

Crystal structure of Cas9 in complex with guide RNA and target DNA

Microbes have the CRISPR-Cas adaptive immune system as a defense against invading foreign nucleic acids, such as phages and plasmids. The CRISPR-Cas system is composed of CRISPR (clustered, regularly interspaced short palindromic repeats) loci in the genome, and a cluster of Cas (CRISPR-associated) genes. The CRISPR loci consist of identical repeating sequences (referred to as repeats) interspaced by short sequences (referred to as spacers) derived from previously infected foreign nucleic acids. There are three types of CRISPR-Cas systems (types I–III). In the type II CRISPR-Cas system, dual noncoding RNAs, crRNA (CRISPR RNA), and tracrRNA (*trans*-activating crRNA), are transcribed from CRISPR loci, and bind to the effector nuclease Cas9 to form a Cas9–crRNA–tracrRNA ternary complex [1]. The ternary complex recognizes the target double-stranded DNAs through sequence complementarity between the target DNA and the 20-nt guide sequence in the bound crRNA, inducing a DNA double-strand break in the target DNA (Fig. 1). In addition to base pairing between the sgRNA and the target DNA, a short nucleotide motif located adjacent to the cleavage site of the target DNA (referred to as PAM (protospacer adjacent motif)) is necessary for the Cas9-catalyzed DNA cleavage [2]. PAMs differ among the Cas9 proteins. For instance, Cas9 from the pathogenic bacterium *Streptococcus pyogenes*, which is widely utilized for genome-editing technology,

recognizes 5'-NGG-3' as the PAM. The 20-nt guide sequences in crRNAs are derived from previously infected foreign nucleic acids, so that crRNAs serve as molecular memories.

In 2012, biochemical studies revealed that Cas9 is an RNA-guided DNA endonuclease with two nuclease domains (HNH and RuvC) [2] (Fig. 2(a)). The HNH domain cleaves the DNA strand complementary to the 20-nt guide sequence in the crRNAs (cDNA), while the RuvC domain cleaves the noncomplementary DNA strand (ncDNA) (Fig. 1). A synthetic single guide RNA (referred to as sgRNA), in which crRNA and tracrRNA are fused with a tetraloop, can also direct Cas9 to DNA cleavage (Fig. 1). Since the discovery in 2013 that the Cas9–sgRNA system can induce site-specific DNA double-strand breaks in the genome [3], the Cas9–sgRNA system has been attracting much attention as a new, versatile genome-editing technology, which works effectively in various types of cells and organisms. Catalytically dead or inactive Cas9 (referred to as dCas9) can serve as an RNA-guided genome-targeting platform, and dCas9-based new technologies, such as those for transcription regulation and chromatin imaging, have also been developed. Despite the rapid advancements in Cas9-based technologies, the mechanism by which the Cas9–sgRNA complex recognizes and cleaves the target DNA remains elusive since Cas9 shares no sequence similarity with other known proteins, except for the two nuclease domains.

To understand this unprecedented, RNA-guided DNA cleavage mechanism, we sought to solve the crystal structure of the Cas9–sgRNA–cDNA ternary complex. *S. pyogenes* Cas9 was expressed in *Escherichia coli*, and then purified to homogeneity by column chromatography. sgRNA was transcribed *in vitro* using T7 RNA polymerase, and then purified by denaturing polyacrylamide gel electrophoresis. The reconstituted Cas9–sgRNA–cDNA ternary complex was purified on a size-exclusion column and subsequently crystallized by the vapor diffusion method. X-ray diffraction data were collected at beamlines BL32XU and BL41XU, and the crystal structure of the ternary complex was determined at 2.5-Å resolution by the Se-SAD method [4]. The crystal structure revealed that Cas9 adopts a bilobed architecture (Fig. 2(b)). One lobe comprises a number of α helices and is responsible for the recognition of sgRNA and cDNA. Thus, we named it the recognition (REC) lobe. Because the other lobe contains the

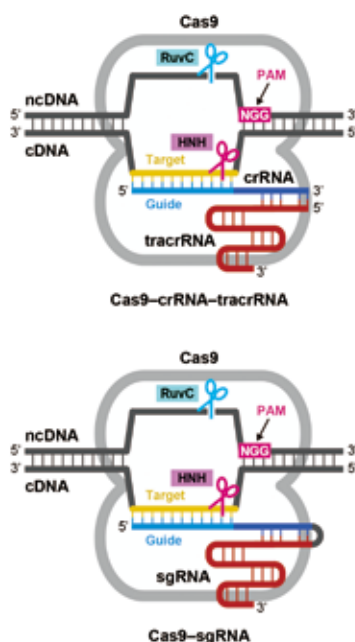


Fig. 1. RNA-guided DNA cleavage by Cas9.

HNH and RuvC nuclease domains, we named it the nuclease (NUC) lobe. The REC and NUC lobes are connected by a long α helix (named the bridge helix) and a flexible linker. In the structure, the HNH domain is located away from the cDNA cleavage site, indicating that the present structure represents an inactive form, and that the HNH domain undergoes a conformational change to approach and cleave the cDNA during catalysis. The negatively charged sgRNA:cDNA heteroduplex is accommodated in a positively charged groove between the REC and NUC lobes, and is recognized by Cas9 via the sugar-phosphate backbone of the heteroduplex. This explains the sequence-independent recognition by Cas9. The NUC lobe also contains a carboxy-terminal domain, which is located at a position suitable for interacting with the PAM. We named this the PAM-interacting (PI) domain. Indeed, our mutational analysis revealed that the PI domain participates in PAM recognition. As anticipated from their amino acid sequences, the REC lobe and the PI domain adopt new protein folds.

The ternary complex structure also reveals that the sgRNA forms a T-shaped architecture, which is composed of the sgRNA:cDNA heteroduplex, the repeat:anti-repeat duplex, and three stem loops (stem loops 1–3), where stem loops 1 and 2 are connected

by a single-stranded linker (Fig. 2(b)). The repeat:anti-repeat duplex adopts a partial duplex architecture, which is extensively recognized by the REC lobe. The PAM-proximal ‘seed’ region in the 20-nt guide sequence is extensively recognized by a cluster of conserved arginine residues on the bridge helix (Fig. 2(b)), which is consistent with the previous data showing the importance of the base pairing between the ‘seed’ region in sgRNA and the cDNA for Cas9-catalyzed cleavage [2].

Based on the structural finding that the tetraloop and stem loop 2 in sgRNA do not directly contact Cas9 and are exposed to the solvent in the ternary complex [4] (Fig. 2(b)), we recently developed a potent, dCas9-based transcription activator [5]. We generated an engineered sgRNA, in which the MS2-interacting hairpin aptamers are fused to the tetraloop and stem loop 2, and showed that three components—dCas9, the engineered sgRNA, and the MS2 protein fused with transcription activator domains—are able to mediate efficient, targeted transcriptional activation of endogenous genes.

In summary, the high-resolution structure of the Cas9–sgRNA–cDNA complex provides mechanistic insight into the RNA-guided DNA cleavage by Cas9, and thus, paves the way for the rational design of new, Cas9-based technologies.

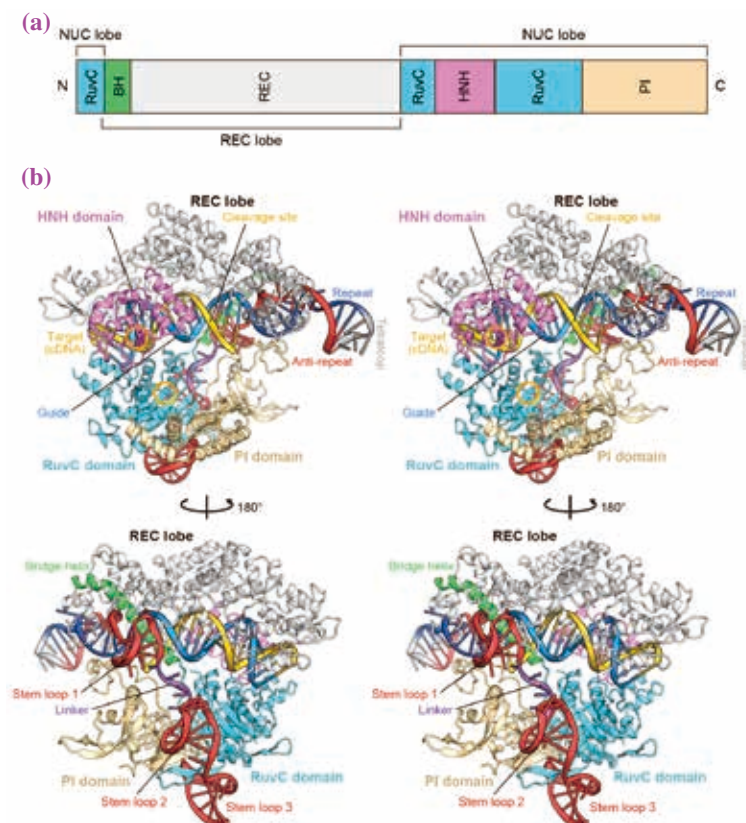


Fig. 2. (a) Domain structure of Cas9. (b) Crystal structure of the Cas9–sgRNA–cDNA ternary complex (stereoview). The active sites are indicated by orange circles. Structural image was prepared with CueMol (<http://www.cuemol.org>).

Hiroshi Nishimasu^{a,b,*} and Osamu Nureki^a

^a Department of Biological Sciences,
The University of Tokyo

^b JST/PRESTO

*E-mail: nisimasu@bs.s.u-tokyo.ac.jp

References

- [1] E. Deltcheva *et al.*: Nature **471** (2011) 602.
- [2] M. Jinek *et al.*: Science **337** (2012) 816.
- [3] L. Cong *et al.*: Science **339** (2013) 819.
- [4] H. Nishimasu, F.A. Ran, P.D. Hsu, S. Konermann, S.I. Shehata, N. Dohmae, R. Ishitani, F. Zhang and O. Nureki: Cell **156** (2014) 935.
- [5] S. Konermann *et al.*: Nature **517** (2015) 583.