学位論文

# Structural analysis of Cpf1, a novel nuclease of a CRISPR-Cas system

# （CRISPR-Cas 系に関わる新規ヌクレアーゼ Cpf1 の X 線結晶構造解析）

平成 29 年 12 月博士（理学）申請

東京大学大学院理学系研究科

生物科学専攻

山野　峻

# Contents

## Chapter 3: Structural basis for the altered PAM recognition by engineered CRISPR-Cpf1 ......................65

## Chapter 4: Structural basis for the canonical and non-canonical PAM recognition by CRISPR-Cpf1 . 97

# Abstract

CRISPR (clustered regularly interspaced short palindromic repeat)-Cas (CRISPR-associated protein) systems are microbial adaptive immune systems against invading foreign genetic elements accompanying with viral infections and plasmids transfers. Cas proteins form effector ribonucleoprotein complexes with guide RNA, and the effector complexes recognize and degrade the target nucleic acids complementary to the guide sequences. CRISPR-Cas systems are divided into class 1 and class 2. Class 2 CRISPR-Cas systems use a single Cas endonuclease as an effector element, and they are further divided into type II, V, and VI. Cas9, an effector protein of class 2 type II CRISPR-Cas system has been harnessed for gene editing tools. Cpf1 (Cas12a) is an effector endonuclease of class 2 type V CRISPR-Cas systems. Cpf1s derived from *Acidaminoncoccus sp.* BV3L6 (AsCpf1) and *Lachnospiraceae bacterium* ND2006 (LbCpf1) exhibit robust DNA targeting and cleavage activities in eukaryotic cells, so they also have been harnessed for gene editing tools.

Cpf1 shows some unique features different from Cas9 in the aspects of the target DNA cleavage and the PAM recognition. Therefore, it was predicted that the target DNA recognition and cleavage mechanisms of Cpf1 are different from Cas9. There was no structural information of Cpf1, so the reasons why Cpf1 exhibits such unique features remained elusive. Here in this study, we determined the crystal structures of Cpf1–crRNA–target DNA complexes, and succeeded in revealed the working mechanisms of Cpf1.

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

Although the RuvC nuclease domain was predicted from the amino acids sequences of Cpf1, no other known domain structures were predicted. To reveal the working mechanisms of Cpf1, we determined the crystal structure of AsCpf1–crRNA–target DNA complex in 2.8 Å resolution. AsCpf1 adopts a bilobed architecture, and the RNA–DNA heteroduplex bound in the central channel. The TTTV PAM recognition mechanism of AsCpf1 was unveiled from this structure. AsCpf1 recognizes the shape of the PAM duplex and the bases of PAM nucleotides, which are referred to as shape readout and base readout respectively. Moreover, the mutation analysis and the target DNA cleavage assays revealed the target DNA cleavage mechanism. The results of assays indicated that both strands of the target DNA are cleaved by the RuvC domain,

and the Nuc domain guides the target DNA strand to the catalytic center of the RuvC domain.

## Structural basis for the altered PAM recognition by engineered CRISPR-Cpf1

Cpf1–crRNA complex requires the TTTV PAM for target recognition. When Cpf1 is used for gene editing tools, the requirement of the TTTV PAM may limit the availability of the suitable target sites. To address this limitation, two AsCpf1 variants (RVR and RR) which recognize alternative PAMs are created by a structure-guided mutation screen. The RVR and RR variants recognize the TATV and TYCV PAMs respectively. However, how the Cpf1 variants recognize the altered PAM remains unknown. We determined the crystal structures of the RVR and RR variants in complex with crRNA and target DNA in 2.0 Å resolution. The new interactions were formed mainly between the altered PAM complementary nucleotides and the substituted residues. These high-resolution structures revealed the altered PAM recognition mechanisms and the roles of the substitutions of the Cpf1 variants.

## Structural basis for the canonical and non-canonical PAM recognition by CRISPR-Cpf1

Recently PAM preferences of Cpf1s have been revealed. LbCpf1 and AsCpf1 prefer the canonical TTTV PAM. In addition, they also recognize non-canonical C-containing PAMs. To elucidate the non-canonical PAM recognition mechanisms of Cpf1, we determined the crystal structures of LbCpf1 in complex with crRNA and target DNA containing either TTTA, TCTA, TCCA, or CCCA as the PAM in 2.4–2.5 Å resolution. The canonical PAM recognition mechanism of LbCpf1 are similar to that of AsCpf1, and the PAM duplex in the TTTA PAM complex is highly distorted. In the C-containing PAM complex structures, the C-containing PAM duplexes adopt less distorted conformations compared to that of the TTTA complex. A structural comparison between the four complexes revealed altered hydrogen-bonding interactions between the PI domains and the PAM nucleotides, accompanying the displacement of their PI domains. These structures revealed the importance of the flexibility of the PI domains and the interactions between the PI domains and the PAM duplexes for the distinct PAM recognitions.

# Table of abbreviations

| Abbreviation | Full name |
|---|---|
| DNA | deoxyribonucleic acid |
| RNA | ribonucleic acid |
| ss | single-stranded |
| ds | double-stranded |
| DTT | dithiothreitol |
| EDTA | ethylenediaminetetraacetic acid |
| kDa | kilodalton |
| IPTG | isopropyl β-D-thiogalactopyranoside |
| LB | Luria-Bertani |
| Ni-NTA | nickel-nitrilotriacetic acid |
| PCR | polymerase chain reaction |
| PDB | Protein Data Bank |
| PEG | polyethylene glycol |
| SAD | single wavelength anomalous diffraction |
| SDS-PAGE | sodium dodecyl sulfate poly-acrylamide gel electrophoresis |
| SeMet | selenomethionin |
| TEV | tobacco etch virus |
| Tris | tris(hydroxymethyl)aminomethane |
| WT | wild type |

## Table of amino acid and nucleic acid base abbreviations

| Abbreviation | Full name |
|---|---|
| A, Ala | alanine |
| C, Cys | cysteine |
| D, Asp | aspartic acid |
| E, Glu | glutamic acid |
| F, Phe | phenylalanine |
| G, Gly | glycine |
| H, His | histidine |
| I, Ile | isoleucine |
| K, Lys | lysine |
| L, Leu | leucine |
| M, Met | methionine |
| N, Asn | asparagine |
| P, Pro | proline |
| Q, Gln | glutamine |
| R, Arg | arginine |
| S, Ser | serine |
| T, Thr | threonine |
| V, Val | valine |
| W, Trp | tryptophan |
| Y, Tyr | tyrosine |
| A | adenine |
| T | thymine |
| G | guanine |
| C | cytosine |
| U | uracil |

# Chapter 1: General introduction

## 1.1 Overview of CRISPR-Cas systems

Prokaryotes are constantly threatened by phage infection and invasion by mobile genetic elements through conjugation and transformation. CRISPR-Cas (clustered regularly interspaced short palindromic repeats and CRISPR associated) systems provide adaptive immunity for bacteria and archaea against foreign genetic elements (Doudna and Charpentier, 2014; Marraffini, 2015; Mohanraju et al., 2016; Wright et al., 2016) (Figure 1-1). A CRISPR locus is composed of CRISPR arrays and *cas* genes. CRISPR arrays contain variable spacer regions and conserved repeat regions, and *cas* genes are organized in operons. This system is conserved in about 50% of bacteria and almost all archaea (Sorek et al., 2013). Spacer sequences often match fragments of previously encountered foreign genetic elements, and they play a key role in memories of the invasions. Cas proteins, translated from the *cas* operons, and crRNAs (CRISPR RNAs), transcribed from the CRISPR array, play critical roles for the eliminations of the invading foreign genetic elements.

The working mechanism of CRISPR-Cas system is constituted with three steps, “adaptation”, “expression”, and “interference”. In the first “adaptation” stage, invading foreign genetic elements were processed into short fragments and integrated into CRISPR arrays as new spacer sequences by the integration activity of the Cas1–Cas2 hetero hexamer (Heler et al., 2015; Nuñez et al., 2014, 2015a, 2015b; Wang et al., 2015; Wei et al., 2015; Wright et al., 2017; Xiao et al., 2017). In the second “expression” stage, crRNA precursors are transcribed from the CRISPR array, and mature crRNAs containing one repeat region and one spacer region are processed from the precursors. The mature crRNAs form effector complexes with effector proteins translated from *cas* operons. In the third “interference” stage, when foreign genetic elements with the same sequences invade again, the effector complexes recognize and degrade them by the sequence complementarity.

For gene targeting by effector complexes, a PAM (protospacer adjacent motif), a short specific sequence, should exist at the proximal region of the target site, in addition to the sequence complementarity between the guide RNA and target DNA. A PAM plays an important role for the effector nucleases to distinguish the self-CRISPR array and non-self invading sequences (Deveau et al., 2008; Marraffini and Sontheimer, 2010;

Mojica et al., 2009; Sashital et al., 2012; Westra et al., 2013). These mechanisms are conserved features in all CRISPR-Cas systems.

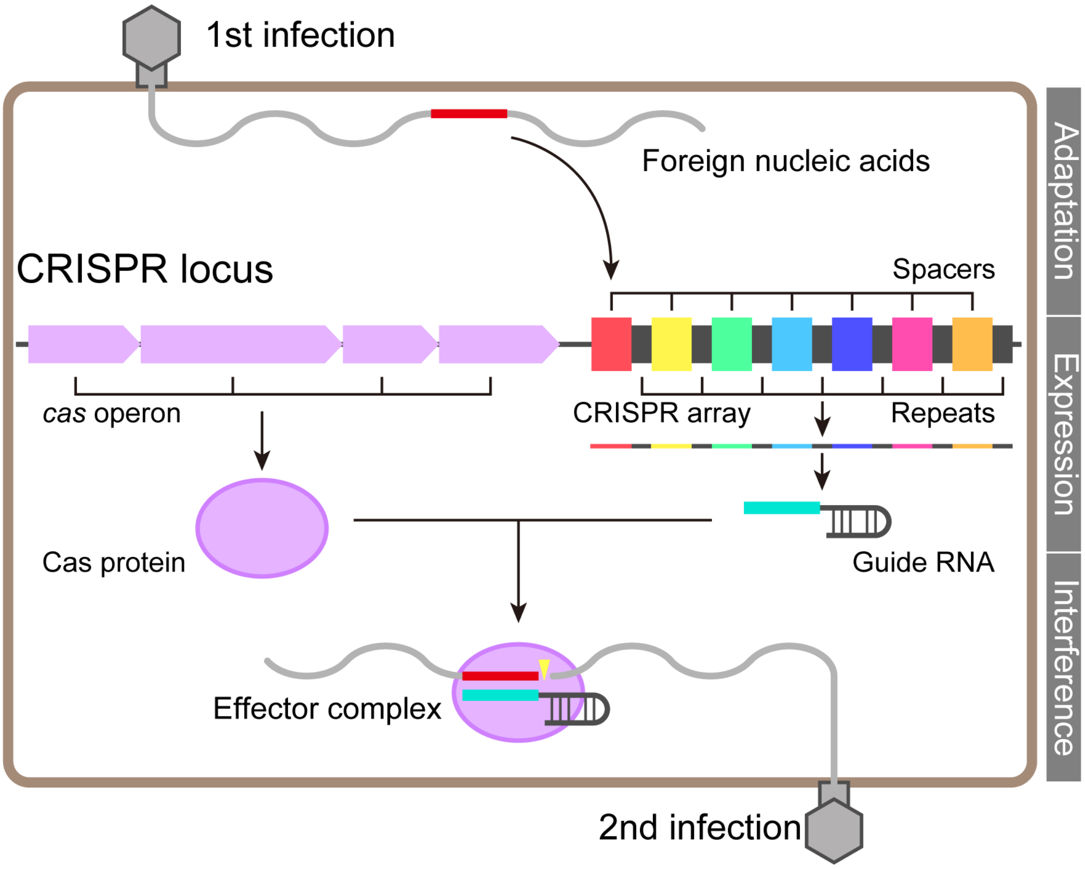


**Figure 1-1. Overview of CRISPR-Cas systems**
The model of the working mechanisms of CRISPR-Cas systems.

## 1.2 Discovery of CRISPR-Cas system

In 1987, Japanese research group discovered characteristic roughly palindromic repeated sequences in *Escherichia coli* when they sequenced the *iap* gene (Ishino et al., 1987). However, at that time, the functions of this repeat cluster were unknown. In 2002, it is discovered that CRISPR arrays exist nearby some conserved genes, and short non-coding RNAs are transcribed from this repeat cluster (Tang et al., 2002). Arrays of these repeated sequences and associating genes were named as CRISPR and *cas* genes respectively (Jansen et al., 2002). As a result of accumulation of prokaryotic genome information, bio-informatic sequence analysis revealed the sequence similarities between the spacer sequences and mobile genetic elements of phages and plasmids. This result suggested that the CRISPR array work as memories of invasion by foreign genetic elements (Mojica et al., 2005).

In 2007, this hypothesis was proofed experimentally. A research group in Danisco, a bio-based company, investigated the genome of the phage resistant *Streptococcus thermophiles*, and found that the CRISPR array of *S. thermophiles* contain the same sequences to the phage genome, thereby elucidating the role and the importance of the spacer sequences in the adaptive immune system (Barrangou et al., 2007). Moreover, in 2008, some basal key features of CRISPR-Cas systems are characterized. First, Cas proteins form ribonucleoproteins with the short non-cording RNA (crRNA) transcribed from the CRISPR array, and Cas proteins are active in the complexes. Second, most of the effector complexes target the dsDNA derived from phages and plasmids. Third, the selections of the target sites mainly depend on the guide sequences of crRNA, therefore the effector complexes are programmable DNA endonucleases (Brouns et al., 2008; Marraffini and Sontheimer, 2008).

CRISPR-Cas systems are divided into class 1 and class 2 based on the construction of the effector complexes (Makarova et al., 2015; Nishimasu and Nureki, 2017; Shmakov et al., 2017). In class 1 CRISPR-Cas systems, Cas proteins form multi-subunit complex, and it form an effector complex with a crRNA (Figure 1-2). Class 1 CRISPR-Cas systems are further divided into type I, III and IV. The effector complex of class 1 type I CRISPR-Cas system is Cascade (CRISPR-associated complex for antiviral defence), and it is studied very well biochemically and structurally (Jackson et al., 2014; Jore et al., 2011; Mulepati et al., 2014; Wiedenheft et al., 2011; Zhao et al., 2014) (Figures 1-2A and 1-2B). The effector complexes of class 1 type III CRISPR-Cas system are named as Cmr and Csm complex (Figures 1-2C and 1-2D). These complexes can target both DNA and RNA (Jiang et al., 2016b; Osawa et al., 2015; Samai et al., 2015; Taylor et al., 2015). In contrast, class 2 CRISPR-Cas systems, a single endonuclease form a ribonucleoprotein complex with crRNA and works as an effector complex. Cas9 is the most well characterized effector protein in class 2 type II CRISPR-Cas system (Figure 1-3). In recent years, novel CRISPR-Cas systems are discovered one after another. In class 2 type V CRISPR-Cas system, Cpf1 (Cas12a) and C2c1 (Cas12b) are identified as effector proteins (Shmakov et al., 2015; Zetsche et al., 2015a). Whereas they target dsDNA, they exhibit some differences compared to Cas9. In class 2 type VI CRISPR-Cas system, C2c2 (Cas13a) is the effector protein, and this is the single protein which target ssRNA (Abudayyeh et al., 2016; East-Seletsky et al., 2016; Knott et al., 2017; Liu et al., 2017a, 2017b).

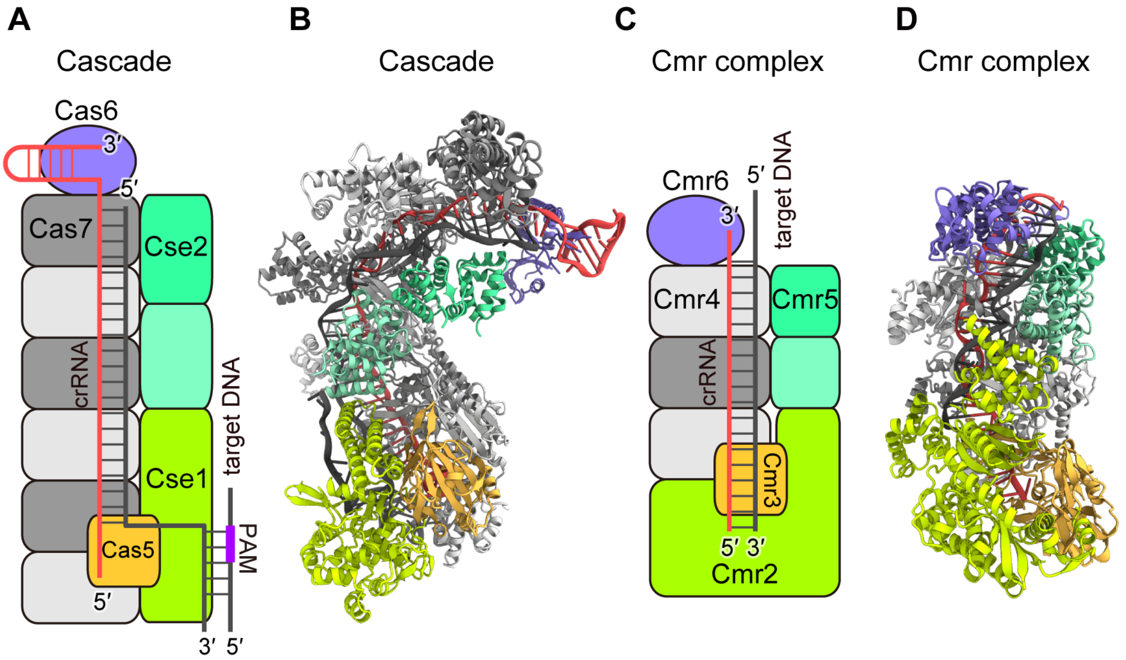


**Figure 1-2. Structures of effector complexes in class1 CRISPR-Cas systems**
(A) Model of the crystal structure of Cascade.
(B) Crystal structure of Cascade bound to a dsDNA target (PDB: 5H9F).
(C) Model of the crystal structure of Cmr complex.
(D) Crystal structure of chmeric Cmr complex bound to a crRNA and non-cleaved ssDNA target (PDB: 3X1L).

## 1.3 Cas9

Cas9 is an effector endonuclease classified into the class 2 type II CRISPR-Cas system (Figures 1-3A and 1-3B). Cas9 has two nuclease domains, the RuvC and HNH domain (Barrangou et al., 2007), targets dsDNA, and creates a blunt end in the PAM proximal region (Garneau et al., 2010). In CRISPR-Cas systems, many effector proteins are single RNA-guided DNA endonucleases, cooperating with crRNAs. However, in 2011, the additional guide RNA of Cas9 was discovered in the CRISPR locus of *Streptococcus pyogenes*. The second guide RNA was necessary for the nuclease activity of Cas9, and it was named as tracrRNA (trans-activating crRNA) (Deltcheva et al., 2011) (Figure 1-3C). This discovery revealed that Cas9 is a dual RNA-guided DNA endonuclease, cooperating with crRNA and tracrRNA. SpCas9 (*Streptococcus pyogenes* Cas9) was characterized in 2012. The HNH and RuvC domains cleave the target DNA strand and non-target DNA strand, respectively. SpCas9 recognizes NGG

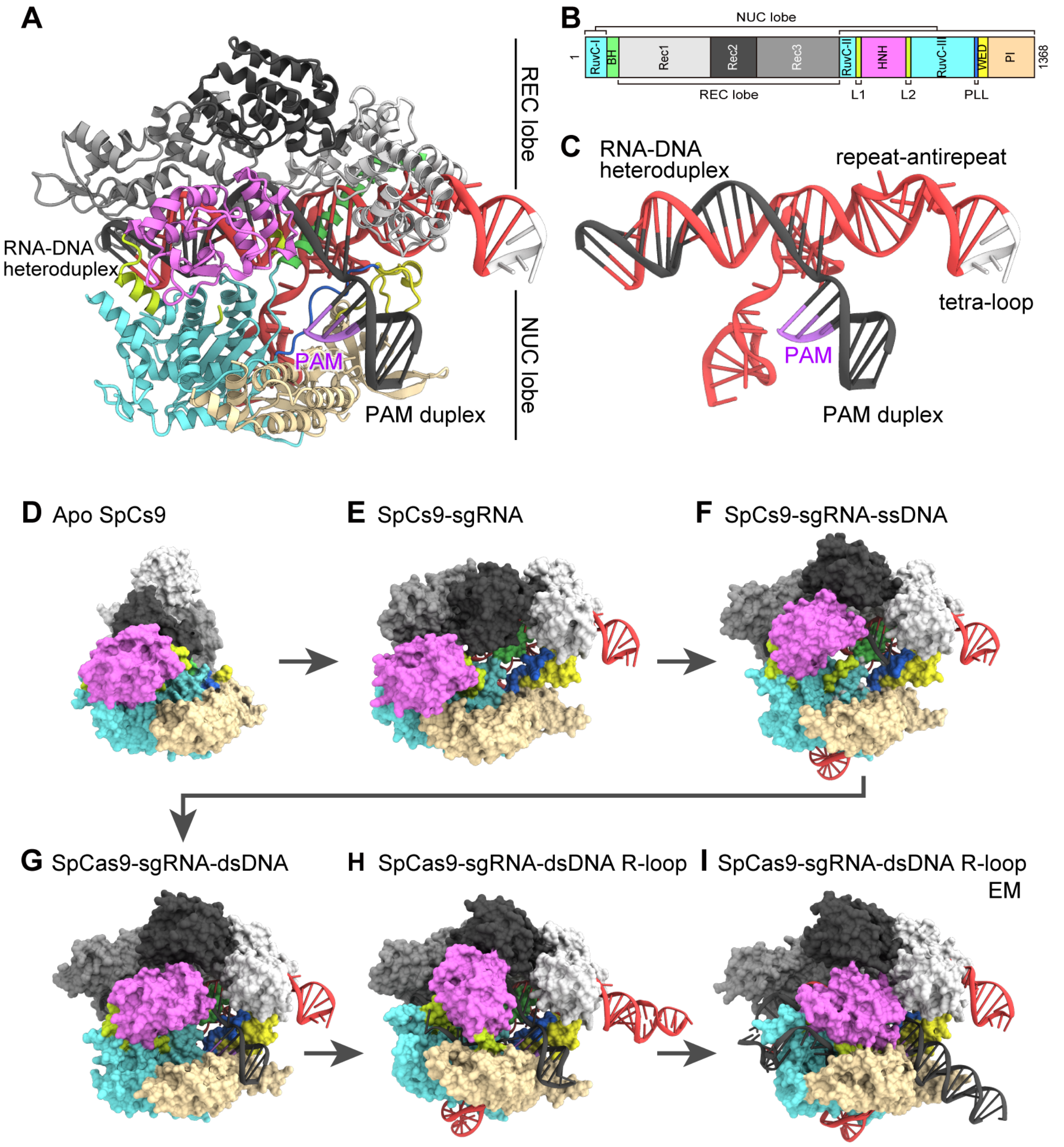


**Figure 1-3. Structures of SpCas9**

(A) Crystal structure of SpCas9 in complex with sgRNA and target DNA (4UN3).

(B) The domain organization of SpCas9.

(C) The structure of sgRNA and target DNA. The tetra-loop of sgRNA is colored white.

(D–I) Crystal structure of SpCas9 in different states. Apo-SpCas9 (4CMP) (D), SpCas9-sgRNA binary complex (4ZT0) (E), SpCas9-sgRNA-ssDNA complex (4OO8) (F), SpCas9-sgRNA-dsDNA complex (4UN3) (G), SpCas9-sgRNA-dsDNA R-loop complex (5F9R) (H), cryo-EM structure of SpCas9-sgRNA-taget DNA complex (5Y36) (I). The conformational changes accompanying with sgRNA and target DNA recognitions and cleavages are indicated by gray arrows.

PAM (N is A, T, G or C). Cas9 is active with sgRNA (single-guide RNA) which is created by connecting crRNA and tracrRNA with an artificial linker loop (Jinek et al., 2012). This character highlights the potential to exploit the system for programmable gene editing.

To elucidate the unique working mechanisms of Cas9, structural information of Cas9 is indispensable. The structures of SpCas9 in six different states have been reported (Apo-SpCas9, SpCas9–sgRNA binary complex, SpCas9–sgRNA–ssDNA ternary complex, and SpCas9–sgRNA–dsDNA complex) (Anders et al., 2014; Huai et al., 2017; Jiang et al., 2015, 2016a; Jinek et al., 2014; Nishimasu et al., 2014) (Figure 1-3D–1-3I). Cas9 is a bilobed structure constructed with the REC lobe and the NUC lobe. The REC lobe is constituted by the bridge helix (BH), the Rec1, Rec2 and Rec3 domains, and the NUC lobe is constituted by the WED, PI, RuvC and HNH domains, respectively. The guide region of sgRNA and the target DNA form the RNA–DNA hetero duplex, and it is accommodated in the central channel between the two lobes. Based on these structures, the sgRNA, target DNA, and NGG PAM recognition mechanisms are revealed. Furthermore, the structural comparisons suggest the conformational changes of Cas9 accompanying with binding the sgRNA and the target DNA. The conformational changes of Cas9 were elucidated by FRET and AFM (Shibata et al., 2017; Sternberg et al., 2015).

According to specific combinations of *cas* genes, class 2 type II systems are divided into three subtypes, type II-A, II-B, and II-C. The crystal structures of SpCas9 and SaCas9 (*Staphylococcus aureus* Cas9) in type II-A, FnCas9 (*Francisella novicida* Cas9) in type II-B, and CjCas9 (*Campylobacter jejuni* Cas9) in type II-C are determined (Hirano et al., 2016a; Nishimasu et al., 2015; Yamada et al., 2017). These structures were determined in complex with sgRNA and target DNA containing PAMs. The structural comparisons revealed the structural diversities and functional similarities of Cas9. Whereas there are no similarities in the amino acid sequences and structures of sgRNA, they are commonly bilobed structures and the hetero duplexes are accommodated in the central channels. In their working mechanisms, the requirement of PAM for dsDNA targeting and the target DNA cleavage mechanisms are conserved among the Cas9 family members.

## 1.4 Application of Cas9 for gene editing tools

The effector complexes of CRISPR-Cas systems are programmable endonucleases, therefore they were expected to being applied for gene targeting tools (Marrafini et al., 2008). Notably gene editing by Cas9 was demonstrated *in vitro* in 2012 and *in vivo* in 2013 for the first time (Cong et al., 2013; Gasiunas et al., 2012; Jinek et al., 2012, 2013; Hwang et al., 2013; Jiang et al., 2013; Mali et al., 2013). The target sites of Cas9 are changed easily by changing the guide sequence of sgRNA. When the cleavage site is repaired, insertions and deletions are induced mistakenly. This is the core mechanism of gene editing, and it is applied for the knock down of the target gene and the knock in of specific sequences. Cas9 is a more advanced, inexpensive and easier genome editing tool compared to ZFNs (zinc finger nucleases) (Urnov et al., 2010) and TALENs (transcription activator-like effector nucleases) (Joung and Sander, 2013), making the gene editing technology more familiar for all scientists. This is the revolution in the field of life sciences and medical sciences.

The gene editing technologies relating to Cas9 have improved rapidly (Barrangou and Doudna, 2016; Komor et al., 2016a). Whereas the guide sequences of sgRNAs are changed flexibly, the PAM should exist next to the target sites. It limits the selectivity of the target sites. To address this problem, Cas9 variants with altered PAM specificities are created. Originally WT SpCas9 recognizes NGG PAM and WT SaCas9 recognizes NNGRRT PAM. The engineered VQR, EQR and VRER SpCas9 variants recognize the NGA, NGAG and NGCG PAMs, respectively (Anders et al., 2016; Hirano et al., 2016b; Kleinstiver et al., 2015a). The engineered KKH SaCas9 variant recognizes the NNNRRT PAM (Kleinstiver et al., 2015b). Furthermore, the FnCas9 variant was created by structure-guided engineering, and the PAM preference was changed from NGR to YG (Hirano et al., 2016a). These Cas9 variants have broadened the range of the target sequences selectivity.

In addition, the targeting specificity of Cas9 is one of key factors for the application of Cas9 for gene editing tools. Cas9 sometimes targets the off-target sites which resemble the on-target sequences (Fu et al., 2013; Hsu et al., 2013). This is the one of the problems which impair the expansion of Cas9 for therapeutic usage. The high-fidelity Cas9 with almost no detectable genome-wide off-target effects was created by structure-guided engineering (Kleinstiver et al., 2016a; Slaymaker et al., 2015). To control the timing of activation, split SpCas9s are created. The sprit SpCas9s are assembled by sgRNA, fused FKBP (FK506 binding protein 12) and FRB (FKBP rapamycin binding) domains, or fused photoinducible dimerization domains, and they maintained the activity comparable to WT SpCas9 (Nihongaki et al., 2015; Wright et al., 2015; Zetsche et al., 2015b). Furthermore, Cas9 fused with GFP or other fluorescence protein enables the imaging of target DNAs (Chen et al., 2013). Cas9 fused with transcription activation factors (Konermann et al., 2014; Nishimasu et al., 2015) and fused with deaminases (Nishida et al., 2016; Komor et al., 2016) are applied for the programmable gene editing without double stranded breaks. These Cas9 relating tools are indispensable in the field of life science.

## 1.5 Cpf1

In 2015, Cpf1, a novel effector complex, was discovered, which is classified into the class 2 type V CRISPR-Cas system (Zetsche et al., 2015a). Whereas both of Cpf1 and Cas9 are effector endonucleases belonging to the class 2 CRISPR-Cas systems, there are some differences between them (Figure 1-4). First, Cpf1 is a single-RNA guided DNA endonuclease, guided by a crRNA, whereas Cas9 is a dual-RNA guided DNA endonuclease, guided by a pair of crRNA and tracrRNA (Deltcheva et al., 2011). Second, Cpf1 recognizes a T-rich PAM, whereas Cas9 recognizes the G-rich PAM (Fonfara et al., 2014; Karvelis et al., 2015). Third, Cpf1 generates staggered ends in its PAM-distal target site (Zetsche et al., 2015a), 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., 2015a), 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). Fifth, Cpf1 can process its own crRNA array to generate the mature crRNAs (Fonfara et al., 2016). Sixth, Cpf1 exhibits higher targeting specificity in mammalian cells, as compared with Cas9 (Kim et al., 2016a; Kleinstiver et al., 2016b). There was no sequence similarity between Cpf1 and Cas9 except the RuvC domains, so the prediction of the overall structure of Cpf1 was impossible. Similar to Cs9, Cpf1 shows gene targeting activities in eukaryotic cells, so is has been harnessed for novel gene editing tools.

To reveal the working mechanisms of Cpf1, the structural information of Cpf1 is indispensable. Moreover, it would be possible to generate novel gene editing tools by the structural-guided protein engineering.

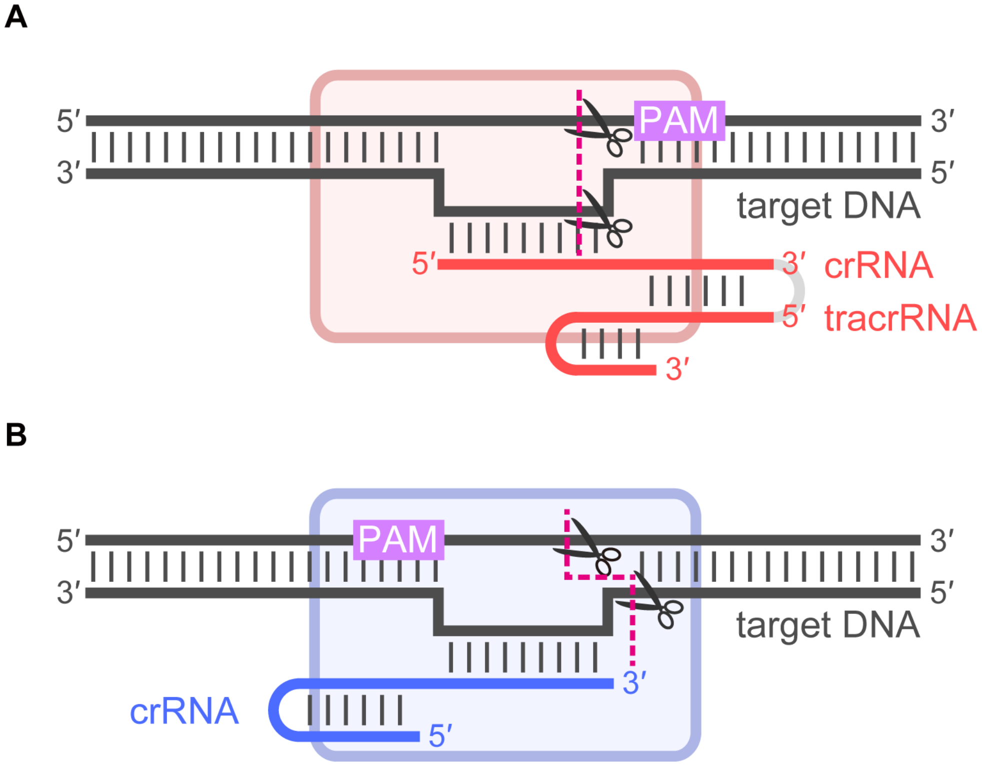


**Figure 1-4. Comparison between cas9 and Cpf1**

(A) The taget interference mechanism of Cas9.

(B) The taget interference mechanism of Cpf1.

# Chapter 2: Crystal structure of Cpf1 in complex with guide RNA and target DNA

## 2.1 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 novel nuclease supporting domain. The RuvC domain are responsible for the cleavage of both the non-target and target strands and the nuclease supporting domain assists the cleavage of the target strand, respectively, and jointly generate staggered DNA double-strand breaks. AsCpf1 recognizes the TTTV 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.

## 2.2 Introduction

A second class 2 (type V) effector protein, Cpf1, has been harnessed for genome editing (Zetsche et al., 2015a). 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., 2015a), 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., 2015a), whereas Cas9 uses the HNH and RuvC endonuclease domains to cleave the target and non-target DNA strands, respectively (Gasiunas et al., 2012; Jinek et al., 2012). Especially the third character of Cpf1 is expected to be convenient for non-homologous end joining (NHEJ)-based gene insertion. Generating the staggered cut with an overhang could enable to identify the proper orientations of the DNA inserts. Together, these observations imply major differences in the target DNA recognition and cleavage mechanisms between Cas9 and Cpf1.

## 2.3 Research aims

To clarify how Cpf1 recognizes and cleaves DNA targets, we determined the crystal structure of AsCpf1 in complex with the crRNA and its double-stranded DNA target containing the TTTV PAM (V is A, G or C). Moreover, to reveal the structural similarity and differences between the two class 2 effector proteins, we compared the structure of AsCpf1 and Cas9. The structural comparison explained the distinct functionalities and suggested the functional convergence.

## 2.4 Materials and methods

### 2.4.1 Plasmid construction

The gene encoding full-length *Acidaminococcus sp.* Cpf1 (AsCpf1, residues 1–1307) was PCR-amplified using the pcDNA3.1-AsCpf1 plasmid (Zetsche et al., 2015a) as the template, and cloned between the *Nde*I and *Xho*I sites of the modified pE-SUMO vector (LifeSensors). The original vector used as the template of PCR was received from Dr. Feng Zhang of MIT. The modified pE-SUMO vector contains a N-terminal His$_6$-tag followed by a SUMO tag and a tobacco etch virus (TEV) protease cleavage site. The plasmid DNAs were amplified in *Escherichia coli* Mach (Thermo Fisher Scientific), cultured in LB medium (Nacalai Tesque) at 37°C overnight.

### 2.4.2 Expression and purification

The AsCpf1-expressing *E. coli* Rosetta2 (DE3) (Novagen) cells were cultured at 37°C in LB medium (containing 20 mg/l kanamycin) until the OD$_{600}$ reached 0.8, and protein expression was then induced by the addition of 0.1 mM isopropyl β-D-thiogalactopyranoside (IPTG) (Nacalai Tesque). The *E. coli* cells were further cultured at 20°C for 18 h, and harvested by centrifugation at 5,000 g for 10 min. The *E. coli* cells were resuspended in buffer A (50 mM Tris-HCl, pH 8.0, 20 mM imidazole, 300 mM NaCl and 3 mM 2-mercaptoethanol), lysed by sonication, and then centrifuged at 40,000 g for 30 min. The supernatant was mixed with 5 ml Ni-NTA Superflow (QIAGEN) equilibrated with buffer A for about 1 hour at 4°C, and the mixture was loaded into an Econo-Column (Bio-Rad). The resin was washed with 5 column volumes of buffer A, 5 column volumes of buffer B (50 mM Tris-HCl, pH 8.0, 20 mM imidazole, 1 M NaCl and 3 mM 2-mercaptoethanol) and 3 column volumes of buffer A. The protein was eluted with buffer C (50 mM Tris-HCl, pH 8.0, 300 mM imidazole, 300 mM NaCl and 3 mM 2-mercaptoethanol). The protein was loaded onto a HiTrap SP HP column (GE Healthcare) equilibrated with buffer D (20 mM Tris-HCl, pH 8.0 and 200 mM NaCl). The column was washed with buffer D, and the protein was then eluted with a linear gradient of 200–1,000 mM NaCl. To remove the His$_6$-SUMO-tag, the eluted protein was mixed with TEV protease (home made), and was dialyzed at 4°C for 12 h against buffer E (20 mM Tris-HCl, pH 8.0, 40 mM imidazole, 300 mM NaCl and 3 mM 2-mercaptoethanol). The protein was passed through the Ni-NTA column equilibrated with buffer E. The protein was concentrated using an Amicon Ultra 10K filter (Millipore), and was further purified by chromatography on a HiLoad Superdex 200 16/60 column (GE Healthcare) equilibrated with buffer F (10 mM Tris-HCl, pH 8.0, 150 mM NaCl and 1 mM DTT). The purified AsCpf1 proteins were stored at −80°C until use. 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 concentrated using an Amicon Ultra 10K filter. The complex was purified by gel filtration chromatography on a Superdex 200 Increase column (GE Healthcare) equilibrated with buffer F.

### 2.4.3 crRNA and target DNA construction

Two crRNAs with different length of the guide sequences (20-nt or 24-nt) were designed based on the previous study (Zetsche et al., 2015a). The crRNAs were purchased from Gene Design. PAM containing target DNAs were designed to form the PAM duplex. We predicted that the length of PAM duplex affects the crystallization of the complex, so six target DNAs with different length of the target sequences (20-nt or 24-nt) and the PAM duplex (8-bp, 9-bp, or 10-bp) were prepared. The target and non-target DNA strands were purchased from Sigma-Aldrich.

### 2.4.4 Crystallization

The peak fractions of the purified AsCpf1–crRNA–target DNA complex were concentrated by Amicon Ultra 10K filter. Using the concentrated complex solution ($A_{260\text{ nm}}$ = 10), the initial crystallization screening was performed at 20°C by the sitting-drop vapor diffusion method.

The screening kits used for the initial screening

Crystal Screen, PEG/Ion (Hampton Research)
JBScreen Classic 1, 2, 4, 5 (Jena Bioscience)
MemGold, MemGold2, JCSG-*plus*, PACT *premier* (Molecular Dimensions)
Wizard Classic 1 and 2 (Emerald Biosystems)

For crystallization optimization, Additive Screen (Hampton Research) was added to the reservoir solutions, in addition to the optimization of the concentration of precipitations and salt, and pH of the buffers. After the optimization by the sitting-drop vapor diffusion method, further optimization was performed by the hanging-drop vapor diffusion method. The crystallization drops were formed by mixing 1 μl of complex solution ($A_{260 \text{ nm}} = 10$) and 1 μl of reservoir solution, and then were incubated against 0.5 ml of reservoir solution. The crystals were improved by micro-seeding using Seed Bead (Hampton Research). In the sitting-drop vapor diffusion method, the crystallization conditions were prepared using PLATEMASTER P220 (GILSON) and Mosquito crystallization robot (TTP Labtech).

### 2.4.5 X-ray diffraction analysis and data processing

The native crystals were cryoprotected in a solution consisting of 11% PEG3,350, 100 mM sodium acetate (pH 4.5), 15% 1,6-hexanediol and 30% ethylene glycol. To reduce radiation damage, all X-ray-diffraction data were collected at 100 K on the beamlines BL32XU and BL41XU at SPring-8, and PXI X06SA at the Swiss Light Source. A diffraction data set was collected from a single crystal using a X-ray beam at a wavelength of 1.000 Å, an oscillation range of 180° (0.1° per image), an exposure time of 1.0 s per image. The X-ray-diffraction data were processed using DIALS (Waterman et al., 2013) and AIMLESS (Evans and Murshudov, 2013).

### 2.4.6 Crystallization of SeMet-labeled AsCpf1

The selenomethionine (SeMet)-labeled AsCpf1 protein was expressed in *E. coli* B834 (DE3) (Novagen), and purified using a similar protocol as that for the native protein. The SeMet-labeled complex was reconstituted and crystallized by mixing 1 μl of complex solution ($A_{260 \text{ nm}} = 10$) and 1 μl of reservoir solution under similar conditions to the native protein. The SeMet-labeled crystals were cryoprotected in a solution consisting of 35% PEG400, 100 mM sodium acetate (pH 4.0), 200 mM lithium sulfate and 150 mM NaCl. All X-ray-diffraction data were collected at 100 K on the beamlines BL41XU at SPring-8, and PXI X06SA at the Swiss Light Source. A diffraction data set was collected from a single crystal using a X-ray beam at a wavelength of 0.979 Å, an oscillation range of 180° (0.1° per image), an exposure time of 1.0 s per image.

### 2.4.7 Phase determination, model building and structure refinement

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). Structural figures were prepared using CueMol (http://www.cuemol.org).

### 2.4.8 Generation of the AsCpf1 mutants

This experiment was performed partly in collaboration with Dr. Feng Zhang (MIT). The human codon-optimized AsCpf1 mutants were cloned using the Golden Gate strategy (Engler et al., 2009). Briefly, WT 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.

### 2.4.9 Cleavage activity of AsCpf1 in HEK293FT cells

This experiment was performed partly in collaboration with Dr. Feng Zhang (MIT). 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 (Epicentre). Indels were analyzed by deep sequencing, as previously described (Hsu et al., 2013).

### 2.4.10 Synthesis of crRNA

This experiment was performed partly in collaboration with Dr. Feng Zhang (MIT). 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 hours and then the RNA was purified using Agencourt RNAClean XP beads (Beckman Coulter).

### 2.4.11 Preparation of AsCpf1-containing cell lysate

This experiment was performed partly in collaboration with Dr. Feng Zhang (MIT). HEK293 cells, growing in 6-well plates, were transfected with AsCpf1 expression plasmids (2 μg) using the Lipofectamine 2000 reagent. After 48 hours, the cells were harvested by washing with DPBS (Life Technologies) and then were resuspended in 250 ml of lysis buffer (20 mM HEPES (pH 7.5), 100 mM KCl, 5mM $MgCl_2$, 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.

### 2.4.12 *In vitro* cleavage assay

This experiment was performed partly in collaboration with Dr. Feng Zhang (MIT). 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 (1x CutSmart® buffer (NEB), 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 TTTA 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).

## 2.5 Results

### 2.5.1 Expression and purification

WT AsCpf1 protein was stable and the expression level was high. Furthermore, WT AsCpf1 was highly purified by the combination of the Ni-NTA chromatography, the cation exchange chromatography, and the gel filtration chromatography. The gel filtration chromatogram peak showed monodispersity (Figures 2-1). The final yield of the WT AsCpf1 was 17 mg per 1 L culture medium.

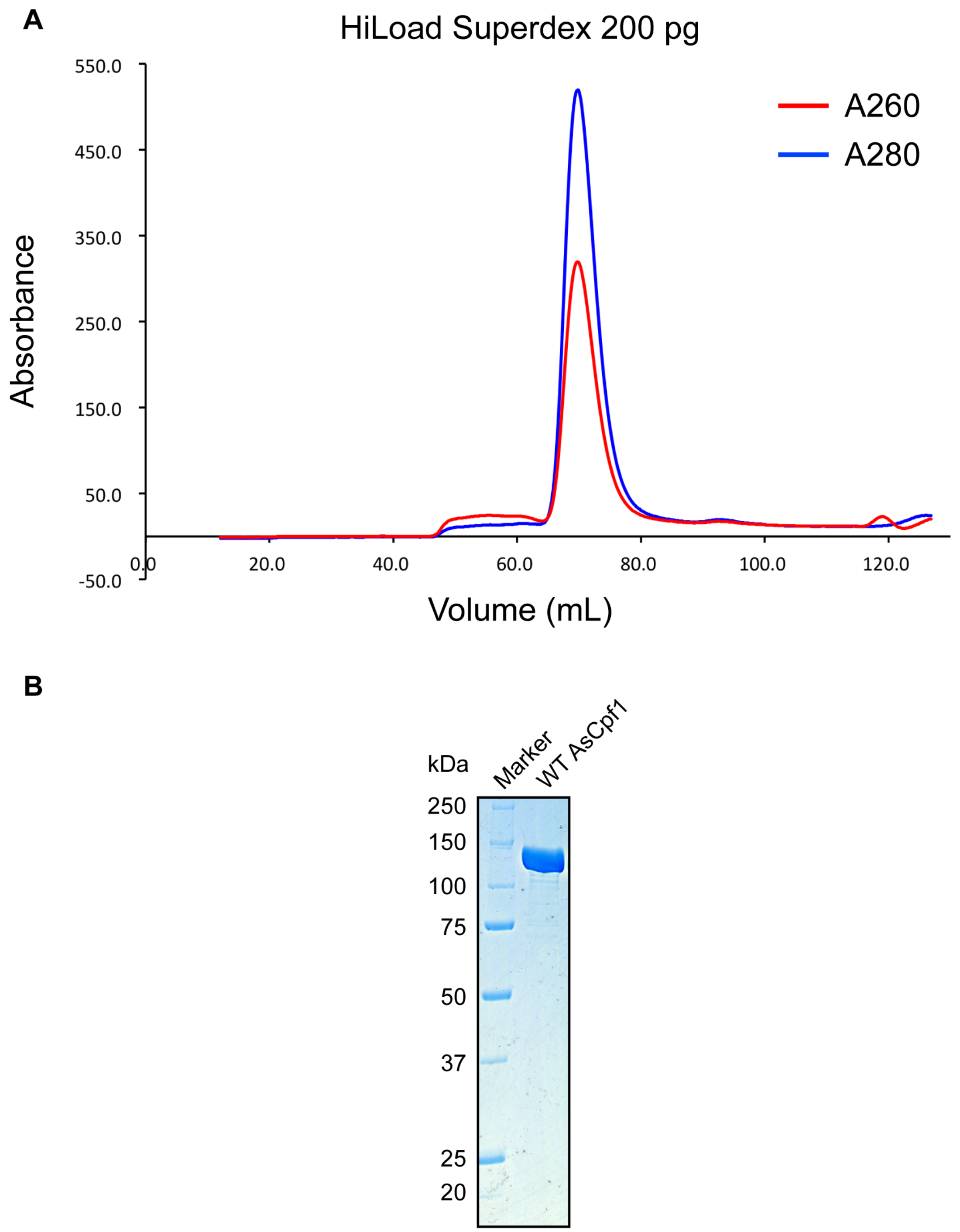


**Figure 2-1. Size-exclusion chromatogram and SDS-PAGE**

(A) Chromatogram of full-length AsCpf1.

(B) SDS-PAGE analysis with SimplyBlue SafeStain. Left lane, molecular-weight marker (labeled in kDa); right lane, WT AsCpf1.

### 2.5.2 Crystallization and X-ray diffraction analysis

The WT AsCpf1 in complex with a 43-nt crRNA containing 24-nt guide sequence, a 34-nt target DNA strand, and a 10-nt non-target DNA strand was suitable for crystallization (Figure 2-2). In the initial screening, the initial crystals of the complex were obtained under the conditions of No. 27 of JB Screen Classic 1, 2, 4, 5 (8% PEG4,000, 100 mM sodium acetate pH 4.6) and No. 88 of MemGold (28% PEG400, 200 mM lithium sulfate, 100 mM sodium citrate pH 3.5) by sitting-drop vapor diffusion method. As a result of the optimization of the condition, the crystals suitable for the X-ray diffraction analysis were obtained under the two conditions, conditions A and B (A : 8–10% PEG3,350, 100 mM sodium acetate (pH 4.5), 10–15% 1,6-hexanediol, B : 27–30% PEG400, 100 mM sodium acetate (pH 4.0), 200 mM lithium sulfate) (Figure 2-3). The 2.8 Å resolution data sets were collected from both crystals.

### 2.5.3 Expression and purification of SeMet-labeled AsCpf1

The WT AsCpf1 contains 18 methionines, so the single-wavelength anomalous diffraction (SAD) method using the SeMet-labeled AsCpf1 is suitable for the phase determination. The final yield of the SeMet-labeled AsCpf1 was 2.8 mg per 1 L culture medium, which is decreased enormously compared to the native AsCpf1.

### 2.5.4 Crystallization of SeMet-labeled AsCpf1

The crystals of the SeMet-labeled complex were obtained under the condition B (27–30% PEG400, 100 mM sodium acetate (pH 4.0), and 200 mM lithium sulfate) (Figure 2-4).

### 2.5.5 Phase determination, model building and structure refinement

Selenium anomalous dispersions were observed from the crystals of the SeMet-labeled complex, so the phase determination was performed by Se-SAD method. When the two electron density maps from two crystals are compared each other, the map calculated from the data sets of crystal B (obtained under the condition B) contains more disorder regions. Therefore, the final structure was refined using the data sets collected from crystal A (obtained under the condition A) (2.8 Å resolution, $R_{work}/R_{free}$ = 0.216/0.255) (Figures 2-5). Data collection and refinement statistics are shown in

Table 2-1.

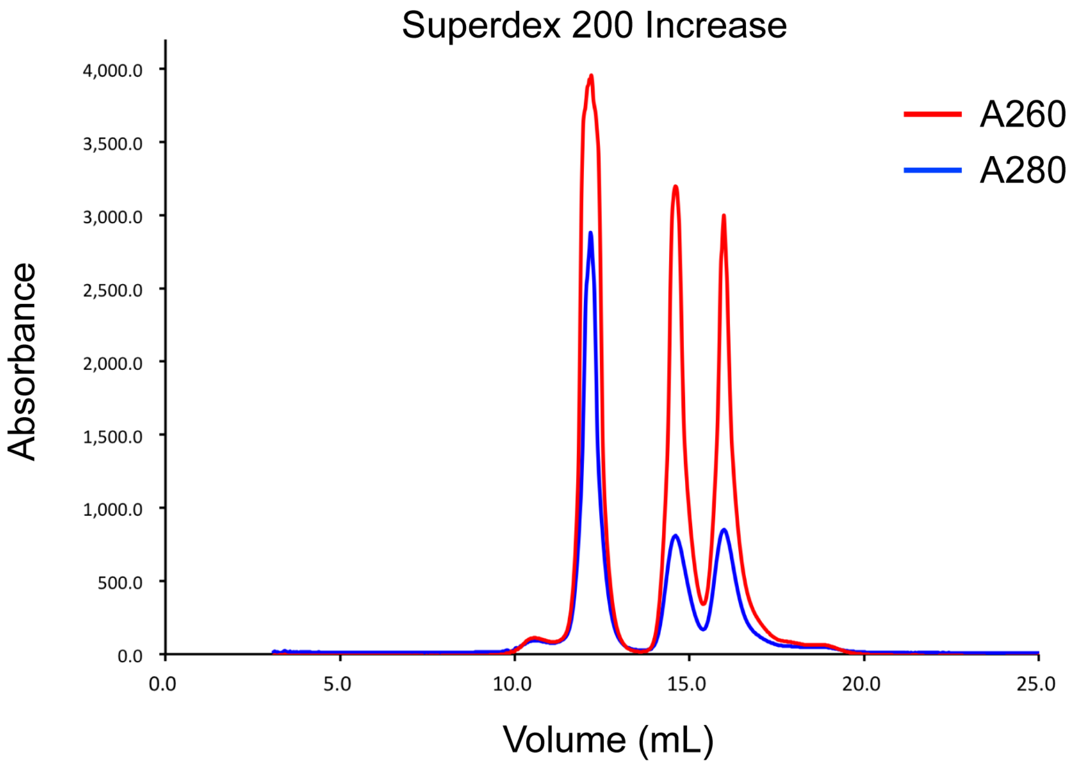


**Figure 2-2. Size-exclusion chromatogram**

Chromatogram of AsCpf1-crRNA-target DNA complex.

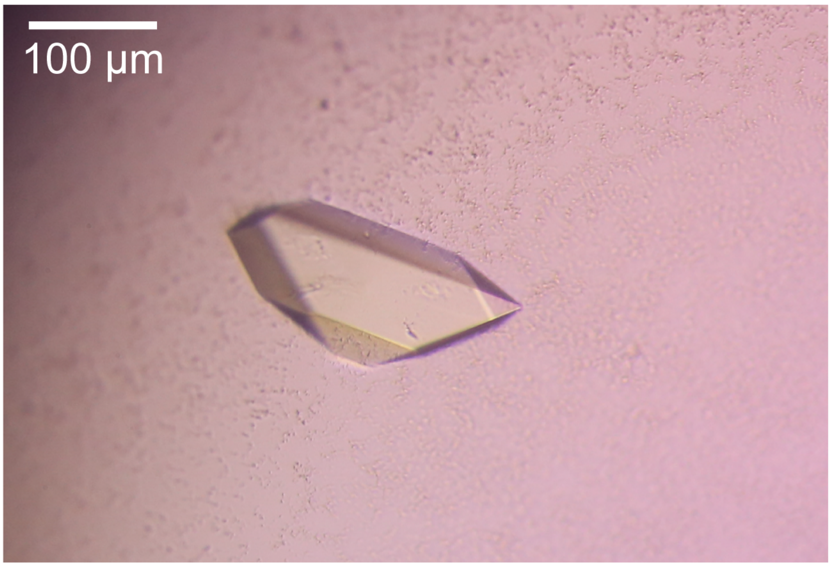


**Figure 2-3. Native crystal of AsCpf1-crRNA-target DNA complex**
Native crystal of AsCpf1-crRNA-taget DNA complex.

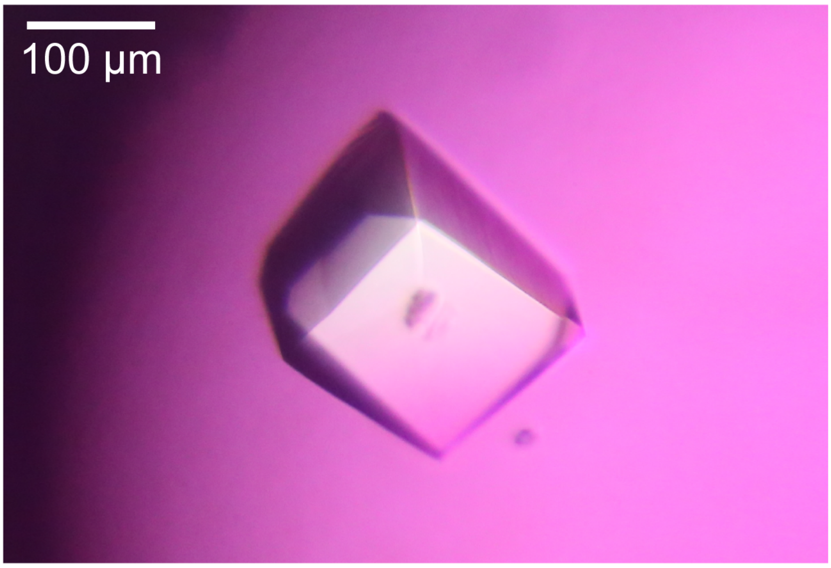


**Figure 2-4. SeMet crystal of AsCpf1-crRNA-target DNA complex**
SeMet labeled crystal of AsCpf1-crRNA-taget DNA complex.

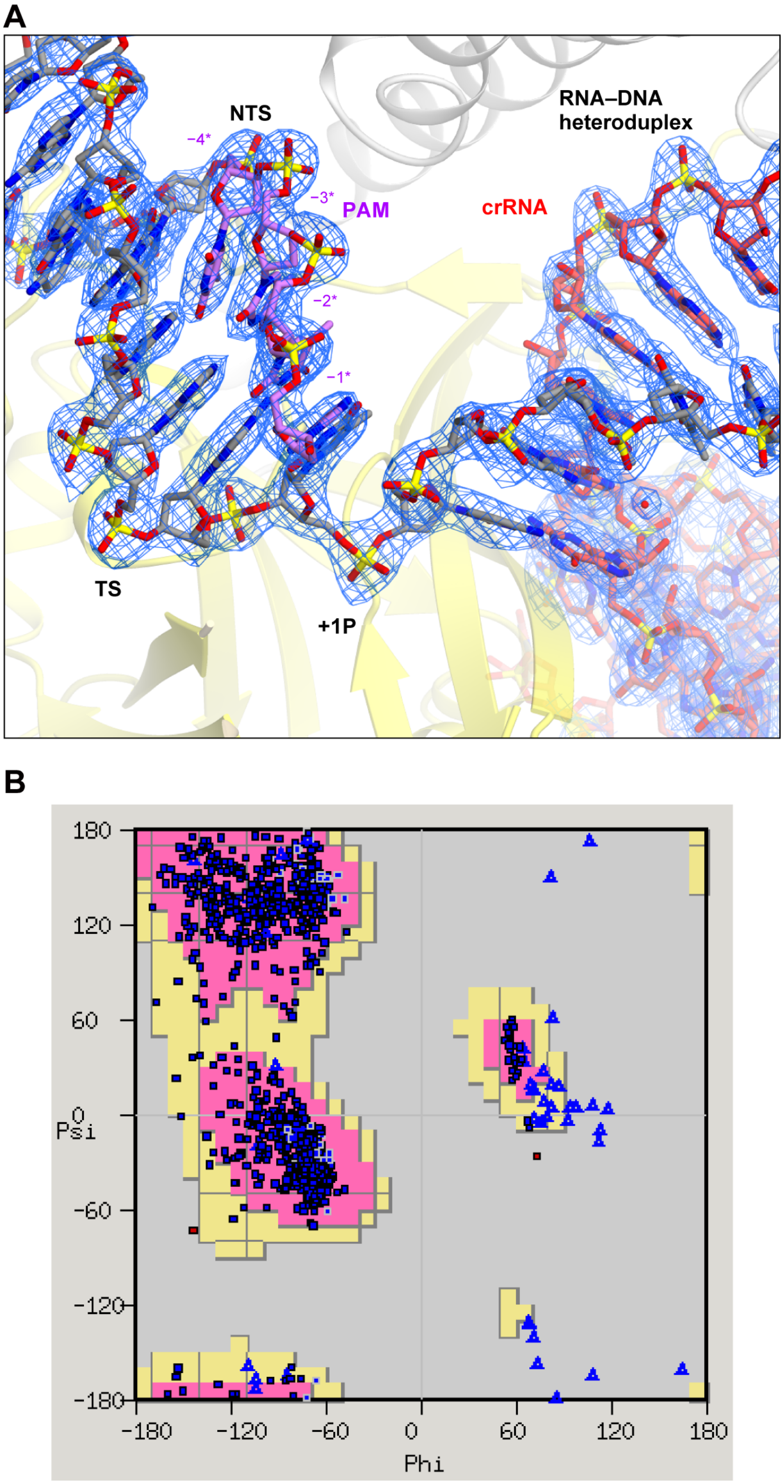


**Figure 2-5. Electron density map**

(A) The $2mF_{\mathrm{o}} - DF_{\mathrm{c}}$ electron density map (contoured at 1.5 σ) for the bound nucleic acids is shown as a blue mesh. +1P, +1 phosphate.

(B) Ramachandran plots of native AsCpf1.

**Table 2-1. Data Collection and Refinement Statistics.**

| | 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) |
| $R_{\mathrm{merge}}$ | 0.089 (0.32) | 0.155 (2.08) |
| $R_{\mathrm{pim}}$ | 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 | |
| $R_{\mathrm{work}}$ / $R_{\mathrm{free}}$ | 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.

### 2.5.6 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 TTTA PAM, by the single-wavelength anomalous diffraction (SAD) method (Figures 2-6, 2-7 and 2-8). 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 2-6C, 2-6D and 2-7). 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 2-6C).

A Dali search (Holm and Rosenström, 2010) detected no structural similarity between the REC1, REC2, as well as 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 domains of Cpf1 have no detectable homologs outside the Cpf1 protein family, and appear to adopt novel structural folds (Figures 2-6C and 2-8). The REC1 domain comprises 13 α helices, while the REC2 domain comprises 10 α helices and 2 β strands that form a small antiparallel sheet (Figures 2-8A and 2-8B). 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; Hirano et al., 2016a; Nishimasu et al., 2015), 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 2-6A, 2-8A and 2-8C). The WED domain can be divided into a core subdomain comprising a 9-stranded, distorted antiparallel β sheet (β1–β8 and β11) flanked by 7 α helices (α1–α6 and α9), and a subdomain comprising 2 β strands (β9–β10) and 2 α helices (α7 and α8) (Figures 2-8A and 2-8C). Examination of the Cpf1 sequence alignment revealed that helices α7 and α8 are not conserved among Cpf1 homologs (Zetsche et al., 2015a) (Figure 2-9). The PI domain comprises

7α 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 2-6A and 2-8A and 2-8B). As discussed previously (Zetsche et al., 2015a), 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 2-6A, 2-6C and 2-6D). The Nuc domain is inserted between the RuvC-II and RuvC-III motifs.

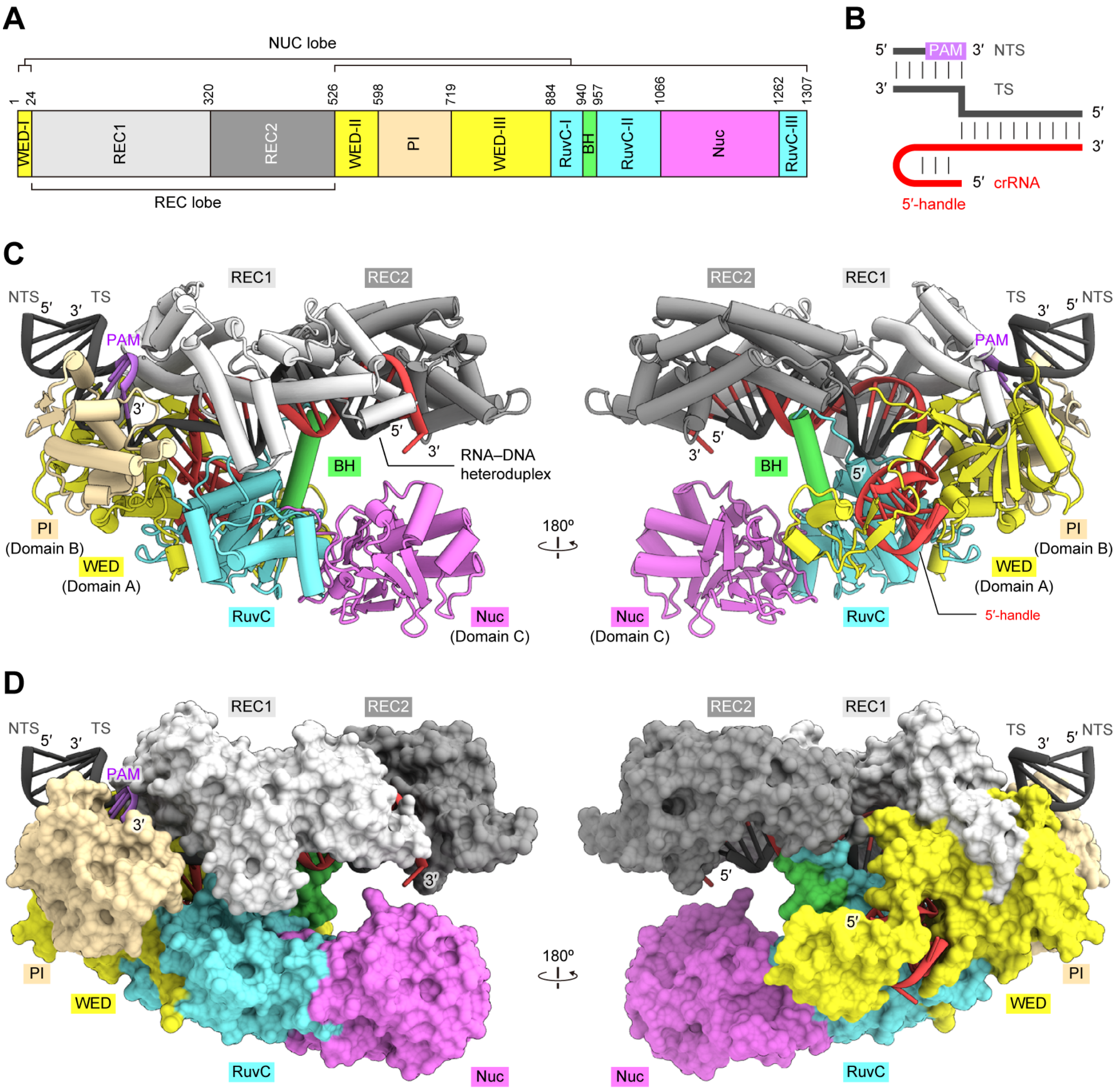


**Figure 2-6. Overall structure of the AsCpf1–crRNA–target DNA complex**

(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).

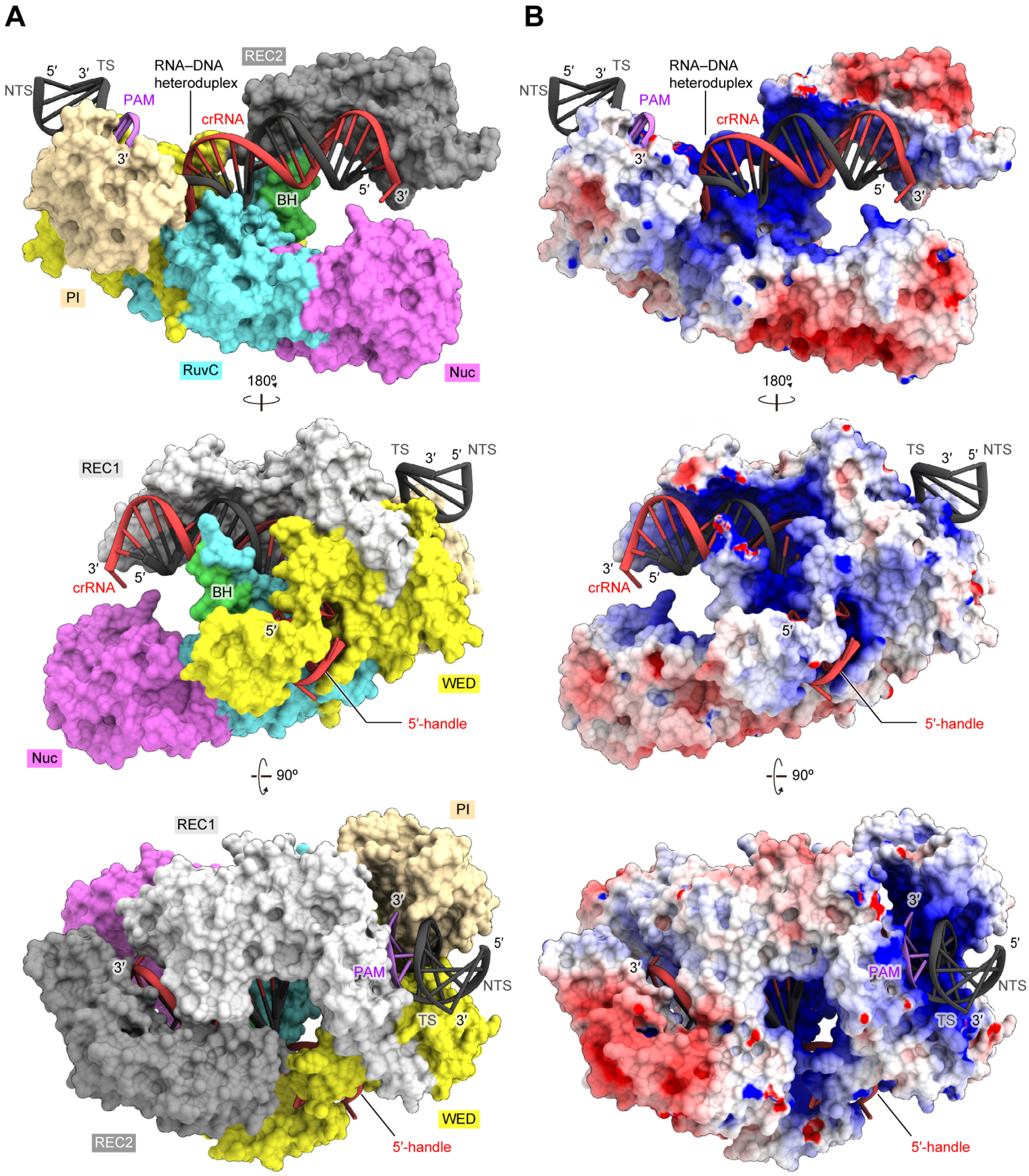


**Figure 2-7. Molecular surface of AsCpf1**

(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.

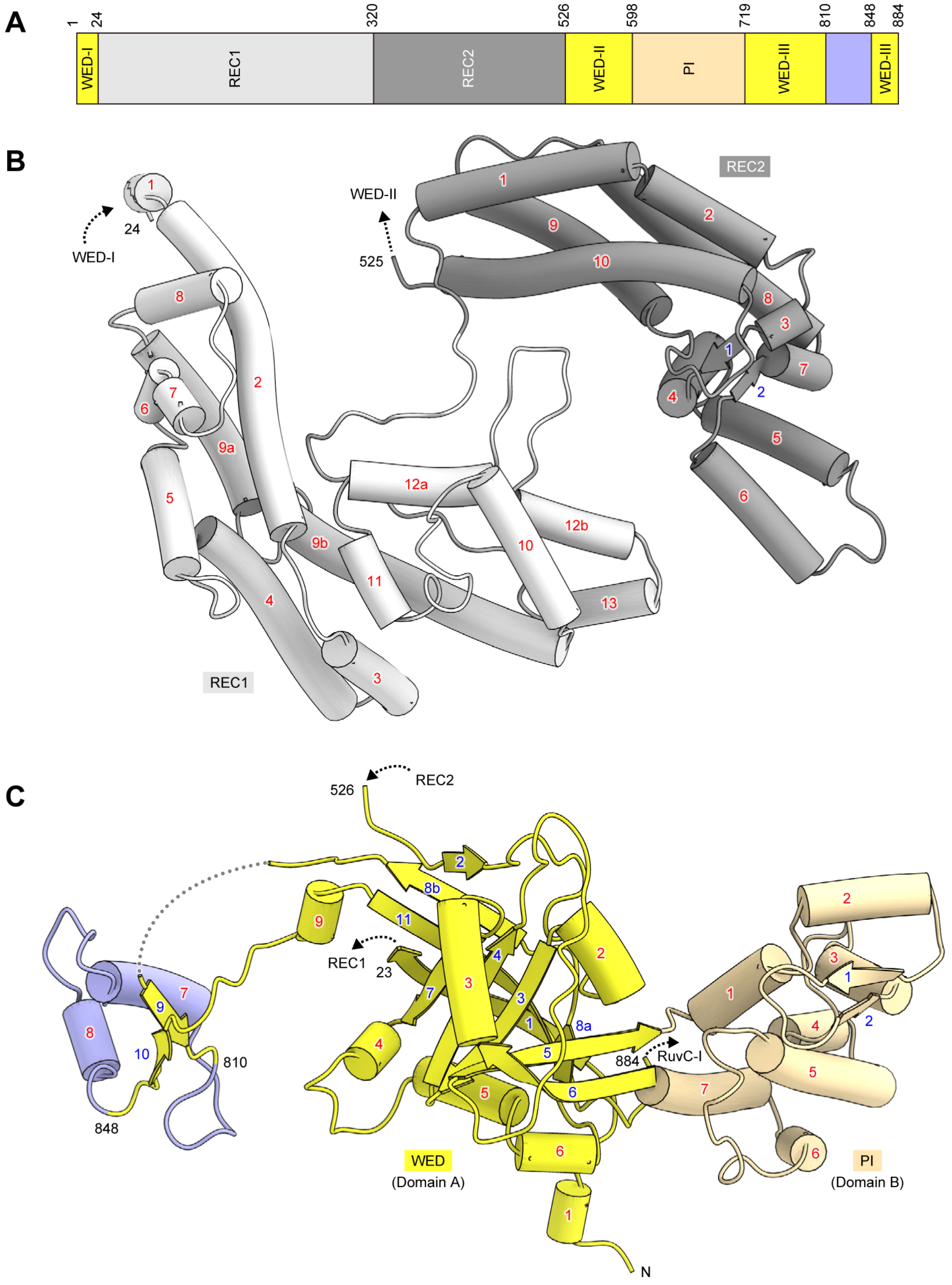


**Figure 2-8. Structure of REC1, REC2, WED and PI domains**

(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.

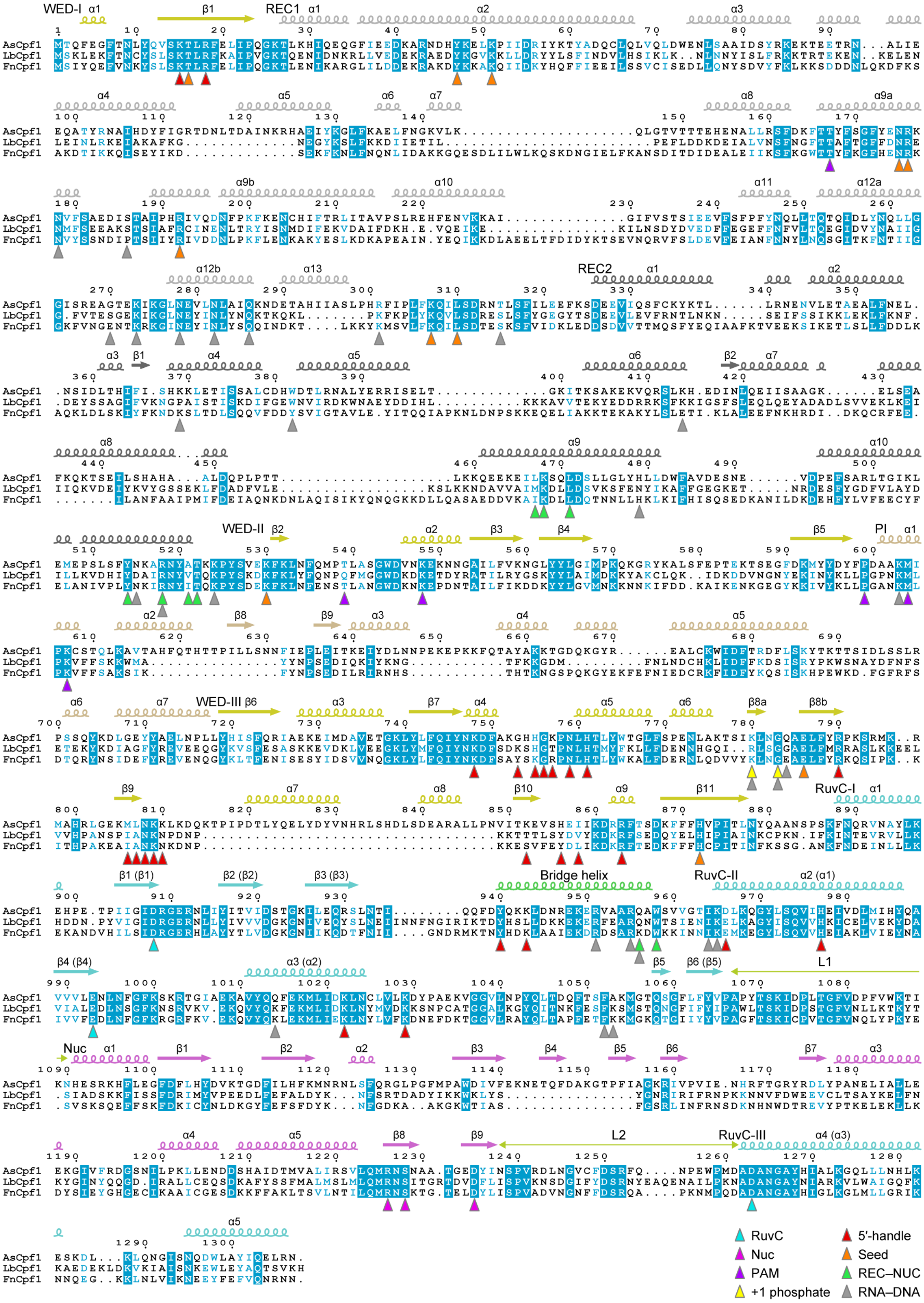

WED-I
REC1
REC2
WED-II
PI
WED-III
RuvC-I
Bridge helix
RuvC-II
L1
Nuc
L2
RuvC-III
AsCpf1
LbCpf1
FnCpf1
RuvC
Nuc
PAM
+1 phosphate
5′-handle
Seed
REC–NUC
RNA–DNA

**Figure 2-9. Multiple sequence alignment of Cpf1 proteins**

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).

### 2.5.7 Structure of 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 2-10A and 2-10B). The nucleotides G1–C20 in the crRNA and dC1–dG20 in the target DNA strand form the 20-bp RNA–DNA heteroduplex (Figures 2-10A and 2-10B). 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 2-10A and 2-10B).

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., 2015a) (Figures 2-10A and 2-10C). 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 2-11A). U(−10) and A(−18) form a reverse Hoogsteen A•U base pair, and participate in pseudoknot formation. The O4 and the 2′-OH of U(−10) hydrogen bond with the 2′-OH and the N1 of A(−19), respectively (Figure 2-11B). 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 2-11C). Importantly, U(−1), U(−10), U(−16) and A(−18) in the crRNA are conserved among the CRISPR-Cpf1 systems (Zetsche et al., 2015a), indicating that Cpf1 crRNAs form similar pseudoknot structures.

### 2.5.8 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 2-11D). 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 2-11E). 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., 2015a). The N6 of 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 2-11F). Moreover, the phosphodiester backbone of the 5′-handle forms an extensive network of interactions with the WED and RuvC domains (Figure 2-12). The residues involved in the crRNA 5′-handle recognition are largely conserved in the Cpf1 protein family (Zetsche et al., 2015a) (Figure 2-9), highlighting the functional relevance of the observed interactions between AsCpf1 and the crRNA.

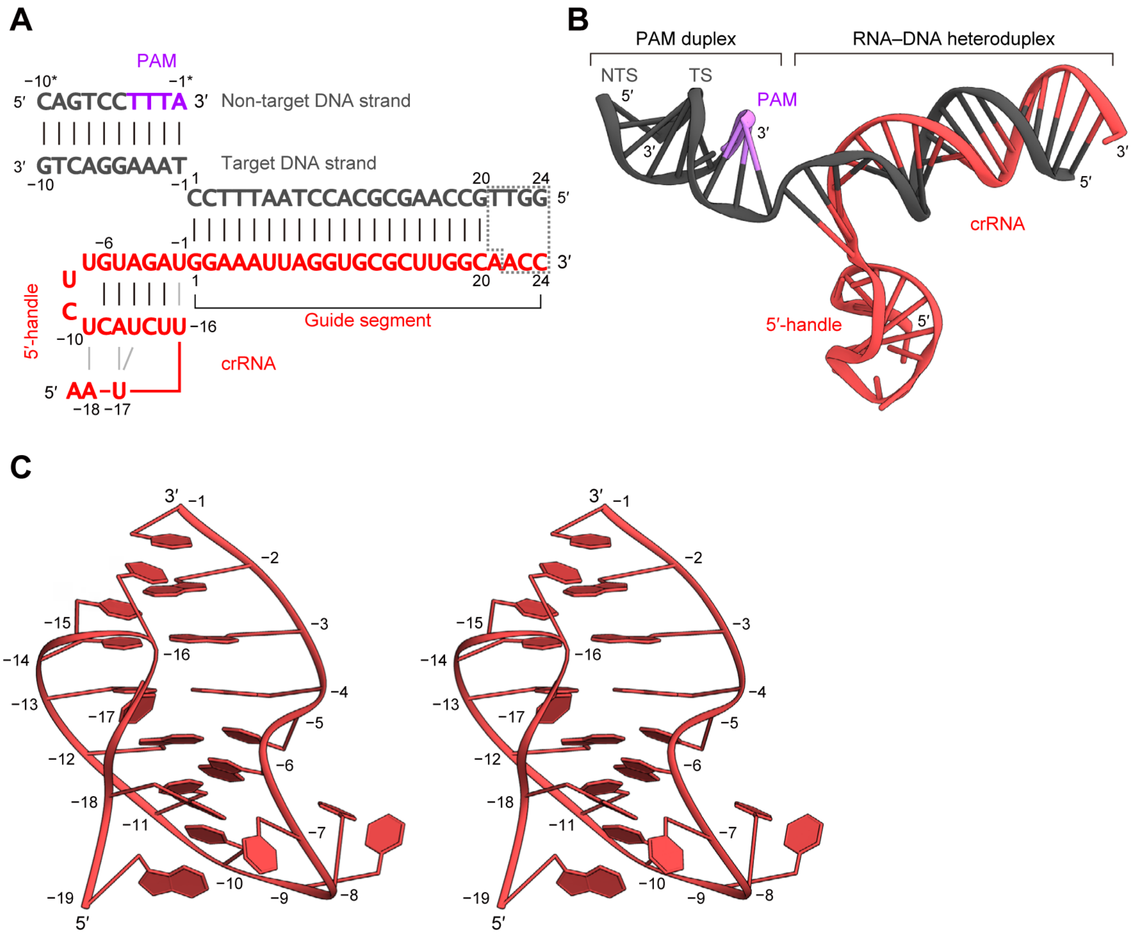


**Figure 2-10. Structure of the crRNA and target DNA**

(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).

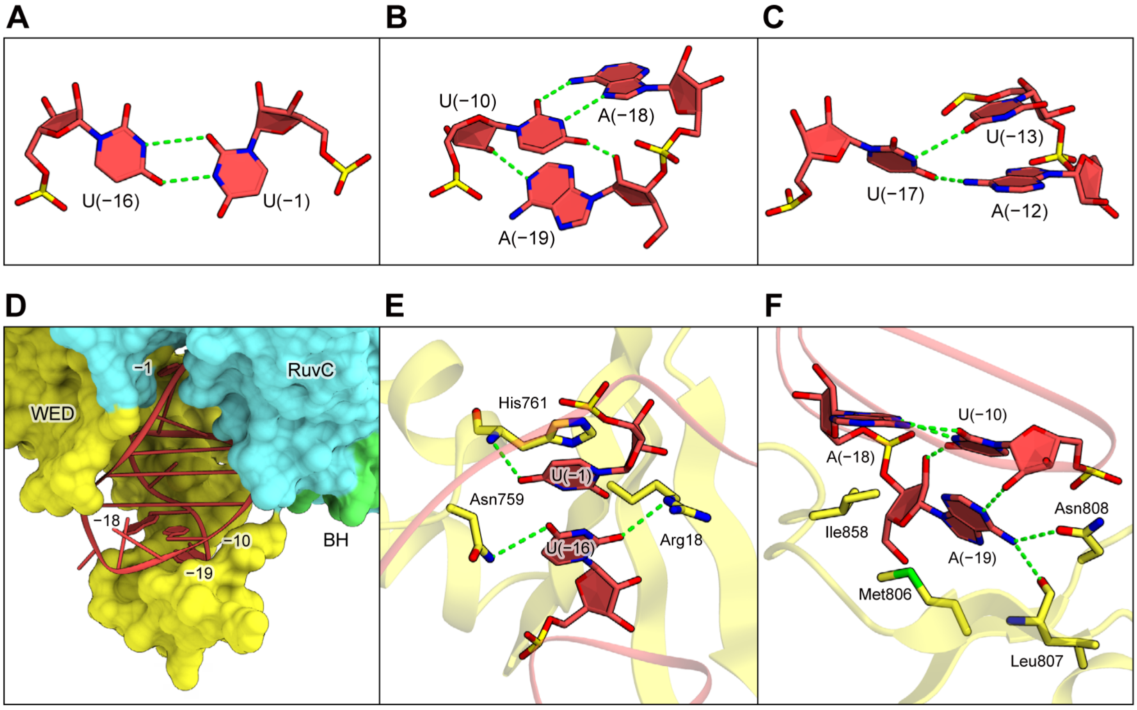


**Figure 2-11. Recognition of the 5′-handle**

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

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

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

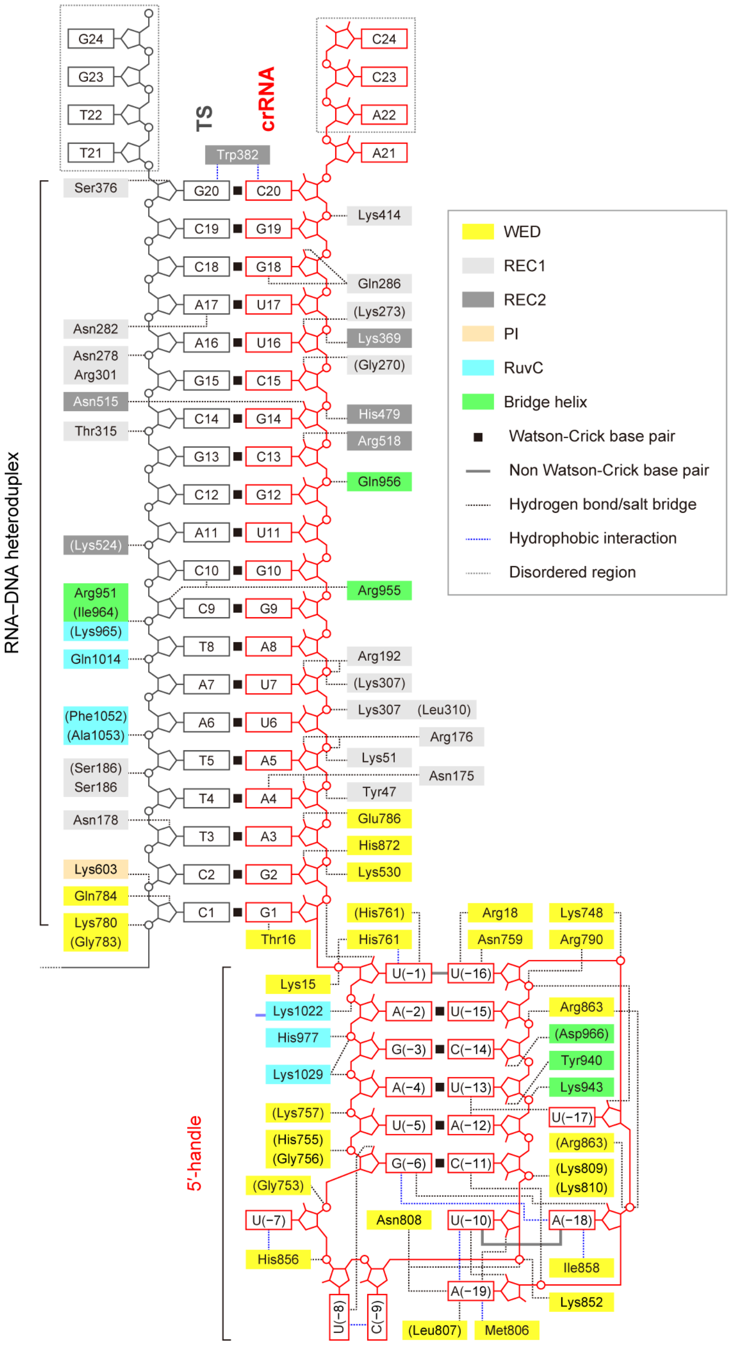


**Figure 2-12. Schematic of the nucleic acid recognition by Cpf1**

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.

### 2.5.9 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 2-7, 2-12 and 2-13). 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 2-12, 2-13 and 2-14). Arg951 and Arg955 in the bridge helix and Lys968 in the RuvC domain, which interact with the phosphate backbone of the target DNA strand (Figure 2-13B), are conserved among the Cpf1 family members (Zetsche et al., 2015a) (Figure 2-9). Notably, the sugar-phosphate backbone of the nucleotides G1–A8 in the crRNA forms multiple contacts with the WED and REC1 domains (Figures 2-12 and 2-14A), and the base pairing within the 5-bp PAM-proximal, "seed" region is important for Cpf1-mediated DNA cleavage (Zetsche et al., 2015a). 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 2-14A). 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., 2015a) (Figure 2-9), and the R176A, R192A, G783P and R951A mutants exhibited reduced activities (Figure 2-14B), 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 2-13A). 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 2-14C). Indeed, the W382A mutant showed reduced activity (Figure

2-14B), 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., 2015a) (Figure 2-9). These observations indicate that Cpf1 recognizes the 20-bp RNA–DNA heteroduplex, and can explain the previous finding that the *Francisella novicida* U112 Cpf1 (FnCpf1) cleaved the target DNA in a similar manner, using either the 20-nt or 24-nt guide-containing crRNA (Zetsche et al., 2015a).

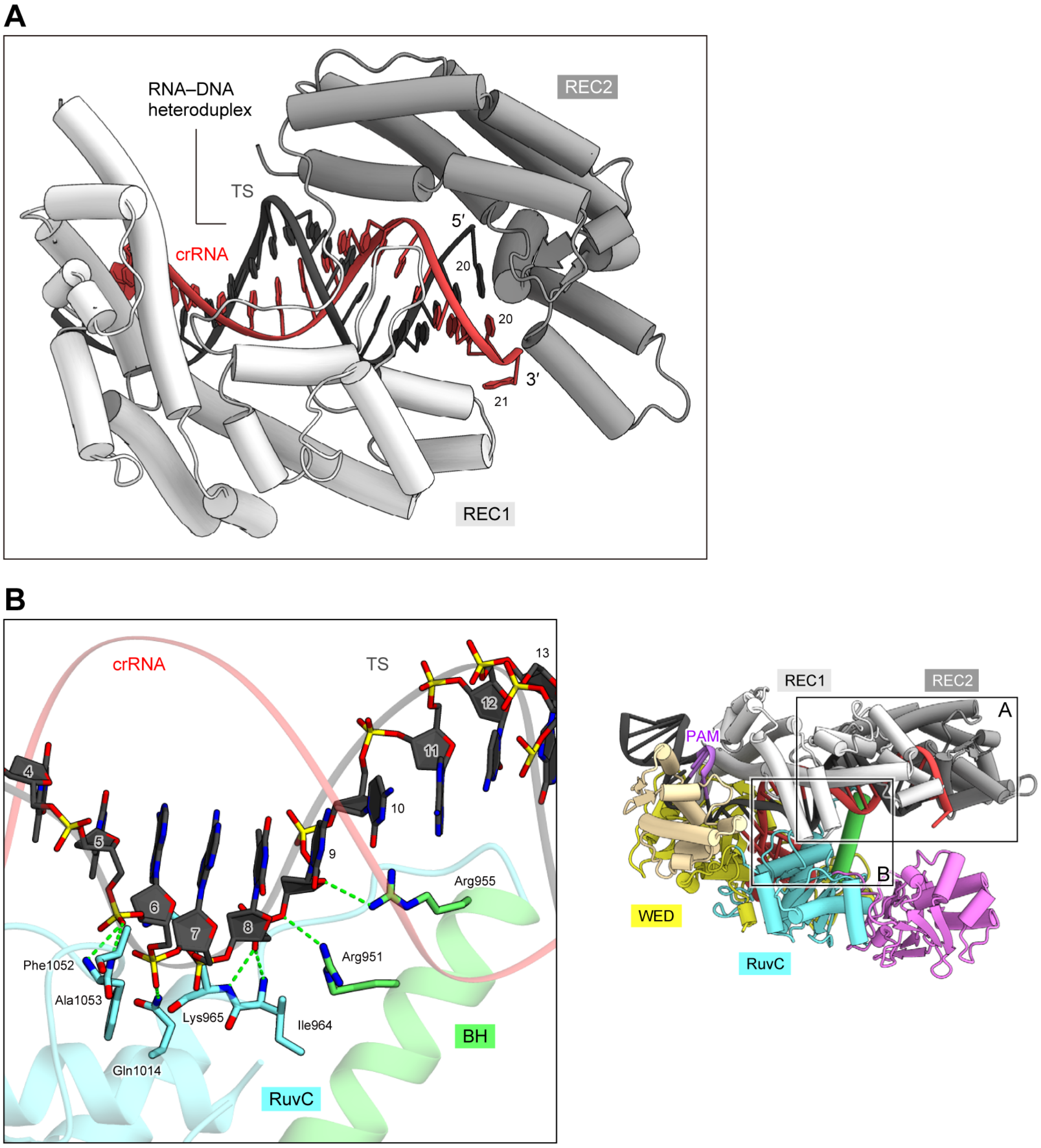


**Figure 2-13. Recognition of the crRNA–target DNA heteroduplex**

(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.

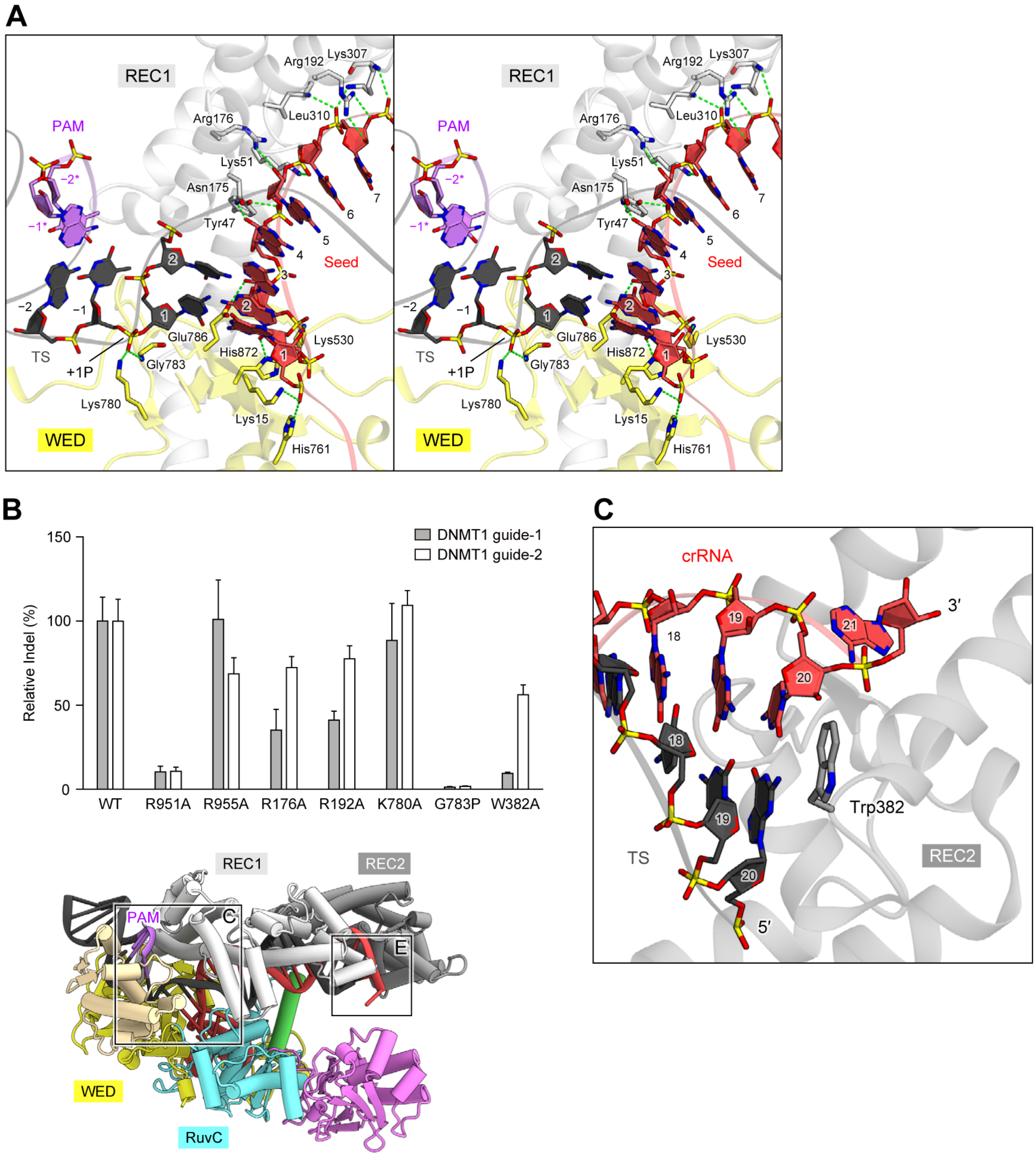


**Figure 2-14. Recognition of the crRNA–target DNA heteroduplex**

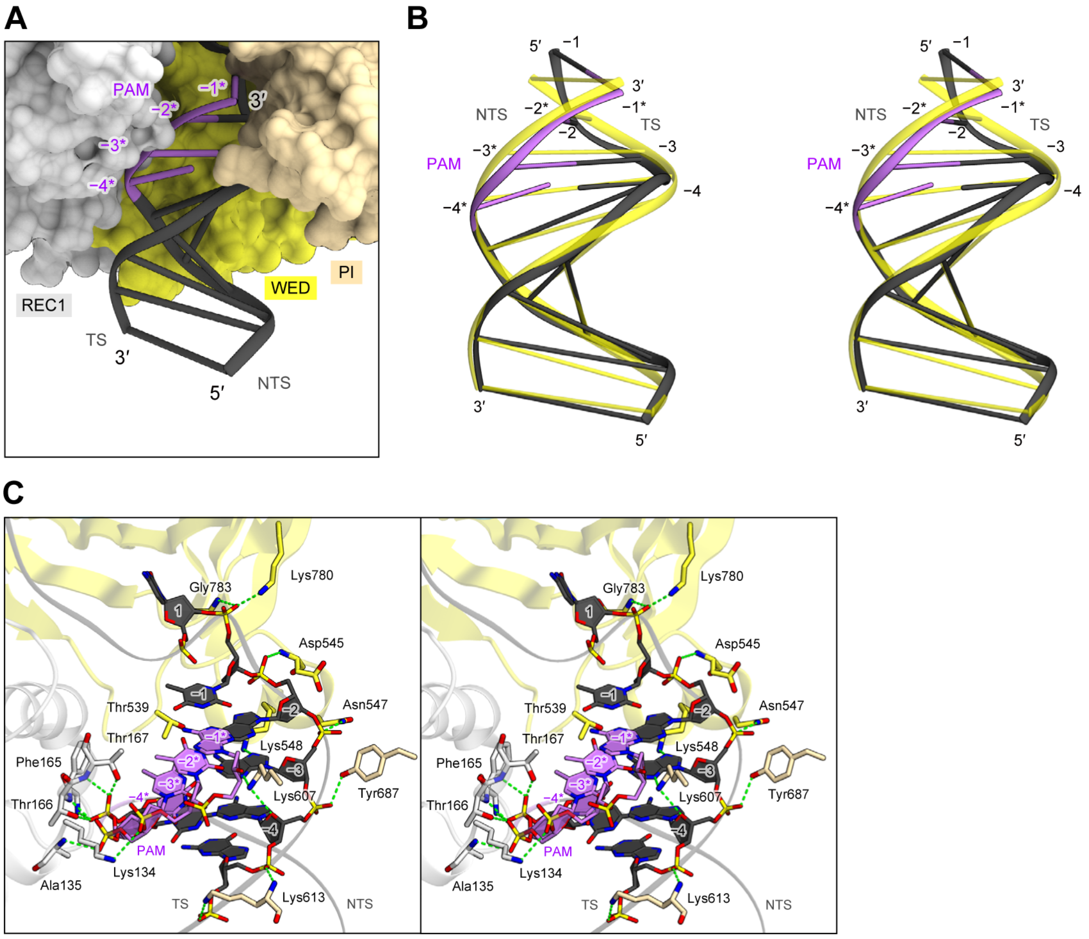
(A) Recognition of the crRNA seed region and the +1 phosphate group (+1P) (stereo view). Hydrogen bonds are shown as dashed lines.

(B) 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).

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

### 2.5.10 Recognition of the TTTV PAM

The PAM duplex adopts a distorted conformation with a narrow minor groove, as often observed in AT-rich DNA (Rohs et al., 2009; Stella et al., 2010), and is bound to the groove formed by the WED, REC1 and PI domains (Figures 2-15A and 2-15B). The PAM duplex is recognized by the WED, REC1 and PI domains from the major and minor groove sides, respectively (Figure 2-15C). The dT(−1):dA(−1*) base pair in the PAM duplex does not form base-specific contacts with the protein (Figures 2-15). Although the distorted shape of the PAM duplex is predicted to affects the preference for V in the 4th position, the mechanisms for the specificity was not fully explained based on this structure. Lys607 in the PI domain is inserted into the narrow minor groove, and plays critical roles in the PAM recognition (Figure 2-15C). 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 2-16A). 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 2-16D), 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 3rd T in the TTTV 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 2-16B). 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 2-16E). These observations are consistent with the requirement of the 2nd 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 2-16C). 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 2-16C). 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 2-16F). These structural observations are consistent with the requirement of the 1st T in the PAM. The K548A and M604A mutants exhibited reduced activities (Figure 2-16G), confirming that Lys548 and Met604 participate in the PAM recognition. More importantly, the K607A mutant showed almost no activity (Figure 2-16G), indicating that Lys607 is critical for the PAM recognition. Together, these results demonstrate that AsCpf1 recognizes the TTTV 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., 2015a) (Figure 2-9). 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.

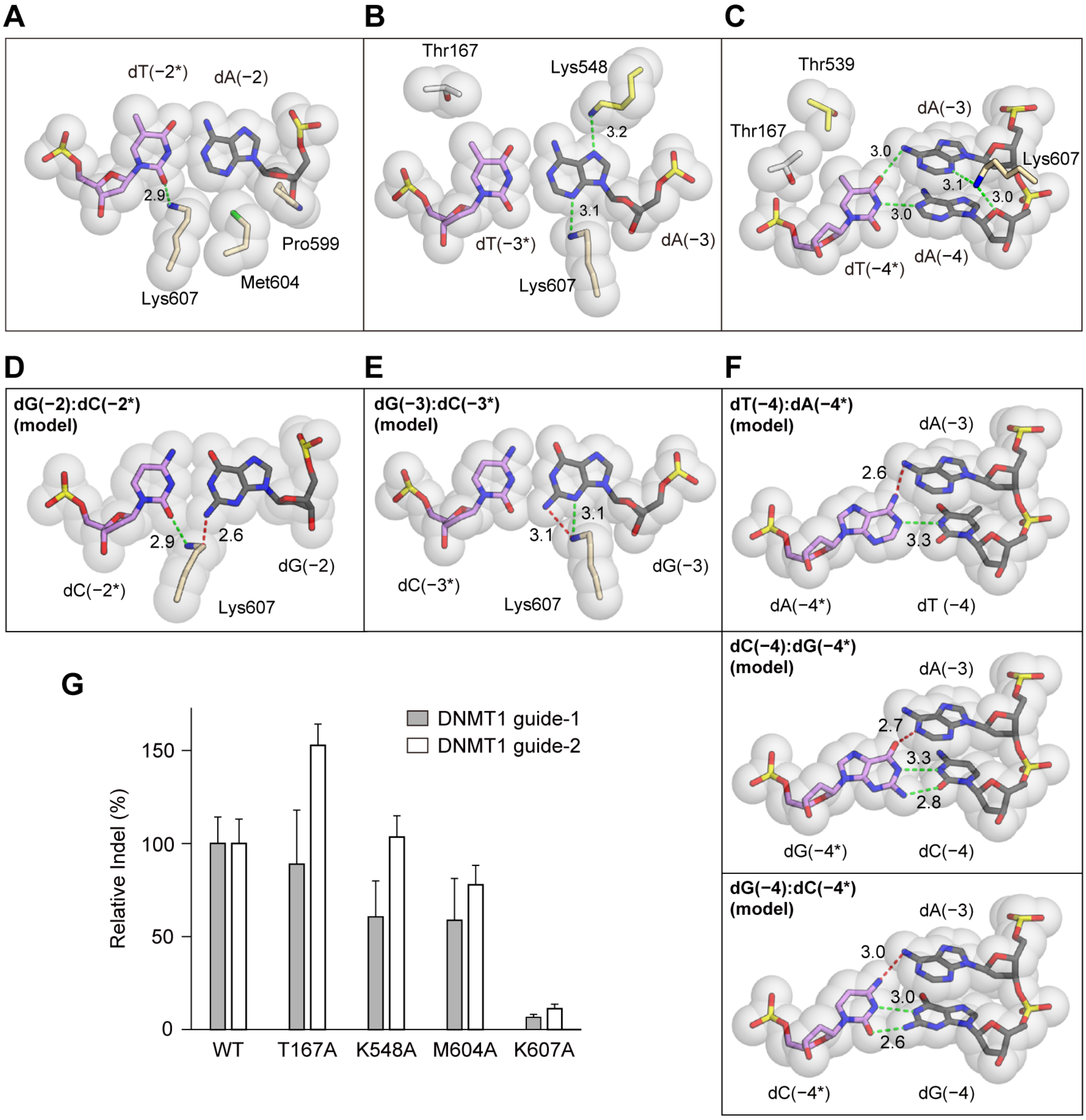


## Figure 2-15. Recognition of the TTTV PAM

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

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

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



### Figure 2-16. Recognition of the TTTV PAM

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

(D) 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.

(E) 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.

(F) 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 (D), (E) and (F), potential favorable and unfavorable interactions are depicted as green and red dashed lines, respectively.

(G) 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).

### 2.5.11 The RuvC-like endonuclease and the supporting domain

The RuvC domain comprises a typical RNase H fold, consisting of a 5-stranded mixed β-sheet (β1–β5) flanked by 3 α helices (α1–α3), and 2 additional α helices and a β strand (Figure 2-17A). 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 2-17B). As observed in FnCpf1 (Zetsche et al., 2015a), the D908A and E993A mutants had almost no activity, whereas the D1263A mutant exhibited significantly reduced activity (Figure 2-17C), 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 2-17A, 2-18A and 2-18B). 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 2-18B). 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 2-18B). These residues, with the exceptions of Leu467 and Ala521, are highly conserved among the Cpf1 family members (Zetsche et al., 2015a) (Figure 2-9), and the W958A mutant exhibited reduced activity (Figure 2-17C). 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 2-17A). The Nuc domain comprises 5 α helices and 9 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 partially conserved Ser1228, are clustered in the proximity of the active site of the RuvC domain (Zetsche et al., 2015a) (Figures 2-9 and 2-17B). The S1228A mutant showed DNA cleavage activity comparable to that of the WT AsCpf1 (Figure 2-17C). In contrast, the D1235A mutant exhibited reduced activity, whereas the R1226A mutant showed almost no activity (Figure 2-17C), 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 2-18C), indicating that the RuvC domain cleave the non-target DNA strand and the Nuc domain is indispensable for the cleavage of the target DNA strand (Figure 2-18A). As in FnCpf1 (Zetsche et al., 2015a), the mutations of the catalytic residues in the AsCpf1 RuvC domain abolished the cleavage of both DNA strands (Figure 19). The Nuc domain is predicted not a nuclease domain, suggesting that the Nuc domain do not cleave the target DNA strand, but guide the strand to the catalytic center of the RuvC domain. However, further functional and structural studies are required to fully characterize the RNA-guided DNA cleavage mechanism of Cpf1.

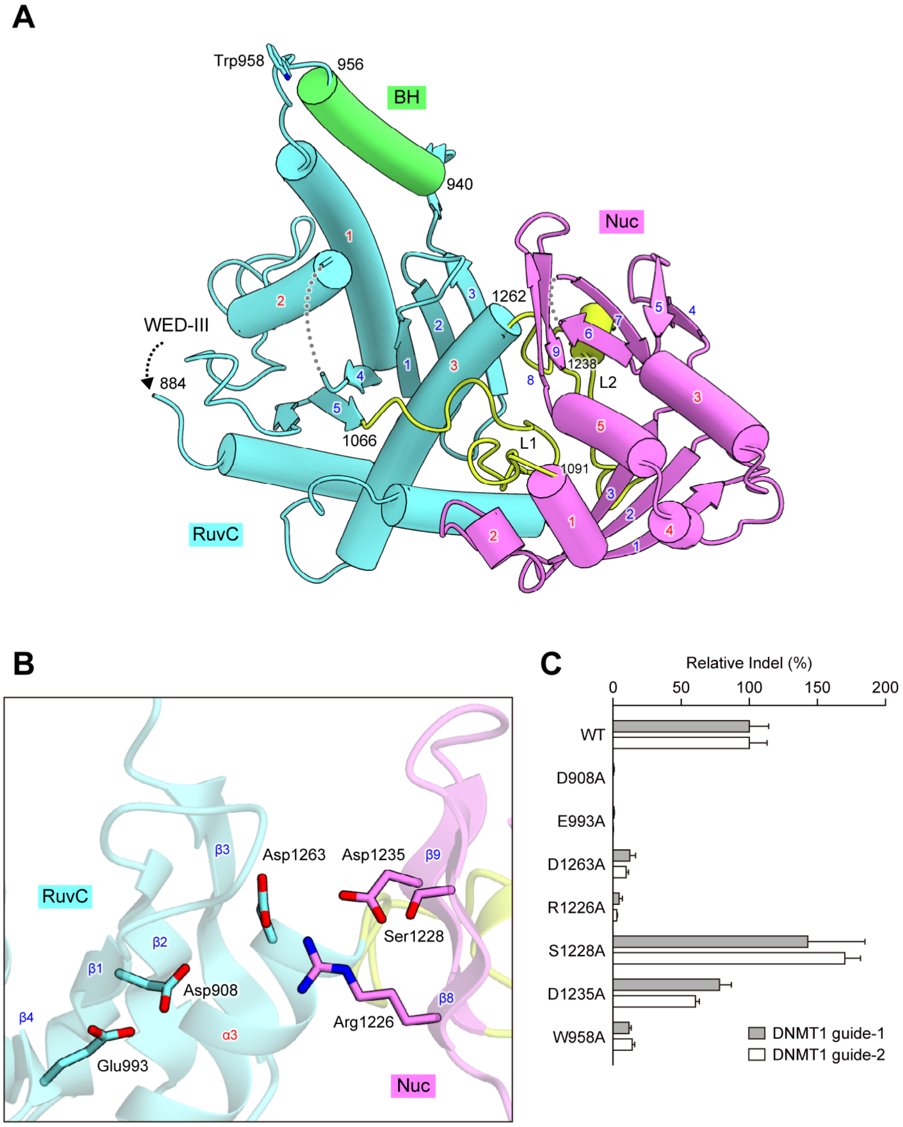


**Figure 2-17. The RuvC and Nuc nuclease domains**

(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 WT AsCpf1.

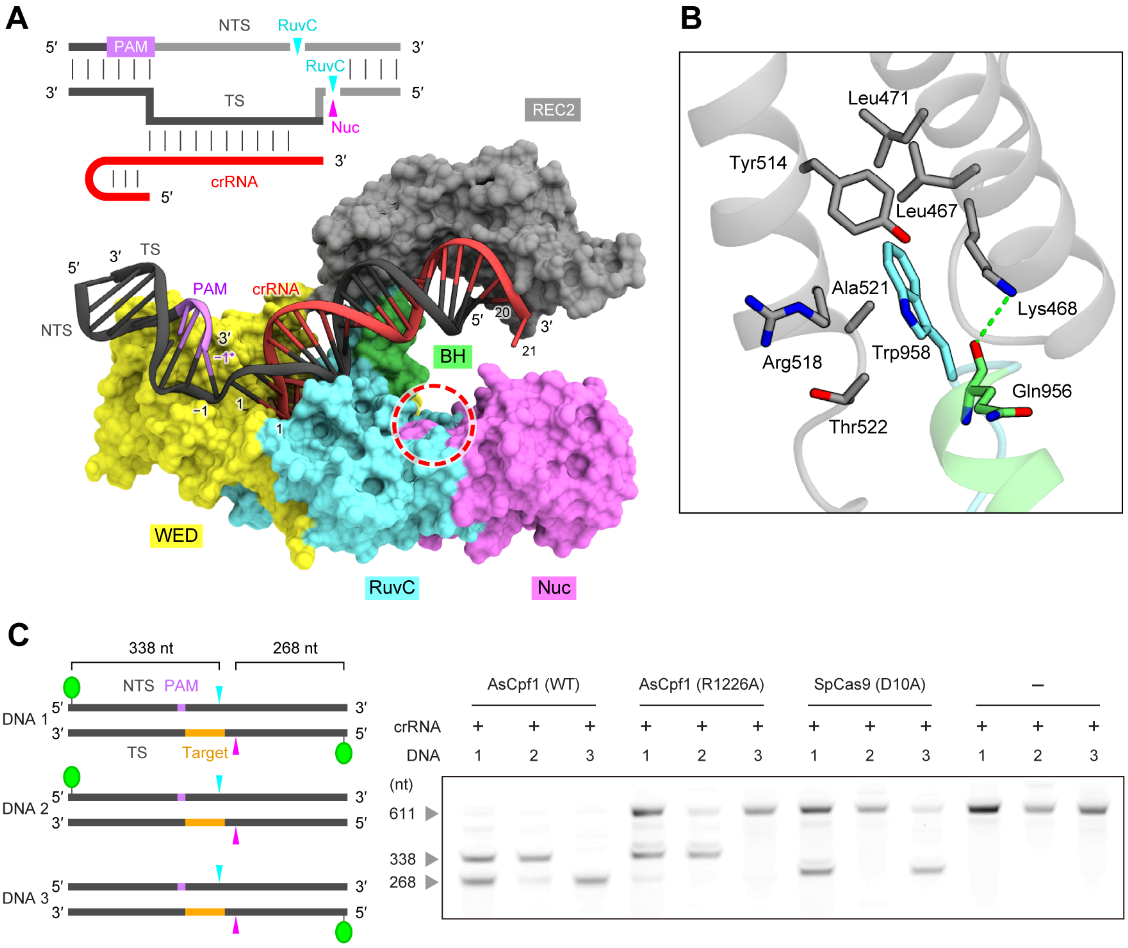


**Figure 2-18. The RuvC and Nuc nuclease domains**

(A) 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.

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

(C) 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.

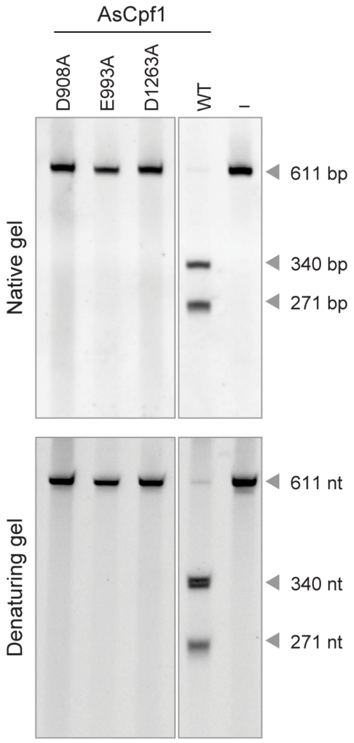


**Figure 2-19. Mutational analysis of the RuvC catalytic residues**
The wild type or mutants of AsCpf1 were incubated with the crRNA and double-stranded DNA target, and the reaction products were resolved on native TBE and denaturing TBE-Urea polyacrylamide gels. The gels were stained with SYBR Gold (Invitrogen). The mutations of the RuvC catalytic residues (D908A, E993A and D1263A) abolished the cleavage of both the target and non-target DNA strands.

## 2.6 Discussion

The present structure of the AsCpf1–crRNA–target DNA complex provides mechanistic insights into the 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 2-20A–2-20D). 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 2-20A and 2-20B). 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 2-20A and 2-20C). In Cas9, the HNH domain is inserted between strand β4 and helix α1 of the RNase H fold in the RuvC domain (Nishimasu et al., 2014; Anders et al., 2014) (Figure 2-20E). 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 2-20F). The Nuc 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). Our mutational analysis suggested that the Nuc domain is a ssDNA guiding domain responsible for the target DNA strand cleavage. Notably, the position of the Nuc domain of Cpf1 is reasonable to explain the role as a guide for the single-stranded region of the target DNA strand outside the heteroduplex and creating the staggered cleavage site with the 4 base overhang (Figures 2-20B and 2-20D), whereas the HNH domain of Cas9 cleaves the target DNA strand within the heteroduplex (Jinek et al., 2012; Gasiunas et al., 2012) (Figure 2-20C). These structural differences can also 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., 2015a). Unlike

*Streptococcus pyogenes* Cas9 (SpCas9), in which inactivation of the RuvC nuclease turns the enzyme into a nickase cleaving the target strand, an active RuvC domain is required for the cleavage of both strands by AsCpf1 (Figure 2-18C), suggesting further conformational changes of Cpf1 to cleave both DNA strands by single nuclease domain. Together, these findings indicate that, despite the overall structural similarity, 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. Especially, the previous report showed that the cleavage site of the target DNA strand is defined strictly (Zetsche et al., 2015a), and the guiding mechanism by the Nuc domain is predicted insufficient to explain the selection of the cleavage site of the target DNA strand. In SpCas9, the HNH–RuvC III linker acts as switch, and it communicates the conformational changes of the HNH domain to the RuvC domain, which is necessary to activate the RuvC domain for cleavage (Sternberg et al., 2015). In AsCpf1, although the conformational changes induced by the target DNA cleavage are not reported yet, there are possibilities that conformational changes and communications between domains may exist and assist in defining the cleavage site strictly and cleaving the both DNA strands by single nuclease domain.

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 2-21A). 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 2-21B). 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 2nd G and 3rd G in the NGG PAM are recognized by Arg1333 and Arg1335 in the PI domain, via bidentate hydrogen bonds, respectively (Anders et al., 2014) (Figure 2-21A and 2-22A). 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; Stella et al., 2010), 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 2-21B and 2-22B). 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, while the widely different members of the Cpf1 family recognize similar T-rich PAMs (Zetsche et al., 2015a).

In summary, the present structure of AsCpf1, combined with the mutational analysis of the RuvC and Nuc 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.

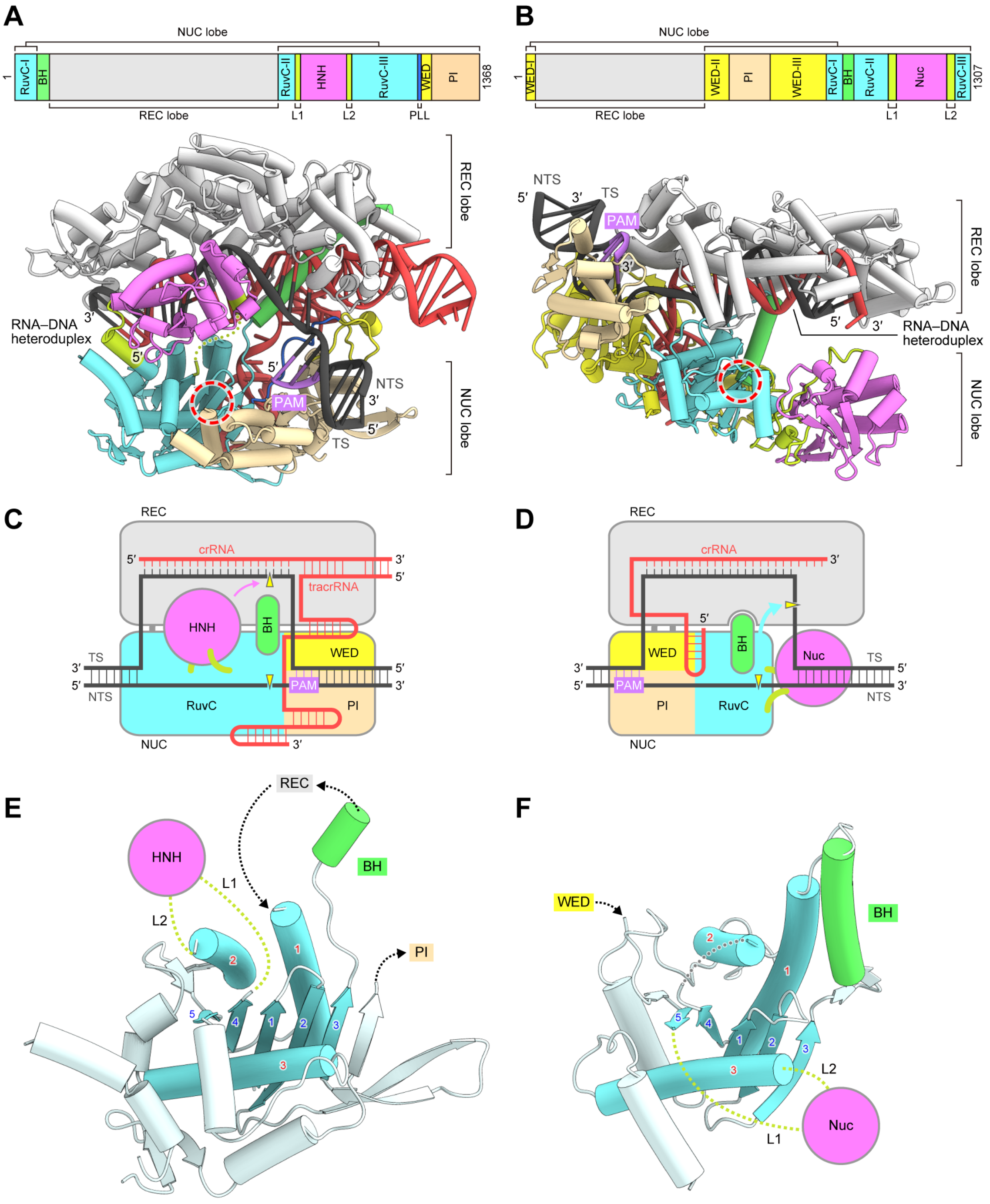


**Figure 2-20. Comparison between Cas9 and Cpf1**

(A and B) Comparison of the domain organizations and overall structures between SpCas9 (PDB ID 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 ID 4UN3) (E) and AsCpf1 (F). The secondary structures of the conserved RNase H fold are numbered.

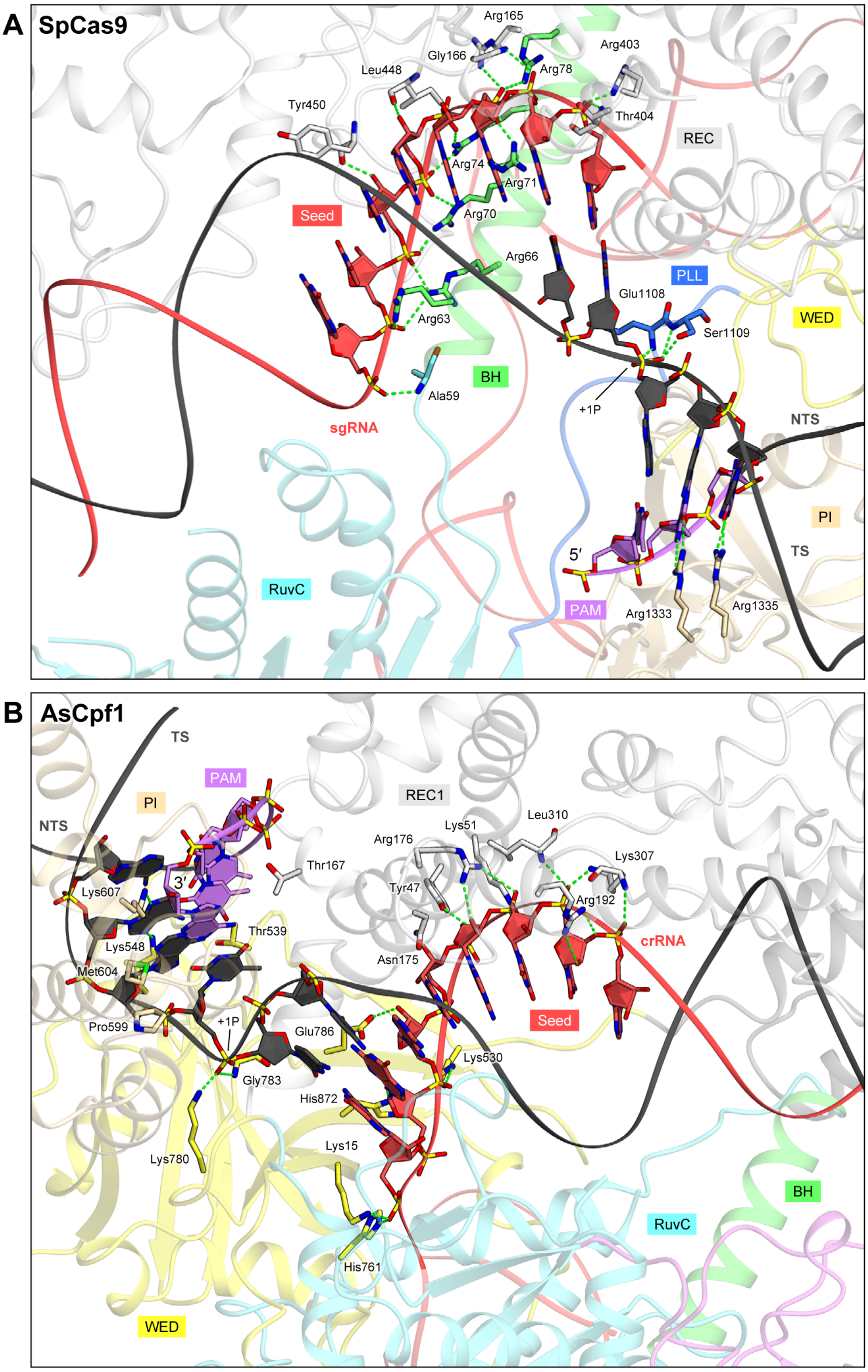


**Figure 2-21. RNA-guided DNA targeting mechanisms of SpCas9 (PDB: 4UN3) (A) and AsCpf1 (B)**

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.

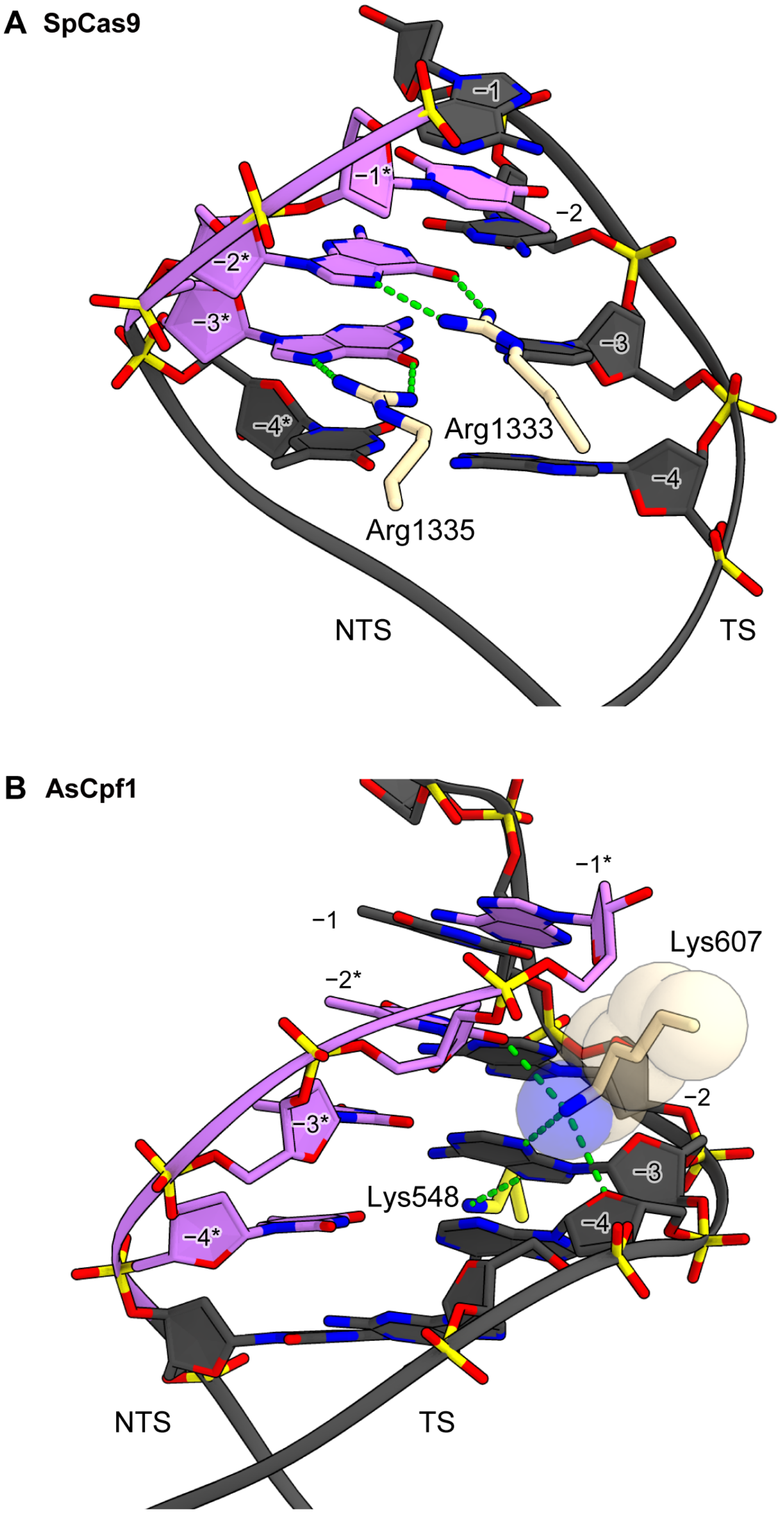


**Figure 2-22. PAM recognition mechanisms of SpCas9 (PDB: 4UN3) (A) and AsCpf1 (B)**

Key protein residues and nucleotides in the PAM duplex are shown as stick models. Hydrogen bonds are shown as dashed lines.

# Chapter 3: Structural basis for the altered PAM recognition by engineered CRISPR-Cpf1

## 3.1 Summary

The RNA-guided Cpf1 nuclease cleaves double-stranded DNA targets complementary to the CRISPR RNA (crRNA), and has been harnessed for genome editing technologies. Recently, *Acidaminococcus sp. BV3L6* (AsCpf1) was engineered to recognize altered DNA sequences as the protospacer adjacent motif (PAM), thereby expanding the target range of Cpf1-mediated genome editing. Whereas WT AsCpf1 recognizes the TTTV PAM, the RVR (S542R/K548V/N552R) and RR (S542R/K607R) variants can efficiently recognize the TATV and TYCV PAMs, respectively. However, their PAM recognition mechanisms remained unknown. Here, we present the 2.0 Å resolution crystal structures of the RVR and RR variants bound to a crRNA and its target DNA. The structures revealed that the RVR and RR variants primarily recognize the PAM-complementary nucleotides *via* the substituted residues. Our high-resolution structures delineated the altered PAM recognition mechanisms of the AsCpf1 variants, providing a basis for the further engineering of CRISPR-Cpf1.

## 3.2 Introduction

The crystal structure of the AsCpf1–crRNA–target DNA complex provided mechanistic insights into the crRNA-guided DNA recognition and cleavage (Yamano et al., 2016; Gao et al., 2016). AsCpf1 adopts a bilobed architecture that accommodates the crRNA–target DNA heteroduplex. The PAM-containing DNA duplex adopts a distorted conformation characteristic of an AT-rich DNA duplex, and is recognized by AsCpf1 *via* the base and shape readout mechanisms (Yamano et al., 2016). The RuvC and Nuc domains are located at positions suitable to induce staggered DNA double-strand breaks at the PAM-distal positions. A structural comparison of the AsCpf1–crRNA–target DNA ternary complex (Gao et al., 2016; Yamano et al., 2016) with the LbCpf1–crRNA binary complex (Dong et al., 2016) indicated a structural rearrangement accompanying the crRNA–target DNA heteroduplex formation. Furthermore, a structural comparison of Cpf1 with Cas9 explained their distinct functionalities, and suggested the functional convergence between the class 2 CRISPR-Cas effector nucleases (Yamano et al., 2016).

Recently, a structure-guided mutagenesis screen identified two AsCpf1 variants with altered PAM specificities (Gao et al., 2017). The RVR variant contains three substitutions (S542R/K548V/N552R), and efficiently cleaves target sites with the non-canonical TATV PAM, in addition to those with the canonical TTTV PAM. In contrast, the RR variant contains two substitutions (S542R/K607R), and cleaves target sites with the non-canonical TYCV PAM, including the T-less CCCC PAM. Importantly, these two AsCpf1 variants showed robust activities in human cells, thus contributing to the expansion of target spaces in Cpf1-mediated genome editing; however, the altered PAM recognition mechanisms of these variants remained unknown. It is particularly interesting to determine how the K607R substitution contributes to the altered PAM recognition by the RR variant, since Lys607 in the PI (PAM-interacting) domain is critical for the TTTV PAM recognition by WT AsCpf1 (Yamano et al., 2016). It is also unknown how the S542R substitution functions in the distinct PAM recognitions by the RVR and RR variants. In addition, the means by which the K548V and N552R substitutions participate in the altered PAM recognition by the RVR variant are not readily predictable.

## 3.3 Research aims

Here, to reveal how the AsCpf1 variants achieve the altered the PAM recognition, we present the high-resolution crystal structures of the RVR and RR variants of AsCpf1 in complexes with the crRNA and its target DNA with the altered PAMs. Furthermore, we try to reveal similarities and differences in the altered PAM recognition mechanisms between the AsCpf1 and SpCas9 variants by a structural comparison.

## 3.4 Materials and methods

### 3.4.1 Sample preparation

WT AsCpf1 and the RVR and RV variants were prepared essentially as described previously (Yamano et al., 2016). The gene encoding full-length AsCpf1 (residues 1–1307) was cloned into the modified pE-SUMO vector (LifeSensors), and the mutations (S542R, K548V, N552R and K607R) were introduced by a PCR-based method. The plasmid DNAs were amplified in *Escherichia coli* Mach (Thermo Fisher Scientific), cultured in LB medium (Nacalai Tesque) at 37°C overnight. The AsCpf1-expressing *E. coli* Rosetta2 (DE3) cells were cultured at 37°C in LB medium (containing 20 mg/l kanamycin) until the $OD_{600}$ reached 0.8, and protein expression was then induced by the addition of 0.1 mM isopropyl-β-D-thiogalactopyranoside (IPTG) (Nacalai Tesque). The *E. coli* cells were further cultured at 20°C for 18 h, and harvested by centrifugation at 5,000 g for 10 min. The *E. coli* cells were resuspended in buffer A (50 mM Tris-HCl, pH 8.0, 20 mM imidazole, 300 mM NaCl and 3 mM 2-mercaptoethanol), lysed by sonication, and then centrifuged at 40,000 g for 30 min. The supernatant was mixed with 5 ml Ni-NTA Superflow (QIAGEN) equilibrated with buffer A for about 1 hour at 4°C, and the mixture was loaded into an Econo-Column (Bio-Rad). The resin was washed with 5 column volumes of buffer A, 5 column volumes of buffer B (50 mM Tris-HCl, pH 8.0, 20 mM imidazole, 1 M NaCl and 3 mM 2-mercaptoethanol) and 3 column volumes of buffer A. The protein was eluted with buffer C (50 mM Tris-HCl, pH 8.0, 300 mM imidazole, 300 mM NaCl and 3 mM 2-mercaptoethanol). The protein was loaded onto a HiTrap SP HP column (GE Healthcare) equilibrated with buffer D (20 mM Tris-HCl, pH 8.0 and 200 mM NaCl). The column was washed with buffer D, and the protein was then eluted with a linear gradient of 200–1,000 mM NaCl. To remove the $His_6$-SUMO-tag, the eluted protein was mixed with TEV protease (home made), and was dialyzed at 4°C for 12 h against buffer E (20 mM Tris-HCl, pH 8.0, 40 mM imidazole, 300 mM NaCl and 3 mM 2-mercaptoethanol). The protein was passed through the Ni-NTA column equilibrated with buffer E. The protein was concentrated using an Amicon Ultra 10K filter (Millipore), and was further purified by a HiLoad Superdex 200 16/60 column (GE Healthcare) equilibrated with buffer F (10 mM Tris-HCl, pH 8.0, 150 mM NaCl and 1 mM DTT). The purified AsCpf1 proteins were stored at −80°C until use. The crRNA and target DNA were purchased from Gene Design and Sigma-Aldrich, respectively. The purified AsCpf1 protein was mixed with the crRNA, the target DNA strand and the non-target DNA strand (molar ratio, 1:1.5:2.3:2.3), and the reconstituted complex was concentrated using an Amicon Ultra 10K filter. The AsCpf1–crRNA–target DNA complex was purified by gel filtration chromatography on a Superdex 200 Increase column (GE Healthcare) equilibrated with buffer F.

### 3.4.2 *In vitro* cleavage assay

*In vitro* cleavage experiments were performed as previously described (Nishimasu et al., 2015), with minor modifications. The target pUC119 plasmids with the different PAMs were generated by a PCR-based method. The plasmid DNAs were amplified in *Escherichia coli* Mach (Thermo Fisher Scientific), cultured in LB medium (Nacalai Tesque) at 37°C overnight. The *Eco*RI-linearized pUC119 plasmid (100 ng), containing the 24-nt target sequence and the PAMs, was incubated at 37°C for 5 or 10 min with the AsCpf1–crRNA complex (100 nM), in 10 μl of reaction buffer containing 20 mM HEPES-NaOH, pH 7.5, 100 mM KCl, 2 mM $MgCl_2$, 1 mM DTT and 5% glycerol. The reaction was stopped by the addition of a solution containing EDTA (40 mM final concentration) and Proteinase K (10 μg). Reaction products were resolved on an ethidium bromide-stained 1% agarose gel, and then visualized using an Amersham Imager 600 (GE Healthcare). *In vitro* cleavage experiments were performed at least three times, and representative results were shown.

### 3.4.3 Crystallization

The peak fractions of the purified AsCpf1–crRNA–target DNA complex were concentrated by Amicon Ultra 10K filter ($A_{260\ \mathrm{nm}}$ = 10). The 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 ($A_{260\ \mathrm{nm}}$ = 10) and 1 μl of reservoir solution (7–10% PEG 3,350, 100 mM sodium acetate, pH 4.5, and 10% 1,6-hexanediol), and then were incubated against 0.5 ml of reservoir solution.

### 3.4.4 X-ray diffraction analysis and data processing

The crystals were cryoprotected in a solution consisting of 9–10% PEG 3,350, 100 mM sodium acetate, pH 4.5, 10–15% 1,6-hexanediol and 30% ethylene glycol. To reduce radiation damage, all X-ray-diffraction data were collected at 100 K on the beamline BL41XU at SPring-8. A diffraction data set was collected from a single crystal using a X-ray beam at a wavelength of 1.000 Å, an oscillation range of 180° (0.1° per image), an exposure time of 1.0 s per image. The X-ray-diffraction data were processed using DIALS (Waterman et al., 2013) and AIMLESS (Evans and Murshudov, 2013).

### 3.4.5 Phase determination, model building and structure refinement

The structures were determined by molecular replacement with Molrep (Vagin and Teplyakov, 2010), using the coordinates of WT AsCpf1 (PDB: 5B43) (Yamano et al., 2016) as the search model. The model building and structural refinement were performed using COOT (Emsley and Cowtan, 2004) and PHENIX (Adams et al., 2010), respectively. Structural figures were prepared using CueMol (http://www.cuemol.org).

## 3.5 Results

### 3.5.1 Sample preparation

The RVR and RR variants were highly purified by the same strategy for the purification of WT AsCpf1. The gel filtration chromatogram peaks showed monodispersity (Figures 3-1). The final yield of the RVR and RR variants were 11.4 mg and 14.9 mg per 1 L culture medium, respectively.

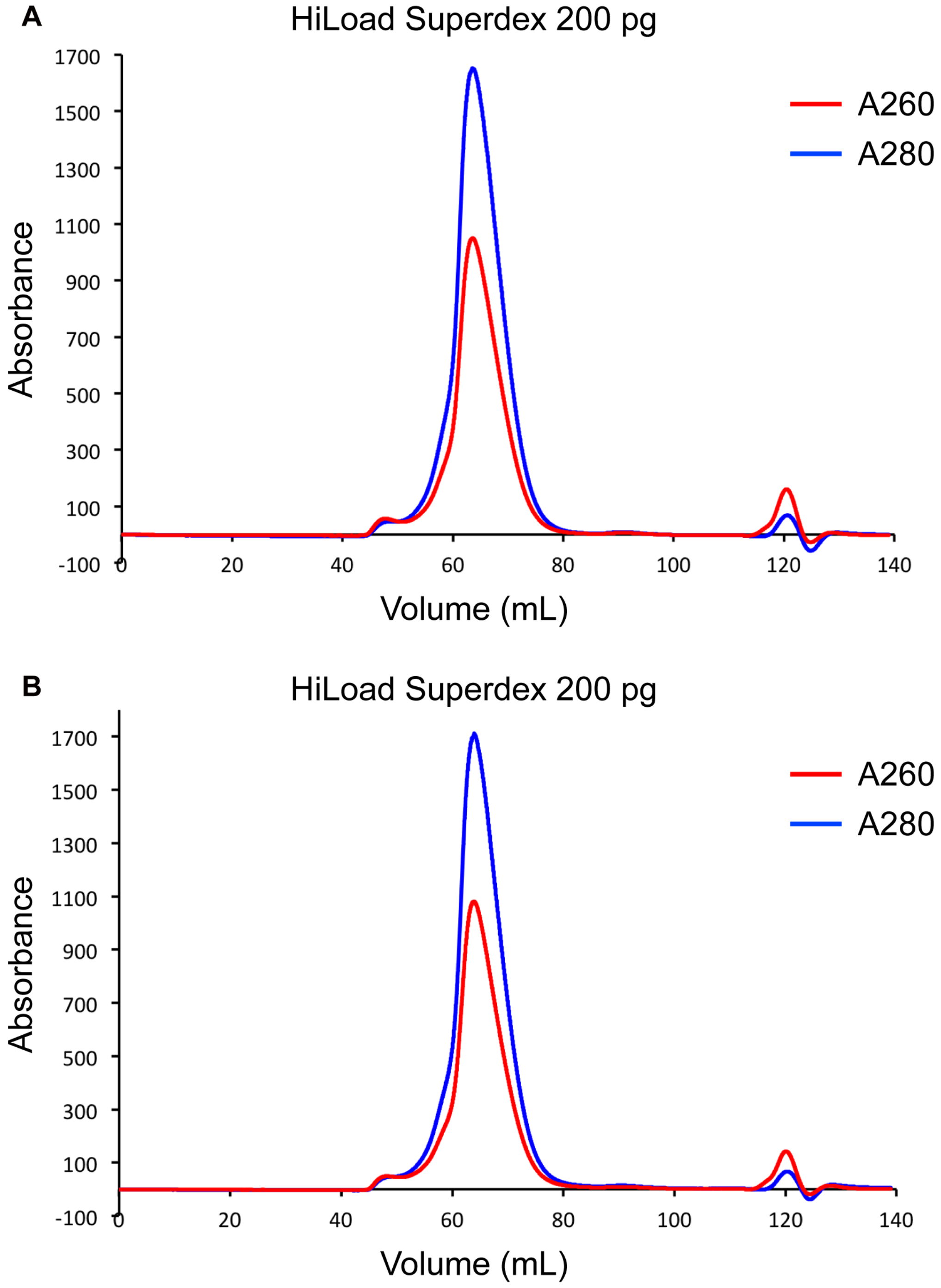


**Figure 3-1. Size-exclusion chromatogram**

(A) Chromatogram of AsCpf1 RVR variant.

(B) Chromatogram of AsCpf1 RR variant.

### 3.5.2 *In vitro* cleavage activities of the AsCpf1 variants

In a previous study, the PAM specificities of the two AsCpf1 variants were determined by a PAM identification assay, in which AsCpf1-expressing HEK293T cell lysates were incubated with the crRNA and a library of plasmids containing a constant target sequence and a degenerate PAM (Gao et al., 2017). We thus evaluated the *in vitro* cleavage activities of the purified RVR and RR variants toward plasmid DNA substrates containing a 24-nt target sequence and different potential PAMs.

Since the RVR variant efficiently cleaved target sites with the TATV PAM (Gao et al., 2017), we examined the ability of the RVR variant to cleave 13 plasmid DNA targets with either NATA, TNTA, TANA or TATN as the potential PAM (Figures 3-2A and 3-3A). The RVR variant efficiently cleaved the TATA target site, as compared with the VATA (V is A, G or C) target sites (Figures 3-2A and 3-3A), confirming the preference for the first T in the TATV PAM. In addition to the altered TATA PAM, the RVR variant efficiently recognized the canonical TTTA PAM (Figures 3-2A and 3-3A), consistent with a previous study (Gao et al., 2017) (also described below). The RVR variant was almost inactive toward the TSTA (S is G or C) and TARA (R is A or G) sites, and less active toward the TACA site (Figures 3-2A and 3-3A). These results confirmed the strong preference of the second A and the third T for the TATV PAM recognition by the RVR variant. The RVR variant was less active toward the TATT site, as compared with the TATV sites (Figures 3-2A and 3-3A), indicating the preference for the fourth V. These results demonstrated that the RVR variant efficiently recognizes TATV and TTTV as the PAM, consistent with a previous study (Gao et al., 2017).

Since the RR variant efficiently cleaved target sites with the TYCV PAM (Y is T or C) (Gao et al., 2017), we examined the RR variant for the ability to cleave 13 plasmid DNA targets with either NCCC, TNCC, TCNC or TCCN as the potential PAM (Figures 3-2B and 3-3B). The RR variant efficiently cleaved the TCCC site, as compared to the VCCC sites (Figures 3-2B and 3-3B), indicating the preference for the first T in the TYCV PAM. The RR variant efficiently cleaved the TYCC sites, but not the TRCC sites (Figures 3-2B and 3-3B), confirming the preference of the second Y for the PAM recognition by the RR variant. The RR variant was almost inactive toward the TCDC PAMs (D is A, T or G) (Figures 3-2B and 3-3B), confirming the strong preference of the third C for the PAM recognition by the RR variant. The RR variant was much less active toward the TCCT site, as compared with the TCCV sites (Figures 3-2B and 3-3B), indicating the preference for the fourth V. These results confirmed that the RR variant efficiently recognizes TYCV as the PAMs, consistent with earlier observations (Gao et al., 2017).

The RVR variant recognizes the TTTV PAM as well as the TATV PAM, whereas the RR variant is less active towards the TTTV PAM (Gao et al., 2017). We thus examined the *in vitro* cleavage activities of WT AsCpf1 and the RVR and RV variants towards four plasmid targets with the TTTN PAMs (Figures 3-2C and 3-3C). WT AsCpf1 and the RVR variants efficiently cleaved the TTTV sites, as compared with the TTTT site (Figures 3-2C and 3-3C), consistent with previous studies (Gao et al., 2017; Kim et al., 2016b; Zetsche et al., 2015a). The RR variant was less active toward the TTTV sites, as compared with WT AsCpf1 and the RVR variant (Figures 3-2C and 3-3C). In stark contrast to the RVR and RR variants, WT AsCpf1 exhibited no or little activity toward the TATV and TYCV sites (Figures 3-4), highlighting the substantial differences in the PAM specificities between WT AsCpf1 and the two AsCpf1 variants. Together, our *in vitro* cleavage experiments confirmed that, unlike WT AsCpf1, the RVR and RR variants efficiently recognize the TATV and TYCV PAMs, respectively.

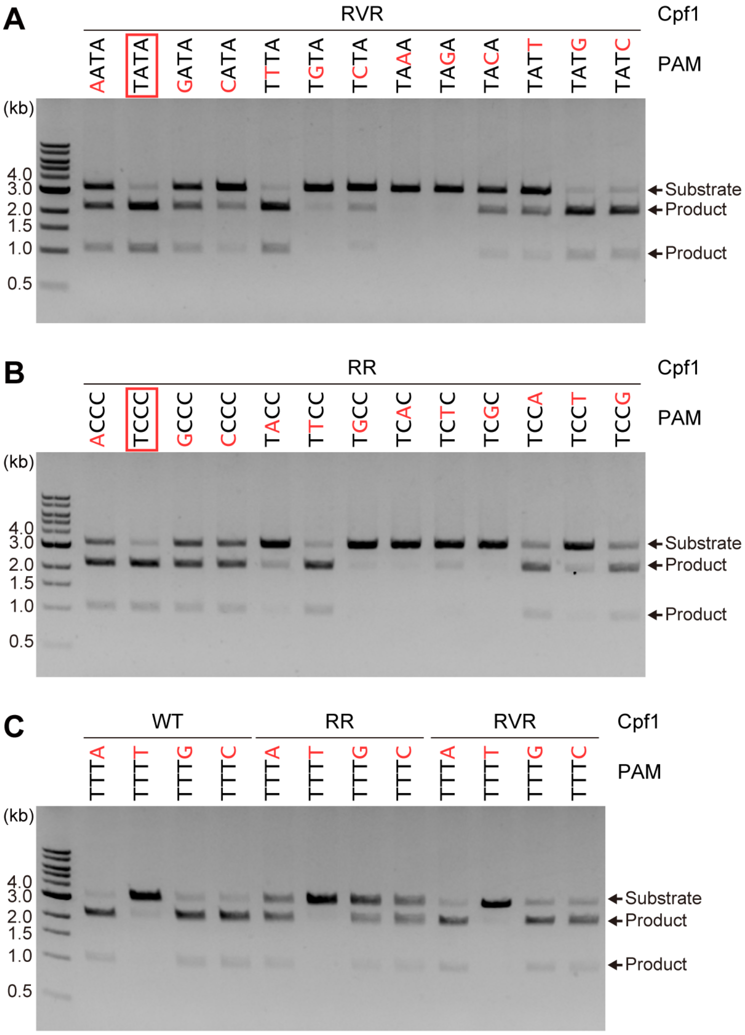


**Figure 3-2. In vitro cleavage activities of WT AsCpf1 and AsCpf1 variants**

(A and B) PAM specificities of the RVR (A) and RR (B) variants. The AsCpf1-crRNA complex (100 nM) was incubated at 37°C for 5 min with a linearized plasmid target with the different PAMs. The favorable PAMs for the RVR (TATA) and RR (TCCC) variants are boxed in red. The substituted nucleotides are colored red.

(C) Fourth PAM nucleotide preferences of WT AsCpf1 and the RVR and RR variants. The AsCpf1-crRNA complex (100 nM) was incubated at 37°C for 5 min with a linearized plasmid target with the TTTN PAMs. The substituted nucleotides are colored red.

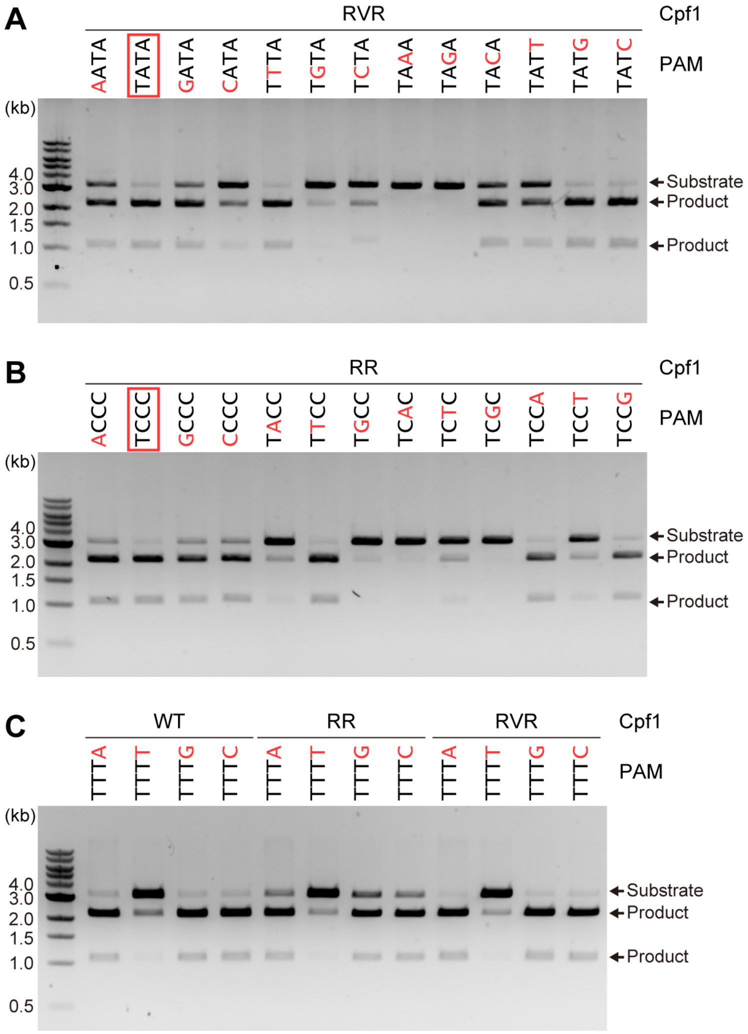


**Figure 3-3. In vitro cleavage activities of WT AsCpf1 and AsCpf1 variants**

(A and B) PAM specificities of the RVR (A) and RR (B) variants. The AsCpf1-crRNA complex (100 nM) was incubated at 37°C for 10 min with a linearized plasmid target with the different PAMs.

(C) Fourth PAM nucleotide preferences of WT AsCpf1 and the RVR and RR variants. The AsCpf1-crRNA complex (100 nM) was incubated at 37°C for 10 min with a linearized plasmid target with the TTTN PAMs.

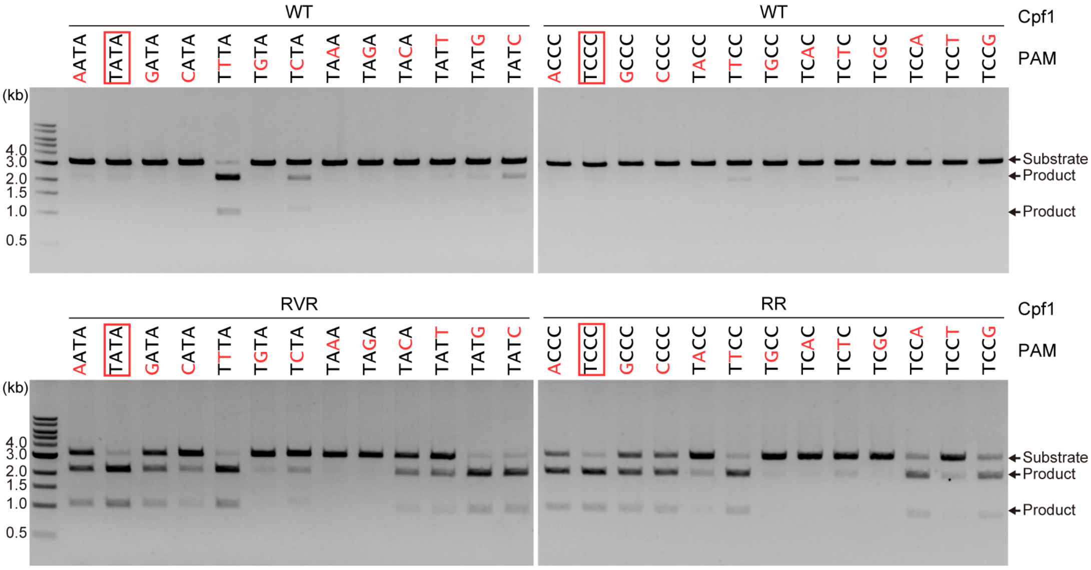


**Figure 3-4. Comparison of the PAM specificities of WT AsCpf1 and AsCpf1 variants**

The AsCpf1-crRNA complex (100 nM) was incubated at 37°C for 5 min with a linearized plasmid target with the different PAMs. For comparison, the cleavage data for the RVR (Figure 3-2A) and RR (Figure 3-2B) variants are shown below those for WT AsCpf1.

### 3.5.3 Crystallization and X-ray diffraction analysis

The RVR and RR variants in complex with a 43-nt crRNA containing 24-nt guide sequence, a 34-nt target DNA strand, and a 10-nt non-target DNA strand was crystallized under the condition A (Figures 3-5 and 3-6). The 2.0 Å resolution data sets were collected from both crystals.

### 3.5.4 Phase determination, model building and structure refinement

Molecular replacement was performed using the crystal structure of AsCpf1–crRNA–target DNA complex as the search model, and thus interpretable electron density maps were obtained. The final structures were refined using the data sets respectively (RVR: 2.0 Å resolution, $R_{work}/R_{free}$ = 0.175/0.210, RR: 2.0 Å resolution, $R_{work}/R_{free}$ = 0.183/0.214) (Figures 3-7 and 3-8). Data collection and refinement statistics are shown in Table 3-1.

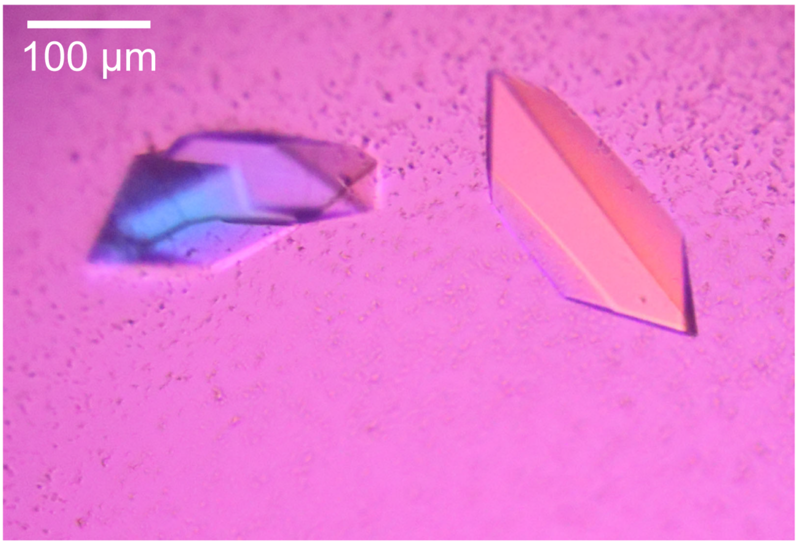


**Figure 3-5. Crystal of AsCpf1 RVR variant**
Crystal of AsCpf1 RVR variant in complex with crRNA and taget DNA.

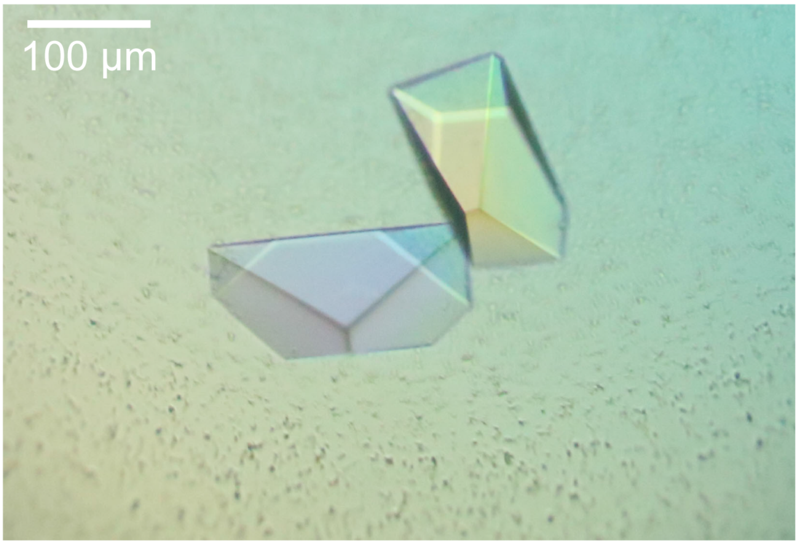


**Figure 3-6. Crystal of AsCpf1 RR variant**
Crystal of AsCpf1 RR variant in complex with crRNA and taget DNA.

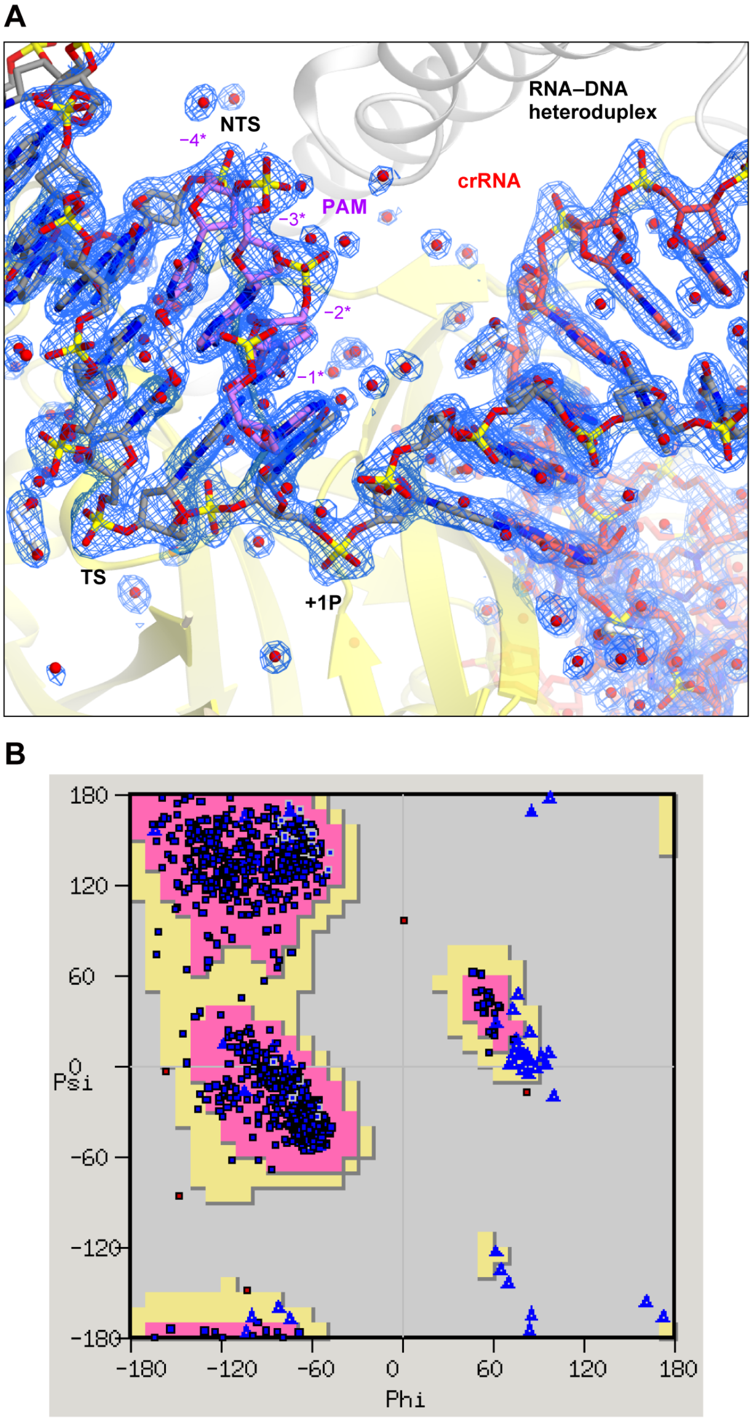


**Figure 3-7. Electron density map of AsCpf1 RVR variant**

(A) The $2mF_o - DF_c$ electron density map (contoured at 1.5 σ) for the bound nucleic acids in RVR-crRNA target DNA complex is shown as a blue mesh. +1P, +1 phosphate.

(B) Ramachandran plots of AsCpf1 RVR variant.

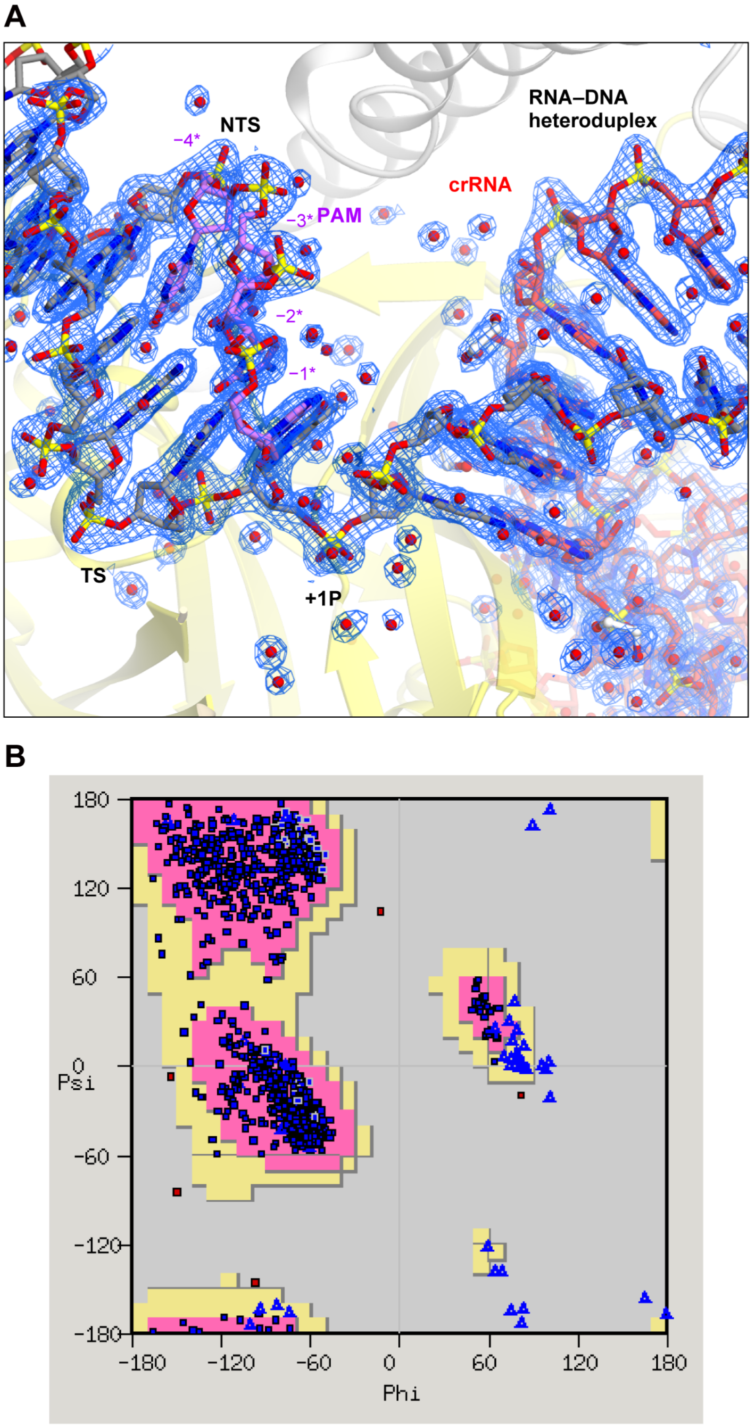


**Figure 3-8. Electron density map of AsCpf1 RR variant**

(A) The $2mF_O - DF_C$ electron density map (contoured at 1.5 σ) for the bound nucleic acids in RR-crRNA target DNA complex is shown as a blue mesh. +1P, +1 phosphate.

(B) Ramachandran plots of AsCpf1 RR variant.

**Table 3-1 Data collection and refinement statistics.**

| | RVR (TATA PAM) | RR (TCCA PAM) |
|---|---|---|
| **Data collection** | | |
| Beamline | SPring-8 BL41XU | SPring-8 BL41XU |
| Wavelength (Å) | 1.0000 | 1.0000 |
| Space group | $P2_12_12_1$ | $P2_12_12_1$ |
| Cell dimensions | | |
| *a*, *b*, *c* (Å) | 81.2, 133.7, 199.6 | 80.7, 133.3, 200.0 |
| *a*, *b*, *g* (°) | 90, 90, 90 | 90, 90, 90 |
| Resolution (Å)* | 40.0–2.0 (2.03–2.00) | 40.0–2.0 (2.03–2.00) |
| $R_{\text{pim}}$ | 0.021 (0.396) | 0.034 (0.466) |
| *I*/s*I* | 17.1 (2.0) | 10.5 (1.5) |
| Completeness (%) | 98.2 (98.3) | 99.9 (99.9) |
| Multiplicity | 6.6 (6.6) | 6.4 (6.4) |
| CC(1/2) | 0.999 (0.813) | 0.999 (0.841) |
| | | |
| **Refinement** | | |
| Resolution (Å) | 40.0–2.0 (2.02–2.00) | 40.0–2.0 (2.02–2.00) |
| No. reflections | 145,851 (4,771) | 145,319 (4,787) |
| $R_{\text{work}}$ / $R_{\text{free}}$ | 0.175 / 0.210 (0.272 / 0.333) | 0183 / 0.214 (0.256 / 0.304) |
| No. atoms | | |
| Protein | 10,455 | 10,431 |
| Nucleic acid | 1,639 | 1,639 |
| Ion | 6 | 6 |
| Solvent | 731 | 747 |
| *B*-factors (Å$^2$) | | |
| Protein | 51.8 | 49.5 |
| Nucleic acid | 50.2 | 48.2 |
| Ion | 50.1 | 45.2 |
| Solvent | 52.3 | 49.6 |
| R.m.s. deviations | | |
| Bond lengths (Å) | 0.007 | 0.006 |
| Bond angles (º) | 0.843 | 0.834 |
| Ramachandran plot (%) | | |
| Favored region | 98.27 | 98.12 |
| Allowed region | 1.65 | 1.80 |
| Outlier region | 0.08 | 0.08 |
| MolProbity score | | |
| Clashscore | 2.68 | 2.48 |
| Rotamer outlier | 2.20 | 2.30 |

*Values in parentheses are for the highest resolution shell.

### 3.5.5 Crystal structures of the AsCpf1 variants

To elucidate the altered PAM recognition mechanisms of the AsCpf1 variants, we determined the crystal structures of (1) the RVR variant bound to the crRNA and its target DNA with the TATA PAM at 2.0 Å resolution, and (2) the RR variant bound to the crRNA and its target DNA with the TCCA PAM at 2.0 Å resolution (Figures 3-9 and Table 3-1). The overall structures of the RVR and RR variants are essentially identical to that of WT AsCpf1 (Yamano et al., 2016) (root-mean-square deviations are 0.53/0.60 Å for the equivalent Cα atoms between WT AsCpf1 and the RVR/RR variants) (Figure 3-9C). The AsCpf1 variants adopt a bilobed architecture consisting of a recognition (REC) lobe and a nuclease (NUC) lobe, in which the crRNA–target DNA heteroduplex is bound to the central channel between the two lobes (Figure 3-9C). In the two structures, the target DNA strand (nucleotides −10 to −1) and the PAM-containing non-target DNA strand (nucleotides −10* to −1*) form the PAM duplex, which is bound to the narrow channel formed by the WED, REC1 and PI domains (Figure 3-9C). The S542R/K548V/N552R and K607R substitutions, identified by random mutagenesis screening for 60 amino acid residues around the PAM duplex, are located in the WED and PI domains, respectively (Figures 3-10). Lys548 and Lys607 are conserved among the Cpf1 family proteins, and participate in the PAM recognition in the WT AsCpf1 structure (Yamano et al., 2016) (Figure 3-10B). In contrast, Ser542 and Asn552 are not well conserved, and do not directly contact the PAM duplex in the WT AsCpf1 structure.

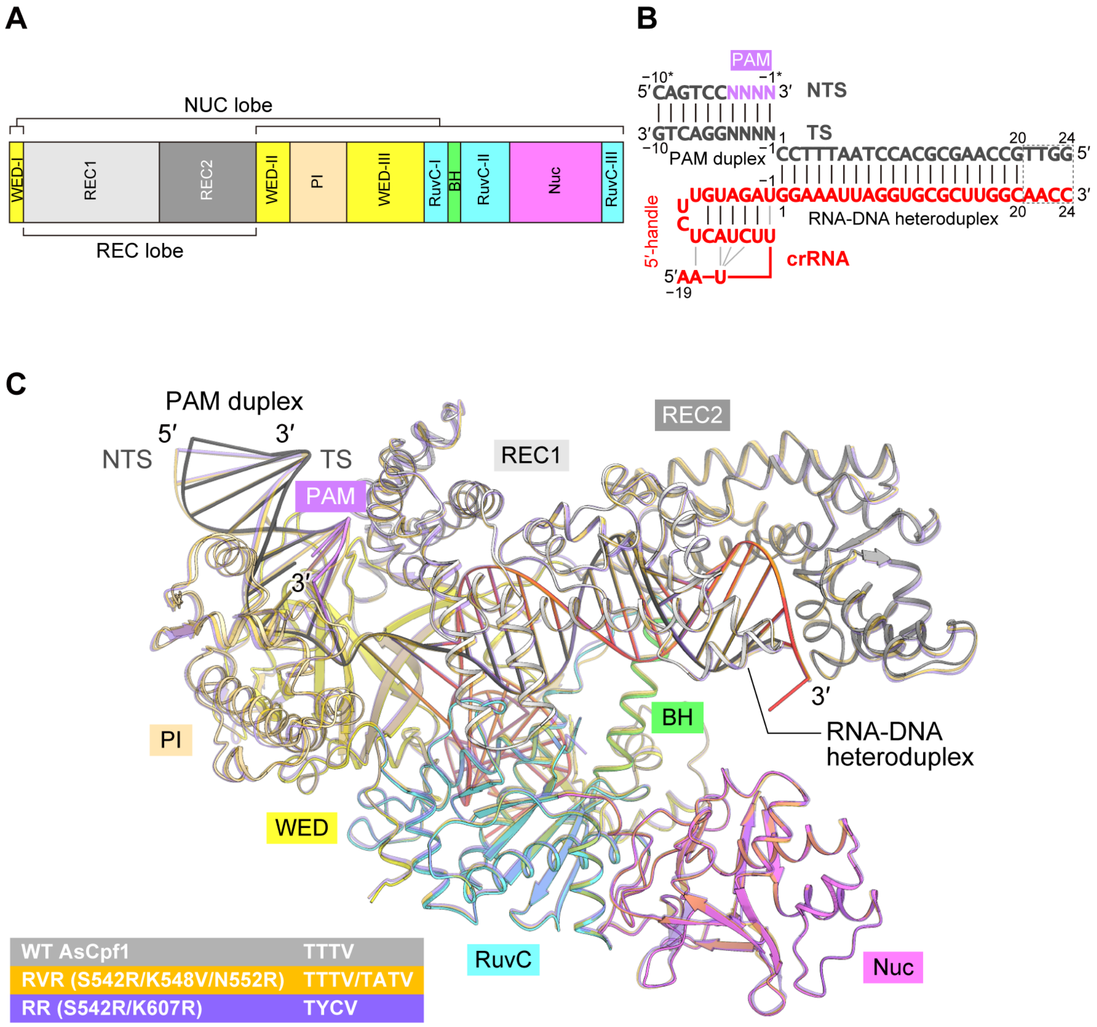
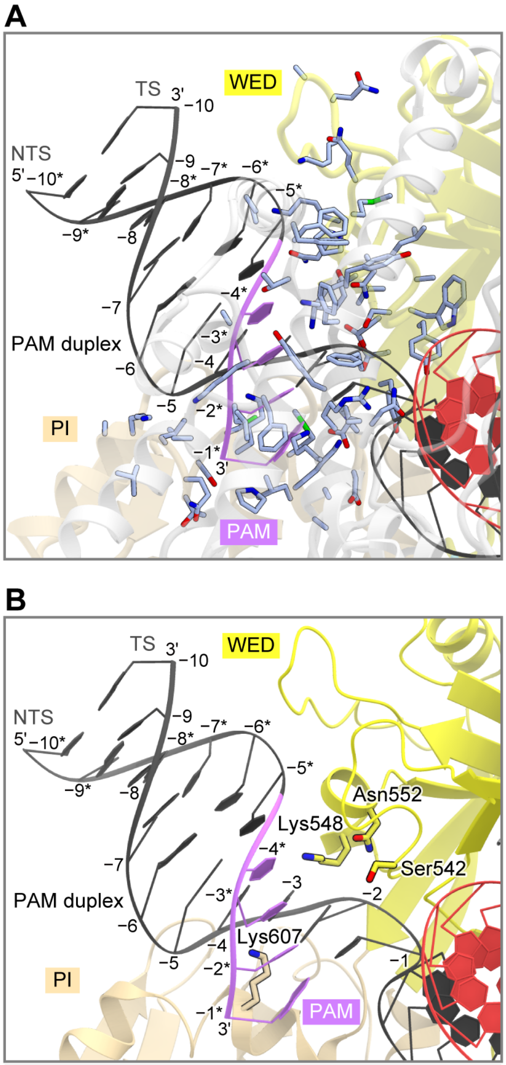


**Figure 3-9. Overall structure of the AsCpf1 variant**

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

(B) Nucleotide sequences of the crRNA and the target DNA. The PAM nucleotides are TATA and TCCA in the RVR and RR variant structures, respectively. The disordered nucleotides in the variant structures are surrounded by dashed lines. TS, target DNA strand; NTS, non-target DNA strand.

(C) Superimposition of the crystal structures of WT AsCpf1 (Yamano et al., 2016) (PDB: 5B43) (colored as in A and B) and the RVR (orange) and RR (purple) variants.



| | |
|---|---|
| WT AsCpf1 | TTTV |
| RVR (S542R/K548V/N552R) | TTTV/TATV |
| RR (S542R/K607R) | TYCV |

**Figure 3-10. Amino acid residues around the PAM duplex of the AsCpf1 variant**
(A) The 60 amino acid residues around the PAM duplex in the WT AsCpf1 structure (Yamano et al., 2016) (PDB: 5B43).
(B) PAM duplex in the WT AsCpf1 structure (Yamano et al., 2016) (PDB: 5B43). The substituted residues are shown as stick models.

### 3.5.6 TTTA PAM recognition by WT AsCpf1

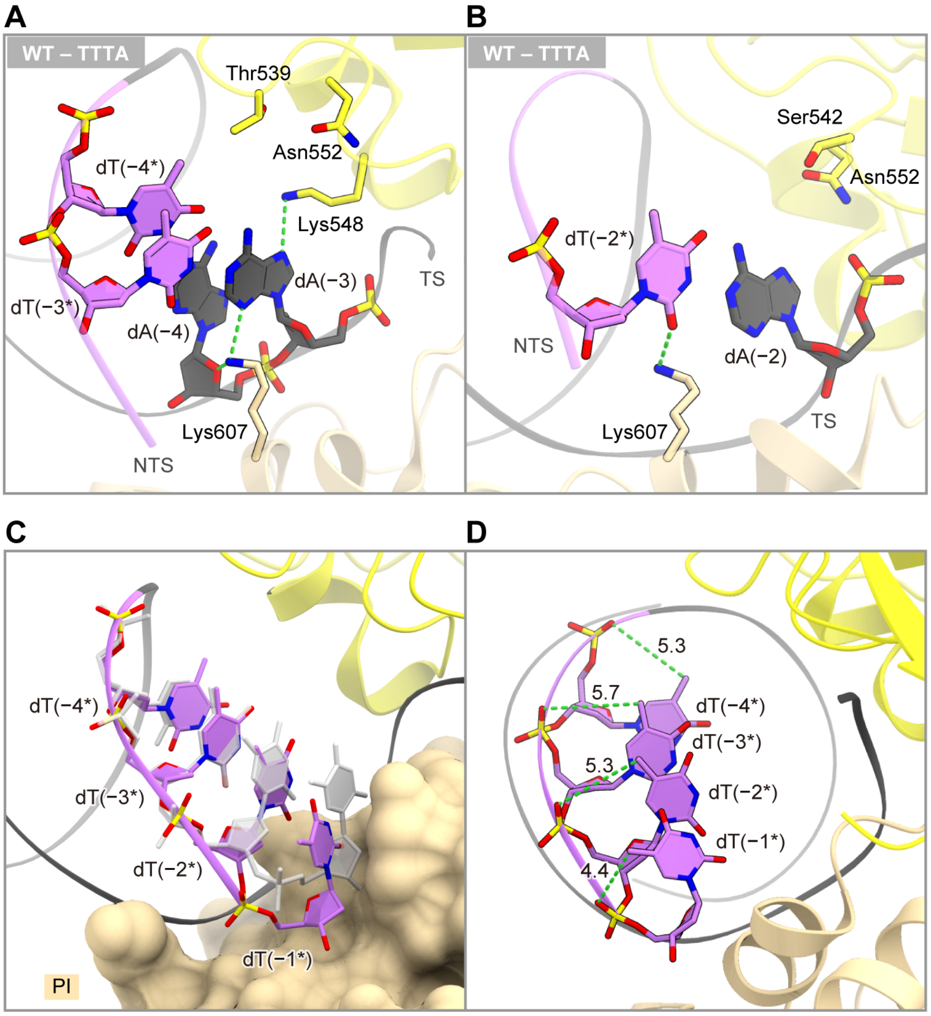
In the WT AsCpf1 structure with the TTTA PAM, the PAM DNA duplex adopts a distorted conformation with a narrow minor groove, and is recognized by the WED, REC1 and PI domains (Yamano et al., 2016) (Figures 3-11A and 3-11B). Notably, the conserved Lys607 residue in the PI domain forms multiple interactions with the PAM duplex from the minor groove side. Lys607 forms hydrogen bonds with the O4′ of dA(−4), the N3 of dA(−3) and the O2 of dT(−2*) (Figures 3-11A and 3-11B). Moreover, Lys548 in the WED domain hydrogen bonds with the N7 of dA(−3) from the major groove side (Figure 3-11A). In the WT AsCpf1 structure, the dT(−1):dA(−1*) base pair does not form base-specific contacts with the AsCpf1 protein (Yamano et al., 2016); nonetheless, our *in vitro* cleavage data and previous studies (Zetsche et al., 2015a; Kim et al., 2016b; Gao et al., 2016a) showed that AsCpf1 prefers the V (V is A, G or C) nucleotides at the fourth PAM position (Figures 3-2C and 3-3C). To clarify structural basis for the fourth V preference, we modeled a T nucleotide at the fourth PAM position (dT(−1*)) in the WT AsCpf1 structure. The modeling indicated that the fourth PAM nucleotide adopts a distinct conformation, due to the interaction with the PI domain (Figure 3-11C), and that the 5-methyl group of dT(−1*) in the non-target strand is located closer (4.4 Å) to the neighboring backbone phosphate group, as compared with those of dT(−2*) (5.3 Å), dT(−3*) (5.7 Å) and dT(−4*) (5.3 Å) (Figure 3-11D). In addition, the modeling indicated that the dA(−1) in the target strand does not form unfavorable interactions with the protein. These observations suggested that AsCpf1 disfavors the fourth T in the PAM, likely due to the relatively shorter distance between its 5-methyl group and the backbone phosphate group.

### 3.5.7 TATA PAM recognition by the RVR variant

In the RVR (S542R/K548V/N552R) variant structure with the TATA PAM, the O4 of dT(−4*) forms a water-mediated hydrogen bond with Arg552 (N552R) (Figure 3-12A), explaining the preference for the first T in the TATV PAM. The N6 and N7 of dA(−3*) are recognized by Thr167 and Thr539 *via* a water-mediated hydrogen-bonding network (Figure 3-12A). Notably, the PAM-complementary dT(−3) is extensively recognized by the protein (Figures 3-12A, 3-12C and 3-12D). The O2 and O4 of dT(−3) form hydrogen bonds with Lys607 and Arg552 (N552R), respectively. In addition, the 5-methyl group of dT(−3) forms a hydrophobic interaction with the side chain of Val548 (K548V) (Figure 3-12C). These structural findings can explain the strong preference of the second A for the TATV PAM recognition by the RVR variant. As in the WT AsCpf1 structure, the O2 of dT(−2*) hydrogen bonds with Lys607 (Figure 3-12D). In addition, the N7 of dA(−2) hydrogen bonds with Arg552 (N552R) (Figures 3-12B and 3-12D). These structural observations are consistent with the preference for the third T in the TATV PAM. In the RVR variant, Arg542 (S542R) does not contact the PAM duplex. Together, these structural findings explain the mechanism of TATV PAM recognition by the RVR variant.

### 3.5.8 TCCA PAM recognition by the RR variant

In the RR (S542R/K607R) variant structure with the TCCA PAM, the O4 of dT(−4*) forms a water-mediated hydrogen bond with Lys548, while the N3 of dA(−4) hydrogen bonds with Arg607 (K607R) (Figure 3-13A). These observations explain the preference of the RR variant for the first T in the TYCV PAM. It is likely that, in WT AsCpf1, Lys548 forms a similar water-mediated interaction with dT(−4*) and contributes to the preference for the first T in the TTTV PAM, although such a water molecule was not resolved in the previous WT AsCpf1 structure at a lower resolution (2.8 Å) (Yamano et al., 2016). dC(−3*) does not directly contact the protein. Instead, the O6 and N7 of dG(−3) hydrogen bond with Lys548, while the N3 of dG(−3) forms a water-mediated hydrogen bond with Arg607 (K607R) (Figure 3-13A). It is likely that the N3 and N7 of the A nucleotide at this position are recognized by Arg607 (K607R) and Lys548, respectively. These observations explain the preference of the second Y for the TYCV PAM recognition by the RR variant. Notably, the O6 and N7 of dG(−2) are recognized by Arg542 (S542R) *via* bidentate hydrogen-bonding interactions, whereas dC(−2*) does not directly contact the protein (Figures 3-13B and 3-13C). These structural findings can explain the strong preference of the third C in the TYCV PAM recognition by the RR variant. Moreover, the side chain of Arg607 (K607R) inserts into the minor groove of the PAM duplex, and interacts with the ribose moieties of dA(−4), dC(−2*) and dA(−1*) (Figure 3-13D). Together, these structural findings explain the mechanism of TYCV PAM recognition by the RR variant.



## Figure 3-11. PAM recognition by WT AsCpf1

(A and B) TTTA PAM recognition by WT AsCpf1 (Yamano et al., 2016) (PDB: 5B43). The interactions with the nucleotides at the first and second PAM positions are shown in (A). The interactions with the nucleotides at the third PAM position are shown in (B).

(C) Conformational differences between the PAM nucleotides. The dT(−1*) nucleotide was modeled into the WT AsCpf1 structure. Superimposition of the nucleotides −5* to −2* (gray) onto the nucleotides −4* to −1* (purple) highlights the displacement of the fourth PAM nucleotide (at −1* position), due to the interaction with the PI domain (shown as a surface representation).

(D) Differences in the distances between the 5-methyl group of the T nucleotide and its adjacent phosphate group at each PAM position. The dT(−1*) nucleotide was modeled into the WT AsCpf1 structure (Yamano et al., 2016) (PDB: 5B43). The distances are given in Å.

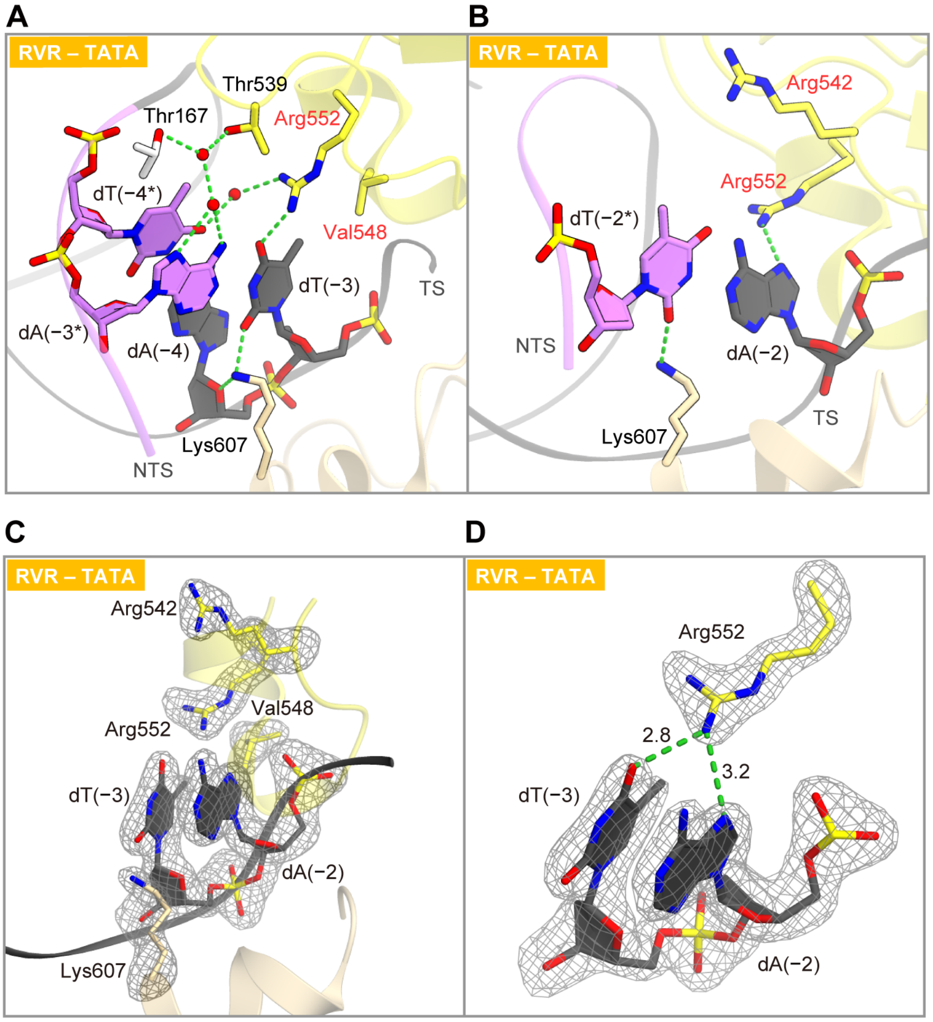


**Figure 3-12. PAM recognition by AsCpf1 RVR variants**

(A and B) TATA PAM recognition by the RVR variant. The interactions with the nucleotides at the first and second PAM positions are shown in (A). The interactions with the nucleotides at the third PAM position are shown in (B). In (A) and (B), the substituted residues are highlighted by red labels. In (A), water molecules are depicted by red spheres.

(C) $mF_O – DF_C$ omit electron density map for the key residues and nucleotides in the RVR variant (contoured at 4σ).

(D) Hydrogen-bonding interactions between Arg552 and the PAM duplex. The $mF_O – DF_C$ omit electron density map is shown as a gray mesh (contoured at 5σ). Hydrogen bonds are shown as dashed lines, and the distances are given in Å.

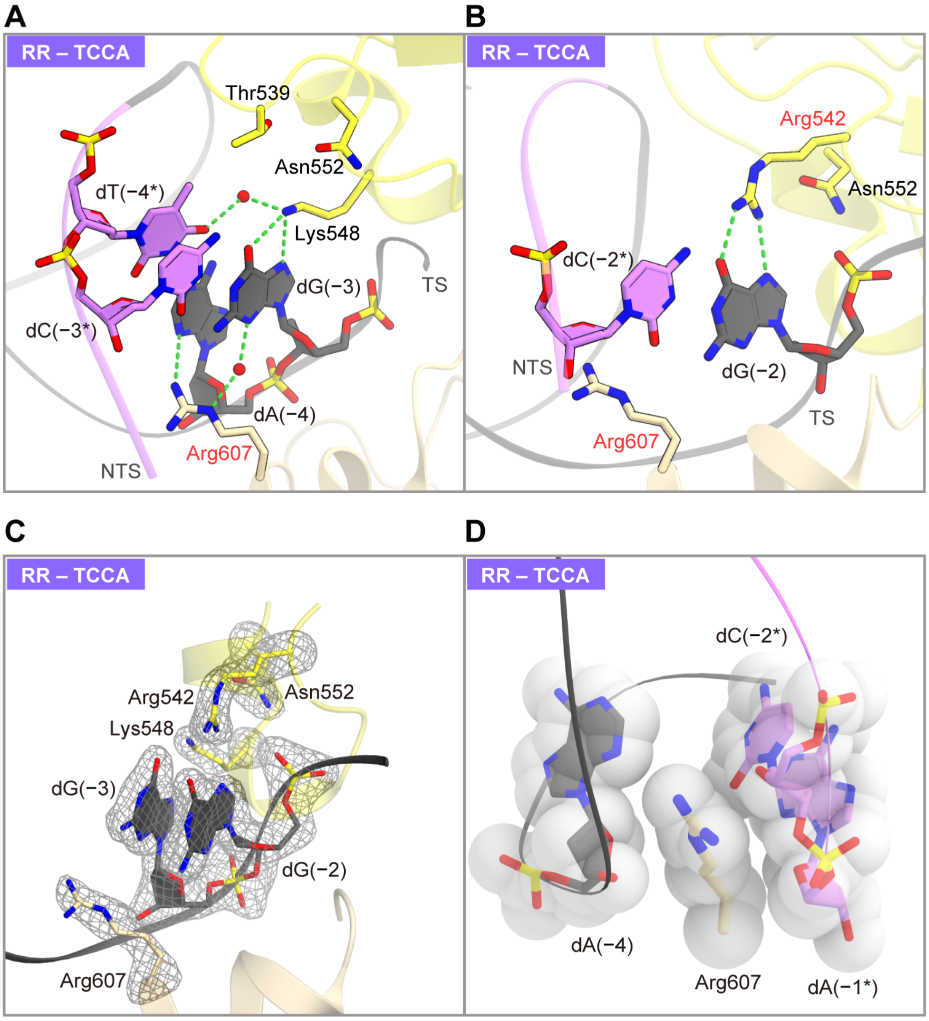


**Figure 3-13. PAM recognition by AsCpf1 RR variants**

(A and B) TCCA PAM recognition by the RR variant. The interactions with the nucleotides at the first and second PAM positions are shown in (A). The interactions with the nucleotides at the third PAM position are shown in (B). In (A) and (B), the substituted residues are highlighted by red labels. In (A), water molecules are depicted by red spheres.

(C) $mF_O - DF_C$ omit electron density map for the key residues and nucleotides in the RR variant (contoured at 4σ).

(D) Hydrophobic interactions between Arg607 and the PAM duplex.

### 3.5.9 Conformational differences in the PAM duplex

A structural comparison between WT AsCpf1 and the variants revealed the conformational differences in their PAM duplexes (Figure 3-14). In the WT AsCpf1 structure with the TTTA PAM, the PAM duplex adopts a distorted conformation characteristic of a T-rich DNA duplex, in which Lys607 forms multiple interactions with the minor-groove edge of the PAM duplex (Yamano et al., 2016) (Figure 3-14). In contrast, in the structures of the RVR (with the TATA PAM) and RR (with the TCCA PAM) variants, the PAM duplexes adopt B-form-like conformations (Figure 3-14), supporting the notion that the distorted conformation of the PAM duplex in the WT AsCpf1 structure is due to the three successive T nucleotides. Unlike the RR variant, the RVR variant efficiently recognizes the TTTV PAM (Figures 3-2C and 3-3C), and the location of the Lys607 residue is similar to that in the WT AsCpf1 structure (Figure 3-14). These observations suggested that the RVR variant recognizes the TTTV PAM in a similar manner to that of WT AsCpf1, and highlighted the importance of Lys607 for the TTTV PAM recognition.

### 3.5.10 Cooperative structural rearrangements induced by the substitutions

A structural comparison between WT AsCpf1 and the two variants also revealed conformational differences in the AsCpf1 proteins. In the structures of the RVR and RR variants, Arg542 (S542R) adopts distinct conformations and plays different functional roles (Figure 3-15A). In the RR variant structure, Arg542 forms bidentate hydrogen bonds with dG(−2) in the target DNA strand, and plays a critical role in the TYCV PAM recognition (Figure 3-15A). In contrast, in the RVR variant structure, Arg542 in the WED domain interacts with Thr167 and Ser170 in the REC1 domain (Figure 3-15A). Our *in vitro* cleavage experiments revealed that the VR (K548V/N552R) variant exhibits reduced activities, as compared with the RVR (S542R/K548V/N552R) variant (Figure 3-16), indicating the functional importance of the Arg542-mediated inter-domain interaction. Given that Arg542 is located far away from the PAM duplex in the RVR structure, it is likely that Arg542 does not interact directly with the PAM duplex and contributes to the structural maintenance of the PAM-duplex channel, thereby enhancing the PAM recognition.

Our high-resolution structures further revealed unexpected conformational rearrangements induced by the N552R substitution in the RVR variant (Figure 3-15B). In the structures of WT AsCpf1 and the RR variant, the side chain of Asn552 hydrogen bonds with the side chain of Thr539 (Figure 3-15B). In the RR variant structure, the side chain of Asn552 also interacts with the backbone phosphate group between dA(−2) and dT(−1) (Figure 3-15B). In contrast, in the RVR variant structure, the side chains of Thr539 and Asn551 adopt distinct conformations, as compared with those in the WT AsCpf1 and RR variant structures, and interact with the side chain of Arg552 (N552R) (Figure 3-15B). Arg552 (N552R) forms a water-mediated interaction with the backbone phosphate group between dA(−2) and dT(−1), while Asn551 interacts with the backbone phosphate group between dC(−6*) and dC(−5*).

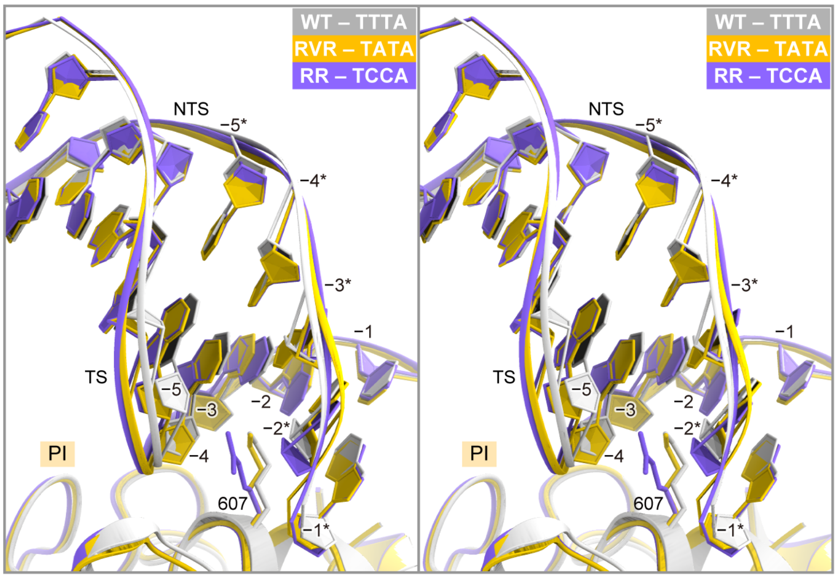


**Figure 3-14. Structural differences in PAM duplex between WT AsCpf1 and AsCpf1 variants**

Conformational differences in the PAM duplexes in the structures of WT AsCpf1 (PDB: 5B43) (stereo view). WT AsCpf1 and the RVR and RR variants are colored gray, orange and purple, respectively.

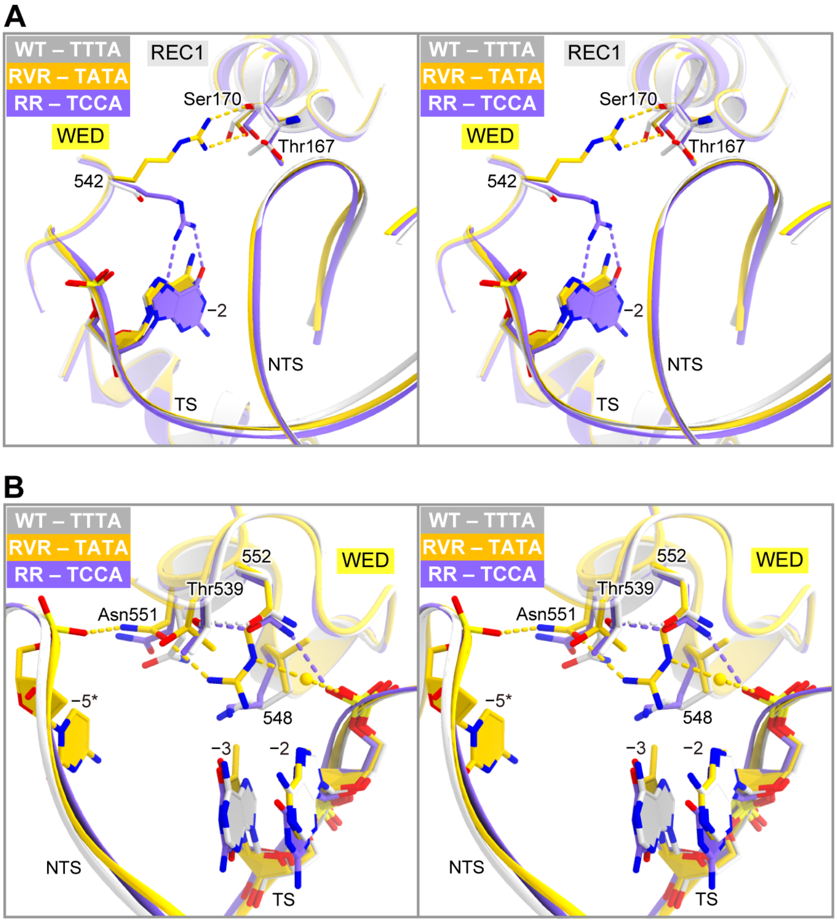
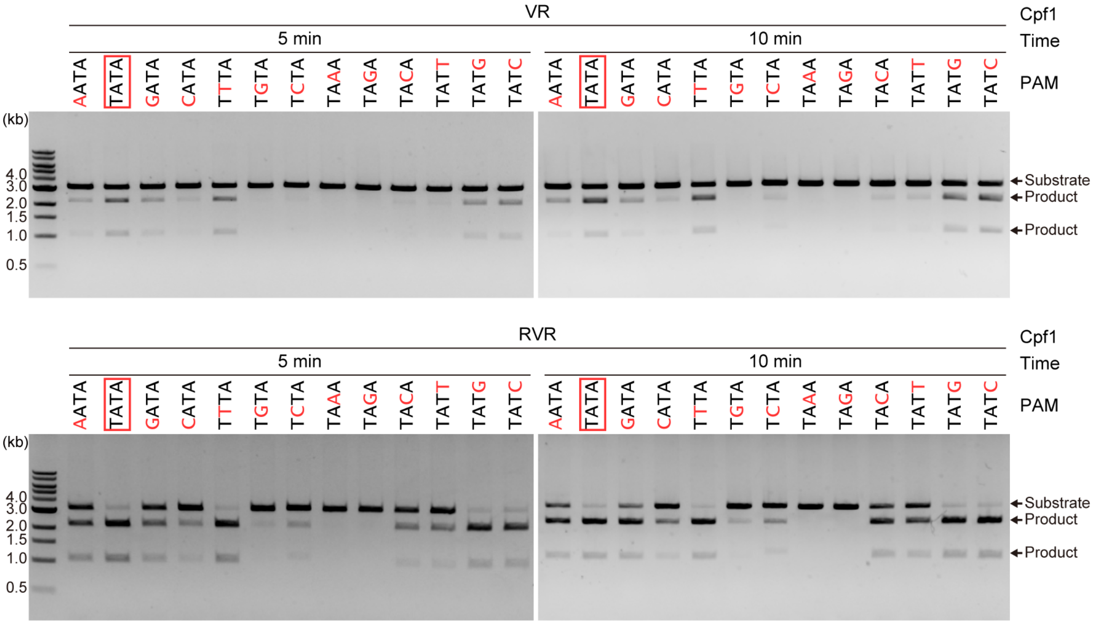


**Figure 3-15. Structural differences between WT AsCpf1 and AsCpf1 variants**

(A) Structural differences in Arg542 (S542R) between the RVR and RR variants (stereo view). (B) Structural rearrangements around Arg552 (N552R) in the RVR variant (stereo view). A water molecule is shown as a sphere. WT AsCpf1 and the RVR and RR variants are colored gray, orange and purple, respectively.



## Figure 3-16. In vitro cleavage activity of the VR variant

The AsCpf1-crRNA complex (100 nM) was incubated at 37°C for 5 or 10 min with a linearized plasmid target with the different PAMs. For comparison, the cleavage data for the RVR variant (Figures 1A and S1A) are shown below those for the VR variant.

## 3.6 Discussion

The present high-resolution structures reveal the altered PAM recognition mechanisms of the RVR and RR variants, and also provide detailed insights into the functional mechanism of WT AsCpf1. WT AsCpf1 recognizes the TTTV PAM mainly *via* multiple interactions between Lys607 and the minor-groove edge of the PAM duplex (Yamano et al., 2016). In contrast, the RVR and RR variants achieve the altered PAM recognition *via* newly formed interactions with the major-groove edges of the PAM-complementary nucleotides in the target strand, rather than the altered PAM nucleotides in the non-target strand. In the RVR variant, Val548 (K548V) and Arg552 (N552R) form base-specific contacts with the T nucleotide complementary to the altered second A in the TATV PAM. In the RR variant, Arg542 (S542R) forms bidentate hydrogen bonds with the G nucleotide complementary to the altered third C nucleotide in the TYCV PAM. This Arg–G interaction is frequently observed in Cas9-mediated PAM recognition, such as those in SpCas9 (Arg1333–G2 and Arg1335–G3 in the NGG PAM) (Anders et al., 2014), SaCas9 (Arg1015–G3 in the NNGRRT PAM) (Nishimasu et al., 2015) and *Francisella novicida* Cas9 (Arg1585–G2 and Arg1556–G3 in the NGG PAM) (Hirano et al., 2016a). In addition, in the RR variant, Arg607 (K607R) donates hydrogen bonds and van der Waals contacts with the PAM duplex, thereby compensating for the loss of the interactions between Lys607 and the PAM duplex observed in WT AsCpf1.

A structural comparison of the AsCpf1 variants with the previously reported SpCas9 variants, such as VQR (D1135V/R1335Q/T1337R) and VRER (D1135V/G1218R/R1335E/T1337R) (Kleinstiver et al., 2015), reveal striking differences in their altered PAM recognition mechanisms. Whereas the third G in the NGG PAM is recognized by Arg1335 in WT SpCas9 (Anders et al., 2014), the third A in the NGA PAM and the third C in the NGCG PAM are recognized by Gln1335 (R1335Q) in the VQR variant and Glu1335 (R1335E) in the VRER variant, respectively (Anders et al., 2016; Hirano et al., 2016b). Thus, the altered PAM recognition by the SpCas9 variants mainly relies on the replacement of the Arg1335–G3 interaction in WT SpCas9 with the altered base-specific interactions (*i.e.*, the Gln1335–A3 interaction in the VQR variant and the Glu1335–C3 interaction in the

VRER variant). In contrast, the altered PAM recognition by the AsCpf1 variants relies on newly formed interactions between the substituted residues and the altered PAM-complementary nucleotides (*i.e.*, Val548/Arg552–A2-complementary T2 in the RVR variant and Arg542–C3-complementary G3 in the RR variant). These differences are reflected by the distinct PAM recognition mechanisms of SpCas9 (base readout from the major-groove side) (Anders et al., 2014) and AsCpf1 (base and shape readout from the minor- and major-groove sides) (Yamano et al., 2016).

The present structures reveals that Arg542 (S542R) plays distinct roles in the RVR and RR variants. Arg542 forms the inter-domain interactions and may reinforce the PAM-duplex binding channel in the RVR variant, whereas Arg542 forms the base-specific contacts with the PAM duplex in the RR variant. These observations demonstrate that the amino-acid substitutions that do not provide interactions with the PAM duplex can contribute to the engineering of the Cpf1's PAM specificity. This contrasts with the altered PAM recognition by the SpCas9 variants, in which the substituted residues provide new contacts with the PAM duplex. These differences also highlight the mechanistic differences in the PAM recognition between SpCas9 (*via* the PAM-binding groove within the PI domain) (Anders et al., 2014) and AsCpf1 (*via* the PAM-binding channel formed by the WED, REC1 and PI domains) (Yamano et al., 2016). Furthermore, the present findings provide important clues for the Cpf1 engineering, and suggest that amino-acid substitutions that reinforce the PAM-binding channel could contribute to the alteration of the Cpf1's PAM specificity.

There are also mechanistic similarities in the altered PAM recognition by the SpCas9 and AsCpf1 variants. In the SpCas9 and AsCpf1 variants, unexpected structural rearrangements play important roles in the altered PAM recognition, thus highlighting the power of structure-guided random mutagenesis approaches. In the SpCas9 variant structures, the direct hydrogen-bonding interactions between the altered third PAM nucleotides and the substituted residues (Gln1335 and Glu1335) are enabled by the unexpected displacement of the PAM duplex, which is cooperatively induced by the other substitutions (D1135V and T1337R) (Anders et al., 2016; Hirano et al., 2016b). In the AsCpf1 RR variant structure, the PAM duplex undergoes a conformational change, partly due to the replacement of Lys607 (K607R). Moreover, in the AsCpf1 RVR variant structure, the N552R substitution induces local conformational changes in Thr539 and Asn551, thus rearranging the interactions with the PAM duplex. These cooperative structural rearrangements are not readily predictable from the WT AsCpf1 structure (Yamano et al., 2016), and thus confirm the power of the combination of structural information and molecular evolution for the engineering of the CRISPR-Cas nucleases.

In summary, our structural studies reveal the altered PAM recognition mechanisms of the recently engineered AsCpf1 variants. Furthermore, the structural comparison between the AsCpf1 and SpCas9 variants enhance our understanding of the PAM recognition mechanisms of class 2 CRISPR-Cas nucleases, and provide a framework for the future engineering of the CRISPR-Cpf1 toolbox.

# Chapter 4: Structural basis for the canonical and non-canonical PAM recognition by CRISPR-Cpf1

## 4.1 Summary

The RNA-guided Cpf1 (also known as Cas12a) nuclease associates with a CRISPR RNA (crRNA), and cleaves the double-stranded DNA target complementary to the crRNA guide. The two Cpf1 orthologs from *Acidaminococcus sp.* (AsCpf1) and *Lachnospiraceae bacterium* (LbCpf1) have been harnessed for eukaryotic genome editing. Cpf1 requires a specific nucleotide sequence, called a protospacer adjacent motif (PAM), for the target recognition. Besides the canonical TTTV PAM, Cpf1 recognizes suboptimal C-containing PAMs. Here, we report four crystal structures of LbCpf1 in complex with the crRNA and its target DNA, containing either TTTA, TCTA, TCCA or CCCA as the PAM. These structures revealed that, depending on the PAM sequences, LbCpf1 undergoes conformational changes to form altered interactions with the PAM-containing DNA duplexes, thereby achieving the relaxed PAM recognition. Collectively, the present structures improve our mechanistic understanding of the PAM-dependent crRNA-guided DNA cleavage by the Cpf1 family nucleases.

## 4.2 Introduction

AsCpf1 and Lbpf1 were identified and biochemically characterized, and they have been harnessed for genome editing in eukaryotic cells (Hur et al., 2016; Kim et al., 2016a, 2016b, 2016c; Kleinstiver et al., 2016b; Tang et al., 2017; Zetsche et al., 2015a, 2016). AsCpf1 and LbCpf1 share 34% sequence identity, while they lack similarity with Cas9 outside their RuvC domains. Previous structural studies provided mechanistic insights into the crRNA–guided DNA recognition and cleavage by the Cpf1 family nucleases (Dong et al., 2016; Yamano et al., 2016; Gao et al., 2016). The crystal structures of the LbCpf1–crRNA binary complex (Dong et al., 2016) and the AsCpf1–crRNA–target DNA ternary complex (Yamano et al., 2016; Gao et al., 2016) revealed the bilobed architectures of Cpf1 and the crRNA recognition mechanism. The AsCpf1–crRNA–DNA structures further revealed the crRNA–guided DNA targeting and the PAM recognition mechanisms (Yamano et al., 2016; Gao et al., 2016). In addition, these structures identified the Nuc domain next to the RuvC nuclease domain, and biochemical data demonstrated that the Nuc domain is involved in the cleavage of the target DNA strand (Yamano et al., 2016). Moreover, a structural comparison between apo-LbCpf1 and the LbCpf1–crRNA complex indicated the crRNA-induced structural rearrangements in Cpf1 (Dong et al., 2016), and a structural comparison between the AsCpf1–crRNA–DNA ternary complex (Yamano et al., 2016; Gao et al., 2016b) and the LbCpf1–crRNA binary complex (Dong et al., 2016) indicated a structural rearrangement accompanying the crRNA–target DNA heteroduplex formation.

Previous *in vitro* cleavage experiments suggested that, whereas LbCpf1 and AsCpf1 prefer the TTTV PAM, they also recognize C-containing sequences as suboptimal PAMs (Zetsche et al., 2015a). Indeed, a recent study demonstrated that LbCpf1 and AsCpf1 can modify target sites with the non-canonical C-containing PAMs, such as CTTA, TCTA and TTCA, in mammalian cells, albeit with lower efficiencies than those with the canonical TTTV PAM (Kim et al., 2016b). However, the mechanism by which Cpf1 recognizes both the canonical and non-canonical PAMs has remained elusive. Moreover, the mechanism of PAM recognition by LbCpf1 also remains unknown, due to the lack of structural information about the LbCpf1–crRNA–DNA complex.

## 4.3 Research aims

In this study, we present that AsCpf1 and LbCpf1 can recognize both the canonical and non-canonical PAMs *in vitro* and *in vivo*. To explain the PAM preferences of Cpf1, we determined four crystal structures of LbCpf1 in complex with the crRNA and its target DNA, containing either TTTA, TCTA, TCCA or CCCA as the PAM, at 2.4–2.5 Å resolutions. These structures revealed that LbCpf1 undergoes conformational changes to form distinct interactions with the PAM-containing DNA duplex, depending on the PAM sequences. A structural comparison of the LbCpf1–crRNA–DNA ternary complex with the LbCpf1–crRNA binary complex revealed the conformational rearrangements in the protein upon the crRNA–target DNA heteroduplex formation. In addition, a structural comparison between LbCpf1 and AsCpf1 revealed similarities and differences in their crRNA–guided DNA cleavage mechanisms. Furthermore, a structural comparison of Cpf1 with Cas9 highlighted the fundamental differences in the PAM recognition mechanisms of the two class 2 effector nucleases.

## 4.4 Materials and methods

### 4.4.1 Sample preparation

The sample preparation was performed as previously described (Yamano et al., 2016), with minor modifications. The gene encoding full-length LbCpf1 (residues 1–1,226) was PCR-amplified using the pcDNA3.1-LbCpf1 plasmid (Zetsche et al. 2015) as the template, and cloned between the *Nde*I and *Xho*I sites of the modified pE-SUMO vector (LifeSensors). The mutations (D832A, E925A, D1180A and R1138A) were introduced by a PCR-based method. The plasmid DNAs were amplified in *Escherichia coli* Mach (Thermo Fisher Scientific), cultured in LB medium (Nacalai Tesque) at 37°C overnight.

The LbCpf1-expressing *E. coli* Rosetta2 (DE3) cells were cultured at 37°C in LB medium (containing 20 mg/l kanamycin) until the $OD_{600}$ reached 0.8, and protein expression was then induced by the addition of 0.1 mM isopropyl-β-D-thiogalactopyranoside (IPTG) (Nacalai Tesque). The *E. coli* cells were further cultured at 20°C for 18 h, and harvested by centrifugation at 5,000 g for 10 min.

The *E. coli* cells were resuspended in buffer A (50 mM Tris-HCl, pH 8.0, 20 mM imidazole, 300 mM NaCl and 3 mM 2-mercaptoethanol), lysed by sonication, and then centrifuged at 40,000 g for 30 min. The supernatant was mixed with 5 ml Ni-NTA Superflow (QIAGEN) equilibrated with buffer A for about 1 hour at 4°C, and the mixture was loaded into an Econo-Column (Bio-Rad). The resin was washed with 5 column volumes of buffer A, 5 column volumes of buffer B (50 mM Tris-HCl, pH 8.0, 20 mM imidazole, 1 M NaCl and 3 mM 2-mercaptoethanol) and 3 column volumes of buffer A. The protein was eluted with buffer C (50 mM Tris-HCl, pH 8.0, 300 mM imidazole, 300 mM NaCl and 3 mM 2-mercaptoethanol). The protein was loaded onto a HiTrap SP HP column (GE Healthcare) equilibrated with buffer D (20 mM Tris-HCl, pH 8.0 and 200 mM NaCl). The column was washed with buffer D, and the protein was then eluted with a linear gradient of 200–1,000 mM NaCl. To remove the His$_6$-SUMO-tag, the eluted protein was mixed with TEV protease (home made), and was dialyzed at 4°C for 12 h against buffer E (20 mM Tris-HCl, pH 8.0, 40 mM imidazole, 300 mM NaCl and 3 mM 2-mercaptoethanol). The protein was passed through the Ni-NTA column equilibrated with buffer E. The protein was concentrated using an Amicon Ultra 10K filter (Millipore), and was further purified by a HiLoad Superdex 200 16/60 column (GE Healthcare) equilibrated with buffer F (10 mM Tris-HCl, pH 8.0, 150 mM NaCl and 1 mM DTT). The purified LbCpf1 proteins were stored at −80°C until use. The crRNA and target DNA were purchased from Gene Design and Sigma-Aldrich, respectively. The purified LbCpf1 protein was mixed with the crRNA, the target DNA strand and the non-target DNA strand (molar ratio, 1:1.5:2.3:2.3), and the reconstituted complex was concentrated using an Amicon Ultra 10K filter. The LbCpf1–crRNA–target DNA complex was purified by gel filtration chromatography on a Superdex 200 Increase column (GE Healthcare) equilibrated with buffer F.

### 4.4.2 *In vitro* cleavage assay

*In vitro* cleavage experiments were performed as previously described (Nishimasu et al., 2015), with minor modifications. The *Eco*RI-linearized pUC119 (100 ng, 4.7 nM), containing the 24-nt target sequence and the PAMs, was incubated with the Cpf1–crRNA complex (50 nM) at 37°C for 5 min, in 20 μl of reaction buffer containing 20 mM HEPES-NaOH, pH 7.5, 100 mM KCl, 2 mM $MgCl_2$, 1 mM DTT and 5% glycerol. The reaction was stopped by the addition of a solution containing EDTA (40 mM final concentration) and Proteinase K (10 μg). Reaction products were resolved on an ethidium bromide-stained 1% agarose gel, and then visualized using an Amersham Imager 600 (GE Healthcare). To examine the time course of DNA cleavage, the *Eco*RI-linearized pUC119 target (600 ng, 4.7 nM) was incubated with the Cpf1–crRNA complex (50 nM) at 37°C in 60 μl reaction buffer. Aliquots (10 μl) were taken at the indicated time points, and the reaction products were then analyzed as described above. To examine whether Cpf1 serves as a nickase, the circular pUC119 target (100 ng, 4.7 nM) was incubated with the Cpf1–crRNA complex (100 nM) at 37°C for 20 min in 10 μl reaction buffer, and the reaction products were then analyzed as described above. *In vitro* cleavage experiments were performed at least three times, and representative results are shown.

### 4.4.3 *In vivo* cleavage assay

This experiment was performed partly in collaboration with Dr. Feng Zhang (MIT). Human embryonic kidney 293T (HEK) cells were maintained in Dulbecco's modified Eagle medium (Gibco), supplemented with 10% fetal bovine serum (FBS), at 37°C under a 5% $CO_2$ atmosphere. HEK cells were seeded at $1.25 \times 10^5$ cells per well in 24-well plates, 24 h prior to transfection. Plasmids encoding humanized LbCpf1 (pY027) or AsCpf1 (pY026) with C-terminal nuclear localization tags and U6-driven crRNAs were transfected at 500 ng per well, using the Lipofectamine 2000 reagent (Life Technologies). Genomic DNA was extracted using 100 μl QuickExtract DNA Extraction Solution (Epicenter), 3 days post-transfection. Insertion/deletion events (indels) were analyzed by a Surveyor nuclease assay, as previously described (Ran et al., 2013). Briefly, the genomic regions flanking the target sites for *DNMT1* or *EMX1* were PCR-amplified, and the products were purified using a QIAQuick PCR purification Kit (QIAGEN) according to the manufacturer's protocol. The purified PCR products (200 ng) were mixed with 1 μl 10 × 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, the products were treated with Surveyor nuclease and Surveyor enhancer

S (IDT), according to the manufacturer's recommended protocol, and analyzed on 10% Novex TBE polyacrylamide gels (Life Technologies). The 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 the relative band intensities. The indel percentage was determined by the formula, $100 \times (1 - (1 - (\boldsymbol{b} + \boldsymbol{c})/(\boldsymbol{a} + \boldsymbol{b} + \boldsymbol{c}))^{1/2})$, where $\boldsymbol{a}$ is the integrated intensity of the undigested PCR product, and $\boldsymbol{b}$ and $\boldsymbol{c}$ are the integrated intensities of each cleavage product. *In vivo* cleavage experiments were performed at three times, and data are shown as mean ± s.e.m ($n$ = 3).

### 4.4.4 Crystallization

Two crRNAs with different length of the guide sequences (20-nt or 24-nt) were designed based on the previous study. The crRNAs were purchased from Gene Design. PAM containing target DNAs were designed to partially form the PAM duplex. The length of PAM duplex is predicted to affect the crystallization of the complex, so five target DNAs with different length of the duplex were prepared (7-bp, 8-bp, 9-bp, 10-bp, or 11-bp). The target and non-target DNA strands were purchased from Sigma-Aldrich. The peak fractions of the purified LbCpf1–crRNA–target DNA complex were concentrated by Amicon Ultra 10K filter. Using the concentrated complex solution ($A_{260\ \mathrm{nm}}$ = 10), the initial crystallization screening was performed at 20°C by the sitting-drop vapor diffusion method.

The screening kits used for the initial screening

Crystal Screen, PEG/Ion (Hampton Research)
JBScreen Classic 1, 2, 4, 5 (Jena Bioscience)
MemGold, MemGold2, JCSG-*plus*, PACT *premier* (Molecular Dimensions)
Wizard Classic 1 and 2 (Emerald Biosystems)

For crystallization optimization, Additive Screen (Hampton Research) was added to the reservoir solutions, in addition to the optimization of the concentration of precipitations and salt, and pH of the buffers. After the optimization by the sitting-drop vapor diffusion method, further optimization was performed by the hanging-drop vapor diffusion method. The crystallization drops were formed by mixing 1 μl of complex solution ($A_{260\,\mathrm{nm}}$ = 10) and 1 μl of reservoir solution (13–18% PEG3,350 and 100 mM MIB buffer, pH 5.0), and then were incubated against 0.5 ml of reservoir solution. The crystals were improved by micro-seeding using Seed Bead (Hampton Research). In the sitting-drop vapor diffusion method, the crystallization conditions were prepared using PLATEMASTER P220 (GILSON) and Mosquito crystallization robot (TTP Labtech).

### 4.4.5 X-ray diffraction analysis and data processing

The crystals of the TCTA, TCCA and CCCA complexes were cryoprotected in a solution consisting of 16–20% PEG3,350, 100 mM MIB buffer, pH 5.0, and 30% ethylene glycol. The crystals of the TTTA PAM complex were cryoprotected in the solution supplemented with 50 mM $MgCl_2$. To reduce radiation damage, all X-ray-diffraction data were collected at 100 K on the beamline BL41XU at SPring-8, and PXI at the Swiss Light Source. A diffraction data set was collected from a single crystal using a X-ray beam at a wavelength of 1.000 Å, an oscillation range of 120° (0.1° per image), an exposure time of 1.0 s per image. The diffraction data were processed using DIALS (Waterman et al., 2013) and AIMLESS (Evans and Murshudov, 2013).

### 4.4.6 Phase determination, model building and structure refinement

The TTTA complex structure was determined by molecular replacement with Molrep (Vagin and Teplyakov, 2010), using the coordinates of LbCpf1 (PDB: 5ID6) (Dong et al., 2016) as the search model. The TCTA, TCCA and CCCA complex structures were determined by molecular replacement, using the TTTA complex structure as the search model. The model building and structural refinement were performed using COOT (Emsley and Cowtan, 2004) and PHENIX (Adams et al., 2010), respectively. Structural figures were prepared using CueMol (http://www.cuemol.org).

## 4.5 Results

### 4.5.1 Sample preparation

WT LbCpf1 protein was stable and the expression level was high. Furthermore, WT LbCpf1 was highly purified by the combination of the Ni-NTA chromatography, the cation exchange chromatography, and the gel filtration chromatography. The gel filtration chromatogram peak showed monodispersity (Figures 4-1). The final yield of the WT LbCpf1 was 14.2 mg per 1 L culture medium.

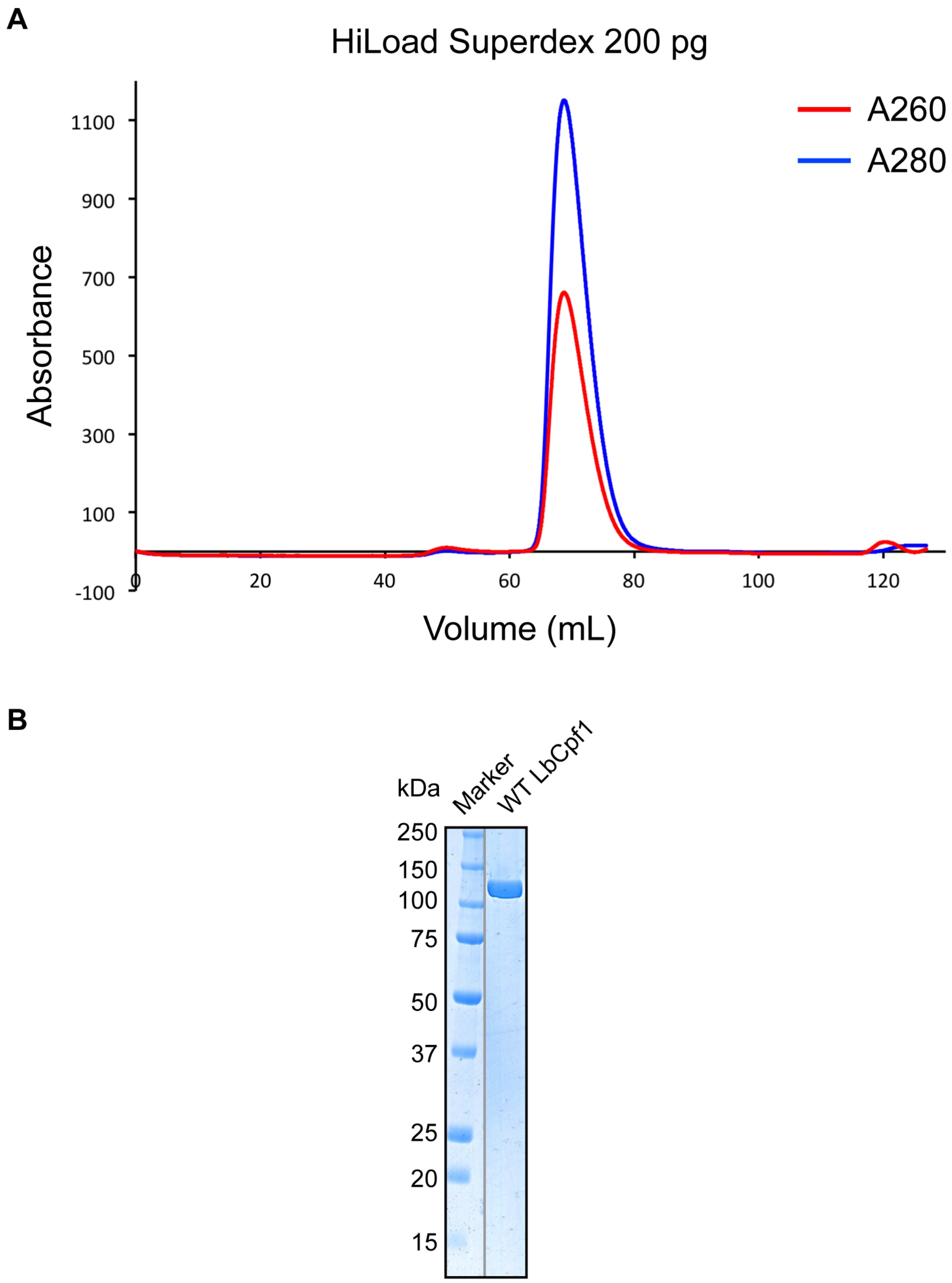


**Figure 4-1. Size-exclusion chromatogram and SDS-PAGE**

(A) Chromatogram of full-length LbCpf1.

(B) SDS-PAGE analysis with SimplyBlue SafeStain. Left lane, molecular-weight marker (labeled in kDa); right lane, WT LbCpf1.

### 4.5.2 DNA cleavage activities of LbCpf1 and AsCpf1

The *in vitro* DNA cleavage activities of purified LbCp1 and AsCpf1 have not been fully investigated, although their activities were compared in mammalian cells (Kim et al., 2016a; Kleinstiver et al., 2016b) and plant cells (Tang et al., 2017). We thus measured the *in vitro* DNA cleavage activities of purified LbCpf1 and AsCpf1, using a plasmid DNA substrate with a 24-nt target sequence and the TTTA PAM (Figure 4-2A). LbCpf1 cleaved the plasmid target slightly more efficiently than AsCpf1 (Figure 4-2A). Previous studies indicated that, while LbCpf1 and AsCpf1 recognize TTTV (V is A, G or C) as the optimal PAM, they also recognize C-containing sequences, such as CTTV, TCTV and TTCV, as suboptimal PAMs (Kim et al., 2016b). To examine their preference for the fourth PAM nucleotide, we measured the cleavage activities of LbCpf1 and AsCpf1 toward four target sites, with either TTTA, TTTT, TTTG or TTTC as the potential PAM (Figure 4-2B). LbCpf1 and AsCpf1 efficiently cleaved the TTTA, TTTG and TTTC sites, but not the TTTT site (Figure 4-2B), confirming their preferences for the fourth V in the TTTV PAM. To further explore their PAM specificities, we measured their cleavage activities toward seven target sites with either TTTA, CTTA, TCTA, TTCA, CCTA, TCCA or CCCA as the potential PAM (Figure 4-2C). LbCpf1 and AsCpf1 cleaved the target sites with either CTTA, TCTA or TTCA as the PAM, albeit with lower efficiencies than the TTTA site (Figure 4-2C). In contrast, LbCpf1 and AsCpf1 were less active toward the CCTA, TCCA and CCCA sites (Figure 4-2C). These results confirmed that LbCpf1 and AsCpf1 recognize CTTV, TCTV and TTCV as the suboptimal non-canonical PAMs.

Furthermore, we examined the activities of LbCpf1 and AsCpf1 toward 42 endogenous target sites, with either TTTV, CTTV, TCTV, TTCV, CCTV, TCCV or CCCV as the potential PAM, in the *DNMT1* and *EMX1* loci in HEK293 cells (Figures 4-3). LbCpf1 and AsCpf1 efficiently modified all six of the target sites with the canonical TTTV PAM (Figures 4-3). In contrast, LbCpf1/AsCpf1 modified the 4/3 CTTV, 3/4 TCTV, 4/3 TTCV, 2/2CCTV, 3/2 TCCV and 1/0 CCCV sites, respectively (Figures 4-3). These results revealed that LbCpf1 and AsCpf1 can edit the endogenous sites with the CTTV, TCTV or TTCV PAM, consistent with our *in vitro* cleavage data and a recent *in vivo* study (Kim et al., 2016b). Together, these results confirmed that, in addition to the canonical TTTV PAM, LbCpf1 and AsCpf1 can target CTTV, TCTV and TTCV as suboptimal non-canonical PAMs.

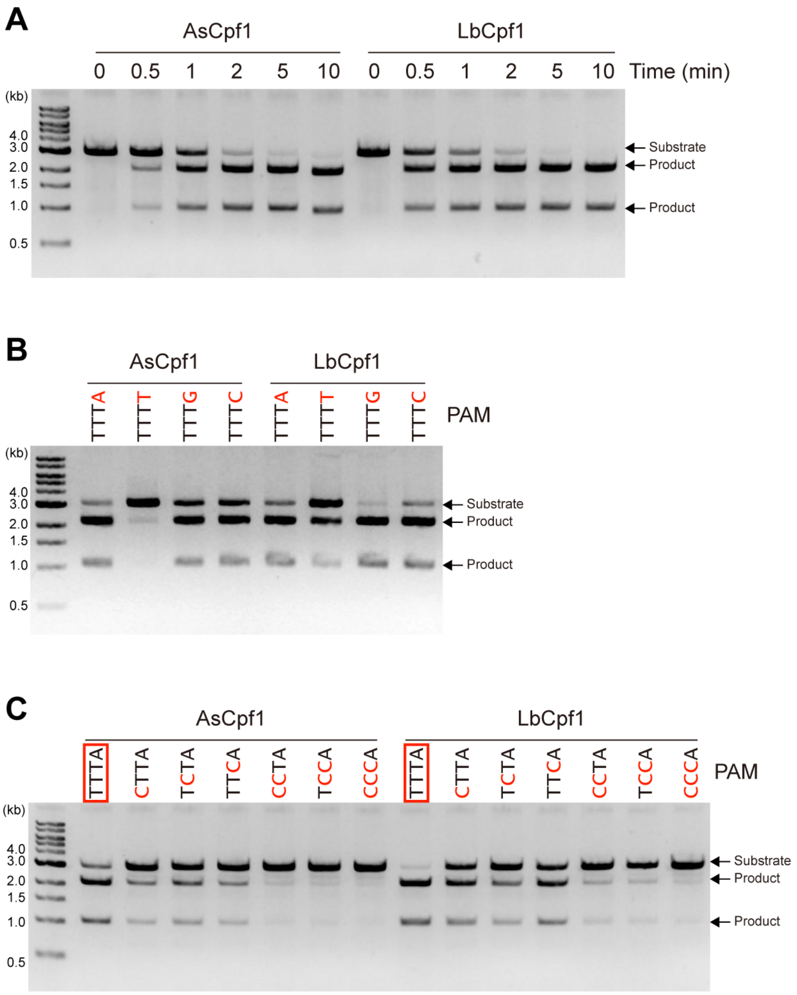


**Figure 4-2. In vitro DNA cleavage activities of LbCpf1 and AsCpf1**

(A) In vitro DNA cleavage activities of LbCpf1 and AsCpf1. The Cpf1–crRNA complex (50 nM) was incubated at 37°C for the indicated time with a linearized plasmid target with the TTTA PAM.

(B) Fourth PAM nucleotide preferences of LbCpf1 and AsCpf1. The Cpf1–crRNA complex (50 nM) was incubated at 37°C for 5 min with a linearized plasmid target with the different PAMs.

(C) PAM nucleotide preferences of LbCpf1 and AsCpf1 at the first second, and third positions. The Cpf1–crRNA complex (50 nM) was incubated at 37°C for 5 min with a linearized plasmid target with the different PAMs.

Data are shown as mean ± s.e.m (n = 3).

In (B) and (C), the canonical PAM is boxed in red, and the substituted nucleotides are colored red.

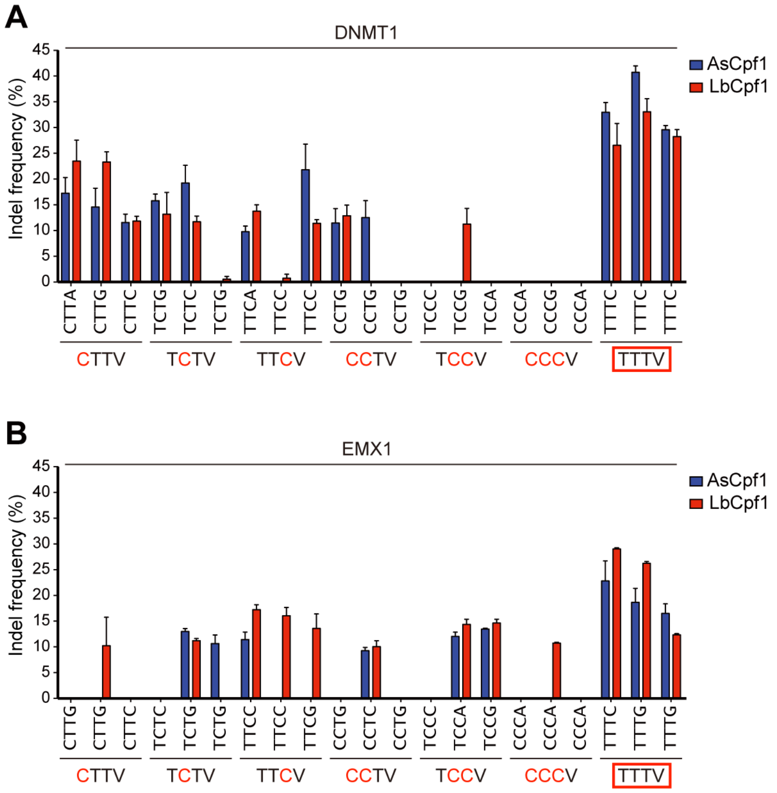


**Figure 4-3. In vivo DNA cleavage activities of LbCpf1 and AsCpf1**

(A and B) In vivo cleavage activities of LbCpf1 and AsCpf1. Indel frequencies for 42 endogenous target sites with the different PAMs were measured in mammalian cells. Target gene is DNMT1 (A) and EMX1 (B). Data are shown as mean ± s.e.m (n = 3).

The canonical PAM is boxed in red, and the substituted nucleotides are colored red.

### 4.5.3 Crystallization and X-ray diffraction analysis

The WT LbCpf1 in complex with a 40-nt crRNA containing 20-nt guide sequence, a 29-nt target DNA strand, and a 9-nt non-target DNA strand containing the TTTA PAM was suitable for crystallization. In the initial screening, the initial crystals of the complex were obtained under the conditions of No. 15 of PACT (25% PEG1,500, 100 mM MIB buffer pH 6.0) by sitting-drop vapor diffusion method. As a result of the optimization of the condition, the crystals suitable for the X-ray diffraction analysis were obtained under the condition C (13–18% PEG3,350, 100 mM MIB buffer (pH 5.0)) (Figure 4-4). The 2.5 Å resolution data sets of the TTTA complex were collected. Although the crystals of TTTA, TCTA, TCCA, and CCCA complexes were obtained under the same condition, the micro seeding was necessary to obtain the crystals of the TCTA, TCCA, and CCCA complexes (Figures 4-5). The data sets of the TCTA, TCCA, and CCCA complexes were collected in 2.4 Å, 2.5 Å and 2.4 Å resolutions, respectively.

### 4.5.4 Phase determination, model building and structure refinement

Molecular replacement was performed using the crystal structure of LbCpf1–crRNA binary complex as the search model, and thus interpretable electron density maps were obtained. The final structures were refined using the data sets respectively (TTTA: 2.5 Å resolution, $R_{work}/R_{free}$ = 0.178/0.228, TCTA: 2.4 Å resolution, $R_{work}/R_{free}$ = 0.193/0.238, TCCA: 2.5 Å resolution, $R_{work}/R_{free}$ = 0.193/0.244, CCCA: 2.4 Å resolution, $R_{work}/R_{free}$ = 0.178/0.229) (Figures 4-6 and 4-7). Data collection and refinement statistics are shown in Table 4-1.

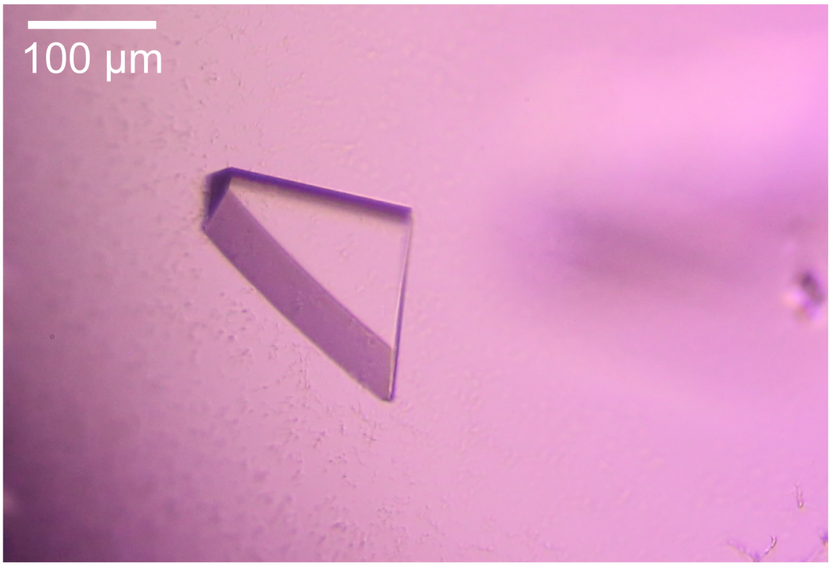


**Figure 4-4. Crystal of LbCpf1**
Crystal of LbCpf1 in complex with crRNA and target DNA.

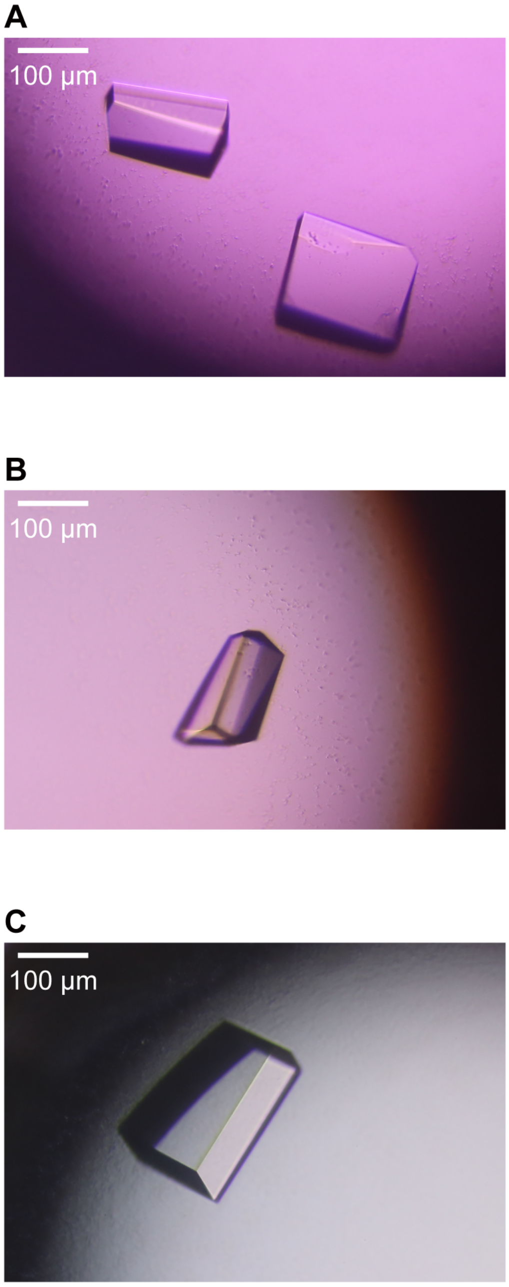


**Figure 4-5. Crystals of  the TCTA, TCCA and CCCA complex**
(A–C) Crystals of the TCTA (A), TCCA (B) and CCCA (C) complexes.

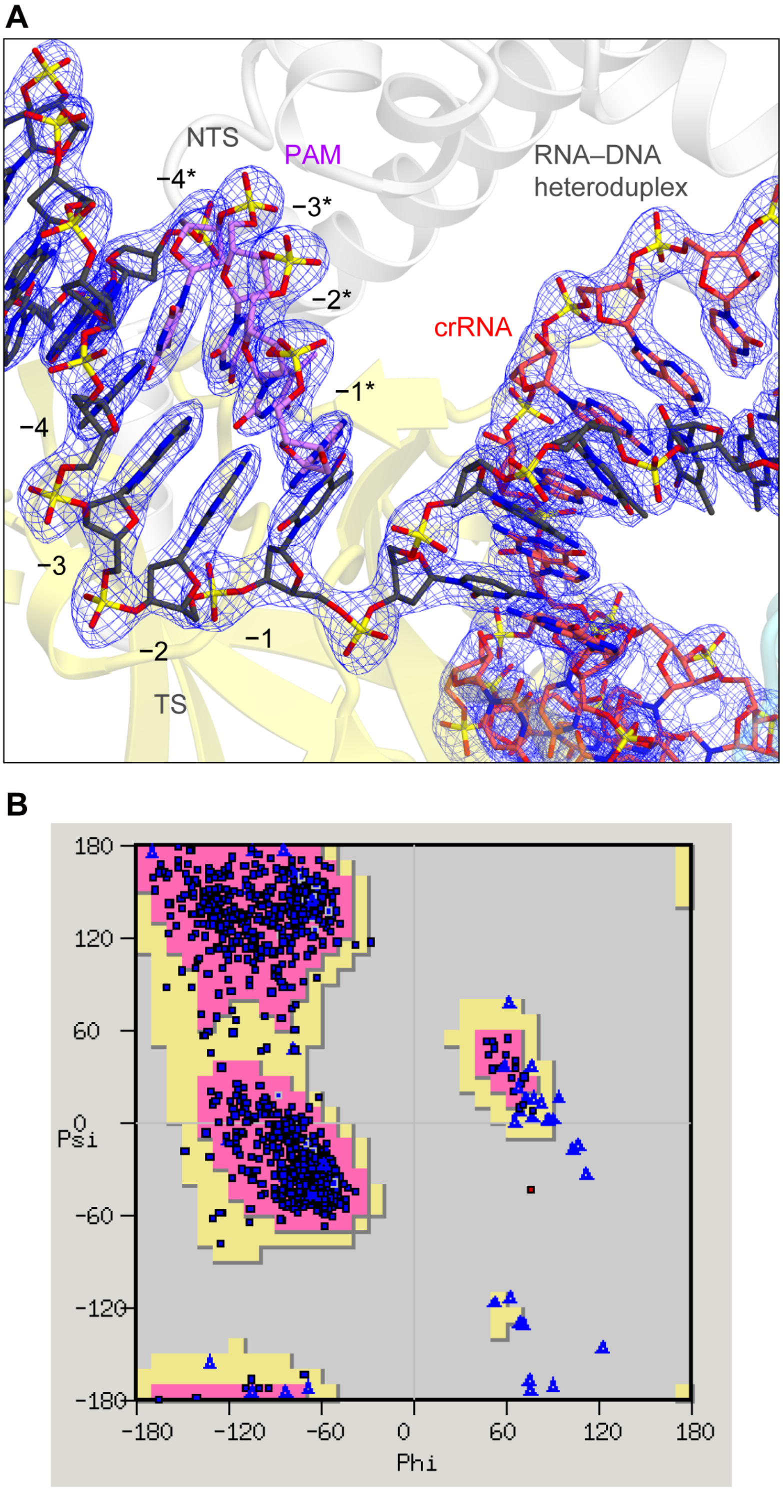


**Figure 4-6. Electron density map of LbCpf1 and Ramachandran plots.**

(A) The $2mF_o – DF_c$ electron density map (contoured at 1.5 σ).

(B) Ramachandran plots of the structure of LbCpf1 in complex with crRNA and target DNA.

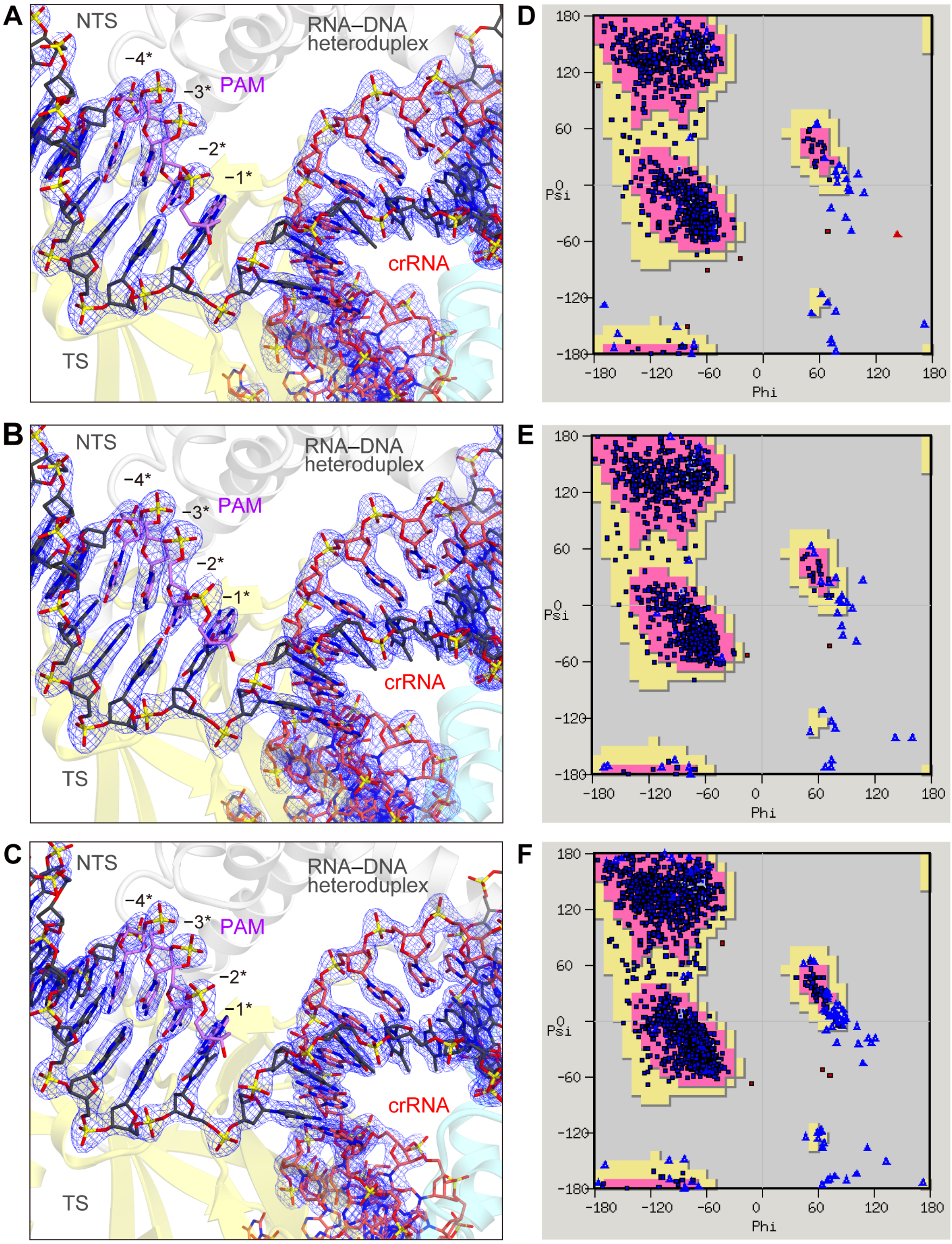


**Figure 4-7. Electron density maps and Ramachandran plots of the TCTA, TCCA and CCCA complexes.**

(A–C) The *2mF*o – *DF*c electron density maps of the TCTA (A), TCCA (B) and CCCA (C) complexes (contoured at 1.5 σ).

(D–F) Ramachandran plots of the TCTA (D), TCCA (E) and CCCA (F) complexes.

**Table 4-1 Data collection and refinement statistics.**

| | TTTA | TCTA | TCCA | TCCC |
|---|---|---|---|---|
| **Data collection** | | | | |
| Beamline | SLS<br>PXI | SPring-8<br>BL41XU | SPring-8<br>BL41XU | SLS<br>PXI |
| Wavelength (Å) | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| Space group | $P4_12_12$ | $P4_12_12$ | $P4_12_12$ | $P2_12_12_1$ |
| Cell dimensions | | | | |
| *a*, *b*, *c* (Å) | 103.2, 103.2, 363.9 | 102.5, 102.5, 373.9 | 102.0, 102.0, 372.5 | 102.0, 103.5, 342.7 |
| *α*, *β*, *γ* (°) | 90, 90, 90 | 90, 90, 90 | 90, 90, 90 | 90, 90, 90 |
| Resolution (Å)* | 49.6–2.5 (2.56–2.50) | 49.4–2.4 (2.45–2.40) | 49.2–2.5 (2.56–2.50) | 49.9–2.4 (2.44–2.40) |
| $R_{\mathrm{pim}}$ | 0.033 (0.436) | 0.014 (0.339) | 0.024 (0.518) | 0.043 (0.162) |
| $I/\sigma I$ | 12.7 (1.6) | 25.0 (2.1) | 15.6 (1.6) | 9.2 (2.8) |
| Completeness (%) | 99.9 (100) | 96.3 (97.8) | 99.8 (100) | 97.3 (82.8) |
| Multiplicity | 12.6 (12.7) | 8.9 (7.0) | 8.6 (9.0) | 4.0 (3.4) |
| CC(1/2) | 0.997 (0.677) | 0.999 (0.700) | 0.998 (0.543) | 0.993 (0.889) |
| | | | | |
| **Refinement** | | | | |
| Resolution (Å) | 49.7–2.5 (2.53–2.50) | 49.4–2.4 (2.43–2.40) | 49.2–2.5 (2.53–2.50) | 49.9–2.4 (2.43–2.40) |
| No. reflections | 69,046 (2,583) | 75,435 (2,649) | 69,014 (2,617) | 138,080 (3,632) |
| $R_{\mathrm{work}}$ / $R_{\mathrm{free}}$ | 0.178 / 0.228 (0.289 / 0.394) | 0.193 / 0.238 (0.307 / 0.383) | 0.193 / 0.244 (0.340 / 0.383) | 0.178 / 0.229 (0.216 / 0.278) |
| No. atoms | | | | |
| Protein | 9,776 | 9,794 | 9,749 | 19,707 |
| Nucleic acid | 1,618 | 1,618 | 1,618 | 3,236 |
| Ion | 5 | 2 | 2 | 5 |
| Solvent | 143 | 53 | 27 | 471 |
| *B*-factors (Å$^2$) | | | | |
| Protein | 83.4 | 81.6 | 88.8 | 60.5 |
| Nucleic acid | 71.3 | 76.4 | 83.5 | 52.0 |
| Ion | 75.7 | 77.9 | 79.2 | 38.8 |
| Solvent | 66.8 | 60.1 | 68.8 | 46.2 |
| R.m.s. deviations | | | | |
| Bond lengths (Å) | 0.007 | 0.007 | 0.007 | 0.007 |
| Bond angles (°) | 0.913 | 0.904 | 0.924 | 0.883 |
| Ramachandran plot (%) | | | | |
| Favored region | 96.83 | 95.62 | 95.94 | 96.56 |
| Allowed region | 3.01 | 3.96 | 3.89 | 3.20 |
| Outlier region | 0.17 | 0.41 | 0.17 | 0.25 |

*Values in parentheses are for the highest resolution shell.

### 4.5.5 Overall structure of the LbCpf1–crRNA–target DNA complex

To clarify the crRNA-guided DNA cleavage mechanism of LbCpf1, we determined the crystal structure of LbCpf1 (residues 1–1,226) in complex with a 40-nt crRNA, a 29-nt target DNA strand and a 9-nt non-target DNA strand containing the TTTA PAM, at 2.5-Å resolution (Figures 4-6 and 4-8A–4-8D and Table 4-1). LbCpf1 adopts a bilobed architecture consisting of a recognition (REC) lobe and a nuclease (NUC) lobe (Figure 4-8C). The REC lobe includes the REC1 and REC2 domains, whereas the NUC lobe comprises the Wedge (WED), PAM-interacting (PI), RuvC and Nuc domains. A characteristic α helix (referred to as the bridge helix) between the RuvC-I and RuvC-II regions interacts with the REC2 domain. The crRNA consists of the 20-nt 5′-handle and the 20-nt guide sequence (Figure 4-8D). The crRNA guide sequence and the target DNA strand form the 20-bp RNA–DNA heteroduplex, which is accommodated in the central channel between the two lobes (Figures 4-8C, 4-8D and 4-9). Two $Mg^{2+}$ ions are bound to the crRNA 5′-handle (Figure 4-8E). One $Mg^{2+}$ is coordinated by six water molecules that interact with the phosphate groups of A(−13), U(−14) and A(−19) and the nucleobases of A(−13) and U(−18), as observed in the LbCpf1–crRNA binary complex (Dong et al., 2016). The other $Mg^{2+}$ is coordinated by the main-chain carbonyl group of Thr716, the phosphate group of A(−4) and four water molecules that interact with the phosphate groups of G(−3) and U(−5) and the side chains of Asp708 and Asn718 (Figure 4-8E).

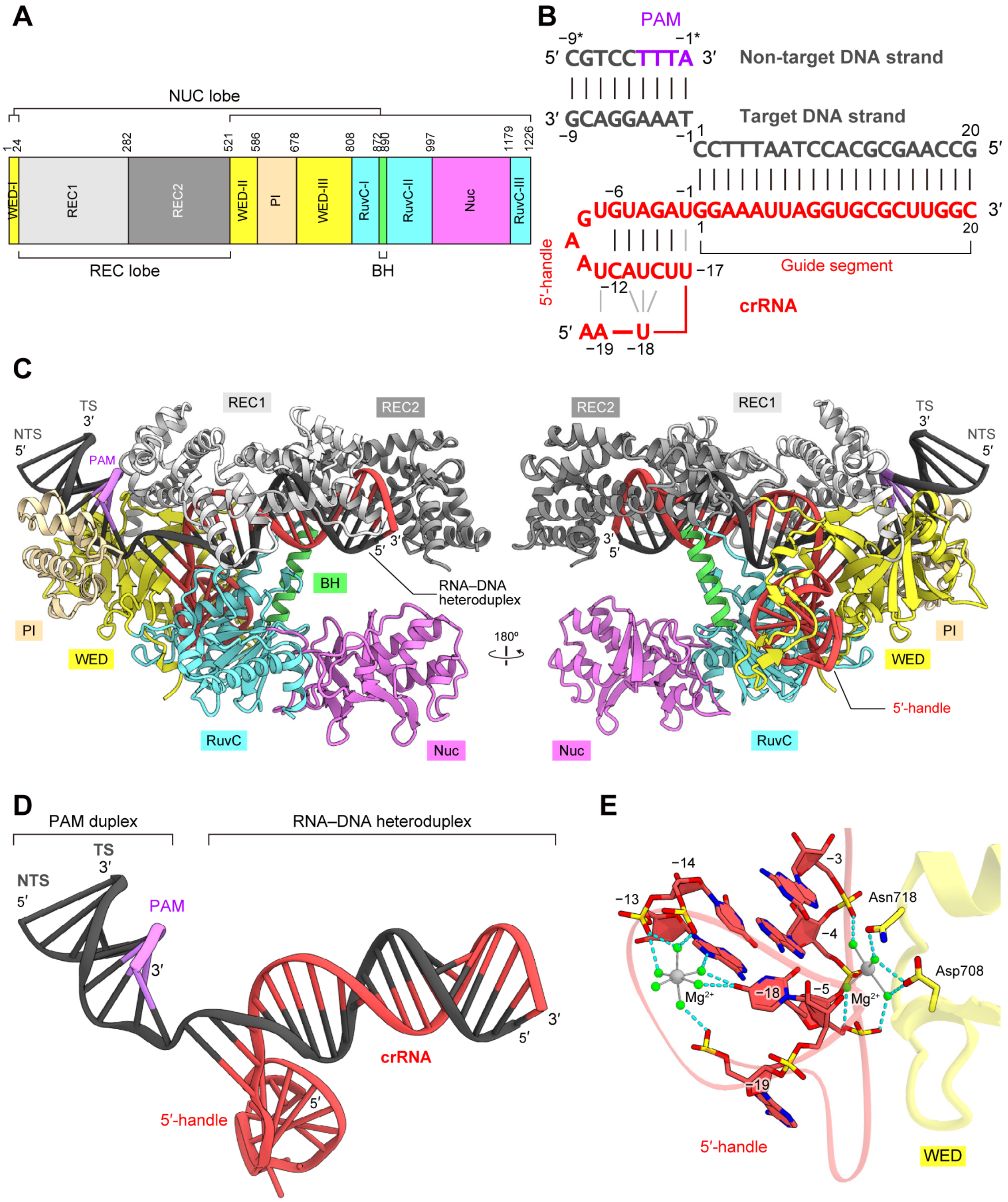


**Figure 4-8. Structure of the LbCpf1–crRNA–target DNA complex**

(A) Domain organization of LbCpf1. BH, bridge helix.

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

(C) Crystal structure of LbCpf1 in complex with the crRNA and its target DNA.

(D) Structure of the crRNA and its target DNA.

(E) Binding of $Mg^{2+}$ ions to the crRNA. The bound $Mg^{2+}$ ions and water molecules are indicated by gray and green spheres, respectively. Hydrogen bonds are shown as dashed lines.

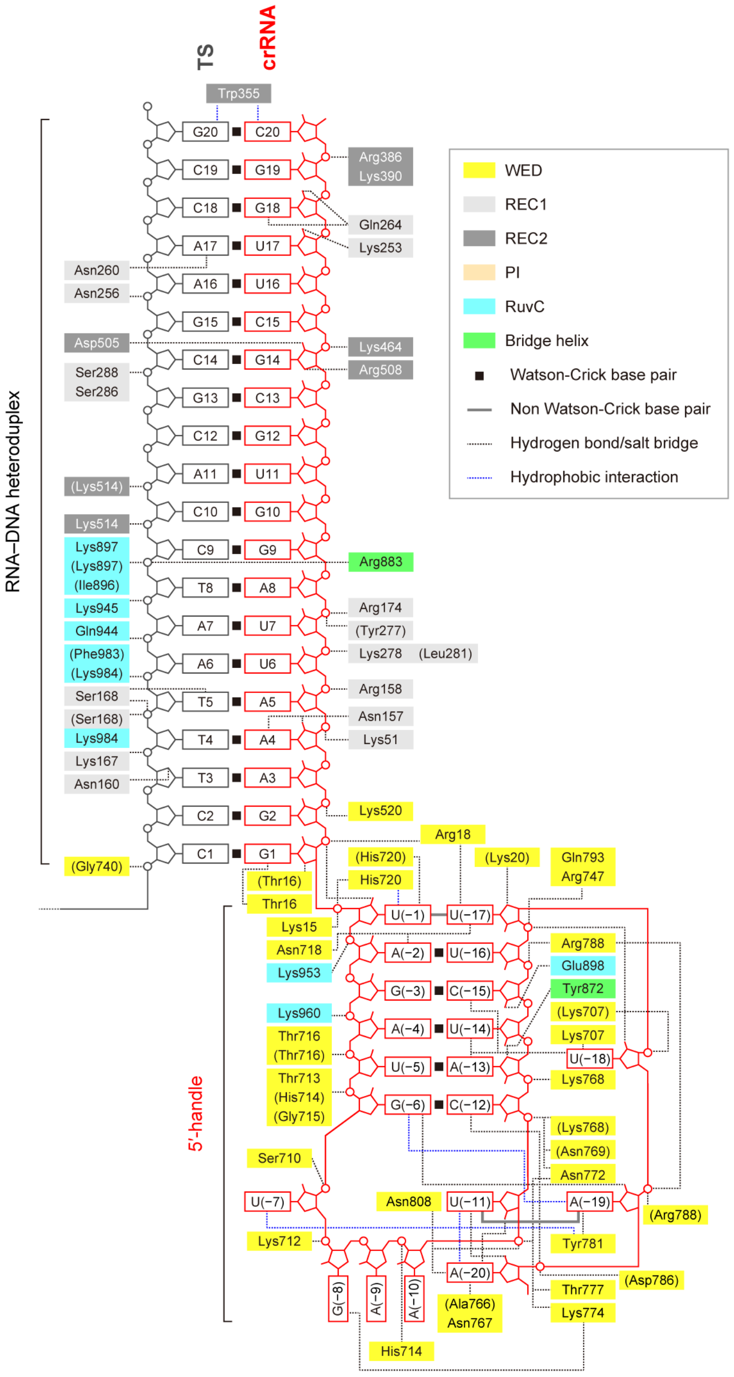


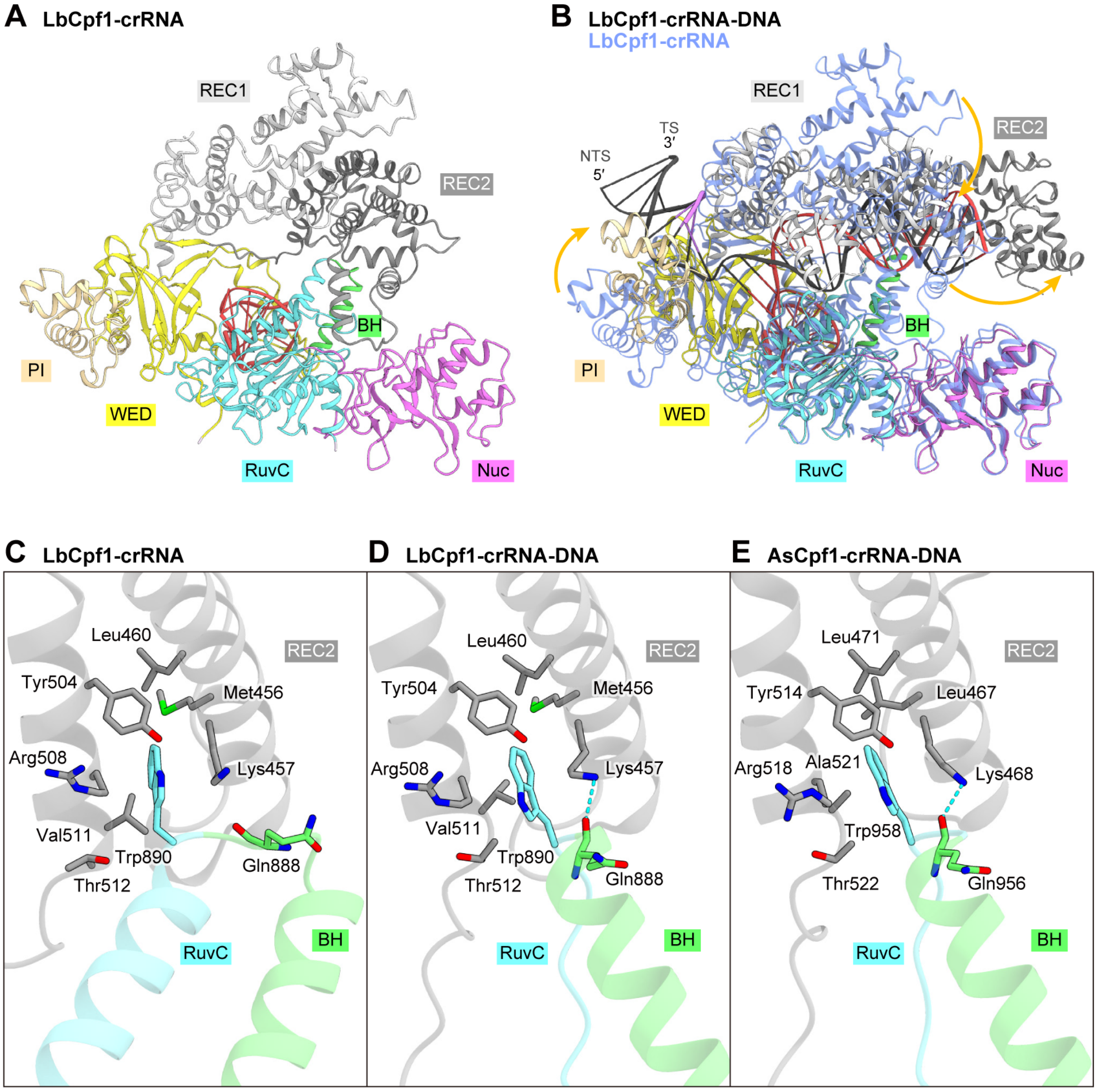
**Figure 4-9. Schematic of the nucleic acid recognition by LbCpf1**

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.

### 4.5.6 Structural rearrangement upon target DNA binding

A structural comparison between the LbCpf1 binary and ternary complexes revealed that, while the individual domains are structurally similar, the REC1, REC2 and PI domains undergo conformational rearrangements upon target DNA binding (Figures 4-10A, 4-10B and 4-11). These structural rearrangements are consistent with a previous prediction, based on a comparison of the LbCpf1 binary complex with the AsCpf1 ternary complex (Gao et al., 2016). As compared to the binary complex, the REC1 and REC2 domains move toward and away from the NUC lobe in the ternary complex, respectively, to form the central channel that accommodates the crRNA–target DNA heteroduplex (Figures 4-10A, 4-10B and 4-11B). In the ternary complex, the PI domain moves toward the REC1 and WED domains to form the PAM-binding channel (Figures 4-10B and 4-11C). In addition, whereas the crRNA guide segment is disordered in the binary complex (Dong et al., 2016), the crRNA guide is ordered and forms the heteroduplex with the target DNA in the ternary complex (Figures 4-10B and 4-11D).

A structural comparison between the binary and ternary complexes also revealed a local conformational change in the bridge helix and the RuvC-II region. In the binary complex, residues 872–885/890–918 and 886–889 adopt α-helical and loop conformations, respectively (Dong et al., 2016) (Figure 4-10C). In contrast, in the ternary complex, residues 885–889 and 890–896 adopt α-helical and loop conformations, respectively (Figure 4-10D). In the binary and ternary complexes, Trp890 is inserted into a hydrophobic pocket in the REC2 domain (Figures 4-10C and 4-10D). In addition, in the ternary complex, the main-chain carbonyl group of Gln888 hydrogen bonds with the side chain of Lys457 in the REC2 domain (Figure 4-10D). These interdomain interactions are conserved in the AsCpf1 ternary complex (Figure 4-10E), and the W958A mutation reduced the DNA cleavage activity (Yamano et al., 2016). These observations highlighted the pivotal role of the conserved tryptophan residue in the RuvC domain for the structural transition from the binary to the ternary complex.



## Figure 4-10. Comparison between the binary and ternary complex structures of LbCpf1

(A) Crystal structure of the LbCpf1–crRNA complex (PDB: 5ID6) (Dong et al., 2016).

(B) Superimposition of the LbCpf1–crRNA–target DNA complex (colored) and the LbCpf1–crRNA complex (blue). Structural changes are indicated by orange arrows.

(C–E) Interactions between the REC and NUC lobes in the LbCpf1 binary complex (C), the LbCpf1 ternary complex (D), and the AsCpf1 ternary complex (PDB: 5B43) (Yamano et al., 2016) (E). Hydrogen bonds are shown as dashed lines.

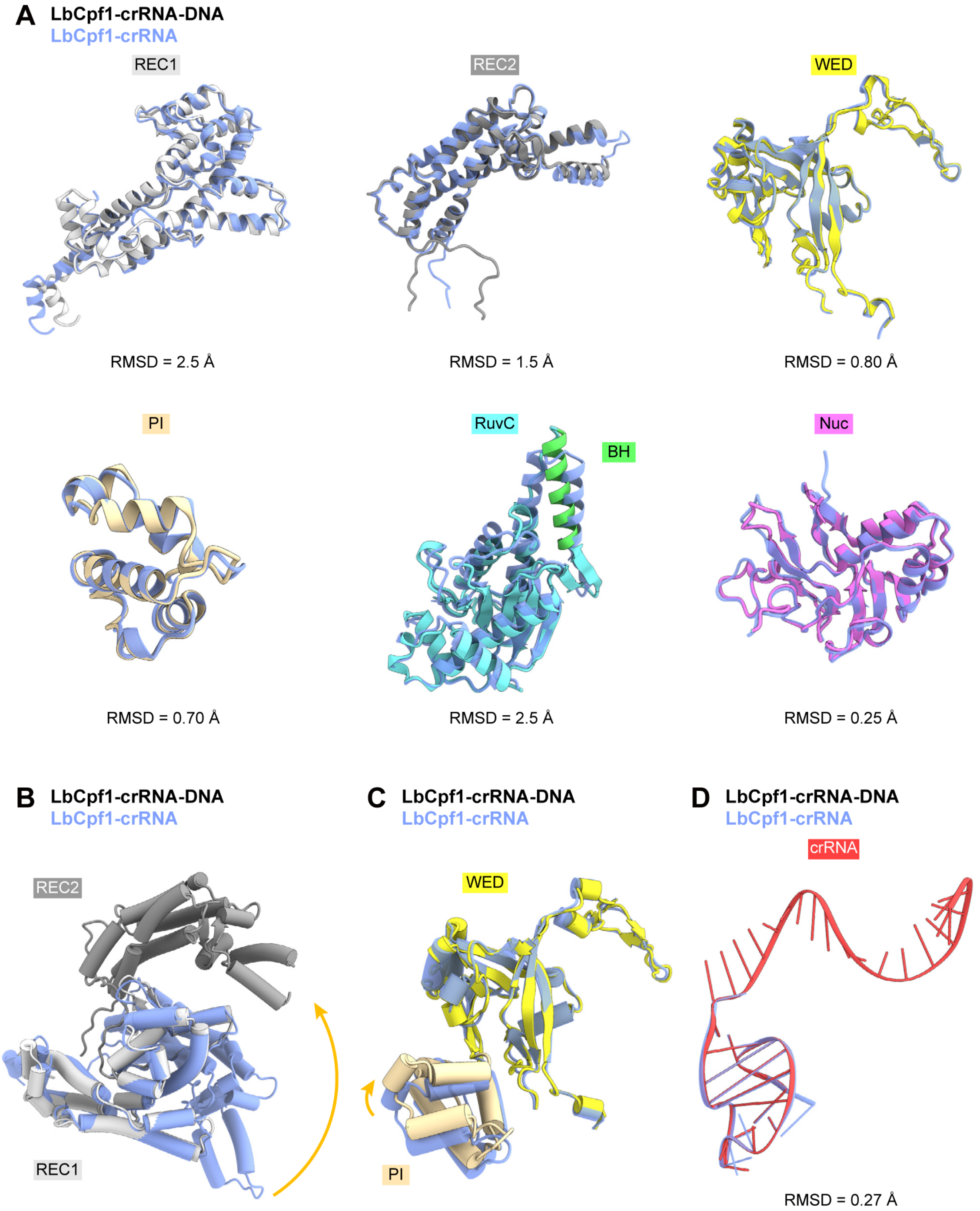


**Figure 4-11. Structural comparison between the individual domains of the LbCpf1 binary and ternary complexes**

(A) Superimposition of the individual domains of LbCpf1 binary (PDB: 5ID6) (blue) and ternary (colored) complexes. Root mean square deviation (RMSD) values for equivalent Cα atoms are shown below the structures.

(B) Superimposition of the REC lobe and WED-PI domains of LbCpf1 binary (PDB: 5ID6) (blue) and ternary (colored) complexes based on their REC1 and WED domains, respectively.

### 4.5.7 RNA-DNA heteroduplex recognition mechanism

A structural comparison between the ternary complexes of AsCpf1 and LbCpf1 revealed that they share a bilobed architecture and recognize the RNA–DNA heteroduplex in similar manners (Figures 4-12A and 4-12B). In the LbCpf1 ternary complex, the 20-bp RNA–DNA heteroduplex is recognized by the REC lobe, in which Trp355 in the REC2 domain stacks with the C20:dG20 base pair in the heteroduplex (Figure 4-12C). Similarly, in the AsCpf1 ternary complex, the 20-bp heteroduplex is accommodated within the REC lobe, in which Trp382, equivalent to Trp355 in LbCpf1, stacks with the C20:dG20 base pair, while AsCpf1 was crystallized with a 24-nt-guide-containing crRNA and its complementary target DNA (Yamano et al., 2016) (Figure 4-12D). Consistent with these structural findings, the crRNA–DNA base pairing in the PAM-distal region is dispensable for the Cpf1-mediated DNA cleavage (Kim et al., 2016a, 2016b; Kleinstiver et al., 2016b; Zetsche et al., 2015a). Together, these observations indicated that LbCpf1 and AsCpf1 recognize the 20-bp crRNA–target DNA heteroduplex.

### 4.5.8 RuvC and Nuc domains

The RuvC and Nuc domains of LbCpf1 are structurally similar to those of AsCpf1 (Figures 4-12E and 4-12F). In the AsCpf1 structure, the RuvC active site is formed by the conserved acidic residues Asp908, Glu993 and Asp1263 (Yamano et al., 2016) (Figure 4-12F). The D908A, E993A and D1263A mutations abolished the *in vitro* cleavage activities, indicating that the RuvC domain is involved in the cleavage of both strands (Zetsche et al., 2015; Yamano et al., 2016). In contrast, Arg1226 in the Nuc domain participates in the target strand cleavage (Yamano et al., 2016). In the LbCpf1 structure, Asp832, Glu925, Asp1180 and Arg1138, equivalent to Asp908, Glu993, Asp1263 and Arg1226 of AsCpf1, are similarly arranged (Figures 4-12E and 4-13). In addition, our *in vitro* cleavage assays confirmed that the D832A, E925A and D1180A mutations abolish the DNA cleavage activity of LbCpf1, while the R1138A mutant functions as a nickase, as in the case of AsCpf1 (Yamano et al., 2016; Zetsche et al., 2015a) (Figure 4-14). These observations indicated that LbCpf1 and AsCpf1 cleave the target DNA *via* similar mechanisms, in which the RuvC domain participates in the cleavage of both strands, while the Nuc domain is involved in the target strand cleavage.

Recently, the crystal structures of C2c1 (also known as Cas12b) from *Alicyclobacillus acidoterrestris* (AaC2c1), which was identified as a type V-B effector nuclease (Shmakov et al., 2015), elucidated its action mechanism and offered clues toward understanding the DNA cleavage mechanism of the type V-A Cpf1 nucleases (Liu et al., 2016; Yang et al., 2016). Despite their limited sequence similarity, Cpf1 and C2c1 have comparable domain architectures and share the RuvC domain, in which the catalytic residues are similarly arranged (Figures 4-15A–4-15D). As in Cpf1, the Nuc domain is located next to the RuvC domain in C2c1, although their Nuc domains adopt distinct folds (Figures 4-15A and 4-15B). In the AaC2c1 structure, Arg911 in the Nuc domain interacts with the backbone phosphate group of the target strand, thereby guiding the target strand into the RuvC active site (Figure 4-15D). These structural observations suggested that C2c1 uses the RuvC active site to cleave both the target and non-target strands (Yang et al., 2016). Notably, LbCpf1 Arg1138 and AsCpf1 Arg1226 are located at positions analogous to that of AaC2c1 Arg911 (Figures 4-15C and 4-15D), and they are involved in the target strand cleavage (Yamano et al., 2016) (Figure 4-14). Thus, these observations suggested that the Cpf1 Nuc domain plays a role in guiding the target strand into the RuvC active site, rather than catalyzing the cleavage of the target strand. Supporting this notion, a recent study indicated that *Francisella novicida* Cpf1 cleaves the target and non-target DNA strands, using the same RuvC domain active site (Swarts et al., 2017).

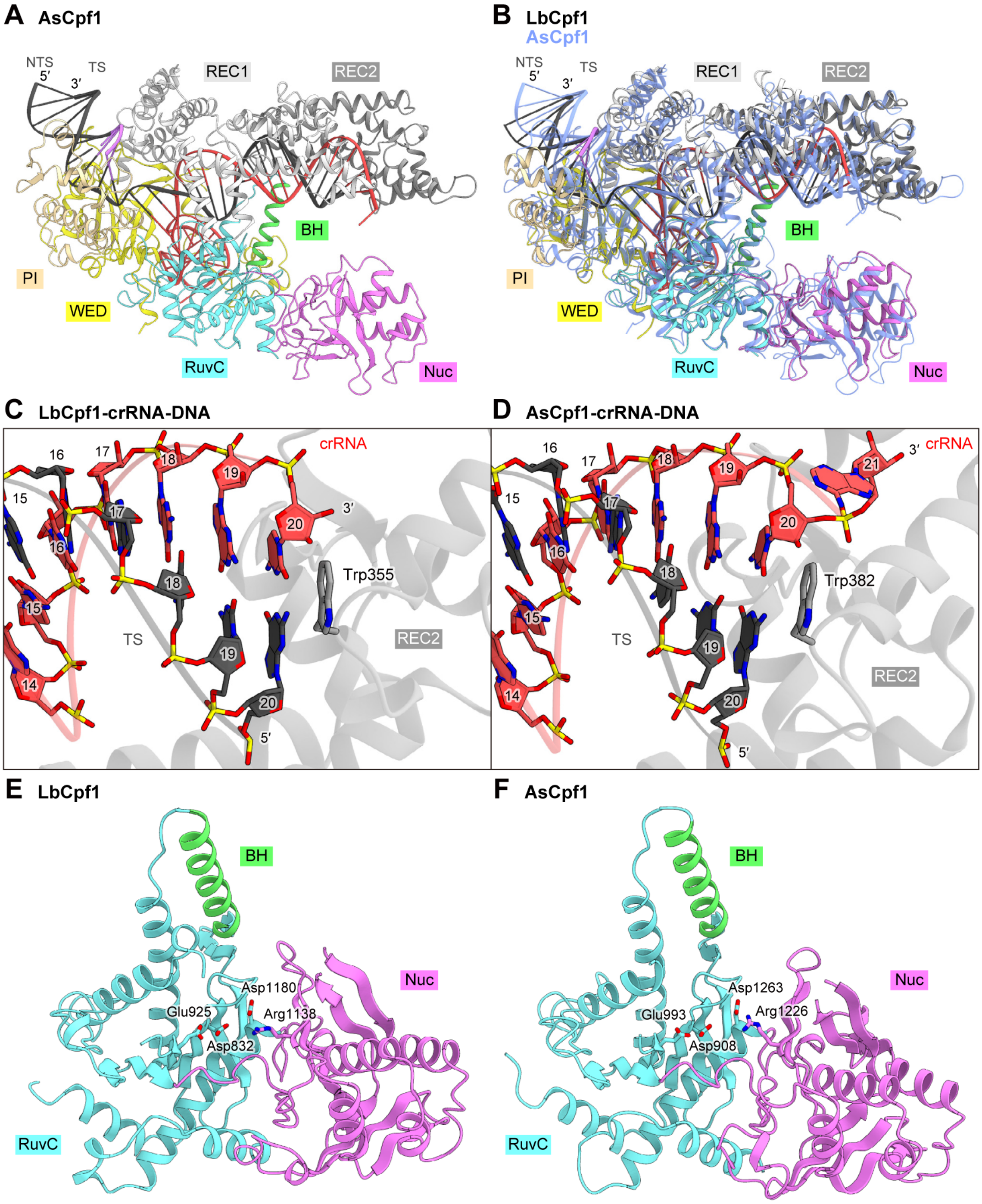


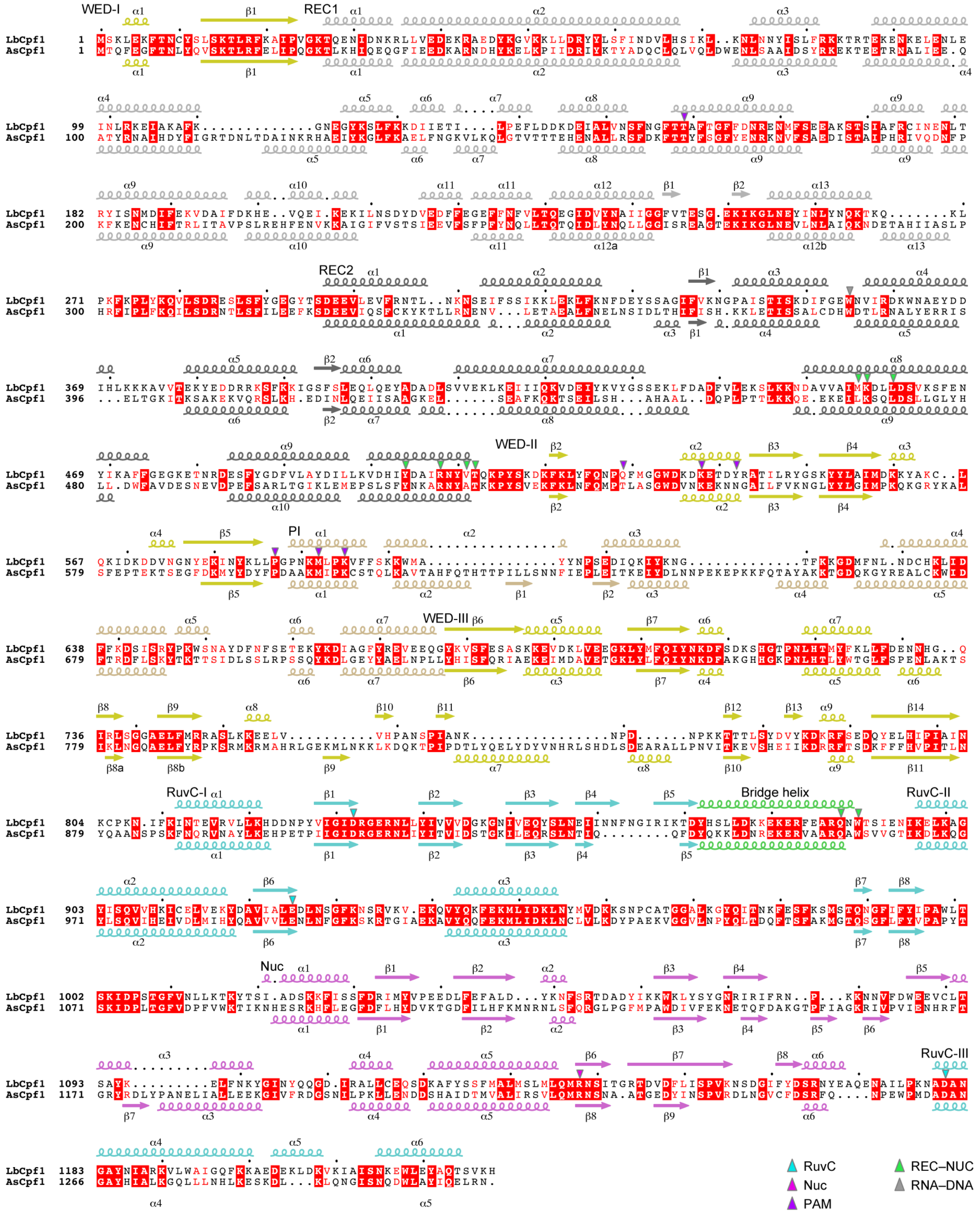
**Figure 4-12. Comparison between the ternary complex structures of LbCpf1 and AsCpf1**

(A) Crystal structure of the AsCpf1–crRNA–DNA complex (PDB: 5B43) (Yamano et al., 2016).

(B) Superimposition of the LbCpf1–crRNA–DNA complex (colored) and the AsCpf1–crRNA–DNA complex (blue).

(C and D) RNA-DNA heteroduplex recognition by LbCpf1 (C) and AsCpf1 (D).

(E and F) RuvC and Nuc domains of LbCpf1 (E) and AsCpf1 (F).



## Figure 4-13. Sequence Alignment of Cpf1 Orthologs

The figure was prepared using ClustalW (McWilliam et al., 2013) and ESPript (Gouet et al., 2003). As, *Acidaminococcus sp.* BV3L6; Lb, *Lachnospiraceae bacterium* ND2006.

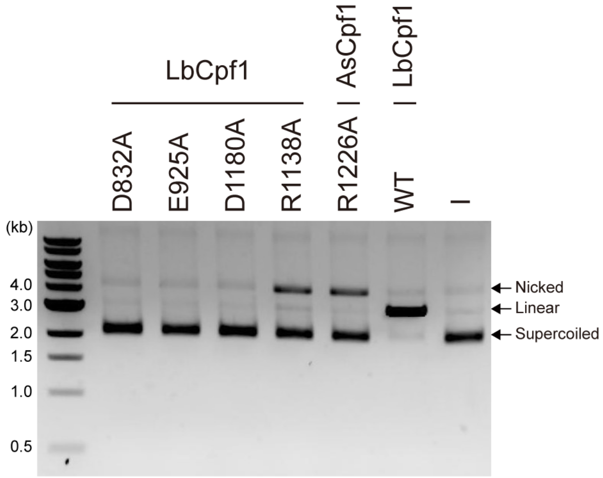
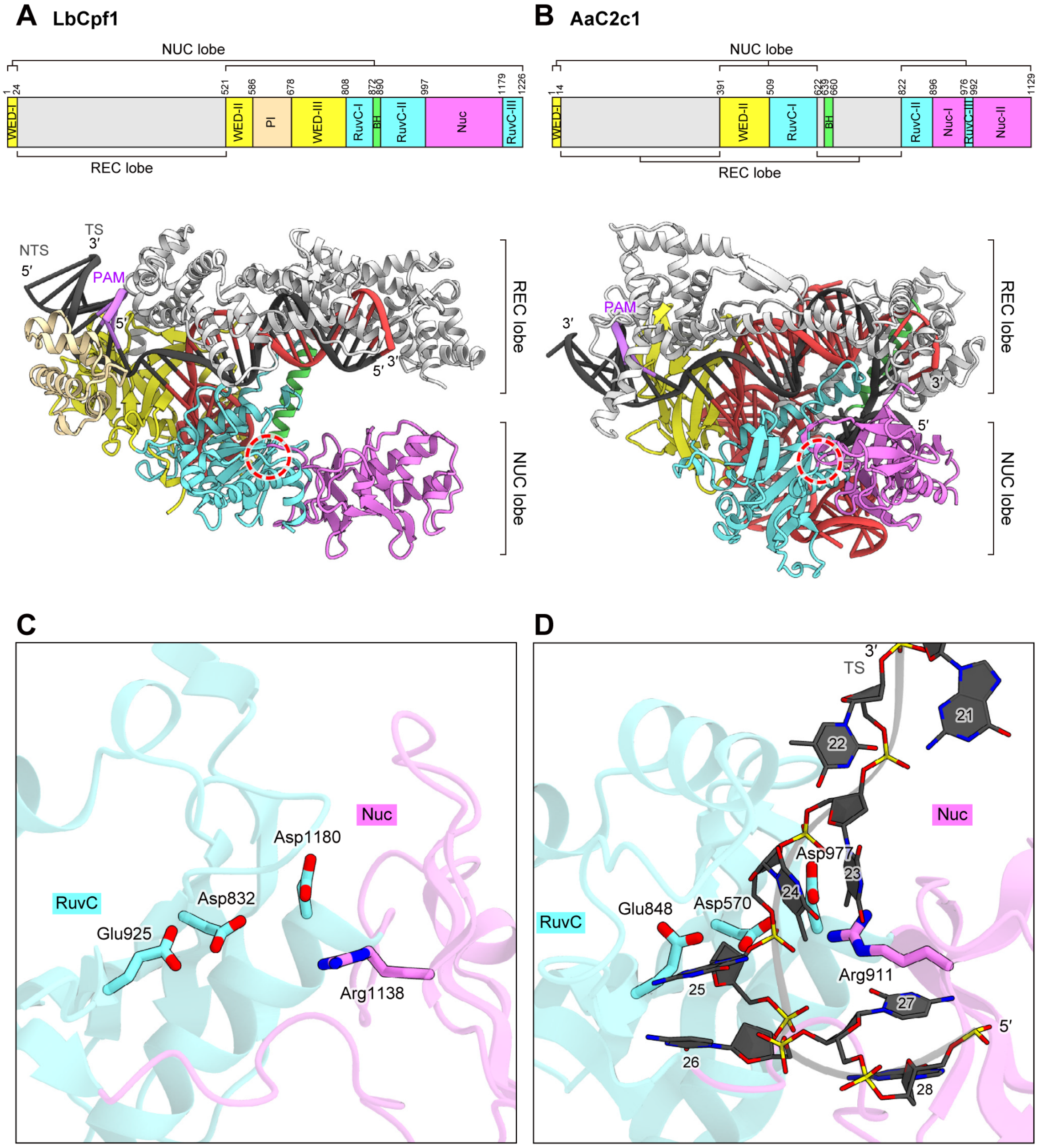


**Figure 4-14. Mutation Analysis of the Catalytic Residues**

In vitro cleavage activity of the wild type or mutants of Cpf1s. The Cpf1-crRNA complex (100 nM) was incubated at 37°C for 20 minutes with a circular plasmid target with the TTTA PAM. The mutation of the RuvC catalytic residues (D832A, E925A and D1180A) abolished the cleavage activity, and the R1138A mutant nicked the plasmid target.



## Figure 4-15. Comparison between Cpf1 and C2c1

(A and B) Comparison of the domain organizations and overall structures between LbCpf1 (A) and AaC2c1 (PDB: 5U30) (B). The catalytic centers of the RuvC domain are indicated by red circle.

(C and D) Comparison of the active sites between LbCpf1 (C) and AaC2c1 (PDB: 5U30) (D). The side chains of Asp570, Glu848 and Asp977 in AaC2c1 are modeled.

### 4.5.9 Recognition mechanism of the canonical TTTV PAM

In the LbCpf1 ternary complex, the TTTA PAM duplex is bound to the channel formed by the REC1, WED and PI domains (Figures 4-16A and 4-16B). The PAM duplex adopts a distorted conformation with a narrow minor groove, as compared with the canonical B-form DNA (root-mean-square deviation (RMSD) is 1.4 Å for 18 equivalent phosphorus atoms), as observed in the AsCpf1 ternary complex (Yamano et al., 2016) (Figure 4-16C). The dT(−1):dA(−1*) base pair in the PAM duplex does not form base-specific contacts with the protein (Figure 4-16B). The O2 of dT(−2*) forms a hydrogen bond with Lys595, whereas the N7 of dA(−2) and the backbone phosphate group between dT(−1) and dA(−2) form hydrogen bonds with Tyr542 (Figure 4-17A). dA(−2) also forms van der Waals interactions with Pro587, Met592 and Lys595. The 5-methyl group of dT(−3*) is in the vicinity of the side-chain methyl group of Thr149, whereas the N3 and N7 of dA(−3) form hydrogen bonds with Lys595 and Lys538, respectively (Figure 4-17B). The 5-methyl group of dT(−4*) is surrounded by the side chains of Thr149 and Gln529, whereas the O4′ of dA(−4) forms a hydrogen bond with Lys595 (Figure 4-17C).

A structural comparison of LbCpf1 with AsCpf1 revealed the high conservation between their PAM recognition mechanisms. In AsCpf1, Lys548 and Lys607, equivalent to Lys538 and Lys595 of LbCpf1, similarly interact with the PAM duplex (Yamano et al., 2016) (Figures 4-13 and 4-17D–4-17F). Nonetheless, their PAM recognition mechanisms are slightly different. Whereas Tyr542 of LbCpf1 forms two hydrogen bonds with dA(−2) (Figure 4-17A), Asn552 of AsCpf1, equivalent to Tyr542 of LbCpf1, does not interact with the PAM duplex (Figure 4-17D).

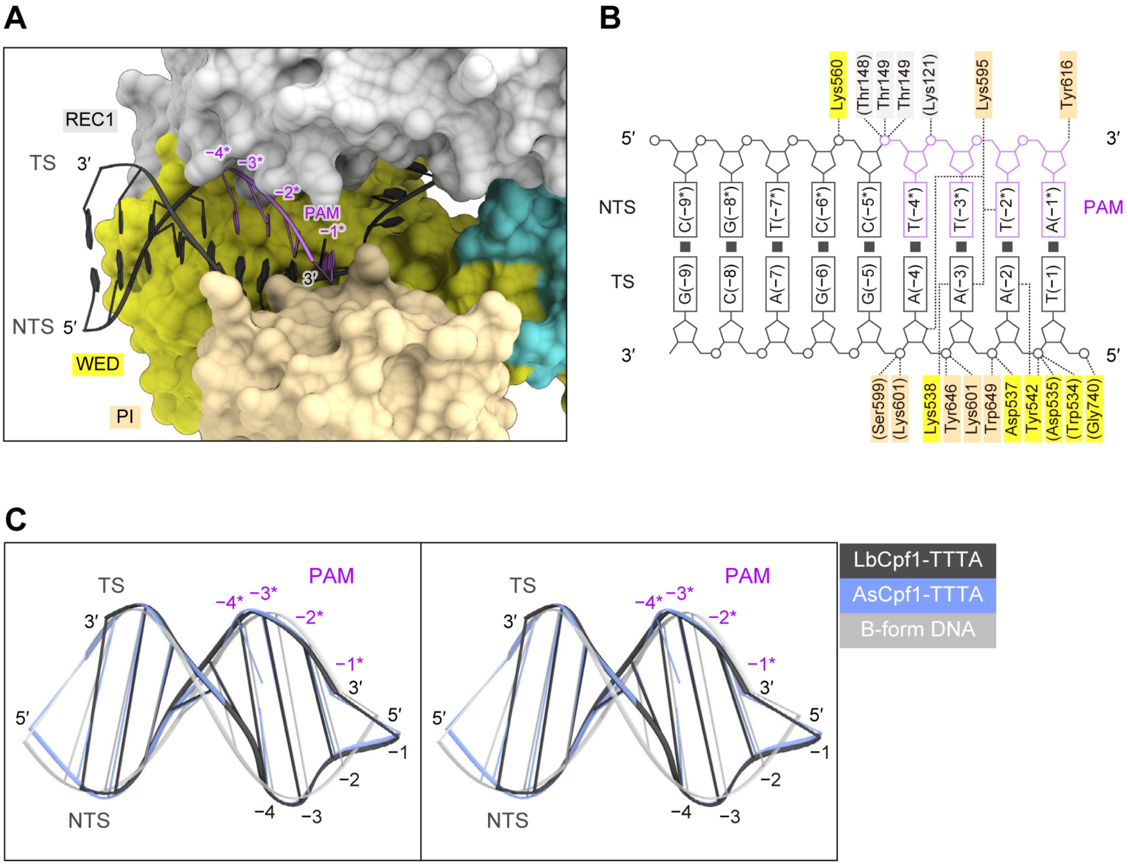


**Figure 4-16. Structure of canonical PAM duplex**

(A) PAM-duplex binding in the LbCpf1 ternary complex.

(B) Schematic of the PAM-duplex recognition by LbCpf1.

(C) Superimposition of the PAM duplexes in LbCpf1 and AsCpf1 (PDB: 5B43) (Yamano et al., 2016) onto the B-form DNA duplex (stereo view).

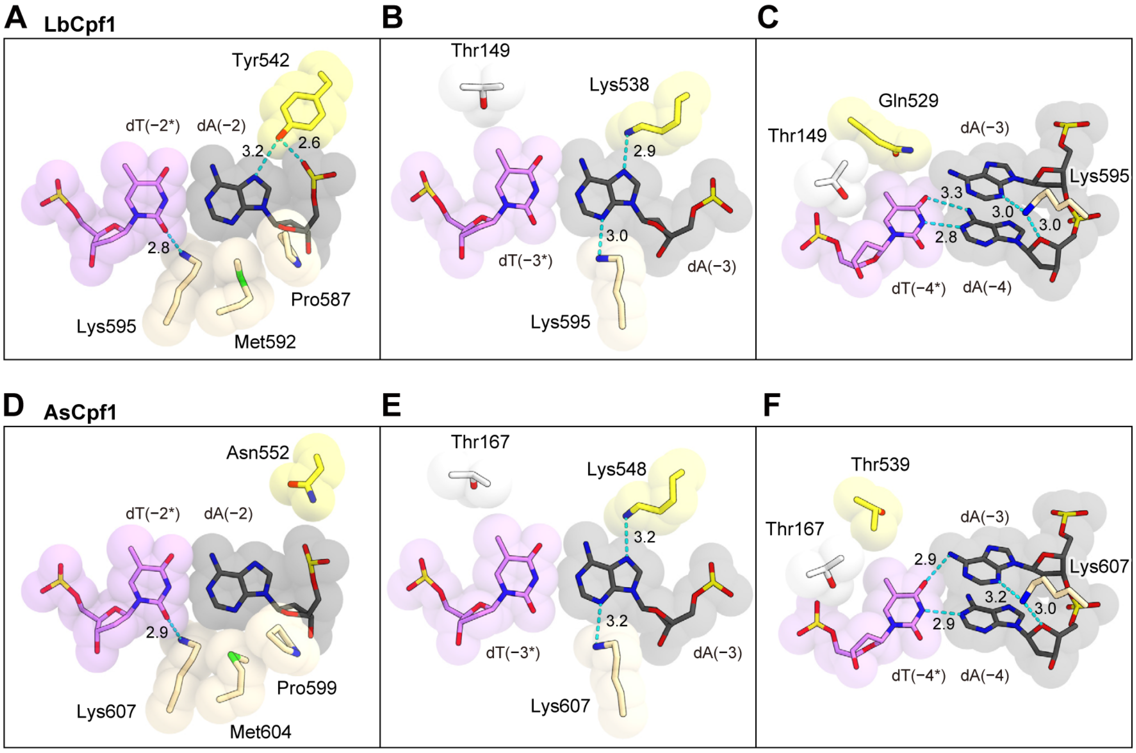


**Figure 4-17. Canonical PAM recognition mechanism**

(A–C) Recognition of dA(−2):dT(−2*) (A), dA(−3):dT(−3*) (B), and dA(−4):dT(−4*) (C) by LbCpf1.

(D–F) Recognition of dA(−2):dT(−2*) (D), dA(−3):dT(−3*) (E), and dA(−4):dT(−4*) (F) by AsCpf1.

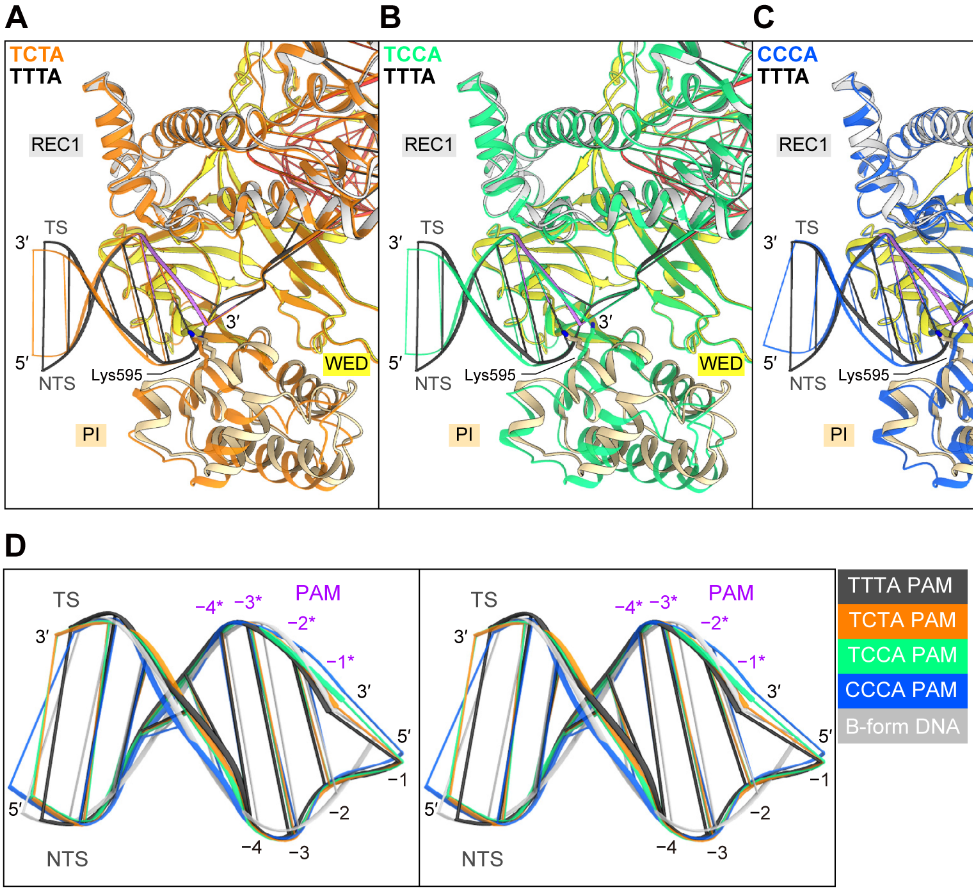
Hydrogen bonds are shown as dashed lines.

### 4.5.10 Recognition mechanisms of the non-canonical PAMs

To investigate the non-canonical PAM recognition mechanism, we determined the crystal structures of the LbCpf1 ternary complex containing either TCTA, TCCA or CCCA as the PAM (Table 4-1). As in the TTTA complex, the TCTA and TCCA complexes crystallized in the space group $P4_12_12$, with one complex molecule in the asymmetric unit. In contrast, the CCCA complex crystallized in the space group $P2_12_12_1$, with two complex molecules in the asymmetric unit, probably due to the slightly different conformations of the bound DNA molecules. Since the two CCCA complex molecules are essentially identical (RMSD is 1.3 Å for equivalent Cα atoms), we will refer to one complex molecule for the following discussion.

In the three complexes, the PAM duplexes are recognized by the REC1, WED and PI domains, as in the TTTA complex (Figures 4-18A–4-18C), and they adopt less distorted conformations, as compared with that in the TTTA complex, probably due to the presence of the G:C base pair(s) (Figure 4-18D). A comparison of the four complex structures revealed that, whereas the REC1 and WED domains are similarly arranged, the PI domains undergo outward displacement in the TCTA, TCCA and CCCA complexes, as compared with that in the TTTA complex (Figures 4-18A–4-18C). The RMSD values for the equivalent Cα atoms in the PI domain between the TTTA complex and the TCTA/TCCA/CCCA complexes are 2.2, 2.4 and 2.2 Å, respectively, by superimposition based on the regions except for the PI domain. The PI domain displacement results in the opening of the PAM-binding channel, thereby allowing the binding of PAM duplexes with distinct conformations. Consistent with its conformational flexibility, the PI domain exhibits higher *B*-factor values than the other domains in the four complexes (Figures 4-19A–4-19D). Furthermore, the PI domain in the TTTA complex displays lower *B*-factor values than those in the TCTA and TCCA complexes (Figures 4-19A–4-19C). These structural observations suggest that LbCpf1 binds the TTTA PAM duplex more stably than the non-canonical PAM duplexes, thereby explaining the preference of LbCpf1 for the canonical TTTV PAM.

In the four complex structures, Lys538 and Tyr542 similarly interact with the second and third PAM-complementary nucleotides (Figures 4-20). In contrast, Lys595 interacts with the PAM nucleotides in distinct manners among the four complex structures, due to the conformational differences in the PI domain and the PAM duplex (Figures 4-20E–4-20H). In the TTTA complex, Lys595 forms hydrogen bonds with the N3 of dA(−3) and the O2 of dT(−2*) (Figure 4-20E). In contrast, in the TCTA complex, Lys595 does not hydrogen bond with the N3 of dG(−3) (Figure 4-20F). Notably, in the TCCA and CCCA complexes, Lys595 is not inserted into the minor groove of the PAM duplex (Figures 4-20G and 4-20H), probably due to steric hindrance between Lys595 and the N2 of dG(−2) (Figure 4-19E). These structural observations are consistent with the fact that LbCpf1 recognizes the non-canonical C-containing PAMs less efficiently than the canonical TTTV PAM. In addition, the side chain of Lys595 is less ordered in the TCTA, TCCA and CCCA complexes, as compared with that in the TTTA complex (Figures 4-20A–4-20D), consistent with the weaker interactions between Lys595 and the non-canonical PAM duplexes. Given the highly conserved PAM recognition mechanisms, AsCpf1 likely recognizes the non-canonical PAMs in similar manners. Together, these structural findings revealed the previously undescribed mechanisms of the canonical and non-canonical PAM recognition by the Cpf1 nucleases.



## Figure 4-18. Comparison between the TTTA, TCTA, TCCA and CCCA complex structures

(A) Superimposition of the TCTA (orange) and TTTA (colored) complexes.

(B) Superimposition of the TCCA (green) and TTTA (colored) complexes.

(C) Superimposition of the CCCA (blue) and TTTA (colored) complexes.

(D) Superimposition of the PAM duplexes in the TTTA, TCTA, TCCA and CCCA complex structures onto the B-form DNA duplex (stereo view).

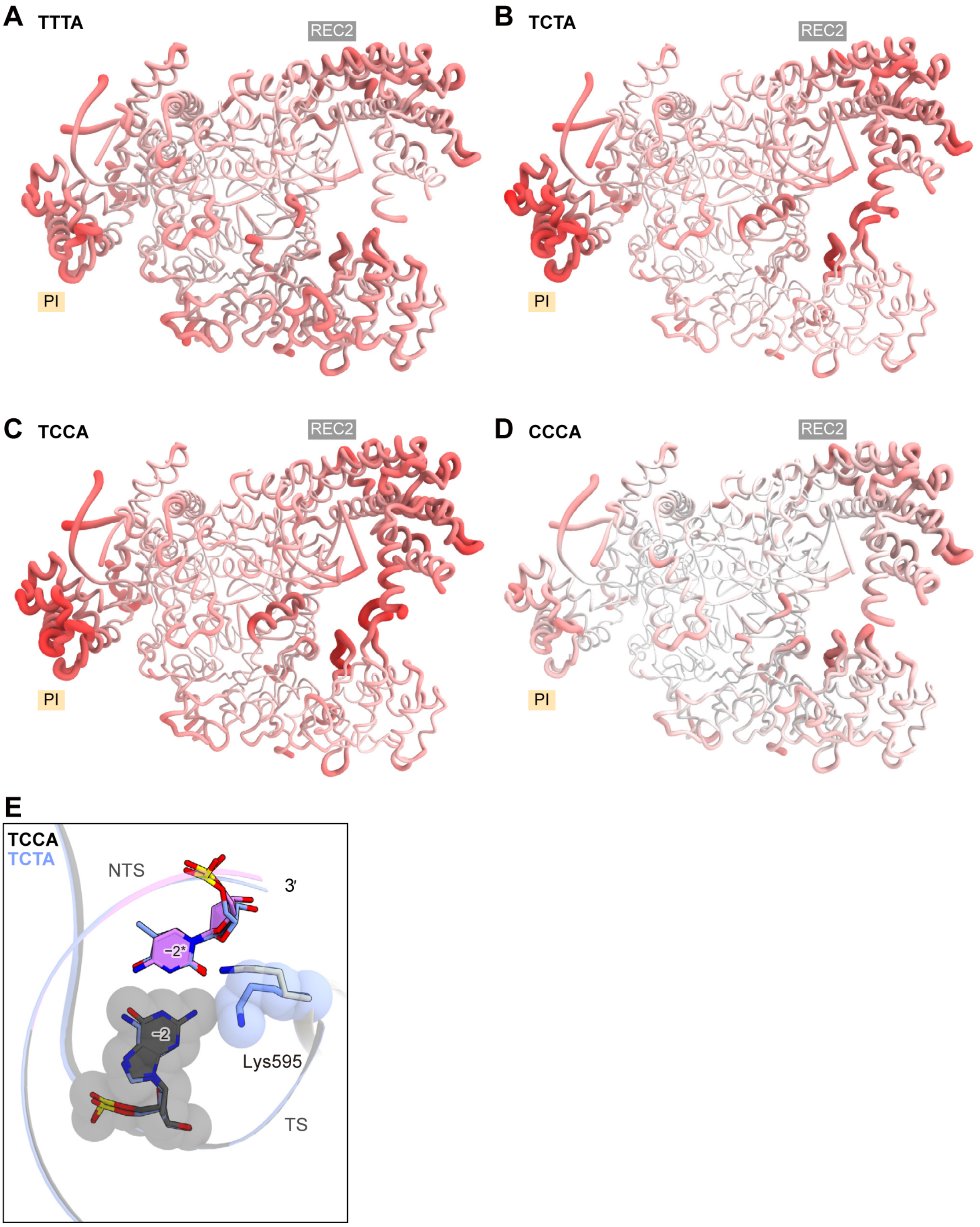


**Figure 4-19. Flexibility of the PI domain**

(A–D) *B*-factor distributions of the TTTA (A), TCTA (B), TCCA (C) and CCCA (D) complexes. The *B*-factor values are colored from white (40 Å$^2$) to red (200 Å$^2$). Note that the CCCA complex exhibits lower *B*-factor values, due to the different crystal packing interactions.

(E) Superimposition of the TCTA and TCCA complexes.

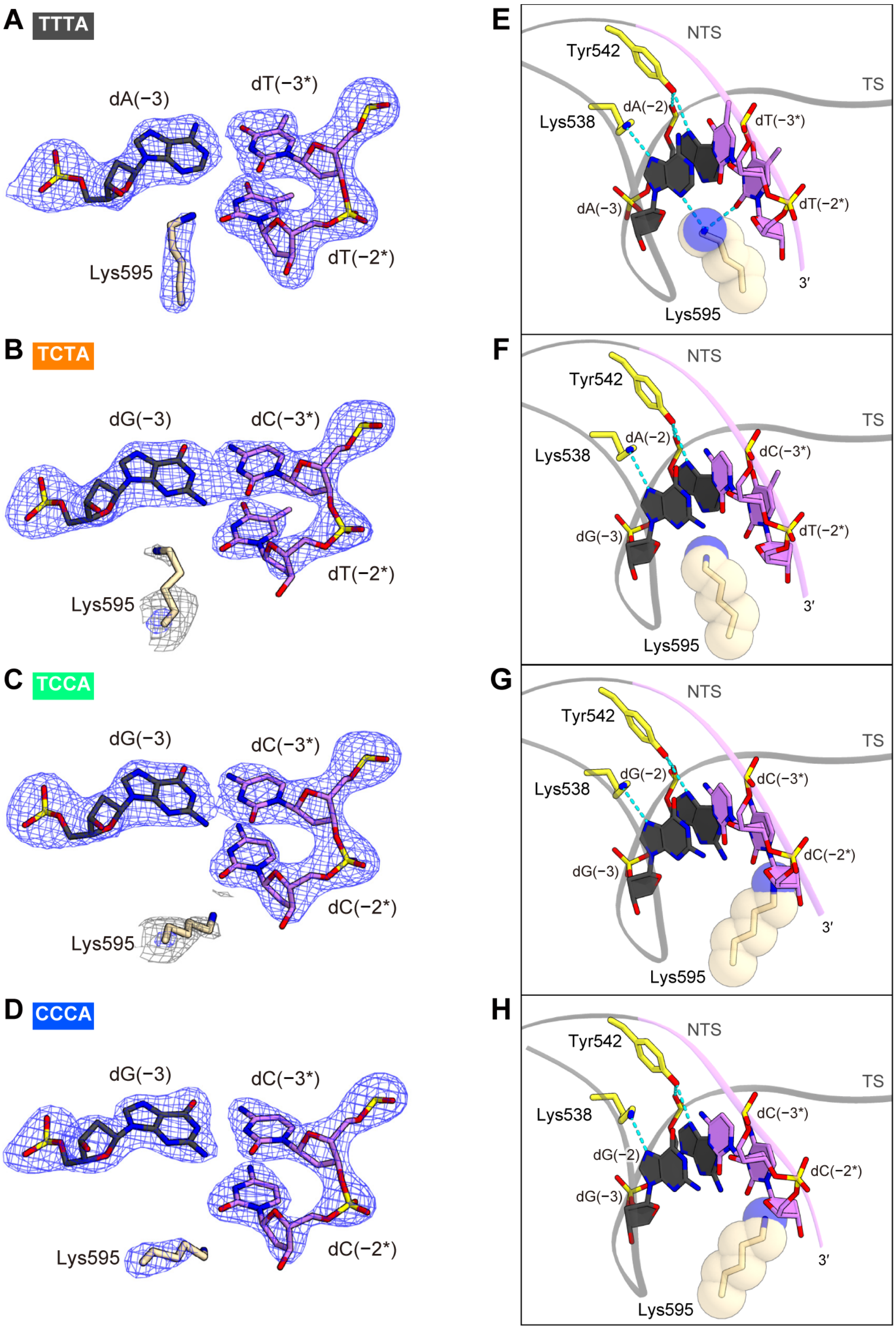


**Figure 4-20. Non-canonical PAM recognition mechanism**

(A–D) The *mF*o – *DF*c omit electron density maps for Lys595 and the key nucleotides in the TTTA (A), TCTA (B), TCCA (C) and CCCA (D) complexes (blue, contoured at 5.0σ). In (B) and (C), the electron density map contoured at 2.0σ (gray) is also shown for Lys595.

(E–H) Recognitions of the TTTA (E), TCTA (F), TCCA (G) and CCCA (H) PAMs. Hydrogen bonds are shown as dashed lines.

## 4.6 Discussion

In this study, we showed that LbCpf1 and AsCpf1 recognize TTTV and CTTV/TCTV/TTCV as the canonical and non-canonical PAMs, respectively, consistent with a recent study (Kim et al., 2016b). We further determined the LbCpf1–crRNA–DNA structures with either the TTTA, TCTA, TCCA or CCCA PAM. A structural comparison of LbCpf1 with AsCpf1 highlighted the mechanistic similarities in the crRNA-guided DNA recognition and cleavage among the Cpf1 nucleases. The TTTA complex structure revealed that LbCpf1 recognizes the canonical TTTV PAM *via* the shape and base readout mechanism, in which Lys595 inserts into the minor groove of the PAM duplex. Lys607 of AsCpf1, equivalent to Lys595 of LbCpf1, forms similar interactions with the PAM duplex (Yamano et al., 2016), and these lysine residues are conserved among the Cpf1 family members (Zetsche et al., 2015a) (Figure 4-13), suggesting that the Cpf1 orthologs recognize their T-rich PAMs in similar manners. Moreover, the present structures revealed that Lys595 is also important for the discrimination between the canonical and non-canonical PAMs.

The present structures also highlighted the fundamental differences in the PAM recognition mechanisms between the type-II Cas9 and type-V Cpf1 effector nucleases (Figures 4-21). In Cas9, the PAM duplex is accommodated within the PAM-binding groove in the PI domain, in which the major-groove edges of the PAM nucleotides are recognized by distinct sets of PAM-interacting residues *via* hydrogen-bonding interactions (*i.e.*, base readout mechanism) (Anders et al., 2014; Nishimasu et al., 2015; Hirano et al., 2016; Yamada et al., 2017). A structural comparison between the binary and ternary complexes of Cas9 suggested that the PAM-binding groove is pre-organized for the PAM recognition (Nishimasu et al., 2014; Anders et al., 2014; Jiang et al., 2015) (Figure 4-21B). In contrast, in Cpf1, the PAM duplex is enveloped within the PAM-binding channel formed by the WED, REC1 and PI domains, in which both the sequence and conformation of the PAM duplex are primarily recognized by the two conserved lysine residues (LbCpf1 Lys538/Lys595 and AsCpf1 Lys548/Lys607) (*i.e.*, base and shape readout mechanism). Importantly, the present LbCpf1 structures revealed that, in contrast to the PAM-binding groove of Cas9, the PAM-binding channel of Cpf1 has conformational flexibility, which allows the recognition of the canonical and non-canonical PAMs (Figure 4-21A).

The tolerant PAM recognition of Cpf1 is unfavorable for the gene editing, and it makes difficult to select target sites and design guide RNAs. However, it may confer some advantages in the role of physiological immune systems in prokaryote. There is an arms race between bacteria and phages. For phages, alteration of the PAM is one of the strategies to avoid the threat of CRISPR-Cas systems. There is a possibility that the tolerant PAM recognition may be valuable to defend the host cells form the phages with altered PAM.

Recently, AsCpf1 has been engineered to recognize altered PAM sequences, using a structure-guided mutation screen (Gao et al., 2017), and the crystal structures of the AsCpf1 variants revealed their altered PAM recognition mechanisms (Nishimasu et al., 2017). Thus, the present findings advance our mechanistic understanding of the CRISPR-Cpf1 family nucleases, and will facilitate the engineering of Cpf1, including the development of variants with altered PAM specificities.

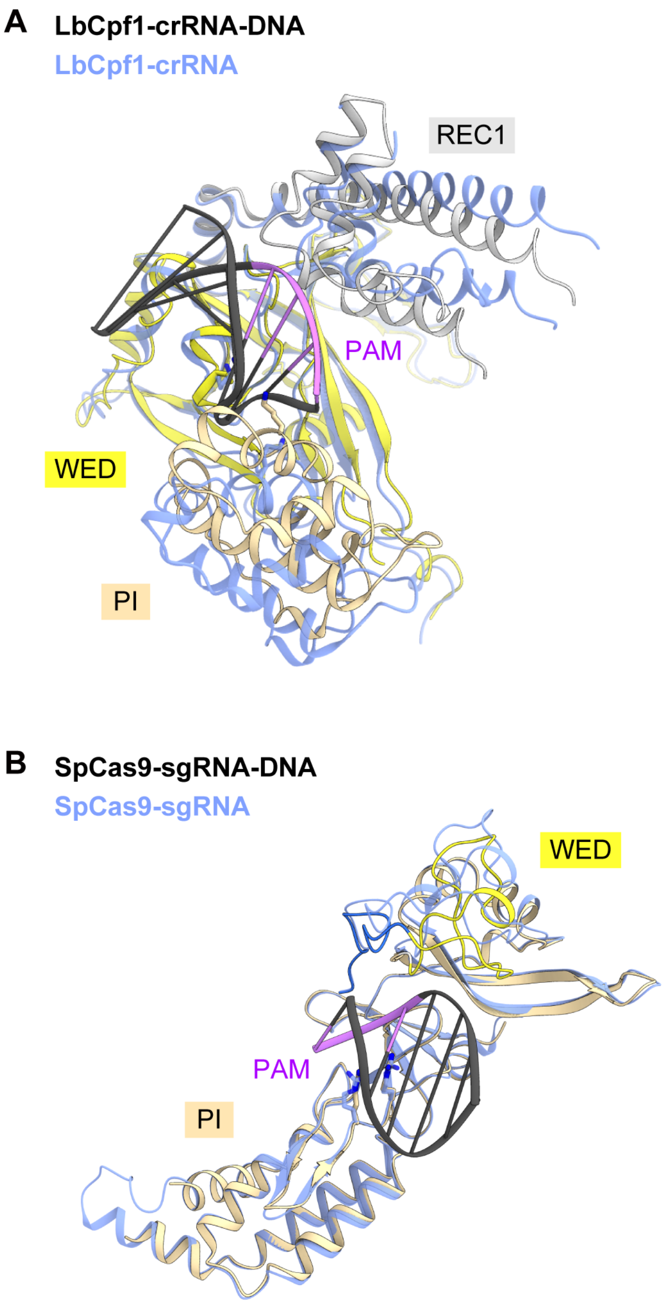


**Figure 4-21. Structural comparisons of PAM recognition channels between binary and ternary complexes**

(A) Superimposition of the LbCpf1-crRNA binary complex (blue) and the LbCpf1-crRNA-DNA ternary complex (colored).

(B) Superimposition of the SpCas9-sgRNA binary complex (blue) and the SpCas9-sgRNA-DNA ternary complex (colored).

# Chapter 5: General discussion

## 5.1 Conservations of the overall structure in Cpf1 family

In this study, we described the conservation of the bilobed structures among the Cpf1 family members by the structural comparisons between the ternary complexes of AsCpf1 and LbCpf1. Recently, the crystal structure of FnCpf1–crRNA binary complex was reported (Swarts et al., 2017). The amino acid sequence similarities between LbCpf1 and FnCpf1 are 40%, and the structural comparison of these binary complexes also reveals the structural similarities (Figures 5-1A and 5-1B). This observation supports the structural conservations among the Cpf1 family members, suggesting the conservations of the conformational changes accompanying the target DNA recognitions.

## 5.2 Structural insights into the target DNA cleavage mechanism of Cpf1

Dong et al. reported the crystal structure of LbCpf1–crRNA binary complex in 2.4 Å resolutions and the electron microscopy structure of apo-LbCpf1 and the binary complex, and Swarts et al. reported the crystal structures of FnCpf1–crRNA binary complex in 3.3 Å resolutions and FnCpf1–crRNA–target DNA R-loop complex in 2.5 Å resolutions. The R-loop complex contains a full-length double-stranded target DNA. The structural comparisons of these structures provide clues about the conformational changes of Cpf1 accompanying the crRNA and target DNA binding, and the target DNA cleavage mechanisms. The apo-Cpf1 structure showed by negative staining electron microscopy displayed an extended conformation, and the particles suggest the structural flexibility of the REC and NUC lobes. Both the crystal structure and the EM structure of LbCpf1–crRNA binary complexes display triangle shapes. The structural comparisons indicate the conformational changes from the extended shape to the triangle shape accompanying with crRNA binding. In the structure of LbCpf1–crRNA complex, the seed region of the crRNA is not ordered, whereas in FnCpf1–crRNA complex five nucleotides in the seed region are ordered and adopt an A-form-like helical conformation. The seed region is solvent exposed, and poised for hybridization with target DNA strand. However, the overall structures are identical each other

(Figures 5-1A and 5-1B), suggesting the recognition of the 5′-handle of crRNA, not the pre-ordering of the seed region, trigger the conformational change of Cpf1 and a formation of a preliminary central channel. After that, the target DNA binding induces the structural rearrangements of the Rec1 and Rec2 domains in the REC lobe, leading to form the central channel as described before (Figures 5-1C and 5-1D).

In the R-loop structure of FnCpf1, the unwound target dsDNA forms a PAM-distal DNA duplex between the Rec2 and Nuc domains (Figure 5-1E). Compared to the Cpf1–crRNA–target DNA complexes, the Nuc domain positioned further from the Rec2 domain to accommodate the PAM-distal DNA duplex. While we predicted that this R-loop structure mimic the physiological DNA cleaving state, the distance between the catalytic site of the RuvC domain and the cleavage site of dsDNA was too far. Further conformational changes of Cpf1 are suggested from this structure to cleave the target DNA. Cpf1 cleave the dsDNA by single nuclease domain, thus the cleavage mechanisms are different from Cas9. The cleaving order of the target and non-target DNA strand and the conformational changes accompanying the cleavages remain elusive. We predict that the other two intermediate states exist which the RuvC domain cleaves the target DNA strand and non-target DNA strand.

Recently, there were technical advances in the single particle analysis by cryo-electron microscopy (cryo-EM). Cryo-EM is a powerful tool to determine the structure of proteins which are not suitable for crystallization. The structure of SpCas9–sgRNA–tagret DNA complex containing full dsDNA was determined by cryo-EM, and it reveal further conformational changes of the HNH domain compared to the SpCas9 R-loop structure (Huai et al., 2017; Jiang et al., 2016a). Previous studies suggest that the crystallization of Cpf1–crRNA–target DNA complex containing full dsDNA is difficult. When the structure is determined by cryo-EM, there is a possibility that we can observe some unexpected conformation of Cpf1 and reveal the cleaving order of the target DNA.

## 5.3 The PAM recognition mechanisms of Cpf1

The crystal structure of FnCpf1 R-loop structure revealed the PAM recognition mechanism of FnCpf1. According to the previous reports, FnCpf1 recognizes TTV PAM (Zetsche et al., 2015a). The PAM duplex of FnCpf1 was highly distorted, and surrounded by the Rec1, WED, and PI domains (Figures 5-2A and 5-2B). The minor groove of the distorted duplex was narrower than the minor groove of the canonical B form dsDNA and the PAM duplex of AsCpf1 and LbCpf1 (Figure 5-2C). The dT(−1):dA(−1*) base pair in the PAM duplex does not form base-specific contacts with the protein. The O2 of dT(−2*) forms a hydrogen bond with Lys671. dA(−2) also forms van der Waals interactions with Pro663, Met668 and Lys671 (Figure 5-3A). The 5-methyl group of dT(−3*) is in the vicinity of the side-chain methyl group of Thr177, whereas the N3 and N7 of dA(−3) form hydrogen bonds with Lys671 and Lys613, respectively (Figure 5-3B). The 5-methyl group of dT(−4*) is surrounded by the side chains of Thr177, whereas the O4′ of dA(−4) forms a hydrogen bond with Lys671 (Figure 5-3C). As observed in AsCpf1 and LbCpf1 complexes, the PAM recognition mechanisms are conserved in FnCpf1, which is recognized by the combination of the shape readout and the base readout (Figures 5-3D–5-3I). The side chain of Lys671, which equivalent to Lys607 of AsCpf1 and Lys595 of LbCpf1, is inserted into the narrow minor groove and recognizes the distorted shape of the PAM duplex. Moreover, Lys671 and Lys613, which equivalent to Lys548 of AsCpf1 and Lys538 of LbCpf1, recognize the bases specifically.

FnCpf1 recognizes TTV PAM, which is simpler than TTTV PAM recognized by AsCpf1. Whereas the PAMs are recognized by FnCpf1 and AsCpf1 in same manners and all key residues and interactions are conserved among them (Figures 5-3A–5-3F), Thr604 in the WED domain, which equivalent to Thr539 of AsCpf1, positioned slight further from the methyl group of dT(−4*) compared to Thr539 (Figures 5-3C and 5-3F). Thr604 does not interact with dT(−4*), and this position makes space around the nucleotide. It eliminates the strong preference for a thymine and enables FnCpf1 to accept any nucleotides as this position. Furthermore, whereas dT(−4*) in AsCpf1 complex tilts to the complementary dA(−4) and form hydrogen bonds with dA(−4) and dA(−3), dT(−4*) and dA(−4) form canonical AT base pair in the FnCpf1 ternary complex (Figures 5-3C and 5-3F). It contributes the TTV PAM recognition mechanism of FnCpf1. LbCpf1 prefers the TTTV PAM. However, it also recognizes the C-containing PAM, especially the CTTV PAM as a second PAM. In the LbCpf1 ternary complex, dT(−4*) and dA(−4) form canonical AT base pair similar to the FnCpf1 ternary complex, and Gln529 in the WED domain equivalents to Thr539 of AsCpf1 and Thr604 of FnCpf1, forms van del Waals interactions with dT(−4*) (Figure

5-3I). When a purine nucleotide exists at the first position of the PAM, the steric hindrance is predicted between the nucleotide and Gln529, suggesting limits the selectivity of nucleotide for a pyrimidine at the first position of the PAM. These comparisons of PAM recognition mechanisms suggest the possibility of further engineering of Cpf1. When Gln529 of LbCpf1 is substituted to another amino acid residue with unbulky side chain, there is some possibility of changing the PAM preference of LbCpf1 from TTTV to TTV.

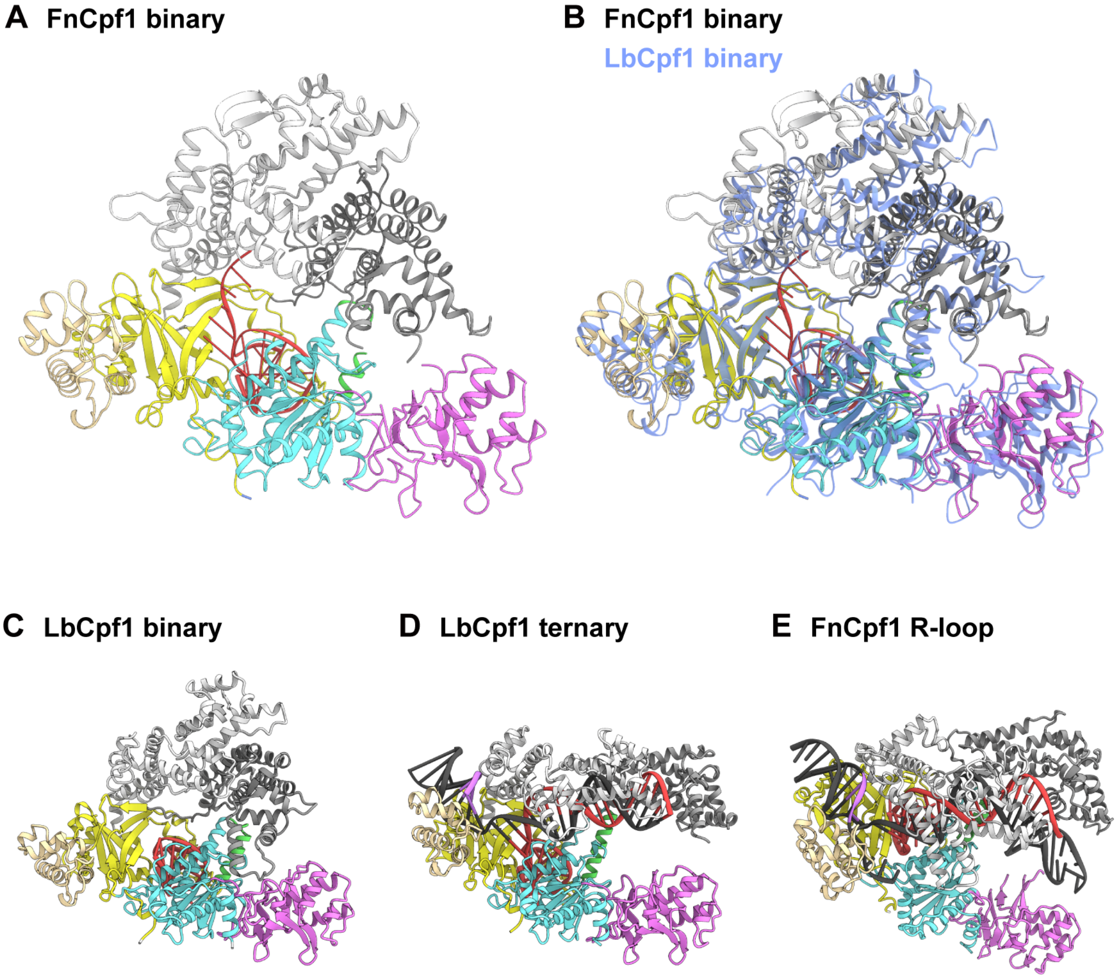


**Figure 5-1. Structural comparison between FnCpf1 and LbCpf1**
(A) Crystal structure of the FnCpf1–crRNA binary complex (PDB: 5NG6).
(B) Superimposition of the FnCpf1–crRNA binary complex(colored) and the LbCpf1–crRNA binary complex (PDB: 5ID6) (blue).
(C-E) Strucutural comparisons between the LbCpf1-binary complex, the LbCpf1-crRNA target DNA complex (5XUS) and the FnCpf1-crRNA target DNA R-loop complex (5NFV).

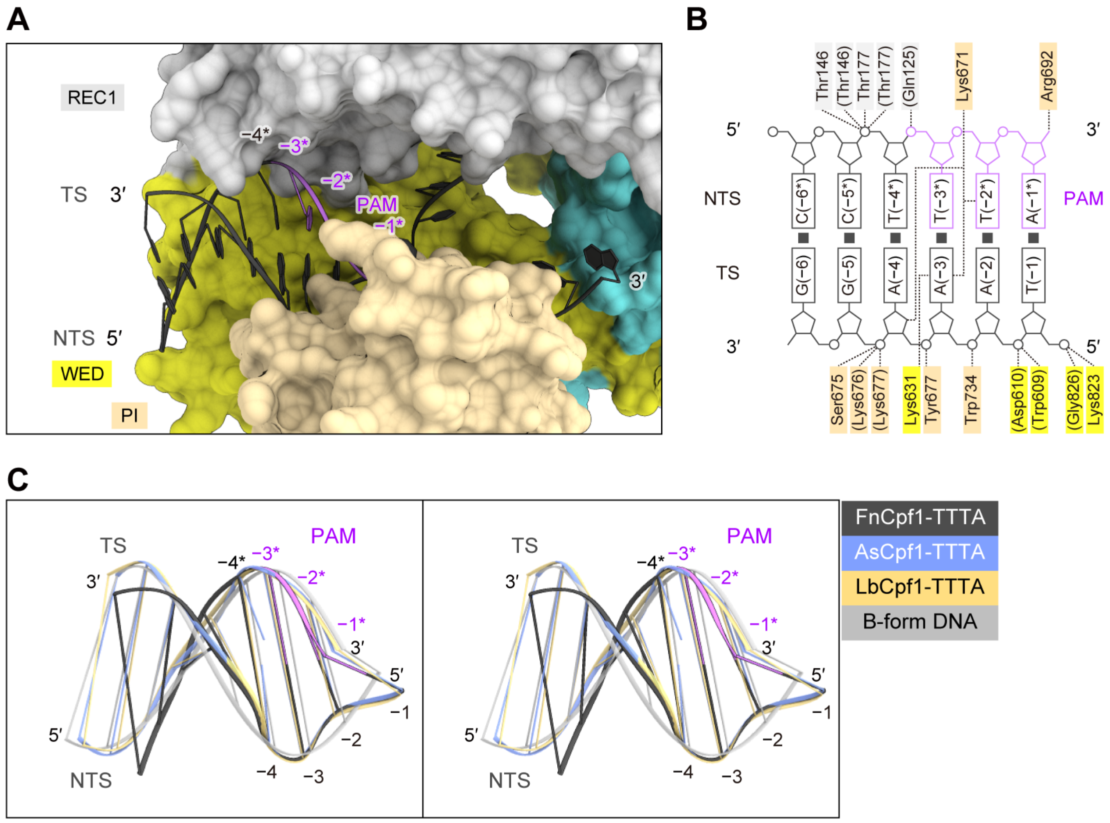


**Figure 5-2. PAM recognition mechanism of FnCpf1**

(A) PAM-duplex binding in the FnCpf1 R-loop complex.

(B) Schematic of the PAM-duplex recognition by FnCpf1.

(C) Superimposition of the PAM duplexes in FnCpf1 (5NFV), LbCpf1 (PDB: 5XUS) and AsCpf1 (PDB: 5B43) onto the B-form DNA duplex (stereo view).

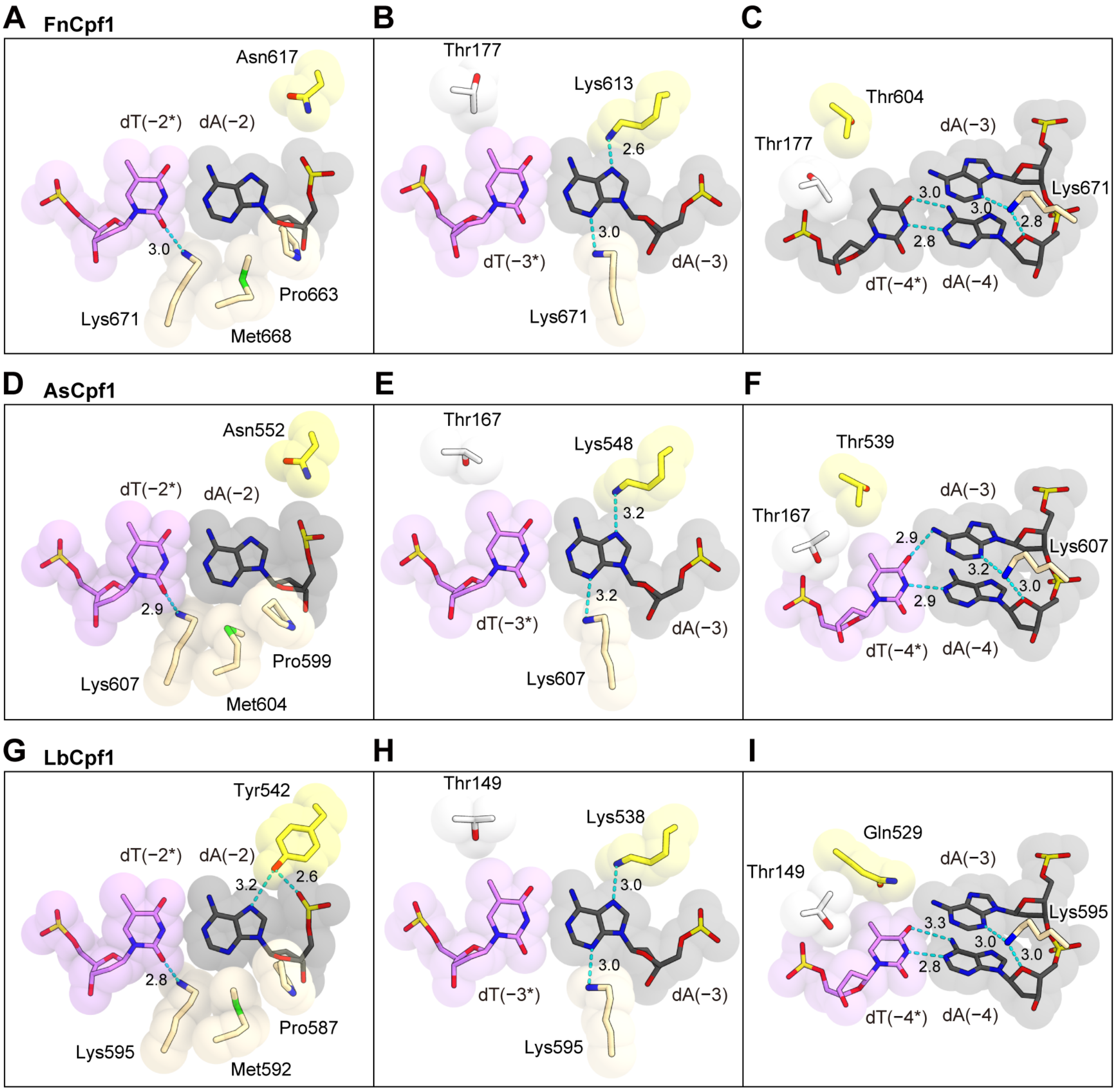


**Figure 5-3. Comparison between the PAM recognition mechanisms of FnCpf1, AsCpf1 and LbCpf1**

(A–C) Recognition of dA(−2):dT(−2*) (A), dA(−3):dT(−3*) (B), and dA(−4):dT(−4*) (C) by FnCpf1 (PDB: 5NFV).

(D–F) Recognition of dA(−2):dT(−2*) (D), dA(−3):dT(−3*) (E), and dA(−4):dT(−4*) (F) by AsCpf1 (PDB: 5B43).

(G–I) Recognition of dA(−2):dT(−2*) (G), dA(−3):dT(−3*) (H), and dA(−4):dT(−4*) (I) by LbCpf1 (5XUS).

In (A–I), hydrogen bonds are shown as dashed lines.

# References

Abudayyeh, O.O., Gootenberg, J.S., Konermann, S., Joung, J., Slaymaker, I.M., Cox, D.B., Shmakov, S., Makarova, K.S., Semenova, E., Minakhin, L., et al. (2016). C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. Science *5573* 1–9.

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. Sect. 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.

Anders, C., Bargsten, K., and Jinek, M. (2016). Structural Plasticity of PAM Recognition by Engineered Variants of the RNA-Guided Endonuclease Cas9. Mol. Cell *61*, 895–902.

Barrangou, R., and Doudna, J.A. (2016). Applications of CRISPR technologies in research and beyond. Nat. Biotechnol. *34*, 933–941.

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.

Brouns, S.J., Hayday, A.C., Volkmann, A., Raulet, D.H., Knowles, G.C., Wiest, D.L., Beermann, F., Clevers, H., Held, W., Mccormick, F., et al. (2008). Small CRISPR RNAs guide antiviral defence in prokaryotes. Science *321*, 960–964.

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

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 System. Science *339*, 819–824.

Cowtan, K. (2006). The Buccaneer software for automated model building. 1. Tracing protein chains. Acta Crystallogr. Sect. 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.

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

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.

Doudna, J.A., and Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. Science *346*, 1258096–1258096.

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

Emsley, P., and Cowtan, K. (2004). Coot: Model-building tools for molecular graphics. Acta Crystallogr. Sect. 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 ils restriction enzymes. PLoS One *4*, 1–9.

Evans, P.R., and Murshudov, G.N. (2013). How good are my data and what is the resolution? Acta Crystallogr. Sect. 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.

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.

Fu, Y., Foden, J.A., Khayter, C., Maeder, M.L., Reyon, D., Joung, J.K., and Sander J.D. (2013). High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nat. Biotechnol. *31*, 822–826.

Gao, L., Cox, D.B.T., Yan, W.X., Manteiga, J.C., Schneider, M.W., Yamano, T., Nishimasu, H., Nureki, O., Crosetto, N., and Zhang, F. (2017). Engineered Cpf1 variants with altered PAM specificities. Nat. Biotechnol. *35*, 789–792.

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

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

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. *109*, E2579-E2586.

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.

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. (2016a). Structure and Engineering of Francisella novicida Cas9. Cell *164*, 950–961.

Hirano, S., Nishimasu, H., Ishitani, R., and Nureki, O. (2016b). Structural Basis for the Altered PAM Specificities of Engineered CRISPR-Cas9. Mol. Cell *61*, 886–894.

Holm, L., and Rosenström, P. (2010). Dali server: Conservation mapping in 3D. Nucleic Acids Res. *38*, 545–549.

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.

Huai, C., Li, G., Yao, R., Zhang, Y., Cao, M., Kong, L., Jia, C., Yuan, H., Chen, H., Lu, D., et al. (2017). Structural insights into DNA cleavage activation of CRISPR-Cas9 system. Nat. Commun. *8*, 1375, 1–9.

Hur, J.K., Kim, K., Been, K.W., Baek, G., Ye, S., Hur, J.W., Ryu, S.-M., Lee, Y.S., and Kim, J.-S. (2016). Targeted mutagenesis in mice by electroporation of Cpf1 ribonucleoproteins. Nat. Biotechnol. *34*, 807–808.

Hwang, W.Y.. Fu, Y., Reyon, D., Maeder, M.L., Tsai, S.Q., Sander, J.D., Peterson, R.T., Yeh, J.R., and Joung, J.K. (2013). Efficient genome editing in zebrafish using a CRISPR-Cas system. Nat. Biotechnol. *31*, 227–229.

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.

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

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

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. (2016a). Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. Science *351*, 867–871.

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.

Jiang, W., Samai, P., and Marraffini, L.A. (2016b). Degradation of Phage Transcripts by CRISPR-Associated RNases Enables Type III CRISPR-Cas Immunity. Cell *164*, 710–721.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A Programmable Dual-RNA – Guided DNA Endonuclease in Adaptice Bacterial Immunity. Science *337*, 816–822.

Jinek, M., East, A., Cheng, A., Lin, S., Ma, E., and Doudna, J. (2013). RNA-programmed genome editing in human cells. Elife *2013*, 1–9.

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.

Jore, M.M., Lundgren, M., van Duijn, E., Bultema, J.B., Westra, E.R., Waghmare, S.P., Wiedenheft, B., Pul, Ü., Wurm, R., Wagner, R., et al. (2011). Structural basis for CRISPR RNA-guided DNA recognition by Cascade. Nat. Struct. Mol. Biol. *18*, 529–536.

Joung, J.K., and Sander, J.D. (2013). TALENs: a widely applicable technology for targeted genome editing. Nat. Rev. Mol. Cell Biol. *14*, 49–55.

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.

Kim, D., Kim, J., Hur, J., Been, K.W., Yoon, S., and Kim, J.S. (2016a). Genome-wide target specificities of Cpf1 nucleases in human cells. Nat. Biotechnol. *34*, 863–868.

Kim, H.K., Song, M., Lee, J., Menon, A.V., Jung, S., Kang, Y.M., Choi, J.W., Woo, E., Koh, H.C., Nam, J.W., et al. (2016b). In vivo high-throughput profiling of CRISPR–Cpf1 activity. Nat. Methods. *14*, 153–159.

Kim, Y., Cheong, S.A., Lee, J.G., Lee, S.W., Lee, M.S., Baek, I.J., and Sung, Y.H.

(2016c). Generation of knockout mice by Cpf1-mediated gene targeting. Nat. Biotechnol. *34*, 808–810.

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. (2015a). Engineered CRISPR-Cas9 nucleases with altered PAM specificities. Nature *523*, 481–485.

Kleinstiver, B.P., Prew, M.S., Tsai, S.Q., Nguyen, N.T., Topkar, V. V, Zheng, Z., and Joung, J.K. (2015b). Broadening the targeting range of Staphylococcus aureus CRISPR-Cas9 by modifying PAM recognition. Nat. Biotechnol. *33*, 1–7.

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

Kleinstiver, B.P., Tsai, S.Q., Prew, M.S., Nguyen, N.T., Welch, M.M., Lopez, J.M., McCaw, Z.R., Aryee, M.J., and Joung, J.K. (2016b). Genome-wide specificity of CRISPR-Cas Cpf1 nucleases in human cells. Nat. Biotechnol. *34*, 869–874.

Knott, G.J., East-Seletsky, A., Cofsky, J.C., Holton, J.M., Charles, E., O'Connell, M.R., and Doudna, J.A. (2017). Guide-bound structures of an RNA-targeting A-cleaving CRISPR–Cas13a enzyme. Nat. Struct. Mol. Biol. *24*, 825–833.

Komor, A.C., Badran, A.H., and Liu, D.R. (2016a). CRISPR-Based Technologies for the Manipulation of Eukaryotic Genomes. Cell *168*, 1–17.

Komor, A.C., Kim, Y.B., Packer, M.S., Zuris, J.A., and Liu, D.R. (2016b). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature *533*, 420–424.

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. (2014). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex.

Nature *517*, 583–588.

Liu, L., Chen, P., Wang, M., Li, X., Wang, J., Yin, M., and Wang, Y. (2016). C2c1-sgRNA Complex Structure Reveals RNA-Guided DNA Cleavage Mechanism. Mol. Cell *65*, 310–322.

Liu, L., Li, X., Wang, J., Wang, M., Chen, P., Yin, M., Li, J., Sheng, G., and Wang, Y. (2017a). Two Distant Catalytic Sites Are Responsible for C2c2 RNase Activities. Cell *168*, 121–134.e12.

Liu, L., Li, X., Ma, J., Li, Z., You, L., Wang, J., Wang, M., Zhang, X., and Wang, Y. (2017b). The Molecular Architecture for RNA-Guided RNA Cleavage by Cas13a. Cell *170*, 714–726.e10.

Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J., Barrangou, R., Brouns, S.J.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.

Marraffini, L.A., and Sontheimer, E.J. (2008). CRISPR Interference Limits Horizontal Gene Transfer in Staphylococci by Targeting DNA. Science *322*, 1843–1845.

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

Mohanraju, P., Makarova, K.S., Zetsche, B., Zhang, F., Koonin, E. V, and Van der

Oost, J. (2016). Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. Science *353*, aad5147.

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.

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

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

Nihongaki, Y., Kawano, F., Nakajima, T., and Sato, M. (2015). Photoactivatable CRISPR-Cas9 for optogenetic genome editing. Nat. Biotechnol. *9*, 1–8.

Nishida Keiji , Takayuki Arazoe, Nozomu Yachie, Satomi Banno, Mika Kakimoto, Mayura Tabata, Masao Mochizuki, Aya Miyabe, Michihiro Araki, K.Y.H. (2016). Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. Science *8729*.

Nishimasu, H., and Nureki, O. (2017). Structures and mechanisms of CRISPR RNA-guided effector nucleases. Curr. Opin. Struct. Biol. *43*, 68–78.

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.

Nishimasu, H., Yamano, T., Gao, L., Zhang, F., Ishitani, R., and Nureki, O. (2017). Structural Basis for the Altered PAM Recognition by Engineered CRISPR-Cpf1. Mol. Cell *67*, 139–147.e2.

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.

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

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

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

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.

Samai, P., Pyenson, N., Jiang, W., Goldberg, G.W., Hatoum-Aslan, A., and Marraffini, L.A. (2015). Co-transcriptional DNA and RNA cleavage during type III CRISPR-cas immunity. Cell *161*, 1164–1174.

Sashital, D.G., Wiedenheft, B., and Doudna, J.A. (2012). Mechanism of Foreign DNA Selection in a Bacterial Adaptive Immune System. Mol. Cell *46*, 606–615.

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.

Shmakov, S., Abudayyeh, O.O., Makarova, K.S., Wolf, Y.I., Gootenberg, J.S., Semenova, E., Minakhin, L., Joung, J., Konermann, S., Severinov, K., et al. (2015). Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. Mol. Cell *60*, 385–397.

Shmakov, S., Smargon, A., Scott, D., Cox, D., Pyzocha, N., Yan, W., Abudayyeh, O.O., Gootenberg, J.S., Makarova, K.S., Wolf, Y.I., et al. (2017). Diversity and evolution of class 2 CRISPR–Cas systems. Nat. Rev. Microbiol. *15*, 169–182.

Slaymaker, I.M., Gao, L., Zetsche, B., Scott, D.A., Yan, W.X., and Zhang, F. (2015). 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*, 244–248.

Sorek, R., Lawrence, C.M., and Wiedenheft, B. (2013). CRISPR-Mediated Adaptive Immune Systems in Bacteria and Archaea. Annu. Rev. Biochem. *82*, 237–266.

Stella, S., Cascio, D., and Johnson, R.C. (2010). The shape of the DNA minor groove directs binding by the DNA-bending protein Fis. Genes Dev. *24*, 814–826.

Sternberg, S.H., LaFrance, B., Kaplan, M., and Doudna, J.A. (2015). Conformational control of DNA target cleavage by CRISPR-Cas9. Nature *527*, 1–14.

Swarts, D.C., van der Oost, J., and Jinek, M. (2017). Structural Basis for Guide RNA Processing and Seed-Dependent DNA Targeting by CRISPR-Cas12a. Mol. Cell *66*, 221–233.e4.

Tang, T.H., Bachellerie, J.P., Rozhdestvensky, T., Bortolin, M.L., Huber, H., Drungowski, M., Elge, T., Brosius, J., and Huttenhofer, A. (2002). Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus. Proc. Natl. Acad. Sci. *99*, 7536–7541.

Tang, X., Lowder, L.G., Zhang, T., Malzahn, A.A., Zheng, X., Voytas, D.F., Zhong, Z., Chen, Y., Ren, Q., Li, Q., et al. (2017). A CRISPR–Cpf1 system for efficient genome editing and transcriptional repression in plants. Nat. Plants *3*, 17018.

Taylor, D.W., Zhu, Y., Staals, R.H., Kornfeld, J.E., Shinkai, A., van der Oost, J., Nogales, E., and Doudna, J.A. (2015). Structures of the CRISPR-Cmr complex reveal mode of RNA target positioning. Science *348*, 581–585.

Urnov, F.D., Rebar, E.J., Holmes, M.C., Zhang, H.S., and Gregory P.D. (2010). Genome editing with engineered zinc finger nucleases. Nat. Rev. Genet. *11*, 636–646.

Vagin, A., and Teplyakov, A. (2010). Molecular replacement with MOLREP. Acta Crystallogr. Sect. D Biol. Crystallogr. *66*, 22–25.

Wang, J., Li, J., Zhao, H., Sheng, G., Wang, M., Yin, M., and Wang, Y. (2015). Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems. Cell *163*, 840–853.

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 intefration software. CCP4 Newsletter *49*, 16–19.

Wei, Y., Terns, R.M., Terns, M.P., Terns, M.P., and Terns, M.P. (2015). Cas9 function and host genome sampling in type II-A CRISPR–cas adaptation. Genes Dev. *29*, 356–361.

Westra, E.R., Semenova, E., Datsenko, K.A., Jackson, R.N., Wiedenheft, B., Severinov, K., and Brouns, S.J.J. (2013). Type I-E CRISPR-Cas Systems Discriminate Target from Non-Target DNA through Base Pairing-Independent PAM Recognition. PLoS Genet. *9*.

Wiedenheft, B., Lander, G.C., Zhou, K., Jore, M.M., Brouns, S.J.J., van der Oost, J., Doudna, J.A., and Nogales, E. (2011). Structures of the RNA-guided surveillance complex from a bacterial immune system. Nature *477*, 486–489.

Wright, A. V., Sternberg, S.H., Taylor, D.W., Staahl, B.T., Bardales, J. a., Kornfeld, J.E., and Doudna, J.A. (2015). Rational design of a split-Cas9 enzyme complex. Proc. Natl. Acad. Sci. 201501698.

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.

Wright, A. V., Liu, J.J., Knott, G.J., Doxzen, K.W., Nogales, E., and Doudna, J.A. (2017). Structures of the CRISPR genome integration complex. Science *357*, 1113–1118.

Xiao, Y., Ng, S., Hyun Nam, K., and Ke, A. (2017). How type II CRISPR-Cas establish immunity through Cas1-Cas2-mediated spacer integration. Nature *550*, 137–141.

Yamada, M., Watanabe, Y., Gootenberg, J.S., Hirano, H., Ran, F.A., Nakane, T., Ishitani, R., Zhang, F., Nishimasu, H., and Nureki, O. (2017). Crystal Structure of the Minimal Cas9 from Campylobacter jejuni Reveals the Molecular Diversity in the CRISPR-Cas9 Systems. Mol. Cell *65*, 1109–1121.e3.

Yamano, T., Nishimasu, H., Zetsche, B., Hirano, H., Slaymaker, I.M., Li, Y., Fedorova, I., Nakane, T., Makarova, K.S., Koonin, E. V., et al. (2016). Crystal Structure of Cpf1 in Complex with Guide RNA and Target DNA. Cell *165*, 949–962.

Yang, H., Gao, P., Rajashankar, K.R., and Patel, D.J. (2016). PAM-Dependent Target DNA Recognition and Cleavage by C2c1 CRISPR-Cas Endonuclease. Cell *167*, 1814–1828.e12.

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. (2015a). Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. Cell *163*, 759–771.

Zetsche, B., Volz, S.E., and Zhang, F. (2015b). A split-Cas9 architecture for inducible genome editing and transcription modulation. Nat. Biotechnol. *33*, 139–142.

Zetsche, B., Heidenreich, M., Mohanraju, P., Fedorova, I., Kneppers, J., DeGennaro, E.M., Winblad, N., Choudhury, S.R., Abudayyeh, O.O., Gootenberg, J.S., et al. (2016). Multiplex gene editing by CRISPR–Cpf1 using a single crRNA array. Nat. Biotechnol. *35*, 31–34.

Zhao, H., Sheng, G., Wang, J., Wang, M., Bunkoczi, G., Gong, W., Wei, Z., and Wang, Y. (2014). Crystal structure of the RNA-guided immune surveillance Cascade complex in Escherichia coli. Nature *515*, 147–150.

# Original papers

*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
"Crystal structure of Cpf1 in complex with guide RNA and target DNA"
***Cell***, 165, 949-962, April 21, 2016
*equally contribution

Linyi Gao, David B T Cox, Winston X Yan, John C Manteiga, Martin W Schneider, Takashi Yamano, Hiroshi Nishimasu, Osamu Nureki, Nicola Crosetto, and Feng Zhang,
"Engineered Cpf1 variants with altered PAM specificities"
***Nature Biotechnology***, 35, 789-792, June 5, 2017

*Hiroshi Nishimasu, *Takashi Yamano, Linyi Gao, Feng Zhang, Ryuichiro Ishitani, and Osamu Nureki
"Stuructural basis for the altered PAM recognition by engineered CRISPR-Cpf1"
***Molecular Cell***, 67, 139-147, June 5, 2017
*equally contribution

Takashi Yamano, Bernd Zetsche, Ryuichiro Ishitani, Feng Zhang, Hiroshi Nishimasu, and Osamu Nureki
"Structural basis for the canonical and non-canonical PAM recognition by CRISPR-Cpf1"
***Molecular Cell***, 67, 633-645, August 17, 2017

# Acknowledgements

First of all, the author would like to express my greatest appreciation to Professor Osamu Nureki for his supervision during the Ph.D. course. Besides Professor Nureki, the author would like to thank Dr. Ryuichiro Ishitani, Dr. Hiroshi Nishimasu, Dr. Tomohiro Nishizawa, and Dr. Takanori Nakane for technical advices and experimental supports, and Ms. Rieko Yamazaki for secretarial assistances.

This study was performed in collaboration with many scientists. The author is deeply grateful to Dr. Feng Zhang, Mr. Bernd Zetsche, Dr. Ian M. Slaymaker, Dr. Yinqing Li, Ms. Iana Fedorova and Mr. Linyi Gao in Massachusetts Institute of Technology (MIT) and Broad Institute of MIT and Harvard for providing the original vectors of Cpf1, performing the functional analysis of Cpf1, and sharing a lot of beautiful and critical data. The author received generous supports from Dr. Eugene V. Koonin and Dr. Kira S. Makarova in National Center for Biotechnology Information (NCBI) in National Institute of Health (NIH) for writing the papers and improving the manuscripts from the view points of the evolutionary biology. The author owes the deepest gratitude to Dr. Hiroshi Nishimasu for teaching all of the basic techniques and knowledge necessary for crystallography and structural biology, and instruction to write the sophisticated papers. Dr. Ishitani and Dr. Nakane supported the author for the structure determinations and model buildings. Mr. Hisato Hirano gave the author critical comments to improve the manuscripts. The author would like to thank all members in Nureki lab for having a wonderful day together.

Finally, the author would like to thank his family for continuous supports through his life.

Takashi Yamano