Doctoral Dissertation

博士論文

Crystal structures of CRISPR-Cas9 and RNA methyltransferase
（CRISPR-Cas9 および RNA メチル化酵素の結晶構造解析）

A Dissertation Submitted for the Degree of Doctor of Philosophy
December 2019

令和元年１２月（理学）申請

Department of Biological Sciences, Graduate School of Science,
The University of Tokyo

東京大学大学院理学系研究科

生物科学専攻

Seiichi Hirano

平野 清一

Contents

Abstract

The RNA-guided DNA endonuclease Cas9 cleaves double-stranded DNA targets bearing a protospacer adjacent motif (PAM) and complementarity to a single-guide RNA (sgRNA). The widely-used *Streptococcus pyogenes* Cas9 (SpCas9) recognizes the NGG PAM which constrains the targetable sites in genome editing applications. Recently, SpCas9 has been engineered by directed evolution to exhibit altered PAM specificities. The VQR, EQR, and VRER mutants recognize the NGA, NGAG, and NGCG PAMs, respectively. However, the PAM recognition mechanisms of the SpCas9 mutants remain unknown. Here, I determined the crystal structures of the VQR, EQR and VRER mutants in complexes with their sgRNAs and DNA targets. A structural comparison of the SpCas9 mutants with wild-type SpCas9 revealed that, in the mutants, the multiple mutations induce an conformational change in the sugar-phosphate backbone of the DNA duplex, thereby enabling the altered PAM recognition. The structural findings clarify the mechanisms of the altered PAM recognition by the SpCas9 mutants, and provide a structural basis for further Cas9 engineering.

A variety of Cas9 orthologs have been identified and functionally characterized. Previous structural studies revealed that the Cas9 orthologs recognize their cognate sgRNAs and PAMs in distinct manners. *Corynebacterium diphtheriae* Cas9 (CdCas9) recognizes the NNRHHHY PAM, which is promiscuous as compared with the PAMs for other Cas9 orthologs, such as the NGG PAM for SpCas9. However, the mechanism of the promiscuous PAM recognition by CdCas9 remains unknown. Here, I determined the crystal structure of CdCas9 in complex with its sgRNA and target DNA, at 2.9 Å resolution. The biochemical and structural analyses revealed that CdCas9 recognizes the promiscuous

A-rich PAM, via a combination of hydrogen-bonding and van der Waals interactions. The structural findings provide insights into the mechanistic diversity in the PAM recognition by Cas9 enzymes.

N^6-methyadenosine (m^{6}A) modifications in eukaryotic mRNAs are associated with the fate of mRNAs in cells. The m^{6}A modification is related to the writer, reader, and eraser proteins, as a reversible and functional mark of mRNA. A newly identified writer, cap-specific adenosine *N*^{6}-methyltransferase (CAPAM), mediates the *N*^{6}-methylation of m^{6}A$_m$ in the 5′ cap of eukaryotic mRNAs. However, the molecular mechanism of the cap-specific *N*^{6}-methylation by CAPAM is unknown. Here, I determined the crystal structure of CAPAM in complex with m^{7}G-capped RNA and a cofactor analog, *S*-adenosyl homocysteine, at 2.0 Å resolution. The structure revealed that the core region of CAPAM consists of a canonical methyltransferase domain and a novel helical domain, thereby forming a positively charged groove, which could bind m^{7}G-capped RNA. Structural comparison of CAPAM with other m^{6}A methyltransferases, METTL3-METTL14 and METTL16, suggested that CAPAM catalyzes the *N*^{6}-methylation through a conserved mechanism, but recognizes its RNA substrate in a manner different from those of the other m^{6}A writers. Taken together, the structural findings provide mechanistic insights into the eukaryotic mRNA modification.

Abbreviations

Nucleic Acids

Abbreviation	Formal Name
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
N	Any
R	Adenine or Guanine
Y	Cytosine or Thymine
H	Adenine, Thymine, or Cytosine
B	Thymine, Guanine, or Cytosine

Other Abbreviations

Abbreviation	Formal Name
DNA	Deoxyribonucleic acid
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
HEPES	4-(2-Hydroxyethyl) piperazine-1-ethanesulfonic acid
GST	Glutathione *S*-transferase
PAGE	Polyaclylamidegel electrophoresis
PCR	Polymerase chain reaction
PEG	Polyethylene glycol
RNA	Ribonucleic acid
SDS	Sodium dodecylsulfate
SUMO	Small Ubiquitin-related(like) modifier
Tris	Tris(hydroxymethyl)aminomethane
TEV	Tobacco etch virus

Chapter 1: General Introduction

CRISPR-Cas system

CRISPR (clustered regularly interspaced short palindromic repeat)-Cas (CRISPR-associated) system is an adaptive immune system which is conserved in almost all archaea and one-third of bacteria[1]. The CRISPR-Cas system protects the bacterial and archaeal hosts from phage and plasmid. Once the host cell is infected with phage and plasmid, Cas1 (CRISPR-associated protein 1)-Cas2 (CRISPR-associated protein 2) complex binds the viral DNA from phage and plasmid through the recognition of specific sequence[2] (Fig. 1). The Cas1-Cas2 complex cleaves the viral DNA at the site of specific sequence and obtains the protospacer DNA[2] (Fig. 1). Thus, the specific sequence is designated as PAM (protospacer adjacent motif). The Cas1-Cas2 complex integrates the protospacer DNA into the CRISPR array consisting of repeat and spacer regions[3,4] (Fig. 1). The CRISPR array is transcribed and processed into crRNA (CRISPR RNA) (Fig. 1). While the type I, III, and IV CRISPR-Cas systems utilize crRNA in complex with multiple effector proteins, the type II, V, and VI CRISPR-Cas systems utilize crRNA in complex with single effector proteins[5]. In the type II CRISPR-Cas system, which is most widely used for genome editing, Cas9 (CRISPR-associated protein 9) in complex with crRNA and tracrRNA (*trans*-activating crRNA) targets the viral DNA[6,7] (Fig. 1). Cas9 selectively targets the protospacer sequence with the PAM in the viral DNA, but not the spacer sequence of the CRISPR array in the host DNA, due to the absence of the PAM in the spacer-flanking repeat region[8] (Fig. 1). Thus, the PAM is utilized for self and non-self discrimination in the adaptive immune system.

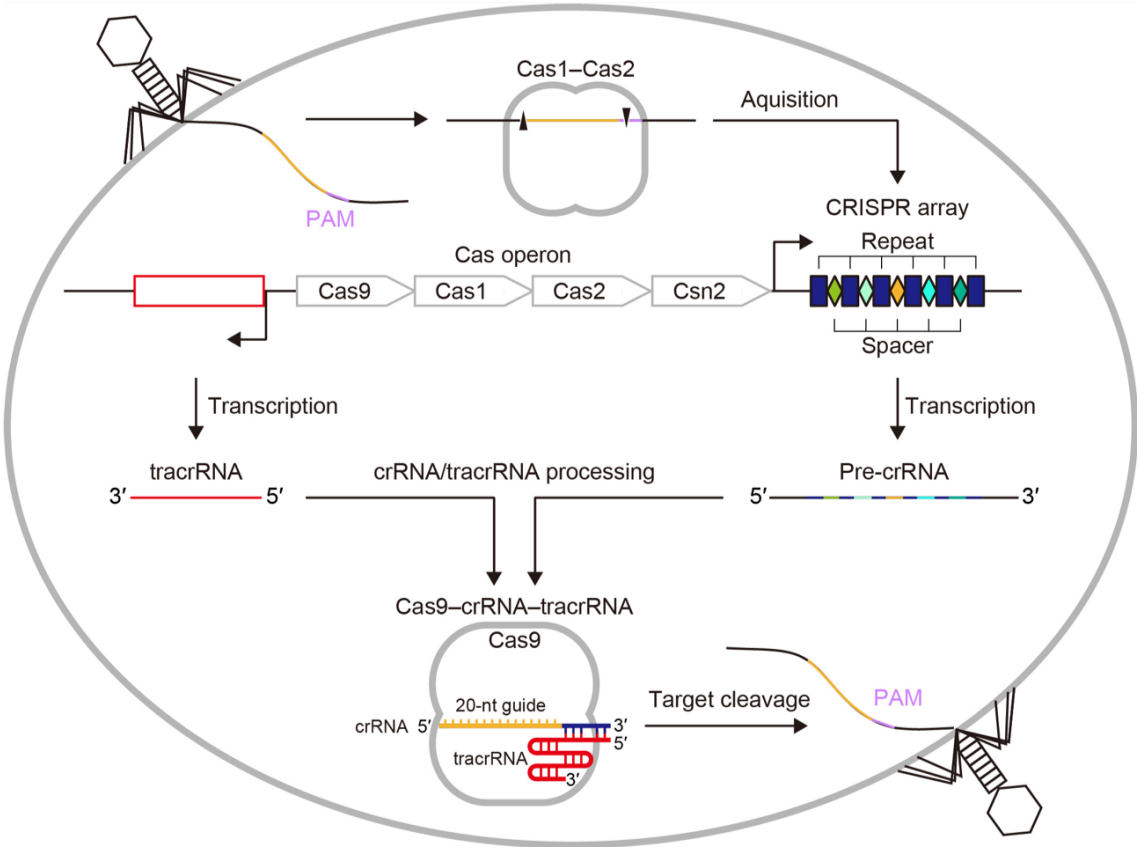

Fig. 1. The type II CRISPR-Cas system.

Cas9

Cas9s are identified in more than 1000 bacterial species[9]. Two DNA endonuclease domains, which are designated as the RuvC and HNH domains, are conserved in Cas9s[10,11]. Other regions except for the RuvC and HNH domains are highly divergent among Cas9s[10,11]. The sequence divergence of Cas9s is related to the sequence specificity of their cognate crRNAs and tracrRNAs[11,12]. In the biochemical studies, Cas9 associates with dual guide RNAs (crRNA and tracrRNA) or a synthetic single-guide RNA (sgRNA) and cleaves double-stranded DNA targets complementary to the crRNA[13,14] (Fig. 2). Besides the crRNA-target DNA complementarity, DNA recognition by Cas9 requires the PAM immediately downstream of the target DNA sequence[15,16] (Fig. 2). The RNA-guided DNA endonuclease Cas9 is repurposed to genome editing[17]. The human genome contains 3 billion base pairs which are less than 4^{16} possible combination of 16 A:T/T:A/G:C/C:G base pairs. Assuming that the human genome does not contain repetitive sequences, 16-bp target search can find one specific site in the human genome. The widely used Cas9 from *Streptococcus pyogenes* Cas9 (SpCas9) recognizes 20-bp target and 2-bp PAM[13,14]. The two-component system consisting of SpCas9 and its sgRNA can target endogenous genomic sites in a wide range of cell types and organisms, and thus has been applied to genome editing[18,19]. SpCas9 recognizes an NGG sequence as the PAM, thus constraining the targetable sites in genome editing applications[20]. Numerous Cas9s and their cognate sgRNAs are characterized for genome editing in eukaryote cells[18,19,21–26] (Table 1). These Cas9s recognize the GC-contained PAM, thus cannot target the AT-rich regions in genome editing.

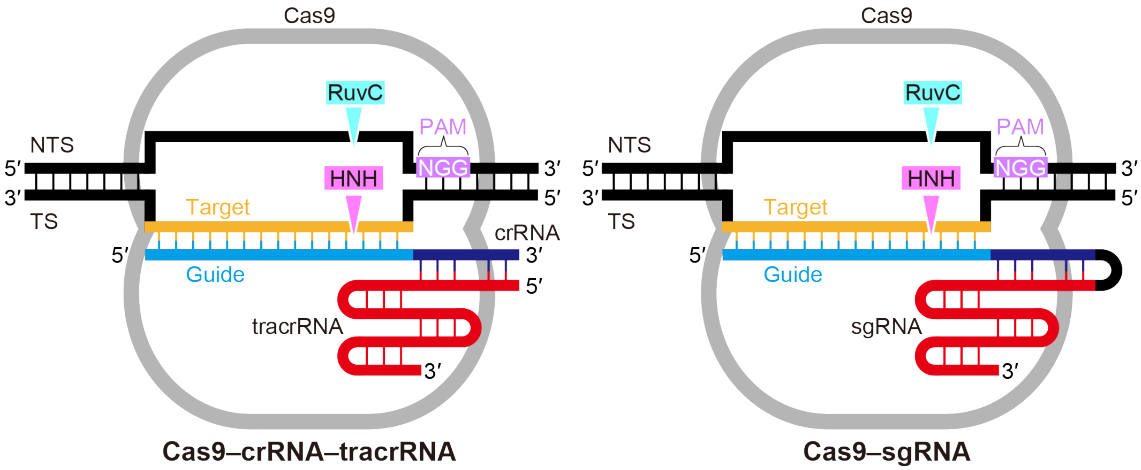

Fig. 2. Schematics of RNA-guided DNA cleavage by Cas9-sgRNA and Cas9-crRNA-tracrRNA complexes.

TS, target strand; NTS, non-target strand.

Table 1. Cas9 orthologs used for genome editing in eukaryote cells		
Organisms	Size (a.a.)	PAM *
Francisella novicida	1,629	NGG
Streptococcus thermophilus **	1,394	NGGNG
Streptococcus pyogenes	1,368	NGG
Streptococcus thermophilus ***	1,121	NNRGAAD
Brevibacillus laterosporus	1,092	NNNNCNAA
Geobacillus stearothermophilus	1,087	NNNNCNAA
Neisseria meningitidis	1,082	NNNNGATT
Staphylococcus aureus	1,052	NNGRRT
Campylobacter jejuni	984	NNNVRYAC
* N=A/T/G/C, D=A/T/G, R=A/G, V=A/G/C, Y=C/T		
** This Cas9 is encoded in CRISPR locus 3		
*** This Cas9 is encoded in CRISPR locus 1		

Cas9 action mechanism

The crystal structures of SpCas9 provided mechanistic insights into the RNA-guided DNA cleavage by Cas9[27–32]. Cas9 comprises recognition (REC) and nuclease (NUC) lobes and accommodates the guide RNA–target DNA heteroduplex in a central channel between the two lobes[33] (Fig. 2A and 2B). The REC lobe mainly consists of α helices and participates in the recognition of the RNA–DNA heteroduplex and sgRNA scaffold (Fig. 2B). The NUC lobe consists of the RuvC, HNH, WED, and PAM-interacting (PI) domains[33] (Fig. 2A and 2B). The PAM-containing duplex (PAM duplex) is bound between the WED and PI domains, where the PAM nucleotides are recognized by a specific combination of amino-acid residues in the PI domain[30] (Fig. 2A and 2B). The PAM recognition induces the unwinding of the double-stranded DNA target, thereby initiating the base-pairing between the crRNA guide and target DNA sequence[30]. The HNH and RuvC domains cleave the DNA target strands complementary (target strand) and non-complementary (non-target strand) to the guide region of sgRNA, respectively[32] (Fig. 2C). The crystal structures of SpCas9 explain the conserved mechanism of RNA-guided DNA cleavage by Cas9. However the diverse mechanism of guide RNA and PAM recognition by Cas9 remains to be fully elucidated, due to the limited sequence identities among Cas9s.

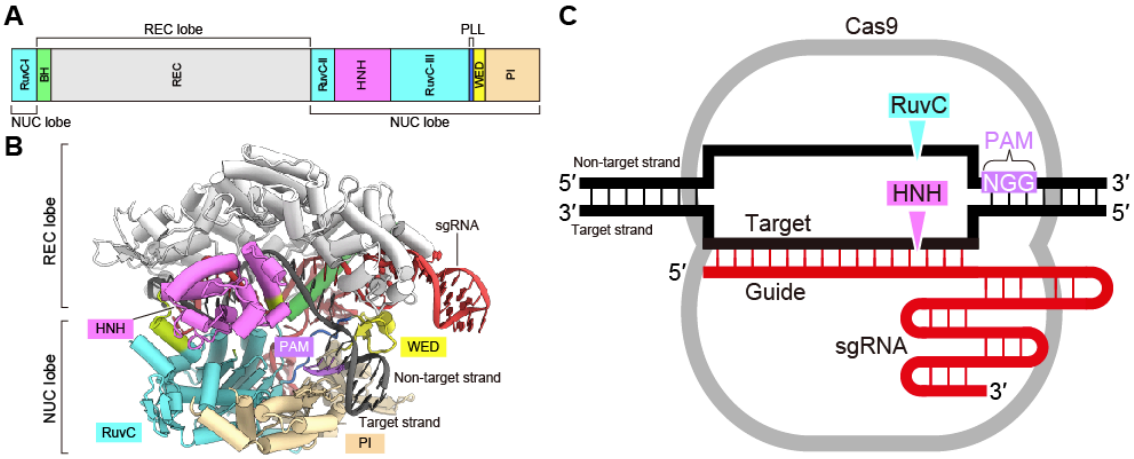

Fig. 3. Cas9 structure and function.

(A) Domain organization of Cas9. BH, bridge helix; PLL, phosphate lock loop.

(B) Crystal structure of SpCas9 (PDB: 4UN3).

(C) Schematics of RNA-guided DNA cleavage by SpCas9.

DNA readout mechanism by Cas9

The mechanism of DNA readout by protein is well characterized by more than 1500 structures of protein-DNA complexes, such as transcriptional factors in complex with promoter DNAs[34]. Proteins recognize DNA sequences in base readout and/or shape readout manners (Fig. 4). In the base readout manner, proteins recognize DNA base edges in major or minor groove. The combinations of hydrogen acceptor, hydrogen donor, and methyl group are shown on the DNA groove surface. The amino acid residues of proteins recognize DNA bases via hydrogen-bonding interactions and hydrophobic interactions. The most representative combination of amino acid-base is arginine and guanine[35] (Table 2). In the shape readout manner, proteins recognize DNA sugar-phosphate backbones. DNA shape adopts A/B/Z-form, is bending/kinked, depending on sequence and protein-interaction.

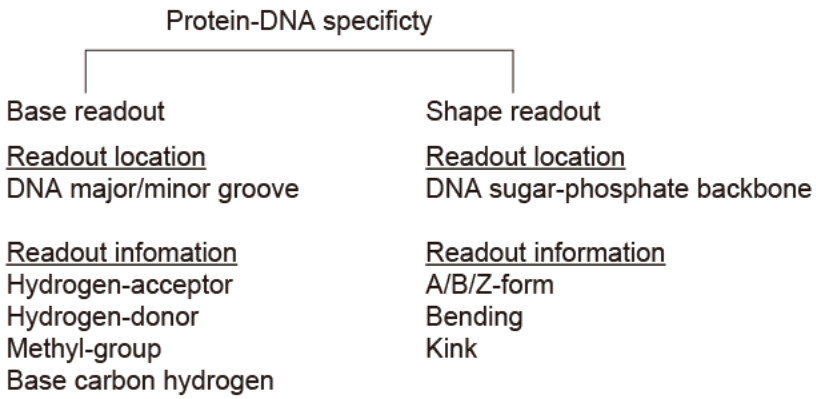

Fig. 4. DNA readout mechanism by protein.

Table 2. The numbers of the observed hydrogen-bonds in 129 protein-DNA complex structures.				
Amino acids (3 letter code)	DNA bases			
	Thymine	Cytosine	Adenine	Guanine
Arg	24	8	19	98
Lys	9	6	4	30
Ser	3	2	1	12
Thr	5	3	4	0
Asn	7	10	18	7
Gln	2	2	16	6
Gly	1	4	0	6
His	0	1	1	12
Tyr	0	2	0	1
Ala	1	1	0	1
Glu	0	10	1	1
Ile	0	0	0	3
Asp	0	5	2	2
Val	0	0	0	0
Cys	0	1	0	0
Phe	0	0	0	1
Leu	0	0	0	0
Met	1	0	0	0
Trp	0	0	0	0
Pro	0	1	0	0
Total	53	56	66	180

The Cas9 protein recognizes the DNA sequence as a PAM. In SpCas9, the two guanines are read out from the major groove via the most representative protein-DNA interactions, arginine-guanine interactions[30]. While Cas9 recognizes the specific DNA sequence in a similar manner to transcriptional factors, the function of the PAM recognition by Cas9 is distinct from those of the promoter recognition by transcriptional factors. The PAM binding induces the DNA unwinding due to the displacement of target DNA base-paired with guide RNA[30], while the DNA-binding of transcription factors does not induce enzymatic activity. As (1) the prediction of protein-DNA specificity is not well-established (2) the DNA-binding function of Cas9 is distinct from those of other proteins (3) the PAM specificity of Cas9 is diverse, the PAM recognition mechanism needs to be investigated.

N^6-methyladenosine (m^{6}A) modification

Eukaryotic mRNAs contain a variety of chemical modifications on riboses and nucleobases[36] (Fig. 5). The most abundant and well-characterized mRNA internal modification is *N*^{6}-methyladenosine (m^{6}A) modification, in which 6-amino group of adenosine is methylated[37]. The technological advances in deep sequencing and transcriptome-wide mapping reveals that the distributions of m^{6}A modification[38]. The m^{6}A modification is found in more than 7000 human gene transcripts[39]. The m^{6}A modification is abundant in the coding and 3′-untranslated regions of mRNAs[39,40]. The m^{6}A modification sites are located at the center of DRACH sequence motif and near the transcription start sites and the stop codons[41].

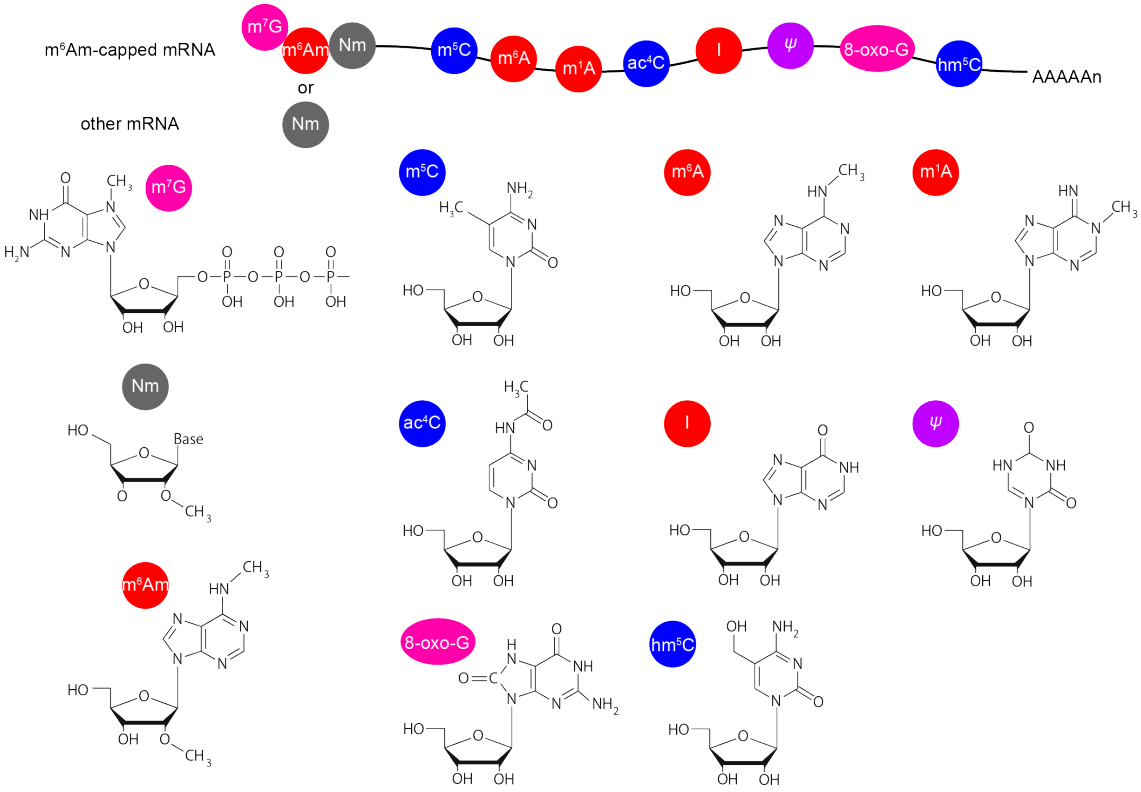

Fig. 5. Chemical modifications in eukaryote mRNA.

Physiological role of m^{6}A modification

The m^{6}A modification is associated with the fate of mRNAs via the interactions with three types of enzymes; writer (methyltransferase), eraser (demethylase), and reader (m^{6}A-binding protein)[38] (Fig. 6). The METTL3-METTL14-WTAP complex is mainly responsible for the m^{6}A modification in mRNAs[37,42]. METTL3, METTL14, and WTAP are the catalytic, regulatory, and auxiliary subunits of the writer complex, respectively. The METTL3-METTL14 heterodimer mediates *N*^{6}-adenosine methylation of the DRAH motif-containing RNA, consistent with the DRACH-motif of m^{6}A modification sites in mRNAs[43]. WTAP localizes the METTL3-METTL14-WTAP complex to the nuclear speckles[38]. FTO and ALKBH5 demethylate *N*^{6}-methylated adenosine in mRNA, although the substrate specificities of FTO and ALKBH5 remain to be investigated[44,45]. The YTH domain-containing proteins bind *N*^{6}-methylated adenosine in mRNA and are associated with molecular machinery in the cellular events, such as splicing, nuclear export, degradation, and translation[46] (Fig. 6). YTHDC1, one of the YTH domain-containing proteins, preferentially binds GG(m^{6}A)C motif-containing RNA, which may be related to the DRACH-motif of m^{6}A modification sites in mRNAs[47]. The perturbations of m^{6}A writer/reader/eraser genes cause the phenotypes related to circadian rhythm, cancer progression, neuronal function, and sex determination[48,49]. Furthermore, the biogenesis and function of m^{6}A modification via the interactions with m^{6}A writer/reader/eraser proteins depend on environmental conditions and developmental stages[49,50]. Taken together, physiological role of m^{6}A modification is related to sequence-dependent, protein-mediated, and dynamic manners. To elucidate the complicated pathways from m^{6}A modification to physiological role, a simple model which excludes the diverse manners (e.g. the sequence-dependent manner does not exist), needs to be investigated as a milestone in epitranscriptomics.

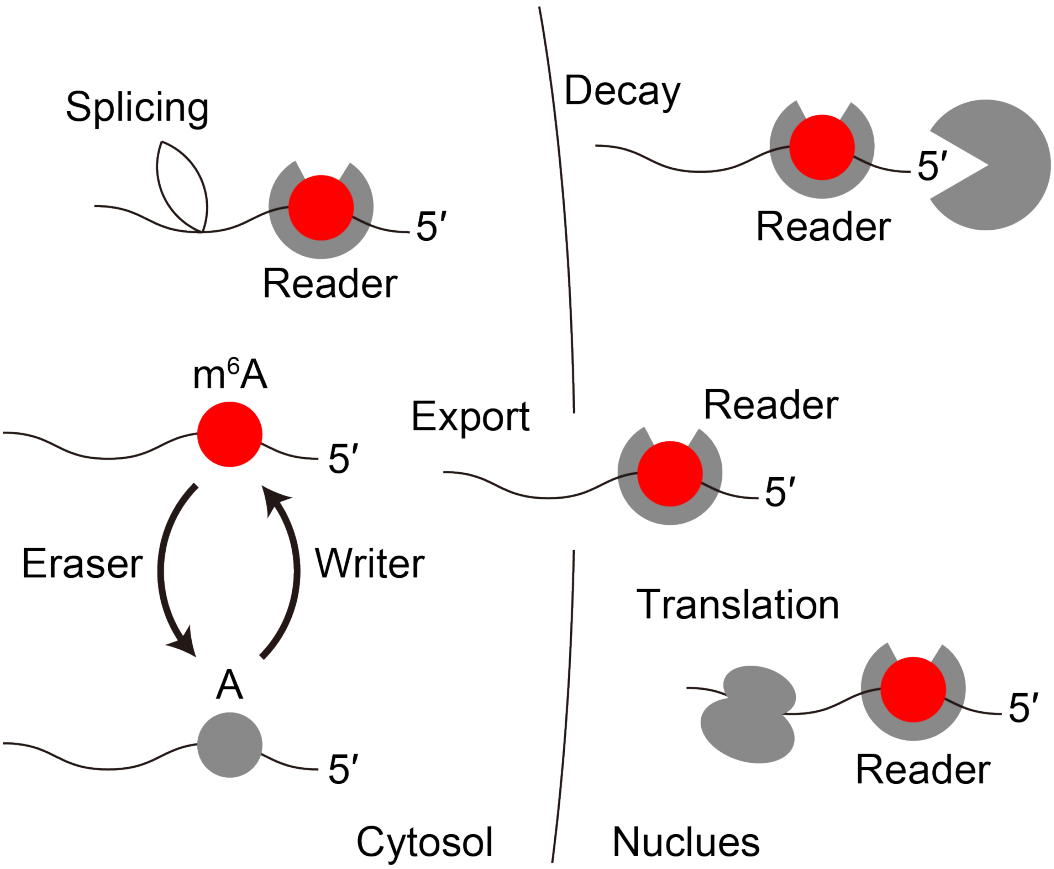

Fig. 6. m^6A biogenesis and function in cells.

Research overview

In genome editing application, Cas9 cannot target the genomic regions containing AT-rich sequence because the widely-used Cas9s recognizes the GC-containing PAM. To address the target limitation of Cas9, engineered Cas9s with altered PAM specificity are developed[24,51–53], and a Cas9 ortholog with distinct PAM specificity is identified[18,19,21–26]. However, the PAM recognition mechanism and the detailed PAM characterization remain to be investigated due to the lack of structural information. I determined the crystal structures of these Cas9s, revealed the PAM recognition mechanism by these Cas9s, and provided the framework for further Cas9 engineering.

To elucidate the complicated pathways of m^6A modification in epitranscriptomics field, I focused on the highly site-specific m^6A modification, which may function in a simple manner. This m^6A modification is introduced by a novel RNA methyltransferase, which affects a large set of mRNAs and is related to the stress-responsive phenotype[54–57]. However, the mRNA recognition and m^6A modification remains to be elucidated. I determined the crystal structure of m^6A RNA methyltransferase and revealed the unexpected mechanism of site-specific m^6A modification by a novel folding domain. The structural insights provided the framework for functional investigation of the m^6A modification pathway.

Both CRISPR-associated endonuclease Cas9 and m^6A RNA methyltransferase CAPAM explore genome and transcriptome and introduce small changes at target sites. These Cas9 and CAPAM molecules cause a variety of phenotypic changes through gene repair and regulation mechanisms, respectively. I examined these molecules biochemically and provided mechanistic insights into the

genome and transcriptome targeting in the cellular complicated environment. The mechanistic insights highlighted the importance of specific interactions between amino-acids and nucleic-acids and provided the structural basis for the development of novel tools inducing conversion from genotypes to phenotypes.

Chapter 2: Crystal structures of SpCas9 mutants

2.1 Introduction

One of the problems in genome editing is the limitation of targetable sites due to the PAM specificity of Cas9. The widely-used SpCas9 recognizes the NGG PAM and other Cas9 orthologs recognize the GC-containing PAMs (Table 1). To expand the targetable sites in genome editing, SpCas9 has been engineered by rational design, based on the crystal structure of SpCas9[30]. In the crystal structure of SpCas9 in complex with its sgRNA and target DNA, the 20 residues of SpCas9 forms direct interactions with the target DNA in non-base-specific manners, and the 2 residues of SpCas9 (Arg1333 and Arg1335) forms direct interactions with the NGG PAM nucleobases of target DNA in base-specific manners[30]. In a previous study, the R1333A/R1335A mutant of SpCas9 was designed for target DNA binding without the PAM recognition (NGG to NNN conversion of the PAM)[30]. Also, the R1333Q/R1335Q mutant of SpCas9 was designed for target DNA binding with the altered PAM recognition (NGG to NAA conversion of the PAM)[30]. However, these mutants did not show the cleavage activity of the target DNA, indicating that the base-specific interactions between two guanines and two arginines are necessary for the DNA targeting of SpCas9[30].

Instead of the rational design strategy, SpCas9 has been engineered by molecular evolution strategy[51] (Fig. 7). The bacterial cells were transformed with the first plasmid containing toxic gene with the SpCas9-target site and the altered PAM (the NGA or NGC PAM). Then, the bacterial cells were transformed with the second plasmid encoding randomly mutagenized SpCas9 and its sgRNA. The bacterial cells were screened for SpCas9 mutants which cleaves the target site with the altered PAM. The screening revealed that The VQR (D1135V/R1335Q/T1337R), EQR (D1135E/R1335Q/T1337R), and VRER (D1135V/G1218R/R1335E/T1337R) mutants recognize the NGA, NGAG, NGCG PAMs, respectively[51]. These SpCas9 mutants showed robust genome editing activities for endogenous target sites with altered PAMs in human cells, and thus expanded the target space in Cas9-mediated genome editing[51]. The SpCas9 single mutants (R1335Q and R1335E) showed no cleavage activity towards any targets, while the SpCas9 multiple mutants showed the cleavage activities towards the targets with altered PAMs[51]. These data indicated that the multiple mutations of SpCas9 cooperatively function for the altered PAM specificity.

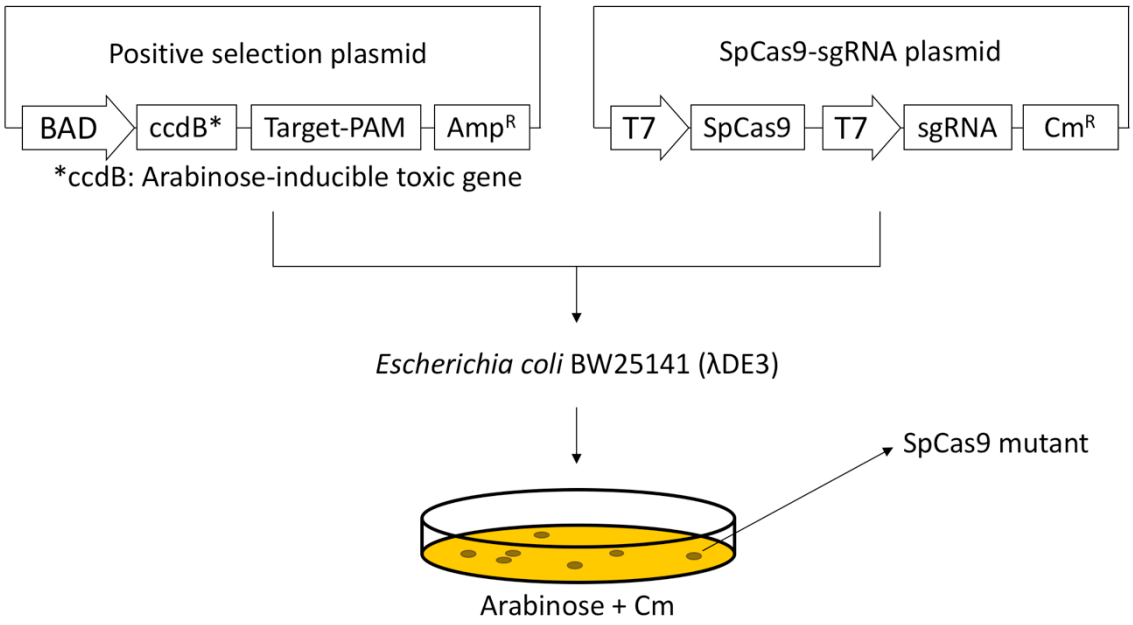

Fig. 7. Positive selection of SpCas9 mutants with altered PAM specificities.

BAD, P_{BAD} promoter; Amp, Ampicillin; T7, T7 promoter; Cm, Chloramphenicol.

To reveal the altered PAM recognition mechanism, I modeled the VQR, EQR, and VRER mutants in complex with their sgRNAs and target DNAs. The crystal structure of the wild-type SpCas9 showed that the DNA duplex containing PAM is bound to the groove between the WED and PI domains in the PAM recognition[30]. The mutated residues of the VQR, EQR, and VRER mutants are located at the PAM-binding groove. Thus, these residues at the positions of 1135, 1218, 1335, and 1337 are predicted to contribute to the altered PAM specificities from the structural viewpoint. In protein-DNA complex structures, adenine and cytosine bases are recognized by glutamine and glutamate residues, respectively[35]. The VQR/EQR and VRER mutants recognize the NGA/NGAG and NGCG PAMs, respectively[51]. Thus, the introduced glutamine and glutamate residues (R1335Q and R1335E) are predicted to recognize the adenine and cytosine bases in the PAMs, respectively. However, the models based on the wild-type SpCas9 structure cannot explain the altered PAM recognition mechanism. Given that the side chain lengths of glutamine and glutamate are shorter than that of arginine, the glutamine and glutamate residues (R1335Q and R1335E) cannot form the hydrogen-bonding interactions with the adenine and cytosine bases, respectively. Furthermore, the role of the introduced one valine and two arginine residues (D1135V, G1218R, and T1337R) cannot be explained due to the lack of the structures of the SpCas9 mutants in complex with their sgRNAs and target DNAs.

2.2 Research aims

The prediction of protein-DNA specificity is difficult because the amino-acid residues and nucleic acid residues are flexible in protein-DNA complexes. I determined the crystal structures of engineered SpCas9 mutants, elucidated the mechanism of altered PAM recognition, and provided the framework for the further engineering of Cas9.

2.3 Methods

2.3.1 Sample preparation

The wild-type SpCas9 and SpCas9 mutants used in the crystallization and *in vitro* cleavage experiments contain the C80L/C574E mutations for the solubilization. The D10A/H840A mutations were introduced into the SpCas9 mutants for preventing the potential DNA cleavage during the crystallization. The His_6-GST-tagged SpCas9 was expressed at 20 °C overnight in *Escherichia coli* Rosetta 2 (DE3) (Novagen). The bacterial lysate containing recombinant SpCas9 was batched with Ni-NTA Superflow resin (Qiagen) in buffer A (50 mM Tris-HCl, pH 8.0, 20 mM imidazole, and 1 M NaCl) and eluted with buffer B (50 mM Tris-HCl, pH 8.0, 300 mM imidazole, and 0.3 M NaCl). The His_6-GST-tagged SpCas9 was mixed with TEV protease targeting the cleavage site between the His_6-GST-tag and SpCas9. The mixture was dialyzed at 4 °C overnight in buffer C (20 mM Tris-HCl, pH 8.0, 40 mM imidazole, and 0.5 M NaCl). After removing the His_6-GST-tag by a Ni-NTA column, SpCas9 was purified through a HiTrapSP HP column (GE Healthcare), using buffer D (20 mM Tris-HCl, pH 8.0) and buffer E (20 mM Tris-HCl, pH 8.0 and 2 M NaCl). In a previous study, the SpCas9 (wild-type)–sgRNA–target DNA (NGG PAM) complex was crystallized, using the 81-nt sgRNA, the 28-nt target DNA strand, and the 8-nt non-target DNA strand[30]. The 81-nt sgRNA was transcribed *in vitro*, using T7 RNA polymerase and a PCR-amplified DNA template. The transcribed RNA was purified by 10% acrylamide denaturing Urea PAGE. I reconstituted the purified SpCas9 mutant, the purified sgRNA, the 28-nt target DNA strand, and the 8-nt non-target DNA strand at a molar ratio of 1:1.5:2.3:2.7. The SpCas9-sgRNA-DNA complex was isolated through a Superdex 200 Increase column (GE Healthcare), using buffer F (20 mM HEPES-NaOH, pH 7.5, 250 mM KCl, 5 mM $MgCl_2$, 1 mM dithiothreitol).

2.3.2 *in vitro* cleavage assay

The pUC119 plasmid containing the 20-nt target sequence and the PAM was linearized with BamHI digestion and used as the substrate for *in vitro* cleavage assays. The BamHI-linearized plasmid (100 ng, 5 nM) was incubated at 37 °C for 5 min with the SpCas9-sgRNA complex (10–250 nM) in 10 μL of reaction buffer containing 20 mM HEPES-NaOH, pH 7.5, 100 mM KCl, 2 mM $MgCl_2$, 1 mM dithiothreitol, and 5% glycerol. The reaction was stopped by the addition of quench buffer containing EDTA (40mM final concentration) and proteinase K (10 μg). Reaction products were resolved on 1% agarose gel containing ethidium bromide and then visualized using a Typhoon FLA 9500 scanner (GE Healthcare).

2.3.3 Crystallography

The SpCas9-sgRNA-DNA complex was crystallized at 20 °C, using the hanging-drop vapor diffusion method. Crystals were obtained by mixing 1 μL of complex solution ($A_{260\ nm}$ = 15) and 1 μL of reservoir solution (100 mM Tris-acetate, pH 8.0, 400 mM KSCN, 15–17% PEG3,350). The crystals were improved by microseeding, using Seed Bead (Hampton Research). The crystals were cryoprotected in a buffer containing 100 mM Tris-acetate, pH 8.0, 400 mM KSCN, 30% PEG3,350 and 10% ethylene glycol. The X-ray diffraction data were collected at 100 K on beamline BL41XU at SPring-8 and processed using XDS[58] and AIMLESS[59]. The structure was determined by molecular replacement with MOLREP[60], using the wild-type SpCas9 structure (PDB: 4UN3) as the search model. The structure models were built using COOT[61], and refined using PHENIX[62]. Structural figures were prepared using CueMol (http://www.cuemol.org).

2.4 Results

2.4.1 Sample preparation for the biochemical analysis

I made plasmid constructs of the C80L/C574E SpCas9 (designated as wild-type SpCas9), the C80L/C574E/D1135V/R1335Q/T1337R SpCas9 (designated as VQR mutant), the C80L/C574E/D1135E/R1335Q/T1337R SpCas9 (designated as EQR mutant), and the C80L/C574E/D1135V/G1218R/R1335E/T1337R SpCas9 (designated as VRER mutant) (Fig. 8A). The 10–20% acrylamide SDS-PAGE analysis showed that the SpCas9 proteins were produced in high purity. The 10 % acrylamide denaturing Urea PAGE analysis showed that the SpCas9 sgRNA containing 20-nt target sequence was produced in high purity (Fig. 8B).

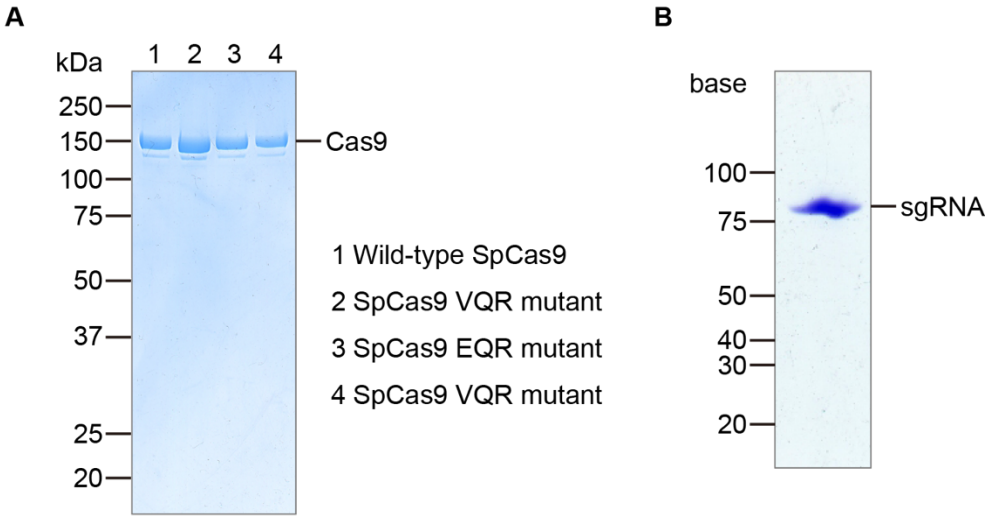

Fig. 8. Sample preparation for the biochemical analysis.

(A) SDS-PAGE analysis of the wild-type and mutant SpCas9 proteins.

(B) Denaturing Urea PAGE analysis of the SpCas9 sgRNA.

2.4.2 Biochemical characterization of SpCas9 mutants

To confirm the cleavage activities of SpCas9 mutants shown in *in vivo* experiments[51], I performed the *in vitro* cleavage assays using purified SpCas9 mutants. The wild-type SpCas9 efficiently cleaved the target plasmid with TGG PAM, rather than TGA and TGCG PAMs (Fig. 9A and 9B). The VQR mutant efficiently cleaved the target plasmid with TGA PAM, rather than TGG PAM (Fig. 9A). The EQR mutant efficiently cleaved the target plasmid with TGAG PAM, rather than TGGG PAM (Fig. 9B). The VRER mutant efficiently cleaved the target plasmid with TGCG PAM, rather than TGGG PAM (Fig. 9C). The VQR mutant preferred the TGAG target to the TGAH targets (H is A, T, or C), while the EQR and VRER mutants are specific to the TGAG and TGCG PAMs, respectively (Fig. 9D). Taken together, I confirmed that the VQR, EQR, and VRER mutants recognize the NGA, NGAG, and NGCG PAM, respectively, consistent with the data of previous experiments in mammalian cells[51].

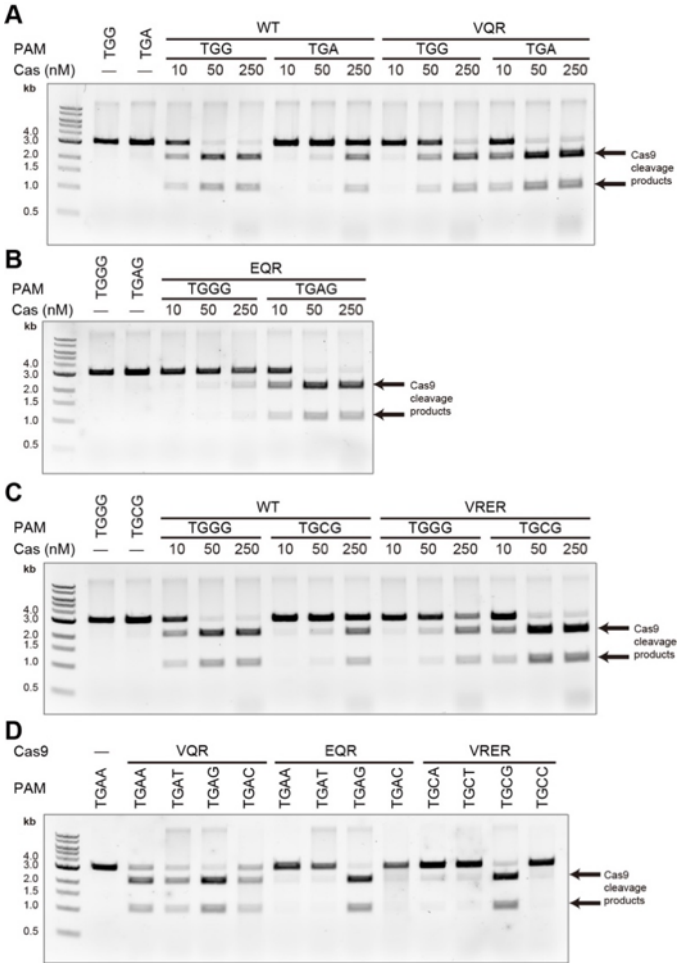

Fig. 9. *in vitro* cleavage activities of the SpCas9 mutants.

(A) *in vitro* cleavage activities of wild-type and VQR SpCas9. The SpCas9-sgRNA complex (10, 50 and 250 nM) was incubated at 37 °C for 5 min with the target plasmid containing either the TGG or TGA PAM.

(B) *in vitro* cleavage activities of wild-type and VRER SpCas9. The SpCas9-sgRNA complex (10, 50 and 250 nM) was incubated at 37 °C for 5 min with the target plasmid containing either the TGGG or TGCG PAM.

(C) *in vitro* cleavage activity of EQR SpCas9. The SpCas9-sgRNA complex (10, 50 and 250 nM) was incubated at 37 °C for 5 min with the target plasmid containing either the TGGG or TGAG PAM.

(D) Preference of the VQR, EQR and VRER mutants for the 4th PAM nucleotides. The SpCas9-sgRNA complex (50 nM) was incubated at 37 °C for 5 min with the target plasmid containing either the TGAN PAM (VQR and EQR) or the TGCN PAM (VRER).

2.4.3 Crystallization and structural determination of SpCas9 mutants

To prevent the potential DNA cleavage, I introduced the D10A/H840A mutations into the VQR, EQR, and VRER mutants. Similar to the sample preparation for the biochemical analysis, the SpCas9 proteins and their cognate sgRNA were produced in high yield and purity. The chromatography profiles showed that all the SpCas9–sgRNA–target DNA complexes of the VQR, EQR, and VRER mutants are stable in the solution (Fig. 10). I obtained about 300-μm length crystals of the VQR, EQR, and VRER mutants in complex with their sgRNAs and target DNAs containing the TGAG, TGAG, and TGCG PAMs, respectively (Fig. 11A). The analysis of X-ray diffraction experiments showed that the collected data of the VQR, EQR, and VRER mutants were at 2.0, 2.2, and 2.2 Å resolutions, respectively (Fig. 11B and 11C) (Table 3). The refinement and validation software, PHENIX, showed that the R_{work} / R_{free} values of the final VQR, EQR, and VRER structure models are 0.200 / 0.230, 0.207 / 0.231, and 0.200 / 0.230, respectively (Table 3).

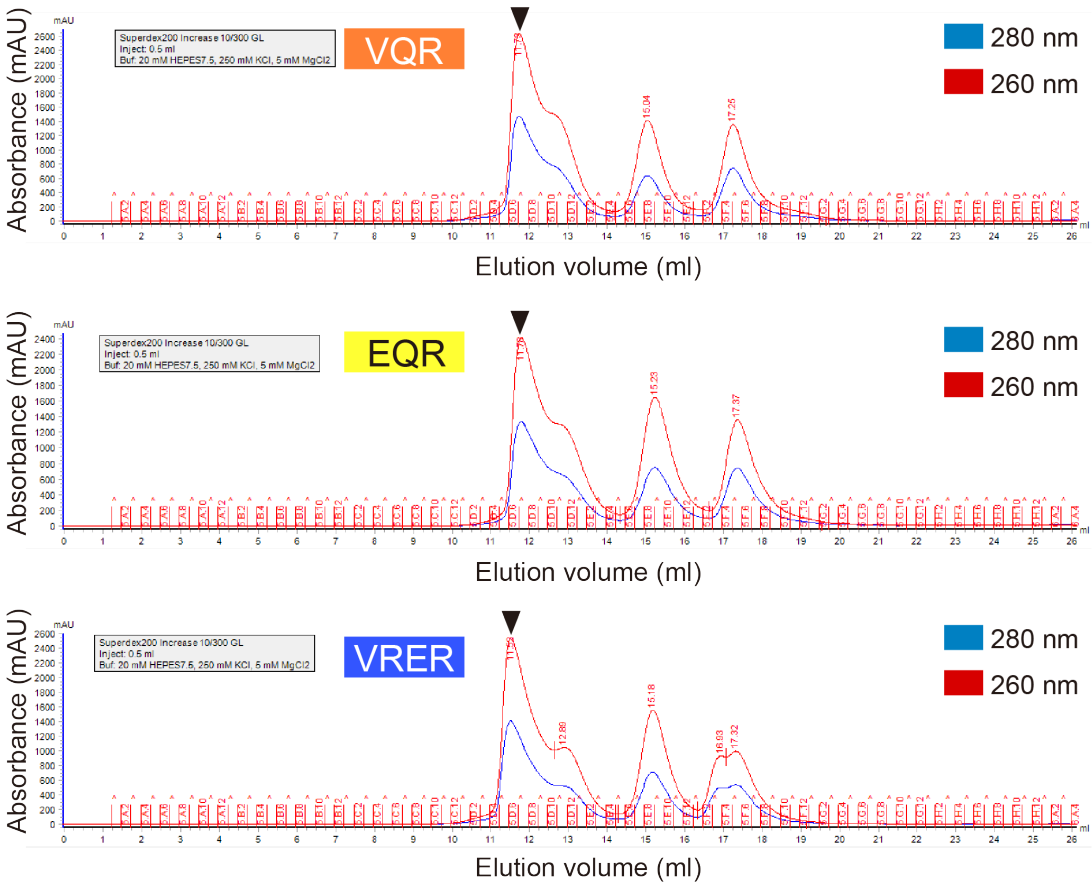

Fig. 10. The size exclusion chromatography profile of the SpCas9–sgRNA–target DNA complexes The original raw data of this profile had been lost due to the laboratory renovation. Instead of the re-calculated raw data, the copy of the data in the laboratory note is shown in this thesis. The triangles shows the putative SpCas9–sgRNA–target DNA complex fractions, based on the molecular weights of the complexes.

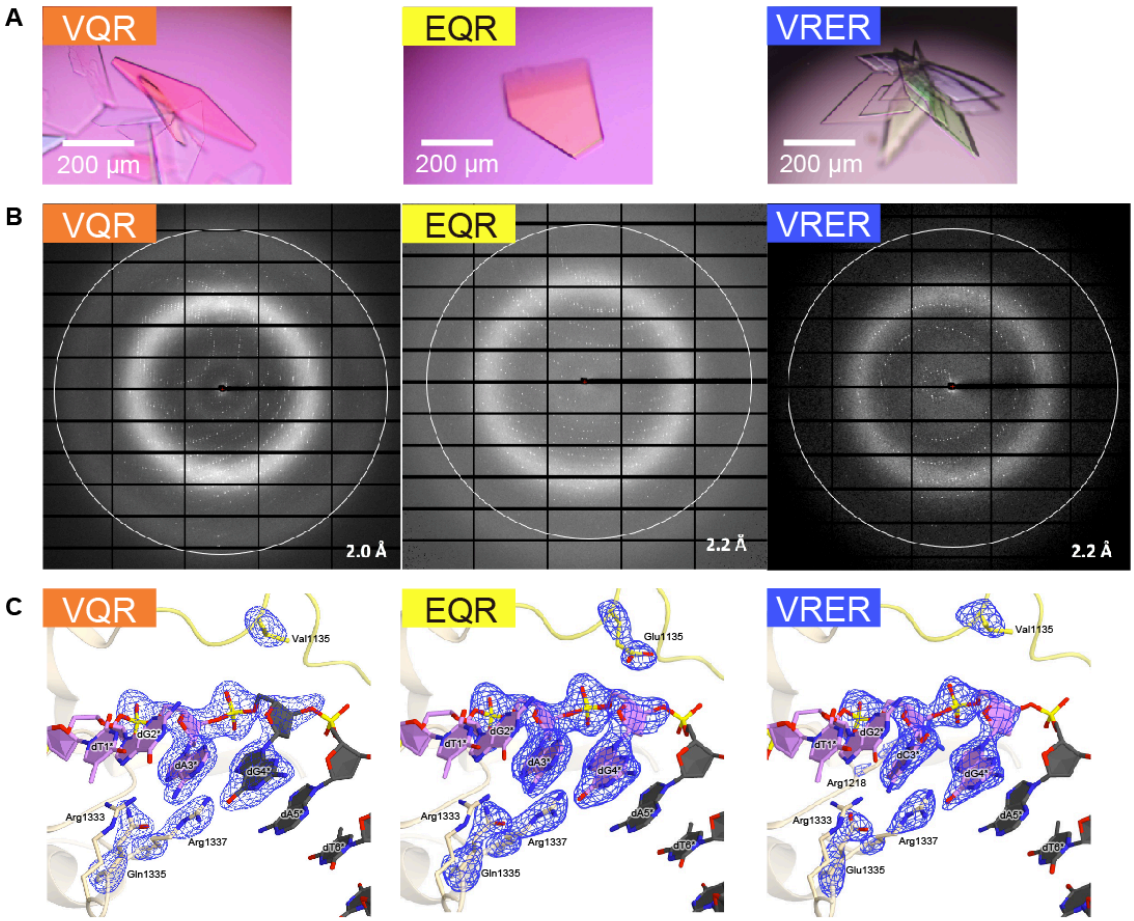

Fig. 11. X-ray crystallography of SpCas9 mutants.

(A) Crystals of the VQR, EQR, and VRER mutants.

(B) Diffraction images obtained from the VQR, EQR, and VRER mutant crystals.

(C) $mF_O - DF_C$ omit electron maps of the VQR, EQR, and VRER mutants (contoured at 4.0σ).

Table 3. Data collection and refinement statistics.

	VQR	EQR	VRER
Data collection			
Beamline	SPring-8 BL41XU	SPring-8 BL41XU	SPring-8 BL41XU
Wavelength (Å)	1.0000	1.0000	1.0000
Space group	*C*2	*C*2	*C*2
Cell dimensions			
a, *b*, *c* (Å)	177.8, 67.7, 187.6	178.0, 67.8, 187.6	177.0, 69.5, 188.2
α, β, γ (°)	90, 111.2, 90	90, 111.1, 90	90, 109.7, 90
Resolution (Å)*	49.4–2.0 (2.03–2.00)	48.0–2.2 (2.24–2.20)	49.0–2.2 (2.24–2.20)
R_{merge}	0.070 (1.016)	0.071 (0.664)	0.051 (0.535)
R_{pim}	0.044 (0.635)	0.050 (0.452)	0.041 (0.430)
$I/\sigma I$	10.2 (1.8)	8.0 (1.5)	11.2 (1.8)
Completeness (%)	97.6 (97.4)	97.8 (98.1)	94.7 (95.8)
Multiplicity	3.5 (3.5)	2.9 (2.9)	2.4 (2.4)
CC(1/2)	0.996 (0.557)	0.913 (0.699)	0.995 (0.615)
Refinement			
Resolution (Å)	47.9–2.0	48.0–2.2	49.0–2.2
No. reflections	136,985	119,122	103,608
R_{work}/ R_{free}	0.201 / 0.231	0.212 / 0.249	0.194 / 0.234
No. atoms			
Protein	10,545	10,706	10,714
Nucleic acid	2,468	2,464	2,464
Ion	7	8	7
Solvent	854	841	805
B-factors (Å^2)			
Protein	53.5	60.7	59.2
Nucleic acid	52.2	55.2	53.9
Ion	29.6	36.8	42.0
Solvent	44.0	49.8	48.7
R.m.s. deviations			
Bond lengths (Å)	0.005	0.002	0.002
Bond angles (º)	0.785	0.490	0.475
Ramachandran plot (%)			
Favored region	97.0	97.3	97.2
Allowed region	2.7	2.6	2.7
Outlier region	0.3	0.1	0.1

*Values in parentheses are for the highest resolution shell.

2.4.4 Crystal structures of SpCas9 mutants

Overall structures of the SpCas9 mutants were similar to that of the wild-type SpCas9, indicating that the mutations have no effect on overall structures of SpCas9 (Fig. 12A–C). In the SpCas9 mutants, the sgRNA-target DNA heteroduplexes are accommodated at the central channel between the REC and NUC lobes (Fig. 12A and 12C). The PAM-contained DNA duplexes is bound to the groove between the WED and PI domains (Fig. 12A and 12C). The mutated residues 1135, 1218, 1335, and 1337 at the PAM-binding groove interact with the non-target strand DNA (Fig. 12B). In the SpCas9 mutants, the PAM nucleotides of the non-target strand are recognized by three residues 1333, 1335, and 1337 in base-specific manners (Fig. 13). For the clarity, I represent the nucleotide of the PAM in abbreviation, unless otherwise stated. For example, the guanine nucleotide at the 2nd position of the PAM in the non-target strand DNA is designated as dG2*. In the wild-type SpCas9, the dG2* and dG3* nucleobases form hydrogen bonds with Arg1333 and Arg1335, respectively (Fig. 13). In the VQR mutant, the dG2*, dA3*, and dG4* nucleobases form hydrogen bonds with Arg1333, Gln1335, and Arg1337, respectively (Fig. 13). In the EQR mutant, the dG2*, dA3*, and dG4* nucleobases form hydrogen bonds with Arg1333, Gln1335, and Arg1337, respectively (Fig. 13). In the VRER mutant, the dG2*, dC3*, and dG4* nucleobases form hydrogen bonds with Arg1333, Glu1335, and Arg1337, respectively (Fig. 13). In the VQR mutant and wild-type SpCas9, the dA3* and dG3* nucleobases form bidentate hydrogen bonds with residue 1335, while, in the VRER mutant, the dC3* nucleobase forms only a single hydrogen bond with residue 1335 (Fig. 13). These results indicate that, in the VRER mutant, the dG4*-Arg1337 interaction compensates for the dC3*-Glu1335 interaction, resulting in the requirement of 4th G in the NGCG PAM (Fig. 13). Taken together, the SpCas9 mutants recognize their cognate PAMs through hydrogen-bonding interactions.

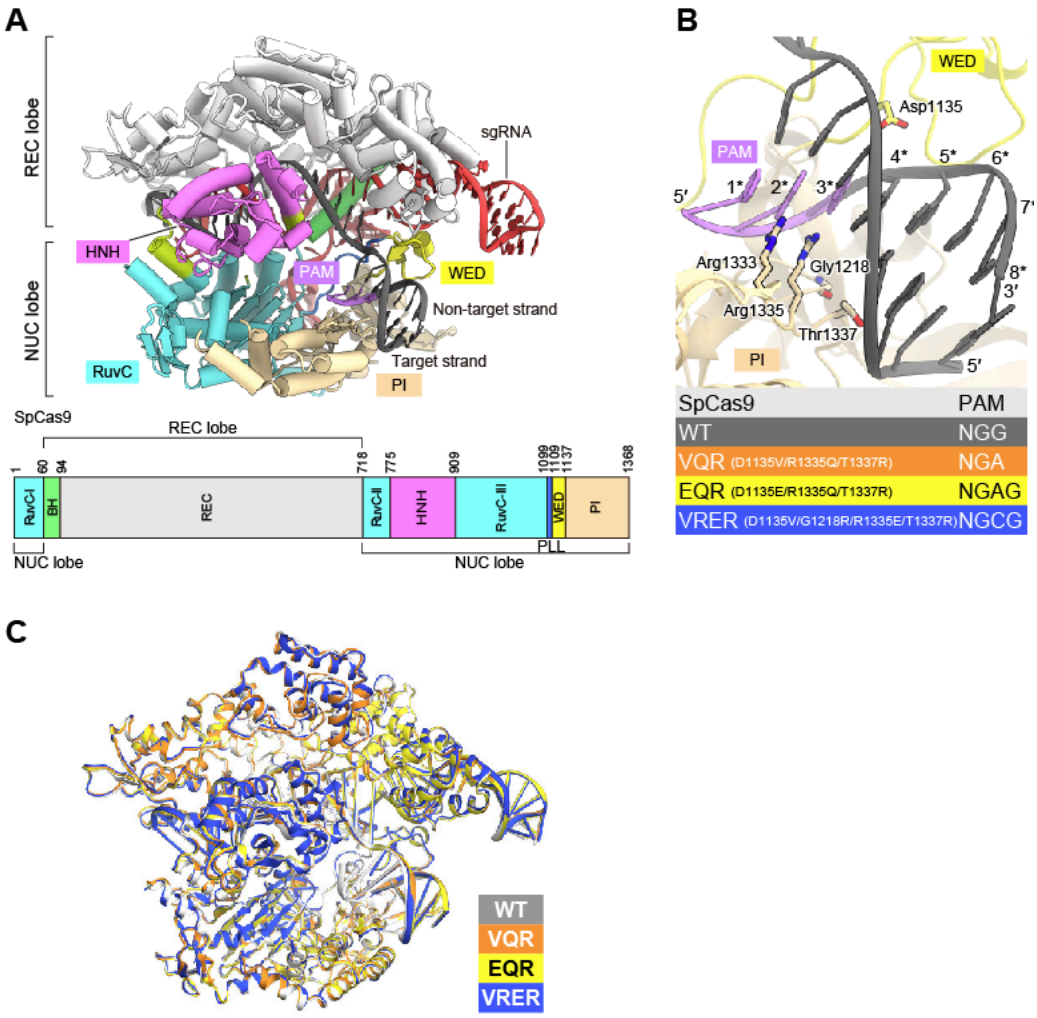

Fig. 12. Crystal structures of the SpCas9 mutants.

(A) Overall structure of the VRER mutant in complex with the sgRNA and the target DNA. BH, bridge helix; PLL, phosphate lock loop.

(B) Location of the mutational residues mapped in the wild-type SpCas9 structure (PDB: 4UN3).

(C) Superimposition of the wild-type SpCas9 (PDB: 4UN3) (gray), VQR (orange), EQR (yellow), VRER (blue) mutants.

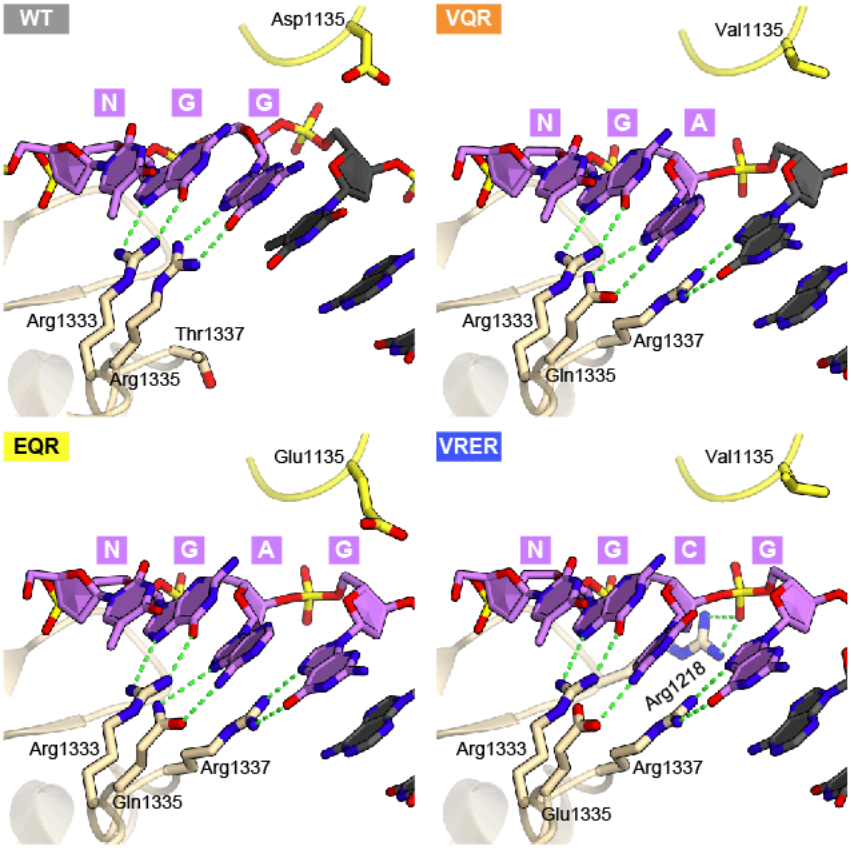

Fig. 13. PAM recognition by the wild-type SpCas9 (PDB: 4UN3), VQR, EQR, and VRER mutants. The PAM is colored in purple. Hydrogen bonds are shown as dashed lines.

2.4.5 DNA strand displacement in the PAM recognition

Structural comparison of the three SpCas9 mutants with the wild-type SpCas9 revealed that, in all the three mutants, the DNA strand is displaced and the third nucleobase in the PAM moves towards the residue 1335 (Fig. 13A). The DNA strand displacement enables the hydrogen-bonding interaction between residue 1335 and the third nucleobase in the PAM (Fig. 13B). A previous study showed that the additional mutations as well as the R1335Q/R1335E mutation are necessary for the altered PAM recognition by engineered SpCas9[51]. These results indicate that the multiple mutations cooperatively induce the DNA strand displacement and enable the altered PAM recognition. In the three SpCas9 mutants, Arg1337 interacts with the fourth PAM nucleobase, inducing the DNA strand displacement (Fig. 13B). Val1135 and Glu1135 interact with the ribose of fourth PAM nucleotide, facilitating the DNA strand displacement in the VQR/VRER and EQR mutants, respectively (Fig. 13C). Glu1135 in the EQR mutant interacts with the PAM nucleotide more tightly than Val1135 in the VQR mutant, defining the specific interaction between Arg1337 and dG4* nucleobase (Fig. 13C). Consistently, the VQR mutant recognizes the relaxed NGAN PAM, while the EQR mutant recognizes the specific NGAG PAM (Fig.13C). In the VRER mutant, Arg1218 interacts with the phosphate backbone, stabilizing the displaced DNA strand (Fig. 13C). Taken together, the structural comparison revealed the DNA strand displacement and the structural findings explained the role of each mutation in the PAM recognition.

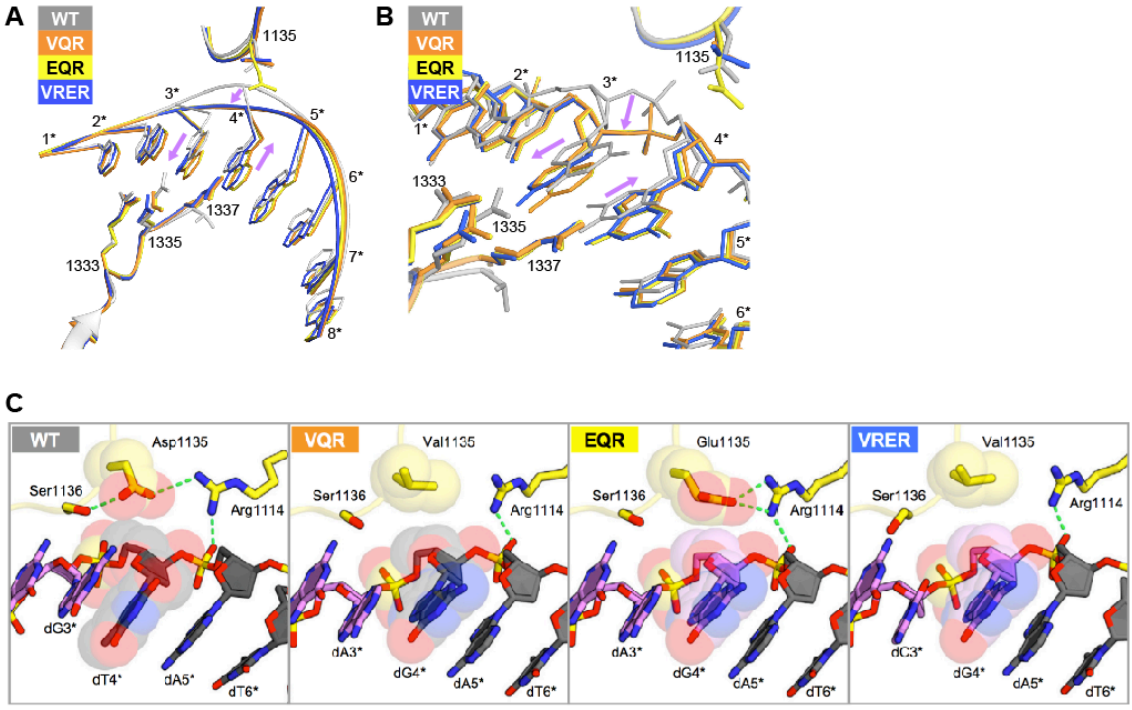

Fig. 13. DNA strand displacement by the multiple mutations.

(A) The phosphate backbone displacement in the PAM. The wild-type SpCas9 (PDB: 4UN3) (gray), VQR (orange), EQR (yellow), VRER (blue) mutants. are superimposed.

(B) The displacement of the third and fourth PAM nucleotides. The wild-type SpCas9 (PDB: 4UN3) (gray), VQR (orange), EQR (yellow), VRER (blue) mutants. are superimposed.

(C) Interactions between the residues 1135 and the fourth PAM nucleotides.

2.5 Discussion

This study provided the mechanistic insights into the PAM recognition which is distinct from the conventional DNA recognition, such as the promoter DNA recognition by transcriptional factors. The *in vitro* cleavage assay showed the base-specific DNA-binding is coupled with the DNA cleavage, highlighting the functionality of PAM recognition. The biochemical and structural analyses showed that the glutamine/glutamate residues contribute to the G to A/C specificity conversions not only in target DNA-binding but also in target DNA-cleavage. Therefore, this study showed the unique property of DNA recognition by Cas9.

The present crystal structures of the VQR, EQR, and VRER mutants showed that the multiple mutations induce the target DNA displacement. In the back-to-back study, another research group determined the crystal structures of the VQR, EQR, and VRER mutants in complex with their sgRNAs and target DNAs containing the NGA, NGAG, and NGCG PAMs, respectively[63]. Their crystal structures and my crystal structures are similar, validating my explanation about the altered PAM recognition mechanism.

The 3D structures of the displaced DNA base and sugar-phosphate backbone is informative for further protein engineering of SpCas9. The displaced DNA sugar-phosphate backbone is similar in the VQR and VRER structures. Thus, the additional arginine (G1218R) is predicted to form the hydrogen-bonding interactions with the DNA sugar-phosphate backbone in the VQR mutant as well as in the VRER mutant. Indeed, the VRQR mutant (D1135V/G1218R/R1335Q/T1337R) showed the higher activities towards the target sites containing the NGA(A/T/C) PAM, rather than the VQR mutant[64]. In

the VRER structures, the cytosine in the NGCG PAM forms the hydrogen-bonding interaction with the glutamate (R1335E). Given that the thymine forms the hydrogen-bonding interaction with the asparagine in protein-DNA complexes[35] (Table 2), the VRNR mutant (D1135V/G1218R/R1335N/T1337R) may cleave the target DNA containing the NGTG PAM, although the side chain length of asparagine is shorter than that of glutamate.

The displaced DNA bases and sugar-phosphate backbones are recognized by Cas9 protein in a shape readout manner, which is common in protein-DNA complexes[34]. This study showed that the general concept in protein-DNA specificity can be applied to Cas9. In the previous study assuming that the target DNA shape is not shifted, the rational design of Cas9 (e.g. the SpCas9 R1333Q/R1335Q mutant) did not work[30]. However, this study showed that the new design strategy of DNA displacement is workable. This strategy worked with the development of the VRQR mutant[64]. This strategy supports for the rational design of further DNA displacement. Based on the crystal structure of the VQR mutant, the five mutations (G1218R/N1286Q/I1331F/D1332K/T1337R) were predicted to displace the 2nd–4th DNA nucleotides, and the three mutations (R1333Q/R1335Q/T1337R) were predicted to recognize the 2nd–4th DNA nucleotides as an altered PAM[63]. The SpCas9 G1218R/N1286Q/I1331F/D1332K/R1333Q/R1335Q/T1337R mutant cleaved the target DNA with the NAAG PAM[63]. Taken together, this study provided the framework for further engineering of Cas9 with altered PAM specificity.

This study showed the design strategy of Cas9 with altered PAM specificity by analyzing each role of the mutations in the VRER mutant from the structural viewpoint. The structural analyses of the VQR

and EQR mutants support that of the VRER mutant. In the VRER mutant, the valine/arginine residues (D1135V/G1218R) form the hydrophobic/hydrogen-bonding interactions with the DNA sugar-phosphate backbone, stabilizing the protein-DNA interactions. These two mutations (D1135V and G1218R) not only induce the DNA displacement in assistance with R1335E and T1337R, but also reinforce the protein-DNA binding. The design strategy of reinforced protein-DNA binding was applied to the SpCas9 L1111R/D1135V/G1218R/E1219F/A1322R/R1335V/T1337R mutant which recognizes the NG PAM and was designated as SpCas9-NG[53]. Based on the crystal structure of the VRER mutant, the SpCas9-NG was designed to contain the two mutations (D1135V and G1218R) which are shared with the VRER mutant, the three mutations (L1111R, E1219F, and A1322R) which reinforce the protein-DNA binding, and the two mutations (R1335V and T1337R) which alter the PAM specificity[53]. The design strategy of reinforced protein-DNA binding works in Cas9 orthologs. *Staphylococcus aureus* Cas9 (SaCas9) recognizes the NNGRRT PAM, while the SaCas9 E782K/N968K/R1015H mutant recognizes the NNNRRT PAM[29,65]. The introduced two lysine residues (E782K/N968K) are predicted to form the interactions with the target DNA, supporting the altered PAM specificity by the introduction of histidine residue (R1015H)[29,65]. *Francisella novicida* Cas9 (FnCas9) recognizes the NGG PAM, while FnCas9 E1369R/E1449H/R1556A mutant recognizes the YGN PAM[24]. According to the crystal structure of the FnCas9 mutant, the introduced arginine and histidine residues (E1369R/E1449H) form the interactions with the target DNA, supporting the altered PAM specificity by the introduction of alanine residue (R1556A)[24]. Taken together, the design strategy of Cas9, which was proposed in this study, works in not only SpCas9 but also other Cas9 orthologs.

Structural comparison of the VQR mutant with the SpCas9-NG mutant showed the different roles of the introduced arginine residue (T1337R) in the two mutants[53]. In the VQR mutant, Arg1337 interacts with the fourth nucleotide in the PAM and induces the DNA displacement (Fig. 14). In the SpCas9-NG mutant, Arg1337 interacts with the third and fourth nucleotides in the PAM and contributes to the PAM preference of the SpCas9-NG mutant[53] (Fig. 14). In contrast to the SpCas9-NG mutant, Arg1337 is stacked with the residue at the 1335 position, so that the fixed side chain of Arg1337 contributes to the DNA displacement. Given that the arginine residues form a variety of hydrogen-bonding interactions with adenine/thymine/guanine/cytosine in protein-DNA complexes[35] (Table 2), in the VQR mutant, the side chain conformations of Arg1337 interacting with adenine/thymine/cytosine need to be investigated by X-ray crystallography.

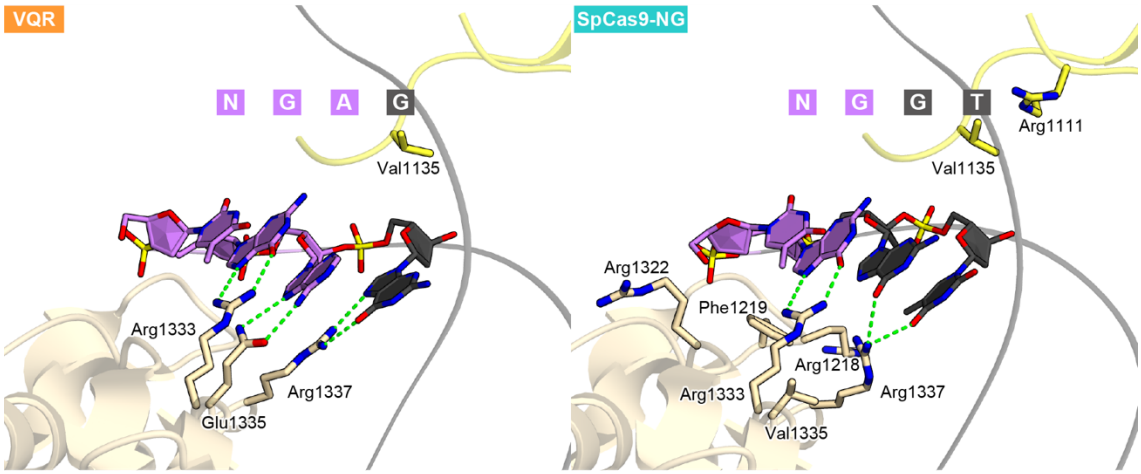

Fig. 14. Structural comparison of the VQR mutant with the SpCas9-NG mutant (PDB: 6AI6).

Chapter 3: Crystal structure of CdCas9

3.1 Introduction

Among bacteria and archaea, the CRISPR-Cas system is evolved though horizontal gene transfers[66]. The natural context may evolve Cas9 by DNA shuffling and DNA point mutation. The laboratory experiments of random mutagenesis and site-directed mutagenesis evolve Cas9 only by DNA point mutation. Protein engineering of SpCas9 altered the NGG PAM to the NGA, NGCG, and NG PAMs[24,51,53,65]. These PAM conversions were limited to only one nucleotide. To explore the distinct PAM specificity from the NGG PAM specificity, I focused on Cas9 orthologs discovered form the natural diversity. More than 1000 Cas9-like proteins containing the RuvC and HNH domains has been discovered from bacterial genome[9,11,22]. The biochemical studies characterized 19 Cas9 orthologs with their cognate guide RNA and PAM[9,11,22]. Currently, 9 out of 19 Cas9 orthologs showed their activities in eukaryote cells[18,19,21–26] (Table 1). Among the other 10 Cas9 orthologs, I focused on *Corynebacterium diphtheriae* Cas9 (CdCas9) with important properties in genome editing application[22,67].

In genome editing application, there are four important topics; (1) targetability, (2) versatility, (3) efficiency, and (4) precision. As for (1) targetability, the PAM recognition constrains the targetable sites in genome. CdCas9 recognizes the NNRHHHY PAM including the NNAAAAT PAM, suggesting the possibility of CdCas9-mediated genome editing in the AT-rich regions[22]. As for (2) versatility, the small size of Cas9 is necessary for efficient AAV-delivery of Cas9 into specific tissues in therapeutic application. In contrast to the size of SpCas9 (1,368 a.a.), the size of CdCas9 (1,084 a.a.) is comparable to that of SaCas9 (1,052 a.a.), which is used in the efficient Cas9-delivery system[22]. As for (3) efficiency, the genome editing efficiency depends on the DNA cleavage efficiency of Cas9, the chromatin accessibility of target site, and the effect of DNA repair machinery[9,68]. In a previous study, CdCas9 showed the weak DNA cleavage activity[22,67]. As for (4) precision, it is not investigated whether CdCas9 cleaves the target DNA containing the mismatches with guide RNA or not. Taken together, CdCas9 has two advantageous properties in (1) targetability and (2) versatility, and one disadvantageous property in (3) efficiency for genome editing.

The crystal structures of SpCas9, *Staphylococcus aureus* Cas9 (SaCas9), *Francisella novicida* Cas9 (FnCas9), *Campylobacter jejuni* Cas9 (CjCas9), and *Neisseria meningitidis* Cas9 (NmCas9) revealed the mechanistic divergence in the guide RNA and PAM recognition[24,29,30,69,70]. The sgRNAs of these Cas9 orthologs contains the distinct structural features, such as an additional stem loop (SpCas9 sgRNA), U-shaped linker (FnCas9 sgRNA), and RNA triplex (CjCas9 sgRNA)[24,33,70]. These Cas9 orthologs recognize these structural features and sequences in distinct manners. SpCas9, SaCas9, FnCas9, CjCas9, and NmCas9 recognize the NGG, NNGRRT, NGG, NNNVRYMC, and NNNGATT PAMs, respectively, using distinct sets of amino-acid residues[24,29,30,69,70]. CdCas9 shares the low sequence similarity with SpCas9 (20%), SaCas9 (20%), FnCas9 (15%), CjCas9 (19%), and NmCas9 (20%). Therefore, the recognition mechanism of the distinct guide RNA and the promiscuous NNRHHHY PAM by CdCas9 remains unknown.

3.2 Research aims

Although CdCas9 has the useful properties in genome editing applications and the distinct features in the CRISPR-Cas9 system, the characterization of CdCas9 was not fully investigated. The A-rich PAM recognition mechanism of CdCas9 remains unknown. Therefore, I performed the biochemical and structural analyses of CdCas9 towards the CdCas9-mediated genome editing and the further understanding of the CRISPR-Cas divergence.

3.3 Methods

3.3.1 Sample preparation

The residues 498–663 of CdCas9 (1,084 a.a.) were truncated and connected by a GGGSGG linker because the residues were predicted to be a flexible HNH domain hampering the crystallization. I designate the truncated CdCas9 mutant as CdCas9-ΔHNH. The D10A mutation was introduced into the CdCas9-ΔHNH for preventing the potential DNA cleavage during the crystallization. The CdCas9-ΔHNH DNA sequence was inserted into the pE-SUMO vector for stable expression in bacterial cells. The His_6-SUMO-tagged CdCas9-ΔHNH was expressed at 20 °C overnight in *Escherichia coli* B834 (DE3) (Novagen). For the *in vitro* cleavage experiments, the His_6-SUMO-tagged full-length CdCas9 was expressed at 20 °C overnight in *E. coli* Rosetta 2 (DE3) (Novagen). For the crystallization and in vitro cleavage experiments, the bacterial lysate containing recombinant CdCas9 was batched with Ni-NTA Superflow resin (Qiagen) in buffer A (50 mM Tris-HCl, pH 8.0, 20 mM imidazole, and 1 M NaCl) and eluted with buffer B (50 mM Tris-HCl, pH 8.0, 300 mM imidazole, and 0.3 M NaCl). The His_6-SUMO-tagged CdCas9 was mixed with TEV protease targeting the cleavage site between the His_6-SUMO-tag and CdCas9. The mixture was dialyzed at 4 °C overnight in buffer C (20 mM Tris-

HCl, pH 8.0, 40 mM imidazole, and 0.5 M NaCl) and purified by a Ni-NTA column removing the His_6-SUMO-tag. To remove the non-specifically bound nucleic acids, CdCas9 was purified through a HiTrapHeparin HP column (GE Healthcare), using buffer D (20 mM Tris-HCl, pH 8.0) and buffer E (20 mM Tris-HCl, pH 8.0 and 2 M NaCl). In a previous study, the CdCas9 crRNA and CdCas9 tracrRNA were identified[22]. I designed the 112-nt CdCas9 sgRNA which was the fusion of crRNA and CdCas9 tracrRNA with the GAAA linker. The 112-nt sgRNA was transcribed in vitro, using T7 RNA polymerase and a PCR-amplified DNA template. The transcribed RNA was purified by 8% acrylamide denaturing Urea PAGE. In the crystal structure of the SpCas9-sgRNA-DNA complex, the 28-nt target DNA strand and the 8-nt non-target DNA strand contributed to the crystallization[30]. I reconstituted the CdCas9-ΔHNH, the purified sgRNA, the 28-nt target DNA strand, and the 8-nt non-target DNA strand at a molar ratio of 1:1.5:2.3:2.7. The CdCas9-ΔHNH-sgRNA-DNA complex was isolated through a Superdex 200 Increase column (GE Healthcare), using buffer G (10 mM Tris-HCl, pH 8.0 and 150 mM NaCl).

3.3.2 *in vitro* cleavage assay

The pUC119 plasmid, containing the 24-nt target sequence and the PAM was linearized with EcoRI digestion and used as the substrate for *in vitro* cleavage assays. The EcoRI-linearized plasmid (100 ng, 5 nM) was incubated at 37 °C for 0.25–5 min with the CdCas9-sgRNA complex (50 nM) in 10 μL of reaction buffer containing 20 mM HEPES-NaOH, pH 7.5, 100 mM KCl, 2 mM $MgCl_2$, 1 mM dithiothreitol, and 5% glycerol. The reaction was stopped by the addition of quench buffer containing EDTA (40 mM final concentration) and proteinase K (4 μg). Reaction products were resolved, visualized, and analyzed, using a MultiNA microchip electrophoresis device (Shimadzu).

3.3.3 PAM discovery assay

Zhang and colleagues performed this assay, using the purified CdCas9 which I prepared. To generate a PAM library, ssDNA oligonucleotides (Integrated DNA Technologies), containing seven randomized nucleotides downstream of the 20-nt target sequence, were converted to dsDNA via fill-in with the large Klenow fragment (New England Biolabs) and cloned into the pUC19 plasmid by Gibson cloning (New England Biolabs). The plasmid library was cleaved *in vitro* with the CdCas9-sgRNA complex targeting the PAM-containing region. The uncut plasmid was isolated with 2% agarose gel and Zymoclean Gel DNA Recovery Kit (Zymo Research), PCR-amplified, and sequenced on a MiSeq sequencer (Illumina). The resulting sequence data were analyzed by extracting the seven nucleotide PAM regions, counting the individual PAMs, and normalizing the PAM to the total reads for each sample. For a given PAM sequence, enrichment was calculated as the $\log_2$ ratio compared to a no-protein control, with a 0.01 pseudocount adjustment. The PAMs above an enrichment threshold set to 0.3 were used to generate sequence logos[71].

3.3.4 Indel analysis in mouse zygotes

Hatada and colleagues performed this experiment, using the purified CdCas9 and its sgRNAs which I prepared. All animal procedures were approved by the Animal Care and Experimentation Committee at Gunma University and performed in accordance with approved guidelines. For the superovulation, female B6D2F1 mice (8–10 weeks old, CLEA Japan) were injected with 7.5 units of pregnant mare's serum gonadotropin (PMSG; ASKA Pharmaceutical). Furthermore, 48 hours after PMSG infection, the female B6D2F1 mice were injected with 7.5 units of human chorionic gonadotrophin (hCG; ASKA Pharmaceutical). The superovulated B6D2F1 female mice were mated with B6D2F1 male mice

overnight. Zygotes were collected from oviducts 21 hours after the hCG injection. Then, pronuclei-formed zygotes were placed into the M2 medium. Microinjection of CdCas9-sgRNA into the mouse zygotes was performed, using a microscope equipped with a microinjector (Narishige). The CdCas9-sgRNA complexes were assembled by mixing the purified CdCas9 (40 ng/μL) and its sgRNA (50 ng/μL), targeting the mouse *Tet1EX4* or *Tet1EX12* locus. The CdCas9-sgRNA complexes (1 pL) were injected into the pronuclei of the zygotes. After injection, the zygotes were cultured in M16 medium for 4 days. To detect indels, the targeted region was amplified by PCR, using the genomic DNA extracted from each blastocyst. The PCR products were digested with EcoT14I that cleaves the CdCas9 target site of the unmodified genomes, and then were resolved on agarose gel.

3.3.5 Crystallography

The CdCas9-ΔHNH-sgRNA-DNA complex was crystallized at 20 °C, using the hanging-drop vapor diffusion method. Crystals were obtained by mixing 1 μL of complex solution ($A_{260\,nm}$ = 15) and 1 μL of reservoir solution (0.1 M Tris-HCl, pH 8.0, 22–25% PEG3,350, 0.2 M lithium sulfate, 0.3 M potassium fluoride). The crystals were cryoprotected in a buffer containing 0.1 M Tris-HCl, pH 8.0, 25% PEG3,350, 0.2 M lithium sulfate, 0.3 M potassium fluoride, and 20% ethylene glycol. The X-ray diffraction data were collected at 100 K on beamline BL41XU at SPring-8 and processed using DIALS[72] and AIMLESS[59]. The structure was determined by the Se-SAD method, using PHENIX AutoSol[62]. The structure model was automatically built using Buccaneer[73], followed by manual model building using COOT[61] and structural refinement using PHENIX[62]. Structural figures were prepared using CueMol (http://www.cuemol.org).

3.4 Results

3.4.1 Sample preparation for the biochemical analysis

I prepared the CdCas9 proteins and the CdCas9 sgRNAs. The 10–20% acrylamide SDS-PAGE analysis showed that the CdCas9 proteins were produced in high purity (Fig. 15A). The 8 % acrylamide denaturing Urea PAGE analysis showed that the CdCas9 sgRNAs were produced in high purity (Fig. 15B). The 112–116-nt CdCas9 sgRNAs with 20–24 nt guide were observed as 75–100-nt RNAs in the denaturing Urea PAGE. The observation indicated that the CdCas9 sgRNAs formed the secondary structures even in the denaturing polyacrylamide gel.

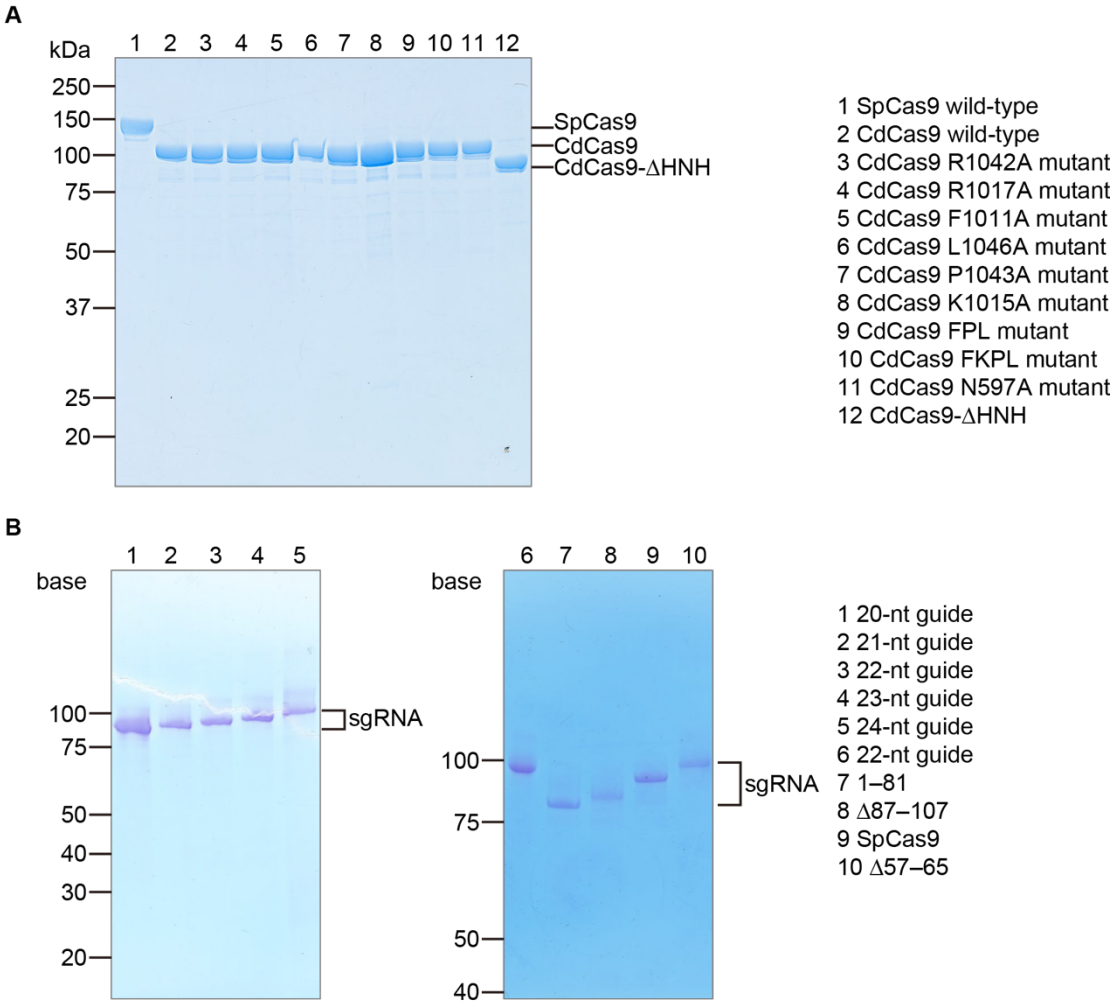

Fig. 15. Sample preparation for the biochemical and structure analyses.

(A) SDS-PAGE analysis of the CdCas9 proteins.

FPL, F1011A/P1043A/L1046A; FKPL, F1011A/K1015A/P1043A/L1046A.

(B) Denaturing Urea PAGE analysis of the CdCas9 sgRNAs.

20/21/22/23/24-nt guide, the full-length CdCas9 sgRNA with 20/21/22/23/24-nt guide; 1–81, the CdCas9 sgRNA, in which nucleotides 82–112 were truncated; Δ87–107, the CdCas9 sgRNA, in which nucleotides 87–107 were replaced with GAAA; SpCas9, the SpCas9 sgRNA with 22-nt guide; Δ57–65, the CdCas9 sgRNA, in which nucleotides 57–65 were replaced with GAAA.

3.4.2 CdCas9 PAM specificity

I examined the optimal guide RNA length, using sgRNA with 20–24-nt guide sequences. The *in vitro* cleavage data showed that the sgRNA with 22-nt guide sequence induces the target cleavage by CdCas9 most efficiently (Fig. 16A). The cleavage activity of CdCas9 is comparable to that of SpCas9 (Fig. 16B). Zhang and colleagues showed that CdCas9 recognizes the NNRHHHY PAM through the PAM discovery assay, using the NNNNNNN PAM library (Fig. 16C). I examined the cleavage activity of CdCas9, using the 28 plasmid targets containing the 28 PAMs (Fig. 16D). CdCas9 shows the comparable activity towards the NGGAAAC, GNGAAAC, GGRAAAC, GGGHAAC, GGGAAAY PAMs (Fig. 16D). CdCas9 is less active towards the GGYAAAC, GGGGAAC, GGGABAC, GGGAABC, GGGAAAR PAMs (Fig. 16D). The substation with three thymines/cytosines and guanines at the positions 4–6 in the GGGAAAC PAM reduced and abolished the cleavage activity of CdCas9, respectively (Fig. 16D). The substation with two guanines at the positions 4–6 in the GGGAAAC PAM reduced or abolished the cleavage activity of CdCas9 (Fig. 16D). CdCas9 prefers A-rich sequence and tolerates single G substitution, but not multiple G substitution, at the positions 4–6 in the PAM (Fig. 16D). Taken together, CdCas9 recognizes the NNRHHHY PAM in a promiscuous manner.

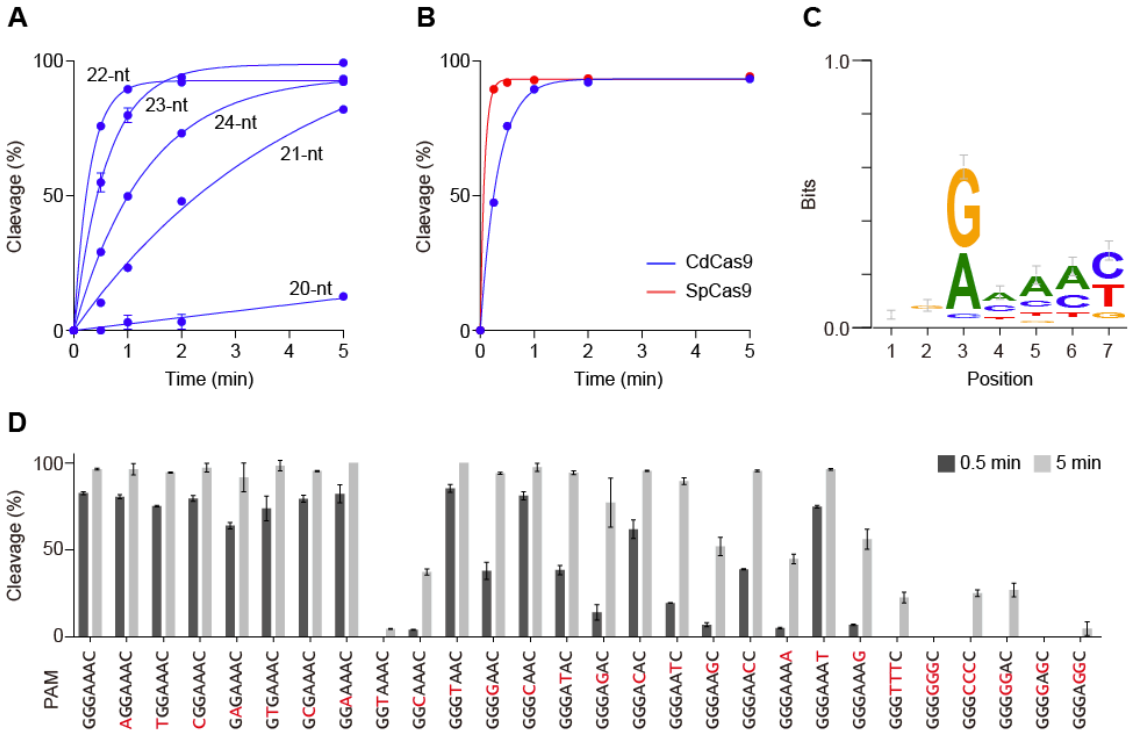

Fig. 16. *in vitro* cleavage activities of CdCas9.

(A) *in vitro* cleavage activities of CdCas9 with the sgRNAs containing 20–24-nt guide sequences. The target plasmid containing the GGGAAAC PAM was incubated with the CdCas9-sgRNA complex at 37 °C for 0.5, 1, 2, and 5 min.

(B) *in vitro* cleavage activities of CdCas9 and SpCas9. CdCas9 and SpCas9 were programmed with their 22- and 20-nt guide sgRNAs, respectively. The target plasmid containing the GGGAAAC PAM was incubated with the CdCas9-sgRNA complex at 37 °C for 0.25, 0.5, 1, 2, and 5 min.

(C) Motif obtained from the PAM discovery assay. Zhang and colleagues performed this assay.

(D) PAM preference of CdCas9. The target plasmids containing the different PAMs were incubated with the CdCas9-sgRNA (22-nt guide) complex at 37 °C for 0.5 and 5 min. Error bars represent s.d. from $n = 3$ replicates.

3.4.3 CdCas9-mediated genome editing

In a previous study, CdCas9 fails to show its activity towards the target site in the genome, using its sgRNA with 20-nt guide[22]. In this study, the biochemical data showed the sgRNAs with 22-nt and 24-nt guides induce CdCas9-mediated DNA cleavage more efficiently than the sgRNA with 20-nt guide. To examine whether CdCas9 shows its activity in cells, using the optimal guide RNA, Hatada and colleagues performed the indel analysis of CdCas9-mediated genome editing in mouse zygotes. The microinjection of the CdCas9-sgRNA complexes into mouse zygotes induced the insertion and deletion (indel) at the target sites in the genome. At the *Tet1EX4* target site with the GTATAAT PAM, the CdCas9-sgRNA complex with 22-nt guide induced the indel, while the CdCas9-sgRNA complex with 20/24-nt guide did not induce the indel (Table 4). At the *Tet1EX12* target site with the TGGTAAT PAM, the CdCas9-sgRNA complex with 22/24-nt guide induced the indel, while the CdCas9-sgRNA complex with 20-nt guide did not induce the indel (Table 4). Notably, at the *Tet1EX12* target site, the sgRNA with 24-nt guide induced CdCas9-mediated genome editing more efficiently than the sgRNA with 22-nt guide, although 22-nt is the most optimal guide length in the biochemical experiment (Table 4). Therefore, the findings in the biochemical experiment enabled CdCas9-mediated genome editing in mammalian cells.

Table 4. CdCas9-mediated genome editing in mouse zygotes

Locus	Target seuqence	PAM	Indel (%)
Tet1EX4	TTGGTCCTGCCCCAAGGTGT (20-nt)	GTATAAT	0 (0/23 embryos)
Tet1EX4	ACTTGGTCCTGCCCCAAGGTGT (22-nt)	GTATAAT	5 (1/21 embryos)
Tet1EX4	ACACTTGGTCCTGCCCCAAGGTGT (24-nt)	GTATAAT	0 (0/17 embryos)
Tet1EX12	ACCCTTACCCTGGAGTTCCA (20-nt)	TGGTAAT	0 (0/19 embryos)
Tet1EX12	TCACCCTTACCCTGGAGTTCCA (22-nt)	TGGTAAT	8 (2/24 embryos)
Tet1EX12	TGTCACCCTTACCCTGGAGTTCCA (24-nt)	TGGTAAT	56 (10/18 embryos)

3.4.4 Crystallization and structural determination of CdCas9 mutants

I made a plasmid construct of CdCas9-ΔHNH containing the D10A mutation. Similar to the wild-type CdCas9 protein, the CdCas9-ΔHNH protein and its cognate sgRNA were produced in high yield and purity (Fig. 15A and 15B). The chromatography profiles showed that the CdCas9–sgRNA–target DNA complex is stable in the solution (Fig. 17). I obtained about 200-μm length crystals of CdCas9-ΔHNH protein in complex with their sgRNAs and target DNAs containing the GGGTAAT PAM (Fig. 18A). The analysis of X-ray diffraction experiment showed that the collected data of CdCas9-ΔHNH was at 2.9 Å resolution (Fig. 18B, 18C, and Table 4). The refinement and validation software, PHENIX, shows that the R_{work} / R_{free} value of the final CdCas9-ΔHNH structure model is 0.221 / 0.254 (Table 5).

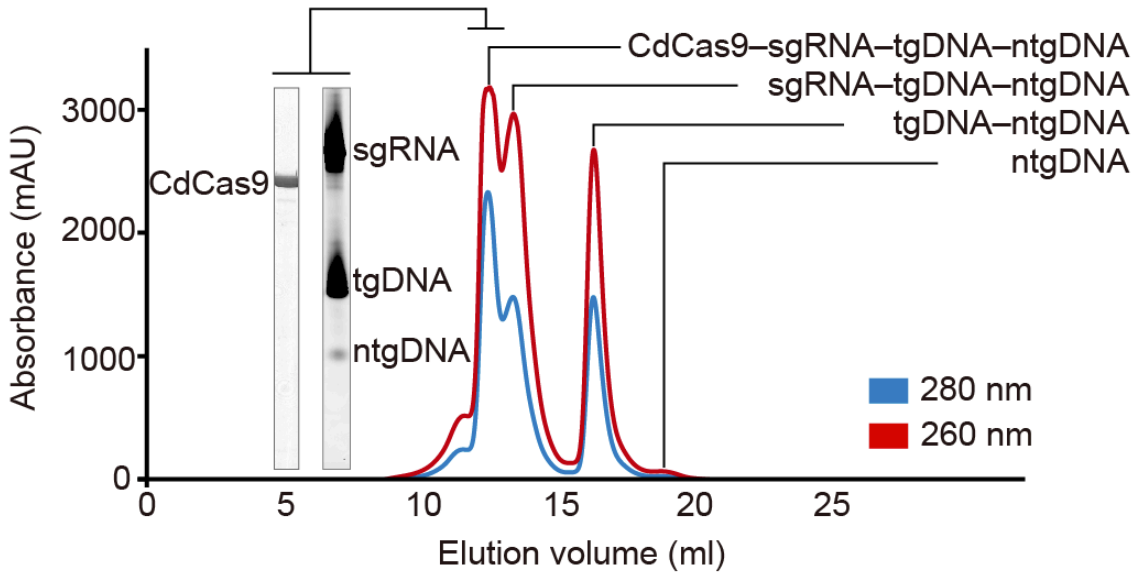

Fig. 17. The size exclusion chromatography profile of the CdCas9–sgRNA–target DNA complex.

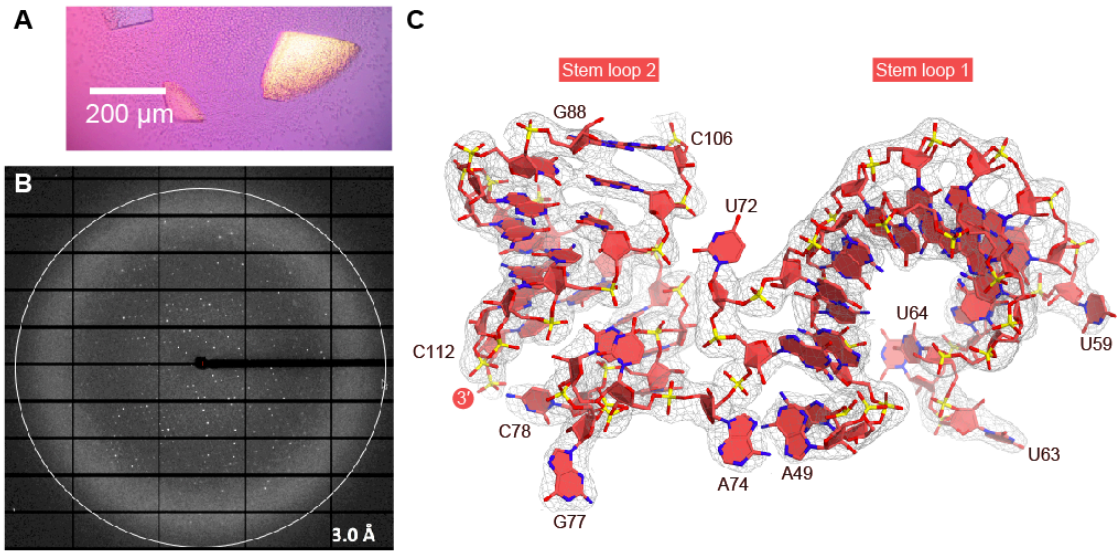

Fig. 18. X-ray crystallography of CdCas9.

(A) The CdCas9 crystal.

(B) A diffraction image obtained from the CdCas9 crystal.

(C) $2mF_O - DF_C$ omit electron maps of the CdCas9 sgRNA scaffold (contoured at 1.5σ).

Table 5. Data collection and refinement statistics.

Data collection	
Beamline	SPring-8 BL41XU
Wavelength (Å)	0.9790
Space group	*C*2
Cell dimensions	
a, *b*, *c* (Å)	139.0, 119.0, 116.3
α, *β*, *γ* (°)	90, 113.6, 90
Resolution (Å)[a]	106.6–2.9 (3.03–2.90)
R_{merge}	0.168 (3.024)
R_{pim}	0.047 (0.844)
$I/\sigma I$	10.2 (1.4)
Completeness (%)	100.0 (100.0)
Multiplicity	13.4 (13.5)
CC(1/2)	0.999 (0.802)
Refinement	
Resolution (Å)	67.9–2.9 (3.00–2.90)
No. reflections	38,462 (3,795)
R_{work} / R_{free}	0.221 / 0.254 (0.399 / 0.469)
No. atoms	
Protein	6,292
Nucleic acid	2,755
Ion	1
Solvent	11
B-factors (Å^2)	
Protein	116.1
Nucleic acid	112.0
Ion	121.5
Solvent	72.8
R.m.s. deviations	
Bond lengths (Å)	0.003
Bond angles (°)	0.54
Ramachandran plot (%)	
Favored region	96.94
Allowed region	2.94
Outlier region	0.12
MolProbity score	
Clashscore	6.44
Rotamer outlier	5.00

[a]Values in parentheses are for the highest resolution shell.

3.4.5 Crystal structure of CdCas9

The obtained crystal structure of CdCas9-ΔHNH showed that CdCas9 adopts a bilobed architecture consisting of REC lobe (residues 86–448) and NUC lobe (residues 1–51 and 449–1084) (Fig. 19A and 19C). The two lobes are connected by a bridge helix (residues 52–85) (Fig. 19C). The REC lobe comprises the REC1 (residues 86–235) and REC2 domains (residues 236–448) (Fig. 19C). The NUC lobe comprises the RuvC (residues1–51, 449–497, and 664–807), WED (residues 821–904), and PI (residues 905–1084) domains(Fig. 19C). The HNH domain (residues498–663) was truncated for crystallization (Fig. 19A). The RuvC and WED domains are connected by a phosphate lock loop (residues 808–820) (Fig. 19C). The sgRNA consists of the guide region (G1–C20), repeat:anti-repeat duplex (A21:U48–G32:C37), tetraloop (G33–A36), stem loop 1 (A50–G73), a single-stranded linker (A74–C81), and stem loop 2 (G82–C112) (Fig. 19B and 19D). The guide region and the target DNA strand forms the guide:target heteroduplex (Fig. 19D). The repeat:anti-repeat duplex consists of ten Watson-Crick base pairs (C22:G47–G24:C45 and G26:C43–G32:C37) and two non-canonical base pairs (A21:U48 and G25:U44) (Fig. 19D). Stem loop 1 consists of seven Watson-Crick base pairs (C52:G70–G57:C65 and C58:C61), two non-canonical base pairs (A50:G73 and G51:U71) and six unpaired nucleotides (U59, C60, U62–U64, and U72) (Fig. 19D). The basal region of stem loop 2 consists of six Watson-Crick base pairs (G82:C112–G86:C108 and G88:C106) and a non-canonical base pair (U87:G107) (Fig. 19D). The upper region of stem loop 2 (C89–G105) exhibits no electron density, indicating that this region is flexible (Fig. 18C and 19D). The sgRNA-target DNA heteroduplex is accommodated in the central channel between the REC and NUC lobes (Fig. 19C). The PAM-containing DNA duplex is bound to the groove between the WED and PI domains (Fig. 19C). The sgRNA consisting of three stem loops is recognized extensively by all the domains except for the REC2 domain (Fig. 19C). Overall structure of CdCas9 in the sgRNA/target DNA-bound state is similar to those of other Cas9 orthologs, indicating that the RNA-guided DNA targeting mechanism is conserved in CdCas9.

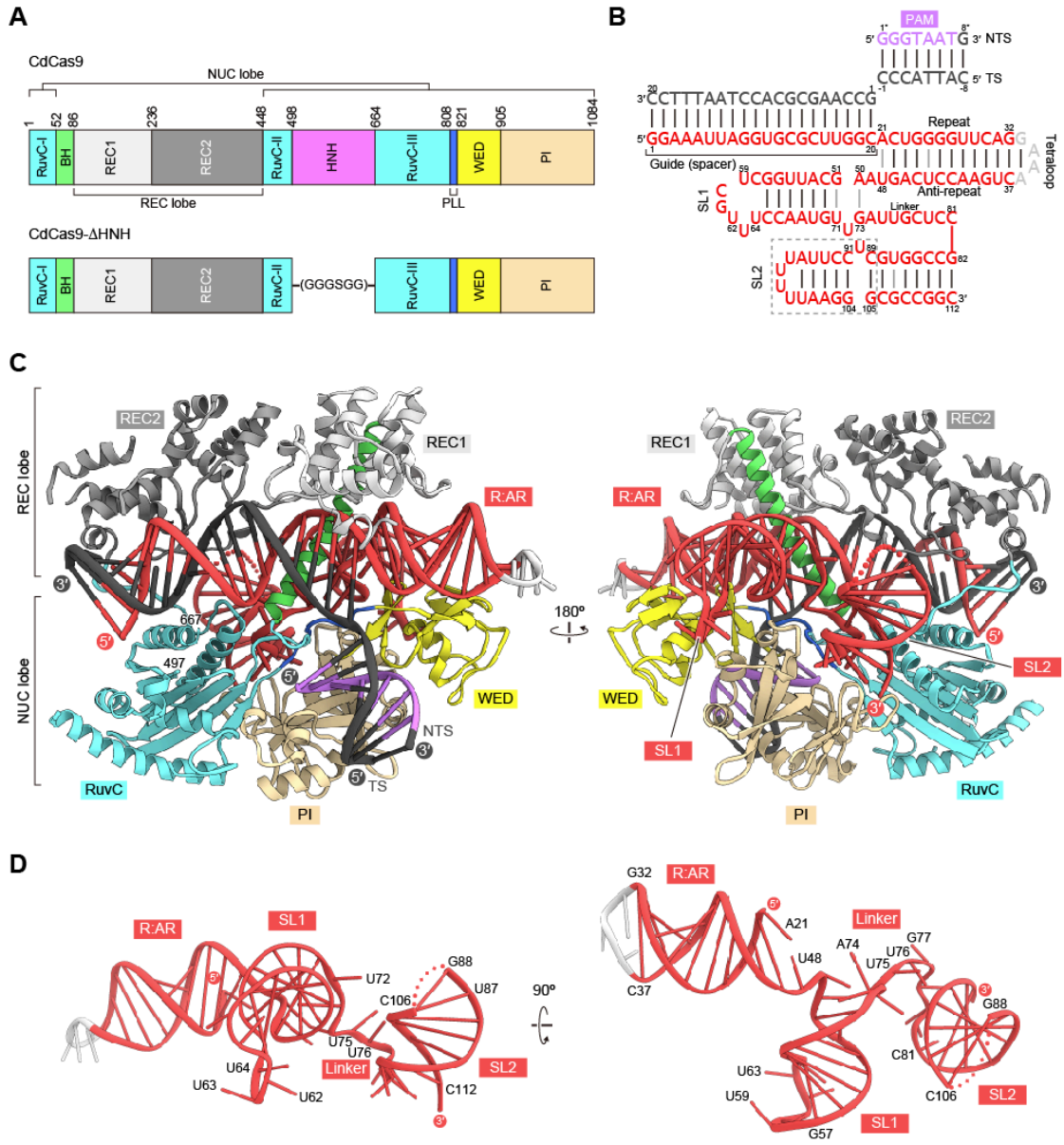

Fig. 19. Crystal structure of CdCas9-sgRNA-target DNA complex.

(A) Domain structure of CdCas9. BH, bridge helix; PLL, phosphate-lock loop.

(B) Schematics of the sgRNA and the target DNA. The disordered region is shown in a dashed box. SL1, stem loop 1; SL2, stem loop 2; TS, target strand; NTS, non-target strand.

(C) Overall structure of CdCas9-ΔHNH in complex with the sgRNA and the target DNA. The disordered region of the sgRNA is shown as a dotted line. R:AR, repeat:anti-repeat duplex; SL1, stem loop 1; SL2, stem loop 2.

(D) Structure of the sgRNA scaffold. The guide region is omitted for clarity. The disordered region is shown as a dotted line.

3.4.6 Recognition of the CdCas9 sgRNA

CdCas9 recognizes the basal and upper regions of stem loop 1 (A50–G51, U62–U64, and G70–G73) (Fig. 20 and 21A). The deletion of nucleotides 57–65 in the sgRNA reduced the CdCas9-mediated DNA cleavage activity, indicating that the upper region contact is important for the protein-RNA interaction (Fig. 21B). CdCas9 recognizes the basal regions of repeat:anti-repeat duplex and stem loop 2 (A21–U28, C43–U48, and C109–C112), while CdCas9 does not recognize the upper regions of repeat:anti-repeat duplex and stem loop 2 (U29–A40 and G86–C108) (Fig. 20 and 21A). The sgRNA truncation of nucleotides 82–112, but not nucleotides 87–107, reduced the CdCas9-mediated DNA cleavage activity, indicating that the basal region of stem loop 2 is recognized by CdCas9 and the upper region of stem loop 2 is solvent-exposed (Fig. 21A and 21B). In CdCas9, the basal region in the repeat:anti-repeat duplex and stem loop 1 is recognized by the REC1/WED domains, the bridge helix, and the phosphate lock loop (Fig. 20 and 21A). The kinked linker and the basal region of stem loop 2 are recognized by the binding surface of the RuvC/PI domains, the bridge helix, and the phosphate lock loop in base-specific manners (Fig. 20 and 21A). The flipped-out G77 and C78 nucleobases form the hydrogen-bonding interactions with Asn977 and Arg1070, respectively (Fig. 20 and 21C). The stacked U79, C80, and C81 nucleobases form the hydrogen-bonding interactions with Asp939, His1076, and His1076, respectively (Fig. 20 and 21C). Furthermore, CdCas9 did not cleave the target plasmid in the presence of the SpCas9 sgRNA (Fig. 21B and 21D). Taken together, these results showed that CdCas9 recognize its distinct scaffold sgRNA in an orthogonal manner.

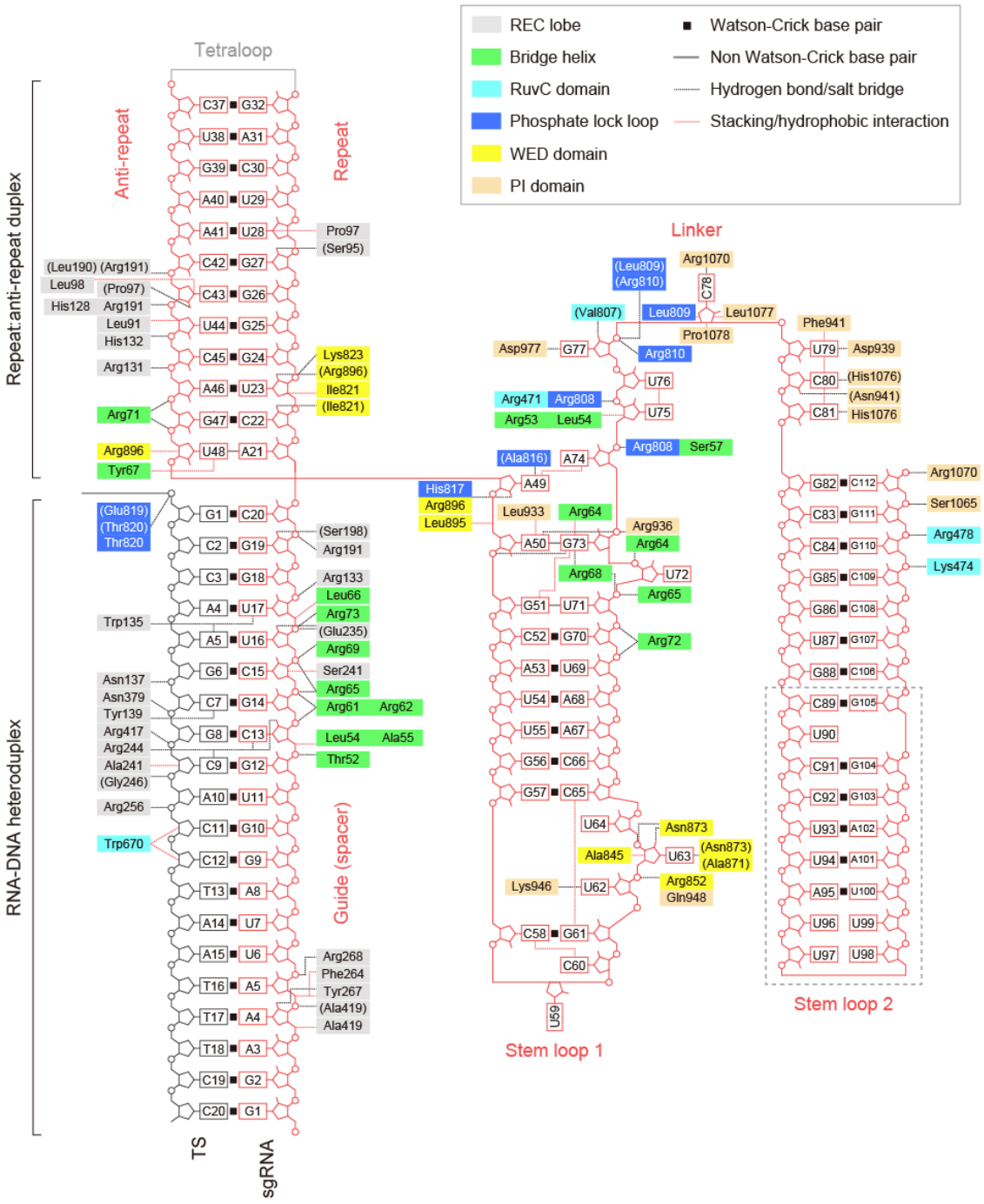

Fig. 20. Schematic of the nucleic acid recognition by CdCas9.

The residues that interact with the sgRNA and target DNA via their main chains are shown within parentheses. The disordered region is enclosed in a dashed box.

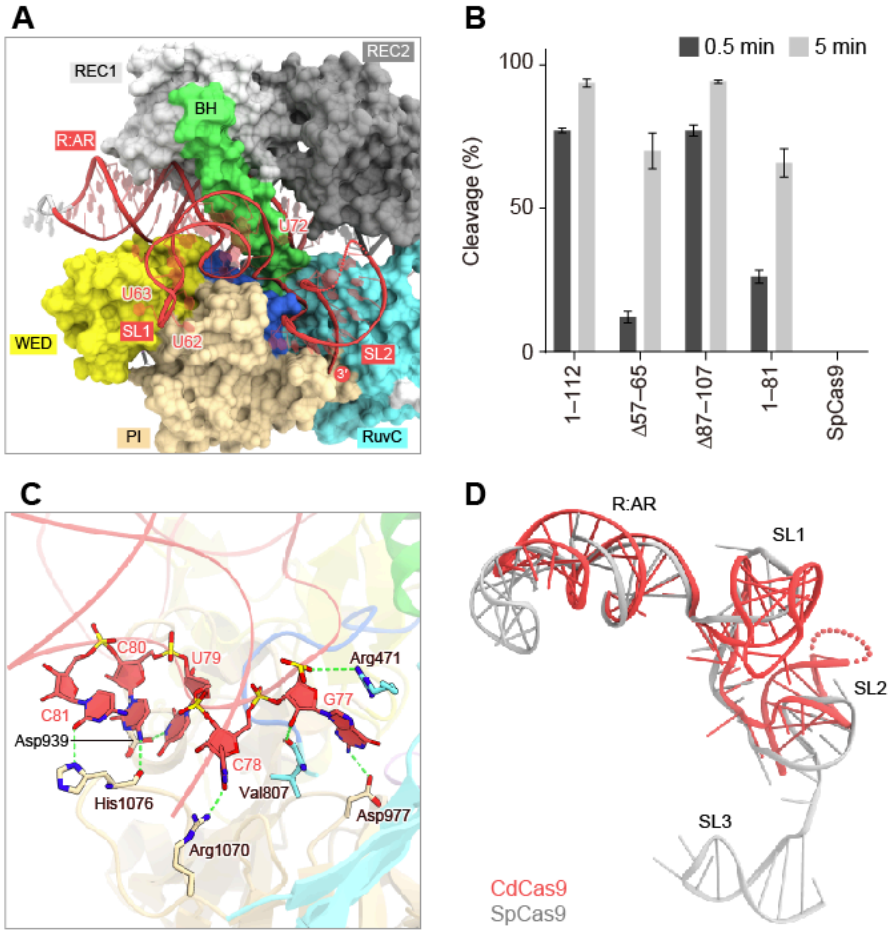

Fig. 21. Recognition of the CdCas9 sgRNA.

(A) Recognition of the sgRNA scaffold. The disordered region is shown as a dotted line. R:AR, repeat:anti-repeat duplex; SL1, stem loop 1; SL2, stem loop 2.

(B) In vitro cleavage activities of the truncated sgRNAs. The target plasmids containing the different PAMs were incubated with the CdCas9-sgRNA (22-nt guide) complex at 37 °C for 0.5 and 5 min. 1–114, the full-length CdCas9 sgRNA; Δ57–65, the CdCas9 sgRNA, in which nucleotides 57–65 were replaced with GAAA; Δ87–107, the CdCas9 sgRNA, in which nucleotides 87–107 were replaced with GAAA; 1–81, the CdCas9 sgRNA, in which nucleotides 82–114 were truncated; SpCas9, the SpCas9 sgRNA with 22-nt guide. Error bars represent s.d. from n = 3 replicates.

(C) Base-specific recognition of the sgRNA linker region. Hydrogen bonds are shown as dashed lines.

(D) Superposition of the CdCas9 sgRNA with the SpCas9 sgRNA (PDB: 4OO8). The disordered region is shown as a dotted line.

3.4.7 Recognition of the NNRHHHY PAM

The GGGTAAT PAM sequence is read from the major groove side in the PAM-binding groove (Fig. 22A). Both the PAM nucleobases (dG2* and dG3*) and the PAM-complemental nucleobases (dA(−4), dT(−5), dT(−6), and dA(−7)) are recognized by the PI domain (Fig. 22B and 22C). The dG1* nucleobase has no interaction with the CdCas9 protein, consistent with no preference of the first position in the PAM (Fig. 16C, 16D, and 22C). The dG2* nucleobase forms bidentate hydrogen-bonds with Arg1042 (Fig. 22C). The interaction may be required for the suboptimal PAM, but not the optimal NNRHHHY PAM, because CdCas9 shows no preference of the second position in the GNGAAAC PAM (Fig. 16D). The dG3* nucleobase forms a single hydrogen-bond with Arg1017, consistent with the third R requirement in the PAM (Fig. 16C, 16D, and 22C). The dA(−4), dT(−5), and dT(−6) nucleobases has no hydrogen-binding interaction with the CdCas9 protein (Fig. 22C). The modeling three adenines, thymines, guanines, and cytosines into the nucleobases shows that the modeled nucleobases are in the vicinity of the Phe1011, Lys1015, Pro1043, and Leu1046 (I designate these residues as a hydrophobic patch) (Fig. 23). The modeling suggested that the hydrophobic patch forms the hydrophobic interactions with methyl groups of thymines and crashes with the 4-amino groups of cytosines, explaining the observation that CdCas9 prefers the A-rich PAMs and rejects the G-rich PAMs (Fig. 16D and 23). The dA(−7) nucleobase forms a single hydrogen-bond with Lys1015, consistent with the seventh Y requirement in the PAM (Fig. 16C, 16D, and 22C). The mutational assay confirmed that the set of residues (Phe1011, Lys1015, Arg1017, Pro1043, and Leu1046) are important for the PAM recognition by CdCas9 (Fig. 22D). Taken together, CdCas9 recognizes the NNRHHHY PAM through hydrogen-bonding and hydrophobic interactions.

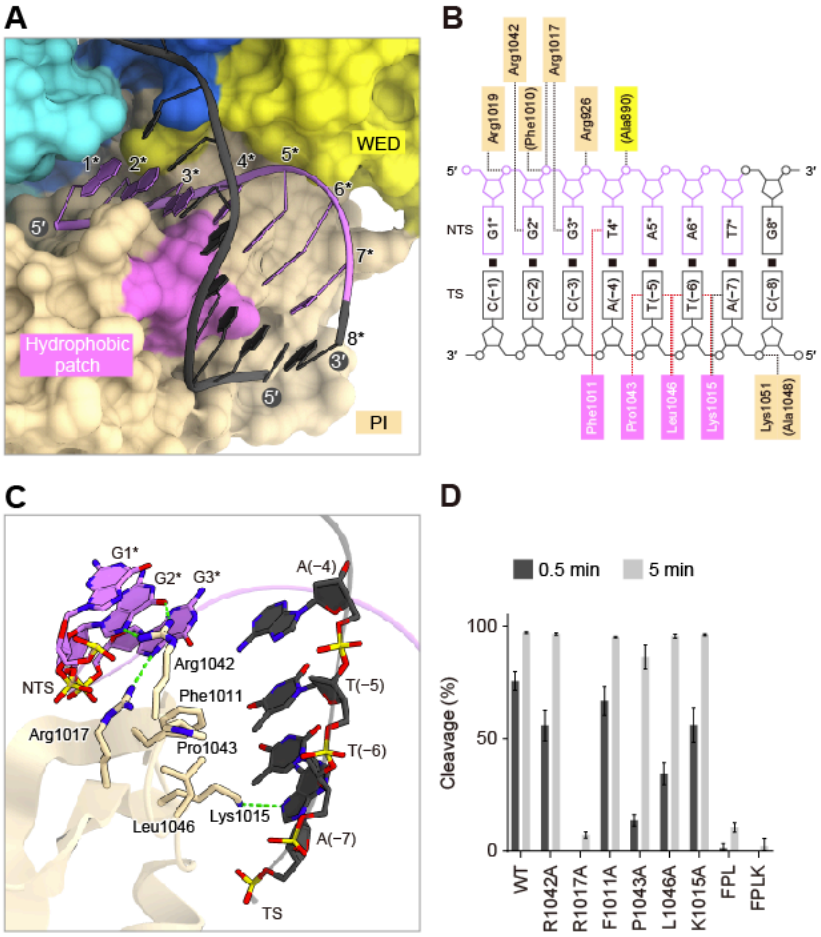

Fig. 22. PAM recognition by CdCas9.

(A) Binding of the PAM duplex to CdCas9. The PAM in the target strand is colored in purple.

(B) Schematics of the PAM recognition by CdCas9. Hydrogen bonds and hydrophobic interactions are depicted by black and red dashed lines, respectively. TS, target strand; NTS, non-target strand.

(C) Recognition of the GGGTAAT PAM. Hydrogen bonds are shown as dashed lines.

(D) Effects of the mutations on the PAM-interacting residues. The target plasmids containing the different PAMs were incubated with the CdCas9-sgRNA (22-nt guide) complex at 37 °C for 0.5 and 5 min. FPL, F1011A/P1043A/L1046A; FPLK, F1011A/K1015A/P1043A/L1046A. Error bars represent s.d. from n = 3 replicates.

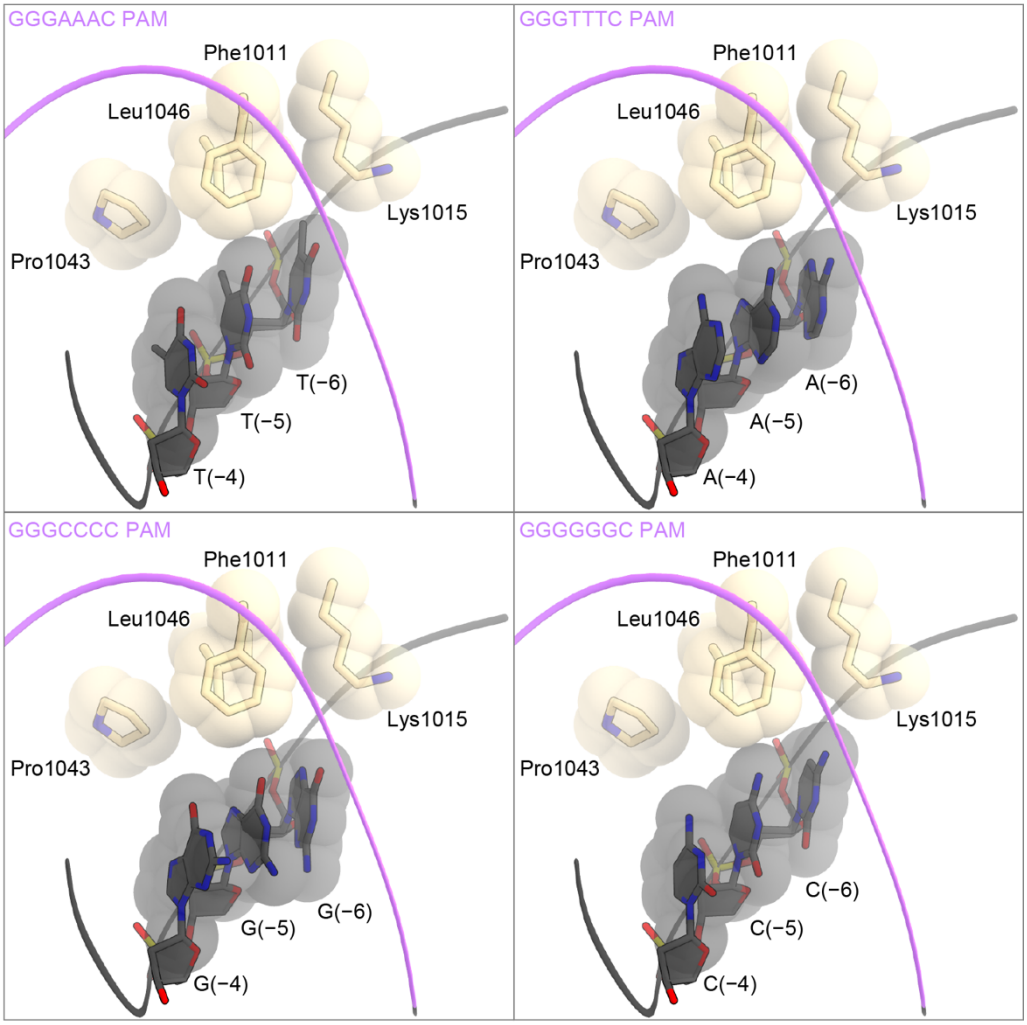

Fig. 23. PAM recognition by the hydrophobic patch.

Based on the CdCas9 structure with the GGGTAAT PAM, AAA, TTT, CCC, and GGG were modeled at positions 4–6 in the target strand.

3.5 Discussion

Structural comparison of CdCas9 with other Cas9 orthologs highlighted the functional convergence and divergence of the CRISPR-Cas9 systems and provided the insights into the CdCas9 properties related to the genome editing application. Overall structures of CdCas9, SpCas9, SaCas9, FnCas9, CjCas9, and NmCas9 in complex with their cognate sgRNAs and target DNAs are similar, indicating that CdCas9 shared the RNA-guided DNA cleavage mechanism with other Cas9s[24,29,30,69,70] (Fig. 24). The REC1 domains of CdCas9, SaCas9, NmCas9, and CjCas9 are more compact than those of SpCas9 and FnCas9, due to the lack of the insertions in the REC1 domains (Fig. 24). The PI domains of CdCas9, SaCas9, FnCas9 NmCas9, and CjCas9 are more compact than that of SpCas9, due to the lack of the insertion in the PI domain (Fig. 24). The WED domains of CdCas9, SaCas9, and NmCas9 are more compact than that of FnCas9, due to the distinct domain folding of the WED domain (Fig. 24). The WED domain shrinks in SpCas9 and CjCas9 (Fig. 24). The WED domain folding of CdCas9 are similar to those of SaCas9, and NmCas9, while the WED domain position of CdCas9 is different from those of SaCas9 and NmCas9 (Fig. 24). The different positions of the WED domains are related to the sgRNA recognition by these Cas9s[29,69]. The Cas9 sizes, which are important in genome editing applications, mainly depend on the REC1, WED, and PI domains. The crystal structure of CdCas9 showed that CdCas9 is miniaturized in a similar manner to SaCas9 and NmCas9 and CdCas9 expands structural repertoire in the WED domain position.

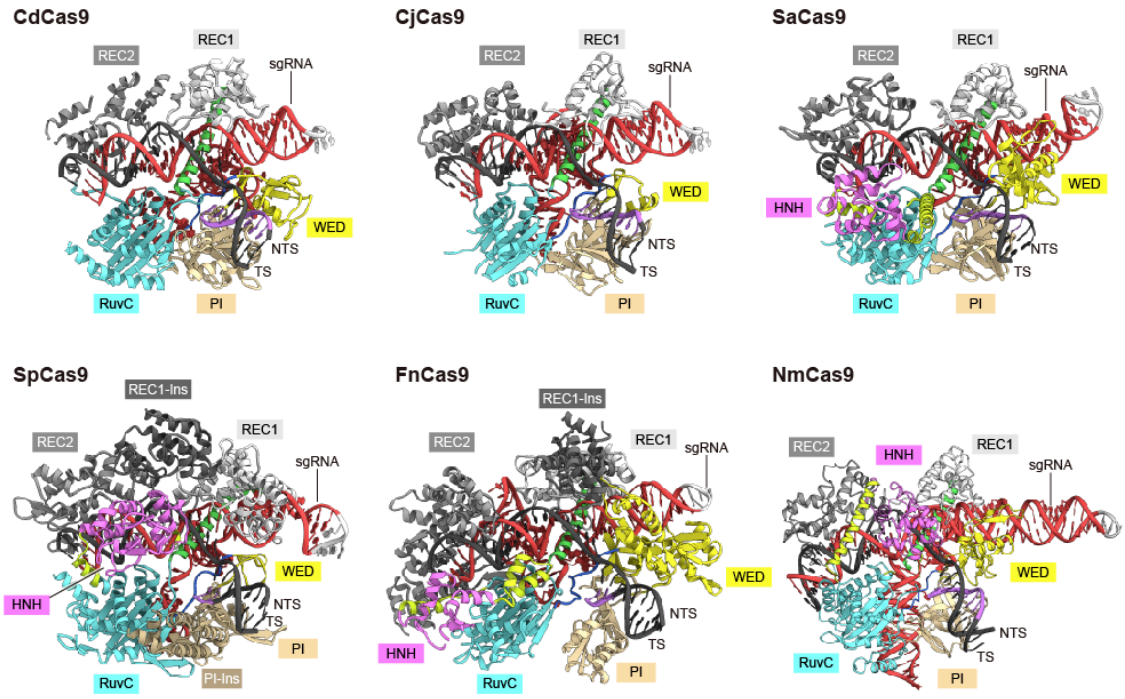

Fig. 24. Overall structures of CdCas9, CjCas9 (PDB: 5X2G), SaCas9 (PDB: 5CZZ), SpCas9 (PDB: 4UN3), FnCas9 (PDB: 5B2O), NmCas9 (PDB: 6JDV).

The biochemical experiment showed that the CdCas9 sgRNA with 22-nt guide length induces the CdCas9-mediated target cleavage most efficiently (Fig. 16A). The mouse zygote experiment showed that the CdCas9 sgRNA with 22–24-nt guide length induces the CdCas9-mediated genome editing efficiently (Table 4). I explain the DNA cleavage mechanism of CdCas9 and the reason why the optimal guide length of the CdCas9 sgRNA is 22–24-nt.

Before explaining the DNA cleavage mechanism of CdCas9, I explain three topics about the DNA cleavage mechanism of SpCas9 and NmCas9. (1) The optimal guide lengths are different between SpCas9 and NmCas9. (2) The linker helixes between the RuvC and HNH domains are important for DNA cleavage. (3) The insertion helix in the RuvC of NmCas9 may be important for the DNA cleavage. As for (1) the optimal guide length, a previous study showed that the sgRNAs with 20-nt and 30-nt guides induce SpCas9-mediated genome editing at the same efficiency[74]. The 30-nt guide is trimmed into the 20-nt guide in mammalian cells[74]. Another study showed that the sgRNAs with 22–24-nt guides, rather than 20-nt guide, induce NmCas9-mediated genome editing efficiently[75]. These studies indicated that the optimal guide lengths of SpCas9 and NmCas9 are 20-nt and 22–24-nt, respectively. As for (2) the linker helix, previous studies showed that the HNH and RuvC domains of Cas9 allosterically cleave the target DNA strand and the non-target DNA strand via linker helixes, respectively[76]. The structures of SpCas9 and NmCas9 in the catalytic states showed that the linker helixes stabilize the HNH domains at the target cleavage site[32,69] (Fig. 25A and 25B). These studies showed that the linker helix is important for stabilizing the active state. As for (3) the insertion helix, in the NmCas9 structure, the insertion helix adjacent to the RuvC active site forms the interaction network with its sgRNA and the linker helix, stabilizing the active state of NmCas9[69] (Fig. 25A).

Notably, more than 22-nt guide of NmCas9 sgRNA is required for forming the interaction network, consistent with the optimal guide length of NmCas9 is 22–24-nt[69,75] (Fig. 25A). This interaction network was not observed in the SpCas9 structure due to the lack of the RuvC insertion helix in SpCas9[32] (Fig. 25B). The structural feature is consistent with the optimal guide length of SpCas9 is 20-nt, suggesting that SpCas9 may not require the interaction network for stabilizing the active state.

The RuvC insertion helix is conserved in CdCas9 (Fig. 25C), suggesting the possibility of the interaction network formation among the insertion helix, the linker helix, and the CdCas9 sgRNA in the active state of CdCas9 (Fig. 25D). The HNH domain and the linker helixes, which were truncated for the crystallization, are conserved in CdCas9 (Fig. 19A). Given that the optimal guide length of CdCas9 is 22–24-nt, CdCas9 may utilize its sgRNA with 22–24-nt guide length for forming the interaction network and stabilizing the active state (Fig. 16A and 25D). Taken together, this study highlights the utility of the guide RNA optimization with a mechanical hypothesis.

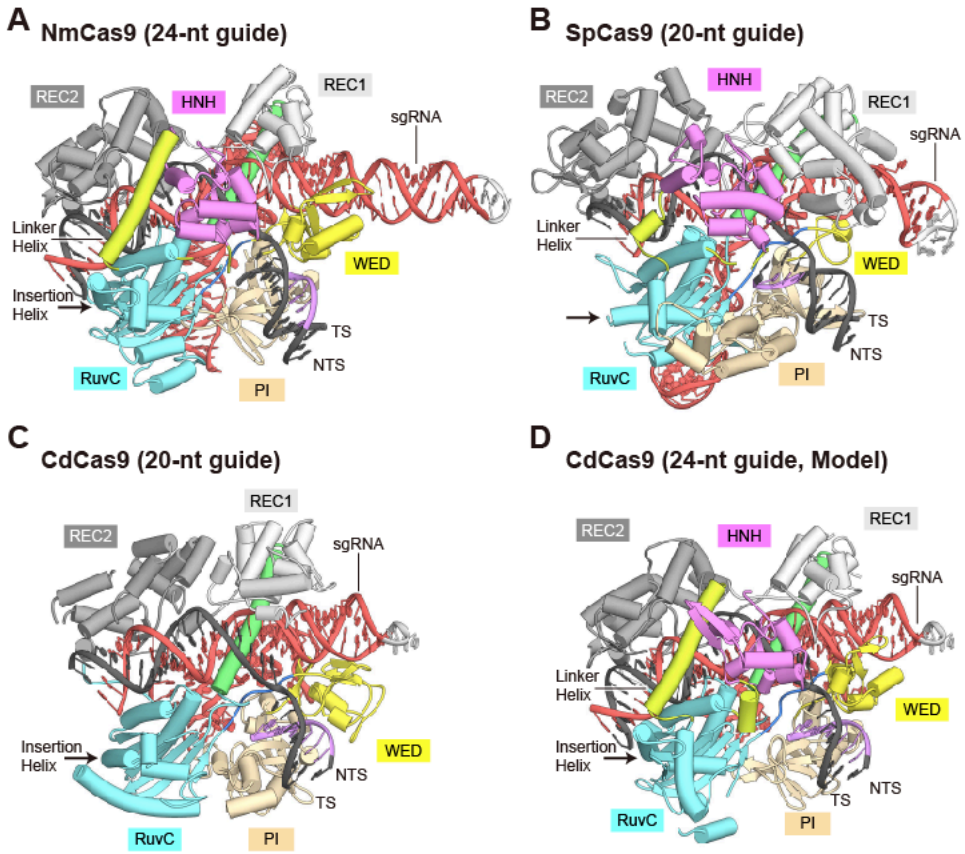

Fig. 25. Active states of Cas9s.

(A) Crystal structure of NmCas9 in the active state (PDB: 6JDV).

(B) Crystal structure of SpCas9 in the active state (PDB: 6O0Y).

(C) Crystal structure of CdCas9 in the inactive state.

(D) Homologymodel of CdCas9 based on the active state structure of NmCas9 (PDB: 6JDV).

Structural comparison of CdCas9 with other Cas9s revealed the distinct mechanism of guide RNA recognition by CdCas9. Cas9 orthologs recognize their cognate sgRNAs consisting of the guide region, the repeat:anti-repeat region, and the 3′-terminal region in distinct manners. The 3′-terminal region of CdCas9 sgRNA containing stem loop 1, stem loop 2, and SL1-SL2 linker, is distinct from that of SpCas9 sgRNA containing stem loop 1, stem loop 2, SL1-SL2 linker, and stem loop 3, and that of CjCas9 sgRNA containing RNA triple helix instead of stem loops[33,70] (Fig.). The 3′-terminal region composition of CdCas9 sgRNA is similar to those of SaCas9, FnCas9, and NmCas9 sgRNAs[24,29,69] (Fig. 24). In CdCas9, the basal and upper regions of stem loop 1 is recognized by the bridge helix and the WED domain, respectively. In SaCas9 and NmCas9, the basal and upper regions of stem loop 1 is recognized by the bridge helix, extensively . The SL1-SL2 linker of CdCas9 sgRNA is single-stranded, while that of FnCas9 sgRNA is U-shaped. These differences of sgRNA scaffolds and recognition manners contribute to the RNA-binding affinities of Cas9s. The biochemical studies showed that CdCas9 bind its sgRNA less tightly, as compared to SpCas9. Due to the lack of comprehensive analyses for the correlation between the binding affinities and the Cas9–sgRNA interactions, this study cannot identify the structural features which may hamper the CdCas9–sgRNA complex formation and the robust genome editing in mammalian cells. Taken together, the crystal structure of CdCas9 indicates the possibility of strategy for the CdCas9-mediated robust genome editing, such as enhancing the sgRNA-binding affinity.

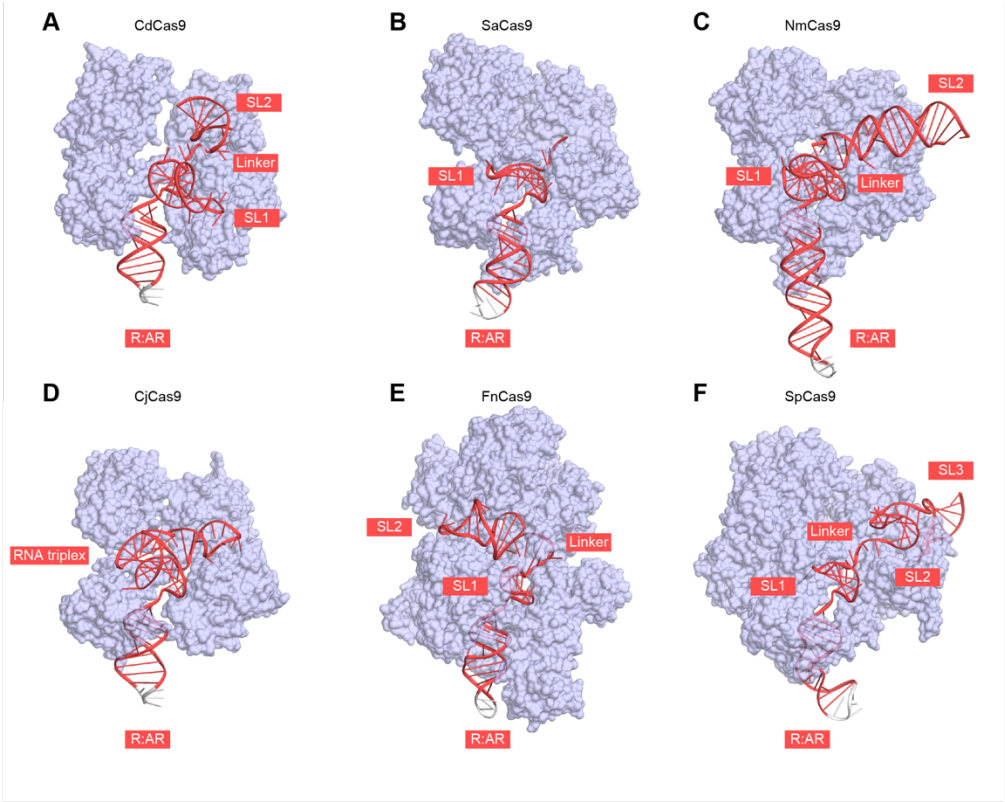

Fig. 26. guide RNA recognition by CdCas9 (A), SaCas9 (PDB: 5CZZ) (B), NmCas9 (PDB: 6JDV) (C), CjCas9 (PDB: 5X2G) (D), FnCas9 (PDB: 5B2O) (E), SpCas9 (PDB: 4UN3) (F).

The guide region of sgRNA is omitted for clarity. R:AR, repeat:anti-repeat duplex; SL1, stem loop 1; SL2, stem loop 2; Linker, SL1-SL2 linker.

Structural comparison of CdCas9 with other Cas9s revealed the distinct mechanism of PAM recognition by CdCas9. Among Cas9 orthologs, the PI domains share a conserved core fold consisting of seven β-strands and read the PAM sequence by the diverse residues on the β5–7 strands[24,29,30,69,70] (Fig. 27A–F). SpCas9, SaCas9, and FnCas9 recognize the G-rich sequence through bidentate hydrogen-bonding interactions with arginine residues (Fig. 27B, 27C, and 27F), which show high specificity in the protein-DNA complexes[35]. NmCas9 recognizes the specific PAM through bifurcate and bidentate hydrogen-bonding interactions with histidine and glutamate residues, respectively (Fig. 27D). CjCas9 recognizes the less specific PAM through single hydrogen-bonding interactions (Fig. 27E). These Cas9 orthologs recognize their PAMs via only hydrogen-bonding interactions (Fig. 27B–F). In contrast, CdCas9 recognize its PAM via a combination of single hydrogen-bonding and van der Waals interactions, resulting in the promiscuous PAM recognition (Fig. 27A). The hydrophobic interactions are observed in the thymine nucleotide recognition by protein-DNA complexes[35]. Together, these structural features elucidated the mechanistic diversity in the PAM recognition by Cas9 orthologs.

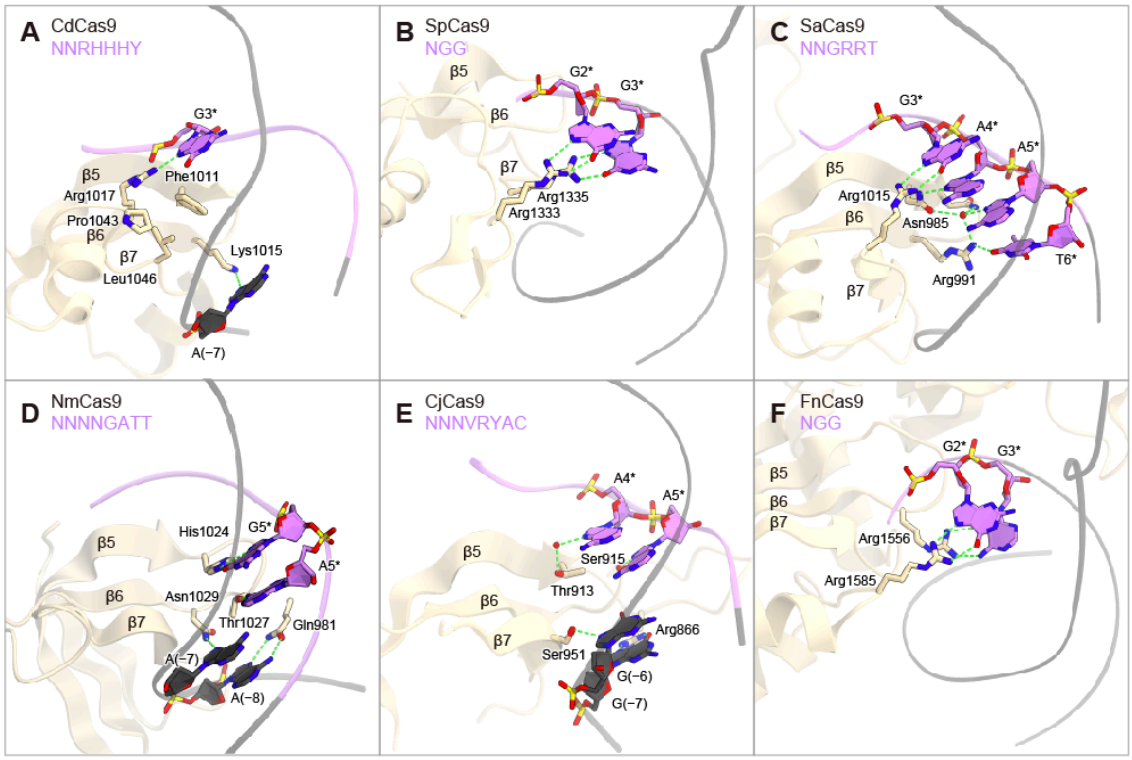

Fig. 27. PAM recognition by the Cas9 orthologs.

(A–F) PAM recognition by CdCas9 (A), SpCas9 (PDB: 4UN3) (B), SaCas9 (PDB: 5CZZ) (C), NmCas9 (PDB: 6JDV) (D), CjCas9 (PDB: 5X2G) (E), and FnCas9 (PDB: 5B2O) (F). The PAMs are colored in purple. Hydrogen bonds are shown as dashed lines. Water molecules are depicted as red spheres.

The crRNA and sgRNA contain spacer and repeat sequences derived from the CRISPR array in the CRISPR locus. Because repeat sequences adjacent to spacer sequences do not contain the PAM in the CRISPR locus, Cas9 can distinguish self (host DNA without the PAM) and non-self (viral DNA with the PAM) sequences in the adaptive immune system[8]. CdCas9 recognizes the NNRHHHY PAM promiscuously while rejecting the three guanines at the forth to sixth PAM positions (Fig. 16D). Intriguingly, the repeat sequence of CdCas9 sgRNA indicates that the protospacer adjacent sequence (5′-ACTGGGG-3′) contains the three guanines (Fig. 19B). This notion suggests that CdCas9 can distinguish self from non-self due to the anti-determinant sequence of three guanines.

Chapter 4: Crystal structure of RNA methyltransferase

4.1 Introduction

N^6,2′-*O*-dimethyladenosine (m^6Am) modification

Nearly half a century ago, N^6,2′-*O*-dimethyladenosine (m^6Am) was identified at the 5′-end of eukaryotic mRNAs[77]. The 5′-cap structure of eukaryotic mRNA consists of the 7-methylguanosine (m^7G), a triphosphate linker, and the 2′-*O*-methylated nucleotide (Nm) (Fig. 5). The nascent transcript is modified by RNGTT/RNMT-mediated m^7G-capping and CMTR1-mediated 2′-*O*-methylation[78,79]. Furthermore, if the first nucleotide of nascent transcript is adenosine, the 5′-capped transcript is methylated at the N^6 position of the first nucleotide[80] (Fig. 5). However, cap-specific adenosine N^6-methyltransferase has not been identified.

Identification of cap-specific adenosine N^6-methyltransferase (CAPAM)

The m^6Am in the 5′-cap of transcript is conserved in the tested vertebrate organisms (human, mouse, and zebrafish), but not in other tested organisms (yeast, nematode, and fly)[56]. As a candidate of cap-specific adenosine N^6-methyltransferase, four research groups examined PCIF1, which is conserved in human, mouse, and zebrafish, but not in yeast and nematode[54–56,81]. PCIF1 consists of an RNAPII-interacting WW domain and an adenine N^6-methyltransferase domain, suggesting that the PCIF could be recruited to RNAPII and mediate the modification of 5′-capped transcripts[82,83]. Consistent with other studies[54–56], Suzuki and colleagues showed that *PCIF1* knockout causes the loss of m^6Am in poly(A)$^+$ RNAs in human cells[57]. Purified PCIF1 protein catalyzed the N^6-methylation of 5′-capped RNA in the presence of *S*-adenosylmethionine[57]. Suzuki and colleagues designated PCIF1 as cap-specific adenosine N^6-methyltransferase (CAPAM).

Physiological roles of m^{6}Am modification and CAPAM

The physiological roles of RNA modification and its enzyme depend on cellular locations, cell types, environmental conditions, and developmental stages[49]. Suzuki and colleagues showed that the growth of *CAPAM* KO cells is sensitive to oxidative stress (Fig. 28). The *ATF4* transcript followed by stress-responsive translation starts with m^{6}Am[55]. To understand the physiological basis, three research groups examined transcriptome and proteome in *CAPAM* KO cells[55,56,81] (Table 6 and Fig. 29). Suzuki and colleagues showed that the level of m^{6}Am-contained transcripts, compared to that of other transcripts, slightly increased upon KO of *CAPAM*[57] (Table 6 and Fig. 28). In the study using other cell types and modified reference data, the level of m^{6}Am-contained transcripts did not significantly change upon KO of *CAPAM*[56] (Table 6 and Fig. 29). In the study classifying transcripts into two groups of upper and lower gene expression, upon KO of *CAPAM*, the level of m^{6}Am-contained transcripts in the lower groups decreased markedly and the level of m^{6}Am-contained transcripts in the upper groups did not significantly decrease[55] (Table 6 and Fig. 29). To evaluate the translation efficiency of transcripts, Suzuki and colleagues performed ribosome profiling. The translation level of m^{6}Am-contained transcripts, compared to that of other transcripts, significantly decreased upon KO of *CAPAM*[57] (Table 6 and Fig. 28). In the study using modified reference data, the translation level of m^{6}Am-contained transcripts, compared to that of other transcripts, slightly increased upon KO of *CAPAM*[55] (Table 6 and Fig. 29). Proteomics analysis showed that the protein expression of almost only m^{6}Am-contained transcripts increased significantly upon KO of *CAPAM*[56] (Table 6 and Fig. 29). Taken together, CAPAM-mediated m^{6}Am modification is not related to the transcript stability and affects the translation in both positive and negative manners. Cell phenotypes of stress-response may be related to gene regulation through CAPAM-mediated m^{6}Am modification.

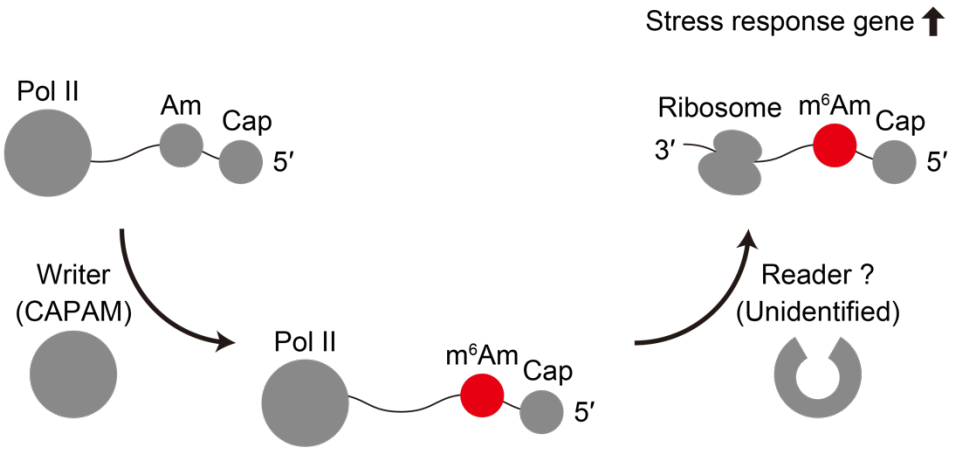

Fig. 28. Biogenesis and function of m^{6}Am modification, proposed by Suzuki and colleagues.

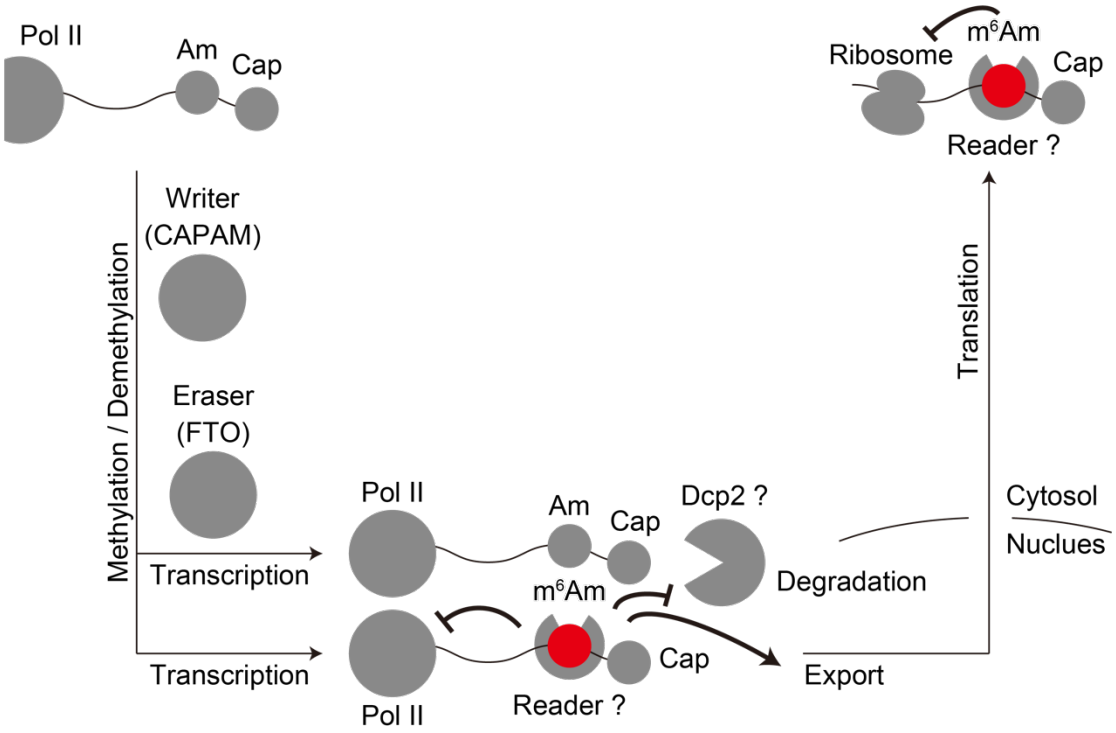

Fig. 29. Current opinion of putative physiological role of m^{6}Am modification.

Table 6. Transcriptomic and proteomic changes upon knockout of *CAPAM*.

Studies	Cell-type	Reference data (m^6Am-mapped data)	Transcriptomic change	Proteomic change
Akichika et al. (2019)	HEK293T	miCLIP data Linder et al. (2015)	Relative abundance of m^6Am-transcripts slightly increases.	Relative abundance of m^6Am-transcripts binding to Ribosome significantly decreases. (Ribosome-profiling)
Sendinc et al. (2019)	MEL624	m^6Am-Exo-seq data Sendinc et al. (2019)	Relative abundance of m^6Am-transcripts does not significantly change.	Relative abundance of m^6Am-transcripts binding to Ribosome slightly increases. (Ribosome-profiling)
Boulias et al. (2019)	HEK293T	Revised miCLIP data, using *CAPAM* KO cells Boulias et al. (2019)	Upper half (strongly expressed transcripts) Relative abundance of m^6Am-transcripts does not significantly change. Lower half (weakly expressed transcripts) Relative abundance of m^6Am-transcripts markedly decrease.	Relative abundance of protein translated from m^6Am-transcripts significantly increases. (Proteomics)

4.2 Research aims

The mechanism of CAPAM-mediated m^6Am modification is important for the physiological understanding of m^6A and m^6Am modifications. Crystal structures of METTL3-METTL14 complex and METTL16 revealed the mechanism of internal m^6A modification[43,84]. However, the mechanism of terminal m^6Am modification by CAPAM is unknown. To elucidate the mechanism of m^6Am modification by CAPAM, I determined the crystal structures of CAPAM.

4.3 Methods

4.3.1 Sample preparation

The full-length human CAPAM (residues 1–704), the core region of human CAPAM (residues 174–668), and the core region of zebrafish CAPAM (residues 178–673) were cloned into the pE-SUMO vector (LifeSensors). The His_6-SUMO-tagged CAPAM was expressed at 20 °C overnight in *Escherichia coli* Rosetta 2 (DE3) (Novagen). The bacterial lysate containing the recombinant CAPAM was batched with Ni-NTA Superflow resin (Qiagen) in buffer A (50 mM Tris-HCl, pH 8.0, 20 mM imidazole, and 1 M NaCl) and eluted with buffer B (50 mM Tris-HCl, pH 8.0, 300 mM imidazole, and 0.3 M NaCl). The His_6-SUMO-tagged CAPAM was mixed with TEV protease and dialyzed at 4 °C overnight in buffer C (20 mM Tris-HCl, pH 8.0, 40 mM imidazole, and 0.5 M NaCl). After removing the His_6-SUMO-tag by a Ni-NTA column, CAPAM was purified through a HiTrapHeparin HP column (GE Healthcare), using buffer D (20 mM Tris-HCl, pH 8.0) and buffer E (20 mM Tris-HCl, pH 8.0 and 2 M NaCl). The CAPAM was purified through a Superdex 200 Increase column (GE Healthcare), using buffer H (10 mM Tris-HCl, pH 8.0 and 150 mM NaCl, 1 mM dithiothreitol).

4.3.2 *in vitro* methyltransferase assay

Suzuki and colleagues performed this assay, using the purified CAPAM that I prepared. The m^7GpppAm-capped 110-nt RNA (3 μM) was incubated at 37 °C for 15 min with the CAPAM protein (0.1 μM) in 10 μL of reaction buffer containing 20 mM HEPES-NaOH, pH 7.5, 1 mM dithiothreitol, 36.5 μM *S*-[methyl-^{14}C]-adenosyl methionine (PerkinElmer). Aliquots were spotted on Whatman 3MM filter paper and the filter papers were immediately soaked in 5% trichloroacetic acid. Radioactivity remaining on the filter papers was measured, using the Tri-Carb 2910TRl liquid scintillation counter (PerkinElmer). The m^6A(m) formation rate was measured as the molar ratio of the incorporated methyl group calculated from the ^{14}C radioactivity to the substrate RNA.

4.3.3 Crystallography

Purified human CAPAM (residues 174–668) was crystallized at 20 °C using the sitting-drop vapor diffusion method. Crystals were obtained by mixing 0.2 μl of protein solution (6 mg/ml) and 0.2 μl of reservoir solution containing 0.1 M bis-Tris propane, pH 8.0, 9–12% PEG3,350, and 0.2–0.4 M sodium nitrate. Crystals in the *S*-adenosyl homocysteine (SAH)-bound form were obtained by the soaking method. After growing to full size, the crystals were soaked in solution containing 0.1 M bis-Tris propane, pH 8.0, 12% PEG3,350, 0.4 M sodium nitrate, 30% glycerol, 2 mM SAH (Sigma-Aldrich), and 2 mM m7GpppA (NEB), and then incubated at 20°C for 3 min. The human CAPAM crystals in the apo form were cryoprotected in solution containing 0.1 M bis-Tris propane, pH 8.0, 12% PEG3,350, 0.4 M sodium nitrate, and 30% glycerol.

Purified zebrafish CAPAM (residues 178–673) was crystallized at 4 °C using the hanging-drop vapor

diffusion method. Crystals were obtained by mixing 1 μl of protein solution (7 mg/ml) and 1 μl of reservoir solution containing 0.1 M bis-Tris propane, pH 7.5, 11–15% PEG3,350, and 0.1 M KSCN. The crystals were then improved by micro-seeding using Seed Bead (Hampton Research). SeMet-labeled zebrafish CAPAM was crystallized under similar conditions, using native zebrafish CAPAM crystals as seed stocks. The crystals of zebrafish CAPAM in the SAH-bound, m7GpppA/SAH-bound, and m7GpppAmG/SAH-bound forms were obtained by the soaking method. After growing to full size, the crystals were soaked in solution containing 0.1 M bis-Tris propane, pH 7.5, 15% PEG3,350, 0.1 M KSCN, 25% ethylene glycol, and the ligands including 2 mM SAH (the SAH-bound form), 2 mM m7GpppA/2 mM SAH (the m7GpppA/SAH-bound form), and 10 mM m7GpppAmG (Trilink)/2 mM SAH (the m7GpppAmG/SAH-bound form). The crystals were then incubated at 20 °C for 3 h. The SeMet-labeled zebrafish CAPAM crystals in the apo form were cryoprotected in solution containing 0.1 M bis-Tris propane, pH 7.5, 15% PEG3,350, 0.1 M KSCN, and 25% glycerol. The X-ray diffraction data were collected at 100 K on beamlines BL41XU at SPring-8 and PXI at Swiss Light Source. The X-ray diffraction data were processed using XDS[58] and AIMLESS[59]. The structure of zebrafish CAPAM in the apo form was determined by the Se-SAD method, using SHELXD[85], autoSHARP[86], and PHENIX AutoBuild[87]. The structures of human CAPAM in the apo and SAH-bound forms were determined by the molecular replacement method with MOLREP[60], using the structure of zebrafish CAPAM in the apo form as a search model. The structures of zebrafish CAPAM in the ligand-bound forms were detemined by the molecular replacement method with MOLREP, using the structure of zebrafish CAPAM in the apo form as a search model. Model building and refinement were performed using COOT[61] and PHENIX[62], respectively. Structural figures were prepared using CueMol (http://www.cuemol.org).

4.4 Results

4.4.1 Crystallization and structural determination of CAPAM

The RNAPII-interacting WW domain was connected to the core region of CAPAM by the 93 amino-acids flexible region (residues 81–173), which is predicted to hamper the crystallization (Fig. 30A). I truncated the N-terminal region (residues 1–173) of human CAPAM and designated the protein as the ΔWW mutant. Suzuki and colleagues performed the *in vitro* methyltransferase. In the *in vitro* methyltransferase experiment, the full-length CAPAM and the ΔWW mutant showed the comparable activities of $m^6A(m)$ formation, while the catalytically dead N553A mutant showed the reduced activity of $m^6A(m)$ formation, indicating that the WW domain does not have effect on the $m^6A(m)$ formation (Fig. 30B). For crystallization, I prepared the core regions of human CAPAM (residues 174–668) and zebrafish CAPAM (residues 178–673) (Fig. 30C). The 10–20% acrylamide SDS-PAGE analysis showed that the CAPAM proteins were produced in high yield and purity (Fig. 31). I obtained about 200-μm length crystals of the human and zebrafish CAPAMs (Fig. 32A). The analysis of X-ray diffraction experiments showed that the collected data of the CAPAMs were at 1.8–2.9 Å resolutions (Fig. 32B and Table 7). The refinement and validation software, PHENIX, showed that the R_{work} / R_{free} values of the final structure models of human CAPAM in the apo and SAH-bound forms and zebrafish CAPAM in the apo, SAH-bound, m7GpppA/SAH-bound, and m7GpppAmG/SAH-bound forms are 0.214 / 0.252, 0.208 / 0.236, 0.188 / 0.214, 0.179 / 0.203, 0.181 / 0.210, and 0.177 / 0.215, respectively (Table 7). Overall structures of the six CAPAM structures are similar (Fig. 33A), and thus I describe the structure of zebrafish CAPAM in the m7GpppA/SAH-bound forms unless otherwise stated.

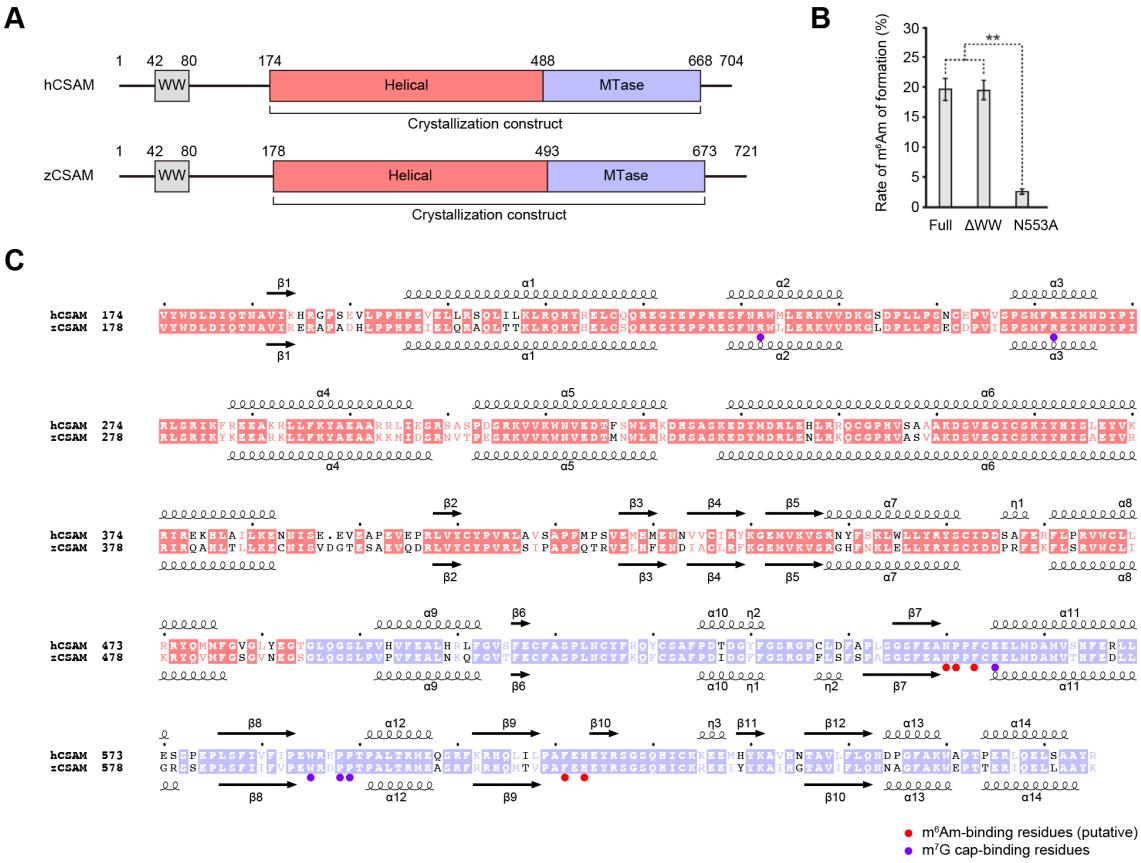

Fig. 30. Crystallization constructs of CAPAM.

(A) Domain structures of human CAPAM (hCAPAM) and zebrafish CAPAM (zCAPAM).

(B) *in vitro* methyltransferase activities of the CAPAM proteins. Full, the full-length hCAPAM; ΔWW, the WW domain-truncated hCAPAM; N553A, the N553A hCAPAM mutant. Suzuki and colleagues performed this assay.

(C) Sequence alignment of hCAPAM and zCAPAM. Key residues are indicated by circles.

The figure was prepared using Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/) and ESPript (http://espript.ibcp.fr/ESPript/ESPript/).

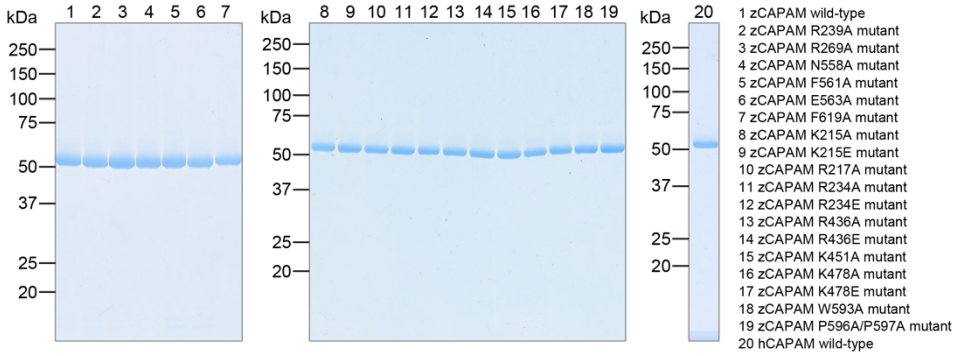

Fig. 31. SDS-PAGE analysis of the CdCas9 proteins for the biochemical and structure analyses.

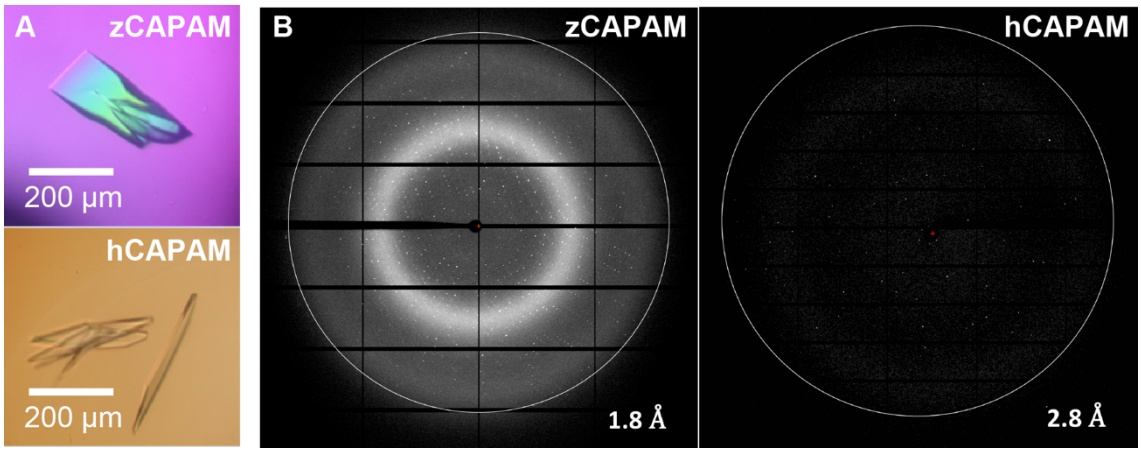

Fig. 32. X-ray crystallography of CAPAM.

(A) The crystals of human CAPAM (hCAPAM) and zebrafish CAPAM (zCAPAM).

(B) Diffraction images obtained from the CAPAM crystals.

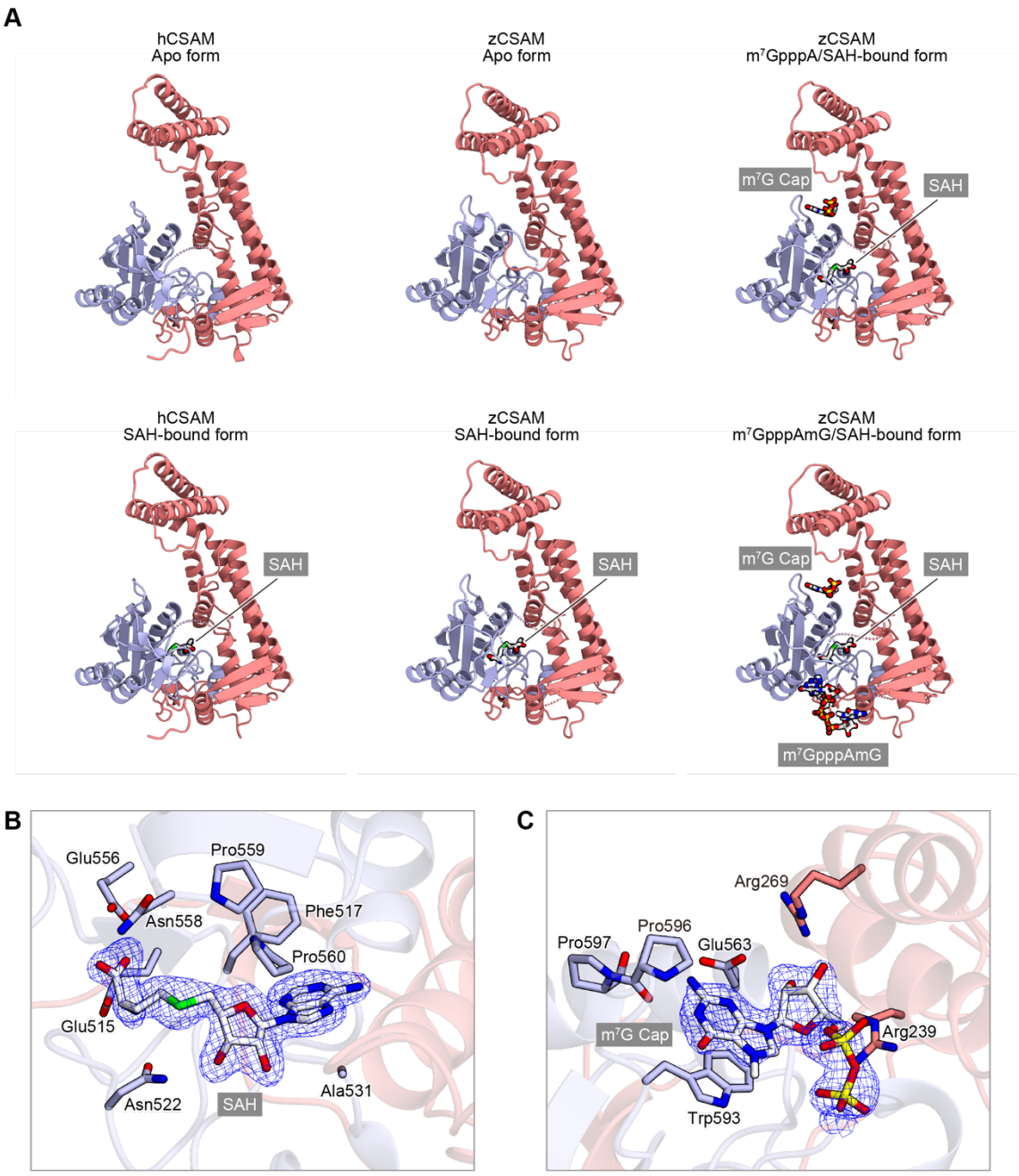

Fig. 33. Crystal structures of CAPAM.

(A) Overall structures of hCAPAM (apo and SAH-bound forms) and zCAPAM (apo, SAH-bound, m^7GpppA/SAH-bound, and m^7GpppAmG/SAH-bound forms).

(B and C) The $2mF_o - DF_c$ electron density maps of SAH and the m^7G cap (contoured at 1.2σ) in the m^7GpppA/SAH-bound form.

Table 7. Data collection and refinement statistics.

	hCSAM apo	**hCSAM** SAH	**zCSAM** apo (SeMet)	**zCSAM** SAH	**zCSAM** SAH/m^{7}GpppA	**zCSAM** SAH/m^{7}GpppAmG
Data collection						
Beamline	SPring-8 BL41XU	SPring-8 BL41XU	SLS PXIII	SPring-8 BL41XU	SLS PXIII	SPring-8 BL41XU
Wavelength (Å)	1.0000	1.0000	0.9790	1.0000	1.0000	1.0000
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$
Cell dimensions						
a, b, c (Å)	70.7, 120.3, 156.8	70.2, 120.4, 157.0	71.5, 84.1, 93.1	71.4, 82.3, 92.8	71.7, 83.2, 93.1	74.2, 84.0, 93.0
α, β, γ (°)	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 90
Resolution (Å)*	48.12–2.70 (2.82–2.70)	47.99–2.90 (3.06–2.90)	47.03–2.00 (2.05–2.00)	46.62–1.80 (1.84–1.80)	46.90–2.00 (2.05–2.00)	47.74–1.80 (1.84–1.80)
R_{merge}	0.110 (0.751)	0.117 (0.682)	0.126 (1.028)	0.040 (0.587)	0.063 (0.654)	0.050 (0.576)
R_{pim}	0.046 (0.316)	0.049 (0.282)	0.054 (0.429)	0.017 (0.241)	0.026 (0.265)	0.021 (0.234)
$I/\sigma I$	10.6 (2.4)	10.6 (2.6)	10.1 (2.1)	21.2 (2.7)	16.1 (2.9)	19.0 (3.3)
Completeness (%)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	99.1 (98.3)
Multiplicity	6.7 (6.6)	6.7 (6.8)	12.4 (12.9)	6.7 (6.8)	6.8 (7.1)	6.8 (6.9)
CC(1/2)	0.998 (0.900)	0.998 (0.912)	0.996 (0.913)	0.999 (0.938)	0.998 (0.917)	0.999 (0.911)
Refinement						
Resolution (Å)	48.13–2.70 (2.80–2.70)	47.99–2.90 (3.00–2.90)	47.03–2.00 (2.07–2.00)	46.62–1.80 (1.86–1.80)	46.90–2.00 (2.07–2.00)	47.74–1.80 (1.86–1.80)
No. reflections	37,449 (3,692)	30,174 (2,965)	38,621 (3,794)	51,336 (5,044)	38,062 (3,736)	53,942 (5,272)
R_{work} / R_{free}	0.217/0.256 (0.325/0.373)	0.218/0.249 (0.304/0.295)	0.193/0.217 (0.230/0.255)	0.177/0.204 (0.271/0.308)	0.190/0.213 (0.267/0.325)	0.177/0.214 (0.222/0.266)
No. atoms						
Protein	8,002	7,978	3,930	3807	3,808	3,808
Ligand	0	52	0	34	71	165
Solvent	64	19	199	244	163	314
B-factors (Å^2)						
Protein	61.4	65.0	44.5	47.3	51.3	39.7
Ligand	—	39.1	—	34.4	55.4	56.2
Solvent	44.8	36.9	43.5	47.3	47.0	42.6
R.m.s. deviations						
Bond lengths (Å)	0.002	0.002	0.003	0.006	0.002	0.006
Bond angles (°)	0.51	0.47	0.56	0.80	0.55	0.82
Ramachandran plot (%)						
Favored region	97.24	97.13	97.28	98.70	98.91	99.35
Allowed region	2.76	2.87	2.72	1.30	1.09	0.65
Outlier region	0.00	0.00	0.00	0.00	0.00	0.00
MolProbity score						
Clashscore	5.39	3.80	1.67	1.71	2.60	1.53
Rotamer outlier	1.27	1.16	0.23	0.48	0.48	0.24

*Values in parentheses are for the highest resolution shell.

4.4.2 Crystal structures of CAPAM

The core region of CAPAM consists of the Helical and methyltransferase (MTase) domains (Fig. 34A and 34B). The Helical domain consists of three-helix bundles (α1-α6-α8 and α4-α5-α6), four-helix bundle (α1-α2-α3-α6), and β-sheets (β1-β2 and β3-β4-β5) (Fig. 34B). A Dali search detected no structural similarity between the Helical domain and any of the known protein structures. The MTase domain adopts a canonical Rossman fold containing a conserved catalytic motif (residues 558–561) (Fig. 34B). SAH is bound to a catalytic pocket of CAPAM (Fig. 34C). The m^7G cap is bound in the "m^7G site" located between the Helical and MTase domains (Fig. 34D). The ribose and guanine moieties of m^7G are recognized by Arg239/Arg269/Glu563 and Glu563/Trp593/Pro595/Pro596, respectively (Fig. 34D). Suzuki and colleagues performed the *in vitro* methyltransferase assay, using the CAPAM proteins that I prepared. The mutational assay confirmed the importance of these residues for the m^7G cap recognition (Fig. 34F). The m^7G cap, but not the target adenosine adjacent to the m^7G cap, was visible in the electron density map (Fig. 34E), suggesting that the target nucleotide is disordered in the crystal structure. Based on the reported structure of M.TaqI methyltransferase bound to a DNA substrate[88], we modeled a 2′-*O*-methyladenosine (Am) at the active site of CAPAM (Fig. 35A–C). The model suggested that the adenine moiety of Am forms hydrogen-bonding interactions with Asn558/Pro559/Phe561 and π-stacking interactions with Phe561/Phe619, and the ribose moiety of Am forms van der Waals interactions with His621 (Fig. 35B). Suzuki and colleagues performed the mutational assay, using the purified CAPAM proteins that I prepared. The mutational assay indicated that these residues are important for target nucleotide recognition (Fig. 34F). Notably, the Helical domain forms a positively charged groove in the vicinity of the active site, suggesting that the RNA strand following the 5′-cap binds to this positive groove (Fig. 36A and 36B). Suzuki and colleagues

performed the mutational assay, using the purified CAPAM proteins that I prepared. The mutational assay revealed the functional significance of the positively charged groove (Fig. 36C). In the m^7GpppAmG/SAH-bound form, in addition to the m^7G site, the m^7GpppAmG bound between three CAPAM molecules in the crystallographic asymmetric unit (Fig. 33, lower right), suggesting the single-strand RNA substrate may bind to the positively charged groove. Overall, the structural and mutational data provided mechanistic insights into m^7G-capped RNA recognition and m^6Am modification by CAPAM.

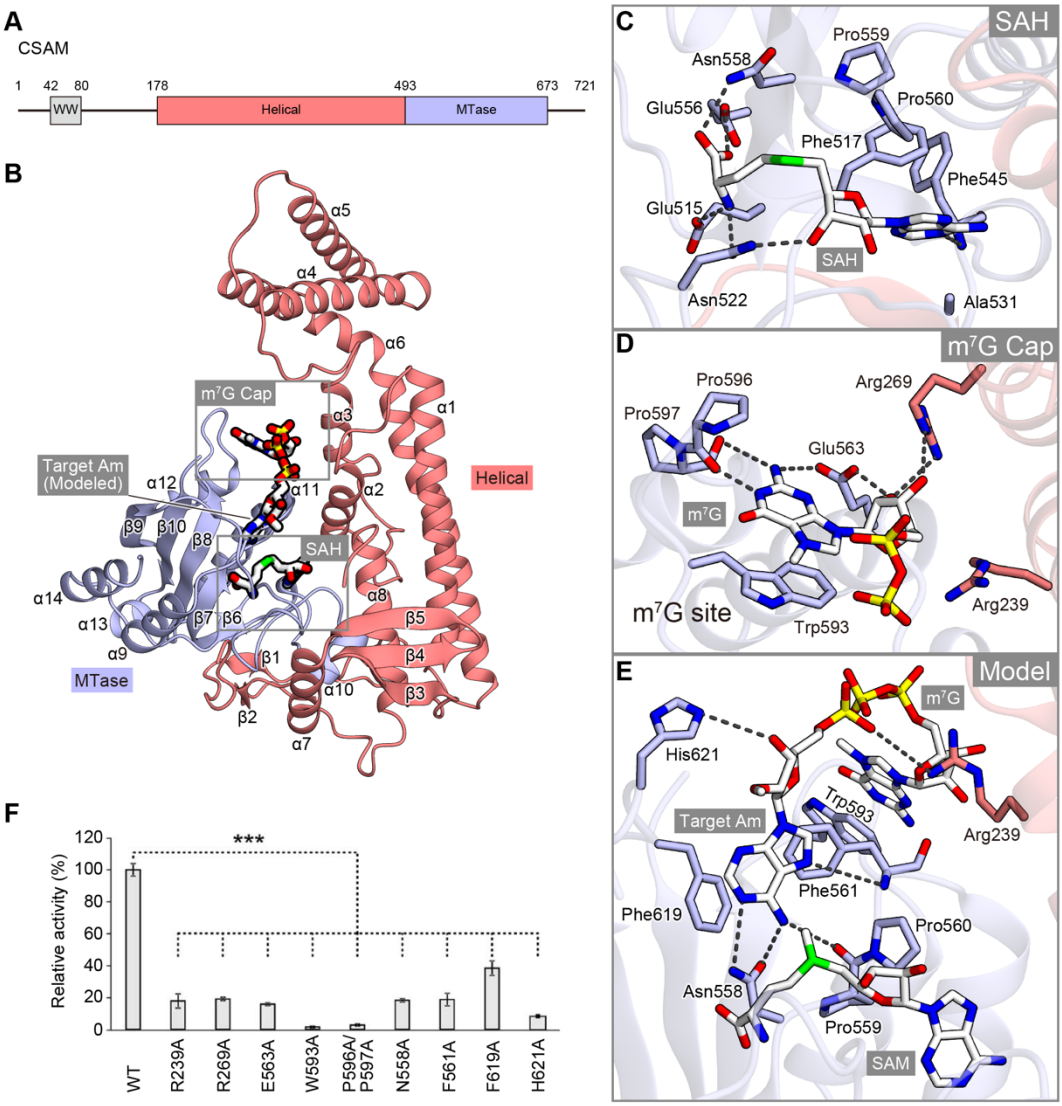

Fig. 34. Structure of CAPAM in complex with SAH and m^{7}G Cap.

(A) Domain structure of zCAPAM.

(B) Overall structure of zCAPAM in complex with SAH and m^{7}G Cap.

(C and D) Recognition of SAH and the m^{7}G Cap. Hydrogen bonds are shown as dashed lines.

(E) Putative binding site of the target Am nucleotide.

(F) *in vitro* methyltransferase assay for examining effects of the mutations on the 5′-cap-interacting residues. The m^{7}GpppAm-capped RNA was incubated with the zCAPAM proteins at 37 °C for 15 min. Error bars represent s.d. from n = 4 replicates. $^{**}P < 1.0\times10^{-6}$ by Student's t-test. Suzuki and colleagues performed this assay, using the purified CAPAM proteins that I prepared.

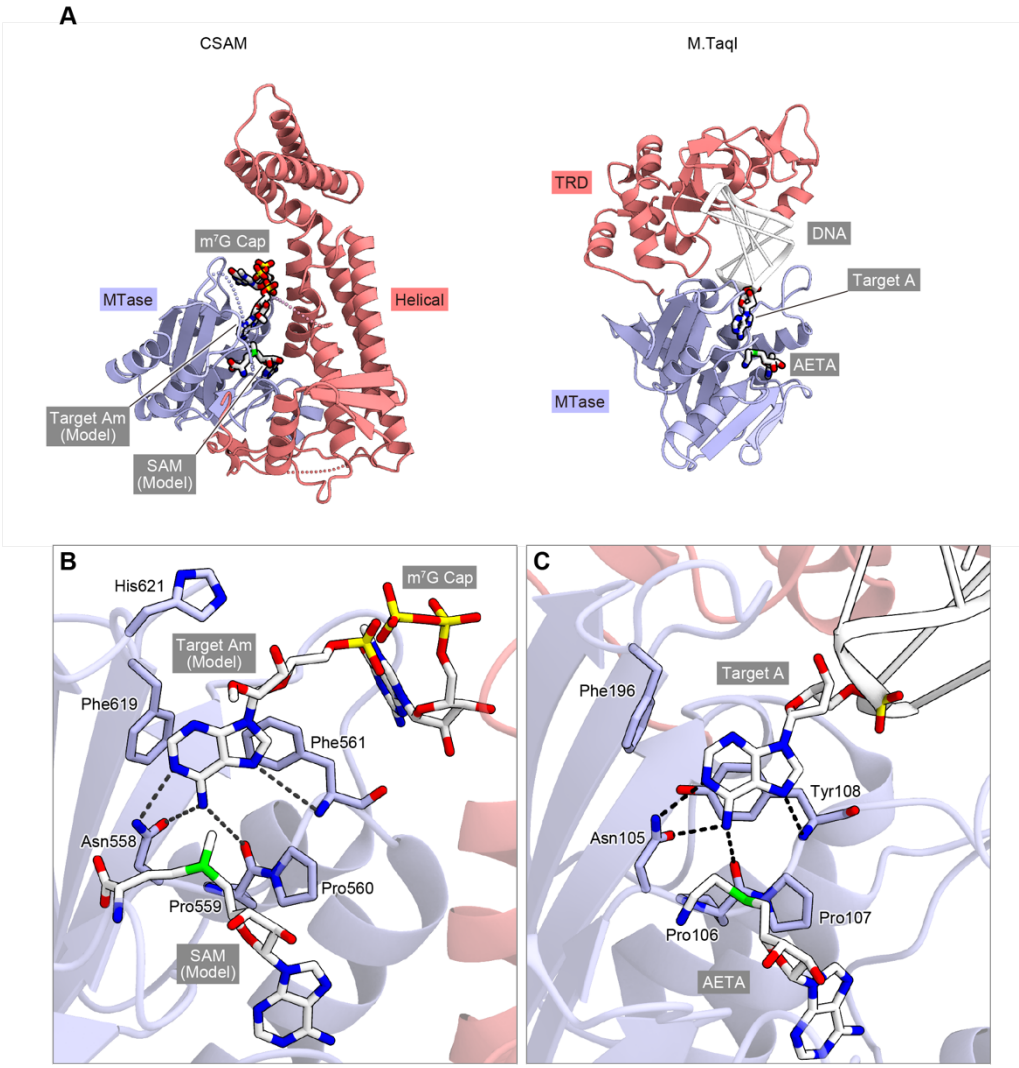

Fig. 35. Structural comparison with DNA m^6A methyltransferase

(A) Structural comparison of zCAPAM with M.TaqI (PDB: 1G38). In the zCAPAM-m^7GpppA/SAH complex structure, *S*-adenosyl methionine (SAM) and 2′-*O*-methyladenosine (Am) are modeled as the methyl donor and acceptor, respectively. In the structure of the M.TaqI in complex with its target DNA and a cofactor analog, 5′-[2-(amino)ethylthio]-5′-deoxyadenosine (AETA), the target DNA is bound between the target recognition domain (TRD) and the MTase domain. The target A nucleotide is flipped out and docked into the active site.

(B and C) Active sites of zCAPAM (B) and M.TaqI (C). Hydrogen bonds are shown as dashed lines.

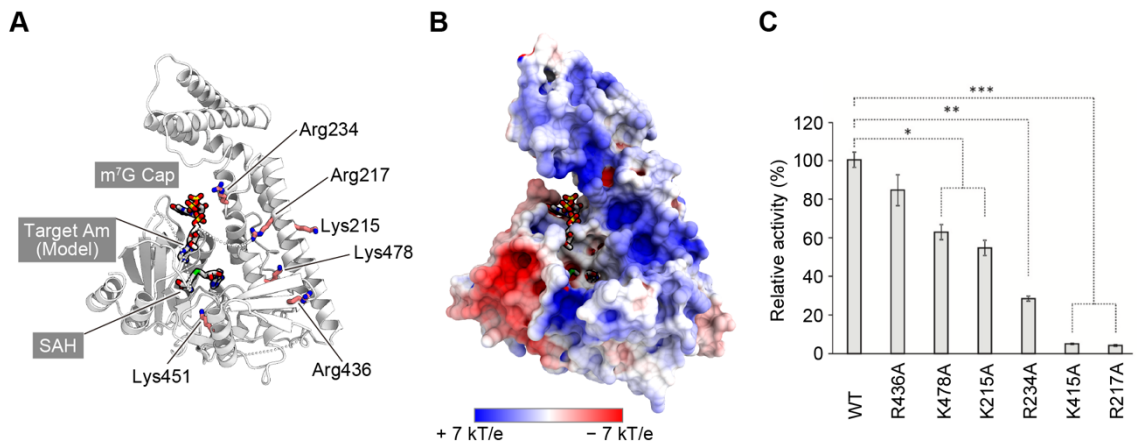

Fig. 36. Putative RNA-binding groove in CAPAM.

(A) Ribbon representation of zCAPAM in the m^{7}GpppA/SAH-bound form. 2′-*O*-methylated A nucleotide (Am) is modeled as a methyl acceptor. Positively charged residues on the protein surface are mapped on the structure.

(B) Electrostatic surface potential of zCAPAM in the m^{7}GpppA/SAH-bound form.

(C) *in vitro* methyltransferase assay for examining effects of the mutations on the putative RNA-binding residues. The m^{7}GpppAm-capped RNA was incubated with the zCAPAM proteins at 37 °C for 15 min. Error bars represent s.d. from $n = 4$ replicates. $^{**}P < 1.0\times10^{-6}$ by Student's *t*-test. Suzuki and colleagues performed this assay, using the CAPAM proteins that I prepared.

4.5 Discussion

The present structure of zebrafish CAPAM in complex with the m^7G cap and SAH provided mechanistic insights into the m^6Am modification in the 5′ cap. Recent structures of the other m^6A writers (METTL3-METTL14 complex and METTL16) suggested the diverse mechanisms of RNA substrate recognition and m^6A modification[43,84]. Structural comparison of CAPAM with these m^6A writers revealed that m^6A writers share a conserved MTase domain with a Rossman fold and have an additional diverse domain/subunit (Fig. 37A and 37B). In CAPAM, the Helical domain forms a positively charged groove which could bind the 5′-capped single-strand RNA (Fig. 37B). In METTL3-METTL14 complex, the dimer interface forms a positively charged groove which could bind the single-strand RNA containing the consensus motif (Fig. 37B). In METTL16, the N-terminal region forms a positively charged wide groove which binds the structured RNA (Fig. 37B). These findings show that the distinct structures additional diverse domains/subunits define the target specificities of m^6A writers.

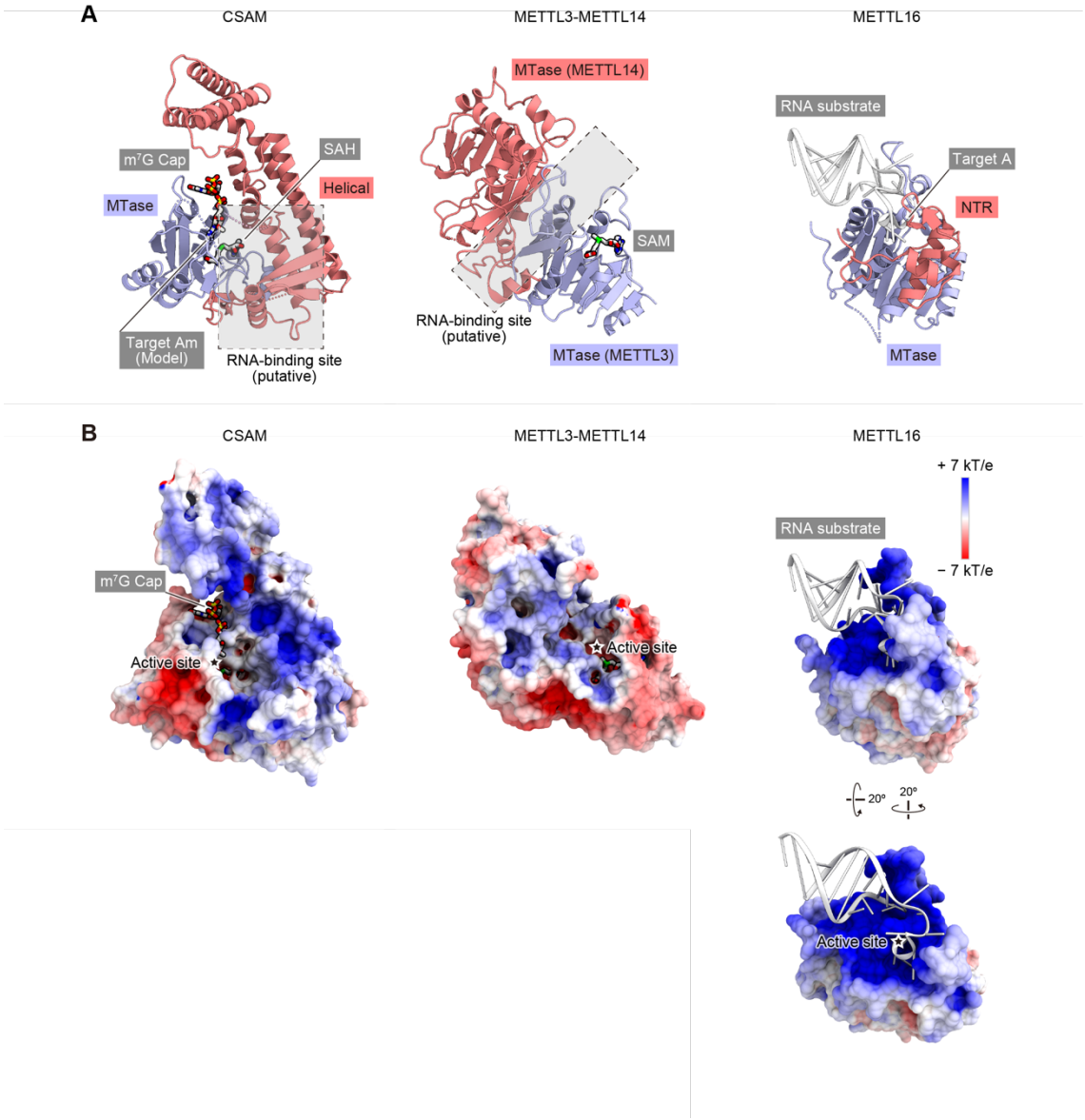

Fig. 37. Structural comparison of m^6A writers.

(A) Structures of CAPAM (m^7GpppA/SAH-bound form), the METTL3-METTL14 complex (PDB: 5IL1), and METTL16-RNA complex (PDB: 6DU4). The Am nucleotide is modeled at the active site of CAPAM. The MTase domains of the m^6A writers are structurally aligned. NTR, N-terminal region. (B) Electrostatic surface potential of CAPAM, the METTL3-METTL14 complex (PDB: 5IL1), and the METTL16-RNA complex (PDB: 6DU4).

The findings about the 5′-cap m^6A modification by CAPAM advances the epitranscriptomics field because the structural information revealed that, in m^6A writers, the 5′-cap m^6A modification by CAPAM is distinct from the internal m^6A modification by the METTL3-METTL14 complex. In contrast to m^6A writers, due to the lack of structural information, it remains to be clear whether a major m^6A eraser, FTO, demethylates the 5′-cap m^6A modification or the internal m^6A modification. The biochemical study showed that FTO prefers the 5′-cap m^6A nucleotide to the internal m^6A nucleotide[89]. The cellular study showed that FTO prefers the internal m^6A nucleotide to the 5′-cap m^6A nucleotide in the nucleus[90]. The discrepancy in the FTO preference of m^6A nucleotide will be clarified by the structural information of the FTO-5′-capped RNA complex. In the CAPAM structure in complex with the 5′-cap, the 5′-cap is accommodated in the pocket of the MTase and Helical domains (Fig. 37A and 37B). In contrast, the internal m^6A nucleotide are predicted to be bound to the surface of the METTL3-METTL14 complex[43] (Fig. 37A and 37B). In other words, the 5′-cap m^6A nucleotide is modified in the cave of CAPAM, while the internal m^6A nucleotide is modified on the bridge of the METTL3-METTL14 complex. The structural features of CAPAM shows that the substrate of CAPAM is only the 5′-cap m^6A nucleotide. Taken together, the findings about the specificity of CAPAM provided the structural basis towards the understanding of the complicated m^6A modification pathways.

Chapter 5: General Discussion

From the viewpoint that the origin of life is gene, genetic study shows the relationship between genotype and phenotype. The invention of the conversion tool (genome editing) produced a paradigm shift in genetics. The sequencing and mapping technological advancement greatly expands the concept of genetics to genomics, transcriptomics, epigenetics, and epitranscriptomics. In the advanced genetics field, there are two important proteins, Cas9 and CAPAM. This study provide the insights into the action mechanisms of RNA-guided DNA endonuclease Cas9 and RNA methyltransferase CAPAM. I discuss two topics, programmability and modularity, in targeting genome and transcriptome by these proteins.

In an application of genome editing, the RNA-guided DNA endonuclease Cas9 has a programmability which is divided into three topics, efficiency, precision, and targetability. The efficiency and precision are related to which Cas9 orthologs/mutants are used and how long guide sequences are used in the experiments. CdCas9 shows the highest activity with 22-nt guide sequence length (Fig. 16A), while SpCas9, SaCas9, and CjCas9 requires the 20, 21, 22-nt guide sequence length, respectively[22,25]. The truncation of guide sequence in SpCas9 sgRNA reduces the cleavage activity towards not only the on-target sites but also the off-target sites[91]. The base-paired guide RNA and target DNA heteroduplex is sensed by the REC2 domain, in which the interacting residues differ among Cas9 orthologs[92]. To understand the relationship between efficiency/precision and guide sequence length, further structural and functional studies are needed, such as the catalytic states of Cas9 orthologs.

The targetability is related to PAM. To address the PAM limitation in genome editing, protein engineering and natural diversity are used for the development of the tools. Protein engineering including site-directed mutagenesis and random mutagenesis can introduce the small change of PAM recognition into Cas9. This strategy can be applied to the well-characterized Cas9s, such as SpCas9, SaCas9, and FnCas9. The optimal cell-types and expression methods of these Cas9s have been already characterized, thus further characterization is minimal. The natural evolution including both the random mutagenesis and the homologous recombination can introduce the big change of PAM recognition into Cas9. The strategy utilizing natural diversity can explore novel Cas9, which requires the labor of characterization for genome editing application. Both the protein engineering and natural diversity strategies have advantages and disadvantages in the development of genome editing toolbox. This study reveals the structural basis of proteins derived from these strategies and shows the validity of both strategy for further expanding the targetable sites. Both strategies may develop engineered or discovered Cas9s recognize the C-rich or T-rich PAM.

In many cases, one protein has one function, thus, fusion protein can be engineered with multiple functions. For example, the fusion protein of the catalytically dead Cas9 (dCas9, binding-module) and histone acetyltransferase p300 (effector-module) introduces the H3K27 acetylation in the vicinity of the dCas9-binding site[93]. Protein can be regarded as a module which is connectable with other modules, while domain cannot be regarded as a module. In this study, the PI domain of CdCas9 forms extensive interactions with the RuvC and WED domains (Fig. 19C), indicating that the PI domain swapping with other Cas9 orthologs can abolish the solubility of chimeric Cas9s. The MTase domain of CAPAM forms extensive interactions with the different additional domain from other m^6A writers (Fig. 37A).

Crystal structures of CdCas9 and CAPAM indicate that the domain swapping methods have difficulties in altering the PAM specificity and RNA substrate specificity. Using protein module, but not domain module, the fusion protein of dCas9-METTL14-METTL3 in complex with an sgRNA and a short DNA duplex containing PAM introduces internal m^6A modification into the dCas9-binding site of endogenous mRNAs[94]. The fusion protein of dCas9-CAPAM can introduce 5′-terminal m^6A modification into dCas9-binding endogenous mRNAs. This system will elucidate the functional role of 5′-terminal m^6A modification in specific transcripts and in specific conditions.

References

1. Faure, G. *et al.* CRISPR–Cas in mobile genetic elements: counter-defence and beyond. *Nat. Rev. Microbiol.* **17**, 513–525 (2019).
2. Wang, J. *et al.* Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems. *Cell* **163**, 840–853 (2015).
3. Xiao, Y., Ng, S., Hyun Nam, K. & Ke, A. How type II CRISPR-Cas establish immunity through Cas1-Cas2-mediated spacer integration. *Nature* **550**, 137–141 (2017).
4. Wright, A. V. *et al.* Structures of the CRISPR genome integration complex. *Science (80-.).* eaao0679 (2017). doi:10.1126/science.aao0679
5. Mohanraju, P. *et al.* Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. *Science (80-.).* **353**, aad5147 (2016).
6. Deltcheva, E. *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602–607 (2011).
7. Sapranauskas, R. *et al.* The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. *Nucleic Acids Res.* **39**, 9275–9282 (2011).
8. Sternberg, S. H., Richter, H., Charpentier, E. & Qimron, U. Adaptation in CRISPR-Cas Systems. *Mol. Cell* **61**, 797–808 (2016).
9. Hsu, P. D., Lander, E. S. & Zhang, F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157**, 1262–1278 (2014).
10. Chylinski, K., Makarova, K. S., Charpentier, E. & Koonin, E. V. Classification and evolution of type II CRISPR-Cas systems. *Nucleic Acids Res.* **42**, 6091–6105 (2014).
11. Fonfara, I. *et al.* Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. *Nucleic Acids Res.* **42**, 2577–2590 (2014).
12. Esvelt, K. M. *et al.* Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nat. Methods* **10**, 1116–1123 (2013).
13. Jinek, M. *et al.* A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science (80-.).* **337**, 816–821 (2012).
14. Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl. Acad. Sci.* **109**, E2579–E2586 (2012).
15. Deveau, H. *et al.* Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *J. Bacteriol.* **190**, 1390–1400 (2008).
16. Garneau, J. E. *et al.* The CRISPR/cas bacterial immune system cleaves bacteriophage and

plasmid DNA. *Nature* **468**, 67–71 (2010).

17. Doudna, J. A. & Charpentier, E. The new frontier of genome engineering with CRISPR-Cas9. *Science (80-.).* **346**, 1258096–1258096 (2014).
18. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science (80-.).* (2013). doi:10.1126/science.1231143
19. Mali, P. *et al.* RNA-Guided Human Genome Engineering via Cas9. *Science (80-.).* **339**, 823–826 (2013).
20. Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733–740 (2009).
21. Hou, Z. *et al.* Efficient genome engineering in human pluripotent stem cells using Cas9 from Neisseria meningitidis. *Proc. Natl. Acad. Sci. USA* **110**, 15644–15649 (2013).
22. Ran, F. A. *et al.* In vivo genome editing using Staphylococcus aureus Cas9. *Nature* **520**, 186–191 (2015).
23. Müller, M. *et al.* Streptococcus thermophilus CRISPR-Cas9 systems enable specific editing of the human genome. *Mol. Ther.* **24**, 636–644 (2016).
24. Hirano, H. *et al.* Structure and Engineering of Francisella novicida Cas9. *Cell* **164**, 950–961 (2016).
25. Kim, E. *et al.* In vivo genome editing with a small Cas9 orthologue derived from Campylobacter jejuni. *Nat. Commun.* **8**, 1–12 (2017).
26. Harrington, L. B. *et al.* A thermostable Cas9 with increased lifetime in human plasma. *Nat. Commun.* **8**, 1–8 (2017).
27. Jinek, M. *et al.* Structures of Cas9 Endonucleases Reveal RNA-Mediated Conformational Activation. *Science (80-.).* **343**, 1247997–1247997 (2014).
28. Jiang, F., Zhou, K., Ma, L., Gressel, S. & Doudna, J. A. A Cas9-guide RNA complex preorganized for target DNA recognition. *Science (80-.).* (2015). doi:10.1126/science.aab1452
29. Nishimasu, H. *et al.* Crystal Structure of Staphylococcus aureus Cas9. *Cell* **162**, 1113–1126 (2015).
30. Anders, C., Niewoehner, O., Duerst, A. & Jinek, M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **513**, 569–573 (2014).
31. Jiang, F. *et al.* Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science (80-.).* **351**, 867–871 (2016).
32. Zhu, X. *et al.* Cryo-EM structures reveal coordinated domain motions that govern DNA cleavage

by Cas9. *Nat. Struct. Mol. Biol.* (2019). doi:10.1038/s41594-019-0258-2

33. Nishimasu, H. *et al.* Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**, 935–949 (2014).

34. Rohs, R. *et al.* Origins of Specificity in Protein-DNA Recognition. *Annu. Rev. Biochem.* **79**, 233–269 (2010).

35. Luscombe, N. M. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res.* **29**, 2860–2874 (2001).

36. Li, X., Xiong, X. & Yi, C. Epitranscriptome sequencing technologies: Decoding RNA modifications. *Nat. Methods* **14**, 23–31 (2016).

37. Zaccara, S., Ries, R. J. & Jaffrey, S. R. Reading, writing and erasing mRNA methylation. *Nat. Rev. Mol. Cell Biol.* **20**, 608–624 (2019).

38. Meyer, K. D. & Jaffrey, S. R. Rethinking m 6 A Readers, Writers, and Erasers . *Annu. Rev. Cell Dev. Biol.* **33**, 319–342 (2017).

39. Dominissini, D. *et al.* Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* **485**, 201–206 (2012).

40. Meyer, K. D. *et al.* Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. *Cell* **149**, 1635–1646 (2012).

41. Linder, B. *et al.* Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat. Methods* **12**, 767–772 (2015).

42. Liu, J. *et al.* A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. *Nat. Chem. Biol.* **10**, 93–95 (2014).

43. Wang, X. *et al.* Structural basis of N6-adenosine methylation by the METTL3-METTL14 complex. *Nature* **534**, 575–578 (2016).

44. Jia, G. *et al.* N6-Methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nat. Chem. Biol.* **7**, 885–887 (2011).

45. Zheng, G. *et al.* ALKBH5 Is a Mammalian RNA Demethylase that Impacts RNA Metabolism and Mouse Fertility. *Mol. Cell* (2013). doi:10.1016/j.molcel.2012.10.015

46. Lewis, C. J. T., Pan, T. & Kalsotra, A. RNA modifications and structures cooperate to guide RNA-protein interactions. *Nat. Rev. Mol. Cell Biol.* **18**, 202–210 (2017).

47. Xu, C. *et al.* Structural basis for selective binding of m6A RNA by the YTHDC1 YTH domain. *Nat. Chem. Biol.* **10**, 927–929 (2014).

48. Fustin, J. M. *et al.* XRNA-methylation-dependent RNA processing controls the speed of the circadian clock. *Cell* **155**, 793 (2013).

49. Roundtree, I. A., Evans, M. E., Pan, T. & He, C. Dynamic RNA Modifications in Gene Expression Regulation. *Cell* **169**, 1187–1200 (2017).

50. Engel, M. *et al.* The Role of m6A/m-RNA Methylation in Stress Response Regulation. *Neuron* **99**, 389-403.e9 (2018).

51. Kleinstiver, B. P. *et al.* Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* **523**, 481–485 (2015).

52. Kleinstiver, B. P. *et al.* Broadening the targeting range of Staphylococcus aureus CRISPR-Cas9 by modifying PAM recognition. *Nat. Biotechnol.* **33**, 1293–1298 (2015).

53. Nishimasu, H. *et al.* Engineered CRISPR-Cas9 nuclease with expanded targeting space. *Science (80-.).* **361**, 1259–1262 (2018).

54. Sun, H., Zhang, M., Li, K., Bai, D. & Yi, C. Cap-specific, terminal N 6-methylation by a mammalian m6Am methyltransferase. *Cell Res.* **29**, 80–82 (2019).

55. Boulias, K. *et al.* Identification of the m6Am Methyltransferase PCIF1 Reveals the Location and Functions of m6Am in the Transcriptome. *Mol. Cell* **75**, 631-643.e8 (2019).

56. Sendinc, E. *et al.* PCIF1 Catalyzes m6Am mRNA Methylation to Regulate Gene Expression. *Mol. Cell* **75**, 620-630.e9 (2019).

57. Akichika, S. *et al.* Cap-specific terminal N 6 -methylation of RNA by an RNA polymerase II–associated methyltransferase. *Science (80-.).* **363**, (2019).

58. Kabsch, W. *et al. XDS. Acta Crystallogr. Sect. D Biol. Crystallogr.* (2010). doi:10.1107/S0907444909047337

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

60. Vagin, A. & Teplyakov, A. Molecular replacement with MOLREP. *Acta Crystallogr. Sect. D Biol. Crystallogr.* (2010). doi:10.1107/S0907444909042589

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

62. Adams, P. D. *et al.* PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 213–221 (2010).

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

64. Kleinstiver, B. P. *et al.* High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490–495 (2016).

65. Kleinstiver, B. P. *et al.* Broadening the targeting range of Staphylococcus aureus CRISPR-Cas9

by modifying PAM recognition. *Nat. Biotechnol.* **33**, 1293–1298 (2015).

66. Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.* **13**, 722–736 (2015).

67. Ma, E., Harrington, L. B., O'Connell, M. R., Zhou, K. & Doudna, J. A. Single-Stranded DNA Cleavage by Divergent CRISPR-Cas9 Enzymes. *Mol. Cell* **60**, 398–407 (2015).

68. Chen, F. *et al.* Targeted activation of diverse CRISPR-Cas systems for mammalian genome editing via proximal CRISPR targeting. *Nat. Commun.* **8**, 1–12 (2017).

69. Sun, W. *et al.* Structures of Neisseria meningitidis Cas9 Complexes in Catalytically Poised and Anti-CRISPR-Inhibited States. *Mol. Cell* 1–15 (2019). doi:10.1016/j.molcel.2019.09.025

70. Yamada, M. *et al.* Crystal structure of the minimal Cas9 from Campylobacter jejuni reveals the molecular diversity in the CRISPR-Cas9 systems. *Mol. Cell* **65**, 1109-1121.e3 (2017).

71. Crooks, G., Hon, G., Chandonia, J. & Brenner, S. WebLogo: a sequence logo generator. *Genome Res* **14**, 1188–1190 (2004).

72. Waterman, D. G. *et al.* The DIALS framework for integration software. *CCP4 Newsl. protein Crystallogr.* **49**, 16–19 (2013).

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

74. Ran, F. A., Hsu, P. D., Lin, C. & Gootenberg, J. S. Resource Double Nicking by RNA-Guided CRISPR Cas9 for Enhanced Genome Editing Specificity. *Cell* **154**, 1380–1389 (2013).

75. Lee, C. M., Cradick, T. J. & Bao, G. The neisseria meningitidis CRISPR-Cas9 system enables specific genome editing in mammalian cells. *Mol. Ther.* **24**, 645–654 (2016).

76. Sternberg, S. H., Lafrance, B., Kaplan, M. & Doudna, J. A. Conformational control of DNA target cleavage by CRISPR-Cas9. *Nature* **527**, 110–113 (2015).

77. Wei, C. M., Gershowitz, A. & Moss, B. N6, O2'-dimethyladenosine a novel methylated ribonucleoside next to the 5' terminal of animal cell and virus mRNAs. *Nature* (1975). doi:10.1038/257251a0

78. Moteki, S. & Price, D. Functional coupling of capping and transcription of mRNA. *Mol. Cell* (2002). doi:10.1016/S1097-2765(02)00660-3

79. Bélanger, F., Stepinski, J., Darzynkiewicz, E. & Pelletier, J. Characterization of hMTr1, a human Cap1 2′-O-ribose methyltransferase. *J. Biol. Chem.* (2010). doi:10.1074/jbc.M110.155283

80. Keith, J. M., Ensinger, M. J. & Moss, B. HeLa cell RNA(2'-O-methyladenosine-N6-)-methyltransferase specific for the capped 5'-end of messenger RNA. *J. Biol. Chem.* **253**, 5033–5039 (1978).

81. Akichika, S. *et al.* Cap-specific terminal N 6 -methylation of RNA by an RNA polymerase II–associated methyltransferase. *Science (80-.).* **363**, 1–41 (2019).

82. Fan, H. *et al.* PCIF1, a novel human WW domain-containing protein, interacts with the phosphorylated RNA polymerase II. *Biochem. Biophys. Res. Commun.* **301**, 378–385 (2003).

83. Iyer, L. M., Zhang, D. & Aravind, L. Adenine methylation in eukaryotes: Apprehending the complex evolutionary history and functional potential of an epigenetic modification. *BioEssays* **38**, 27–40 (2016).

84. Mendel, M. *et al.* Methylation of Structured RNA by the m 6 A Writer METTL16 Is Essential for Mouse Embryonic Development. *Mol. Cell* **71**, 986-1000.e11 (2018).

85. Sheldrick, G. M. A short history of SHELX. *Acta Crystallographica Section A: Foundations of Crystallography* (2008). doi:10.1107/S0108767307043930

86. Vonrhein, C., Blanc, E., Roversi, P. & Bricogne, G. Automated structure solution with autoSHARP. *Methods Mol. Biol.* (2007). doi:10.1385/1-59745-266-1:215

87. Terwilliger, T. C. *et al.* Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. in *Acta Crystallographica Section D: Biological Crystallography* (2007). doi:10.1107/S090744490705024X

88. Goedecke, K., Pignot, M., Goody, R. S., Scheidig, A. J. & Weinhold, E. Structure of tile N6-adenine DNA methyltransferase M·TaqI in complex with DNA and a cofactor analog. *Nat. Struct. Biol.* **8**, 121–125 (2001).

89. Mauer, J. *et al.* Reversible methylation of m6 Am in the 5′ cap controls mRNA stability. *Nature* **541**, 371–375 (2017).

90. Wei, J. *et al.* Differential m 6 A, m 6 A m , and m 1 A Demethylation Mediated by FTO in the Cell Nucleus and Cytoplasm. *Mol. Cell* **71**, 973-985.e5 (2018).

91. Fu, Y., Sander, J. D., Reyon, D., Cascio, V. M. & Joung, J. K. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat. Biotechnol.* (2014). doi:10.1038/nbt.2808

92. Chen, J. S. *et al.* Enhanced proofreading governs CRISPR-Cas9 targeting accuracy. *Nature* **550**, 407–410 (2017).

93. Wang, H., La Russa, M. & Qi, L. S. CRISPR/Cas9 in Genome Editing and Beyond. *Annu. Rev. Biochem.* **85**, 227–264 (2016).

94. Liu, X. M., Zhou, J., Mao, Y., Ji, Q. & Qian, S. B. Programmable RNA N 6-methyladenosine editing by CRISPR-Cas9 conjugates. *Nat. Chem. Biol.* **15**, 865–871 (2019).

Acknowledgements

I really appreciate the six years support of Professor Osamu Nureki. When I joined into the Nureki laboratory, he often talked to me about my work and thought about my future concerns together. I felt I was not a lonely researcher and I gained the confidence of my academic career. He gave me a lot of great chances of communicating with other researchers, such as academic meetings in both Japan and other countries. I learned a lot from fruitful environments which Professor Nureki cultivates on his own effort in about 20 years. I also really appreciate the six years support of Associate professor Hiroshi Nishimasu. He taught me about everything which is necessary for science, from how to pipet, how to design an experiment, how to write a paper, how to make a figure, how to accept a paper. For six years, he was patient with my pre-mature manuscripts about 100 times and responded to me kindly. I shared a great time with my groupmates in Nureki laboratory. The senior members gave me a lot of advices and kind words which motivated me every time. We had a fruitful discussions with smart people in Nureki laboratory. I thank for all of my great collaborators. Especially, Mr. Shinichiro Akichika, Dr. Takeo Suzuki, and Dr. Tsutomu Suzuki gave me a chance of a historical research in the epitranscriptome field. Dr. Takuro Horii and Dr. Izuho Hatada gave me a great data of in vivo analysis, which appended the value of genome editing tools to CdCas9 study. Dr. Kazuya Hasegawa often helped me about X-ray diffraction experiments at SPring-8 and I determined the ten crystal structures of valuable proteins. Dr. Omar Abudayyeh, Dr. Jonathan Gootenberg, and Dr. Feng Zhang gave me the precise and valuable data, which provided the insights into the further PAM analysis and resulted in the discovery of promiscuous recognition manner. Finally, I appreciate the 28 years support of my family. I worked efficiently in the laboratory and was relaxed in my house due to the atmosphere of my family and the support of my life.