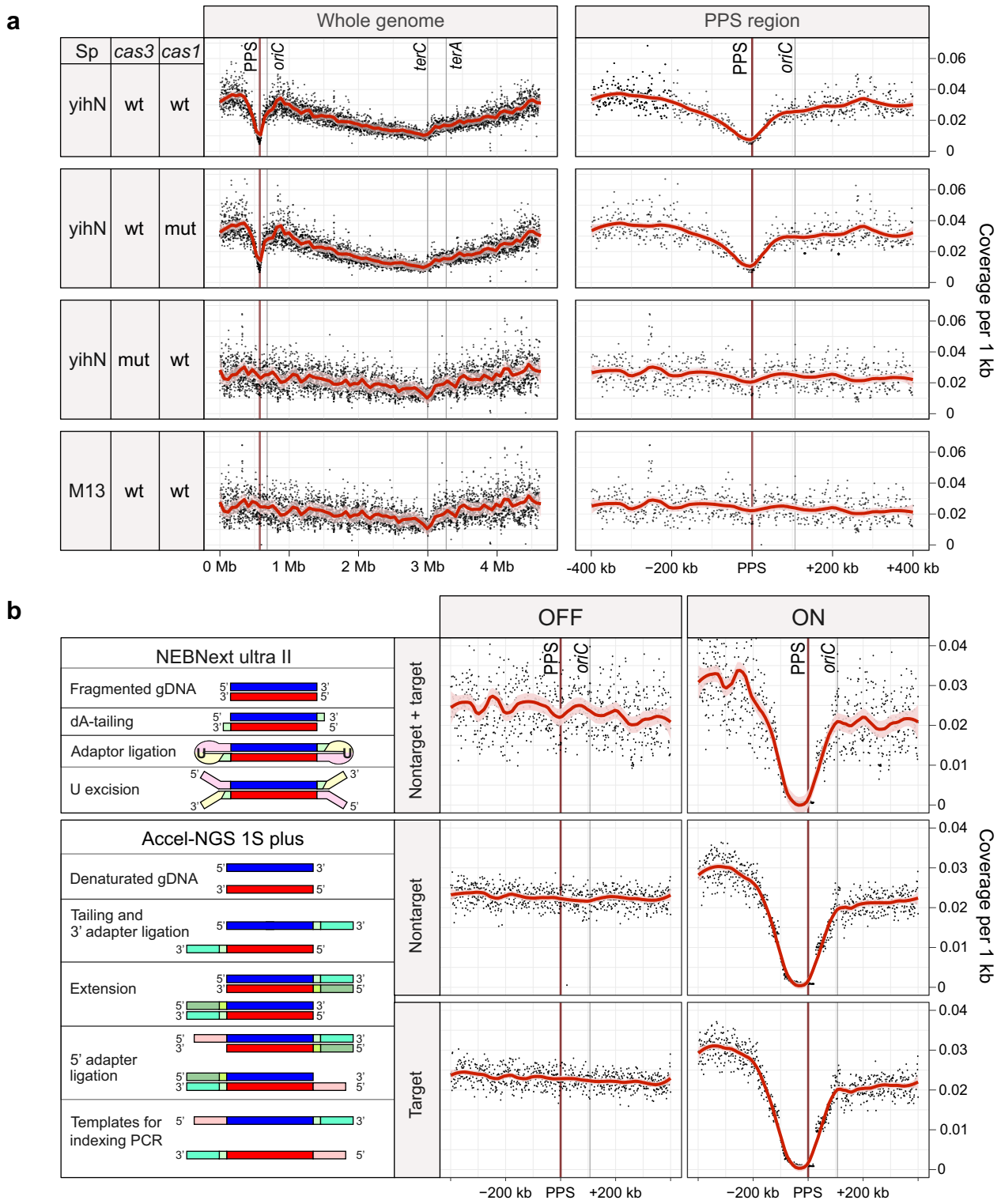
Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

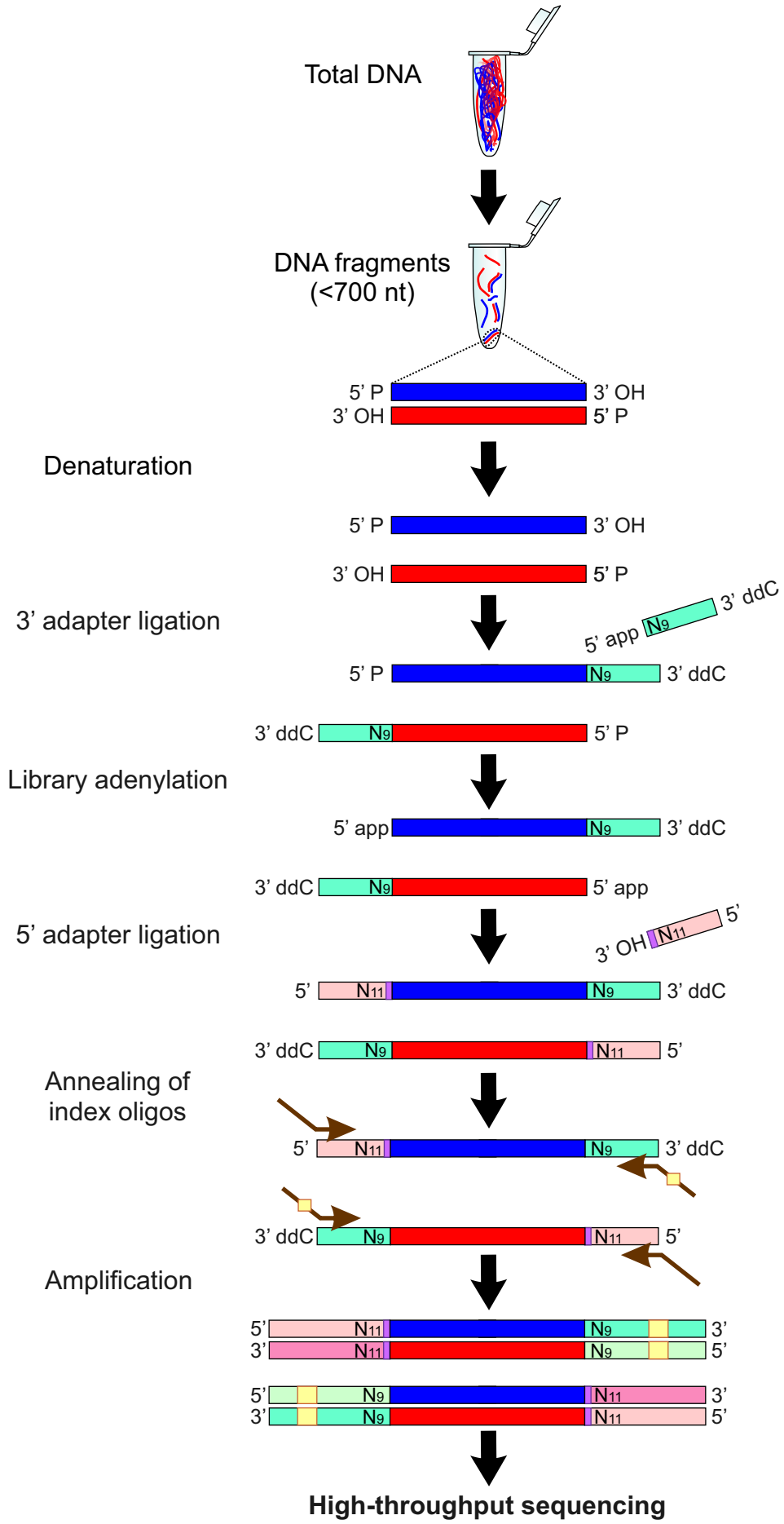
Supplementary Information

Detection of Spacer Precursors Formed *In Vivo* During Primed CRISPR Adaptation

Shiriaeva et al.

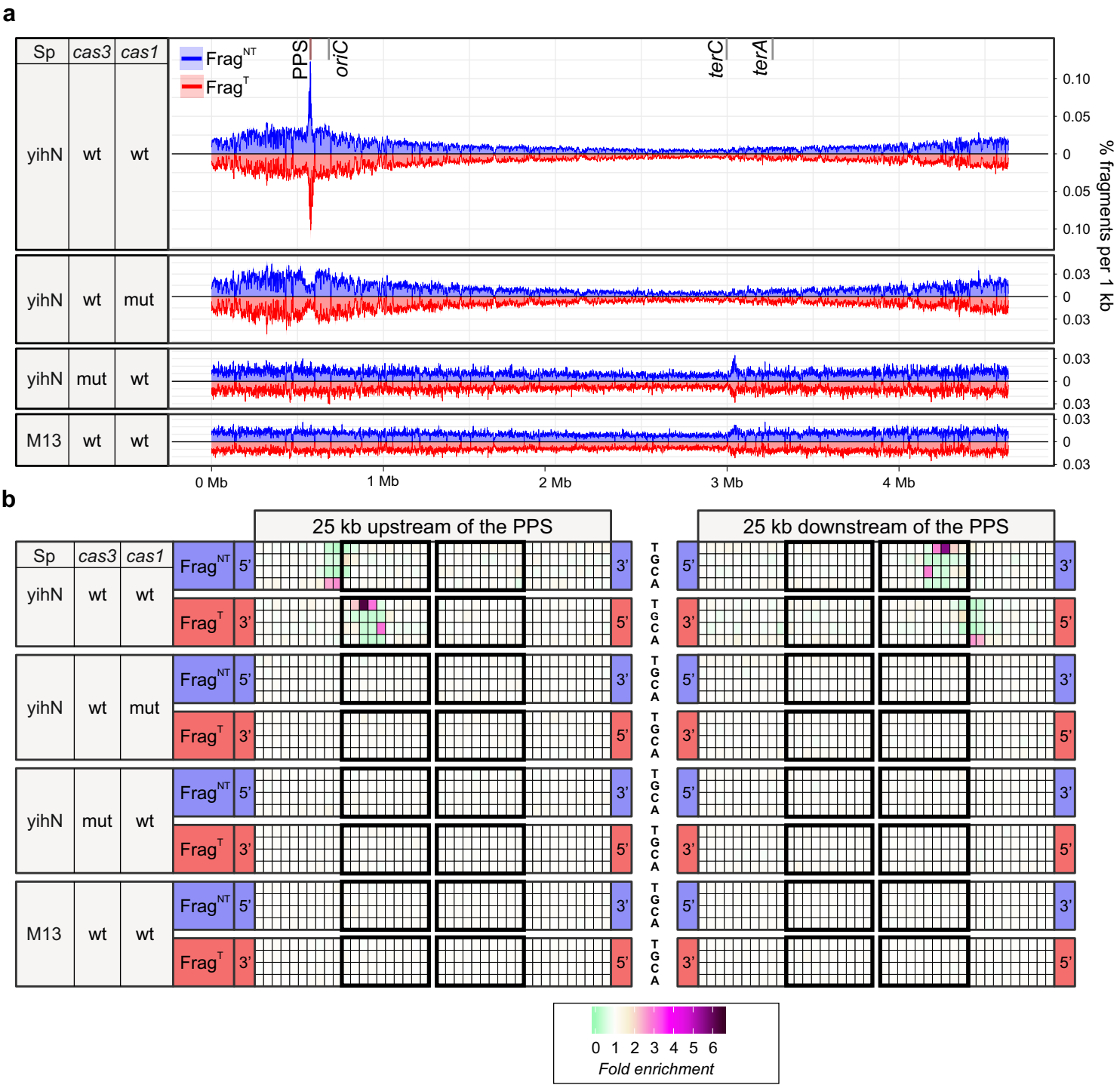


Supplementary Figure 1. CRISPR interference in the type I-E self-targeting system results in loss of chromosomal DNA in the PPS^{yihN} region. a, High-throughput-sequencing analysis of genomic DNA: effects of disruptions in components of interference or adaptation. Graph of sequence coverage per 1 kb for the whole genome (left) or PPS^{yihN} region (right) in the indicated strains. *oriC*, site of replication origin; *terA* and *terC*, sites of replication termination; dot, coverage per 1 kb (mean of 3 biological replicates); red line, Loess smoothing; pink shading, 99% confidence interval. *cas1* mut, gene encoding Cas1^{H208A}, *cas3* mut, gene encoding Cas3^{H74A}. (We note that differences in growth rate likely account for the difference in coverage between *oriC* and the *terC* site in cells undergoing interference vs. cells not undergoing interference; Supplementary Table 3). **b,** High-throughput sequencing analysis of genomic DNA: comparison of library construction methods. Left, steps in library construction using a NEBNext ultra II kit (analysis of double-stranded DNA) or Accel NGS 1S plus kit (analysis of single-stranded DNA). Right, PPS-region coverage plots obtained for wild-type cells.



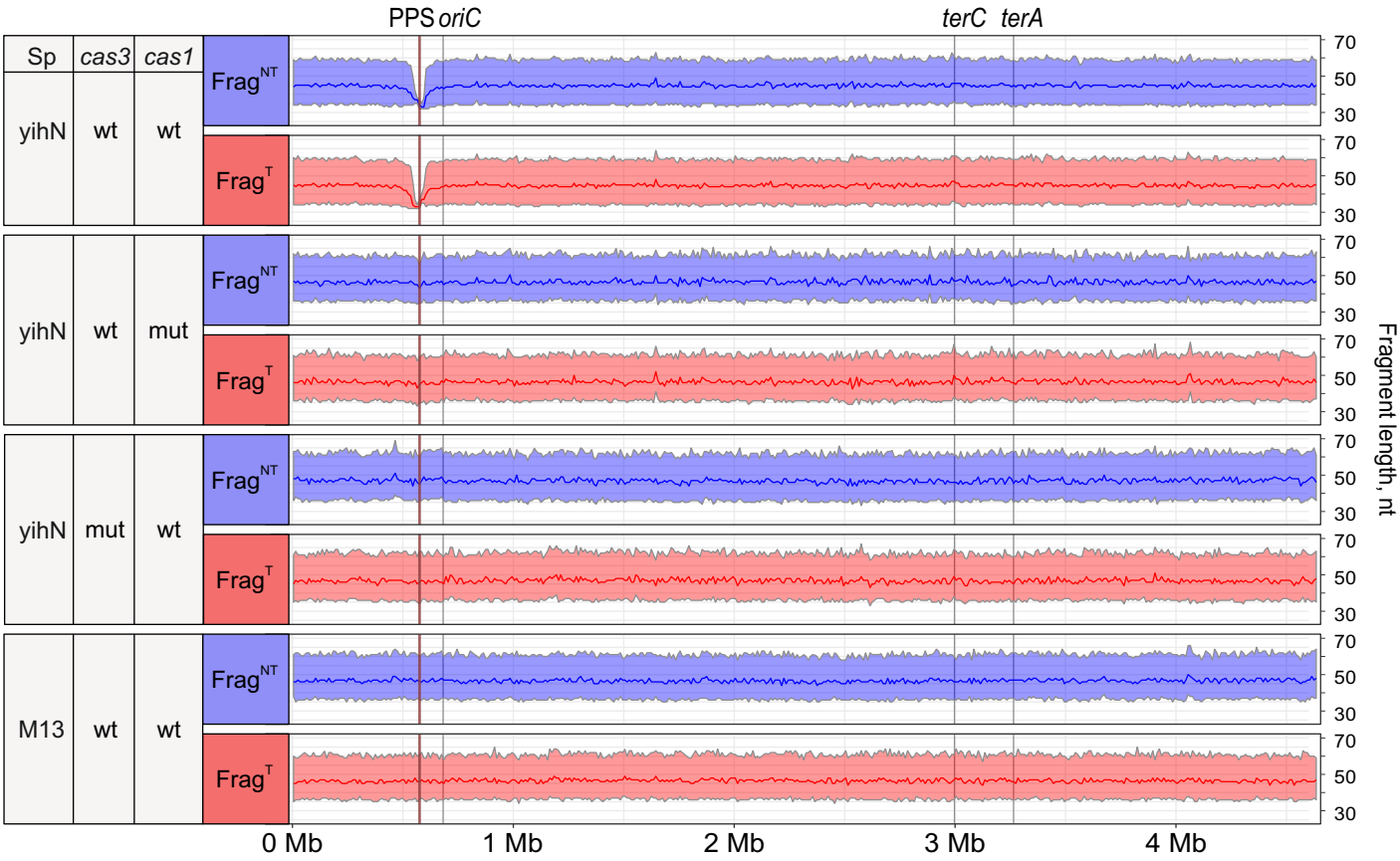
Supplementary Figure 2. Strand-specific, high-throughput sequencing of DNA fragments, FragSeq.

Steps in library construction. 5' app, adenylated 5' end; 3' ddC, blocked 3' end; N9 and N11, unique molecular identifiers on 3' and 5' adapters; purple rectangle, 4-nt barcode on 5' adapter; yellow rectangle, index.



Supplementary Figure 3. FragSeq results for the type I-E self-targeting system: fragment coverage plots and sequence analysis.

a, Genomic coverage plots. Percentage of total DNA fragments per 1 kb for the indicated strains (mean of three biological replicates). Coordinates on the X-axis represent the location on the *E. coli* chromosome. Blue, nontarget-strand-derived fragments ($Frag^{NT}$); red, target-strand-derived fragments ($Frag^{T}$). **b,** Sequence features of PPS-region fragments and adjacent chromosomal region. Plot shows heat map of relative abundance of A, T, C, or G for the indicated fragment 5' or 3' ends. Ten positions of sequences that are detected in fragment 5' or 3' ends are shown in black rectangles. Shading represents enrichment (>1) or depletion (<1) of each nucleotide for sequences associated with PPS-region-derived fragments vs. sequences associated with non-PPS-region-derived fragments.



Supplementary Figure 4. FragSeq results for the type I-E self-targeting system: length distributions.

Length distributions of genome-derived fragments in the indicated strains. Coordinates on the X-axis represent the location on the *E. coli* chromosome. Solid lines represent the median fragment length per 10 kb, shaded areas represent fragment lengths between the first and third quartiles.

Supplementary Table 1. Strains used in this study

Name	Description	Source
KD403	K-12 F^+, *lacUV5-cas3 araBp8-cse1*, CRISPR I: repeat-Sp^{yihN}-repeat, CRISPR II deleted. Sp^{yihN} (TCAAACAACCGACCTTGTTGTTCGCTATTGCC) targets chromosomal protospacer PPS (CCAAACAACCGACCTTGTTGTTCGCTATTGCC) within *yihN* gene forming a mismatch between crRNA and PPS at position +1.	This study
KD518	Like KD403, except Cas1 H208A	This study
KD753	Like KD403, except Cas3 H74A	This study
KD263	Like KD403, except CRISPR I: repeat-Sp^{M13}-repeat. Sp^{M13} (CTGTCTTTCGCTGCTGAGGGTGACGATCCCGC) targets g8 gene of M13 phage.	Shmakov *et al.*[1]
BL21-AI	F^-*ompT hsdSB (rB- mB-) gal dcm araB::T7RNAP-tetA*	Invitrogen
KD675	BL21-AI_ΔCRISPR carrying *Pseudomonas aeruginosa* CRISPR array with a single spacer (ACGCAGTTGCTGAGTGTGATCGATGCCATCAG) and a protospacer with a mismatch at position +1 (TCGCAGTTGCTGAGTGTGATCGATGCCATCAG) preceded by a functional GG PAM introduced into *ompL*/*yihN* intergenic region corresponding to the positions 4372171-4372261 of NC_012947	Vorontsova *et al.*[2]

Supplementary Table 2. Statistics for sequencing total genomic DNA purified from self-targeting strain KD403 (with or without induction of *cas* genes expression)

Library preparation	Strain	Number of reads aligned to the genome	Mean coverage, all genome	Mean coverage, PPS-flanking regions*	Mean coverage, PPS	Mean coverage, *terC*	Ratio of coverage: PPS-flanking regions / PPS	Ratio of coverage: PPS-flanking regions / *terC*
NEBNext® Ultra™ II DNA Library Prep Kit (NEB)	KD403 -I	3413501	47.4	50.1	45.6	40.0	1.1	1.3
Accel-NGS® 1S Plus DNA Library Kit (Swift Biosciences)	KD403 -I	4666914	64.8	70.9	70.5	64.2	1.0	1.1
NEBNext® Ultra™ II DNA Library Prep Kit (NEB)	KD403 +I	2117093	48.5	52.0	0.3	38.2	183.8	1.4
Accel-NGS® 1S Plus DNA Library Kit (Swift Biosciences)	KD403 +I	2974368	41.2	47.8	0.2	37.5	251.9	1.3

*Mean coverage over PPS-flanking regions was calculated as a mean of coverage 200 kb upstream and 100 kb downstream of the PPS

Supplementary Table 3. Statistics for sequencing total genomic DNA purified from induced self-targeting strain and control *cas1* mutant (Cas1 H208A), *cas3* mutant (Cas3 H74A) and nontargeting cells

Libraries were prepared only using NEBNext® Ultra™ II DNA Library Prep Kit for Illumina (NEB)

Strain	Replica	Number of reads aligned to the genome	Mean coverage	Mean coverage, PPS-flanking regions*	Mean coverage, PPS	Mean coverage, *terC*	Ratio of coverage: PPS-flanking regions / PPS**	Ratio of coverage: PPS-flanking regions / *terC****
KD263 (nontargeting)	1	7351683	154.6	166.0	147.6	100.8	1.1	1.6
	2	7313371	155.6	166.8	153.1	99.7	1.1	1.7
	3	5953547	123.4	132.9	117.6	79.8	1.1	1.7
KD753 (*cas3* mutant)	1	5826829	122.0	134.5	105.0	77.2	1.3	1.7
	2	1875434	39.0	43.5	33.7	24.7	1.3	1.8
	3	1887489	40.2	44.9	34.8	25.9	1.3	1.7
KD403 (self-targeting)	1	1148014	22.0	32.7	7.0	11.4	4.7	2.9
	2	886656	16.8	23.9	4.5	8.5	5.3	2.8
	3	632137	11.9	16.0	3.2	6.4	5.0	2.5
KD518 (*cas1* mutant)	1	844872	15.9	23.4	5.2	7.3	4.5	3.2
	2	766838	14.6	25.4	7.0	6.6	3.6	3.8
	3	1108980	20.9	33.5	7.1	9.9	4.7	3.4

*Mean coverage over PPS-flanking regions was calculated as a mean of coverage 200 kb upstream and 100 kb downstream of the PPS

**The ratio of genomic coverage between PPS-flanking regions and PPS is greater in self-targeting wild-type or *cas1* mutant strain compared to self-targeting *cas3* mutant or nontargeting cells in which Sp^{yihN} is replaced by a spacer targeting M13 phage (Sp^{M13}) (Cultures capable of interference vs. cultures incapable of interference, p-value = 0.001, Wilcoxon rank sum test).

***The ratio of genomic coverage between region in proximity to the *oriC* and the *terC* sites is greater in self-targeting wild-type or *cas1* mutant strain compared to self-targeting *cas3* mutant or nontargeting cells in which Sp^{yihN} is replaced by a spacer targeting M13 phage (Sp^{M13}) (Cultures capable of interference vs. cultures incapable of interference, p-value = 0.001, Wilcoxon rank sum test).

Supplementary Table 4. Statistics for sequencing spacers acquired during primed adaptation in self-targeting KD403 strain (number of protospacers on each strand upstream or downstream of the PPS)

Replica	Slipped and flipped AAG-protospacers	Number of newly acquired spacers sequenced	% protospacers from total number of protospacers			
			Protospacers in region 100 kb upstream of the PPS		Protospacers in region 100 kb downstream of the PPS	
			Nontarget strand	Target strand	Nontarget strand	Target strand
1	Included	1031147	57.3	1.2	0.8	39.8
	Removed	1005715	57.6	0.8	0.6	40.1
2	Included	1108683	57	1.2	0.9	40.3
	Removed	1079093	57.3	0.8	0.6	40.6
3	Included	1087616	56.4	1.6	1.3	38.9
	Removed	1058208	57	1	0.9	39.3

Supplementary Table 5. Statistics for sequencing spacers acquired during primed adaptation in self-targeting KD403 strain (% protospacers flanked by AAG PAM on each strand upstream or downstream of the PPS)

Replica	Slipped and flipped AAG-protospacers	% protospacers flanked by AAG PAM in the analyzed region			
		Protospacers in region 100 kb upstream of the PPS		Protospacers in region 100 kb downstream of the PPS	
		Nontarget strand	Target strand	Nontarget strand	Target strand
1	Included	95.4	45.1	52.5	95.8
	Removed	97.2	66.8	76.9	97.5
2	Included	95.2	43.3	49.1	95.7
	Removed	97.3	66.4	73.5	97.5
3	Included	96.4	44	56.8	96.7
	Removed	98	74	85.1	98.3

Supplementary Table 6. Statistics for sequencing short DNA fragments generated *in vivo* in type I-E system

Strain	Replica	Amount of reads before removal of overamplified reads	Amount of reads after removal of overamplified reads	Reads uniquely aligned to the genome
KD263 (nontargeting)	1	5694320	991750	730865
	2	1118248	272398	198314
	3	1068441	538644	436247
KD753 (*cas3* mutant)	1	1792903	564591	430276
	2	4061792	797131	619071
	3	765178	200324	130917
KD403 (self-targeting)	1	6329070	1853456	1479620
	2	652932	427896	283221
	3	4193159	1964365	1650564
KD518 (*cas1* mutant)	1	151277	115932	74883
	2	3292762	1155851	897906
	3	247651	211711	130165

Supplementary Table 7. Statistics for sequencing short DNA fragments generated *in vivo* in type I-E system (reads in 50-kb PPS-containing region*)

Strain	Replica	Total number of reads	Reads mapped to the PPS-containing region*	Reads in PPS-containing region with either 5'-CTTNN-3' or 5'-AAG-3' motif**	Reads mapped to the PPS-containing region, % from total	Reads in PPS-containing region with either 5'-CTTNN-3' or 5'-AAG-3' motif**, % from all reads in the PPS-containing region
KD263 (nontargeting)	1	730865	8377	352	1.15	4.20
	2	198314	2401	103	1.21	4.29
	3	436247	5009	219	1.15	4.37
KD753 (*cas3* mutant)	1	430276	5497	342	1.28	6.22
	2	619071	8030	478	1.30	5.95
	3	130917	1619	88	1.24	5.44
KD403 (self-targeting)	1	1479620	69878	44684	4.72	63.95
	2	283221	12105	8031	4.27	66.34
	3	1650564	57947	34529	3.51	59.59
KD518 (*cas1* mutant)	1	74883	841	42	1.12	4.99
	2	897906	10875	551	1.21	5.07
	3	130165	1415	79	1.09	5.58

* PPS-containing region is a region spanning 25 kb upstream and 25 kb downstream of the PPS

**Reads with 5'-CTTNN-3' motif have this motif on their 3' ends; reads with 5'-AAG-3' motif have either 5'-A/AG-3' or 5'-AA/G-3' motif on their 5' ends

Supplementary Table 8. Statistics for sequencing short DNA fragments generated *in vivo* in type I-E system (reads outside of 50-kb PPS-containing region*)

Strain	Replica	Total number of reads	Reads mapped outside of the PPS-containing region*	Reads outside of PPS-containing region* with either 5'-CTTNN-3' or 5'-AAG-3' motif**	Reads outside of PPS-containing region with either 5'-CTTNN-3' or 5'-AAG-3' motif**, % from all reads outside of the PPS-containing region*
KD263 (nontargeting)	1	730865	722488	27395	3.79
	2	198314	195913	7084	3.62
	3	436247	431238	15956	3.70
KD753 (*cas3* mutant)	1	430276	424779	15639	3.68
	2	619071	611041	21928	3.59
	3	130917	129298	4696	3.63
KD403 (self-targeting)	1	1479620	1409742	66930	4.75
	2	283221	271116	12702	4.69
	3	1650564	1592617	73092	4.59
KD518 (*cas1* mutant)	1	74883	74042	3072	4.15
	2	897906	887031	32005	3.61
	3	130165	128750	4803	3.73

* PPS-containing region is a region spanning 25 kb upstream and 25 kb downstream of the PPS

**Reads with 5'-CTTNN-3' motif have this motif on their 3' ends; reads with 5'-AAG-3' motif have either 5'-A/AG-3' or 5'-AA/G-3' motif on their 5' ends

Supplementary Table 9. Statistics for sequencing short DNA fragments generated *in vivo* in type I-E system (reads mapped to the target strand in 25-kb region upstream of the PPS, "25 kb TS up;" or nontarget strand in 25-kb region downstream of the PPS, "25 kb NS dw")

Strain	Replica	Total number of reads	Reads in 25 kb TS up or 25 kb NS dw region	Reads in 25 kb TS up or 25 kb NS dw region, 36-38 nt	Reads in 25 kb TS up or 25 kb NS dw region, 36-38 nt, with 5'-CTTNN-3' motif*	Reads in 25 kb TS up or 25 kb NS dw region, 36-38 nt, with 5'-CTTNN-3' motif*, % from total number of reads	Reads in 25 kb TS up or 25 kb NS dw region, 36-38 nt, with 5'-CTTNN-3' motif*, % from 36-38 nt reads in this region
KD263 (nontargeting)	1	730865	3903	245	3	4.1E-04	1.22
	2	198314	1129	77	0	0.0E+00	0.00
	3	436247	2527	254	0	0.0E+00	0.00
KD753 (*cas3* mutant)	1	430276	2601	233	45	1.0E-02	19.31
	2	619071	3681	328	29	4.7E-03	8.84
	3	130917	760	62	1	7.6E-04	1.61
KD403 (self-targeting)	1	1479620	29055	16134	13953	9.4E-01	86.48
	2	283221	5131	2911	2586	9.1E-01	88.84
	3	1650564	24337	12638	10761	6.5E-01	85.15
KD518 (*cas1* mutant)	1	74883	419	26	1	1.3E-03	3.85
	2	897906	5242	428	5	5.6E-04	1.17
	3	130165	724	54	4	3.1E-03	7.41

*Reads with 5'-CTTNN-3' motif have this motif on their 3' ends

Supplementary Table 10. Statistics for sequencing short DNA fragments generated *in vivo* in type I-E system (reads mapped to the nontarget strand in 25-kb region upstream of the PPS, "25 kb NS up;" or target strand in 25-kb region downstream of the PPS, "25 kb TS dw")

Strain	Replica	Total number of reads	Reads in 25 kb NS up or 25 kb TS dw region	Reads in 25 kb NS up or 25 kb TS dw region, 32-34 nt	Reads in 25 kb NS up or 25 kb TS dw region, 32-34 nt, with 5'-AAG-3' motif*	Reads in 25 kb NS up or 25 kb TS dw region, 32-34 nt, with 5'-AAG-3' motif*, % from total number of reads	Reads in 25 kb NS up or 25 kb TS dw region, 32-34 nt, with 5'-AAG-3' motif*, % from 32-34 nt reads in this region
KD263 (nontargeting)	1	730865	4459	246	10	1.4E-03	4.07
	2	198314	1269	71	2	1.0E-03	2.82
	3	436247	2477	184	3	6.9E-04	1.63
KD753 (*cas3* mutant)	1	430276	2867	217	56	1.3E-02	25.81
	2	619071	4271	320	87	1.4E-02	27.19
	3	130917	839	58	11	8.4E-03	18.97
KD403 (self-targeting)	1	1479620	40665	23868	21573	1.5E+00	90.38
	2	283221	6937	4166	3846	1.4E+00	92.32
	3	1650564	33479	18777	16994	1.0E+00	90.50
KD518 (*cas1* mutant)	1	74883	420	54	4	5.3E-03	7.41
	2	897906	5523	448	96	1.1E-02	21.43
	3	130165	676	53	12	9.2E-03	22.64

*Reads with 5'-AAG-3' motif have either 5'-A/AG-3' or 5'-AA/G-3' motif on their 5' ends

Supplementary Table 11. Correlation between 32- to 34-nt 5'-AAG-3'-associated fragments and 36- to 38-nt 5'-CTT-3' associated fragments in self-targeting KD403 strain

	Fragments with either 5'-A/AG-3' or 5'-AA/G-3' motif on their 5' ends
Fragments with 5'-CTTNN-3' motif on their 3' ends	r=0.48 (95% confidence interval: 0.45-0.51); t = 27.667, df = 2538, p-value < 2.2e-16

Supplementary Table 12. Correlation between number of spacers and corresponding prespacers (DNA fragments conjugated to respective PAM) in self-targeting KD403 strain

	32- to 34-nt fragments with either 5'-A/AG-3' or 5'-AA/G-3' motif on their 5' ends	36- to 38-nt fragments with 5'-CTTNN-3' motif on their 3' ends
Spacers	r=0.57 (95% confidence interval: 0.54-0.59); t = 36.772, df = 2826, p-value < 2.2e-16	r=0.5 (95% confidence interval: 0.48-0.53); t = 30.976, df = 2826, p-value < 2.2e-16

Supplementary Table 13. Oligonucleotides used for prespacer efficiency assay

#	Transforming oligo names	Transforming oligo sequences
1.	G_33	5'GCCCAATTTACTACTCGTTCTGGTGTTTCTCGT 3' 3'CGGGTTAAATGATGAGCAAGACCACAAAGAGCA 5'
	C_33	
2.	AAG_35	5' AAGCCCAATTTACTACTCGTTCTGGTGTTTCTCGT 3' 3' TTCGGGTTAAATGATGAGCAAGACCACAAAGAGCA 5'
	TTC_35	
3.	G_33	5' GCCCAATTTACTACTCGTTCTGGTGTTTCTCGT 3' 3' AGTTCGGGTTAAATGATGAGCAAGACCACAAAGAGCA 5'
	AGTTC_37	
4.	AG_34	5' AGCCCAATTTACTACTCGTTCTGGTGTTTCTCGT 3' 3' AGTTCGGGTTAAATGATGAGCAAGACCACAAAGAGCA 5'
	AGTTC_37	
5.	G_32	5' GCCCAATTTACTACTCGTTCTGGTGTTTCTCG 3' 3' AGTTCGGGTTAAATGATGAGCAAGACCACAAAGAGCA 5'
	AGTTC_37	
6.	AG_33	5' AGCCCAATTTACTACTCGTTCTGGTGTTTCTCG 3' 3' AGTTCGGGTTAAATGATGAGCAAGACCACAAAGAGCA 5'
	AGTTC_37	

*Nucleotides corresponding to the PAM are written in red

Supplementary Table 14. Prespacer efficiency assay (overall level of adaptation and source of new spacers)

Transforming oligo	Replica	Number of CRISPR arrays	Spacers aligned to genome or pCas1+2	Spacers aligned only to oligo	% of CRISPR arrays elongated due to incorporation of oligo-derived spacer	% of CRISPR arrays elongated due to incorporation of a spacer from the genome or pCas1+2
G_33 + C_33	1	519963	38830	3711	0.7	7.5
	2	379132	26417	3122	0.8	7.0
	3	528544	32241	5595	1.1	6.1
AAG_35 + TTC_35	1	1242907	65814	113924	9.2	5.3
	2	845820	53575	92624	11.0	6.3
	3	847249	42347	116434	13.7	5.0
G_33 + AGTTC_37	1	958220	54203	115681	12.1	5.7
	2	860995	49232	110677	12.9	5.7
	3	1062383	52337	147334	13.9	4.9
AG_34 + AGTTC_37	1	813150	38076	88079	10.8	4.7
	2	773041	31635	80185	10.4	4.1
	3	309799	13061	34515	11.1	4.2
G_32 + AGTTC_37	1	530912	30089	18224	3.4	5.7
	2	496235	28562	21304	4.3	5.8
	3	962226	53517	53987	5.6	5.6
AG_33 + AGTTC_37	1	623827	39669	15692	2.5	6.4
	2	911818	47412	20233	2.2	5.2
	3	370459	19469	9912	2.7	5.3

Supplementary Table 15. Prespacer efficiency assay (insertion of properly processed* oligo only)

Transforming oligo	Replica	Number of CRISPR arrays	Properly processed oligo-derived spacers*	% of CRISPR arrays elongated due to incorporation of a properly processed oligo-derived spacer*	Direct orientation**	Reverse orientation**	Direct orientation**, %	Reverse orientation**, %
G_33 + C_33	1	519963	3385	0.7	1973	1412	58.3	41.7
	2	379132	2878	0.8	1690	1188	58.7	41.3
	3	528544	5140	1.0	2984	2156	58.1	41.9
AAG_35 + TTC_35	1	1242907	103392	8.3	102624	768	99.3	0.7
	2	845820	85060	10.1	84522	538	99.4	0.6
	3	847249	105799	12.5	105173	626	99.4	0.6
G_33 + AGTTC_37	1	958220	86771	9.1	86339	432	99.5	0.5
	2	860995	78527	9.1	78292	235	99.7	0.3
	3	1062383	100267	9.4	99950	317	99.7	0.3
AG_34 + AGTTC_37	1	813150	71646	8.8	71345	301	99.6	0.4
	2	773041	64577	8.4	64380	197	99.7	0.3
	3	309799	26567	8.6	26513	54	99.8	0.2
G_32 + AGTTC_37	1	530912	1875	0.4	1695	180	90.4	9.6
	2	496235	633	0.1	546	87	86.3	13.7
	3	962226	1328	0.1	1147	181	86.4	13.6
AG_33 + AGTTC_37	1	623827	1918	0.3	1639	279	85.5	14.5
	2	911818	1101	0.1	941	160	85.5	14.5
	3	370459	376	0.1	318	58	84.6	15.4

* We define properly processed oligos as the oligos that were processed between an A and a G in the PAM sequence (T and C in PAM-complementary sequence) and integrated as a 33 bp spacer

Properly processed oligos can be integrated in direct (spacer starts with G; **GCCCAATTTACTACTCGTTCTGGTGTTTCTCGT) or reverse (spacer ends with C; ACGAGAAACACCAGAACGAGTAGTAAATTGGG**C**) orientation

Supplementary Table 16. Prespacer efficiency assay (length of oligo-derived spacers)

Transforming oligo	Replica	% oligo-derived spacers of 33 bp length
G_33 + C_33	1	91.2
	2	92.3
	3	91.9
AAG_35 + TTC_35	1	91.5
	2	93.1
	3	92.2
G_33 + AGTTC_37	1	91.2
	2	91.8
	3	90.7
AG_34 + AGTTC_37	1	90.8
	2	91.4
	3	90.7
G_32 + AGTTC_37	1	87.8
	2	87.6
	3	87
AG_33 + AGTTC_37	1	88.5
	2	88.6
	3	86.8

Supplementary Table 17. Statistics for sequencing short DNA fragments generated *in vivo* in type I-F system

Strain	Replica	Amount of reads before removal of overamplified reads	Amount of reads after removal of overamplified reads	Reads uniquely aligned to the genome
KD675 (*E.coli* strain with type I-F self-targeting system)	1	12437786	9739694	7034061
	2	12256224	9792344	7128753

Supplementary Table 18. Statistics for sequencing short DNA fragments generated *in vivo* in type I-F system (reads mapped to the target strand in 5-kb region upstream of the PPS, "5 kb TS up;" or nontarget strand in 5-kb region downstream of the PPS, "5 kb NS dw")

Strain	Replica	Total number of reads	Reads in 5 kb TS up or 5 kb NS dw region	Reads in 5 kb TS up or 5 kb NS dw region, 31-32 nt	Reads in 5 kb TS up or 5 kb NS dw region, 31-32 nt, with 5'-CCA-3' motif*	Reads in 5 kb TS up or 5 kb NS dw region, 31-32 nt, with 5'-CCA-3' motif*, % from 31-32 nt reads in this region
KD675 (*E.coli* strain with type I-F self-targeting system)	1	7034061	5740	696	454	**65.23**
	2	7128753	2665	280	145	**51.79**

*Reads with 5'-CCA-3' motif have 5'CC/A-3' motif on their 5' ends

Supplementary Table 19. Statistics for sequencing short DNA fragments generated *in vivo* in type I-F system (reads mapped to the nontarget strand in 5-kb region upstream of the PPS, "5 kb NS up;" or target strand in 5-kb region downstream of the PPS, "5 kb TS dw")

Strain	Replica	Total number of reads	Reads in 5 kb NS up or 5 kb TS dw region	Reads in 5 kb NS up or 5 kb TS dw region, 37-38 nt	Reads in 5 kb NS up or 5 kb TS dw region, 37-38 nt, with 5'-TGGNNN-3' motif*	Reads in 5 kb NS up or 5 kb TS dw region, 37-38 nt, with 5'-TGGNNN-3' motif*, % from 37-38 nt reads in this region
KD675 (*E.coli* strain with type I-F self-targeting system)	1	7034061	5325	751	260	**34.62**
	2	7128753	2520	383	123	**32.11**

*Reads with 5'-TGGNNN-3' motif have this motif on their 3' ends

Supplementary Table 20. Statistics for sequencing short DNA fragments generated *in vivo* in type I-F system (reads mapped outside of the 10-kb PPS-containing region; 31-32 nt reads)

Strain	Replica	Total number of reads	Reads outside of the PPS-containing region	Reads outside of the PPS-containing region, 31-32 nt	Reads outside of the PPS-containing region, 31-32 nt, with 5'-CCA-3' motif*	Reads outside of the PPS-containing region, 31-32 nt, with 5'-CCA-3' motif*, % from 31-32 nt reads in this region
KD675 (*E.coli* strain with type I-F self-targeting system)	1	7034061	7022770	288270	16846	**5.84**
	2	7128753	7123339	275324	18138	**6.59**

*Reads with 5'-CCA-3' motif have 5'CC/A-3' motif on their 5' ends

Supplementary Table 21. Statistics for sequencing short DNA fragments generated *in vivo* in type I-F system (reads mapped outside of the 10-kb PPS-containing region; 37-38 nt reads)

Strain	Replica	Total number of reads	Reads outside of the PPS-containing region	Reads outside of the PPS-containing region, 37-38 nt	Reads outside of the PPS-containing region, 37-38 nt, with 5'-TGGNNN-3' motif*	Reads outside of the PPS-containing region, 37-38 nt, with 5'-TGGNNN-3' motif*, % from 37-38 nt reads in this region
KD675 (*E.coli* strain with type I-F self-targeting system)	1	7034061	7022770	387537	9720	**2.51**
	2	7128753	7123339	422796	9866	**2.33**

*Reads with 5'-TGGNNN-3' motif have this motif on their 3' ends

Supplementary Table 22. List of primers used for amplification of CRISPR array.

Name	Sequence (5' to 3')	Purpose
LDR-F2	ATGCTTTAAGAACAAATGTATACTTTTAG	Monitoring primed adaptation in KD263 and KD403
Ec_minR	CGAAGGCGTCTTGATGGGTTTG	
LDR-F2	ATGCTTTAAGAACAAATGTATACTTTTAG	High-throughput sequencing of spacers acquired during primed adaptation in KD403
autoSp2_R	AATAGCGAACAACAAGGTCGGTTG	
BLCRdir	GGTAGATTGTGACTGGCTTAAAAAATC	High-throughput sequencing of spacers acquired during prespacer efficiency assay in BL21-AI
BLCRreverse	GTTTGAGCGATGATATTTGTGCTC	

Supplementary Table 23. List of adapters used for FragSeq

Name	Sequence (5' to 3')	Description
i112	GTTCAGAGTTCTACAGTCCGACGATCCTGANNNNNNNNNN	5′ adapter with CTGA barcode and 11N extension used in KD263 short DNA fragments library preparation (barcode is underlined)
i113	GTTCAGAGTTCTACAGTCCGACGATCGACTNNNNNNNNNN	5′ adapter with GACT barcode and 11N extension used in KD753 short DNA fragments library preparation (barcode is underlined)
i114	GTTCAGAGTTCTACAGTCCGACGATCAGTCNNNNNNNNNN	5′ adapter with AGTC barcode and 11N extension used in KD403 short DNA fragments library preparation (barcode is underlined)
i115	GTTCAGAGTTCTACAGTCCGACGATCTCAGNNNNNNNNNN	5′ adapter with TCAG barcode and 11N extension used in KD518 and KD675 short DNA fragments library preparation (barcode sequence is underlined)
i116	Phos/NNNNNNNNNTGGAATTCTCGGGTGCCAAGG/ddC/	3′ adapter with 9N random sequence used in short DNA fragments library preparation

Supplementary Table 24. List of Illumina primers used for amplification of FragSeq libraries.

Name	Sequence (5' to 3')	Sample
RP1	AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA	All samples
RPI3	CAAGCAGAAGACGGCATACGAGAT**GCCTAA**GTGACTGGAGTTCCTTGGCACCCGA GAATTCCA	KD263, replicate 1
RPI4	CAAGCAGAAGACGGCATACGAGAT**TGGTCA**GTGACTGGAGTTCCTTGGCACCCGA GAATTCCA	KD263, replicate 2
RPI5	CAAGCAGAAGACGGCATACGAGAT**CACTGT**GTGACTGGAGTTCCTTGGCACCCGA GAATTCCA	KD263, replicate 3
RPI6	CAAGCAGAAGACGGCATACGAGAT**ATTGGC**GTGACTGGAGTTCCTTGGCACCCGA GAATTCCA	KD753, replicate 1
RPI7	CAAGCAGAAGACGGCATACGAGAT**GATCTG**GTGACTGGAGTTCCTTGGCACCCGA GAATTCCA	KD753, replicate 2
RPI8	CAAGCAGAAGACGGCATACGAGAT**TCAAGT**GTGACTGGAGTTCCTTGGCACCCGA GAATTCCA	KD753, replicate 3
RPI9	CAAGCAGAAGACGGCATACGAGAT**CTGATC**GTGACTGGAGTTCCTTGGCACCCGA GAATTCCA	KD403, replicate 1
RPI10	CAAGCAGAAGACGGCATACGAGAT**AAGCTA**GTGACTGGAGTTCCTTGGCACCCGA GAATTCCA	KD403, replicate 2
RPI11	CAAGCAGAAGACGGCATACGAGAT**GTAGCC**GTGACTGGAGTTCCTTGGCACCCG AGAATTCCA	KD403, replicate 3
RPI12	CAAGCAGAAGACGGCATACGAGAT**TACAAG**GTGACTGGAGTTCCTTGGCACCCGA GAATTCCA	KD518, replicate 1
RPI13	CAAGCAGAAGACGGCATACGAGAT**TTGACT**GTGACTGGAGTTCCTTGGCACCCGA GAATTCCA	KD518, replicate 2
RPI14	CAAGCAGAAGACGGCATACGAGAT**GGAACT**GTGACTGGAGTTCCTTGGCACCCG AGAATTCCA	KD518, replicate 3
RPI20	CAAGCAGAAGACGGCATACGAGAT**GGCCAC**GTGACTGGAGTTCCTTGGCACCCG AGAATTCCA	KD675 replicate 1
RPI21	CAAGCAGAAGACGGCATACGAGAT**CGAAAC**GTGACTGGAGTTCCTTGGCACCCG AGAATTCCA	KD675 replicate 2

PCR was performed with TruSeq Small RNA RP1 primer and one of RPI index primers (index sequence is shown in bold).

Supplementary References

1. Shmakov, S. et al. Pervasive generation of oppositely oriented spacers during CRISPR adaptation. Nucleic Acids Res 42, 5907-5916, (2014).
2. Vorontsova, D. et al. Foreign DNA acquisition by the I-F CRISPR-Cas system requires all components of the interference machinery. Nucleic Acids Res 43, 10848-10860, (2015).

CONCLUSIONS

In this work we biochemically characterized several Cas effectors of different CRISPR-Cas systems types.

These studies show that CRISPR-Cas nucleases, though relying on common mechanisms of invader's DNA recognition through guide RNAs, often have completely different domain organization, modes of action and unique properties, which together determines the areas of their potential biotechnological applications.

The contrast between these enzymes is clearly seen on the example of Cas9 and Cas12 enzymes coming, correspondingly, from Type II and Type V CRISPR-Cas systems. Indeed, the HNH domain responsible for cleavage of the target DNA strand in Cas9 is absent in Cas12 enzymes. Cas12 effectors use additional Nuc domain for DNA strands replacement in the RuvC active site, allowing double-stranded DNA cleavage by a single nuclease domain. Thus, although both Cas9 and Cas12 introduce double stranded DNA breaks, they rely on different domains and moreover, have different modes of action.

Besides structural differences between Type II and Type V effectors, studies of Cas12a enzymes revealed a new feature of these effectors: the ability to process pre-crRNA into mature crRNAs without any help from other enzymes. We used this property of AsCas12a to develop a multiplex gene editing system, which was successfully applied for genome modification in different organisms (108, 123).

The diversity of Cas effectors is naturally smaller between enzymes of the same CRISPR-Cas systems Type. We used high throughput sequencing to show that DpbCas12e and AsCas12a produce very similar DNA cleavage pattern, generating 3-5 nt 5'-overhangs. Similarly to Cas12a, Cas12e can be programmed to produce longer 5'-overhags by the use of crRNAs with shorter spacer segments. Despite on these similarities, Cas12e in contrast to Cas12a doesn't demonstrate target-activated nonspecific ssDNA cleavage (113).

Narrowing down to enzymes of the same subtype of CRISPR-Cas systems, the Type II-C Cas9 proteins characterized in this work, we can see that exploitable Cas effectors diversity persists even at this level. CcCas9, DfCas9, and PpCas9 have characteristic for II-C effectors small sizes but require distinct PAM sequences, guide RNAs and have different temperature preferences.

PpCas9 demonstrated activity in human cells and potentially can be used for genome editing, although additional studies of its efficiency and specificity should be performed. Indeed, PpCas9 showed different targeting efficiency on different genes, which may reflect its preference for certain genome methylation patterns, folding of DNA and/or other factors. Elaborate

comparison of PpCas9 activity and specificity with SpCas9, as well as SauriCas9 and Nme2Cas9, is important for future applications of PpCas9. Due to its small size, PpCas9 can be packaged into all-in-one AAV particles. We continue to work on developing of PpCas9-based genome editing systems by testing the activity of CRISPR-PpCas9 carrying AAV particles in mice.

While PpCas9 can be used for eukaryotic genome editing, DfCas9 and CcCas9 can potentially can find application in microbial biotechnology. In particular, knowing of CcCas9 PAM sequences and crRNA requirements makes possible genome modification of its host, a promising biofuel producer *Clostridium cellulolyticum*, through the use of endogenous CRISPR-Cas system.

Although the characterization of orthologues may appear straight forward, it has many challenges. Indeed, the targeted search for Cas9 enzymes active in human cells conducted by different research groups, when several orthologues are studied in parallel, shows that only a small fraction of Cas9 enzymes with demonstrated *in vitro* activity can introduce indels in eukaryotic genomes, (60). This can be explained by actual low efficiency of the nucleases as well as by improper folding of the enzymes during heterologous expression (when purified recombinant proteins or human cells lysates expressing Cas9 genes are used for *in vitro* DNA cleavage), not optimal sgRNA design, etc. Indeed, in contrast to PpCas9 and DfCas9, which actively cleave DNA in complex with corresponding sgRNAs *in vitro*, CcCas9 fails to cleave DNA with any of sgRNA variants we tested. Possibly, the folding of guide RNAs inside the CcCas9 globule is different from that in PpCas9, DfCas9, and SpCas9, and thus, a common way of introducing several nucleotides linkers does not work in this case. Similar problems with sgRNA design were described by Hu et al., when 30 Cas9 enzymes tested in human cells failed to cleave the eukaryotic genome with initially proposed sgRNAs sequences (98).

Despite all the challenges, in last several years several Cas9 nucleases were successfully applied for eukaryotic genome modification and their number seems to be growing steadily, similarly to the number of restriction enzymes, which were actively characterized and introduced into laboratory practice in the 1970s and 80s. In contrast to restriction enzymes, Cas9 nucleases are programmable, and the set of efficient needed Cas9 enzymes is limited by their PAM requirements to cover the whole range of genomic targets. The improvement of known nucleases properties, such as specificity, PAM requirements and efficiency, by mutagenesis and directed evolution approaches may also reduce the need of extensive Cas9 orthologs searches in the future (124, 125). In the meantime, the quest for most efficient and versatile Cas effectors will continue.

REFERENCES

1. Frost,L.S., Leplae,R., Summers,A.O. and Toussaint,A. (2005) Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.*, **3**, 722–732.

2. Paez-Espino,D., Eloe-Fadrosh,E.A., Pavlopoulos,G.A., Thomas,A.D., Huntemann,M., Mikhailova,N., Rubin,E., Ivanova,N.N. and Kyrpides,N.C. (2016) Uncovering Earth's virome. *Nature*, **536**, 425–430.

3. Clokie,M.R., Millard,A.D., Letarov,A.V. and Heaphy,S. (2011) Phages in nature. *Bacteriophage*, **1**, 31–45.

4. Suttle,C.A. (2005) Viruses in the sea. *Nature*, **437**, 356–361.

5. Koonin,E.V. (2016) Horizontal gene transfer: essentiality and evolvability in prokaryotes, and roles in evolutionary transitions. *F1000Res*, **5**.

6. Cohen,D., Melamed,S., Millman,A., Shulman,G., Oppenheimer-Shaanan,Y., Kacen,A., Doron,S., Amitai,G. and Sorek,R. (2019) Cyclic GMP–AMP signalling protects bacteria against viral infection. *Nature*, **574**, 691–695.

7. Koonin,E.V. (2017) Evolution of RNA- and DNA-guided antivirus defense systems in prokaryotes and eukaryotes: common ancestry vs convergence. *Biol Direct*, **12**, 5.

8. Shmakov,S., Abudayyeh,O.O., Makarova,K.S., Wolf,Y.I., Gootenberg,J.S., Semenova,E., Minakhin,L., Joung,J., Konermann,S., Severinov,K., *et al.* (2015) Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. *Mol. Cell*, **60**, 385–397.

9. Burstein,D., Harrington,L.B., Strutt,S.C., Probst,A.J., Anantharaman,K., Thomas,B.C., Doudna,J.A. and Banfield,J.F. (2017) New CRISPR–Cas systems from uncultivated microbes. *Nature*, **542**, 237–241.

10. Makarova,K.S., Wolf,Y.I., Snir,S. and Koonin,E.V. (2011) Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. *J. Bacteriol.*, **193**, 6039–6056.

11. Soutourina,O. (2019) Type I Toxin-Antitoxin Systems in Clostridia. *Toxins (Basel)*, **11**.

12. Harms,A., Brodersen,D.E., Mitarai,N. and Gerdes,K. (2018) Toxins, Targets, and Triggers: An Overview of Toxin-Antitoxin Biology. *Mol. Cell*, **70**, 768–784.

13. Mruk,I. and Kobayashi,I. (2014) To be or not to be: regulation of restriction-modification systems and other toxin-antitoxin systems. *Nucleic Acids Res.*, **42**, 70–86.

14. Gordeeva,J., Morozova,N., Sierro,N., Isaev,A., Sinkunas,T., Tsvetkova,K., Matlashov,M., Truncaite,L., Morgan,R.D., Ivanov,N.V., *et al.* (2019) BREX system of Escherichia coli distinguishes self from non-self by methylation of a specific DNA site. *Nucleic Acids Res.*, **47**, 253–265.

15. Lisitskaya,L., Aravin,A.A. and Kulbachinskiy,A. (2018) DNA interference and beyond: structure and functions of prokaryotic Argonaute proteins. *Nat Commun*, **9**, 5165.

16. Luria,S.E. (1953) Host-induced modifications of viruses. *Cold Spring Harb. Symp. Quant. Biol.*, **18**, 237–244.

17. Luria,S.E. and Human,M.L. (1952) A nonhereditary, host-induced variation of bacterial viruses. *J. Bacteriol.*, **64**, 557–569.

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

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

20. van Soolingen,D., de Haas,P.E., Hermans,P.W., Groenen,P.M. and van Embden,J.D. (1993) Comparison of various repetitive DNA elements as genetic markers for strain differentiation and epidemiology of Mycobacterium tuberculosis. *J. Clin. Microbiol.*, **31**, 1987–1995.

21. Pourcel,C., Salvignol,G. and Vergnaud,G. (2005) CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology (Reading, Engl.)*, **151**, 653–663.

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

23. Bolotin,A., Quinquis,B., Sorokin,A. and Ehrlich,S.D. (2005) Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology (Reading, Engl.)*, **151**, 2551–2561.

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

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

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

27. Garneau,J.E., Dupuis,M.-È., Villion,M., Romero,D.A., Barrangou,R., Boyaval,P., Fremaux,C., Horvath,P., Magadán,A.H. and Moineau,S. (2010) The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature*, **468**, 67–71.

28. Cong,L., Ran,F.A., Cox,D., Lin,S., Barretto,R., Habib,N., Hsu,P.D., Wu,X., Jiang,W., Marraffini,L.A., *et al.* (2013) Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science*, **339**, 819–823.

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

30. Lange,S.J., Alkhnbashi,O.S., Rose,D., Will,S. and Backofen,R. (2013) CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems. *Nucleic Acids Res.*, **41**, 8034–8044.

31. Savitskaya,E., Lopatina,A., Medvedeva,S., Kapustin,M., Shmakov,S., Tikhonov,A., Artamonova,I.I., Logacheva,M. and Severinov,K. (2017) Dynamics of Escherichia coli type I-E CRISPR spacers over 42 000 years. *Mol. Ecol.*, **26**, 2019–2026.

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

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

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

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

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

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

38. Alkhnbashi,O.S., Shah,S.A., Garrett,R.A., Saunders,S.J., Costa,F. and Backofen,R. (2016) Characterizing leader sequences of CRISPR loci. *Bioinformatics*, **32**, i576–i585.

39. González-Delgado,A., Mestre,M.R., Martínez-Abarca,F. and Toro,N. (2019) Spacer acquisition from RNA mediated by a natural reverse transcriptase-Cas1 fusion protein associated with a type III-D CRISPR-Cas system in Vibrio vulnificus. *Nucleic Acids Res.*, **47**, 10202–10211.

40. Behler,J. and Hess,W.R. (2020) Approaches to study CRISPR RNA biogenesis and the key players involved. *Methods*, **172**, 12–26.

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

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

43. Dillard,K.E., Brown,M.W., Johnson,N.V., Xiao,Y., Dolan,A., Hernandez,E., Dahlhauser,S.D., Kim,Y., Myler,L.R., Anslyn,E.V., *et al.* (2018) Assembly and Translocation of a CRISPR-Cas Primed Acquisition Complex. *Cell*, **175**, 934-946.e15.

44. Shibata,M., Nishimasu,H., Kodera,N., Hirano,S., Ando,T., Uchihashi,T. and Nureki,O. (2017) Real-space and real-time dynamics of CRISPR-Cas9 visualized by high-speed atomic force microscopy. *Nat Commun*, **8**, 1430.

45. Sternberg,S.H., Redding,S., Jinek,M., Greene,E.C. and Doudna,J.A. (2014) DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature*, **507**, 62–67.

46. Gleditzsch,D., Pausch,P., Müller-Esparza,H., Özcan,A., Guo,X., Bange,G. and Randau,L. (2019) PAM identification by CRISPR-Cas effector complexes: diversified mechanisms and structures. *RNA Biol*, **16**, 504–517.

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

48. Swarts,D.C. and Jinek,M. (2019) Mechanistic Insights into the cis- and trans-Acting DNase Activities of Cas12a. *Mol. Cell*, **73**, 589-600.e4.

49. Jiang,F., Taylor,D.W., Chen,J.S., Kornfeld,J.E., Zhou,K., Thompson,A.J., Nogales,E. and Doudna,J.A. (2016) Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science*, **351**, 867–871.

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

51. Jinek,M., Chylinski,K., Fonfara,I., Hauer,M., Doudna,J.A. and Charpentier,E. (2012) A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science*, **337**, 816–821.

52. Gasiunas,G., Barrangou,R., Horvath,P. and Siksnys,V. (2012) Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, E2579-2586.

53. Makarova,K.S., Wolf,Y.I., Iranzo,J., Shmakov,S.A., Alkhnbashi,O.S., Brouns,S.J.J., Charpentier,E., Cheng,D., Haft,D.H., Horvath,P., *et al.* (2020) Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.*, **18**, 67–83.

54. Shmakov,S.A., Makarova,K.S., Wolf,Y.I., Severinov,K.V. and Koonin,E.V. (2018) Systematic prediction of genes functionally linked to CRISPR-Cas systems by gene neighborhood analysis. *Proc Natl Acad Sci USA*, **115**, E5307–E5316.

55. Smargon,A.A., Cox,D.B.T., Pyzocha,N.K., Zheng,K., Slaymaker,I.M., Gootenberg,J.S., Abudayyeh,O.A., Essletzbichler,P., Shmakov,S., Makarova,K.S., *et al.* (2017) Cas13b Is a Type VI-B CRISPR-Associated RNA-Guided RNase Differentially Regulated by Accessory Proteins Csx27 and Csx28. *Mol. Cell*, **65**, 618-630.e7.

56. Makarova,K.S., Wolf,Y.I. and Koonin,E.V. (2018) Classification and Nomenclature of CRISPR-Cas Systems: Where from Here? *CRISPR J*, **1**, 325–336.

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

58. Nishimasu,H., Ran,F.A., Hsu,P.D., Konermann,S., Shehata,S.I., Dohmae,N., Ishitani,R., Zhang,F. and Nureki,O. (2014) Crystal Structure of Cas9 in Complex with Guide RNA and Target DNA. *Cell*, **156**, 935–949.

59. Kim,E., Koo,T., Park,S.W., Kim,D., Kim,K., Cho,H.-Y., Song,D.W., Lee,K.J., Jung,M.H., Kim,S., *et al.* (2017) In vivo genome editing with a small Cas9 orthologue derived from Campylobacter jejuni. *Nat Commun*, **8**, 14500.

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

61. Siksnys,V. and Gasiunas,G. (2016) Rewiring Cas9 to Target New PAM Sequences. *Mol. Cell*, **61**, 793–794.

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

63. Ran,F.A., Hsu,P.D., Lin,C.-Y., Gootenberg,J.S., Konermann,S., Trevino,A.E., Scott,D.A., Inoue,A., Matoba,S., Zhang,Y., *et al.* (2013) Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell*, **154**, 1380–1389.

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

65. Nemudryi,A.A., Valetdinova,K.R., Medvedev,S.P. and Zakian,S.M. (2014) TALEN and CRISPR/Cas Genome Editing Systems: Tools of Discovery. *Acta Naturae*, **6**, 19–40.

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

67. Koonin,E.V. (2018) Open questions: CRISPR biology. *BMC Biol.*, **16**, 95.

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

69. Anzalone,A.V., Randolph,P.B., Davis,J.R., Sousa,A.A., Koblan,L.W., Levy,J.M., Chen,P.J., Wilson,C., Newby,G.A., Raguram,A., *et al.* (2019) Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature*, 10.1038/s41586-019-1711-4.

70. Konermann,S., Brigham,M.D., Trevino,A.E., Joung,J., Abudayyeh,O.O., Barcena,C., Hsu,P.D., Habib,N., Gootenberg,J.S., Nishimasu,H., *et al.* (2015) Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature*, **517**, 583–588.

71. Rees,H.A. and Liu,D.R. (2018) Base editing: precision chemistry on the genome and transcriptome of living cells. *Nat. Rev. Genet.*, **19**, 770–788.

72. Whinn,K.S., Kaur,G., Lewis,J.S., Schauer,G.D., Mueller,S.H., Jergic,S., Maynard,H., Gan,Z.Y., Naganbabu,M., Bruchez,M.P., *et al.* (2019) Nuclease dead Cas9 is a programmable roadblock for DNA replication. *Sci Rep*, **9**, 13292.

73. Wang,H., La Russa,M. and Qi,L.S. (2016) CRISPR/Cas9 in Genome Editing and Beyond. *Annu. Rev. Biochem.*, **85**, 227–264.

74. Shalem,O., Sanjana,N.E., Hartenian,E., Shi,X., Scott,D.A., Mikkelson,T., Heckl,D., Ebert,B.L., Root,D.E., Doench,J.G., *et al.* (2014) Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science*, **343**, 84–87.

75. Mougiakos,I., Mohanraju,P., Bosma,E.F., Vrouwe,V., Finger Bou,M., Naduthodi,M.I.S., Gussak,A., Brinkman,R.B.L., van Kranenburg,R. and van der Oost,J. (2017) Characterizing a thermostable Cas9 for bacterial genome editing and silencing. *Nat Commun*, **8**, 1647.

76. Varshney,G.K., Pei,W., LaFave,M.C., Idol,J., Xu,L., Gallardo,V., Carrington,B., Bishop,K., Jones,M., Li,M., *et al.* (2015) High-throughput gene targeting and phenotyping in zebrafish using CRISPR/Cas9. *Genome Res.*, **25**, 1030–1042.

77. Woo,J.W., Kim,J., Kwon,S.I., Corvalán,C., Cho,S.W., Kim,H., Kim,S.-G., Kim,S.-T., Choe,S. and Kim,J.-S. (2015) DNA-free genome editing in plants with preassembled CRISPR-Cas9 ribonucleoproteins. *Nat. Biotechnol.*, **33**, 1162–1164.

78. Gantz,V.M., Jasinskiene,N., Tatarenkova,O., Fazekas,A., Macias,V.M., Bier,E. and James,A.A. (2015) Highly efficient Cas9-mediated gene drive for population modification of the malaria vector mosquito Anopheles stephensi. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E6736-6743.

79. Platt,R.J., Chen,S., Zhou,Y., Yim,M.J., Swiech,L., Kempton,H.R., Dahlman,J.E., Parnas,O., Eisenhaure,T.M., Jovanovic,M., *et al.* (2014) CRISPR-Cas9 knockin mice for genome editing and cancer modeling. *Cell*, **159**, 440–455.

80. Glemzaite,M., Balciunaite,E., Karvelis,T., Gasiunas,G., Grusyte,M.M., Alzbutas,G., Jurcyte,A., Anderson,E.M., Maksimova,E., Smith,A.J., *et al.* (2015) Targeted gene editing by transfection of in vitro reconstituted Streptococcus thermophilus Cas9 nuclease complex. *RNA Biol*, **12**, 1–4.

81. Colella,P., Ronzitti,G. and Mingozzi,F. (2018) Emerging Issues in AAV-Mediated In Vivo Gene Therapy. *Mol Ther Methods Clin Dev*, **8**, 87–104.

82. Peyvandi,F. and Garagiola,I. (2019) Clinical advances in gene therapy updates on clinical trials of gene therapy in haemophilia. *Haemophilia*, **25**, 738–746.

83. Choudhury,S.R., Fitzpatrick,Z., Harris,A.F., Maitland,S.A., Ferreira,J.S., Zhang,Y., Ma,S., Sharma,R.B., Gray-Edwards,H.L., Johnson,J.A., *et al.* (2016) In Vivo Selection Yields AAV-B1 Capsid for Central Nervous System and Muscle Gene Therapy. *Mol. Ther.*, **24**, 1247–1257.

84. Zincarelli,C., Soltys,S., Rengo,G. and Rabinowitz,J.E. (2008) Analysis of AAV serotypes 1-9 mediated gene expression and tropism in mice after systemic injection. *Mol. Ther.*, **16**, 1073–1080.

85. Grieger,J.C. and Samulski,R.J. (2005) Packaging capacity of adeno-associated virus serotypes: impact of larger genomes on infectivity and postentry steps. *J. Virol.*, **79**, 9933–9944.

86. Lau,C.-H. and Suh,Y. (2017) In vivo genome editing in animals using AAV-CRISPR system: applications to translational research of human disease. *F1000Res*, **6**, 2153.

87. McClements,M.E. and MacLaren,R.E. (2017) Adeno-associated Virus (AAV) Dual Vector Strategies for Gene Therapy Encoding Large Transgenes. *Yale J Biol Med*, **90**, 611–623.

88. Lee,J.K., Jeong,E., Lee,J., Jung,M., Shin,E., Kim,Y.-H., Lee,K., Jung,I., Kim,D., Kim,S., *et al.* (2018) Directed evolution of CRISPR-Cas9 to increase its specificity. *Nat Commun*, **9**, 3048.

89. Slaymaker,I.M., Gao,L., Zetsche,B., Scott,D.A., Yan,W.X. and Zhang,F. (2016) Rationally engineered Cas9 nucleases with improved specificity. *Science*, **351**, 84–88.

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

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

92. Karvelis,T., Gasiunas,G. and Siksnys,V. (2017) Harnessing the natural diversity and in vitro evolution of Cas9 to expand the genome editing toolbox. *Curr. Opin. Microbiol.*, **37**, 88–94.

93. Reisch,C.R. and Prather,K.L.J. (2015) The no-SCAR (Scarless Cas9 Assisted Recombineering) system for genome editing in Escherichia coli. *Sci Rep*, **5**, 15096.

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

95. Karvelis,T., Young,J.K. and Siksnys,V. (2019) A pipeline for characterization of novel Cas9 orthologs. *Meth. Enzymol.*, **616**, 219–240.

96. Acharya,S., Mishra,A., Paul,D., Ansari,A.H., Azhar,M., Kumar,M., Rauthan,R., Sharma,N., Aich,M., Sinha,D., *et al.* (2019) Francisella novicida Cas9 interrogates genomic DNA with very high specificity and can be used for mammalian genome editing. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 20959–20968.

97. Chatterjee,P., Jakimo,N. and Jacobson,J.M. (2018) Minimal PAM specificity of a highly similar SpCas9 ortholog. *Sci Adv*, **4**, eaau0766.

98. Hu,Z., Wang,S., Zhang,C., Gao,N., Li,M., Wang,D., Wang,D., Liu,D., Liu,H., Ong,S.-G., *et al.* (2020) A compact Cas9 ortholog from Staphylococcus Auricularis (SauriCas9) expands the DNA targeting scope. *PLoS Biol.*, **18**, e3000686.

99. Edraki,A., Mir,A., Ibraheim,R., Gainetdinov,I., Yoon,Y., Song,C.-Q., Cao,Y., Gallant,J., Xue,W., Rivera-Pérez,J.A., *et al.* (2019) A Compact, High-Accuracy Cas9 with a Dinucleotide PAM for In Vivo Genome Editing. *Mol. Cell*, **73**, 714-726.e4.

100. Hirano,S., Abudayyeh,O.O., Gootenberg,J.S., Horii,T., Ishitani,R., Hatada,I., Zhang,F., Nishimasu,H. and Nureki,O. (2019) Structural basis for the promiscuous PAM recognition by Corynebacterium diphtheriae Cas9. *Nat Commun*, **10**, 1968.

101. Harrington,L.B., Paez-Espino,D., Staahl,B.T., Chen,J.S., Ma,E., Kyrpides,N.C. and Doudna,J.A. (2017) A thermostable Cas9 with increased lifetime in human plasma. *Nat Commun*, **8**, 1424.

102. Schunder,E., Rydzewski,K., Grunow,R. and Heuner,K. (2013) First indication for a functional CRISPR/Cas system in Francisella tularensis. *Int. J. Med. Microbiol.*, **303**, 51–60.

103. Müller,M., Lee,C.M., Gasiunas,G., Davis,T.H., Cradick,T.J., Siksnys,V., Bao,G., Cathomen,T. and Mussolino,C. (2016) Streptococcus thermophilus CRISPR-Cas9 Systems Enable Specific Editing of the Human Genome. *Mol. Ther.*, **24**, 636–644.

104. Zetsche,B., Gootenberg,J.S., Abudayyeh,O.O., Slaymaker,I.M., Makarova,K.S., Essletzbichler,P., Volz,S.E., Joung,J., van der Oost,J., Regev,A., *et al.* (2015) Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell*, **163**, 759–771.

105. Koonin,E.V. and Makarova,K.S. (2019) Origins and evolution of CRISPR-Cas systems. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, **374**, 20180087.

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

107. Moreno-Mateos,M.A., Fernandez,J.P., Rouet,R., Vejnar,C.E., Lane,M.A., Mis,E., Khokha,M.K., Doudna,J.A. and Giraldez,A.J. (2017) CRISPR-Cpf1 mediates efficient homology-directed repair and temperature-controlled genome editing. *Nat Commun*, **8**, 2024.

108. Wang,M., Mao,Y., Lu,Y., Tao,X. and Zhu,J.-K. (2017) Multiplex Gene Editing in Rice Using the CRISPR-Cpf1 System. *Mol Plant*, **10**, 1011–1013.

109. Tu,M., Lin,L., Cheng,Y., He,X., Sun,H., Xie,H., Fu,J., Liu,C., Li,J., Chen,D., *et al.* (2017) A 'new lease of life': FnCpf1 possesses DNA cleavage activity for genome editing in human cells. *Nucleic Acids Res.*, **45**, 11295–11304.

110. Zetsche,B., Strecker,J., Abudayyeh,O.O., Gootenberg,J.S., Scott,D.A. and Zhang,F. (2019) A Survey of Genome Editing Activity for 16 Cas12a Orthologs. *Keio J Med*, 10.2302/kjm.2019-0009-OA.

111. Ahn,W.-C., Park,K.-H., Bak,I.S., Song,H.-N., An,Y., Lee,S.-J., Jung,M., Yoo,K.-W., Yu,D.-Y., Kim,Y.-S., *et al.* (2019) In vivo genome editing using the Cpf1 ortholog derived from Eubacterium eligens. *Sci Rep*, **9**, 13911.

112. Yamano,T., Nishimasu,H., Zetsche,B., Hirano,H., Slaymaker,I.M., Li,Y., Fedorova,I., Nakane,T., Makarova,K.S., Koonin,E.V., *et al.* (2016) Crystal Structure of Cpf1 in Complex with Guide RNA and Target DNA. *Cell*, **165**, 949–962.

113. Liu,J.-J., Orlova,N., Oakes,B.L., Ma,E., Spinner,H.B., Baney,K.L.M., Chuck,J., Tan,D., Knott,G.J., Harrington,L.B., *et al.* (2019) CasX enzymes comprise a distinct family of RNA-guided genome editors. *Nature*, **566**, 218–223.

114. Yang,H., Gao,P., Rajashankar,K.R. and Patel,D.J. (2016) PAM-Dependent Target DNA Recognition and Cleavage by C2c1 CRISPR-Cas Endonuclease. *Cell*, **167**, 1814-1828.e12.

115. Swarts,D.C., van der Oost,J. and Jinek,M. (2017) Structural Basis for Guide RNA Processing and Seed-Dependent DNA Targeting by CRISPR-Cas12a. *Molecular Cell*, **66**, 221-233.e4.

116. Chen,J.S., Ma,E., Harrington,L.B., Da Costa,M., Tian,X., Palefsky,J.M. and Doudna,J.A. (2018) CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science*, **360**, 436–439.

117. Teng,F., Cui,T., Feng,G., Guo,L., Xu,K., Gao,Q., Li,T., Li,J., Zhou,Q. and Li,W. (2018) Repurposing CRISPR-Cas12b for mammalian genome engineering. *Cell Discov*, **4**, 63.

118. Karvelis,T., Bigelyte,G., Young,J.K., Hou,Z., Zedaveinyte,R., Budre,K., Paulraj,S., Djukanovic,V., Gasior,S., Silanskas,A., *et al.* (2020) PAM recognition by miniature CRISPR-Cas12f nucleases triggers programmable double-stranded DNA target cleavage. *Nucleic Acids Res.*, **48**, 5016–5023.

119. Harrington,L.B., Burstein,D., Chen,J.S., Paez-Espino,D., Ma,E., Witte,I.P., Cofsky,J.C., Kyrpides,N.C., Banfield,J.F. and Doudna,J.A. (2018) Programmed DNA destruction by miniature CRISPR-Cas14 enzymes. *Science*, **362**, 839–842.

120. Yan,W.X., Hunnewell,P., Alfonse,L.E., Carte,J.M., Keston-Smith,E., Sothiselvam,S., Garrity,A.J., Chong,S., Makarova,K.S., Koonin,E.V., *et al.* (2019) Functionally diverse type V CRISPR-Cas systems. *Science*, **363**, 88–91.

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

122. Gootenberg,J.S., Abudayyeh,O.O., Lee,J.W., Essletzbichler,P., Dy,A.J., Joung,J., Verdine,V., Donghia,N., Daringer,N.M., Freije,C.A., *et al.* (2017) Nucleic acid detection with CRISPR-Cas13a/C2c2. *Science*, **356**, 438–442.

123. Li,L., Wei,K., Zheng,G., Liu,X., Chen,S., Jiang,W. and Lu,Y. (2018) CRISPR-Cpf1-Assisted Multiplex Genome Editing and Transcriptional Repression in Streptomyces. *Appl. Environ. Microbiol.*, **84**.

124. Kleinstiver,B.P., Prew,M.S., Tsai,S.Q., Nguyen,N.T., Topkar,V.V., Zheng,Z. and Joung,J.K. (2015) Broadening the targeting range of Staphylococcus aureus CRISPR-Cas9 by modifying PAM recognition. *Nat. Biotechnol.*, **33**, 1293–1298.

125. Kleinstiver,B.P., Prew,M.S., Tsai,S.Q., Topkar,V.V., Nguyen,N.T., Zheng,Z., Gonzales,A.P.W., Li,Z., Peterson,R.T., Yeh,J.-R.J., *et al.* (2015) Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature*, **523**, 481–485.