Biochem. J. (2013) **453**, 155–166 (Printed in Great Britain) doi:10.1042/BJ20130316

multimedia online data

REVIEW ARTICLE

CRISPR interference: a structural perspective

Judith REEKS, James H. NAISMITH[1] and Malcolm F. WHITE[1]

Biomedical Sciences Research Complex, University of St Andrews, St Andrews, Fife KY16 9ST, U.K.

CRISPR (cluster of regularly interspaced palindromic repeats) is a prokaryotic adaptive defence system, providing immunity against mobile genetic elements such as viruses. Genomically encoded crRNA (CRISPR RNA) is used by Cas (CRISPR-associated) proteins to target and subsequently degrade nucleic acids of invading entities in a sequence-dependent manner. The process is known as 'interference'. In the present review we cover recent progress on the structural biology of the CRISPR/Cas system, focusing on the Cas proteins and complexes that catalyse crRNA biogenesis and interference. Structural studies have helped in the elucidation of key mechanisms, including the recognition and cleavage of crRNA by the Cas6 and Cas5 proteins, where remarkable diversity at the level of both substrate recognition and catalysis has become apparent. The RNA-binding RAMP (repeat-associated mysterious protein) domain is present in the Cas5, Cas6, Cas7 and Cmr3 protein families and RAMP-like domains are found in Cas2 and Cas10. Structural analysis has also revealed an evolutionary link between the small subunits of the type I and type III-B interference complexes. Future studies of the interference complexes and their constituent components will transform our understanding of the system.

Key words: antiviral defence, cluster of regularly interspaced palindromic repeats (CRISPR), crystallography, evolution, protein structure, repeat-associated mysterious protein (RAMP).

INTRODUCTION

CRISPRs (cluster of regularly interspaced palindromic repeats) are a prokaryotic defence mechanism against viral infection and horizontal gene transfer. CRISPRs are the largest family of prokaryotic repeats [1] and have been found in 48 % of bacterial and 84 % of archaeal sequenced genomes to date [2]. A CRISPR array consists of a series of short identical repeat sequences separated by similarly short variable sequences known as spacers [3]. Located adjacent to the CRISPR array are clusters of *cas* (CRISPR-associated) genes [4] that encode for the proteins responsible for mediating the CRISPR response to foreign nucleic acids. The spacers are derived from foreign nucleic acids, such as viruses and conjugative plasmids, and provide the host with a 'genetic memory' of threats previously encountered [1,5,6]. New spacers are captured in a poorly understood process known as 'adaptation' and incorporated into the CRISPR locus [7]. The spacers are used to target foreign nucleic acids containing sequences complementary to the spacer, termed protospacers, for degradation [8]; the process is termed 'interference'.

The first step in the interference pathway is the transcription of the CRISPR array from a promoter located in the 'leader' sequence, an AT-rich region located upstream of the CRISPR array [4,9]. The array transcript {pre-crRNA [precursor crRNA (CRISPR RNA)]} is then processed into short crRNAs containing a spacer and flanking repeat fragments (Figure 1) [10]. These crRNAs are subsequently bound by complexes of Cas proteins and used to target homologous foreign dsDNA (double-stranded DNA) or ssRNA (single-stranded RNA) for nucleolytic degradation during CRISPR interference (Figure 1) [8,11].

The CRISPR/Cas systems are divided into three main types (I, II and III) on the basis of the identity and organisation of genes within a *cas* locus [12]. These types are further divided into a total of ten subtypes (I-A, I-B and so on), each of which expresses a different protein complex responsible for interference (Figures 1 and 2). The Cascade (CRISPR-associated complex for antiviral defence) is the effector complex for type I systems [8,13–15]. This name was originally used solely for the type I-E complex [8], which we here call *e*Cascade, but increasingly Cascade is used more as a general term for all type I complexes. Type II systems use a single protein for interference (Cas9) [16], whereas the III-B subtype uses the CMR complex [11]. The interference complex of the III-A subtype has yet to be characterized biochemically, but the similarity of the III-A and III-B operons suggests that interference is indeed mediated by an effector complex rather than a single protein. As a result the putative complex has been termed the CSM complex [12]. Every CRISPR/Cas system apart from the III-B subtype is thought to target dsDNA by forming an R-loop structure, consisting of a heteroduplex between crRNA and the complementary protospacer strand and a ssDNA (single-stranded DNA) non-complementary strand, followed by degradation by the interference nuclease (Figure 1) [8,17–19]. The CMR complex targets ssRNA by forming an RNA duplex, which is subsequently cleaved [11,20].

The mechanisms of adaptation and CRISPR interference have been extensively reviewed (see references [21–26]). In the present review we will focus on the structural biology of the CRISPR system. Crystal structures are available for eight of the 'core' Cas proteins (those found in multiple subtypes) as well as a number of subtype-specific proteins (Figure 2 and Supplementary Table

Abbreviations used: BhCas5c, *Bacillus halodurans* Cas5c; CRISPR, cluster of regularly interspaced palindromic repeats; Cas, CRISPR-associated; Cascade, CRISPR-associated complex for antiviral defence; crRNA, CRISPR RNA; dsDNA, double-stranded DNA; EcoCas3, *Escherichia coli* Cas3; EM, electron microscopy; HD, histidine–aspartate; MjaCas3″, *Methanocaldococcus jannaschii* Cas3″; PaCas6f, *Pseudomonas aeruginosa* Cas6f; PAM, protospacer adjacent motif; PfuCas, *Pyrococcus furiosus* Cas; pre-crRNA, precursor crRNA; RAMP, repeat-associated mysterious protein; RRM, RNA recognition motif; ssDNA, single-stranded DNA; SsoCas, *Sulfolobus solfataricus* Cas; ssRNA, single-stranded RNA; SthCas3, *Streptococcus thermophilus* Cas3; tracrRNA, *trans*-activating crRNA; TtCas, *Thermus thermophilus* Cas.

[1] Correspondence may be addressed to either of these authors (email naismith@st-and.ac.uk or mfw2@st-and.ac.uk).

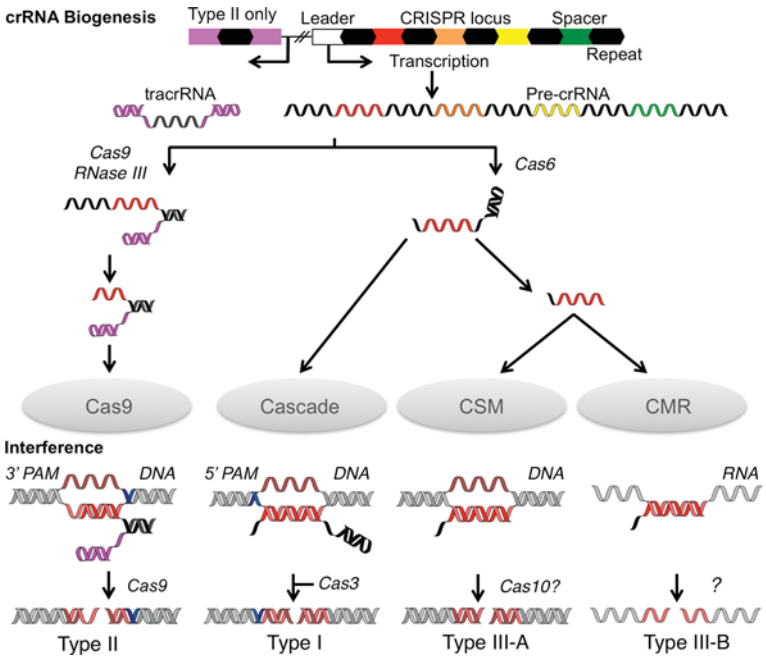

Figure 1 Schematic representation of crRNA biogenesis and CRISPR interference

Processing events involving nucleic acids are coloured; repeats (black), spacers (red–green) and tracrRNA (magenta). For clarity, a single spacer (red) was used to illustrate the processes, although in actual systems all spacers are processed. Targets are shown in other red shades (lighter for the complementary strand and darker for the non-complementary). The PAMs are shown in blue. The pre-crRNA and interference nucleases are indicated along with the interference complexes.

S1 at http://www.biochemj.org/bj/453/bj4530155add.htm). The structures of proteins involved in spacer acquisition have provided interesting insights into their function within the CRISPR/Cas system as well as to similarities to non-Cas proteins, such as the parallels between Cas2 and VapD of the toxin/antitoxin system [27], but will not be discussed further in the present review. EM (electron microscopy) images and structures have been determined for five interference complexes, providing invaluable information on the function of each subunit. CRISPR systems are remarkably diverse and subject to rapid evolutionary change. Analysis of the key structural features of Cas proteins involved in crRNA biogenesis and interference highlights recurring themes and points to evolutionary relationships between apparently distinct protein families.

PRE-crRNA PROCESSING AND crRNA BIOGENESIS

crRNA provides the CRISPR/Cas system with the sequence specificity needed to selectively target foreign nucleic acids. Mature crRNAs are produced from a single long transcript of the CRISPR array (pre-crRNA), which is processed to yield spacers with 5′ and/or 3′ repeat fragments (Figure 1) [10,28,29]. The method and nature of pre-crRNA processing is dependent on the CRISPR/Cas system. Type I and III systems use the Cas6 endonuclease to cleave pre-crRNA within the repeat sequence [8,13,15], with the exception of I-C systems that instead use a catalytic variant of Cas5 [14,30]. The crRNAs from various type III systems are further processed to reduce or remove the repeat sequence at the 3′ end [11,31]. The enzyme responsible for this degradation is not yet known. The type II system uses a very different mechanism, requiring the transcript of an anti-sense near-perfect repeat and flanking sequences [tracrRNA (*trans*-activating crRNA)] located adjacent to the CRISPR array for processing [32]. The duplex formed by pre-crRNA and tracrRNA is bound by Cas9 and cleaved in the repeat sequence by cellular RNase III and then in the spacer by an unknown nuclease to leave a spacer fragment and a 3′ repeat fragment [32].

Cas6 and the catalytic type I-C Cas5 [Cas5c, also confusingly known as Cas5d (Dvulg subtype)] catalyse the same reaction. The pre-crRNA is cleaved upstream of the spacer (8 nt for Cas6, 11 nt for Cas5c) [8,13,14,33,34] generating crRNA with a 5′ repeat-derived sequence known as the ‘5′-handle’ or ‘5′-tag’ that is critical for interference [20,35]. CRISPR repeats can be divided into twelve families based on sequence and secondary structure [36]. Cas5c targets repeats containing hairpin structures, whereas the subfamilies of Cas6 proteins, which broadly align with the CRISPR/Cas system with which they are associated, can cleave either unstructured or hairpin-containing repeats [14,33,37,38].

In type I systems, Cas6 can form an integral part of Cascade or it can exhibit a more transient interaction. Cas6e and Cas6f remain tightly bound to their cleaved products with low or sub-nanomolar affinities, and form part of their respective Cascades [15,37,39,40]. In fact, the type I-F complex (*f*Cascade) assembles specifically around a pre-formed Cas6f/crRNA complex [41]. Cas6 interacts more transiently with the I-A archaeal Cascade (*a*Cascade) [13,42]. Cas6 is not part of the type III-B CMR complex [11,20], and the associations of Cas6 with the type I-B, I-D and III-A complexes are unclear.

The structures of Cas5c and Cas6

Cas5 and Cas6 both belong to the RAMP (repeat-associated mysterious protein) superfamily. These proteins contain one or more RAMP domains, which form ferredoxin-like folds similar to that of the RRM (RNA recognition motif) domain [43], consisting of a four-stranded antiparallel β-sheet (arranged as $\beta_4\beta_1\beta_3\beta_2$) flanked on one face by two α-helices located after β_1 and β_3 in a $\beta\alpha\beta\beta\alpha\beta$ fold (Figure 3A). Five conserved sequence motifs have

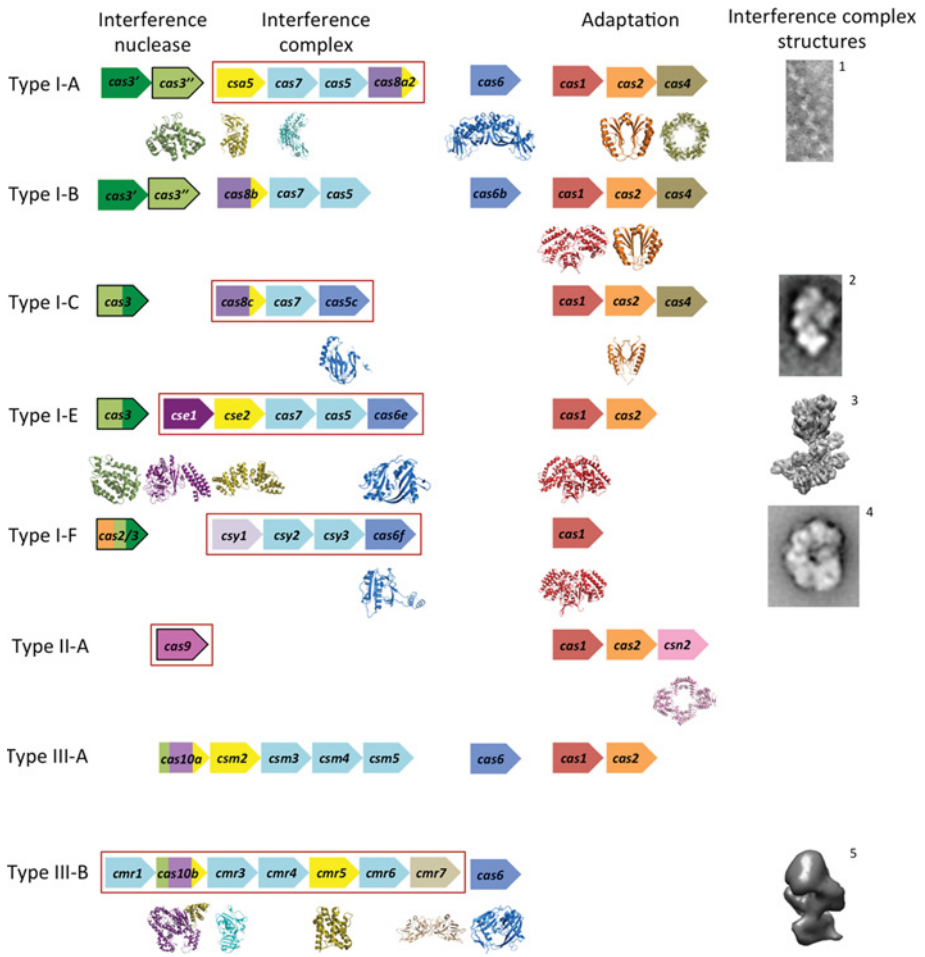

Figure 2 The CRISPR/Cas systems and their respective proteins

Typical gene identities are shown for CRISPR/Cas subtypes according to the recent classification by Makarova et al. [12]. The genes are ordered by function: interference (left) and adaptation (right). The interference proteins are subdivided into the interference nuclease (left, outlined in black), proteins of the interference complex (middle, boxed in red) and pre-crRNA nucleases (right, although some are integral subunits of the interference complexes). The genes are coloured according to conserved domain and protein folds: catalytic RAMPs are shown in blue, non-catalytic RAMPs in light blue, HD nuclease domains in light green, Cas3 helicase domains in dark green, the large subunits in various shades of purple and the small subunits in yellow. Subtypes I-D and II-B are not shown as there is no directly relevant structural data. EM images and structures of the interference complexes (or subcomplex for I-A) are adapted from references [1] [13], [2] [14], [3] EMD-5314, [4] [15] and [5] [20].

been detected in the superfamily; as yet no single protein has been found to contain all five [44].

Cas6 proteins typically contain two sequential RAMP domains with the glycine-rich loop (motif V of the RAMP superfamily sequence motifs) located between $\alpha_{2'}$ and $\beta_{4'}$ of the second (C-terminal) domain (the prime denotes a structural element in the second domain) (Figure 3B) [45–50]. This loop often fits the consensus sequence GΦGXXXXXGΦG, where Φ is a hydrophobic residue, X is any residue and the variable region contains at least one positively charged residue [51]. Other than this motif, the Cas6 proteins exhibit minimal sequence similarity. PaCas6f (*Pseudomonas aeruginosa* Cas6f) is atypical because it contains what is possibly a severely degraded C-terminal RAMP domain (Figure 3C) [33]. The C-terminal domain contains four short β-strands that, although they are orientated to form a RAMP β-sheet, are not aligned to do so (Figure 3C). The RAMP helices are not present, but the glycine-rich loop (albeit differing from the consensus sequence) is located between the correct β-strands. The Cas6 homologues contain additional secondary structure elements relative to the RAMP elements, but only one feature is fully conserved: a β-hairpin connecting $\beta_{2'}$ and $\beta_{3'}$ in the C-terminal domain (we denote this the $\beta_{2'}$–$\beta_{3'}$ hairpin) that extends beyond the β-sheet. This hairpin is even conserved in the abnormal C-terminal domain of PaCas6f.

Cas5c contains an N-terminal RAMP domain and a C-terminal domain consisting of a three-stranded antiparallel β-sheet (Figure 3D) [14,30,52]. The RAMP domain contains a glycine-rich loop that does not match the Cas6 consensus sequence. It also contains a β_2–β_3 hairpin that is joined by another short β-strand to form a β-sheet. In some Cas5c homologues, two helices are inserted into the tip of the hairpin [52]. Due to the hairpin and the glycine-rich loop, this RAMP domain is similar to the Cas6 C-terminal domain, although it also exhibits significant similarity to the N-terminal domain of archaeal Cas6 proteins. In Cas5c, α_2 is not located behind β_4; instead, the shorter β_4 (in other RAMPs, β_4 is longer or is followed by an extended strand) allows α_2 to run antiparallel to β_1 (compare Figure 3B with Figure 3D). This atypical arrangement could correctly position the residues of the active site, which is located at the intersection of α_1 and α_2 at the top of the β-sheet, a location different to that of Cas6 (see below). The β-sheet of the C-terminal domain does not have a RAMP domain arrangement of secondary structure elements. However, $\beta_{1'}$ and

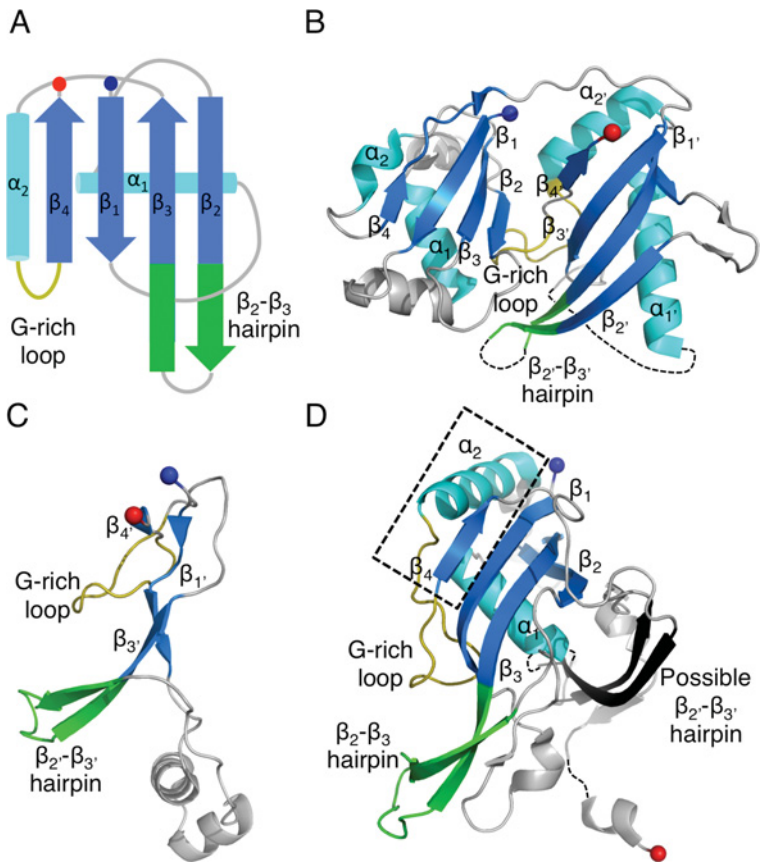

Figure 3 The structures of catalytic RAMP proteins

(**A**) Topology diagram of a RAMP domain. The β-strands are shown in blue and the α-helices in cyan. The glycine-rich loop found in many RAMPs is shown in yellow and the β_2–β_3 hairpin observed in some RAMPs is shown in green. The N- and C-termini are shown as blue and red spheres respectively. (**B**) The structure of TtCas6e (PDB code 1WJ9) highlighting the two RAMP domains that may have arisen from a pseudo-duplication event. Secondary structural elements are labelled as described in the text. Conserved RAMP elements are coloured as in (**A**) and non-conserved elements in grey. Disordered regions are shown as broken black lines. (**C**) The atypical C-terminal domain of PaCas6f (PDB code 2XLK) that probably diverged from the standard RAMP fold. The recognizable features are labelled. (**D**) The structure of BhCas5c (PDB code 4F3M), a catalytic variant of the typically non-catalytic Cas5 family. The short β_4 strand and parallel α_2 helix are boxed in black. The possible $\beta_{2'}$–$\beta_{3'}$ hairpin in the C-terminal domain is shown in black.

$\beta_{2'}$ form an extended β-hairpin reminiscent of the $\beta_{2'}$–$\beta_{3'}$ hairpin of Cas6, although this is the only feature that is potentially RAMP-like. Thus it is not possible to say with certainty whether the C-terminal domain of Cas5c is a highly divergent RAMP domain.

RNA binding and cleavage

Cas5c and Cas6 are both metal-independent ribonucleases that form products with 5′-hydroxyls and 2′,3′-cyclic phosphates [30,33,46,53], indicative of a general acid/base mechanism involving nucleophilic attack by the deprotonated 2′-hydroxyl on the scissile phosphate. The active site of Cas6 is located between α_1 and the glycine-rich loop, although the exact position of the site varies amongst the subfamilies (Figures 4A–4D). Remarkably, the catalytic residues also vary between the proteins and none of the residues are conserved in all of the Cas6 subfamilies. Cas6 enzymes from *Pyrococcus furiosus* (PfuCas6) and *Thermus thermophilus* (TtCas6) possess a catalytic triad of histidine, tyrosine and lysine residues similar to the RNA-splicing endonuclease [37,46,54,55]. The tyrosine residue has been assigned as the general base and the histidine residue as the general acid, with the lysine residue stabilizing the pentacoordinate phosphate intermediate. PaCas6f, however, uses a catalytic dyad of histidine and serine residues, with the histidine residue acting as the general base and the serine residue holding the ribose ring in the correct conformation [41]. Two active Cas6 paralogues from *Sulfolobus solfataricus* contain neither a general acid nor a general base, instead using conserved positively charged residues to correctly orientate the substrate and stabilize the pentacoordinate phosphate intermediate [49,50]. The presence of a catalytic histidine residue in the N-terminal domain had previously been highlighted as a characteristic feature of Cas6s [56], but it is now clear that this is not necessarily the case.

The location of the Cas5c active site is different to that of Cas6, suggesting that the active sites evolved independently of each other. The catalytic triad of BhCas5c (*Bacillus halodurans* Cas5c) consists of a tyrosine residue located in α_1 and histidine and lysine residues in α_2, similar to the PfuCas6 and TtCas6e active sites [14]. The lysine is the only residue of the triad that is invariant across the family; the tyrosine residue can be exchanged for histidine (as in the active Cas5c nucleases from *Mannheimia succiniciproducens* and *Xanthomonas oryzae* [30,52]), phenylalanine or leucine, whereas the catalytic histidine residue can be replaced by other aromatic residues (phenylalanine/tyrosine) (Supplementary Figure S1 at http://www.biochemj.org/bj/453/bj4530155add.htm), but the roles of the residues are not yet understood. None of these supposed catalytic residues are conserved in other Cas5 proteins, perhaps unsurprisingly since only Cas5c is catalytically active.

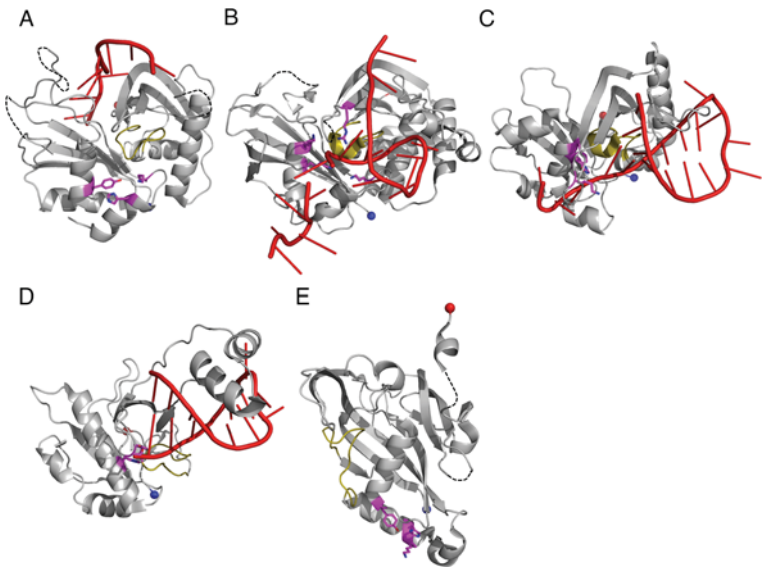

Figure 4 RNA binding and catalysis by Cas6 and Cas5c

The structures of (**A**) PfuCas6 (PDB code 3PKM), (**B**) SsoCas6 (PDB code 4ILL), (**C**) TtCas6e (PDB code 2Y8W) and (**D**) PaCas6f (PDB code 2XLK) in complex with RNA (red). The glycine-rich loop is shown in yellow and the catalytic residues as magenta sticks. (**E**) The structure of BhCas5c (PDB 4F3M) highlighting the position of the active site (magenta). The four structures are shown to the same scale and same orientation. A three-dimensional representation of this Figure is available at http://www.biochemj.org/bj/453/0155/bj4530155add.htm.

As expected for nucleases that process RNA substrates with a range of secondary structures, multiple modes of RNA binding have been observed across the Cas6 family. This perhaps underlies the variation in the position of the active site as the different modes alter the position of the scissile bond. PfuCas6 and its inactive homologue from *Pyrococcus horikoshii* (PhCas6nc) bind unstructured RNA in a 'wrap-around' mechanism where the RNA binds in the cleft between the two domains (Figure 4A) [38,48]. These enzymes bind the 5′ end of the repeat in the cleft between the β-sheets of the two domains and this interaction with the first $\sim$10 nt appears to be the predominant determinant of binding affinity. Although the 3′ end of the substrate, including the scissile phosphate, is disordered in the crystal structures, it is predicted to follow the positively charged cleft into the active site [38]. TtCas6e, PaCaf6f and a homologue from *S. solfataricus* (SsoCas6) bind hairpin RNA with the majority of the contacts formed by the C-terminal domain (Figures 4B–4D). TtCas6e and SsoCas6 bind the hairpin across the helical face of the protein using a series of basic residues to bind the phosphate backbone of the 3′ strand of the hairpin [37,49,55]. The RNA hairpin of SsoCas6 is shorter than that of TtCas6e by 3 bp and is predicted to be unstable in solution [36], meaning that SsoCas6 specifically stabilizes the hairpin conformation. PaCas6f, which shares few C-terminal secondary structure elements with other Cas6 proteins, binds the RNA hairpin between the RAMP β-strands and a helix–loop–helix motif, using the first helix to bind the major groove of the RNA [33]. In all three of these proteins, the $\beta_{2'}$–$\beta_{3'}$ hairpin is inserted into the base of the RNA hairpin, serving to position the scissile phosphate within the active site and, in the case of PaCas6f and SsoCas6, provides key catalytic residues. It seems likely that the $\beta_{2'}$–$\beta_{3'}$ hairpin plays a conserved role across the Cas6 family.

The method of substrate binding in Cas5c must be significantly different to that observed in Cas6 proteins, because the active sites of the two families are in different locations (Figure 4). In Cas5c, RNA is expected to bind to the helical face of the protein, which in all structures is positively charged, particularly adjacent to the active site [14,30]. Both domains of Cas5c are implicated in binding the substrate, including the β-sheet encompassing the putative $\beta_{2'}$–$\beta_{3'}$ hairpin [14,30]. However, neither the β_2–β_3 nor the $\beta_{2'}$–$\beta_{3'}$ hairpin can function by inserting at the base of the RNA hairpin, as this would place the scissile phosphate too far away from the active site. A complex structure of Cas5c and substrate is required to determine the exact mode of binding.

The method of RNA binding for Cas5c and Cas6 differs from typical RRMs, which contain the same ferredoxin-like fold as RAMPs. Typical RRMs possess two conserved sequence motifs located in β_1 and β_3 (termed RNP2 and RNP1 respectively) that are not present in RAMPs (Supplementary Figure S2 at http://www.biochemj.org/bj/453/bj4530155add.htm) [57,58]. These motifs allow RRMs to bind ssRNA or ssDNA across the face of the β-sheet [59,60], although not hairpin or dsRNA (double-stranded RNA), whereas RAMPs bind ssRNA or hairpin RNA through diverse modes of binding.

The active sites appear to have evolved independently for Cas6 and Cas5c, and even within the Cas6 family there is no universally conserved catalytic mechanism. Given that the catalytic rate constants of these enzymes, at 1–5 min^{-1} [37,40], are of the same order as those observed for catalytic RNA [61], these enzymes may be more constrained by the need to recognize pre-crRNA specifically than by a requirement to turn over rapidly.

THE PROTEINS OF THE INTERFERENCE COMPLEXES

Atomic level detail structures are now available for a number of individual proteins that are involved in interference. In addition, EM structures have been solved for a number of the interference complexes (Figure 2). The highest resolution structures available are those of the *Escherichia coli e*Cascade in complex with crRNA and with a crRNA/protospacer RNA duplex at resolutions of 8 and 9 Å (1 Å = 0.1 nm) respectively [39]. Lower resolution images and structures are also available for the *B. halodurans c*Cascade [14], *Ps. aeruginosa f*Cascade [15] and *S. solfataricus* CMR complex [20] as well as the core complex of *S. solfataricus a*Cascade [13]. Although the overall complex topologies can be

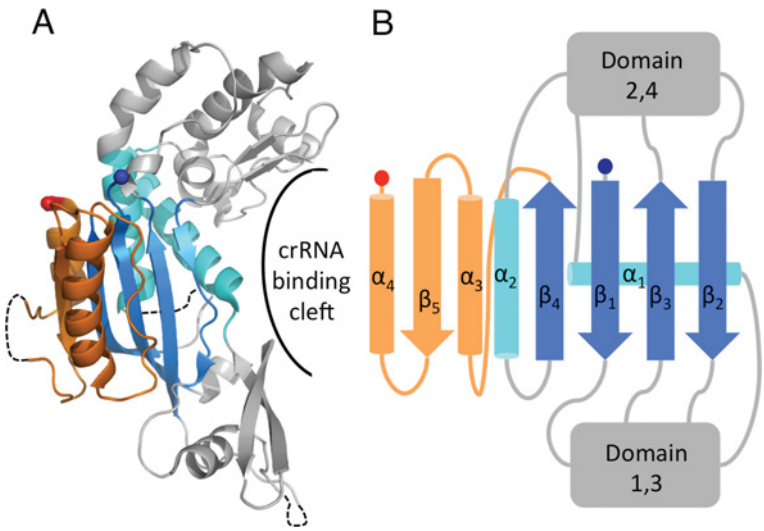

Figure 5 The structure of Cas7, the core subunit of Cascade

(**A**) The structure of SsoCas7 (PDB code 3PS0) where the central RAMP domain is extended by an $\alpha\beta\alpha$ motif (orange) and flanked by two unique domains (grey). The proposed crRNA-binding cleft located across the face of the β-sheet is indicated. (**B**) Topology diagram of SsoCas7 showing the connectivity of the RAMP fold relative to the other domains.

discerned, the resolution of these structures has precluded reliable placement of individual proteins within the complex.

Cas7, the backbone of the type I complex

The structural backbone of Cascade is composed of multiple monomers of Cas7 [13,14,39]. In *e*Cascade, Cas7 assembles into a helical hexameric structure with crRNA binding in a groove formed along the outer face of the oligomer [39]. This helical arrangement is conserved in the core complex of the *S. solfataricus* *a*Cascade, although this complex of Cas5 and Cas7 forms oligomers of variable length [13]. It is possible that further factors are needed to produce a complex of defined length or perhaps *a*Cascade exhibits greater structural plasticity than *e*Cascade. A similar helical arrangement to *e*Cascade was observed in EM images of *c*Cascade [14], and, although it was not possible to unambiguously define the quaternary structure of the complex, it is probable that the six Cas7 subunits of the complex form the same backbone. *f*Cascade contains six Csy3 subunits with a similar twisted topology to both *c*Cascade and *e*Cascade [15]. This, combined with secondary structure predictions and MS fragmentation analysis, has recently led to the hypothesis that Csy3 actually belongs in an expanded Cas7 family [56,62]. Similar structure predictions place Csc2 of *d*Cascade in the Cas7 family [56], suggesting that the Cas7 helical backbone is a conserved and perhaps characteristic feature of all Cascade complexes.

The structure of Cas7 from one of the *S. solfataricus* *a*Cascade complexes [13] (termed SsoCas7) contains a central RAMP fold modified with an additional $\alpha\beta\alpha$ motif located immediately after β_4 (Figure 5A). This motif adds a fifth strand to the β-sheet ($\beta_5\beta_4\beta_1\beta_3\beta_2$) with the two helices on either side of β_5. The loop between α_2 and β_4 is disordered in the structure and is not glycine-rich, a conserved feature of the Cas7 family [56]. Significant insertions are located between each of the four β-strands; these form two distinct regions above and below the β-sheet to form a crescent-shaped molecule (Figure 5B). Residues located in the cleft of SsoCas7 have been implicated in binding crRNA [13]. The structure of *e*Cascade shows that the *E. coli* Cas7 adopts a similar topology to SsoCas7 and that the cleft forms the extended groove along the helical assembly of Cas7 [39]. Given the likely ubiquitous nature of the Cas7 backbone, it is probable that all Cascade complexes bind crRNA in the same manner.

Non-catalytic variants of Cas5

Although Cas5c possesses catalytic activity, the other members of the Cas5 family are non-catalytic and are limited to structural roles. In both *a*Cascade and *e*Cascade, Cas5 interacts stably with Cas7 [13,39]. Cas5e also interacts with Cse1 and Cse2 in *e*Cascade and appears to help stabilize the protospacer-bound conformation of the complex [39]. *c*Cascade contains two copies of Cas5c, which appear to occupy the positions of Cas5 and Cas6e in *e*Cascade [14,39]. Cas5c from *Streptococcus pyogenes* and *X. oryzae* bind dsDNA, which could be mimicking target dsDNA or the heteroduplex of the interference R-loop [52]. Therefore Cas5c seems to be able to function as both a catalytic Cas6 equivalent and a structural Cas5 equivalent.

Of the Cascade complexes, only *d*Cascade and *f*Cascade do not contain Cas5 [12]. On the basis of secondary structure predictions, Makarova et al. [56] predicted that Csc1 (I-D) and Csy2 (I-F) belong to the Cas5 family. EM images and the small-angle X-ray scattering (SAXS) structure of *f*Cascade place Csy2 in a similar position to the structural Cas5s of *c*Cascade and *e*Cascade [14,15,39]. However, the fragmentation patterns of *e*Cascade and *f*Cascade suggest that Csy2 does not interact with Csy3 (probable Cas7 equivalent) in the same manner as Cas5 and Cas7 from *e*Cascade, leading van Duijn et al. [62] to conclude that *f*Cascade does not contain a Cas5 equivalent. Further data are required to settle the relationships between the complexes.

The small subunits of the interference complexes

Several of the interference complexes contain so-called 'small' subunits, which are typically <200 residues. These proteins are Csa5 (I-A), Cse2 (I-E), Csm2 (III-A) and Cmr5 (III-B) and it has been hypothesized that these proteins belong to a single family (Cas11) [56]. Analysis of the structures of Csa5 [63], Cse2 [64,65] and Cmr5 [66] (PDB codes 2OEB and 4GKF) shows that,

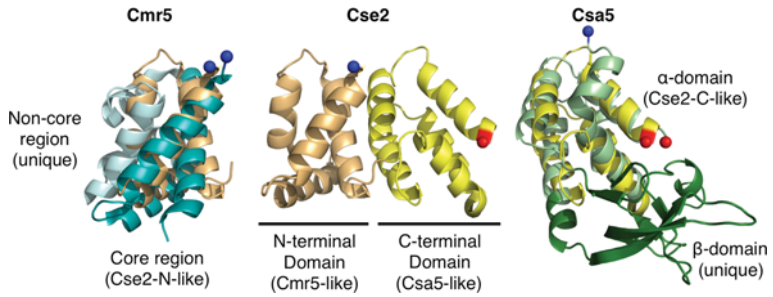

Figure 6 The small subunits of interference complexes

Comparison of *T. thermophilus* Cmr5 (PDB code 2ZOP, left), *T. thermophilus* Cse2 (PDB code 2ZCA, middle) and *S. solfataricus* Csa5 (PDB code 3ZC4, right). The N-terminal domain of Cse2 (light orange) is superimposed on Cmr5 (blue) and the C-terminal domain of Cse2 (yellow) is superimposed on Csa5 (green).

although structural homology can be detected, the evolutionary links between the proteins are complex. Cse2 contains N- and C-terminal domains that consist of four and five α-helices respectively. The N-terminal domain is homologous with the core structure of Cmr5, whereas the C-terminal domain is homologous with one of the domains of Csa5 (Figure 6). Csa5 consists of an α-helical domain (homologous with the Cse2 C-terminal domain) and a β-sheet domain that is not homologous with Cse2 or Cmr5. In fact, this domain is very poorly conserved across the Csa5 family and is likely to vary significantly between homologues.

Possible evolutionary scenarios for the homology include fusion of *csa5* and *cmr5* genes to form *cse2* or the evolution of the three proteins from a single *cse2*-like gene with domain loss to form Csa5 and Cmr5 [63]. Csm2, the remaining small subunit for which there is no structure available, may be critical for determining the likely scenario, although it is certainly possible that Csm2 may not possess any homology with the other small subunits. Makarova et al. [56] suggested that the Cas8 C-terminal domain, which is predicted to be helical, might be homologous with the small subunits, although no experimental structure exists to confirm this.

The Cse2 dimer is an integral part of *e*Cascade [39] and is responsible for stabilizing the R-loop, increasing the affinity of *e*Cascade for dsDNA approximately 10-fold [67]. Cse2 alone binds non-specifically to dsDNA and ssRNA [65]. Conversely, the *S. solfataricus* Csa5 does not stably interact with Cas5/Cas7 in the presence of crRNA or with nucleic acids alone [63]. Cmr5, in contrast with both Csa5 and Cse2, appears to be non-essential to the function of the CMR complex [11]. Thus we conclude that the similarity of the small subunits is structural rather than functional.

The large subunits of the interference complexes

Similarly to the small subunits, each of the type I and III interference complexes contains a 'large' (>500 residues) subunit: Cas8 (I-A, I-B, I-C), Cse1 (I-E), Csy1 (I-F) and Cas10 (I-D, III-A and III-B). Cas10 was originally predicted to be a polymerase (hence the name polymerase cassette for the III-B subtype) on the basis of sequence features typical of a palm domain commonly found in polymerases and cyclases [44]. Subsequently it was proposed that all of the large subunits were homologous and part of a Cas10 superfamily [56]. However, recent structures of a type III-B Cas10 [68,69], denoted Cas10b, show that, although the prediction of the palm domain was correct (albeit more akin to cyclases), no significant structural homology exists with Cse1 [70,71] (PDB codes 4H3T and 4EJ3). This argues against a single common ancestor for all of the large subunits.

Cas10, the large subunit of type III systems

Cas10 is the defining protein of the type III system and consists of an N-terminal HD (histidine–aspartate) phosphohydrolase domain (for which there is no structure) and a C-terminal region (Cas10^{dHD}) that contains the palm domain [56]. Cas10b^{dHD} from *P. furiosus* consists of two adenylate cyclase-like domains (denoted D1 and D3) and two α-helical domains (D2 and D4) (Figures 7A and 7B) [68,69]. D2 is not significantly homologous with known structures, but D4 is structurally homologous with Cmr5 and the N-terminal domain of Cse2, although sequence conservation is minimal and the biological implications of the homology are unclear. A typical adenylate cyclase domain consists of a ferredoxin-like fold with a C-terminal $\alpha_3\beta_5\alpha_4\beta_6\beta_7$ modification, which creates a seven-stranded β-sheet with the two additional helices located on either side of the sheet [72]. D1 and D3 lack some of these key structural elements: D3 lacks α_4 and β_6, whereas D1 lacks every additional element bar α_3. Individually, D1 and D3 are most similar to the type III adenylate cyclase from *Mycobacterium tuberculosis* [72]. However, these bacterial cyclases are typically homodimers, whereas D1 and D3 of Cas10b^{dHD} exist as a pseudoheterodimer more similar to the arrangement of mammalian cyclases [73]. The orientation between D1 and D3 is markedly different to that of typical cyclases which, combined with the loss of key structural and sequence features, is consistent with PfuCas10b^{dHD} lacking a cyclase-like catalytic activity, although D3 retains the ability to bind ADP [68].

In the CMR complex Cas10b interacts with Cmr3, an interaction observed in both *S. solfataricus* and *P. furiosus* [20,74]. The structure of the *P. furiosus* Cas10b^{dHD}–Cmr3 complex shows that the two proteins form a heterodimer with the interface formed by D1 of Cas10b^{dHD} and one face of Cmr3 (see below) [74]. At the interface between the two proteins is a highly positively charged cleft ~50 Å in length, which is suggestive of a role in crRNA binding. The nucleotide bound by D3 in both the Cas10b^{dHD} and Cas10b^{dHD}–Cmr3 structures lies at the centre of this cleft and so could be mimicking crRNA binding by the complex rather than substrate binding by the 'cyclase' domains of Cas10b^{dHD}. This is consistent with the nucleotide binding in a different orientation to that observed in cyclases.

If the Cas10b–Cmr3 complex does bind to part of the crRNA, the remainder of the crRNA must be bound by other subunits of the CMR complex. Three subunits of the complex (Cmr1, Cmr4 and Cmr6) are RAMPs and thus are plausible candidates. Makarova et al. [56] have predicted Cmr4 and Cmr6 to be Cas7 homologues. However, EM structures of the CMR complex (which targets ssRNA and not dsDNA) show that it is more compact than Cascade and lacks a central helical structure [20].

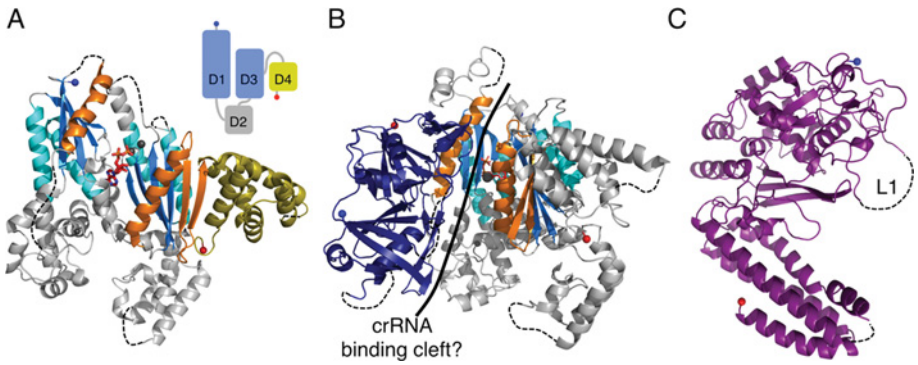

Figure 7 The large subunits of interference complexes

(**A**) The structure of PfuCas10b^{dHD} (PDB code 3UNG) in complex with ADP (red sticks). The ferredoxin-like folds are coloured as for RAMPs and the additional adenylate cyclase elements are shown in orange. D4 is shown in yellow to highlight its homology with the small subunits. The three metal ions are shown as grey spheres. Inset: schematic diagram showing the relative positions of the four domains (D1–D4) with the cyclase-like domains in blue and the small subunit-like domain in yellow. (**B**) The structure of the Cas10b^{dHD}–Cmr3 complex (PDB code 4H4K) with Cmr3 shown in navy blue and Cas10b^{dHD} as in (**A**). The putative crRNA-binding cleft is indicated with a solid black line. (**C**) The structure of Cse1 from *T. thermophilus* (PDB code 4AN8) with the disordered loop L1 indicated.

Cse1, the PAM (protospacer adjacent motif) sensor of *e*Cascade

The structures of Cse1 from *T. thermophilus* [70,71] (PDB code 4EJ3) and *Acidimicrobium ferrooxidans* (PDB code 4H3T) consist of an N-terminal mixed α/β domain with a novel fold and a C-terminal four-helix bundle (Figure 7C). In *e*Cascade, Cse1 is responsible for recognition of the PAM, a short (2–5 nt) conserved sequence located immediately next to the protospacer that is required for interference [75]. Cascade recognizes a PAM located 5′ to the protospacer [75] and, at least for *e*Cascade, PAM recognition uses the complementary strand [76]. Target dsDNA lacking a PAM is bound weakly by *e*Cascade [76,77] and is resistant to cleavage [78], consistent with the observation that mutations in the PAM can prevent interference [15,79].

The N-terminal domain of Cse1 contains a loop (L1, Figure 7C) that is disordered in all of the available crystal structures, but is critical for PAM recognition [70,71]. Analysis of the *e*Cascade structures led Mulepati et al. [70] and Sashital et al. [71] to suggest that L1 binds to the crRNA 5′-handle and PAM in the absence and presence of target DNA respectively. Cse1 is also critical for binding to negatively supercoiled dsDNA, both specifically to a protospacer and also non-specifically, a function that is dependent on the L1 loop [53,70,71]. Sashital et al. [71] have proposed that Cse1 scans dsDNA for PAM sequences and once in contact destabilizes the duplex to allow for target recognition, first through a 5′ seed sequence and then along the remainder of the target.

Other Cascade complexes lack Cse1 and must use a different protein for PAM sensing, although their identities have not been established. Cas8 and Csy1 are candidates as they dissociate easily from their respective complexes (similar to Cse1 and *e*Cascade) and EM images suggest that they are located in a similar position to Cse1 within their complexes [14,39,62].

Cmr3, a type III-B Cas6-like protein

Cmr3 is a RAMP protein of the CMR complex and the structure of PfuCmr3, available only in complex with Cas10b^{dHD}, shows that it contains two RAMP domains arranged in a similar manner to Cas6 (compare Figure 8 with Figure 3B) [74]. The C-terminal domain contains two of the conserved features of Cas6: the $\beta_{2'}$–$\beta_{3'}$ hairpin and the glycine-rich loop, both of which adopt similar conformations to those seen in Cas6 proteins. The Cmr3 glycine-rich loop also exhibits a similar consensus sequence to that of Cas6 (XXXXXGφG, where φ is an aromatic residue, X is any residue and the variable region contains at least one positively charged residue) (Supplementary Figure S3 at http://www.biochemj.org/bj/453/bj4530155add.htm). In the N-terminal domain, a β-strand located after α_2 forms a β-hairpin with β_4, as is also seen in the *Pyrococcus* and *Sulfolobus* Cas6 homologues [46–50], with the turn of the hairpin containing the two conserved glycine residues identified by Makarova et al. [56] as an N-terminal glycine-rich loop. The tip of this loop is disordered, but since it is only three residues in length it acts more as a turn rather than the extended loop seen in many RAMPs.

Cmr3 exhibits two significant deviations from Cas6. $\alpha_{2'}$ is replaced by a short β-strand located immediately prior to the C-terminal glycine-rich loop, similar to the β-strand located before the N-terminal glycine-rich loop. The second difference is the presence of a significant structural insertion located between β_2 and β_3 of the N-terminal domain. This insertion consists of two short helices and seven β-strands and packs against the C-terminal β-sheet. The insertion and the $\beta_{2'}$–$\beta_{3'}$ hairpin together form the interface with Cas10b^{dHD} and line the putative crRNA-binding cleft.

THE INTERFERENCE NUCLEASES

During interference, invading nucleic acids detected by base pairing with crRNA are targeted for degradation by an interference nuclease. In type I systems this is the HD metal-dependent nuclease domain of Cas3, which is recruited to Cascade rather than being an integral component [76]. Type II systems use Cas9 as the sole interference protein with the HNH-like and RuvC-like nuclease domains cleaving the complementary and non-complementary strands of the R-loop respectively [16,80]. The interference nucleases of the type III systems are unknown. The nuclease is within the CMR complex, but Cas10b and Cmr5 have been discounted, as has the *Sulfolobales*-specific protein Cmr7 [11,20,68].

Cas3, the interference nuclease of type I systems

Cas3 is the defining protein of the type I system and consists of an N-terminal HD nuclease domain and a C-terminal superfamily II DExD/H-box helicase domain [12,44,81]. In some systems the

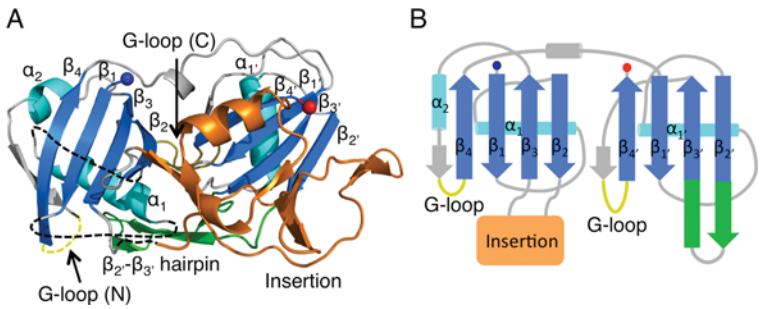

Figure 8 The structure of Cmr3

(**A**) The structure of PfuCmr3 (PDB code 4H4K) showing the RAMP elements and the structural insertion in the N-terminal domain (orange). (**B**) Topology diagram of PfuCmr3 highlighting the conserved RAMP features and the connectivity of the insertion domain.

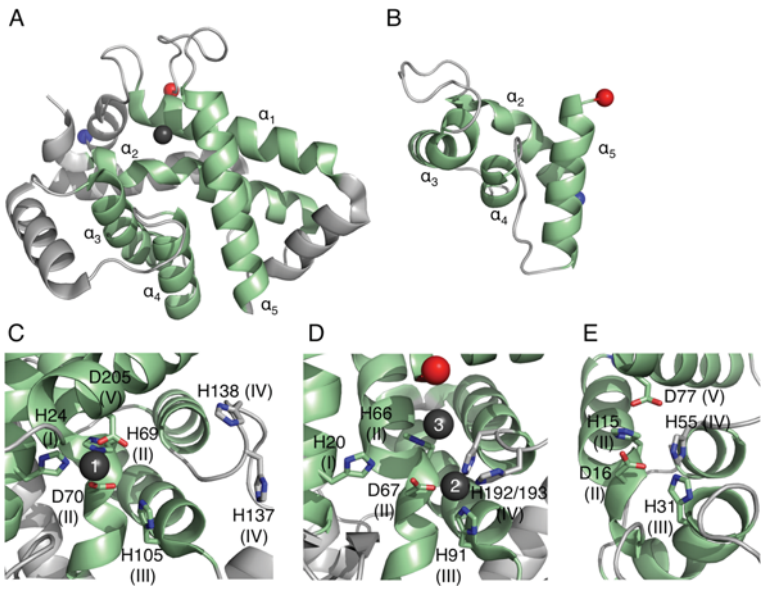

Figure 9 The structures of Cas protein HD domains

(**A**) The structure of TtCas3^{HD} (PDB code 3SKD) with the conserved HD superfamily helices in green and numbered. The Ni^{2+} ion is shown as a dark grey sphere. Residues 222–260 are not shown as they are predicted to belong to the helicase domain. (**B**) A homology model of the HD domain of Cas10a from *S. thermophilus* created using PHYRE2 and consisting of residues 4–79. The four HD domain helices are coloured in green and labelled. (**C**–**E**) Views of the active sites of (**C**) TtCas3^{HD}, (**D**) MjaCas3′′ (PDB code 3S4L) and (**E**) SthCas10a^{HD}. The HD superfamily motifs are shown as sticks with motif numbers in parentheses and the metal ions as grey spheres with site numbers in white.

two domains are expressed as separate proteins (Cas3′′ and Cas3′ respectively); other variations are also known, such as domain fusion to other Cas proteins (for example, Cas3–Cas2 in the I-F subtype and Cas3–Cse1 in some I-E systems) and inversion of the domain order (Figure 2) [12,44,76]. Cas3 is recruited by Cascade after R-loop formation where it catalyses the unwinding and degradation of the invading DNA [76,78].

Cas3 proteins contain all five HD superfamily sequence motifs (H-HD-H-H-D) and the structures of TtCas3^{HD} (HD domain of TtCas3) and MjaCas3′′ (*Methanocaldococcus jannaschii* Cas3′′) revealed eight conserved helices, five of which are characteristic of the HD superfamily (Figure 9A) [82,83]. In the TtCas3^{HD} structure a single Ni^{2+} ion is bound by motifs I, II and V (site 1), whereas site 2 (a binding site formed by motifs II, III and IV) remains unoccupied (Figure 9C). Metal binding at site 2 has been observed in a number of HD domains (for example, see PDB codes 2OGI, 2O08, 2PQ7, 3CCG and 3HC1) and its absence in the TtCas3^{HD} structure is likely to be a crystal artefact. The MjaCas3′′ structure shows a Ca^{2+} ion bound at site 2 as well as a second ion bound by the histidine of motif II (site 3) (Figure 9D). However, the binding at site 3 and the lack of binding at site 1 are likely to be artefacts resulting from the protein engineering required for crystallization.

Characterization of type I-E Cas3 nuclease domains from *T. thermophilus*, *Streptococcus thermophilus,* and *E. coli* and the type I-A Cas3′′ proteins from *M. jannaschii* and *P. furiosus* showed that they are all metal-dependent nucleases specific for ssDNA, although the Cas3′′ proteins also cleave ssRNA *in vitro* [82–84]. These proteins are both endo- and exo-nucleases, with the latter activity proceeding in the 3′→5′ direction. MjaCas3′′, SthCas3 (*Streptococcus thermophilus* Cas3) and EcoCas3 (*E. coli* Cas3) cleave R-loops, the biological substrate of Cas3 and MjaCas3′′ and SthCas3 have been shown to target the non-complementary ssDNA strand specifically [76,78,82]. Structural data is not available for the helicase domain of Cas3, but the type I-E helicase domains of SthCas3 and EcoCas3 catalyse the 3′→5′ Mg^{2+}- and ATP-dependent unwinding of dsDNA and DNA/RNA duplexes [84,85]. Nicking of the non-complementary strand by the HD domain followed by the unwinding of the DNA duplex by the helicase domain would allow for progressive degradation

of the non-complementary strand (Supplementary Figure S4 at http://www.biochemj.org/bj/453/bj4530155add.htm). The complementary strand is also targeted by Cas3 [78] and would occur after dissociation of DNA from the R-loop.

The HD domains of Cas10 proteins

Cas10 proteins contain N-terminal HD domains that are highly divergent from typical HD domains, being both shorter than classical HD proteins and lacking characteristic motifs (Supplementary Figure S5 at http://www.biochemj.org/bj/453/bj4530155add.htm) [86]. A homology model of Cas10a from *S. thermophilus* built using PHYRE2 [87] shows that motifs II, III and IV could co-ordinate a metal ion in a similar way to that of site 2 of Cas3 (Figures 9B and 9E). Therefore this domain could also be catalytically active and might potentially act as the interference nuclease of the CSM complex, although so far experimental confirmation is lacking. In contrast, Cas10b only contains motif II and so is unlikely to be an active nuclease, consistent with the observation that the Cas10b HD domain is not necessary for interference by the CMR complex [68], perhaps unsurprising since this complex targets RNA.

CONCLUDING REMARKS

The structural biology of the CRISPR system provides a wealth of information on the evolution and mechanisms of the proteins involved. It has revealed the underlying relationships between highly divergent proteins that are difficult or impossible to detect using bioinformatic approaches (however heroic) alone. The RAMP (or RAMP-like) domains, present in the Cas2, Cas5, Cas6, Cas7, Cas10 and Cmr3 families, are the leitmotif of the system, providing RNA-binding and -cleavage functionalities that are central to the process. The backbone of all type I complexes is likely to be a helical arrangement of Cas7, and a similar arrangement of Cas7-like RAMP subunits may be found in the CSM complex, given that it, too, targets dsDNA. Key challenges for crystallography include the structure of the Cas9 protein of type II systems, which has so far evaded attempts to place it in a wider context. Structures of the large and small subunits of the various type I and type III-A complexes are expected to clarify the relationships between the different families, and we can look forward to some simplification of the overall picture as these relationships become apparent. Finally, atomic level structural information on the $\sim$400 kDa CRISPR interference complexes remains a grand challenge in molecular biology, one that has been taken up enthusiastically by the structural biology community.

ACKNOWLEDGEMENTS

We thank past and present members of the Naismith and White laboratories for helpful discussions.

FUNDING

The laboratory is funded by grants from the Biotechnology and Biological Sciences Research Council (BBSRC) [grant numbers BB/G011400/1 and BB/K000314/1 (to M.F.W. and J.H.N.)] and a BBSRC-funded studentship to J.R.

REFERENCES

1 Mojica, F. J. M., Diez-Villasenor, C., Garcia-Martinez, J. and Soria, E. (2005) Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J. Mol. Evol. **60**, 174–182

2 Grissa, I., Vergnaud, G. and Pourcel, C. (2007) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinf. **8**, 172

3 Mojica, F. J. M., Diez-Villasenor, C., Soria, E. and Juez, G. (2000) Biological significance of a family of regularly spaced repeats in the genomes of Archaea, bacteria and mitochondria. Mol. Microbiol. **36**, 244–246

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

5 Pourcel, C., Salvignol, G. and Vergnaud, G. (2005) CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology **151**, 653–663

6 Bolotin, A., Ouinquis, B., Sorokin, A. and Ehrlich, S. D. (2005) Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology **151**, 2551–2561

7 Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A. and Horvath, P. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. Science **315**, 1709–1712

8 Brouns, S. J. J., Jore, M. M., Lundgren, M., Westra, E. R., Slijkhuis, R. J. H., Snijders, A. P. L., Dickman, M. J., Makarova, K. S., Koonin, E. V. and van der Oost, J. (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. Science **321**, 960–964

9 Pul, U., Wurm, R., Arslan, Z., Geissen, R., Hofmann, N. and Wagner, R. (2010) Identification and characterization of *E. coli* CRISPR-cas promoters and their silencing by H-NS. Mol. Microbiol. **75**, 1495–1512

10 Tang, T. H., Polacek, N., Zywicki, M., Huber, H., Brugger, K., Garrett, R., Bachellerie, J. P. and Huttenhofer, A. (2005) Identification of novel non-coding RNAs as potential antisense regulators in the archaeon *Sulfolobus solfataricus*. Mol. Microbiol. **55**, 469–481

11 Hale, C. R., Zhao, P., Olson, S., Duff, M. O., Graveley, B. R., Wells, L., Terns, R. M. and Terns, M. P. (2009) RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. Cell **139**, 945–956

12 Makarova, K. S., Haft, D. H., Barrangou, R., Brouns, S. J. J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F. J. M., Wolf, Y. I., Yakunin, A. F. et al. (2011) Evolution and classification of the CRISPR-Cas systems. Nat. Rev. Microbiol. **9**, 467–477

13 Lintner, N. G., Kerou, M., Brumfield, S. K., Graham, S., Liu, H. T., Naismith, J. H., Sdano, M., Peng, N., She, Q. X., Copie, V. 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**, 21643–21656

14 Nam, K. H., Haitjema, C., Liu, X., Ding, F., Wang, H., Delisa, M. P. and Ke, A. (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

15 Wiedenheft, B., van Duijn, E., Bultema, J., Waghmare, S., Zhou, K. H., Barendregt, A., Westphal, W., Heck, A., Boekema, E., Dickman, M. and Doudna, J. A. (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

16 Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A. and Charpentier, E. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science **337**, 816–821

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

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

19 Ivancic-Bace, I., Al Howard, J. and Bolt, E. L. (2012) Tuning in to interference: R-loops and Cascade complexes in CRISPR immunity. J. Mol. Biol. **422**, 607–616

20 Zhang, J., Rouillon, C., Kerou, M., Reeks, J., Brugger, K., Graham, S., Reimann, J., Cannone, G., Liu, H., Albers, S.-V. et al. (2012) Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. Mol. Cell **45**, 303–313

21 Deveau, H., Garneau, J. E. and Moineau, S. (2010) CRISPR/Cas system and its role in phage-bacteria interactions. Annu. Rev. Microbiol. **64**, 475–493

22 Fineran, P. C. and Charpentier, E. (2012) Memory of viral infections by CRISPR-Cas adaptive immune systems: acquisition of new information. Virology **434**, 202–209

23 Horvath, P. and Barrangou, R. (2010) CRISPR/Cas, the immune system of bacteria and archaea. Science **327**, 167–170

24 Marraffini, L. A. and Sontheimer, E. J. (2010) CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nat. Rev. Genet. **11**, 181–190

25 van der Oost, J., Jore, M. M., Westra, E. R., Lundgren, M. and Brouns, S. J. J. (2009) CRISPR-based adaptive and heritable immunity in prokaryotes. Trends Biochem. Sci. **34**, 401–407

26 Wiedenheft, B., Sternberg, S. H. and Doudna, J. A. (2012) RNA-guided genetic silencing systems in bacteria and archaea. Nature **482**, 331–338

27 Kwon, A.-R., Kim, J.-H., Park, S. J., Lee, K.-Y., Min, Y.-H., Im, H., Lee, I., Lee, K.-Y. and Lee, B.-J. (2012) Structural and biochemical characterization of HP0315 from *Helicobacter pylori* as a VapD protein with an endoribonuclease activity. Nucleic Acids Res. **40**, 4216–4228

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

29 Lillestol, R. K., Shah, S. A., Brugger, K., Redder, P., Phan, H., Christiansen, J. and Garrett, R. A. (2009) CRISPR families of the crenarchaeal genus *Sulfolobus*: bidirectional transcription and dynamic properties. Mol. Microbiol. **72**, 259–272

30 Garside, E. L., Schellenberg, M. J., Gesner, E. M., Bonanno, J. B., Sauder, J. M., Burley, S. K., Almo, S. C., Mehta, G. and Macmillan, A. M. (2012) Cas5d processes pre-crRNA and is a member of a larger family of CRISPR RNA endonucleases. RNA **18**, 2020–2028

31 Hatoum-Aslan, A., Maniv, I. and Marraffini, L. A. (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

32 Deltcheva, E., Chylinski, K., Sharma, C. M., Gonzales, K., Chao, Y. J., Pirzada, Z. A., Eckert, M. R., Vogel, J. and Charpentier, E. (2011) CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature **471**, 602–607

33 Haurwitz, R. E., Jinek, M., Wiedenheft, B., Zhou, K. H. and Doudna, J. A. (2010) Sequence- and structure-specific RNA processing by a CRISPR endonuclease. Science **329**, 1355–1358

34 Richter, H., Zoephel, J., Schermuly, J., Maticzka, D., Backofen, R. and Randau, L. (2012) Characterization of CRISPR RNA processing in *Clostridium thermocellum* and *Methanococcus maripaludis*. Nucleic Acids Res. **40**, 9887–9896

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

36 Kunin, V., Sorek, R. and Hugenholtz, P. (2007) Evolutionary conservation of sequence and secondary structures in CRISPR repeats. Genome Biol. **8**, R61

37 Sashital, D. G., Jinek, M. and Doudna, J. A. (2011) An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3. Nat. Struct. Mol. Biol. **18**, 680–687

38 Wang, R. Y., Preamplume, G., Terns, M. P., Terns, R. M. and Li, H. (2011) Interaction of the Cas6 riboendonuclease with CRISPR RNAs: recognition and cleavage. Structure **19**, 257–264

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

40 Sternberg, S. H., Haurwitz, R. E. and Doudna, J. A. (2012) Mechanism of substrate selection by a highly specific CRISPR endoribonuclease. RNA **18**, 661–672

41 Haurwitz, R. E., Sternberg, S. H. and Doudna, J. A. (2012) Csy4 relies on an unusual catalytic dyad to position and cleave CRISPR RNA. EMBO J. **31**, 2824–2832

42 Plagens, A., Tjaden, B., Hagemann, A., Randau, L. and Hensel, R. (2012) Characterization of the CRISPR/Cas subtype I-A system of the hyperthermophilic crenarchaeon *Thermoproteus tenax*. J. Bacteriol. **194**, 2491–2500

43 Wang, R. Y. and Li, H. (2012) The mysterious RAMP proteins and their roles in small RNA-based immunity. Protein Sci. **21**, 463–470

44 Makarova, K. S., Grishin, N. V., Shabalina, S. A., Wolf, Y. I. and Koonin, E. V. (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

45 Ebihara, A., Yao, M., Masui, R., Tanaka, I., Yokoyama, S. and Kuramitsu, S. (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

46 Carte, J., Wang, R. Y., Li, H., Terns, R. M. and Terns, M. P. (2008) Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. Genes Dev. **22**, 3489–3496

47 Park, H.-M., Shin, M., Sun, J., Kim, G. S., Lee, Y. C., Park, J.-H., Kim, B. Y. and Kim, J.-S. (2012) Crystal structure of a Cas6 paralogous protein from *Pyrococcus furiosus*. Proteins: Struct., Funct., Bioinf. **80**, 1895–1900

48 Wang, R., Zheng, H., Preamplume, G., Shao, Y. and Li, H. (2012) The impact of CRISPR repeat sequence on structures of a Cas6 protein-RNA complex. Protein Sci. **21**, 405–417

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

50 Reeks, J., Sokolowski, R. D., Graham, S., Liu, H., Naismith, J. H. and White, M. F. (2013) Structure of a dimeric crenarchaeal Cas6 enzyme with an atypical active site for CRISPR RNA processing. Biochem. J. **452**, 223–230

51 Makarova, K. S., Aravind, L., Grishin, N. V., Rogozin, I. B. and Koonin, E. V. (2002) A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. Nucleic Acids Res. **30**, 482–496

52 Koo, Y., Ka, D., Kim, E.-J., Suh, N. and Bae, E. (2013) Conservation and variability in the structure and function of the Cas5d endoribonuclease in the CRISPR-mediated microbial immune system. J. Mol. Biol., doi:10.1016/j.jmb.2013.02.032

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

54 Calvin, K. and Li, H. (2008) RNA-splicing endonuclease structure and function. Cell. Mol. Life Sci. **65**, 1176–1185

55 Gesner, E. M., Schellenberg, M. J., Garside, E. L., George, M. M. and MacMillan, A. M. (2011) Recognition and maturation of effector RNAs in a CRISPR interference pathway. Nat. Struct. Mol. Biol. **18**, 688–692

56 Makarova, K. S., Aravind, L., Wolf, Y. I. and Koonin, E. V. (2011) Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. Biol. Direct **6**, 38

57 Maris, C., Dominguez, C. and Allain, F. H. T. (2005) The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. FEBS J. **272**, 2118–2131

58 Clery, A., Blatter, M. and Allain, F. H. T. (2008) RNA recognition motifs: boring? Not quite. Curr. Opin. Struct. Biol. **18**, 290–298

59 Oubridge, C., Ito, N., Evans, P. R., Teo, C. H. and Nagai, K. (1994) Crystal-structure at 1.92 angstrom resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. Nature **372**, 432–438

60 Ding, J. Z., Hayashi, M. K., Zhang, Y., Manche, L., Krainer, A. R. and Xu, R. M. (1999) Crystal structure of the two-RRM domain of hnRNP A1 (UP1) complexed with single-stranded telomeric DNA. Genes Dev. **13**, 1102–1115

61 Lilley, D. M. J. (2003) The origins of RNA catalysis in ribozymes. Trends Biochem. Sci. **28**, 495–501

62 van Duijn, E., Barbu, I. M., Barendregt, A., Jore, M. M., Wiedenheft, B., Lundgren, M., Westra, E. R., Brouns, S. J. J., Doudna, J. A., van der Oost, J. and Heck, A. J. R. (2012) Native tandem and ion mobility mass spectrometry highlight structural and modular similarities in clustered-regularly-interspaced shot-palindromic-repeats (CRISPR)-associated protein complexes from *Escherichia coli* and *Pseudomonas aeruginosa*. Mol. Cell. Proteomics **11**, 1430–1441

63 Reeks, J., Graham, S., Anderson, L., Liu, H., White, M. F. and Naismith, J. H. (2013) Structure of the archaeal Cascade subunit Csa5: relating the small subunits of CRISPR effector complexes. RNA Biol. **10**, doi:10.4161/rna.23854

64 Agari, Y., Yokoyama, S., Kuramitsu, S. and Shinkai, A. (2008) X-ray crystal structure of a CRISPR-associated protein, Cse2, from *Thermus thermaphilus* HB8. Proteins: Struct., Funct., Bioinf. **73**, 1063–1067

65 Nam, K. H., Huang, Q. and Ke, A. (2012) Nucleic acid binding surface and dimer interface revealed by CRISPR-associated CasB protein structures. FEBS Lett. **586**, 3956–3961

66 Sakamoto, K., Agari, Y., Agari, K., Yokoyama, S., Kuramitsu, S. and Shinkai, A. (2009) X-ray crystal structure of a CRISPR-associated RAMP superfamily protein, Cmr5, from *Thermus thermophilus* HB8. Proteins: Struct., Funct., Bioinf. **75**, 528–532

67 Westra, E. R., Nilges, B., van Erp, P. B. G., van der Oost, J., Dame, R. T. and Brouns, S. J. J. (2012) Cascade-mediated binding and bending of negatively supercoiled DNA. RNA Biol. **9**, 1134–1138

68 Cocozaki, A. I., Ramia, N. F., Shao, Y., Hale, C. R., Terns, R. M., Terns, M. P. and Li, H. (2012) Structure of the Cmr2 subunit of the CRISPR-Cas RNA silencing complex. Structure **20**, 545–553

69 Zhu, X. and Ye, K. (2012) Crystal structure of Cmr2 suggests a nucleotide cyclase-related enzyme in type III CRISPR-Cas systems. FEBS Lett. **586**, 939–945

70 Mulepati, S., Orr, A. and Bailey, S. (2012) Crystal structure of the largest subunit of a bacterial RNA-guided immune complex and its role in DNA target binding. J. Biol. Chem. **287**, 22445–22449

71 Sashital, D. G., Wiedenheft, B. and Doudna, J. A. (2012) Mechanism of foreign DNA selection in a bacterial adaptive immune system. Mol. Cell **46**, 606–615

72 Sinha, S. C., Wetterer, M., Sprang, S. R., Schultz, J. E. and Linder, J. U. (2005) Origin of asymmetry in adenylyl cyclases: structures of *Mycobacterium tuberculosis* Rv1900c. EMBO J. **24**, 663–673

73 Linder, J. U. and Schultz, J. E. (2003) The class III adenylyl cyclases: multi-purpose signalling modules. Cell. Signalling **15**, 1081–1089

74 Shao, Y., Cocozaki, A. I., Ramia, N. F., Terns, R. M., Terns, M. P. and Li, H. (2013) Structure of the cmr2-cmr3 subcomplex of the cmr RNA silencing complex. Structure **21**, 376–384

75 Mojica, F. J. M., Diez-Villasenor, C., Garcia-Martinez, J. and Almendros, C. (2009) Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology **155**, 733–740

76 Westra, E. R., van Erp, P. B. G., Kunne, T., Wong, S. P., Staals, R. H. J., Seegers, C. L. C., Bollen, S., Jore, M. M., Semenova, E., Severinov, K. et al. (2012) CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. Mol. Cell **46**, 595–605

77 Semenova, E., Jore, M. M., Datsenko, K. A., Semenova, A., Westra, E. R., Wanner, B., van der Oost, J., Brouns, S. J. J. and Severinov, K. (2011) Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc. Natl. Acad. Sci. U.S.A. **108**, 10098–10103
78 Sinkunas, T., Gasiunas, G., Waghmare, S. P., Dickman, M. J., Barrangou, R., Horvath, P. and Siksnys, V. (2013) *In vitro* reconstitution of Cascade-mediated CRISPR immunity in *Streptococcus thermophilus*. EMBO J. **32**, 385–394
79 Deveau, H., Barrangou, R., Garneau, J. E., Labonte, J., Fremaux, C., Boyaval, P., Romero, D. A., Horvath, P. and Moineau, S. (2008) Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. J. Bacteriol. **190**, 1390–1400
80 Gasiunas, G., Barrangou, R., Horvath, P. and Siksnys, V. (2012) Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc. Natl. Acad. Sci. U.S.A. **109**, E2579–E2586
81 Haft, D. H., Selengut, J., Mongodin, E. F. and Nelson, K. E. (2005) A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. PLoS Comput. Biol. **1**, 474–483
82 Beloglazova, N., Petit, P., Flick, R., Brown, G., Savchenko, A. and Yakunin, A. F. (2011) Structure and activity of the Cas3 HD nuclease MJ0384, an effector enzyme of the CRISPR interference. EMBO J. **30**, 4616–4627
83 Mulepati, S. and Bailey, S. (2011) Structural and biochemical analysis of nuclease domain of clustered regularly interspaced short palindromic repeat (CRISPR)-associated protein 3 (Cas3). J. Biol. Chem. **286**, 31896–31903
84 Sinkunas, T., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P. and 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
85 Howard, J. A. L., Delmas, S., Ivancic-Bace, I. and Bolt, E. L. (2011) Helicase dissociation and annealing of RNA-DNA hybrids by *Escherichia coli* Cas3 protein. Biochem. J. **439**, 85–95
86 Galperin, M. Y. and Koonin, E. V. (2012) Divergence and convergence in enzyme evolution. J. Biol. Chem. **287**, 21–28
87 Kelley, L. A. and Sternberg, M. J. E. (2009) Protein structure prediction on the Web: a case study using the Phyre server. Nat. Protoc. **4**, 363–371

Received 4 March 2013/9 April 2013; accepted 17 April 2013
Published on the Internet 28 June 2013, doi:10.1042/BJ20130316

Biochem. J. (2013) **453**, 155–166 (Printed in Great Britain) doi:10.1042/BJ20130316

multimedia online data

SUPPLEMENTARY ONLINE DATA

CRISPR interference: a structural perspective

Judith REEKS, James H. NAISMITH[1] and Malcolm F. WHITE[1]

Biomedical Sciences Research Complex, University of St Andrews, St Andrews, Fife KY16 9ST, U.K.

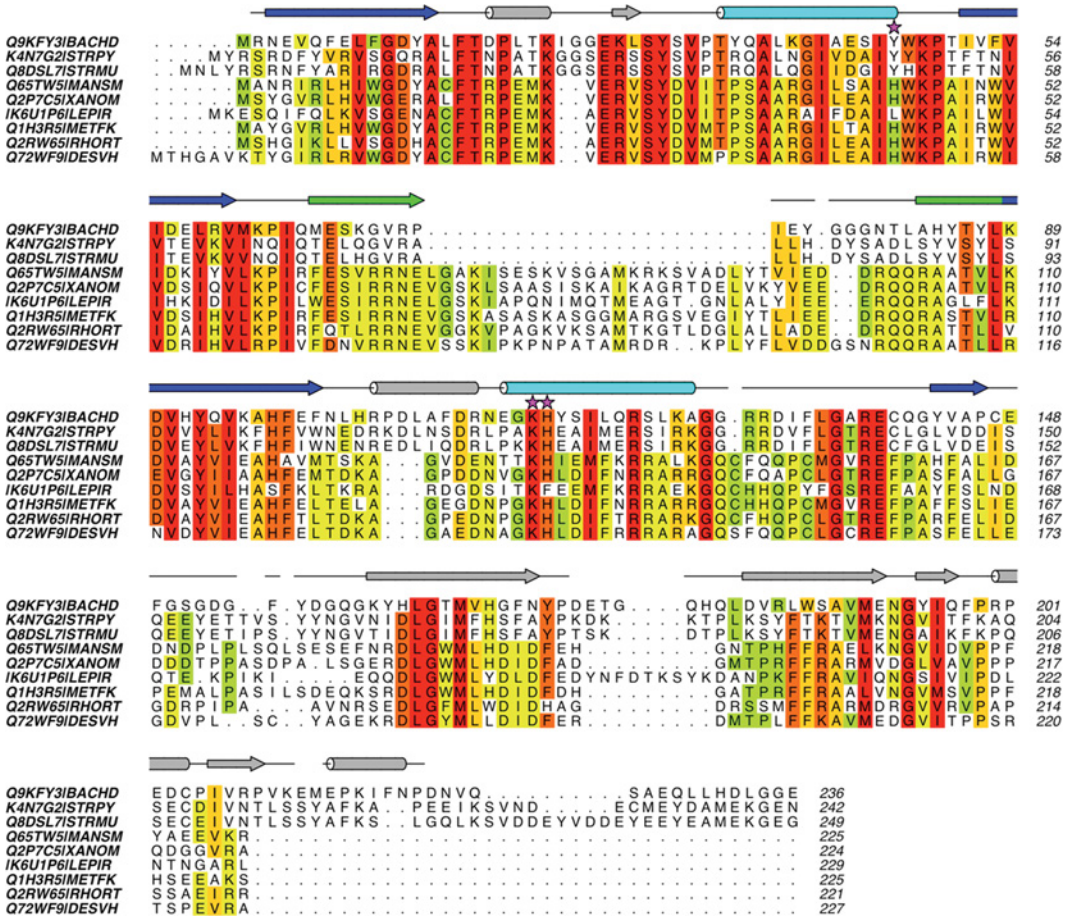

Figure S1 Sequence alignments of Cas5c proteins

Sequence similarity is shaded from red (highest) to green (lowest). The secondary structure elements of BhCas5c are shown above the alignments and are coloured according to Figure 3 of the main text. Gaps in the elements represent disordered residues. The catalytic residues of BhCas5c are indicated by magenta stars.

[1] Correspondence may be addressed to either of these authors (email naismith@st-and.ac.uk or mfw2@st-and.ac.uk).

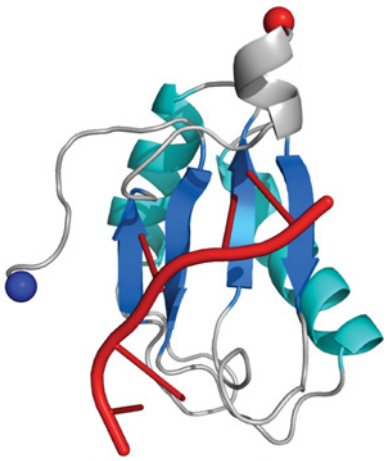

Figure S2 The structure of U1A spliceosomal protein, a typical RRM protein, in complex with RNA (red) (PDB code 1URN)

For clarity, the unbound RNA hairpin is not shown. The two RRM RNA-binding consensus sequences are shown beneath the structure. The RPM domain is coloured in the same manner as the RAMP domain in the main text.

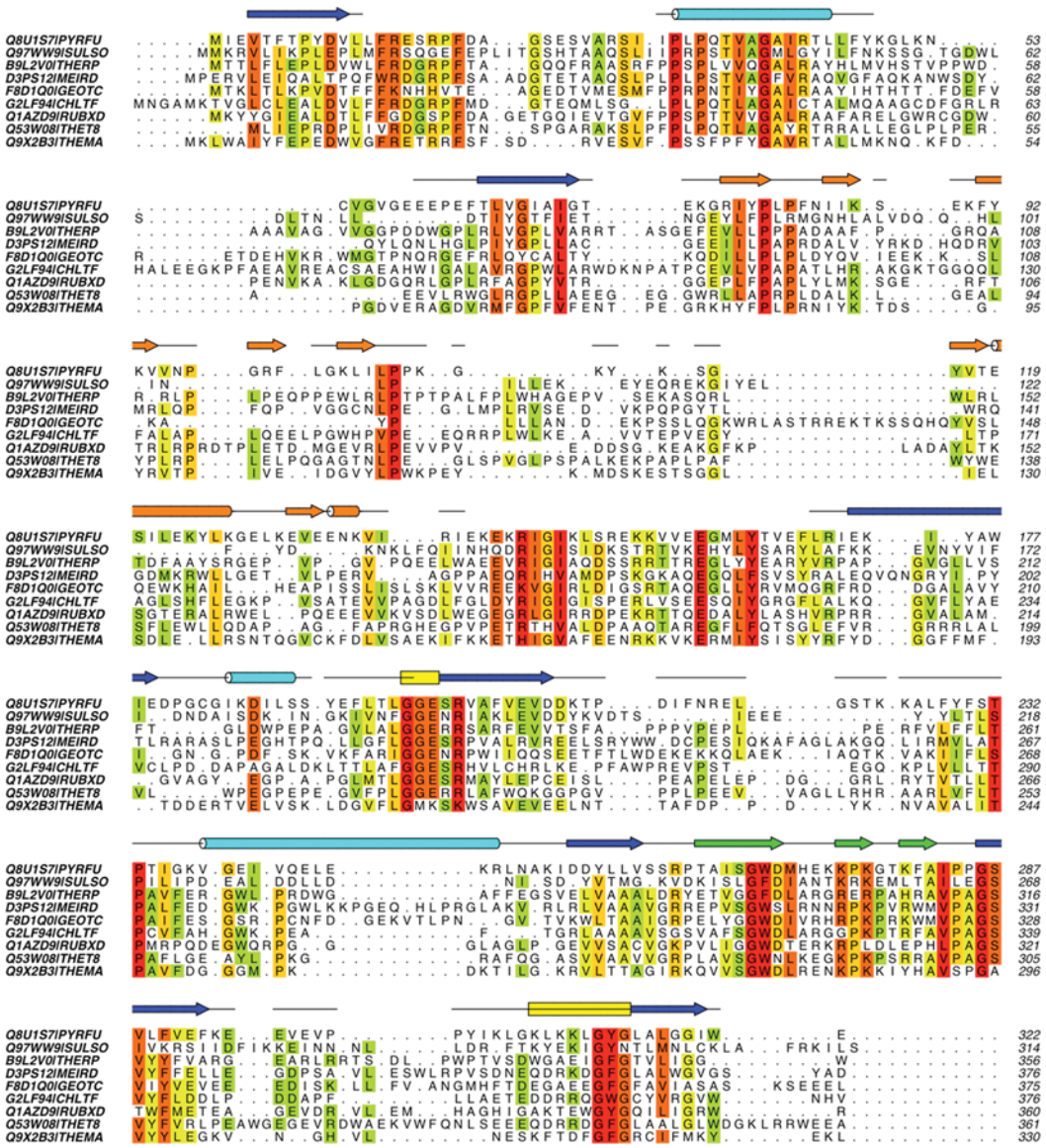

Figure S3 Sequence alignment of Cmr3 proteins

Sequence similarity is shaded from red (highest) to green (lowest). The secondary structure elements from PfuCmr3 are shown above and coloured according to Figure 8 of the main text.

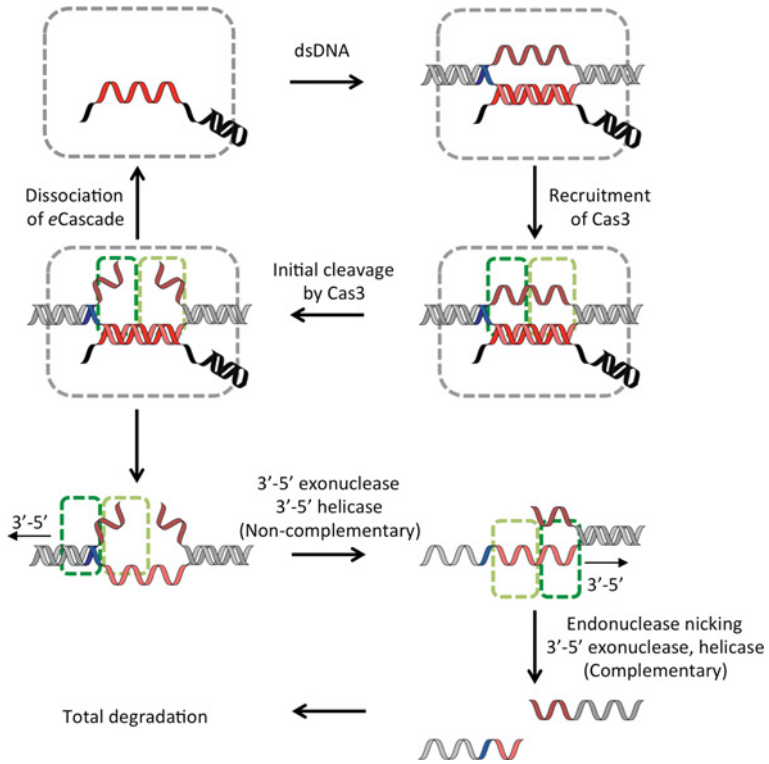

Figure S4 Schematic diagram of dsDNA degradation by Cas3 in the type I-E system

Repeats are shown in black, protospacers and spacers in red, PAMs in blue and DNA in grey. The Cas3 HD domain is represented by a light green dotted line, the Cas3 helicase domain in dark green and *e*Cascade by a grey line.

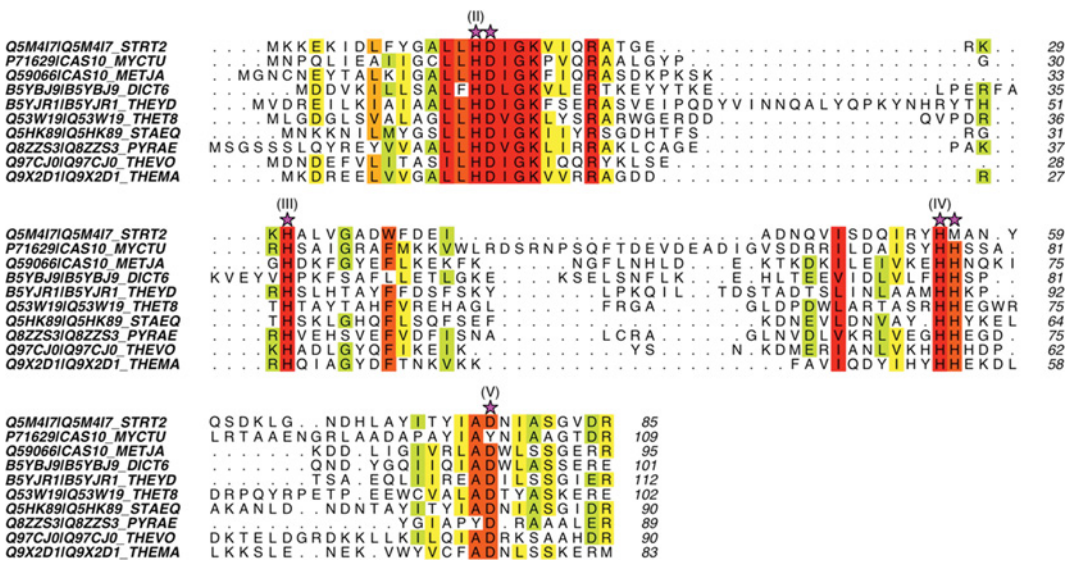

Figure S5 Sequence alignments of the HD domains of Cas10a

The putative HD superfamily sequence motifs are highlighted with magenta stars and the motif number is indicated.

Table S1 Details of all of the crystal structures of Cas proteins available at the time of writing

Protein	Organism	PDB code(s)	Notes
Cas1	*Aquifex aeolicus*	2YZS	
Cas1	*Escherichia coli*	3NKD, 3NKE	
Cas1	*Pseudomonas aeruginosa*	3GOD	
Cas1	*Pyrococcus horikoshii*	3PV9	
Cas1	*Thermotoga maritima*	3LFX	
Cas2	*Bacillus halodurans*	4ES1, 4ES2, 4ES3	
Cas2	*Desulfovibrio vulgaris*	3OQ2	
Cas2	*Pyrococcus furiosus*	2I0X	
Cas2	*S. solfataricus*	2IVY, 2I8E, 3EXC	Two paralogues
Cas2	*Thermus thermophilus*	1ZPW	
Cas3	*T. thermophilus*	3SK9, 3SKD	HD domain only
Cas3″	*Methanocaldococcus jannaschii*	3S4L	
Cas4	*S. solfataricus*	4IC1	
Cas5c	*Mannheimia succiniciproducens*	3KG4	
Cas5c	*Bacillus halodurans*	4F3M	
Cas5c	*Streptococcus pyogenes*	3VZH	
Cas5c	*Xanthomonas oryzae*	3VZI	
Cas6	*E. coli*	4DZD	
Cas6	*P. furiosus*	3I4H, 3PKM, 3UFC	Two paralogues
Cas6	*P. horikoshii*	3QJJ, 3QJL, 3QJP	
Cas6	*S. solfataricus*	3ZFV, 4ILL, 4ILM, 4ILR	Two paralogues
Cas6e	*T. thermophilus*	1WJ9, 2Y8W, 2Y8Y, 2Y9H, 3QRP, 3QRQ, 3QRR	
Cas6f	*Ps. aeruginosa*	2XLI, 2XLJ, 2XLK, 4AL5, 4AL6, 4AL7	
Cas7	*S. solfataricus*	3PS0	
Cas10b^{dHD}	*P. furiosus*	3UNG, 3UR3, 4DOZ, 4H4K	Lacking HD domain, also in complex with Cmr3
Csm6	*S. solfataricus*	3QYF	
Csn2	*Enterococcus faecalis*	3S5U	
Csn2	*Streptococcus agalactiae*	3QHQ	
Csn2	*S. pyogenes*	3TOC, 3V7F	
Csn2	*Streptococcus thermophilus*	3ZTH	
Cmr3	*P. furiosus*	4H4K	In complex with Cas10b^{dHD}
Cmr5	*Archaeglobus fulgidus*	2OEB	
Cmr5	*P. furiosus*	4GKF	
Cmr5	*T. thermophilus*	2ZOP	
Cmr7	*S. solfataricus*	2XVO, 2X5Q	Two paralogues, *Sulfolobales*-specific
Csa3	*S. solfataricus*	2WTE	
Csa5	*S. solfataricus*	3ZC4	
Cse1	*Acidimicrobium ferrooxidans*	4H3T	
Cse1	*T. thermophilus*	4AN8, 4F3E, 4EJ3	
Cse2	*Thermobifida fusca*	4H79	
Cse2	*T. thermophilus*	2ZCA, 4H7A	
Csn2	*Streptococcus thermophilus*	3ZTH	
Csx1	*P. furiosus*	4EOG	
Csx1	*S. solfataricus*	2I71	

Received 4 March 2013/9 April 2013; accepted 17 April 2013
Published on the Internet 28 June 2013, doi:10.1042/BJ20130316