Type II and type V CRISPR effector nucleases from a structural biologist's perspective

Humberto Fernandes[1]
Michal Pastor[1]
Matthias Bochtler[1,2,✉]

[1]Institute of Biochemistry and Biophysics PAS, Warsaw, Poland
[2]International Institute of Molecular and Cell Biology, Warsaw, Poland

✉International Institute of Molecular and Cell Biology, 4 Trojdena St., 02-109 Warsaw, Poland; e-mail: mbochtler@ibb.waw.pl

Received: January 27, 2016
Accepted: May 29 , 2016

Key words: CRISPR; Structural biology; Cas9; Cpf1; RuvC domain; HNH domain

Acknowledgement: This manuscript is dedicated to Alexander Wlodawer on the occasion of his birthday. Work in the authors' laboratory is supported by an NCN grant to M. Bochtler (UMO-2011/02/A/NZ1/00052).

ABSTRACT

The type II and type V CRISPR effector nucleases Cas9 and Cpf1 are "universal" DNA endonucleases, which can be programmed by an appropriate crRNA or sgRNA strand to cleave almost any DNA duplex at a preselected position (constrained only by short, so-called PAMs). In this review, we briefly introduce CRISPR bacterial adaptive immunity as the biological context in which Cas9 and Cpf1 proteins operate, and then present the structural insights that have been obtained in the last two or three years that illustrate the mode of operation of these proteins. We describe the R-loop structures at the core of the Cas9 and Cpf1 complexes, and the structure of the 5′- or 3′-handles that help anchor the nucleic acid complexes to the proteins in a manner that is independent of the target sequence. Next, we describe the molecular architecture of the Cas9 and Cpf1 proteins. We illustrate how Cas9 and Cpf1 proteins scan double stranded DNA for so-called protospacer associated motifs (PAMs), we explain how the phosphate loop (PLL) and basic helix (BH) promote the separation of target and non-target DNA strands and the formation of hybrids between crRNA or sgRNA and the target strand of DNA. We also describe the current understanding of the catalytic mechanisms of RuvC and HNH domains, and a possible, but still very uncertain catalytic role of the Nuc domain. At the end of the review, we briefly summarize key developments that have initiated the field of genomic engineering using Cas9 or Cpf1 nucleases.

INTERACTIONS OF BACTERIA WITH THEIR PATHOGENS

Bacteria defend themselves against nucleic acids from invading phages or conjugating plasmids at multiple levels [1]. Immunity can result from interference with various steps of phage infection of plasmid conjugation, and is mechanistically diverse. Endonucleases play an important role. Generic immunity of bacteria against invading DNA is typically based on a distinction between self- and non-self that relies on the modification status of the DNA [2]. In the prototypical situation, bacteria acquire a restriction modification system consisting of a DNA methyltransferase and an endonuclease of matching or slightly narrower specificity, which is inhibited by the methylation. The methyltransferase protects host DNA against endonuclease cleavage. Invading DNA lacks the modification, and is cleaved by the endonuclease before it is protected by the methyltransferase. Propagation of phages or plasmids in hosts that protect their own DNA by methylation can endow invading DNA with the hallmarks of "self". This has in some cases prompted a reversal of the role of methylation, now treated as a hallmark of non-self. Restriction systems for the defense against modified DNA do not require a protective methyltransferase for the host, and rely on endonucleases that only cleave modified DNA [3]. Either way, immunity is generic, and does not depend on a prior interaction between host and pathogen. In some sense, this type of bacterial immunity is functionally analogous to the "innate" immunity of vertebrates against their pathogens.

CRISPR ADAPTIVE, PATHOGEN SPECIFIC IMMUNITY

Adaptive, pathogen specific immunity of bacteria against invading nucleic acids has recently moved into the focus of interest. It is not based on the modification status of DNA, but instead relies on identification of pathogens based on the storage of their "fingerprints" in the form of a "blacklist" centrally encoded in the bacterial genome [4] (Fig. 1). The "blacklist" is a CRISPR cluster that contains short stretches of pathogen sequence separated by repeats. Prior to the elucidation of their biological role, the short stretches of pathogen sequence were termed "spacers". This terminology has remained in place, even though with hindsight it is now clear that the "spacers" actually carry the important information in CRISPR clusters, whereas the repeats just serve to "separate" spacers representing different pathogens [5]. CRISPR genes are typically associated with a group of so-called CRISPR-associated (Cas) genes, which are both involved in the acquisition of new spacers (i.e. the acquisition of new immunity after encounters with pathogens), as well as with expression of immunity [6,7]. Most bacteria harboring CRISPR-Cas systems share characteristic *cas* genes (typically *cas1-cas6*) [8], which are complemented by additional, less widely distributed *cas* genes [9]. Expression of immunity is always based on the transcrip-

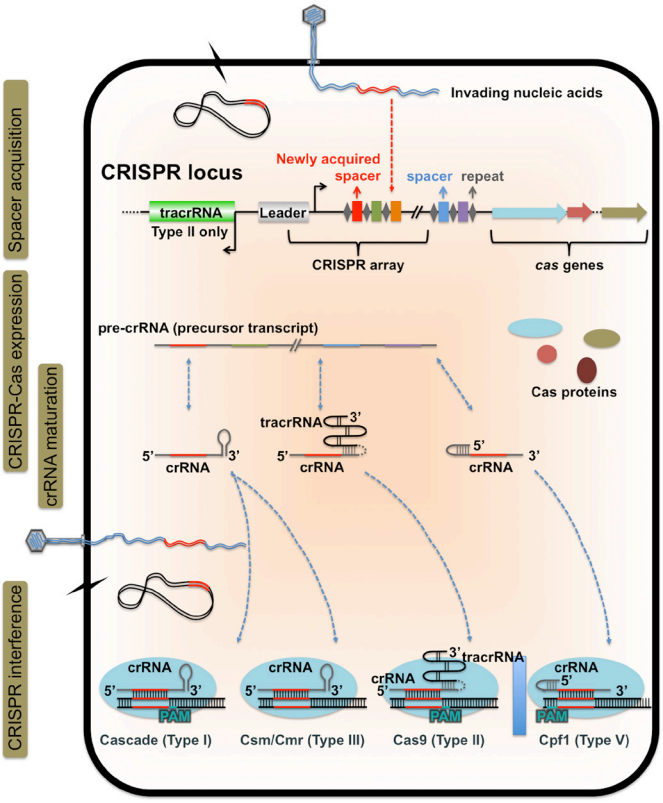

Figure 1. Overview of CRISPR-Cas adaptive immune system. Depicted are the spacer acquisition from invading nucleic acids, the CRISPR-Cas expression and crRNA maturation (RNAses involved not shown), and CRISPR interference by crRNA-guided surveillance. The figure has been adapted from Jiang and Doudna [68].

tion of the CRISPR clusters into so-called pre-crRNA, which is subsequently cleaved by RNA endonucleases into crRNAs, which provide immunity against individual pathogens. The crRNAs then serve as guides that hybridize with target DNA or RNA to direct the cleavage and thus destruction of the invading nucleic acids [10]. In some ways, CRISPR immunity of bacteria is functionally equivalent to adaptive immunity of vertebrates. However, CRISPR target selection by annealing of complementary nucleic acids is a much more predictable process than the interaction of an antibody or B- or T-cell receptor with a peptide epitope. Therefore, clonal selection is not necessary for CRISPR immunity. At the level of bacterial communities, however, bacteria that express "useful" immunity against a pathogen may be selected, in a process reminiscent of clonal selection in vertebrate adaptive immunity. In contrast to vertebrate adaptive immunity, which is not heritable, bacterial adaptive immunity is heritable at the level of cells. At the level of bacterial communities, CRISPR immunity may also spread by horizontal gene transfer (as evidenced by the "sparse" distribution of CRISPR systems across bacterial species), even though the CRISPR systems work of course against horizontal gene transfer, at least at the level of populations [11].

CLASSIFICATION OF CRISPR SYSTEMS

A recent review distinguishes two major classes of CRISPR systems, according to the type of effector nuclease [12,13]. In the class 2 systems, the effector nuclease is a monomer and consists of a single polypeptide chain. This class is further subdivided into the types II and V, and VI. The type II CRISPR systems have a Cas9 endonuclease. This endonuclease has two separate catalytic domains belonging to the RuvC and HNH catalytic groups. The HNH domain cuts the crRNA bound target strand, the RuvC domain the non-target strand [14]. Type V systems have a Cpf1, C2c1 or C2c3 type endonuclease, which shares both homology and analogy with Cas9 type endonucleases [13]. Type VI endonucleases differ from type II and type V endonucleases in that no RuvC motifs have been detected, and are still largely uncharacterized. Until very recently, both type II and type V enzymes were considered to be exclusively DNA endonucleases. However, very recent work has shown that Cpf1 is also involved in processing precursor RNA, using an active site not implicated in DNA cleavage [15]. In contrast to the class 2 effector endonucleases, class 1 effector endonucleases are multi-subunit protein complexes. They can be further classified into type I, type III and type IV, with characteristic signature genes (*cas3* for type I, *cas10* for type III). In type I complexes, a Cascade (or Cascade like) protein complex selects the target, and then recruits the Cas3 helicase/endonuclease, which cleaves target DNA processively [16,17]. Type III complexes can target DNA and RNA [18-23], relying on independent active sites in the complexes. Type IV endonucleases are still poorly studied. In this review, we focus on the recent structures of type II and type V CRISPR effector nucleases, which have been the focus of intense research activity due to their applications in genome editing.

So far, structures are available for Cas9 from *Streptococcus pyogenes* (SpCas9, PDB-accessions 5FW1, 5B2R, 5B2S, 5B2T, 5F9R, 5FQ5, 4ZT0, 4ZT9, 4UN3, 4UN4, 4UN5, 4OO8, 4CMP, 4CMQ) [24-29], *Francisella novicida* (FnCas9, PDB-accessions 5B2O, 5B2P, 5B2Q) [30], *Actinomyces naeslundii* (AnaCas9, PDB-accessions 4OGC, 4OGE) [29] and *Staphylococcus aureus* (SaCas9, PDB-accessions 5FW1, 5AXW, 5CZZ) [31]. Crystal structures include structures of Cas9 alone [29], Cas9 in complex with single guide RNA (sgRNA, a fusion of crRNA and tracrRNA) [32], Cas9 with a DNA-RNA duplex [28], and a large number of structures of Cas9 with the R-loop that can form in the presence of both double stranded target DNA and sgRNA. Very recently, a crystal structure has been recruited which includes not only the double-stranded region of target DNA from the PAM sequence onwards and the sgRNA, but also a long region of single-stranded DNA representing the non-target DNA strand. For Cpf1, crystal structures of the protein from *Lachnospiraceae* bacterium ND2006 (LbCpf1, PDB-accession 5ID6) [33] and from an *Acidaminococcus* sp. (AsCpf1, PDB-accession 5B43) [34] have been published very recently.

Both Cas9 and Cpf1 proteins are complicated, multi-domain proteins. In this review, we will start from the nucleic acid structures, and then proceed to the structures of the proteins, which have evolved to interact with these structures.

AN R-LOOP STRUCTURE AT THE CORE OF THE GENE TARGETING COMPLEXES

The core structural motif of active Cas9 or Cpf1 complexes with substrates is the R-loop, which consists of an RNA-DNA duplex and displaced single DNA strand (Fig. 2A). In the context of Cas9 or Cpf1 complexes, the RNA strand of the R-loop is the spacer fragment of crRNA. The DNA strand that hybridizes with the crRNA strand is known as the target strand (tsD-

NA). Its sequence is complementary to the crRNA sequence. The DNA strand that is unpaired in the R-loop region is known as the non-target strand (ntsDNA). The unpaired region is the protospacer region, with a base sequence identical to the sequence of the crRNA spacer fragment (except of course for the uracil in RNA *versus* thymine in DNA difference). At the 5′- and 3′-ends, a switch-over point marks the transition from the RNA-DNA duplex to the DNA-DNA duplex. Protospacer adjacent motifs (PAMs) are found next to protospacers in DNA substrates, but not next to spacers in CRISPR clusters [5]. The requirement for a PAM in a substrate makes it possible for the effector endonucleases to distinguish between substrates and the CRISPR cluster, preventing auto-destruction of CRISPR arrays. In the complexes with Cas9 or Cpf1 proteins, the PAM regions are located in regions of double stranded DNA, just adjacent to the switch points, either upstream or downstream of the R-loop region ("upstream" and "downstream" are defined based on the direction of crRNA from 5′end to 3′end) (Fig. 2B, 2C). Irrespective of their location upstream or downstream of the R-loop region, PAM sequences always refer to the non-target strand. For downstream PAMs, the first nucleotide of PAM is also the first nucleotide of DNA-DNA Watson-Crick pairing. For upstream PAMs, the last nucleotide of the PAM is also the last nucleotide of DNA-DNA duplex.

HANDLES ON THE crRNA IN THE GENE TARGETING COMPLEXES

The crRNA spacer region is of arbitrary sequence, Cas9 and Cpf1 proteins make only sequence unspecific interactions with the phosphodiester backbone (Fig. 3), but no specific interactions with the DNA bases. Sequence specific contacts defining the register of protein-nucleic acid interactions occur in the repeat regions, but not the spacer regions of the crRNA.

Cpf1 bound crRNAs contain repeat-derived regions upstream of the spacer, which fold into 5′ stem loops. The Lb-

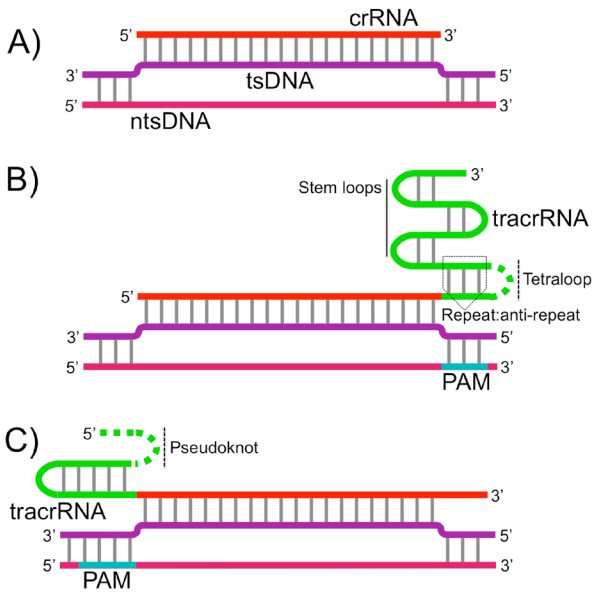

Figure 2. R-loops. A) Strand nomenclature of R-loops. B) Cas9 and C) Cpf1 crRNA and tracrRNA system, adapted from [32,34]. tracrRNA is displayed in short form for simplification. The color code of this figure applies to all other figures as well, unless otherwise stated.

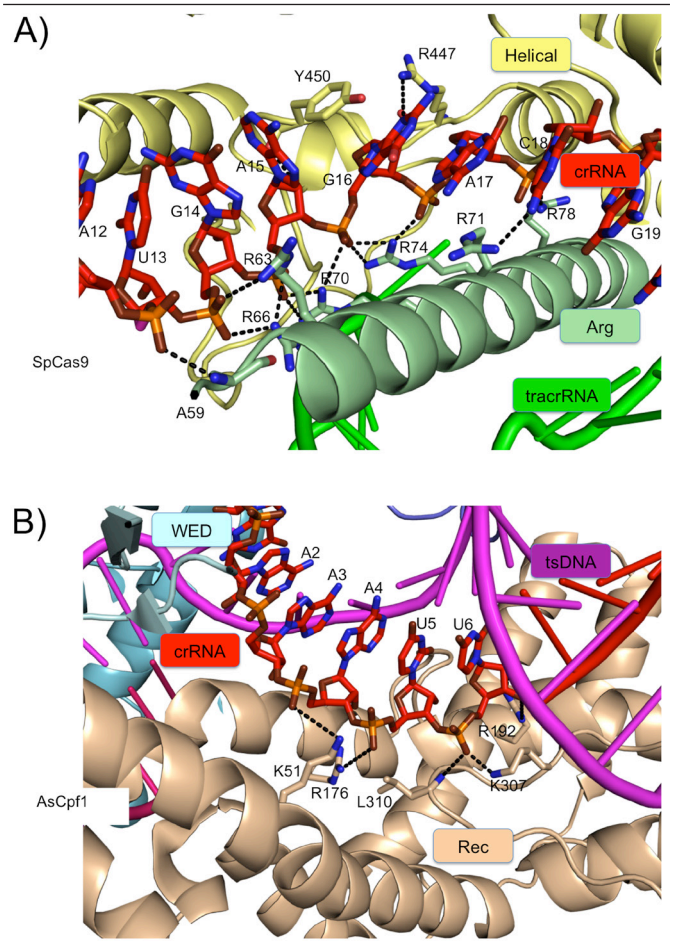

Figure 3. crRNA binding. A) SpCas9 binding of crRNA. Numbering as in PDB ID 4ZTO. The ntsDNA has been truncated (dashed line) for clarity. B) AsCpf1 binding of crRNA. Numbering as in PDB ID 5B43. Figure prepared in PyMol.

Cpf1 bound crRNA indeed forms a simple stem loop. Non-canonical interactions between nucleobases occur at the base of the stem (where RNA-RNA base pairing ends). For AsCpf1, the expected stem loop is also present, but because the most 5′-nucleotides interact with the stem loop, the 5′-handle is better described as a pseudoknot.

Cas9 bound crRNAs are anchored at least in part indirectly, by base pairing interactions between the "repeat" of the crRNA and the (mostly) complementary "anti-repeat" of the trans-encoded crRNA (tracrRNA) [35]. Cas9 interacts with the repeat anti-repeat duplex, as well as with the 3′-handle of the tracrRNA downstream of the anti-repeat. Jinek and colleagues have shown that both the repeat and anti-repeat regions can be truncated (on the 3′-side of the repeat and the 5′-side of the anti-repeat), and that the newly generated ends can be connected by a tetraloop, leading to a sgRNA, which functionally replaces the duplex of crRNA and tracrRNA [36]. Due to its more compact nature, sgRNA has been used instead of crRNA tracrRNA combinations in all reported crystal structures to date.

Crystallographic data confirm that the repeat anti-repeat regions form mostly base paired duplexes, except for a bulge roughly in the middle of the repeat anti-repeat double helix. In the bulge region, non-canonical interactions, for example between a nucleobase and the phosphodiester backbone are

observed. Moreover, several bases are flipped and thus provide "handles" for Cas9 to "hold onto" and help define the register of binding of the repeat anti-repeat region. In the Cas9 complexes, the tracrRNA extends beyond the anti-repeat region and forms a series of stem-loops (numbered according to their position in the tracrRNA sequence from the anti-repeat towards the 3′-end). The degree of conservation of the stem loops decreases towards the 3′-end. Stem loop 1 is partially conserved between *S. pyogenes* and *S. aureus*, but not *F. novicida*. Three stem loops are present in *S. pyogenes*. *S. aureus* tracrRNA or sgRNA have two 3′-stem loops, but for crystallization purposes, the RNA was truncated after stem loop 1. For *F. novicida*, two stem loops are detected, which are both dissimilar in sequence to the ones in *S. pyogenes*. The overall shape of stem loop 1 and also its orientation with respect to the repeat anti-repeat duplex are largely conserved, and together with the repeat anti-repeat duplex, stem loop 1 forms a core structural motif in the complexes. The other stem loops are structurally diverse. Variations in stem loop sequence, structure and overall shape explain why tracrRNAs or sgRNAs tend to be non-interchangeable except between closely related species [37].

At first sight, the indirect anchoring of the crRNA by tracrRNA in the Cas9 complexes seems unnecessary complicated by comparison with the simple attachment of a 5′-handle to the crRNAs in the Cpf1 complexes. However, at second sight, the indirect mode of anchoring the crRNA may have its advantages. First, we note that the repeat anti-repeat region of the crRNA tracrRNA duplex (more precisely the part that is preserved in the sgRNA) interacts with Cas9 proteins, providing an additional handle in addition to the 3′-end of the sgRNA or tracrRNA. We also note that in *S. pyogenes*, the combined length of the repeat of crRNA and tracrRNA is approximately 100 nucleotides, and that the (truncated) sgRNAs for Cas9 from this species still contain about 80 nucleotides of non-spacer sequence. If entire sgRNAs were derived from CRISPR repeats, these would have to be at least 80 and perhaps even 100 nucleotides long, compared to actual CRISPR repeat lengths between 20 and 50 nucleotides [38]. Repeats are generally associated with genetic instability. In CRISPR systems, some of the instability is desirable, because it makes it possible to eliminate spacers (and associated repeats) that have lost their biological usefulness [39]. One may speculate that actually observed repeat lengths suffice to keep CRISPR clusters adaptable, and that additional genetic instability associated with longer repeats may be undesirable. To our knowledge, this explanation has not been tested, but the idea has some support from the distribution of spacer lengths across CRISPR systems from a wide variety of species.

COMPLETE NUCLEIC ACID STRUCTURES IN Cas9 AND Cpf1 TARGETING COMPLEXES

The nucleic acid structures in Cas9 and Cpf1 complexes are rather different from each other. Leaving aside the non-target DNA strand, which is present in all cases, but has so far been tracked crystallographically only in the Cas9 complex [26], the nucleic acid arrangement in the Cpf1 complex can be described as a three-arm structure, with double stranded DNA (with PAM) forming one arm, the crRNA-DNA heteroduplex forming another arm, and finally the 5′-handle of the crRNA forming the third arm. In Cas9 complexes, a four armed structure is formed. The arms are the sgRNA-DNA heteroduplex, the repeat anti-repeat RNA duplex, the (substrate) DNA-DNA duplex, and the 3′-handle of the tracrRNA (Fig. 4).

In the Cas9 complexes, the sgRNA in the spacer and repeat region follows roughly the trajectory that would be expected for it in an A-form duplex (Fig. 3A). As a result the guide-target DNA-RNA heteroduplex and the repeat anti-repeat heteroduplex are approximately co-linear (Fig. 4A). The target DNA strand changes direction drastically at the switchpoint. Therefore the duplex with the non-target DNA strand and the duplex with the spacer region of sgRNA are almost perpendicular to each other. The direction of the 3′-handle is not conserved between different Cas9 complexes. In the SpCas9 complex, this region runs approximately parallel to the DNA duplex, whereas it runs approximately perpendicular to it in the FnCas9 complex. In the Cpf1 complexes, the least perturbed strand is the DNA target strand. Therefore, the DNA duplex region (with PAM) and the RNA-DNA heteroduplex are approximately collinear. The 5′-handle runs approximately perpendicular to this direction. Thus, the nucleic acid structures that need to be created and processed by Cas9 and Cpf1 proteins are quite different, and hence it is no surprise that also the proteins themselves are quite different from each other.

Cas9 AND Cpf1 DOMAIN ORGANIZATION

Cas9 proteins are relatively large, monomeric multi-domain proteins. Their size varies considerably, in the range from about 900 to about 1600 amino acids. SaCas9 with only 1053 amino acids belongs to the "small" Cas9 proteins. FnCas9 with 1629 amino acid is a "large" Cas9. The prototypical SpCas9, which is currently used in most genome engineering applications, consists of 1368 amino acids and thus falls in terms of size somewhere midway between the "small" and "large" Cas9 proteins. All Cas9 proteins studied to date have a bi-lobed architecture. The two lobes are known as the recognition (REC, no relation to helicases) and nuclease (NUC) lobes [28] (Fig. 5A).

In the prototypical SpCas9, the REC lobe itself is built from a long α-helix, termed the bridge helix (BH), and three subdomains, termed REC1, REC2 and REC3. REC1 is directly downstream of the BH, REC2 is an "insertion" in the REC1 domain. REC3 is downstream of REC1, and was included in REC1 in some early descriptions of the structure. The NUC lobe consists of a RuvC-like domain, and HNH like-domain, a wedge domain (WED, included in an adjacent domain in the original description of the structure [28]), and the PAM-interacting domain (PI). The WED domain has been given its name, because it is wedged between the repeat anti-repeat RNA duplex and the target strand non-target strand DNA duplex in the crystal structures. The PAM interacting domain, as its name suggests, is responsible for PAM sequence recognition. In some descriptions, this domain is further subdivided into a topoisomerase-homology domain (TOPO) and a C-terminal domain (CTD) (Fig. 5A). Biochemical experiments have demonstrated that the RuvC and HNH-domains cleave the non-target and target strands, respectively [14]. The entire SpCas9 protein is best thought as a very elaborate version of a RuvC like domain (Fig. 5B, C). The HNH domain and the entire REC lobe are insertions in the RuvC domain, the WED and PI domains are

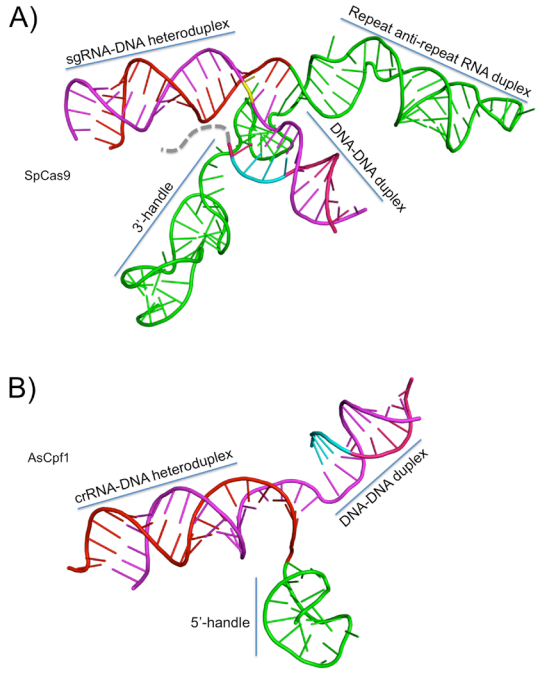

Figure 4. R-loops cartoon representation, and arms assignment. A) Four arms present in the SpCas9 structure (PDB ID 5F9R). B) Three arms present in AsCpf1 structure (5B43). The orientation was chosen so that the strands of the sgRNA-DNA heteroduplex overlap. Figure prepared in PyMol.

downstream of the RuvC domain. The "fragmentation" of the RuvC domain in Cas9 suggests that this domain is the "original" scaffold, which has then be elaborated by the insertion of the REC lobe (itself comprising the BH and REC1, REC2 and REC3 domains) and the HNH domain, as well as by the addition of the phosphate-lock loop (PLL, a short but functionally important part of the structure), and of the WED and PI domains at the C-terminus.

The smaller SaCas9 is generally similar to the architecture of SpCas9, but there are several contractions. First, the recognition lobe spans only 385 residues, compared to 658 residues for the recognition lobe of SpCas9. A major contraction is also seen in the PI, where the SaCas9 CTD comprises 85 residues, compared to 168 amino acids in the SpCas9. On the other hand, the WED domain, which is so compact in SpCas9 that it was not originally considered as an independent domain, is larger in SaCas9 compared to SpCas9.

The large FnCas9 also has essentially the same conserved architecture, but now there are mostly additional insertions compared to the "reference" SpCas9. The REC lobe of FnCas9 is expanded (775 residues) compared to the REC lobe of SpCas9 (658 residues), and also BH, REC1, REC2 and REC3 composition, with REC2 inserted into REC1. However, parts of the REC lobe of FnCas9 are structurally unique and have no counterparts in SpCas9, and surprisingly, the FnCas9 REC2 domain adopts a new fold unrelated to its counterpart in SpCas9. The NUC lobe of FnCas9 is similar in organization to the NUC lobes of the other Cas9 proteins, but the WED domain is larger than in SpCas9 or SaCas9.

Cpf1 proteins are superficially similar in architecture to Cas9 proteins, but most of the similarity results from anal-

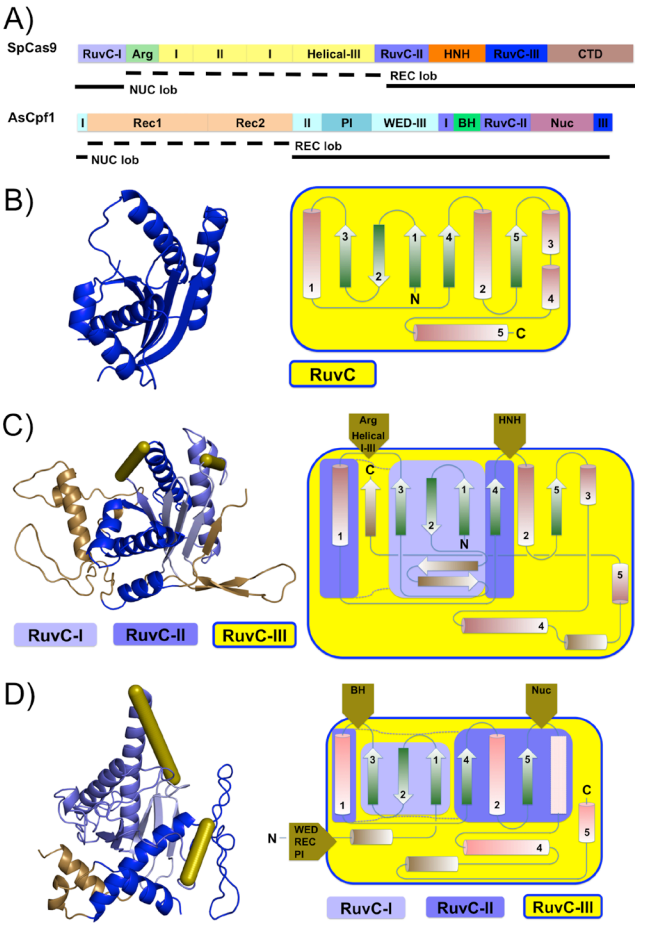

Figure 5. Domain organization, topology and structural aspects of Cas9 and Cpf1. A) 1D domain organization of SpCas9 and AsCpf1 and annotation of the respective proteins lobes, REC and NUC. B) Structure (left panel) and topology arrangements (right panel) of RuvC (PDB ID 4LD0). C) Structure (left panel) and topology arrangements (right panel) of SpCas9 (PDB ID 5F9R). Secondary structure elements inserted on the RuvC fold (blue) are in light brown color, and domain insertion represented by rods/square-arrows in olive color. D) Structure (left panel) and topology arrangements (right panel) of LbCpf1 (PDB ID 5ID6). Secondary structure elements inserted on the RuvC fold (blue) are in light brown color, and domain insertion represented by rods/square-arrows in olive color. Structural panels prepared with PyMol.

ogy rather than homology, as already predicted before the first Cpf1 crystal structures were solved [13] (Fig. 5A). As the LbCpf1 [33] and AsCpf1 [34] were solved and interpreted independently of each other, there is no unified nomenclature for the domains in Cpf1 proteins yet. In fact, due to the timing of crystal structures, and the involvement of different laboratories, the unfortunate situation has now arisen that domains with perceived similar or analogous functions, but no homology carry the same labels, whereas domains that are clearly homologous, but were studied by different groups, carry different labels. We suspect that the nomenclature of Yamano and colleagues (who had DNA in their complex to facilitate interpretation) [34] will prevail and therefore use it preferentially.

The AsCpf1 protein has been described as bi-lobed by the authors of the crystal structure [34]. By analogy with Cas9, the two lobes have been termed the REC and NUC lobes (Fig. 5A). The REC lobe consists of REC1 and REC2 domains, which have similar functions as their counterparts in Cas9 proteins, but are not similar in sequence or struc-

ture (except for strong overrepresentation of α-helices in the fold). The NUC lobe consists of a RuvC domain, and at least three additional domains that are not homologous to domains in Cas9. Based on similar roles in the complexes, two of these domains have been called the WED and PI domains. The third domain may be involved in DNA cleavage and has been called the Nuc domain (not to be confused with the NUC lobe, which contains the Nuc domain among others). There is no HNH domain in Cpf1 proteins. Biochemical data show that inactivation of the RuvC active site in Cpf1proteins blocks cleavage of both DNA strands, and this finding was originally interpreted as evidence that the RuvC domain cleaves both strands [40]. In the AsCpf1 structure, the RuvC domain is not well positioned to cleave the target DNA strand, but the Nuc domain would be well positioned for this task, prompting the hypothesis that the Nuc domain carries out this cleavage step [34]. However, the Nuc domain does not appear to be homologous to known nucleases, and the active site has not yet been identified. Moreover, the mechanistic basis for RuvC control of Nuc activity remains enigmatic. Therefore, the assignment of catalytic activity to the Nuc domain remains tentative.

The LbCpf1 has a similar structure, but the nomenclature is different [33]. The authors of the structure noticed similarity between the LbCpf1 counterpart of the WED domain and a known RNA binding protein, termed small protein B, and termed the domain oligonucleotide-binding domain (OBD). The looped-out helical domain (LHD) in LbCpf1 corresponds to the PI in AsCpf1, but was not identified as PAM interacting, because the crystallized LbCpf1 complex contains only crRNA, but not target DNA. The helix-loop-helix region in LbCpf1 comprises the BH (basic helix), which is interacting with the crRNA. The counterpart of the Nuc domain in AsCpf1 is the "unknown" (UK) domain of LbCpf1, to which the authors of the crystal structure did not ascribe nuclease activity (but also did not exclude it).

PROBABLE EVOLUTION OF Cas9 AND Cpf1 PROTEINS FROM A HOLLIDAY JUNCTION RESOLVASE

The presence of the RuvC domain in both Cas9 and Cpf1 proteins and the large number of insertions in this domain (only the WED domain in Cpf1 is similarly "fragmented") suggests that both Cas9 and Cpf1 proteins have evolved from a RuvC-like Holliday junction resolvase (Fig. 5B, C, D). Superficially, this idea gains additional support from the observation that both RuvC proteins and CRISPR proteins interact with nucleic acid structures that contain single and double stranded regions, as well as cross-over regions between them. Surprisingly, detailed structural comparisons show that the nucleic acid binding modes of the RuvC and Cas9 complexes are not closely related. Despite this difference, it is likely that both Cas9 and Cpf proteins have evolved from RuvC-type Holliday junction resolvases or closely related transposases [12]. In fact, a careful analysis of Cas9 and Cpf1 sequences suggests that the Cas9 and Cpf1 families have evolved from different families of transposon encoded Tnp proteins, containing either only RuvC and HNH or only RuvC domains. An "early" split of Cas9 and Cpf1 proteins is also suggested by the observation that Cpf1 loci contain Cas1, Cas2 and Cas4 proteins more closely related to those in type I than type II CRISPR systems [40,41].

ROLE OF THE PI IN PAM READOUT

Cas9 and Cpf1 both read DNA sequence by protein-DNA interactions (with the PAM) and by nucleic acid hybridization (between target strand and crRNA). Probing DNA for the presence of the PAM consensus is possible in the context of double stranded DNA, whereas sequence comparisons by hybridization require local DNA melting. For efficient scrutiny of long DNA molecules, one can therefore expect that the scan for PAM sequences should be the first step in target recognition. Experiments confirm this expectation [27,42], and crystal structures further demonstrate that the PAM is indeed read out in the context of double stranded DNA [27]. Conceptually, detection of PAM sequences is therefore similar to the detection of the recognition sequence by restriction endonucleases, and one can expect typical motifs for the binding of bases or base pairs, such as arginine guanine interactions and interactions between adenines and glutamine residues. In some cases, the PAMs are partially degenerate, and for example purines (R) or pyrimidines (Y) in one strand, but not a specific base, are required in a given position. From prior experience for amino acid based readout of DNA sequence (as opposed to nucleic acid based readout) one may then expect two possible strategies. First water mediated hydrogen bonds may replace direct hydrogen bonds. As water may act as a donor or acceptor for two hydrogen bonds each, water mediation of hydrogen bonds can select for either a hydrogen bond donor or acceptor. Second, contacts may be mediated by asparagine or glutamine residues, which can serve as hydrogen bond donors or acceptors (of similar shape), depending on the side chain rotamer. Inspection of interactions between PAM interacting domains and PAMs in the available crystal structures shows that sequence specificity for PAMs indeed operates based on the expected biophysical principles.

Cas9 PAM interactions have been studied for several years now, starting from the elucidation of the first Cas9 protein in complex with an R-loop comprising the PAM region [27]. For both SpCas9 and FnCas9, the PAM sequence is NGG. Crystal structures show that the PAM guanines engage in bidentate hydrogen bonding interactions with arginine residues [27,30] (Fig. 6A). For SaCas9, the PAM is NNGRRT (where R stands for purine, i.e. either A or G). The G in the PAM is equivalent to the "last" G in the NGG PAM and interacts via its Hoogsteen edge with the guanidino group of an arginine residue. Selection for the purines (RR) and against pyrimidines (YY) downstream of the G is based on interactions between the bases and asparagine residues, or water mediated hydrogen bonds [31], in agreement with the principles outlined for semi-degenerate recognition of DNA bases outlined above. The PAMs for proteins in the Cas9 family all appear related, and some of the interactions between the proteins and the conserved PAM DNA bases are very similar and probably homologous. However, there are also proteins in the Cas9 family, as noted by Anders and colleagues, which have very different PAMs. For example, the PAM for Cas9 from *Lactobacillus buchneri* is NAAAA [43], The loop in the PAM interacting domain that anchors amino acids that make direct contacts has a very different sequence in the *L. buchneri* Cas9 (QLQ) compared to *S. pyogenes* Cas9 (RYR) [24]. However, a crystal structure that elucidates the

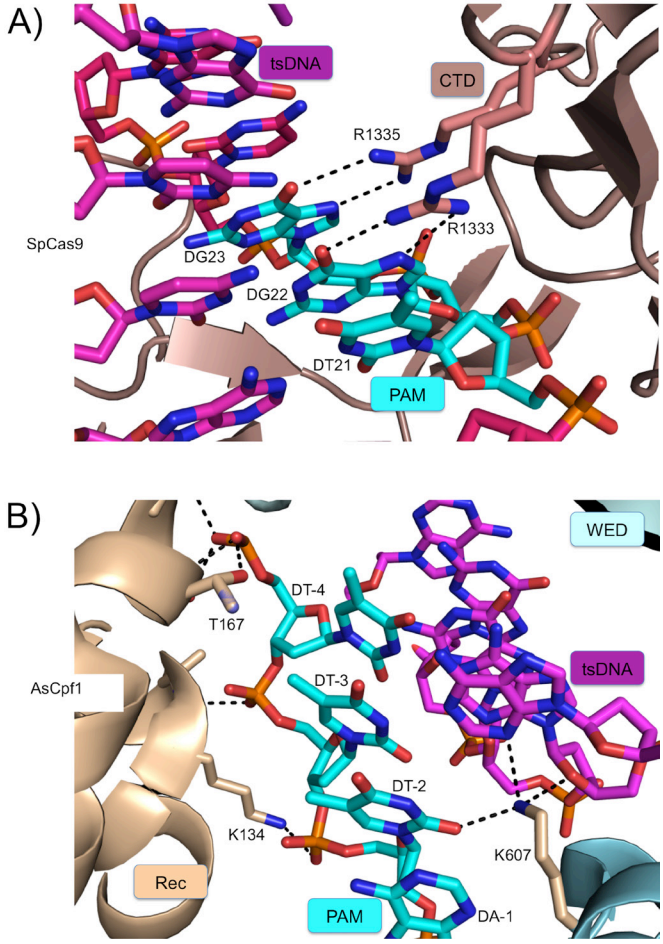

Figure 6. PAM recognition. A) Structural details for the SpCas9 protein. Numbering as in PDB ID 5F9R. B) Structural details for the AsCpf1 protein. Numbering as PDB ID 5B43. Figure prepared in PyMol.

detailed interactions between *Lactobacillus buchneri* Cas9 and the PAM for this species has not yet been solved.

As PAMs constrain uses of RNA guided endonucleases, many efforts have been made to generate Cas9 proteins which accept very generic PAMs, or to make Cas9 variants that recognize novel PAM sequences. The task is not easy, because Cas9 activity goes down when PAM binding is poor and because it is hard to find protein motifs that interact tightly with very short stretches of DNA. Nevertheless, some successes have already been achieved, in some cases by design (the FnCas9 PAM specificity has been altered from NGG to YG) [30], and in other cases by a combination of rational design and selection strategies to identify useful protein variants from large and diverse pools of candidate variants. SpCas9 variants ("VQR", "EQR" and "VRER") have been found that are compatible with NGAN, NGNG and NGCG PAMs [44], and very recent crystal structures have confirmed the expected structural explanations for the altered specificities [24,30].

Cpf1 PAM interactions could not be predicted based on Cas9 PAM interactions, because the PAM interacting domains are not homologous between Cpf1 and Cas9 proteins (as suggested not only by sequence and structure comparisons, but also by the location of the PAM upstream and downstream of the heteroduplex region for Cpf1 and Cas9, respectively). The very recent crystal structure of AsCpf1 offers a glimpse on how Cpf1 proteins interact with PAM regions [34] (Fig. 6B). The PAM for AsCpf1 is TTTN, the AT rich duplex with this sequence assumes a distorted DNA structure with very narrow minor groove. Interactions are from both the major and minor groove sides. Sequence specificity appears to be partly based on hydrogen bonding interactions, and partly also on shape selection (e.g. a guanine NH2 group in the central minor groove would clash with a lysine of AsCpf1, thymine methyl groups occupying a unique position on the outer major groove side not accessible to other bases stack very favorably). The LbCpf1 crystal structure contains only crRNA, but not target DNA, and is therefore not informative about PAM interactions (the LbCpf1 PAM is complicated and may be approximately described as YYYN, where Y stands for T or C, and T is preferred over C in all three positions [40]).

ROLE OF THE PLL AND BH IN DNA STRAND SEPARATION

Adjacent to the PAM, DNA strands have to be separated for interaction with crRNA (or sgRNA), in a sequence-non-specific manner. In the case of R-loop complexes with Cas9, the phosphodiester group linking the last residue of the strand-separated region of DNA to the first residue in double stranded DNA seems to play a key role. This phosphate group engages in several hydrogen bonding interactions with amino acids of a loop in Cas9, which has therefore been termed the "phosphate lock" loop [27]. The lock loop acts to fix the DNA backbone in such a manner that the bases adjacent to the switch-point (between DNA-DNA duplex and DNA-RNA heteroduplex) point essentially in opposite directions, without much affecting the "trajectory" of the phosphodiester backbone (Fig. 7). This drastic deformation of the DNA target strand assures that the bases in the protospacer complementary region are fully accessible for interactions with crRNA (or sgRNA). Mutations to amino acids of Cas9 involved in the "locking" reduce or abolish Cas9 activity, especially when mutations occur in combination [27]. In FnCas9 and SaCas9, the amino acids of the lock loop differ from those in SpCas9, but because the interactions are between phosphate and main-chain, the mode of action of the lock-loop is nonetheless conserved [30,31].

Both Cas9 and Cpf1 proteins do not have known binding sites for ATP or other molecules, which could serve as a source of energy, in order to drive the separation of DNA strands. Thermodynamics suggests that such a source of energy is also not required. RNA-DNA hybrids are typically more stable than DNA-DNA hybrids, but differences are sequence dependent, and in rare cases, the relationship can be reversed [45]. Whether relative RNA-DNA hybrid versus DNA-DNA stability affects the efficacy of targeting does not appear to have been addressed in a systematic manner. However, structures have already shown that Cas9 and Cpf1 proteins facilitate RNA-DNA hybrid formation by pre-binding the crRNA in the spacer region in an A-like conformation that is typical for RNA-DNA duplexes, thus facilitating their formation (Fig. 3). The interactions anchoring the phosphodiester backbone of the spacer region of the crRNA or sgRNA are sequence non-specific and predominantly involve the

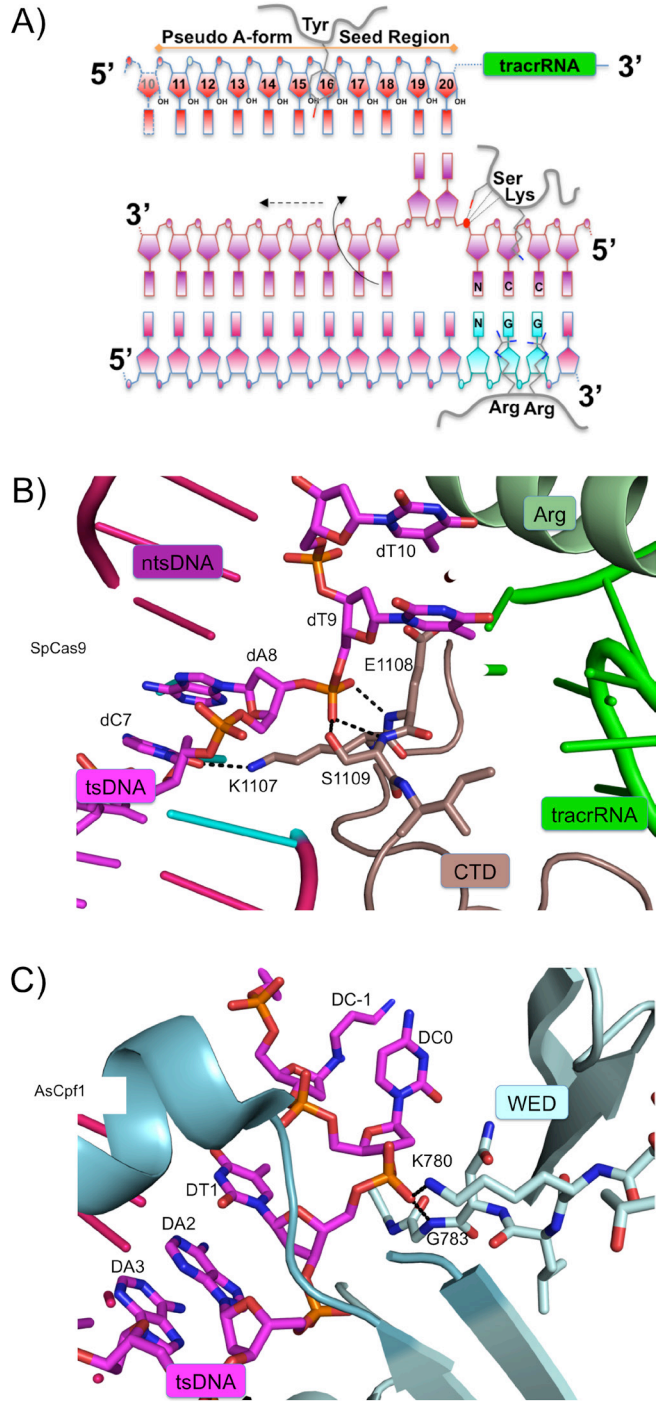

Figure 7. Phosphate-lock loop and base flip. A) Scheme of the phosphate-look loop mechanism and initial steep of RNA-DNA heteroduplex formation for Cas9 proteins. B) Structural detail for the SpCas9 protein. Numbering as in PDB ID 4UN3. C) Structural detail for the AsCpf1 protein. Numbering as in PDB ID 5B43. Panels B and C prepared in PyMol.

phosphates in the RNA backbone and basic residues, typically arginine residues in the BH.

DNA CLEAVAGE

Both Cas9 and Cpf1 cleave the target DNA strand in the heteroduplex region and the non-target strand in the region where it is single stranded. In all cases, cleavages occur on the "downstream" side of the R-loop, close to the PAM for Cas9 and distant to the PAM for Cpf1. Cas9 makes blunt

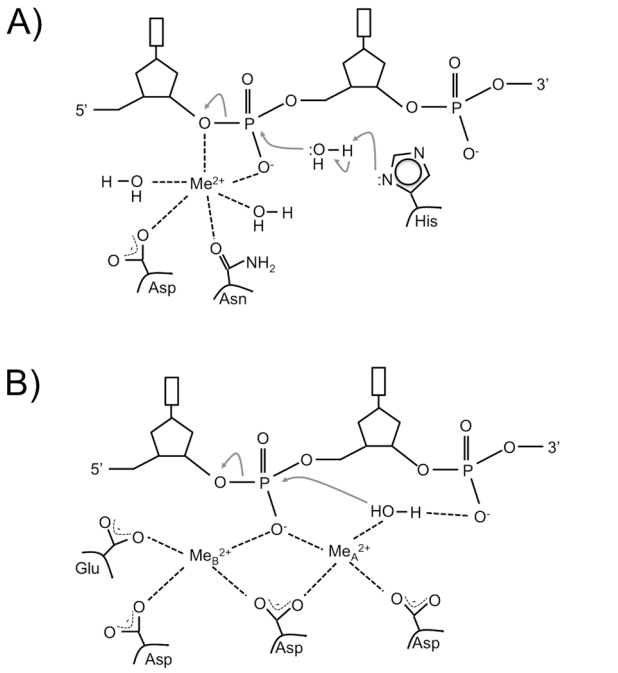

Figure 8. Nuclease catalytic mechanism. A) HNH one metal catalytic center. B) RuvC two metal catalytic center. Arrows indicate the inline attack of the activated water on the phosphorus atom of the phosphate.

and cuts, three nucleotides away from the PAM [14]. Cpf1 cleaves the non-target strand further upstream than the target strand, resulting in staggered DNA ends, with four to five nucleotide 5'-overhangs [40].

For Cas9, the involvement of HNH and RuvC endonuclease domains in DNA cleavage is certain. For Cpf1, the involvement of the RuvC domain in DNA cleavage is also undoubted, but the suggested direct catalytic activity of the Nuc domain is still tentative and not fully understood. Therefore, we will focus here only on the catalytic activity of the RuvC and HNH domains. Both catalyze phosphodiester hydrolysis reactions that lead to 5'-phosphates and free 3'-OH groups. For catalysis, a water molecule has to positioned and activated for inline attack on the phosphorus atom of the phosphate. Moreover, the leaving group (the future 3'-OH) has to be activated.

The most conserved function in DNA nuclease catalysis is the activation of the leaving group [46]. In RuvC (or RNAseH) and HNH nucleases it occurs by direct contact of a divalent metal cation with the leaving group oxygen atom (effectively "acidifying" this group). Positioning and activation of the water (or hydroxide) nucleophile is achieved in more diverse manner. In RuvC type nucleases, which use a two metal ion mechanism [46], the activation and positioning of the nucleophile involves a second divalent metal cation, which is distinct from the one that promotes the departure of the leaving group. In contrast, HNH endonucleases use amino acid based catalysis for this step [47] (Fig. 8).

TARGET STRAND CLEAVAGE

Biochemical experiments indicate that the target strand in the RNA-DNA duplex is cleaved by the HNH domain of

Cas9 [14]. The term HNH for this group of nucleases highlights a pattern of conserved residues in then known members of this protein family. In the meantime, the family has grown, and it has become clear that only the first "H" of the HNH designation is indeed a conserved histidine, whereas the "N" in the HNH designation is frequently not conserved, and the second "H" of the HNH designation may be replaced by an asparagine. The HNH group of nucleases is therefore better described by the alternative designation ββα-Me (where Me stands for metal cation) nuclease family, which refers to a conserved structural motif that is indeed present in all family members. The two β-strands run antiparallel to each other and frequently connected directly to form a hairpin. The α-helix follows immediately after the second β-strand. A histidine, anchored in the first β-strand (the first H in the HNH designation) serves as the general base to activate the nucleophilic water. An aspartate residue directly upstream and a histidine or asparagine residue in the α-helix anchor the divalent cation, typically Mg^{2+} [47] (Fig. 8A).

The ββα motif is present in SpCas9. Sequence comparisons suggest Asp839 and His840 (in the "first" β-strand) and Asn863 (in the α-helix) as key residues in the active site. Mutations of His840 or Asn863 to alanine convert SpCas9 into a nickase, supporting a role of these residues in catalysis [28]. However, a puzzle remains: in multiple crystal structures of SpCas9, Asn863 points away, not towards the predicted location of the bound metal ion, and in this conformation clearly cannot take part in anchoring a metal ion. Moreover, an arginine residue (Arg864) points towards the predicted metal ion location, so that the guanidino group comes close to the location of the metal ion, suggesting that it may functionally replace the metal ion. Alternatively, and perhaps more probably, the "unproductive" orientation of Asn863 may be due to limitations of the crystallographic experiments. First, inactivating variants of Cas9 were used in most crystallization experiments. Second, the HNH domain in most structures is distant from the scissile phosphodiester bond that it should cleave, and it is not uncommon for nuclease active sites to be disordered in the absence of substrate. Third, Mg^{2+} ions, which must have been present in the expression system, may have been lost during purification, leading to a loss of the catalytically competent state. Finally, the limited resolution of the crystallographic data, which is typical for such large structures, makes detailed interpretation of the electron density difficult.

The knowledge (rather than structure) based assignment of active site residues is also supported by conservation (see Fig. 6 of ref [31]]). The equivalent residues to Asp839 and His840 and Asn863 in SpCas9 are Asp556, His557 and Asn580 in SaCas9, Asp581, His582, and Asn606 in AnCas9. In the SaCas9 structure, the Asn580 has been mutated, but AnCas9, Asn606 is present, and points in the expected direction, also suggesting a canonical HNH-type mechanism for Cas9 nucleases.

For Cpf1 proteins, there is currently some uncertainty about the cleavage of the target DNA strand. It may be cleaved either by the RuvC domain, as suggested by the biochemical data [40], or it may be cleaved by the Nuc domain, as suggested by the AsCpf1 crystal structure [34]. As the Nuc domain is only close, but not directly bound to substrate DNA-RNA hybrid, a catalytic mechanism cannot be inferred from the crystal structure. As the Nuc domain is not similar to better characterized nucleases, the mechanism cannot be inferred by homology either, and remains to be elucidated.

NON-TARGET STRAND CLEAVAGE

Biochemical experiments indicate that the non-target strand of DNA is cleaved by the RuvC domain of Cas9 and Cpf1 [14,40]. The RuvC domains belong to a group of nucleases that includes RuvC Holliday junction resolvases, RNaseH, RNaseHII, HIV reverse transcriptase, HIV integrase, some transposases, as well as Argonaute proteins [48]. All of these enzymes are believed to operate by a two metal ion mechanism [46] (Fig. 8B). Both metal ions are believed to help stabilize the negative charge in the transition state (by direct interactions with one of the non-bridging phosphate oxygen atoms). One divalent metal cation (often termed metal cation A) is also involved in positioning and activating the nucleophile, the other metal cation (often termed metal cation B) in promoting the departure of the leaving group. Conserved aspartic and glutamic acids anchor the metal cations. A strictly conserved aspartate residue, typically the most conserved of the conserved catalytic residues, is a direct ligand to both metal ions. In addition, metal ions A and B are contacted by one or two acidic residues, or in some cases also histidine residues.

The putative catalytic residues in the RuvC domain of SpCas9 are Asp10, Glu762, His983, and Asp986. This set is highly reminiscent of the set of catalytic residues in *T. thermophilus* RuvC (Asp7, Glu70, His143 and Asp146). The role of catalytic residues in the *T. thermophilus* is itself somewhat uncertain (due to lack of metal ions and the limited resolution of this structure), but the likely scenario is that SpCas9 Asp10 Cas9 is the acidic residue in direct contact to both metals A and B, Glu762 anchors metal B, and His983 and Asp896 anchor metal A. Unfortunately, none of the Cas9 structures has two metal ions and sufficient resolution to confirm this prediction directly. The FnCas9, solved at a resolution of 1.7 Å, has one metal ion (a Ca^{2+} ion) in the active site, presumably occupying the position of metal A, and a conserved set of active site residues (Asp11, Glu898, His1162, and Asp1165). In SaCas9, the active site residues are conserved (Asp10, Glu477, His701 and Asp704), but no metal is present.

The putative catalytic residues in the RuvC domain of AsCpf1 are Asp908, Glu993 and Asp1263. Based on the order in the amino acid sequence, it is likely that Asp908 contacts both metals A and B, Asp908 metal B, and Glu993 metal A

CONFORMATIONAL FLEXIBILITY

As multi-domain proteins, Cas9s have substantial conformational flexibility, as initially demonstrated by cryo-electron microscopy reconstructions of the complex. The data indicated substantial rearrangements between the two major lobes of Cas9 between the states of the protein in isolation,

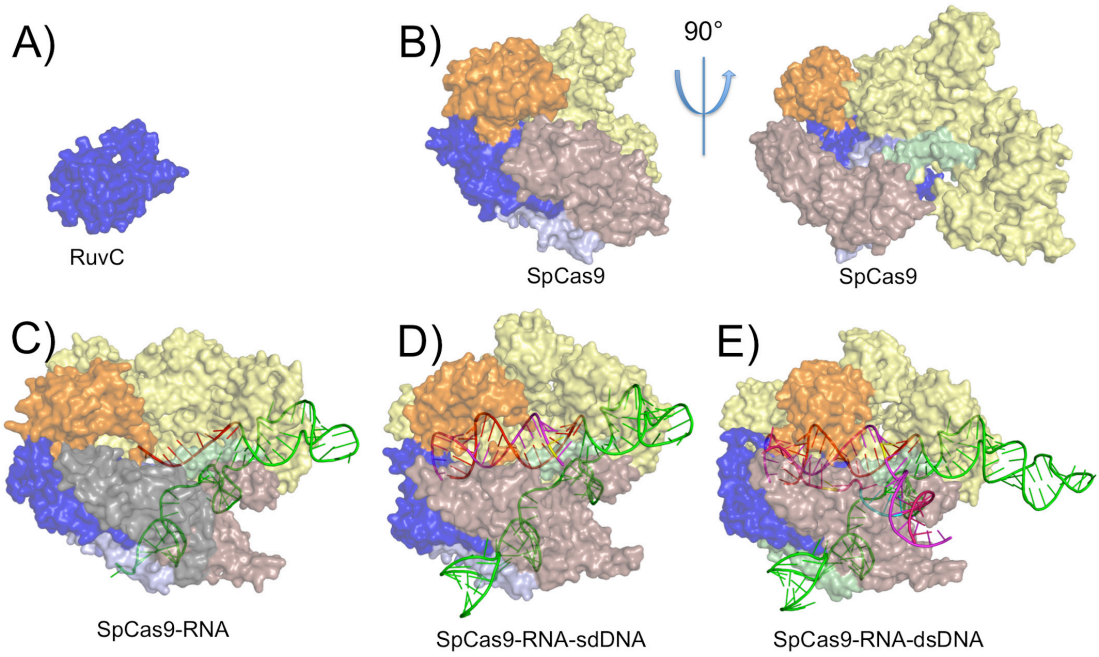

Figure 9. Domain rearrangements of Cas9 proteins upon nucleic acid binding. A) RuvC structure for size and orientation reference (PDB ID 4LD0). All Cas9 structures were superimpose on the RuvC fold in COOT using the SSM protocol. B) Apo SpCas9 (PDB ID4CMP). C) SpCas9-RNA (PDB ID 4ZT0). D) SpCas9-RNA-ssDNA (PDB ID 4OO8). E) SpCas9-RNA-dsDNA (PDB ID 5F9R). Figure prepared in PyMol.

the protein bound to RNA only, and the protein bound to RNA with DNA substrate (presumably forming the R-loop) [29]. Substantial rearrangements were then also demonstrated by comparing the structure of Cas9 bound to sgRNA only ("pre-target") and bound to sgRNA and DNA ("target bound") (Fig. 9). In particular, parts of the REC lobe and the HNH domain were found to change their position drastically between these states [32]. However, it was also clear that even the target bound conformation could not be the productive one, because the active site of the HNH domain, which should cleave the target DNA strand in the heteroduplex was found to be far away from the scissile phosphodiester bond [28,32]. The first Cas9 structure with HNH domain poised to cleave the DNA target strand was obtained when a single stranded portion of the non-target DNA strand was also included in the crystallization mix [26]. The presence of the single stranded region of the non-target strand has very little influence on the arrangement of the nucleic acids, but it induces a drastic reorientation of the HNH domain. As a result, the HNH domain active site comes close the scissile phosphodiester bond that should be cleaved according to the biochemical experiments. Moreover, the non-target DNA single strand also extends towards the RuvC active site, with the scissile phosphodiester bond coming close to it. For Cpf1, a crRNA complex of LbCpf1 and a crRNA DNA complex of AsCpf1 have been trapped by crystallographic methods, but a series of crystal structures of one and the same protein in different functional states is not yet available. Electron microscopy data for LbCpf1 suggest that his Cpf1 proteins may be equally flexible as Cas9 proteins [33].

APPLICATIONS OF CRISPR EFFECTOR NUCLEASES

Prior to the advent of CRISPR based genome engineering, the field was dominated by meganucleases [49], zinc finger nucleases (ZFNs) [50], and transcription activator *like* nucleases (TALENs) [51]. In all cases, a separate protein had to be engineered and validated for every target site. In contrast, Cas9 and Cpf1 proteins have the advantage that a generic endonuclease together with an easily produced RNA could target almost any genomic locus of choice (restricted only by the requirement for a PAM). Breakthroughs on the way towards genome engineering applications of Cas9 were the demonstration that the Cas9 system could work in a heterologous host [52], the discovery of tracrRNA [35,53], and its combination with crRNA into sgRNA [36]. Most importantly, the Cas9 endonucleases turned out to operate also in the context of eukaryotic chromatin [54], even though they do not encounter chromatin in physiological conditions and even though Cas9 activity is inhibited by nucleosomes in vitro [55]. Variants of Cas9 are now available that have improved specificity [56-58], generate a nick instead of a double strand break [59], or lack activity altogether for transcriptional control [60,61]. In cell culture, CRISPR guide RNA libraries are now available that make it possible to carry out genome wide loss of function screens, which were previously only possible using RNA interference [62,63]. CRISPR/Cas9 has also proven to be useful for the generation of transgenic animals [64-66] and plants [67]. Because Cas9 or Cpf1 based targeted endonucleases are relatively cheap and easy to make, the availability of these tools has "democratized" genetic engineering and is likely to cause a revolution in the biosciences, if it has not already done so.

REFERENCES

1. Dy RL, Richter C, Salmond GP, Fineran PC (2014) Remarkable Mechanisms in Microbes to Resist Phage Infections. Annu Rev Virol 1: 307-331
2. Wilson GG, Murray NE (1991) Restriction and modification systems. Annu Rev Genet 25: 585-627

3. Lacks SA, Mannarelli BM, Springhorn SS, Greenberg B (1986) Genetic basis of the complementary DpnI and DpnII restriction systems of *S. pneumoniae*: an intercellular cassette mechanism. Cell 46: 993-1000
4. van der Oost J, Jore MM, Westra ER, Lundgren M, Brouns SJ (2009) CRISPR-based adaptive and heritable immunity in prokaryotes. Trends Biochem Sci 34: 401-407
5. Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Almendros C (2009) Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology 155: 733-740
6. Wiedenheft B, Sternberg SH, Doudna JA (2012) RNA-guided genetic silencing systems in bacteria and archaea. Nature 482: 331-338
7. van der Oost J, Westra ER, Jackson RN, Wiedenheft B (2014) Unravelling the structural and mechanistic basis of CRISPR-Cas systems. Nat Rev Microbiol 12: 479-492
8. Jansen R, Embden JD, Gaastra W, Schouls LM (2002) Identification of genes that are associated with DNA repeats in prokaryotes. Mol Microbiol 43: 1565-1575
9. Haft DH, Selengut J, Mongodin EF, Nelson KE (2005) A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. PLoS Comput Biol 1: e60
10. Marraffini LA, Sontheimer EJ (2010) CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nat Rev Genet 11: 181-190
11. Gophna U, Kristensen DM, Wolf YI, Popa O, Drevet C, Koonin EV (2015) No evidence of inhibition of horizontal gene transfer by CRISPR-Cas on evolutionary timescales. Isme J 9: 2021-2027
12. Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, Barrangou R, Brouns SJ, Charpentier E, Haft DH, Horvath P, Moineau S, Mojica FJ, Terns RM, Terns MP, White MF, Yakunin AF, Garrett RA, van der Oost J, Backofen R, Koonin EV (2015) An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol 13: 722-736
13. Shmakov S, Abudayyeh OO, Makarova KS, Wolf YI, Gootenberg JS, Semenova E, Minakhin L, Joung J, Konermann S, Severinov K, Zhang F, Koonin EV (2015) Discovery and functional characterization of diverse class 2 CRISPR-Cas systems. Mol Cell 60: 385-397
14. Gasiunas G, Barrangou R, Horvath P, Siksnys V (2012) Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc Natl Acad Sci USA 109: E2579-2586
15. Fonfara I, Richter H, Bratovic M, Le Rhun A, Charpentier E (2016) The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. Nature 532: 517-521
16. Sinkunas T, Gasiunas G, Fremaux C, Barrangou R, Horvath P, Siksnys V (2011) Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. EMBO J 30: 1335-1342
17. Sinkunas T, Gasiunas G, Waghmare SP, Dickman MJ, Barrangou R, Horvath P, Siksnys V (2013) *In vitro* reconstitution of Cascade-mediated CRISPR immunity in *Streptococcus thermophilus*. EMBO J 32: 385-394
18. Staals RH, Zhu Y, Taylor DW, Kornfeld JE, Sharma K, Barendregt A, Koehorst JJ, Vlot M, Neupane N, Varossieau K, Sakamoto K, Suzuki T, Dohmae N, Yokoyama S, Schaap PJ, Urlaub H, Heck AJ, Nogales E, Doudna JA, Shinkai A, van der Oost J (2014) RNA targeting by the type III-A CRISPR-Cas Csm complex of *Thermus thermophilus*. Mol Cell 56: 518-530
19. Tamulaitis G, Kazlauskiene M, Manakova E, Venclovas C, Nwokeoji AO, Dickman MJ, Horvath P, Siksnys V (2014) Programmable RNA shredding by the type III-A CRISPR-Cas system of *Streptococcus thermophilus*. Mol Cell 56: 506-517
20. Hale CR, Zhao P, Olson S, Duff MO, Graveley BR, Wells L, Terns RM, Terns MP (2009) RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. Cell 139: 945-956
21. Estrella MA, Kuo F-T, Bailey S (2016) RNA-activated DNA cleavage by the Type III-B CRISPR-Cas effector complex. Genes Dev 30: 460-470
22. Samai P, Pyenson N, Jiang W, Goldberg GW, Hatoum-Aslan A, Marraffini LA (2015) Co-transcriptional DNA and RNA cleavage during type III CRISPR-Cas immunity. Cell 161: 1164-1174
23. Deng L, Garrett RA, Shah SA, Peng X, She Q (2013) A novel interference mechanism by a type IIIB CRISPR-Cmr module in Sulfolobus. Molecular microbiology 87: 1088-1099
24. Anders C, Bargsten K, Jinek M (2016) Structural plasticity of PAM recognition by engineered variants of the RNA-guided endonuclease Cas9. Mol Cell 61: 895-902
25. Hirano S, Nishimasu H, Ishitani R, Nureki O (2016) Structural basis for the altered PAM specificities of engineered CRISPR-Cas9. Mol Cell 61: 886-894
26. Jiang F, Taylor DW, Chen JS, Kornfeld JE, Zhou K, Thompson AJ, Nogales E, Doudna JA (2016) Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. Science 351: 867-871
27. Anders C, Niewoehner O, Duerst A, Jinek M (2014) Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. Nature 513: 569-573
28. Nishimasu H, Ran FA, Hsu PD, Konermann S, Shehata SI, Dohmae N, Ishitani R, Zhang F, Nureki O (2014) Crystal structure of Cas9 in complex with guide RNA and target DNA. Cell 156: 935-949
29. Jinek M, Jiang F, Taylor DW, Sternberg SH, Kaya E, Ma E, Anders C, Hauer M, Zhou K, Lin S, Kaplan M, Iavarone AT, Charpentier E, Nogales E, Doudna JA (2014) Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. Science 343: 1247997
30. Hirano H, Gootenberg JS, Horii T, Abudayyeh OO, Kimura M, Hsu PD, Nakane T, Ishitani R, Hatada I, Zhang F, Nishimasu H, Nureki O (2016) Structure and engineering of *Francisella novicida* Cas9. Cell 164: 950-961
31. Nishimasu H, Cong L, Yan WX, Ran FA, Zetsche B, Li Y, Kurabayashi A, Ishitani R, Zhang F, Nureki O (2015) Crystal structure of *Staphylococcus aureus* Cas9. Cell 162: 1113-1126
32. Jiang F, Zhou K, Ma L, Gressel S, Doudna JA (2015) A Cas9-guide RNA complex preorganized for target DNA recognition. Science 348: 1477-1481
33. Dong D, Ren K, Qiu X, Zheng J, Guo M, Guan X, Liu H, Li N, Zhang B, Yang D, Ma C, Wang S, Wu D, Ma Y, Fan S, Wang J, Gao N, Huang Z (2016) The crystal structure of Cpf1 in complex with CRISPR RNA. Nature 532: 522-526
34. Yamano T, Nishimasu H, Zetsche B, Hirano H, Slaymaker IM, Li Y, Fedorova I, Nakane T, Makarova KS, Koonin EV, Ishitani R, Zhang F, Nureki O (2016) Crystal structure of Cpf1 in complex with guide RNA and Target DNA. DNA. Cell 165: 949-962
35. Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, Eckert MR, Vogel J, Charpentier E (2011) CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature 471: 602-607
36. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337: 816-821
37. Fonfara I, Le Rhun A, Chylinski K, Makarova KS, Lecrivain AL, Bzdrenga J, Koonin EV, Charpentier E (2014) Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. Nucleic Acids Res 42: 2577-2590
38. Darmon E, Leach DR (2014) Bacterial genome instability. Microbiol Mol Biol Rev 78: 1-39
39. Westra ER, Brouns SJ (2012) The rise and fall of CRISPRs--dynamics of spacer acquisition and loss. Mol Microbiol 85: 1021-1025
40. Zetsche B, Gootenberg JS, Abudayyeh OO, Slaymaker IM, Makarova KS, Essletzbichler P, Volz SE, Joung J, van der Oost J, Regev A, Koonin EV, Zhang F (2015) Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell 163: 759-771
41. Makarova KS, Koonin EV (2015) Annotation and Classification of CRISPR-Cas Systems. Methods Mol Biol 1311: 47-75
42. Sternberg SH, Redding S, Jinek M, Greene EC, Doudna JA (2014) DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature 507: 62-67

43. Briner AE, Barrangou R (2014) Lactobacillus buchneri genotyping on the basis of clustered regularly interspaced short palindromic repeat (CRISPR) locus diversity. Appl Environ Microbiol 80: 994-1001
44. Kleinstiver BP, Prew MS, Tsai SQ, Topkar VV, Nguyen NT, Zheng Z, Gonzales APW, Li Z, Peterson RT, Yeh J-RJ, Aryee MJ, Joung JK (2015) Engineered CRISPR-Cas9 nucleases with altered PAM specificities. Nature 523: 481-485
45. Lesnik EA, Freier SM (1995) Relative thermodynamic stability of DNA, RNA, and DNA:RNA hybrid duplexes: relationship with base composition and structure. Biochemistry 34: 10807-10815
46. Yang W (2008) An equivalent metal ion in one- and two-metal-ion catalysis. Nat Struct Mol Biol 15: 1228-1231
47. Sokolowska M, Czapinska H, Bochtler M (2009) Crystal structure of the beta beta alpha-Me type II restriction endonuclease Hpy99I with target DNA. Nucleic Acids Res 37: 3799-3810
48. Nowotny M, Gaidamakov SA, Crouch RJ, Yang W (2005) Crystal structures of RNase H bound to an RNA/DNA hybrid: substrate specificity and metal-dependent catalysis. Cell 121: 1005-1016
49. Stoddard BL (2011) Homing endonucleases: from microbial genetic invaders to reagents for targeted DNA modification. Structure 19: 7-15
50. Urnov FD, Rebar EJ, Holmes MC, Zhang HS, Gregory PD (2010) Genome editing with engineered zinc finger nucleases. Nature Rev Genetics 11: 636-646
51. Boch J (2011) TALEs of genome targeting. Nature Biotechnol 29: 135-136
52. Sapranauskas R, Gasiunas G, Fremaux C, Barrangou R, Horvath P, Siksnys V (2011) The *Streptococcus thermophilus* CRISPR/Cas system provides immunity in *Escherichia coli*. Nucleic Acids Res 39: 9275-9282
53. Karvelis T, Gasiunas G, Miksys A, Barrangou R, Horvath P, Siksnys V (2013) crRNA and tracrRNA guide Cas9-mediated DNA interference in Streptococcus thermophilus. RNA Biol 10: 841-851
54. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, Zhang F (2013) Multiplex genome engineering using CRISPR/Cas systems. Science 339: 819-823
55. Hinz JM, Laughery MF, Wyrick JJ (2015) Nucleosomes inhibit Cas9 endonuclease activity *in vitro*. Biochemistry 54: 7063-7066
56. Kleinstiver BP, Pattanayak V, Prew MS, Tsai SQ, Nguyen NT, Zheng Z, Joung JK (2016) High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. Nature 529: 490-495
57. Ran FA, Hsu PD, Lin C-Y, Gootenberg JS, Konermann S, Trevino AE, Scott DA, Inoue A, Matoba S, Zhang Y, Zhang F (2013) Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. Cell 154: 1380-1389
58. Guilinger JP, Thompson DB, Liu DR (2014) Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. Nature Biotechnol 32: 577-582
59. Shen B, Zhang W, Zhang J, Zhou J, Wang J, Chen L, Wang L, Hodgkins A, Iyer V, Huang X, Skarnes WC (2014) Efficient genome modification by CRISPR-Cas9 nickase with minimal off-target effects. Nat Methods 11: 399-402
60. Konermann S, Brigham MD, Trevino AE, Joung J, Abudayyeh OO, Barcena C, Hsu PD, Habib N, Gootenberg JS, Nishimasu H, Nureki O, Zhang F (2015) Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature 517: 583-588
61. Mali P, Aach J, Stranges PB, Esvelt KM, Moosburner M, Kosuri S, Yang L, Church GM (2013) CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. Nat Biotechnol 31: 833-838
62. Wang T, Wei JJ, Sabatini DM, Lander ES (2014) Genetic screens in human cells using the CRISPR-Cas9 system. Science 343: 80-84
63. Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelsen TS, Heckl D, Ebert BL, Root DE, Doench JG, Zhang F (2014) Genome-scale CRISPR-Cas9 knockout screening in human cells. Science 343: 84-87
64. Port F, Chen HM, Lee T, Bullock SL (2014) Optimized CRISPR/Cas tools for efficient germline and somatic genome engineering in *Drosophila*. Proc Natl Acad Sci USA 111: E2967-2976
65. Jao LE, Wente SR, Chen W (2013) Efficient multiplex biallelic zebrafish genome editing using a CRISPR nuclease system. Proc Natl Acad Sci USA 110: 13904-13909
66. Wang H, Yang H, Shivalila CS, Dawlaty MM, Cheng AW, Zhang F, Jaenisch R (2013) One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. Cell 153: 910-918
67. Bortesi L, Fischer R (2015) The CRISPR/Cas9 system for plant genome editing and beyond. Biotechnology advances 33: 41-52
68. Jiang F, Doudna JA (2015) The structural biology of CRISPR-Cas systems. Curr Opin Struct Biol 30: 100-111

Nukleazy CRISPR typu II i V z punktu widzenia biologa strukturalnego

Humberto Fernandes[1], Michal Pastor[1], Matthias Bochtler[1,2,✉]

[1]Institute of Biochemistry and Biophysics PAS, Warsaw, Poland
[2]International Institute of Molecular and Cell Biology, Warsaw, Poland

✉e-mail: mbochtler@ibb.waw.pl,

Słowa kluczowe: CRISPR; Biologia strukturalna; Cas9; Cpf1; domena RuvC; domena HNH

STRESZCZENIE

Nukleazy efektorowe Cas9 i Cpf1, będące częścią systemu CRISPR typu II i V, są „uniwersalnymi" endonukleazami DNA, które za pomocą odpowiedniej sekwencji crRNA lub sgRNA mogą zostać zaprogramowane do przecięcia niemal każdego dwuniciowego DNA w określonym miejscu (oflankowanym krótkimi sekwencjami PAM). W niniejszym artykule przeglądowym omówiono bakteryjny system odporności swoistej CRISPR jako naturalny kontekst, w którym działają nukleazy Cas9 i Cpf1, i przedstawiono informacje strukturalne dotyczące działania tych białek uzyskane w ciągu ostatnich 2-3 lat. Opisano także strukturę „pętli R" zlokalizowanych w rdzeniu kompleksów Cas9 i Cpf1 oraz „uchwytów" 5′ i 3′ biorących udział w kotwiczeniu kompleksów kwasów nukleinowych na białkach w sposób niezależny od sekwencji docelowej. W artykule omówiono również budowę molekularną białek Cas9 i Cpf1, mechanizm skanowania przez nie dwuniciowego DNA w poszukiwaniu sekwencji PAM (ang. *protospacer associated motifs*), a także sposób rozdzielenia docelowej i niedocelowej nici DNA z udziałem pętli fosforanowej (PLL, *phosphate loop*) i helisy podstawowej (BH, *basic helix*) oraz formowanie się hybryd złożonych z crRNA lub sgRNA i docelowej nici DNA. Ponadto, opisano przyjęty obecnie mechanizm działania domen katalitycznych RuvC i HNH oraz możliwą, a jednocześnie wciąż bardzo niepewną, katalityczną rolę domeny Nuc. W końcowej części niniejszego artykułu przeglądowego znajduje się natomiast krótkie podsumowanie kluczowych osiągnięć, które zapoczątkowały zastosowanie nukleaz Cas9 i Cpf1 w inżynierii genomowej.