Cutting it close: CRISPR-associated endoribonuclease structure and function

Megan L. Hochstrasser[1] **and Jennifer A. Doudna**[1,2,3,4,5]

[1] Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA
[2] Howard Hughes Medical Institute, University of California, Berkeley, CA 94720, USA
[3] California Institute for Quantitative Biosciences, University of California, Berkeley, CA 94720, USA
[4] Department of Chemistry, University of California, Berkeley, CA 94720, USA
[5] Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

Many bacteria and archaea possess an adaptive immune system consisting of repetitive genetic elements known as clustered regularly interspaced short palindromic repeats (CRISPRs) and CRISPR-associated (Cas) proteins. Similar to RNAi pathways in eukaryotes, CRISPR–Cas systems require small RNAs for sequence-specific detection and degradation of complementary nucleic acids. Cas5 and Cas6 enzymes have evolved to specifically recognize and process CRISPR-derived transcripts into functional small RNAs used as guides by interference complexes. Our detailed understanding of these proteins has led to the development of several useful Cas6-based biotechnological methods. Here, we review the structures, functions, mechanisms, and applications of the enzymes responsible for CRISPR RNA (crRNA) processing, highlighting a fascinating family of endonucleases with exquisite RNA recognition and cleavage activities.

CRISPR–Cas systems and crRNA biogenesis

Most prokaryotes employ an RNA-based adaptive immune system known as CRISPR–Cas to identify and eliminate genetic parasites (reviewed in [1–5]). Upon detecting viral or plasmid DNA in the cell, bacteria and archaea with active CRISPR systems respond by integrating short fragments of foreign DNA into the host chromosome at one end of the CRISPR locus (Figure 1) [6–8]. Such loci serve as molecular vaccination cards by maintaining a genetic record of prior encounters with foreign transgressors. The defining feature of CRISPR loci is a series of direct repeats (~20–50 bp) separated by unique spacer sequences of a similar length (Figure 1) [9–11]. Following transcription, CRISPR sequences are processed into short CRISPR-derived RNAs (crRNAs) [12–22]. An interference complex of CRISPR-associated (Cas) proteins uses the mature crRNA as a guide to target and destroy foreign nucleic acids bearing sequence complementarity [7,12,23–25]. In addition to providing adaptive immunity, the CRISPR pathway can also play a role in endogenous gene regulation [26].

CRISPR loci are flanked by a diverse set of *cas* genes that define three major CRISPR types (Types I–III) based on gene conservation and locus organization [27]. The *cas6* gene family encodes a set of RNA endonucleases responsible for crRNA processing in Type I and Type III CRISPR systems [13]. Type II systems use a trans-activating RNA (tracrRNA) together with endogenous RNase III for crRNA maturation [15] (Figure 1), a process that is not covered in this review. The first Cas protein to be crystallized was a Cas6 enzyme from *Thermus thermophilus* [28]. The initial functional characterization of Cas6 was performed in *Pyrococcus furiosus*, where it acts as a CRISPR-specific endoribonuclease (endoRNase) that binds and cleaves within each repeat sequence of the precursor crRNA (pre-crRNA) transcript, generating a library of crRNAs wherein each contains a unique spacer sequence flanked by portions of the adjacent repeats [13].

Further work revealed the basic characteristics of Cas6 endoRNases. Despite minimal sequence homology, Cas6s have several conserved structural features that facilitate binding of both the pre-crRNA and their crRNA product with high affinity [13,16,29,30]. In most CRISPR systems, due to the pseudo-palindromic nature of the repeat sequence, the pre-crRNA adopts a stem loop structure that is bound sequence- and shape-specifically and cleaved at its base [13,14]. Some pre-crRNAs are predicted to be unstructured in solution and thus may be bound differently, although base pairing may be stabilized by protein interactions [10,31,32]. In Type I-B, I-C, I-E, and I-F systems, the endoRNase stays bound to the crRNA and assembles into a complex with other Cas proteins for downstream targeting [12,22,33,34], while in Type I-A and III systems, the crRNA alone is loaded into the targeting complex and Cas6 dissociates [18,35–39] (Figure 1). The Type I interference complex is known as Cascade (CRISPR-associated complex for antiviral defense), and as recently proposed by van der Oost and colleagues [5], the CRISPR subtype from which it derives is denoted with a slash (e.g., the Cascade

Corresponding author: Doudna, J.A. (doudna@berkeley.edu, doudna@uclink.berkeley.edu).
Keywords: CRISPR–Cas; crRNA; Cas5; Cas6; endoribonuclease; structure and function.

0968-0004/

http://dx.doi.org/10.1016/j.tibs.2014.10.007

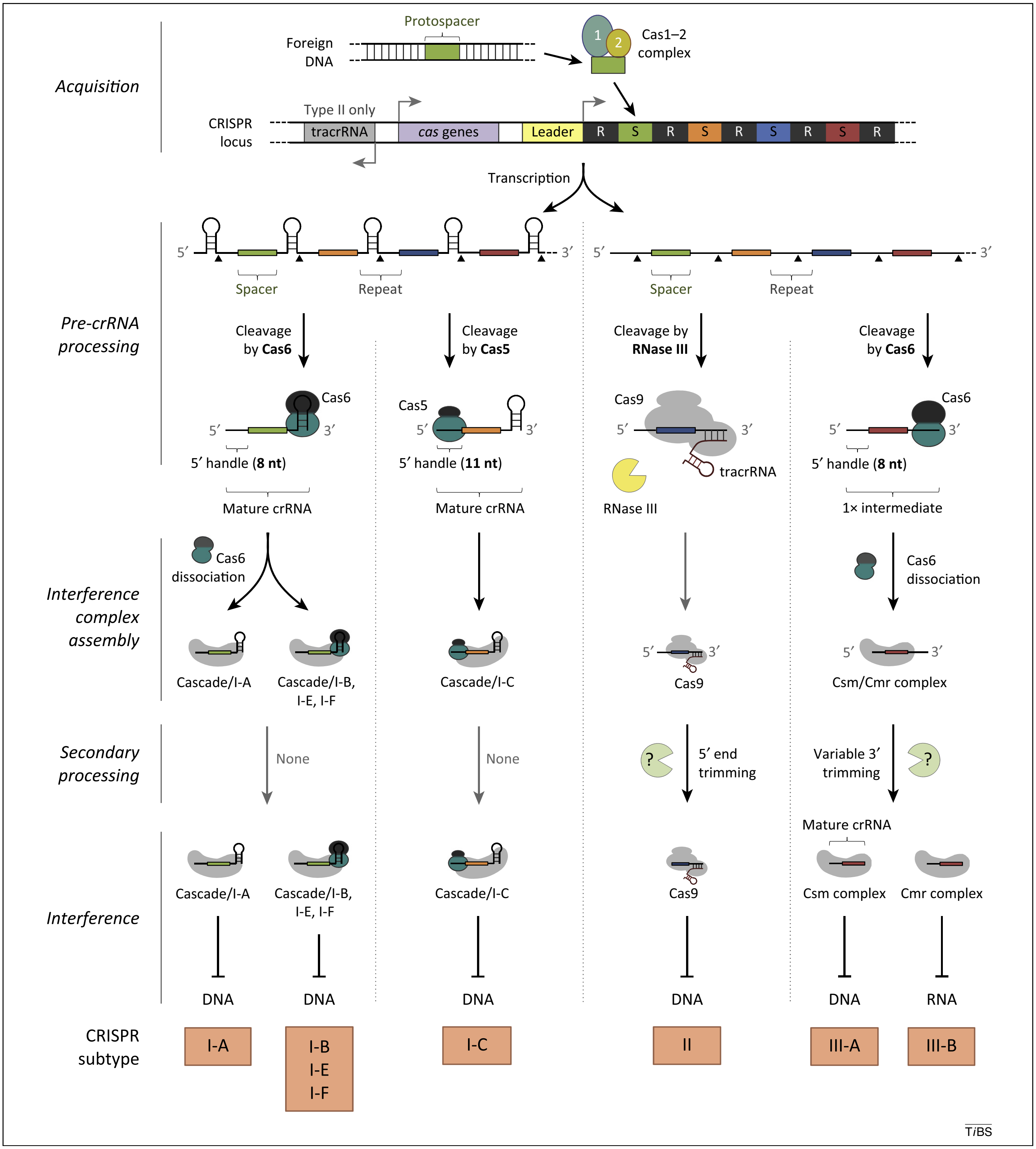

Figure 1. Overview of CRISPR RNA (crRNA) processing and comparison between CRISPR–Cas interference systems. There are three main pathways of CRISPR adaptive immunity (Types I–III) and several subtypes, each typified by a different set of Cas proteins. The first stage of the CRISPR–Cas system is acquisition, in which a foreign DNA sequence is incorporated into the host CRISPR locus. Next, the entire repeat-spacer array is transcribed into a long precursor crRNA (pre-crRNA). A single cleavage within each repeat sequence generates shorter, mature crRNAs. Some crRNAs undergo an additional trimming step. The enzymes responsible for catalysis and exact mode of crRNA processing differ in each system. The crRNA is loaded into an interference complex where it serves as a guide for targeting invasive DNA, or in Type III-B systems, RNA. N- and C-terminal RRM folds are colored teal and gray, respectively, as in subsequent figures. Abbreviations: CRISPR, clustered regularly interspaced short palindromic repeat; Cas, CRISPR-associated; R, repeat; S, spacer; RRM, RNA recognition motif.

Table 1. Summary of the most extensively characterized CRISPR endoribonucleases

Subtype	Name in this review	Organism(s)	Other name(s)	PDB ID(s)[a]	Stoichiometry	Turnover	Refs
I-A	PhoCas6nc	*Pyrococcus horikoshii*	Cas6a	*3QJJ*, *3QJL*, *3QJP*	Dimer (pre-crRNA-dependent)	Non-catalytic	[47]
	SsoCas6-1A	*Sulfolobus solfataricus*	Sso2004 (Cas6-1 family), SsCas6	*4ILL*, **4ILM**, 4ILR	Dimer	ND[b]	[18,32]
	SsoCas6-1B		Sso1437 (Cas6-1 family), SsoCas6	3ZFV	Dimer	Multiple	[51,60]
	SsoCas6-3		Sso1422 (Cas6-3)	–	ND	ND	[51]
I-B	MmaCas6b	*Methanococcus maripaludis*	Mm Cas6b	–	Dimer (pre-crRNA); monomer (crRNA)	ND	[21,55]
I-C	Cas5c	*Bacillus halodurans*, *Mannheimia succiniciproducens*, *Streptococcus pyogenes*, *Xanthomonas oryzae*	Cas5d	4F3M, 3KG4, 3VZH, 3VZI	Monomer	ND	[22,40,41]
I-E	TthCas6e	*Thermus thermophilus*	TTHB192, Cse3	1WJ9, *2Y8W*, *2Y8Y*, *2Y9H*, **3QRP**, *3QRQ*, **3QRR**	Monomer	Single	[16,17,28]
	EcoCas6e	*Escherichia coli*	CasE	4DZD (monomer), **4TVX**, **4U7U**, **4QYZ** (Cascade/I-E)	Monomer	Single	[19,48–50]
	EcoCas5e		CasD	4TVX, 4U7U, 4QYZ (Cascade/I-E)	Monomer	Non-catalytic	[48–50]
I-F	PaeCas6f	*Pseudomonas aeruginosa*	Csy4	*2XLI*, *2XLJ*, *2XLK*, **4AL5**, ***4AL6***, 4AL7	Monomer	Single	[14,29,33,57]
III-A	SepCas6	*Staphylococcus epidermidis*	–	–	ND	ND	[20,37]
III-B	PfuCas6-1	*Pyrococcus furiosus*	PfCas6	3I4H, *3PKM*	Monomer or weak dimer	Single	[13,31,56]
	PfuCas6-3nc		PfCas6-3	3UFC	Monomer	Non-catalytic	[53]
I-B?	TthCas6B	*T. thermophilus*	TTHB231	4C98, **4C9D**	Dimer	Single	[30]
Orphan	TthCas6A		TTHB78	*4C8Y*, **4C8Z**, 4C97	Dimer	Single	

[a] PDB ID key: plain text, apo; *italics*, substrate-bound; **bold**, product-bound; ***bold italics***, mix of substrate and product in crystal; underline, minimal complex (stem loop only).

[b] ND, no data.

complex from a Type I-E CRISPR system is known as Cascade/I-E). Type III-A and III-B systems use the Csm and Cmr complex, respectively (Figure 1).

Cas6 cleavage generates a mature crRNA containing a single spacer sequence with fragments of the CRISPR repeat on either side. The remaining repeat sequence at the 5′ end of the mature crRNA is known as the '5′ handle' or '5′ tag', and in Type III and most Type I systems it is 8 nucleotides (nts) in length (Figure 1) [12,13]. The leftover repeat segment at the 3′ end of the crRNA, which is more variable in length and structure, is sometimes known as the '3′ handle' or '3′ stem loop'. Intriguingly, Cas5, which serves a structural role in the interference complex of most CRISPR subtypes, has evolved to serve as the dedicated endoribonuclease of Type I-C systems, where it generates mature crRNAs with an 11-nt 5′ handle [22,40–42]. In Type III and at least one Type I-A system, further processing by an unknown trimming nuclease removes 3′ portions of the crRNA (Figure 1) [23,35,39,43,44]. Currently, very little is known about Type I-D CRISPR systems, although crRNA expression, processing into a mature species with an 8-nt 5′ handle, and possible 3′ trimming were recently demonstrated *in vivo* in *Synechocystis* sp. PCC6803 [45]. Thus, we speculate that Type I-D crRNA maturation may be most similar to that which occurs in Type III systems.

Subsequent research has defined the scope and mechanism of Cas5 and Cas6 RNA recognition and cleavage, as well as their roles in CRISPR-based immunity. Despite the growing body of work on these enzymes (summarized in Table 1), many aspects of their structure, mechanism, and biological functions remain unaddressed. Here, we review the current understanding of this unique class of endoribonucleases and highlight remaining questions that may be illuminated by future research. We explore crRNA biogenesis in Type I and III CRISPR systems through a thorough discussion of the structures, RNA binding specificity, and catalytic mechanisms of Cas5 and Cas6 proteins. Additionally, we relate these characteristics to the varying roles of CRISPR RNases in delivery of mature crRNAs to interference complexes. We conclude with a summary of the biological applications developed to exploit highly specific RNA binding and cleavage by Cas6.

Core structural elements of CRISPR endoribonucleases

Cas6 family members share surprisingly limited sequence homology. Nevertheless, the many Cas6 crystal structures

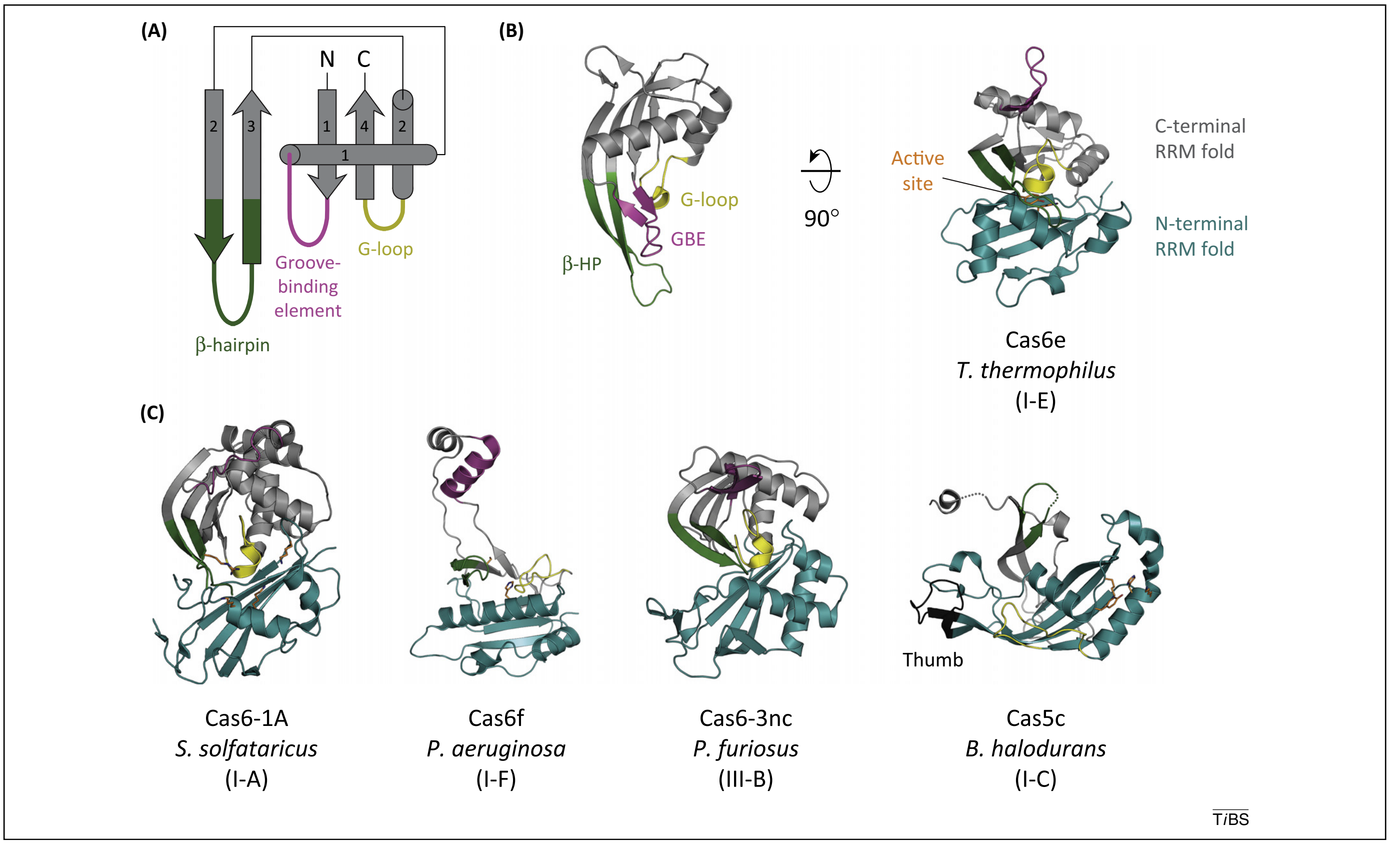

Figure 2. Fundamental structural features of CRISPR endoRNases. **(A)** Topology diagram of a typical Cas6 C-terminal RRM fold with key structural features labeled. **(B)** Two views of *Thermus thermophilus* Cas6e (PDB: 2Y8W) colored as in (A). For clarity, the N-terminal RRM fold has been omitted in the left panel. **(C)** Comparison of structures of Cas6 and Cas5c enzymes associated with different CRISPR subtypes (in parentheses), highlighting shared structural elements, colored as in (A) and (B), with the Cas5 'thumb' in black (PDB: 4ILL, 2XLK, 3UFC, 4F3M). Note that no active site residues are shown for *Pyrococcus furiosus* Cas6-3nc because this protein is non-catalytic. Abbreviations: CRISPR, clustered regularly interspaced short palindromic repeat; Cas, CRISPR-associated; endoRNases, endoribonucleases; RRM, RNA recognition motif.

(Table 1) reveal a common overall fold as well as specific structural features important for crRNA binding. Cas6 enzymes consist solely of two repeat-associated mysterious protein (RAMP) domains that form ferredoxin-like or RNA recognition motif (RRM) folds, a common feature of Cas proteins [4,9,27,46]. Each domain has a βαββαβ secondary structure arrangement, forming a four-stranded antiparallel β-sheet with the two α-helices on one side (Figure 2A). The α-helices of the two RRM folds form one face of the protein, which makes the majority of the crRNA contacts [14,16,17,30–32,47–50]. The active site is typically located between the two domains, and residues from both can participate in catalysis, as described later [13,14,16,17,28,30–32,51].

The C-terminal RRM features three additional elements that are easily identifiable in all Cas6 structures. The first, the glycine-rich loop (G-loop), is the only sequence motif conserved between Cas6 enzymes [9,52]. It typically follows the pattern GhGxxxxxGhG, where 'h' represents a hydrophobic residue and xxxxx contains at least one arginine or lysine [46]. The G-loop is located between α_2' and β_4' (prime indicates features in the C-terminal RAMP domain), and is positioned between the two RRM folds in the Cas6 tertiary structure (Figure 2A,B) [13,14,16,17,28,30–32,47–51,53]. It typically functions in binding and stabilizing the crRNA, sometimes sequence-specifically. In the case of a Cas6 protein from *T. thermophilus* (TthCas6B), it contacts the 2′ hydroxyl immediately upstream of the scissile phosphate, suggesting that this motif participates in the RNA cleavage reaction [30].

The second structural element that is present in all Cas6 structures is a β-hairpin formed by β_2' and β_3' [4]. Compared with those of the N-terminal domain, these two β-hairpin strands are significantly longer, extending far past the other strands in the β-sheet (Figure 2). This β_2'–β_3' hairpin, abbreviated β-HP, inserts into the base of the crRNA stem loop and helps position the scissile phosphate in the active site, sometimes providing catalytic residues [14,32].

The third notable Cas6 structural feature in the second RRM fold is an element that probes the crRNA major groove and often interacts sequence-specifically with these nucleotides. This groove-binding element (GBE) is generally located between β_1' and α_1' in primary sequence and varies in secondary structure (Figure 2). It forms an α-helix in the *Pseudomonas aeruginosa* Cas6 (PaeCas6f, also commonly known as Csy4) [14], a loop in Cas6 enzymes from *Sulfolobus solfataricus* and *Escherichia coli* (SsoCas6-1B and EcoCas6e, formerly known as CasE) [32,48–50], and a β-hairpin in all three Cas6 proteins from *T. thermophilus* (TthCas6e, formerly known as Cse3, TthCas6A, TthCas6B) [16,17,30]. It appears disordered in the catalytic *P. furiosus* Cas6 enzyme (PfuCas6-1) [31] but forms a β-hairpin in a non-catalytic paralog (PfuCas6-3nc) [53].

Most Cas6 enzymes fit this same basic structure, sometimes missing a few elements or possessing additional helices or β-strands. One somewhat anomalous protein is PaeCas6f, which has an intact N-terminal RRM fold but a very different C-terminal domain. While it retains a G-loop, GBE, and β-HP, the C terminus appears to have lost most characteristics of a typical RRM fold (Figure 2C) [14]. Another notable structural variation is the dimerization of some Cas6 proteins, mostly from thermophilic organisms (Table 1). Cas6s associated with the I-A and I-B CRISPR systems in *Pectobacterium atrosepticum*, *Pyrococcus horikoshii*, *S. solfataricus*, *Methanococcus maripaludis*, and *T. thermophilus* are believed to form and/or function as dimers, although the mechanistic effects of dimerization are not clear [30,32,47,51,54,55].

The Cas5 family of CRISPR endoRNases is less understood, although several structures are now available [22,40,41]. Originally known as Cas5d, we herein refer to this enzyme as Cas5c as previously suggested [4], because it is found in the Type I-C CRISPR system. Cas5 proteins from other Type I systems are non-catalytic and serve as structural subunits in interference complexes [48–50]. Interestingly, Cas5c appears to have evolved enzymatic activity in the I-C subtype, which lacks a Cas6 homolog [27]. Similar to Cas6, catalytic Cas5 enzymes have an N-terminal RRM fold, but their C-terminal domain is smaller, consisting of just three antiparallel β-strands, two of which form a β-hairpin, and disordered regions (Figure 2C). Suggestively, the Cas5c active site is not located in the same place as that of Cas6 enzymes, indicating that the catalytic centers may have evolved independently (Figure 2C). Furthermore, Cas5c proteins have a G-loop in their RAMP domain that does not match the consensus sequence of the Cas6 G-loop. Another notable feature of the N-terminal RRM is a 'thumb' between β_2 and β_3 that is highly variable in length and structure [48–50]. It forms a β-hairpin in the *Bacillus halodurans* Cas5 (BhaCas5c) [22], two α-helices in *Xanthomonas oryzae* (XorCas5c) [41], and is disordered in structures from *Mannheimia succiniciproducens* (MsuCas5) [40] and *Streptococcus pyogenes* (SpyCas5c) [41]. Interestingly, the Cas5 thumb may be structurally related to a similar feature found in the Cas7 subunit of Cascade/I-E [48–50]. The possible roles of this structural element and the contributions of other conserved motifs involved in RNA binding are discussed in the next section.

Structural basis of RNA binding and specificity

A hallmark feature of Cas6 enzymes is the high affinity and specificity with which they bind RNA. Binding affinity appears substantially lower for Cas5c enzymes, and there are currently no available structures of Cas5c proteins bound to substrate or product RNA [22]. The structural basis of sequence- and structure-specific pre-crRNA binding by Cas6 enzymes, however, has been well characterized. The RNA-binding regions of Cas6s are positively charged, allowing the proteins to make extensive ionic interactions with the phosphate backbone of the crRNA. A massive network of hydrogen bonds with 2′ hydroxyl groups, bases, and other parts of the crRNA contribute to nonspecific RNA binding as well as sequence-specific recognition [14,16,17,30–32]. Cas6e and Cas6f exhibit extremely high (picomolar) affinities for the cleavage product containing the stem loop and exhibit single-turnover kinetics as a result [16,29,30,56]. Substrate binding and cleavage cause little change in the overall structure of Cas6 enzymes (Figure 3A, compare the first two structures) [16,30,32,47]. In some cases, regions that are disordered in the apo form of the protein undergo conformational ordering upon RNA binding, particularly in the active site [16,30].

As described earlier, the G-loop, GBE, and β-HP provide critical interactions that facilitate RNA binding. All three structural features make nonspecific RNA contacts, while the number of sequence-specific contacts made by each element varies from protein to protein (Figure 3B,C). Many of the critical contacts made by Cas6 enzymes occur in the crRNA stem loop and rely upon both sequence and geometry [14,16,17,30–32,48–50]. PaeCas6f, for example, interacts specifically with multiple base pairs within the stem loop, and both binding and cleavage are severely impaired if the length of the stem is altered [14,29]. Interactions with all four base pairs are important for TthCas6A and three are bound base-specifically [30]. It also sequence-specifically recognizes 2 nts on either side of the stem loop. TthCas6e binds the 3 nts downstream of the cleavage site sequence-specifically [16,17]. Thus, Cas6 enzymes bind structured pre-crRNAs via contacts both within and directly adjacent to the stem loop.

In some Cas6 crystal structures, nucleotides at the 5′ end of the crRNA 3′ handle (upstream of the stem loop in structured crRNAs) are bound in a distal, positively-charged cleft between the two RRM folds (Figure 3A, last image) [30,31,47]. The first example of this mode of substrate recognition was observed in a Type III-B system in *P. furiosus* and led to the 'wrap-around' model, whereby PfuCas6-1 specifically recognizes the first 8 nts of the repeat, and tight binding of this segment tethers the pre-crRNA while directing the downstream scissile phosphate into the distal cleavage site [13,31]. crRNA wrapping is also seen in Cas6 proteins from *P. horikoshii* and *T. thermophilus*, where several of the nucleotides 5′ of the stem loop are recognized base-specifically, and deletions or mutations in this region lead to significant binding and cleavage defects [30,47]. As seen in several recently published crystal structures of the Cascade/I-E targeting complex from *E. coli*, the Cas6e subunit makes sequence-specific contacts with several repeat-derived nucleotides 5′ of the stem loop [48–50]. Thus, in addition to sequence- and structure-specific binding of the stem loop itself, both ionic contacts and base-specific recognition of the 5′ region of the crRNA 3′ handle are critical for some Cas6s.

Extensive interactions with the 5′ end of the 3′ handle may be most important for systems in which the CRISPR repeat sequence is non-palindromic and the crRNA is predicted to be unstructured. While most CRISPR repeats are predicted to form highly stable stem loops, some crRNAs do not appear to possess any secondary structure [10]. Remarkably, the crystal structure of the *S. solfataricus* Cas6-1 family member encoded by SSO2004 (hereafter referred to as SsoCas6-1A) revealed that this protein stabilizes an otherwise unstable stem loop

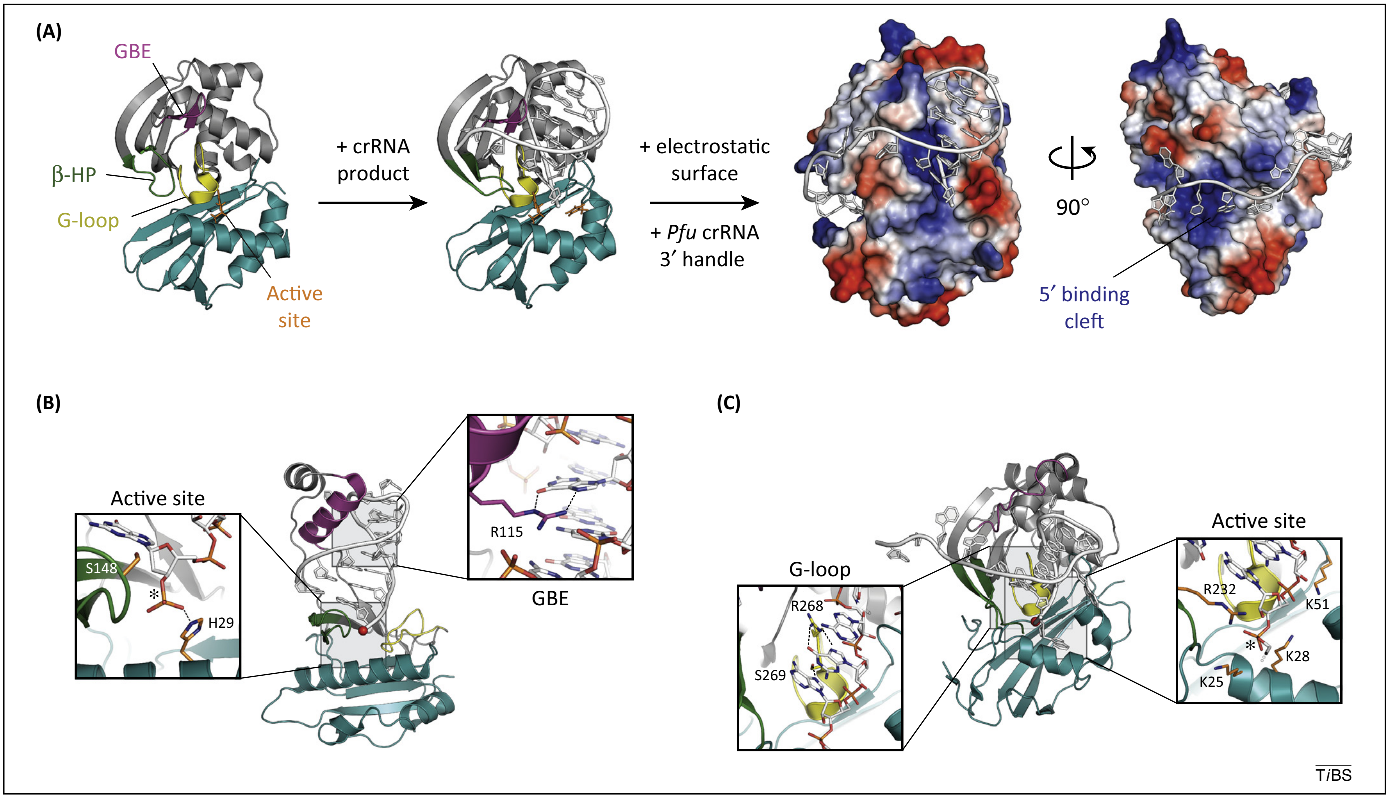

Figure 3. Structure- and sequence-specific RNA binding by Cas6 enzymes. **(A)** First two images: *Thermus thermophilus* Cas6A in the apo form and bound to its product CRISPR RNA (crRNA) (PDB: apo – 4C97, product-bound – 4C8Z). Second two images: electrostatic surface potential rendering of the same enzyme in two views with the first eight nucleotides of the *Pyrococcus furiosus* crRNA 3′ handle (PDB: 3PKM) modeled onto the structure based on alignment of the two proteins, as in Niewoehner *et al.* [30]. For simplicity, only one subunit of the non-crystallographic dimer is shown. **(B)** *Pseudomonas aeruginosa* Cas6f bound to its cognate RNA (PDB: 2XLK). Close-up views highlight the active site and sequence-specific interactions by the groove-binding element. **(C)** *Sulfolobus solfataricus* Cas6-1A bound to its pre-crRNA substrate (PDB: 4ILL). The active site- and sequence-specific contacts made by the glycine-rich loop are shown in detail. For simplicity, only one subunit of the SsoCas6-1A dimer is shown. Abbreviations: CRISPR, clustered regularly interspaced short palindromic repeat; Cas, CRISPR-associated.

structure, consisting of as little as two base pairs, just upstream of the cleavage site [32] (Figure 3C). Thus, extensive crRNA base pairing may not be necessary for recognition and cleavage by some Cas6 proteins, as the enzyme-stabilized secondary structure could be a feature of many non-palindromic repeat-derived crRNAs, which do not adopt a stable fold in solution. PfuCas6-1 and the non-catalytic *P. horikoshii* Cas6 (PhiCas6nc) are thought to recognize unstructured crRNAs, but only the first 8–12 nts are observed in these crystal structures [31,47]. In light of the SsoCas6-1A study and the potential groove-binding β-hairpin in the structure of the non-catalytic paralog PfuCas6-3nc (Figure 2C), it is possible that a short stem loop may form when these proteins bind the crRNA.

The mechanism of RNA binding by Cas5c proteins is largely unknown. In the apo Cas5c structures, it is difficult to identify which structural features may interact with the pre-crRNA, and how these interactions compare with those of Cas6 enzymes. One side of the protein (shown in Figure 2) is conserved, positively charged, and contains the active site, so it is most likely the primary RNA-binding face [22]. The C-terminal β-hairpin in Cas5c may be positioned at the base of the crRNA stem loop, as in Cas6 structures, or the RNA may be bound entirely differently.

As described earlier, the region between β_2 and β_3 is variable in length and structure. In the non-catalytic Cas5e subunit of Cascade/I-E, it forms a β-hairpin with a long loop that acts as a ‘thumb’, docking it into the complex through protein–protein interactions and, importantly, contacts with the 5′ handle [48–50]. If this mode of crRNA binding is common to all Cas5 proteins, then rather than recognizing the stem loop and upstream nucleotides as Cas6 does, interactions with the nucleotides downstream of the stem loop and cut site may be most critical for this family of endonucleases (Figure 1) [48]. In keeping with this idea, Cas5c binding and cleavage activity are sensitive to mutation or truncation of this region of the crRNA [22,40].

The many contacts made by CRISPR endoRNases with both their substrate and product RNAs lead to exquisite binding specificity and affinity. These extensive binding interactions help to position the scissile phosphate in the enzyme active site for cleavage. The mechanism of catalysis by Cas5c and Cas6 enzymes by their variable active site residues is described in detail in the next section.

Mechanism of cleavage and active site plasticity

In addition to highly specific RNA binding, Cas5c and Cas6 enzymes cleave RNA exclusively because their catalytic mechanism requires nucleophilic attack of the scissile phosphate by the 2′ hydroxyl group of the 5′ upstream nucleotide [14,29,56,57]. Structured pre-crRNAs are cleaved 3′ of the stem loop, at or near its base. Replacing the upstream nucleotide with a deoxyribonucleotide prevents cleavage but permits tight binding, making a singly deoxy-substituted pre-crRNA substrate ideal for enzyme crystallography [14,16,17,30–32]. In a manner analogous to the tRNA-splicing endonuclease [58], this metal-independent cleavage reaction typically generates products with 5′ hydroxyl and 2′,3′ cyclic phosphate groups [13,19,22,40]. By contrast, mature crRNAs in the *P. aeruginosa* I-F system have a 3′ phosphate, possibly as a result of 2′,3′ cyclic phosphate opening via hydrolysis by water [33,57]. In the *Staphylococcus epidermidis* III-B system, processed crRNAs have a 3′ hydroxyl group, a currently unexplained observation that potentially suggests a distinct catalytic mechanism [20].

Most Cas6 and possibly Cas5c enzymes employ a general acid–base mechanism in which one residue acts as a general base to deprotonate the 2′ hydroxyl group and another acts as a general acid in protonation of the leaving group. However, there is great variability in their active sites and the catalytic residues are not always conserved, although catalytic histidines are common. For many Cas6 enzymes, mutation of this histidine dramatically reduces cleavage efficiency, but activity can be rescued by addition of imidazole (a histidine mimic), a trick that has been utilized in the development of Cas6-based applications discussed later [30,59].

By further analogy to the tRNA splicing endonuclease, BhaCas5c, PfuCas6-1, and TthCas6e have a catalytic triad of histidine, tyrosine, and lysine, where these residues may function as a general base, a general acid, and in stabilizing the pentacovalent phosphate intermediate, respectively (Figure 2) [13,16,17,22]. PaeCas6f uses a catalytic dyad of histidine and serine (Figure 3B) [14,57]. TthCas6A also has only two catalytic residues, histidine and arginine [30]. TthCas6B and the Cas6b enzyme from *M. maripaludis* each possess two histidine residues and a tyrosine that are important for cleavage [21,30]. Interestingly, mutation of either histidine alone does not abrogate enzyme activity, as it does for Cas6 proteins with single catalytic histidines, suggesting that Cas6b enzymes may have more flexible active sites.

Two recently described endoribonucleases from *S. solfataricus*, Cas6-1A and Cas6-1B (encoded by SSO1437), have a very different arrangement of their catalytic core [32,51]. These enzymes possess three lysine residues and an arginine in their active site (Figures 2C and 3C). This positively-charged center must position the scissile phosphate and carry out catalysis through a distinct mechanism. A detailed understanding of this class of CRISPR endonucleases will require further study.

Cas6 and Cas5c endonucleases have diverse active site arrangements that facilitate cleavage of pre-crRNAs through a metal-independent, general acid–base mechanism. Some of these enzymes have such high affinity for their RNA product that the cleavage reaction is single-turnover, while others may dissociate more readily. These properties have an important influence on the next step in the CRISPR pathway, supplying the mature crRNA to a complex for interference.

Interference complex assembly and crRNA loading

In the extensively studied Type I-E and I-F systems, Cas6 is an integral subunit of the targeting complex, Cascade. Recently, work on the Type I-B system showed that Cas6 is also part of the Cascade/I-B complex [34]. Cas6e and Cas6f remain tightly bound to the crRNA stem loop after cleavage, serving as a nucleation point for subsequent complex assembly [16,29,30,56]. In the *P. aeruginosa* Type I-F system, pre-crRNA processing is the requisite first step in Cascade/I-F formation [33,57]. In Cascade/I-E, a helix from Cas7 tethers the first backbone subunit to Cas6e, fitting snugly into a cleft between the two RRM folds, thereby initiating complex assembly [48–50].

Unlike most Type I systems, Cas6 does not associate stably with archaeal Cascade/I-A or with the Type III targeting complexes [18,35–39]. Instead, it cleaves the pre-crRNA and appears to dissociate before assembly of the Cascade/I-A, Csm, or Cmr complexes is completed. In line with this idea, PfuCas6-1 of the Type III-B system copurifies with a partially processed crRNA that has not yet been trimmed, known as the 1× intermediate, but does not associate with the mature crRNA (Figure 1) [56]. It was recently shown that Type I-A SsoCas6-1B is capable of robust multiple-turnover pre-crRNA processing [60], an activity that had never before been experimentally observed. Multiple-turnover cleavage may indeed be a general property of Cas6 enzymes that do not associate with interference complexes, although confirmation of this idea awaits further *in vitro* enzymatic studies.

Another hallmark characteristic of Type III CRISPR systems is trimming of the crRNA from the 3′ end [23,39,43]. It was recently demonstrated that trimming also occurs in the Type I-A system from *Thermoproteus tenax* [35,44], but it was not observed in the *S. solfataricus* I-A system [18], and thus may not be a general feature of the subtype. Exactly how and when this maturation step occurs remain open questions. The trimming nuclease is entirely unknown, and is not likely to be CRISPR-specific. Some evidence suggests that 3′ trimming happens after targeting complex assembly. In *S. epidermidis* and *Sulfolobus islandicus*, multiple Type III interference complex subunits are necessary for the accumulation of mature crRNAs, but the complexes themselves do not trim crRNAs *in vitro* [37,61]. This suggests that Csm/Cmr complexes assemble around an incompletely processed crRNA, unprotected regions are trimmed by a cellular nuclease, and complex size thereby serves as the determinant of mature crRNA length. Trimming may occur specifically in Type I-A and III CRISPR systems because when Cas6 does not stably associate with the interference complex, the 3′ end of the crRNA is accessible to nuclease activity.

A lingering question in crRNA biogenesis and interference complex loading is how these processes take place in organisms with multiple CRISPR–Cas systems. Many prokaryotes encode more than one Cas6 protein and/or possess CRISPR loci with different repeat sequences. Does a single Cas6 cleave multiple repeat sequences, does each repeat have a dedicated Cas6 enzyme responsible for its processing, or is it variable? In some cases, there has been tight coevolution between each Cas6 and its cognate repeat sequence. Examples include *Synechocystis* sp. PCC6803 and *S. thermophilus*, in which each expressed Cas6 specifically processes one distinct family of repeats [45,62]. Some organisms have more promiscuous Cas6 proteins. In *S. islandicus*, a single Cas6 protein generates crRNA species for three different interference complexes

[61]. *Methanosarcina mazei* has two CRISPR repeat sequences that can be cleaved by either of the two Cas6 enzymes it encodes [63]. TthCas6e cleaves only one of the three repeat sequences, while TthCas6A and TthCas6B can efficiently cleave either of the other two repeats [16,17,30]. Similarly, SsoCas6-1B cleaves one repeat specifically, while substrate selection by SsoCas6-3 is less stringent [60]. Intriguingly, there is a biased distribution of spacers between the three different interference complexes in this system [36,39,60]. Certain crRNAs are preferentially incorporated into the Csm or Cmr complexes, suggesting some sort of coordinated hand-off between each Cas6 and a specific complex. In addition to the general process of crRNA loading into Type I-A and III complexes, the mechanism by which specific crRNAs are sorted into different interference complexes will be an enticing topic to explore in future research.

Concluding remarks

Cas5c and Cas6 proteins make up a class of endoribonucleases required for the generation of mature crRNAs, which are used to target complementary nucleic acids for destruction as part of the CRISPR–Cas adaptive immune system in prokaryotes. Cas6 is the dedicated endoRNase for Type III and most Type I systems. Cas5 proteins are non-catalytic components of the Cascade interference complex in all Type I systems except the I-C subtype, in which the Cas5c variant has evolved to perform the enzymatic function of Cas6. Despite substantial sequence variation, all Cas6 enzymes share several structural motifs that facilitate sequence- and structure-specific pre-crRNA binding. Cas5c is structurally related to Cas6 but its RNA-binding properties are unclear. Both endonucleases catalyze metal-independent cleavage of the pre-crRNA at the base of the stem loop, although some Cas6s cleave substrates predicted to be unstructured. Some Cas endoRNases stay tightly bound to the stem loop product after cleavage and serve as a nucleation point for assembly of Cascade, while others dissociate at some point during crRNA delivery to the targeting complex.

Beyond their biological functions, CRISPR endoRNases are beginning to be applied for both *in vitro* and synthetic biology applications. For example, these enzymes can be used to ensure 5′ end homogeneity of *in vitro* transcribed RNAs [64], generate guide RNAs for genomic engineering in cells [65,66], and have also been adapted to cleave polycistronic transcripts in bacteria [67]. Taking advantage of its high affinity for its substrate, a catalytically inactive version of PaeCas6f can be used for transcript isolation *in vitro*. In this application, transcripts are tagged with the PaeCas6f recognition hairpin at their 5′ end, followed by incubation with cell extract. After binding to immobilized PaeCas6f and washing to remove nonspecific binders, PaeCas6f cleavage is activated by the addition of imidazole to release the RNA together with any high-affinity binding partners [59,68].

Future research will be needed to address outstanding questions about CRISPR endoRNases, including how the crRNA is trimmed in Type III systems and how crRNAs are transferred to the interference machinery in systems where Cas6 does not stably associate with the targeting complex. In addition, it will be interesting to determine the range of RNA structures and sequences recognized by Cas6 enzymes, as well as why some Cas6s dimerize and how this affects crRNA processing. Furthermore, a crystal structure of a Cas5c enzyme bound to its substrate or product could reveal the mechanism by which crRNAs are recognized by this distinctive family of endonucleases. Finally, it will be exciting to determine the scope of activities of Cas6 enzymes in mammalian and other eukaryotic systems, particularly for research and synthetic biology applications.

References

1 Wiedenheft, B. *et al.* (2012) RNA-guided genetic silencing systems in bacteria and archaea. *Nature* 482, 331–338

2 Westra, E.R. *et al.* (2012) The CRISPRs, they are a-changin': how prokaryotes generate adaptive immunity. *Annu. Rev. Genet.* 46, 311–339

3 Sorek, R. *et al.* (2013) CRISPR-mediated adaptive immune systems in bacteria and archaea. *Annu. Rev. Biochem.* 82, 237–266

4 Reeks, J. *et al.* (2013) CRISPR interference: a structural perspective. *Biochem. J.* 453, 155–166

5 van der Oost, J. *et al.* (2014) Unravelling the structural and mechanistic basis of CRISPR–Cas systems. *Nat. Rev. Microbiol.* 12, 479–492

6 Barrangou, R. *et al.* (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315, 1709–1712

7 Garneau, J.E. *et al.* (2010) The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* 468, 67–71

8 Heler, R. *et al.* (2014) Adapting to new threats: the generation of memory by CRISPR–Cas immune systems. *Mol. Microbiol.* 93, 1–9

9 Makarova, K.S. *et al.* (2006) A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol. Direct* 1, 7

10 Kunin, V. *et al.* (2007) Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol.* 8, R61

11 Grissa, I. *et al.* (2007) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* 8, 172

12 Brouns, S.J.J. *et al.* (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* 321, 960–964

13 Carte, J. *et al.* (2008) Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes Dev.* 22, 3489–3496

14 Haurwitz, R.E. *et al.* (2010) Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science* 329, 1355–1358

15 Deltcheva, E. *et al.* (2011) CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* 471, 602–607

16 Sashital, D.G. *et al.* (2011) An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3. *Nat. Struct. Mol. Biol.* 18, 680–687

17 Gesner, E.M. *et al.* (2011) Recognition and maturation of effector RNAs in a CRISPR interference pathway. *Nat. Struct. Mol. Biol.* 18, 688–692

18 Lintner, N.G. *et al.* (2011) Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE). *J. Biol. Chem.* 286, 1–14

19 Jore, M.M. *et al.* (2011) Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nat. Struct. Mol. Biol.* 18, 529–536

20 Hatoum-Aslan, A. *et al.* (2011) Mature clustered, regularly interspaced, short palindromic repeats RNA (crRNA) length is measured by a ruler mechanism anchored at the precursor processing site. *Proc. Natl. Acad. Sci. U.S.A.* 108, 21218–21222

21 Richter, H. *et al.* (2012) Characterization of CRISPR RNA processing in *Clostridium thermocellum* and *Methanococcus maripaludis*. *Nucleic Acids Res.* 40, 9887–9896

22 Nam, K.H. *et al.* (2012) Cas5d protein processes pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvulg CRISPR–Cas system. *Structure* 20, 1574–1584

23 Hale, C. *et al.* (2008) Prokaryotic silencing (psi)RNAs in *Pyrococcus furiosus*. *RNA* 14, 2572–2579

24 Marraffini, L.A. and Sontheimer, E.J. (2008) CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* 322, 1843–1845

25 Jinek, M. *et al.* (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337, 816–821

26 Westra, E.R. *et al.* (2014) CRISPR–Cas systems: beyond adaptive immunity. *Nat. Rev. Microbiol.* 12, 317–326

27 Makarova, K.S. *et al.* (2011) Evolution and classification of the CRISPR–Cas systems. *Nat. Rev. Microbiol.* 9, 467–477

28 Ebihara, A. *et al.* (2006) Crystal structure of hypothetical protein TTHB192 from *Thermus thermophilus* HB8 reveals a new protein family with an RNA recognition motif-like domain. *Protein Sci.* 15, 1494–1499

29 Sternberg, S.H. *et al.* (2012) Mechanism of substrate selection by a highly specific CRISPR endoribonuclease. *RNA* 18, 661–672

30 Niewoehner, O. *et al.* (2014) Evolution of CRISPR RNA recognition and processing by Cas6 endonucleases. *Nucleic Acids Res.* 42, 1341–1353

31 Wang, R. *et al.* (2011) Interaction of the Cas6 riboendonuclease with CRISPR RNAs: recognition and cleavage. *Structure* 19, 257–264

32 Shao, Y. and Li, H. (2013) Recognition and cleavage of a nonstructured CRISPR RNA by its processing endoribonuclease Cas6. *Structure* 21, 385–393

33 Wiedenheft, B. *et al.* (2011) RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proc. Natl. Acad. Sci. U.S.A.* 108, 10092–10097

34 Brendel, J. *et al.* (2014) A complex of Cas proteins 5, 6, and 7 is required for the biogenesis and stability of clustered regularly interspaced short palindromic repeats (CRISPR)-derived RNAs (crRNAs) in *Haloferax volcanii*. *J. Biol. Chem.* 289, 7164–7177

35 Plagens, A. *et al.* (2014) In vitro assembly and activity of an archaeal CRISPR–Cas type I-A Cascade interference complex. *Nucleic Acids Res.* 42, 5125–5138

36 Rouillon, C. *et al.* (2013) Structure of the CRISPR interference complex CSM reveals key similarities with Cascade. *Mol. Cell* 52, 124–134

37 Hatoum-Aslan, A. *et al.* (2013) A ruler protein in a complex for antiviral defense determines the length of small interfering CRISPR RNAs. *J. Biol. Chem.* 288, 27888–27897

38 Hale, C.R. *et al.* (2009) RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* 139, 945–956

39 Zhang, J. *et al.* (2012) Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. *Mol. Cell* 45, 303–313

40 Garside, E.L. *et al.* (2012) Cas5d processes pre-crRNA and is a member of a larger family of CRISPR RNA endonucleases. *RNA* 18, 2020–2028

41 Koo, Y. *et al.* (2013) Conservation and variability in the structure and function of the Cas5d endoribonuclease in the CRISPR-mediated microbial immune system. *J. Mol. Biol.* 425, 3799–3810

42 Punetha, A. *et al.* (2014) Active site plasticity enables metal-dependent tuning of Cas5d nuclease activity in CRISPR–Cas type I-C system. *Nucleic Acids Res.* 42, 3846–3856

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

44 Plagens, A. *et al.* (2012) Characterization of the CRISPR/Cas subtype I-A system of the hyperthermophilic crenarchaeon *Thermoproteus tenax*. *J. Bacteriol.* 194, 2491–2500

45 Scholz, I. *et al.* (2013) CRISPR–Cas systems in the cyanobacterium *Synechocystis* sp. PCC6803 exhibit distinct processing pathways involving at least two Cas6 and a Cmr2 protein. *PLoS ONE* 8, e56470

46 Haft, D.H. *et al.* (2005) A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput. Biol.* 1, e60

47 Wang, R. *et al.* (2012) The impact of CRISPR repeat sequence on structures of a Cas6 protein–RNA complex. *Protein Sci.* 21, 405–417

48 Jackson, R.N. *et al.* (2014) Crystal structure of the CRISPR RNA-guided surveillance complex from *Escherichia coli*. *Science* 345, 1473–1479

49 Zhao, H. *et al.* (2014) Crystal structure of the RNA-guided immune surveillance Cascade complex in *Escherichia coli*. *Nature*. Published online August 12, 2014, http://dx.doi.org/10.1038/nature13733

50 Mulepati, S. *et al.* (2014) Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. *Science* 345, 1479–1484

51 Reeks, J. *et al.* (2013) Structure of a dimeric crenarchaeal Cas6 enzyme with an atypical active site for CRISPR RNA processing. *Biochem. J.* 452, 223–230

52 Makarova, K.S. *et al.* (2002) A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res.* 30, 482–496

53 Park, H-M. *et al.* (2012) Crystal structure of a Cas6 paralogous protein from *Pyrococcus furiosus*. *Proteins* 80, 1895–1900

54 Przybilski, R. *et al.* (2011) Csy4 is responsible for CRISPR RNA processing in *Pectobacterium atrosepticum*. *RNA Biol.* 8, 517–528

55 Richter, H. *et al.* (2013) Comparative analysis of Cas6b processing and CRISPR RNA stability. *RNA Biol.* 10, 700–707

56 Carte, J. *et al.* (2010) Binding and cleavage of CRISPR RNA by Cas6. *RNA* 16, 2181–2188

57 Haurwitz, R.E. *et al.* (2012) Csy4 relies on an unusual catalytic dyad to position and cleave CRISPR RNA. *EMBO J.* 31, 2824–2832

58 Xue, S. *et al.* (2006) RNA recognition and cleavage by a splicing endonuclease. *Science* 312, 906–910

59 Lee, H.Y. *et al.* (2013) RNA–protein analysis using a conditional CRISPR nuclease. *Proc. Natl. Acad. Sci. U.S.A.* 110, 5416–5421

60 Sokolowski, R.D. *et al.* (2014) Cas6 specificity and CRISPR RNA loading in a complex CRISPR–Cas system. *Nucleic Acids Res.* 42, 6532–6541

61 Deng, L. *et al.* (2013) A novel interference mechanism by a type IIIB CRISPR–Cmr module in *Sulfolobus*. *Mol. Microbiol.* 87, 1088–1099

62 Carte, J. *et al.* (2014) The three major types of CRISPR–Cas systems function independently in CRISPR RNA biogenesis in *Streptococcus thermophilus*. *Mol. Microbiol.* 93, 98–112

63 Nickel, L. *et al.* (2013) Two CRISPR–Cas systems in *Methanosarcina mazei* strain Gö1 display common processing features despite belonging to different types I and III. *RNA Biol.* 10, 779–791

64 Salvail-Lacoste, A. *et al.* (2013) Affinity purification of T7 RNA transcripts with homogeneous ends using ARiBo and CRISPR tags. *RNA* 19, 1003–1014

65 Nissim, L. *et al.* (2014) Multiplexed and programmable regulation of gene networks with an integrated RNA and CRISPR/Cas toolkit in human cells. *Mol. Cell* 54, 698–710

66 Tsai, S.Q. *et al.* (2014) Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nat. Biotechnol.* 32, 569–576

67 Qi, L. *et al.* (2012) RNA processing enables predictable programming of gene expression. *Nat. Biotechnol.* 30, 1002–1006

68 Hutin, S. *et al.* (2013) An RNA element in human interleukin 6 confers escape from degradation by the gammaherpesvirus SOX protein. *J. Virol.* 87, 4672–4682