Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig-Maximilians-Universität München

The Backbone of Prokaryotic Adaptive Immunity: the Cas7 Protein Family

Ajla Hrle

aus
Zagreb, Kroatien

2014

Erklärung

Diese Dissertation wurde im Sinne von § 7 der Promotionsordnung vom 28. November 2011 von Frau Professor Dr. Elena Conti betreut.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, den 18.01.2015

Ajla Hrle

Dissertation eingereicht am 01.09.2014

1. Gutachter: Prof. Dr. Elena Conti

2. Gutachter: Prof. Dr. Anita Marchfelder

Mündliche Prüfung am 30.10.2014

Abstract

CRISPR/Cas is the prokaryotic adaptive immune response to viral invasion. Its mechanism is reminiscent of the eukaryotic RNA interference. The host actively incorporates short sequences from invading genetic elements (viruses or plasmids) into a region of its genome that is characterized by clustered regularly interspaced short palindromic repeats (CRISPRs) and a number of CRISPR-associated (*cas*) genes. The molecular memory of previous infections can be transcribed and processed into small RNAs (crRNAs) that guide a multiprotein–nucleic acid interference complex to recognize and cleave incoming foreign genetic material. Three pathways (I, II, III) are defined by their protein machinery and target specificity (DNA vs. RNA). In types I and III, the main protagonist of the interference complex is the Cas7 protein. Up to six copies of Cas7 constitute the complex's main building block that assembles around the crRNA and provides a platform for protein interactions and target binding.

During my PhD work, I solved the crystal structures of two Cas7 orthologs from different archaeal species, at 1.8 Å for *Thermofilum pendens (Tp)* Csc2 and at 2.37 Å for *Methanopyrus kandleri (Mk)* Csm3. The crystal structures of *Mk* Csm3 and *Tp* Csc2 were solved by experimental phasing and revealed a core RRM-like domain with a β_1-α_1-β_2-β_3-α_2-β_4 arrangement of secondary structure elements. The core is flanked by three peripheral domains that are defined by insertions within the core. Structural superposition of the RRM-like core domains of *Mk* Csm3 and *Tp* Csc2 with the representatives of other Cas families (5/6/7) revealed the highest homology beyond the RRM with a Cas7 family homolog. Thus I showed that Cas7 family proteins share equivalent insertions, forming homologous peripheral domains.

Using the information obtained from structural data, I investigated the RNA binding properties *Mk* Csm3, *Tp* Csc2 and a Cas7 protein from subtype I-A, *Thermoproteus tenax (Tt)* Csa2. All orthologs bound RNA in a sequence-independent manner, according to their physiological function of spacer binding. Furthermore, a combined approach consisting of mutation analysis, UV-based protein–RNA crosslinking, mass spectrometry and fluorescence anisotropy mapped the RNA interacting regions to two structurally highly conserved positively charged surfaces.

Taken together, this thesis describes a comprehensive structural study of the Cas7 family, defining the family's structural features. These structural data from single proteins and the mapped RNA binding interfaces agree with protein–RNA interactions observed in the *Escherichia coli* interference complex.

Contents

Abbreviations

cAMP	cyclic AMP
cas	CRISPR-associated
Cascade	CRISPR-associated complex for antiviral defense
Cmr	CRISPR module RAMP
CRISPR	clustered regularly interspaced short palindromic repeats
crRNA	CRISPR guide RNA
crRNP	CRISPR ribonucleoprotein particle
Csc	CRISPR/Cas subtype cyano (based on Cyanobacteria)
Csm	CRISPR/Cas subtype Mtube (based on *M. tuberculosis*)
Ec	*Escherichia coli*
Ec CasC	Cas7 homolog of *Escherichia coli*
EM	electron microscopy
Mk	*Methanopyrus kandleri*
Mk Csc3	Cas7 homolog of *Methanopyrus kandleri*
MS	mass spectrometry
PAM	protospacer adjacent motif
Pf	*Pyrococcus furiosus*
pre-crRNA	precursor CRISPR RNA
protospacers	new spacers on the target DNA
RAMP	repeat associated mysterious protein

RAMP	repeat-associated mysterious proteins
RMSD	root mean square deviation
RNAi	RNA interference
RNP	ribobucleoprotein particles
RRM	RNA recognition motif
sgRNA	single guide RNA
Ss	*Sulfolobus solfataricus*
Ss Csa2	Cas7 homolog of *Sulfolobus sulfataricus*
TEM	transmission electron microscopy
Tp	*Thermofilum pendens*
Tp Csa2	Cas7 homolog of *Thermofilum pendens*
tracrRNA	trans-activating crRNA
Tt	*Thermoproteus tenax*
Tt Csa2	Cas7 homolog of *Thermoproteus tenax*
UV	ultraviolet

Preface

Two decades after scientists first proposed that prokaryotes possess an adaptive immune system, we stand before a young yet quickly evolving field of research. Bacteria and archaea actively incorporate phage DNA or RNA into their genomes within clustered regularly interspaced short palindromic repeats (CRISPRs). These loci form a molecular memory of previous infection that is used to transcribe a guide RNA (crRNA) that can target and silence foreign genetic elements. Three CRISPR pathways have been identified in different organisms, varying by the nature of their targets and their protein machinery. The heart of type I and III CRISPR responses is the crRNA-guided interference complex, composed of up to 11 proteins. It distinguishes self from non-self DNA and specifically targets foreign nucleic acid sequences for degradation via its associated nuclease activity.

In the work that has led to this doctoral thesis, I structurally and functionally characterized Cas proteins belonging to the Cas7 family. These proteins constitute the core of the interference complexes, interacting with all other subunits and the guide RNA. Cas7 proteins achieve this by undergoing versatile structural rearrangements of secondary structural elements in peripheral domains. I focused on Cas7 homologs from three different subtypes and species, applying a combined approach of X-ray crystallography as well as biochemical, biophysical and mass spectrometric methods to comprehensively classify and functionally elucidate Cas7 family features. This led to three publications that will be discussed in the cumulative thesis.

The first chapter is a general introduction to the biological background of the CRISPR pathways. The second chapter includes my results in form of three original manuscripts, all introduced with a short summary. The first two manuscripts constitute my main PhD work. For this thesis, I added unpublished follow-up data to the third manuscript. The last chapters feature a comprehensive discussion that integrates my findings with new insights from recent studies published after my research papers.

1 Introduction

1.1 Phages and prokaryotes: instigators of evolution

Prokaryotic viruses (phages) are the most abundant biological entities on earth [1]. They outnumber their hosts, bacteria and archaea, by an order of magnitude. Phage abundance as well as their genetic and morphological diversity has made these viruses a major evolutionary driving force for bacterial and archaeal communities alike, influencing nearly every ecological setting [2, 3]. Many viruses insert their own genetic information into the host's genome [4]. This horizontal gene transfer is a double-edged sword: on the one hand, viruses can shuttle antibiotic resistance genes between bacteria; on the other hand, viruses exploit their host's resources and may eventually induce cell lysis and death. Consequently, bacteria and archaea have evolved a plethora of innate, multilayered defense strategies against invading genetic elements (Fig. 1) [5]. These include (1) the blockage of phage adsorption by masking or mutation of surface receptors, which prevents attachment and virus entry [6]; (2) the methylation of certain DNA sequence motifs in the host genomes in combination with the expression of corresponding restriction enzymes, which exposes non-self DNA to nucleolytic digestion while endogenous DNA is protected[7]; (3) abortive infection, which induces apoptosis and sacrifices the infected cell to protect other cells of the same species [8].

Each of these well-studied pathways acts at different stages of infection and has evolved species-specific protein machineries [11]. All of these strategies recognize the invader independent of previous encounters. Historically, our appreciation of the microbial immune system has been largely restricted to such innate immune response mechanisms. Conversely, the ability to establish a molecular memory of previous pathogen encounters and to elicit an adaptive immune response upon re-infection was considered a hallmark of higher organisms. Recently, however, a nucleic acid-based prokaryotic adaptive immune system was discovered [12, 13] .

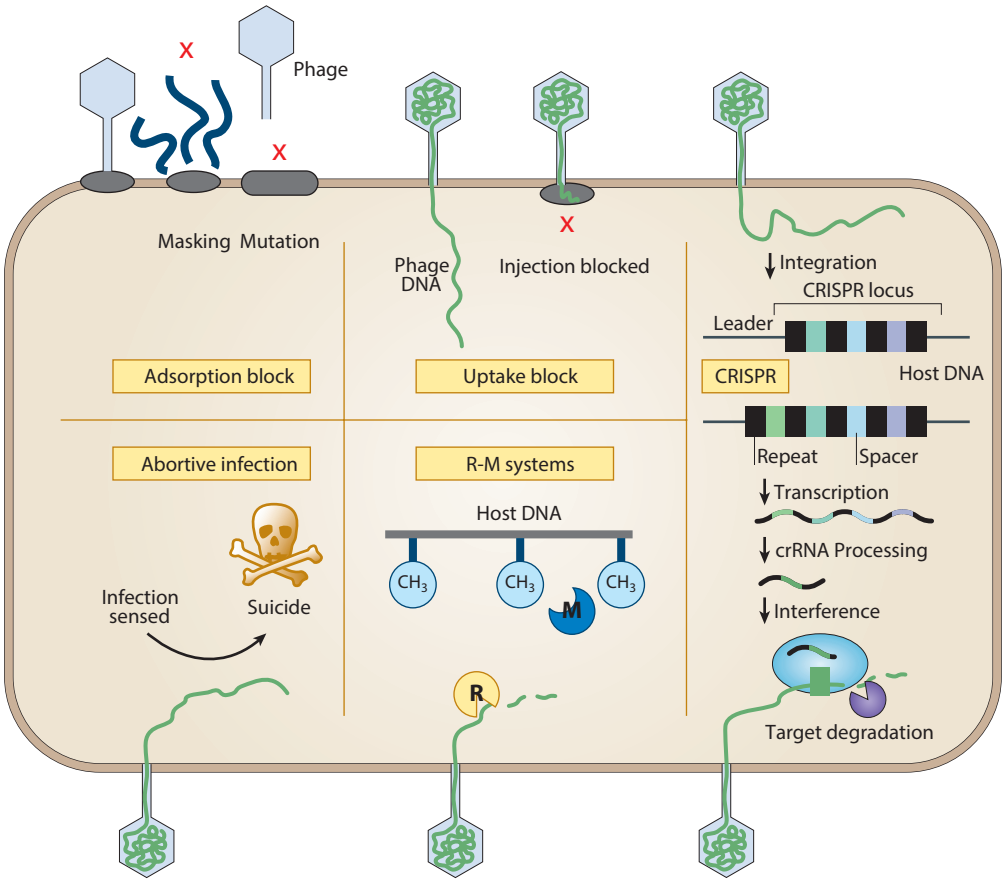


Figure 1: Prokaryotic defense strategies.
Several defense strategies, innate and adaptive, can act independently or in conjunction to prevent phage invasion. Adsorption blockage is achieved by masking the cell surface via post-translational modification, by high mutation rates of the cell surface receptors or blocking phage injection (top left panels). In case of successful phage entry, the invading DNA is directly targeted by the host's restriction enzymes (R); methylation (M, methyltransferase) of host DNA motives prevents self cleavage (mid-bottom panel). A second option is the CRISPR pathway (right panel). The foreign genomic elements are incorporated into the host genome, recalled after a new invasion and used to target complementary sequences for degradation using an elaborate CRISPR machinery (C). The last resort is abortive infection, which sacrifices the cell in benefit of the population, before virus replication can occur (left-bottom panel). Adapted from refs. [9, 10].

1.2 The CRISPR response in a nutshell

Almost all archaea and half of the bacteria were found to contain clustered regularly interspaced short palindromic repeats (CRISPR), in which mobile genetic elements, such as sequences derived from phage DNA/RNA are incorporated (Fig. 1, right panel) [14, 15] . This process requires an arsenal of CRISPR associated (Cas) proteins and establishes a molecular memory in the form of a genomic structure where variable, invader-derived stretches ('spacers') alternate with constant, host-specific repeat sequences. Upon re-infection, the stored information can be accessed in order to destroy invading nucleic acids. To that end, the locus is transcribed and the transcript is processed into a guide RNA with the help of another set of proteins. Finally, an interference complex assembles around the guide RNA. This complex recognizes and targets sequences complementary to the guide for nucleolytic degradation – an approach reminiscent of the eukaryotic RNA inference pathway.

1.3 The CRISPR locus

The machinery required for the CRISPR-mediated immune response is encoded by one contiguous sequence in the prokaryotic genome. The defining feature of a CRISPR locus, the repeat-spacer-repeat pattern (Fig. 2), was first discovered by sequencing a chromosomal fragment of *Escherichia coli* [16]. Two decades later, a vast number of CRISPR arrays in numerous species have been identified and the sheer complexity of the system has become apparent.

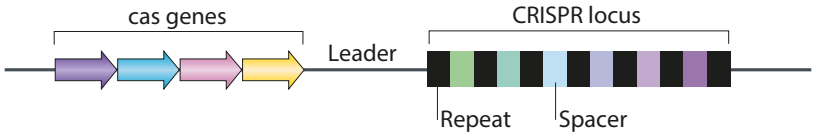


Figure 2: The CRISPR locus.
An A/T-rich leader sequence is followed by a series of repeats (black rectangles), separated by variable spacer sequences derived from invading genetic elements (green–purple). CRISPR-associated (*cas*) genes encode the protein machinery. Adapted from ref. [10].

A combination of computational and molecular biology approaches have shown that all CRISPR loci share a common design and are composed of four universally present elements [17, 18]. The first striking feature is a series of short sequences, termed repeats, ranging from 20–50 base pairs with a conserved sequence at the 3′ end: GAAAN, implicated in protein-binding [19]. The repeats of one locus are almost always identical with respect to size and sequence. However, repeats of different loci vary in sequence, length and secondary structure of their transcripts. Comprehensive studies of prokaryotic CRISPR arrays have classified repeats based on their sequence and found that most bacterial repeats were palindromic, whereas most archaeal repeats were not [19]. The second feature is determined by non-identical spacer sequences of similar length that separate the repeats. Analyses of bacterial, archaeal and viral genome sequences have led to the understanding that the variable spacer elements are virus-derived and confer resistance to the corresponding viruses [20, 21]. The third element is an adenine/thymine (A/T)-rich sequence of approximately 100–500 base pairs that flanks the CRIPSR locus and is termed the leader [14]. Initial observations indicated that new spacers are inserted near the leader sequence. More detailed analyses showed that the leader contains promoter elements and is a binding site for putative regulatory proteins, controlling expression and spacer acquisition [22, 23]. The number of the repeat-spacer-repeat clusters and their individual length is variable within a CRISPR array. Most species contain multiple unrelated loci in their genomes. The total length of each locus spans hundreds to several thousand base pairs and depends solely on previous exposure of the archaea or bacteria to the vast diversity of mobile genetic elements.

1.4 The CRISPR-associated protein machinery

A variable cassette of so called *CRISPR-associated* (*cas*) genes forms the final building block of the CRISPR locus (Fig. 2). The *cas* genes lie adjacent to the CRIPSR array and encode all proteins that are necessary for mediating the adaptive immune response (Fig. 3) [24]. *Cas* genes exhibit an exceptional degree of variation and add to the complexity of the system [25]. Twenty-five *cas* gene products have been defined to date, of which six are generally conserved (*cas1–6*) and only two (*cas1* and *cas2*) are present in all CRISPR loci [14, 26]. *Cas* genes encode a large group of proteins with functions ranging from nucleolytic or helicase enzymatic activity to unique RNA binding properties (Table 1).

Table 1: Overview of major Cas protein families

Cas protein family	Subtype	Name	RRM-like domain	Function
Cas1	I, II, III	Cas1	No	Adaptation
Cas2	I, II, III	Cas2	Yes	Adaptation
Cas4	I, II		No	Adaptation
Csn2	II	Csn2	No	Adaptation
Cas6	I, III	Cas6	Yes	Processing: metal-independent ribonuclease
Cas7	I, III	Subtype-specific nomenclature I-A: Csa2 I-D: Csc2 I-B/C/E: Cas7 I-F:Csy3 III-A: Csm3 III-B: Cmr4	Yes	Interference: backbone of the crRNP complex. Target Cleavage (RNA): catalytic Cas7-like protein Cmr4
Large subunit	I, III	I: Cas8 (I-D: Cas10) III: Cas10 (Csm1/ Cmr2)	Yes	Interference: interaction with Cas7/5, capping of 5′ crRNA end, PAM recognition
Small subunit	I, III	I-A: Csa5 I-E: Cse2 III-A:Csm2 III-B:Cmr5	No	Interference: interaction with Cas7
Cas5	I, III	I:Cas5(a-f) III-A:Csm4 III-B:Cmr3	Yes	Interference: (non-catalytic homologs) interaction with Cas7/8/10. Processing: (catalytic homolog) (I-C) Cas6 substitute
Cas3	I	Cas3 (helicase) Cas3HD	No	Target cleavage (DNA): superfamily 2 helicase and HD-nuclease domain
Cas9	II	Target recognition and nuclease lobes with two active sites (NHN, RuvD)	No	Processing, Interference, Target cleavage (DNA)

These RNA-binding proteins are commonly known as repeat-associated mysterious proteins (RAMPs) [24]. They harbor one or more domains that are reminiscent of the RNA recognition motif (RRM, also known as the ferridoxin-like fold) a structural motif that is ubiquitously found to be involved in nucleic acid interactions in many protein classes [27, 28]. In the case of RAMPs, the RRM-like β_1-α_1-β_2-β_3-α_2-β_4 topology is interrupted by various secondary structure elements. A common feature of all RAMP RRM-like domains is a conserved glycine-rich loop between α_1 and β_2, which has been implicated in RNA binding [29]. Conversely, they have lost the conserved consensus sequences on β_1 and β_3, which often mediate protein–RNA binding in most ribonucleoprotein particles (RNPs) [27]. Other loops and peripheral domains vary between different RAMPs and serve as a basis for classification [30].

1.4.1 CRISPR-associated protein classification

Comprehensive phylogenetic, computational and structural studies have defined three distinct CRISPR systems, each characterized by the presence of a signature gene [30]:

- type I: Cas3, a target-degrading nuclease/helicase.
- type II: Cas9, an RNA-binding and target DNA-degrading nuclease.
- type III: Cas10, a large protein, whose function remains elusive.

The presence and organization of *cas* genes within the operon further define a total of 10 subtypes (I-A–F, II-A–C, III-A/B) [30]. Complementary ways of classifying the different CRISPR machineries are based on features of the repeat sequences themselves or on functional characteristics:

- type I: targets DNA using the 'Cascade' interference complex.
- type II: targets DNA using a single protein, Cas9.
- type III-A: targets DNA using the Csm interference complex.
- type III-B: targets RNA using the Cmr interference complex.

An overview of subtype classification and predicted as well as experimentally determined Cas protein functions is given in Table 1 and Fig. 3.

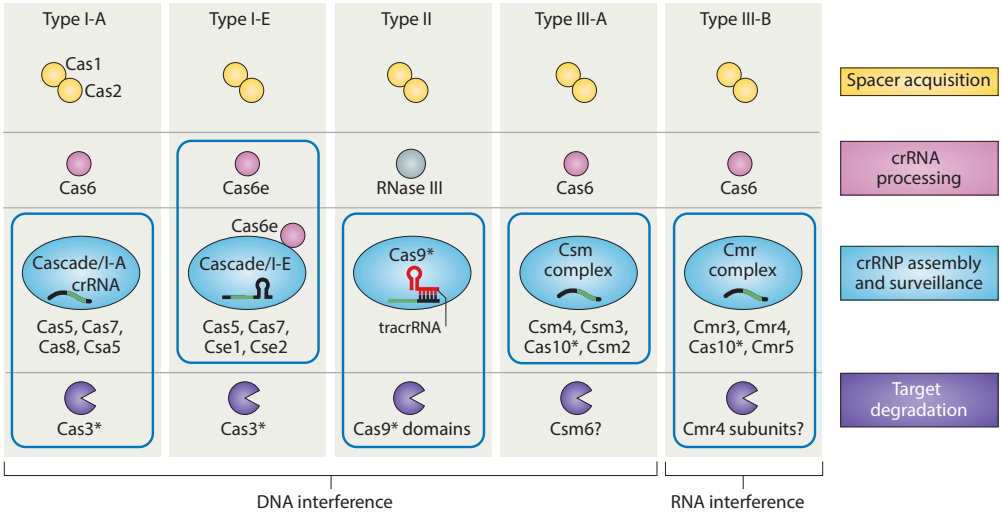


Figure 3: Cas systems.
Based on the presence of a signature Cas gene (indicated by an asterisk) the Cas proteins can be divided into three general types, which coincide with their functional purpose (Type I/II/III-A DNA targeting, III-B RNA targeting). These are further divided into distinct sub-types (a representative selection shown in the five panels). While type I and type III systems contain multiple subunits, the type II system contains a minimalistic set of proteins. Boxes highlight the variation of potential components of the crRNP complexes for each system. Adapted from ref. [10].

1.5 Three steps towards survival

The CRISPR/Cas-mediated immune response pathway can be divided into three general phases, all of which are mediated by single Cas proteins or multiple Cas protein-containing complexes (Fig. 4):

- Acquisition of new spacers: Nucleases and recombinases acquire and incorporate new spacers into the host genome between the repeat sequences.
- crRNA biogenesis and processing: CRISPR RNAs (crRNAs) are transcribed from the repeat and spacer regions and processed into crRNA guides by ribonucleases.
- Target interference and degradation: multi-protein complexes involving RNA-binding proteins assemble around the crRNA guides. They recognize, capture and finally degrade DNA or RNA targets with the help of nucleases.

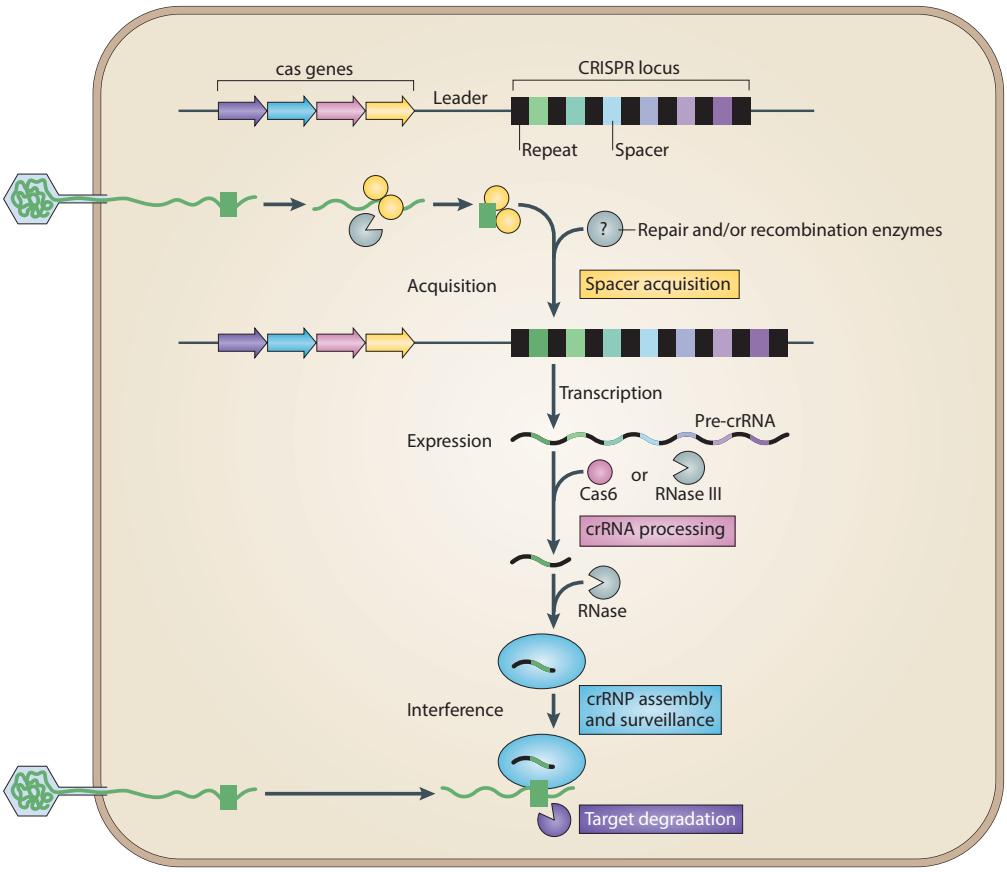


Figure 4: The CRISPR pathway.
Cas genes located adjacent to the repeat-spacer sequences encode a protein machinery (Cas proteins), which administrate the general phases of the immune response. *Cas* genes, Cas proteins as well as other involved proteins (grey) are colored based on their functional contribution to spacer acquisition (yellow), crRNA biogenesis processing (pink), crRNP assembly and target binding (blue), and degradation (purple). In the initial step, invading foreign DNA is recognized and processed fragments (protospacers – green box) are integrated between a repeat of the CRISPR locus. Spacer acquisition is mediated by the universally conserved proteins Cas1 and Cas2 (yellow circles) and has been linked to DNA repair/recombination enzymes (grey circle). As a response to new invasion, the second phase of the CRISPR pathway is initiated: the expression of long primary crRNA transcripts. These are then endonucleolytically processed into mature crRNAs that serve as a template for the assembly of the crRNP. In the final interference step, the fully assembled complex detects the complementary target and initiates its nucleolytic degradation. Adapted from ref. [10]

1.5.1 Step 1: Acquisition of new spacers

It is widely appreciated that the selection of new spacers from the target DNA (protospacers) is driven by a type-specific short sequence of 2–3 nucleotides in length, collectively known as protospacer adjacent motifs (PAMs) [31, 32]. (Fig. 5). Despite being the subject of extensive studies, the detailed molecular understanding of spacer recognition, fragmentation and insertion remain incomplete. In type I systems, these are located at the 3′ end of the protospacer, whereas in type II systems they are located at the 5′ end of the protospacer. Type III systems lack this principle of selection [33]. Upon recognition, protospacers are processed to a defined length and inserted in a PAM-dependent orientation [34]. The place of insertion is located within the first repeat, closest to the leader region [13]. During the process, the leader end repeat is nicked on opposite ends of the two strands, before the protospacer is inserted. After ligation and gap filling, this results in the duplication of the first repeat [35]. This mechanism of leader-directed insertion explains why the order of spacers reflects the chronology of infection. Moreover, initial spacer insertion accelerates subsequent uptake: a positive feedback loop, known as priming, that enhances resistance [36].

'Self' versus 'non-self' discrimination is central to this system, as it prevents cytotoxic autoimmune responses. Occasional insertion of chromosomal (self) DNA fragments have been reported in combination with modified PAMs or systems in which Cas proteins responsible for target recognition are absent [37, 38]. This observation suggests that the ability to distinguish between self and non-self DNA may not only lie within the acquisition but also at the final target interference step.

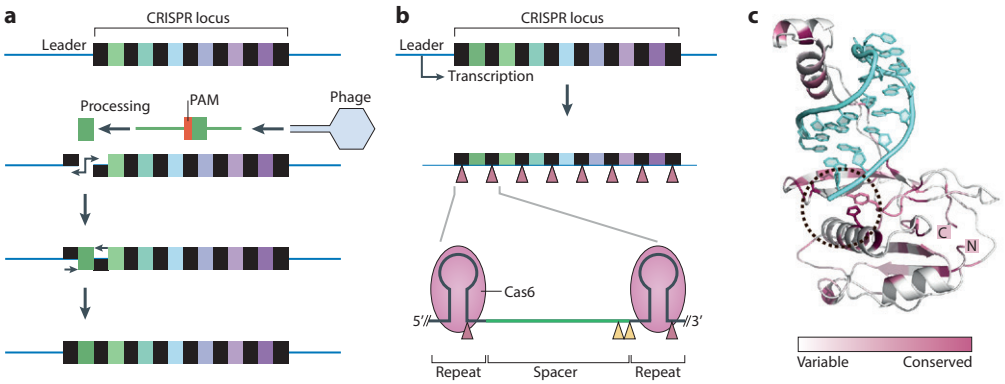


Figure 5: Spacer acquisition, biogenesis and processing.
a. In types I and II, protospacers (green) from the invasive genetic material, in this case phage DNA, are recognized via a short three-nucleotide protospacer adjacent motif (PAM – red). The protospacer is further processed and the leader end repeat is opened in order to allow spacer integration. During this process the initial repeat is duplicated. **b.** The transcription of the CRISPR array produces a long pre-CRISPR RNA (pre-crRNA), which is primarily processed within the repeat sequences (pink triangles). In type I and III systems, this step is catalyzed by the endoribonuclease Cas6 (pink circle). The product is a crRNAs in which spacers are flanked by repeat-derived handles: a 5′ handle of 8 nucleotides, and a longer 3′ handle. The 3′ handle is either a stable hair-pin structure or, if unstructured, it is subject to additional processing (yellow triangles) by unknown ribonucleases. **c.** Crystal structure of a Cas6 family protein in complex with its crRNA substrate. *Thermus thermophilus* ribonucleases Cas6e (from the type I-E system; PDB code 4AL7) binds the stem-loop of the crRNA (blue). The active site, which contains a conserved histidine residue shown as a stick, is encircled. Adapted from ref. [10].

Cas1/2: hallmark proteins of CRISPR

The highly conserved proteins Cas1 and Cas2 together are fundamental for the adaptation phase [35]. Overexpression of both proteins promotes additional spacer insertion, and overexpression of either one leads to a decrease in spacer acquisition. The presence of Cas1 and Cas2 in all subtypes suggests that this mechanism is universally conserved. The tight functional interplay between these two proteins is supported by species-specific fusion of the *cas1*/*cas2* genes and underlined by the recently published crystal structure of the *E. coli* Cas1-Cas2 multi-protein complex [39, 40]. The structural study demonstrates that an intact complex is essential for spacer acquisition *in vivo*. However it is Cas1 that is required for initial processing and spacer integration, suggesting a non-enzymatic role of Cas2 in type I-E. Structural characterization of these two proteins (single and in complex) from different species has shown that they are homodimeric metal-dependent nucleases [40–44]. Cas1 folds into an amino-terminal β-strand domain via which it dimerizes, and a carboxy-terminal α-helical domain which harbors the metal-binding site and residues crucial for DNA cleavage. Cas2 harbors an

RRM-like fold with a typical β_1-α_1-β_2-β_3-α_2-β_4 arrangement. Variable loop regions potentially point to the different substrate preferences and subsequent activity (endoribonuclease, deoxyribonucleases) [29]. The β-sheets from two Cas2 proteins form a conserved metal-binding dimerization interface. However these achievements have not yet shed light on the molecular mechanism of spacer acquisition. Future investigation of accessory factors (RecBCD and RuvB), additional type-specific Cas proteins (Csn2, Cas4, Csa1 and Cas3) and cases of gene fusions (Cas4–Cas1 and Cas1–Cas2, Cas2–Cas3) will contribute to completing the picture [39, 45, 46].

1.5.2 Step 2: crRNA biogenesis and processing

Upon re-infection, the saved information is utilized and the transcription of the repeat-spacer-repeat regions is initiated via the regulatory region within the leader (see Box 1).

The resulting long transcripts, termed precursor CRISPR RNAs (pre-crRNAs), are then endonucleolytically processed in one or more steps by subtype-specific Cas proteins, yielding the mature crRNA [47, 48]. The latter contains the full spacer sequence and elements from the repeat, which can then be recognized by the downstream protein machinery.

Box 1 Expression Regulation.

One explanation for why this unique prokaryotic adaptive immune system remained enigmatic to scientists for such a long time is found in *E. coli*. In this well-studied model organism, the CRISPR machinery is transcriptionally tightly regulated, leading to constitutive, but very low expression levels. Although the details of expression regulation and subtype-specific transcription remain unclear, stress-dependent factors have been observed such as the disturbance of the cell surface as well as simultaneous up-regulation along with restriction and modification proteins and the accumulation of CRISPR transcripts by phage exposure. Factors which possibility control gene expression are small molecules such as cyclic AMP (cAMP) together with putative transcriptional factors such as Csa3 and Cxs1 (contain dinucleotide binding properties).

Cas6 processing

In both types I and III, the endonuclease Cas6 is responsible for the primary processing of the pre-crRNA [49, 50]. It is structurally organized around one or two opposing RRM-like domains, and exhibits strong variation of the additional secondary structure elements [29]. These contribute to RNA binding as well as the positioning of the active site – a consequence of secondary structure diversity of its substrate, the repeat [48]. Common to all Cas6 family proteins is the composition of catalytic residues. A highly conserved histidine joined by a tyrosine or serine and a lysine residue catalyze the nucleophilic attack of a 2′-hydroxyl of a target ribose residue, resulting in cleavage of a single phosphodiester bond within the repeat. The processed crRNAs typically possess a repeat-derived 5′ handle of 8 nucleotides with a free 5′ OH, followed by the spacer and a repeat-derived 3′ handle of variable size with either a 3′ phosphate or a cyclic 2′-3′ phosphate [48]. This handle forms a hairpin structure in some systems [50]. In those systems, Cas6 family proteins (type I-E/type I-F) often exhibit a high affinity towards the hairpin structure of the repeat and stay associated with the product, laying the foundation of the crRNP complex [51].

Maturation

In types I-E and I-F, the processed crRNA is the template for crRNP complex formation [51–53]. In most systems, however, further trimming of the RNA from the 3′-end is required, resulting in a mature crRNA. Interestingly, in type III systems, this 3′-end trimming process removes all repeat-derived nucleotides [22, 33]. In addition, trimming may optionally continue to remove 6 spacer-derived nucleotides, resulting in two distinct crRNA species (ruler mechanism) [54, 55].

The type II CRISPR system combines the basic molecular mechanisms of crRNA processing and later target cleavage. However it predominantly functions with completely different protein machinery and pathway (Box 2).

Box 2 Subtype II.

Selective pressure has driven the evolutionary diversification of the protein machinery [37]. In contrast to the closely related pathways and players of the types I and III, type II possesses the most unique CRISPR locus. Two major differences in the locus architecture distinguish this subtype from the rest: first, a gene that encodes a so-called transactivating crRNA (tracrRNA), consisting of a 25 nucleotide anti-repeat sequence and region which folds into distinct secondary structure elements required for Cas9 recognition [56]. Second, only 3–4 *cas* genes are present, of which 2–3 encode proteins of the adaptation machinery (Cas1, Cas2, Csn2/Cas4) [30]. The largest gene, *cas9*, encodes a multifunctional protein possessing two endonucleolytic activities [57]. This protein alone combines functionalities that are distributed over many individual proteins in other subtypes.

In type II CRISPR-Cas systems, pre-crRNA transcripts hybridize with complementary sequences of Cas9-bound tracrRNA and are initially processed by RNase III [56]. PAM-dependent recognition of the foreign nucleic acid sequence guides the mature Cas9:tracrRNA:crRNA complex to the target [58, 59]. Cas9 then excites its double endonuclease activity and cleaves the complementary strand with its HNH domain and non-complementary strand with its RuvC-like domain [57]. Recently the crystal structures of apo-Cas9 and target-bound Cas9 complex have revealed the extensive nucleic acid interaction interfaces and a unique bi-lobed architecture – one lobe responsible for RNA and target binding the other harboring the nucleases and the PAM interacting domain. [60, 61].

1.5.3 Step 3: Target interference and degradation

The protein machinery of the crRNP interference complexes: commonalities and differences

Upon initial Cas6 processing, one or more copies of four Cas proteins assemble around the guide crRNA [62]. Recent structural and native mass spectrometric studies of types I-A/C/E/F and III-A/B of the multi-subunit crRNP complexes have shown that the basic building blocks are universally present (Fig. 6) [51, 63–67]. Despite subtype-specific variations of crRNP protein composition, a common set of proteins defines the overall architecture of the interference complexes. Central and common to all complexes is the helical arrangement of multiple monomers of Cas7 family proteins around the crRNA [51, 63]. Structural studies of type I-E show how a string of Cas7 monomers cradle the RNA within a central cleft [51, 68]. Atomic-resolution data of single proteins along-side bioinformatics studies have defined the common structural denominators of this protein family: a central RRM-like fold flanked by three distinct peripheral domains, which together define the central ribonucleotide binding cleft and flexible domain, synergistically mediating non-canonical RNA binding [29]. EM studies prove that the length of the guide RNA and further subunits determine the extent of oligomerization and stabilize the Cas7 backbone (ranging from four to six copies in III-B to seven copies in I-E). At the 5′ end, Cas7 interacts with the large subunit protein (in type I a Cas8 homolog; in type III a Cas10 homolog) and a non-catalytic Cas5 family protein [68, 69]. Together these proteins structurally cap the 5′-end of the crRNP complex. Next, small subunits coat the helical backbone [68, 69]. These structurally homologous proteins are helical domains and are subtype-specifically single proteins (Csa5, Cse2 in type I and Csm2, Cmr5 in type III systems) or extensions of the larger subunits (small helical domain of Cmr2, predicted helical C-terminal domain of Cas8 homologs) [29, 30]. Electron microscopy and protein-crosslinking studies of type III-A and -B complexes picture how these small subunit proteins entwine the Cas7-like oligomer (Cmr4) from foot (larger subunit Cmr2/Csm1 and Cas5-like proteins Cmr3/Csm4) to head [65, 69]. Beyond this general organization, prokaryotes have developed a vast phylogenetic divergence amongst the subtypes, which is reflected in the low sequence conservation and structural homology within a protein family [30, 70]. This feature is a consequence of the different functional requirements of the respective system: the nature of the target (DNA/RNA) and its cleavage.

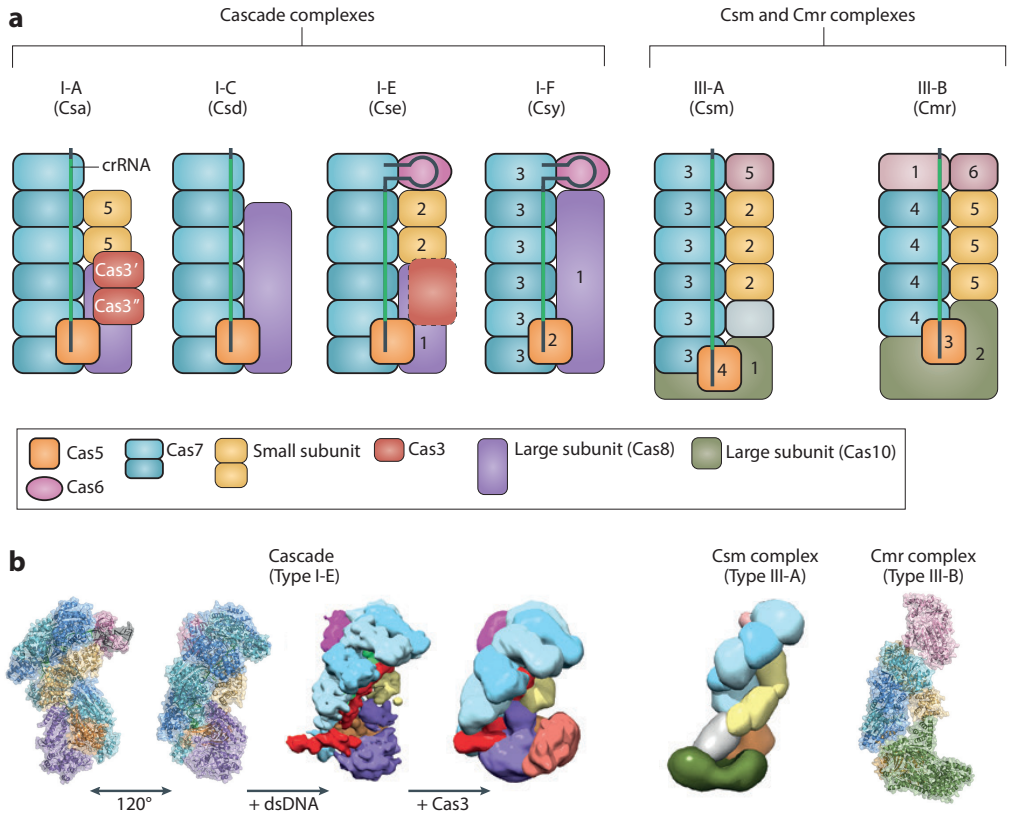


Figure 6: Overview of the crRNP interference complexes.
a. schematic representation of the subunit composition of different type I and III interference complexes. The crRNA spacer is colored green, the flanking repeats black. Conserved Cas protein families as well as the large and small subunits are colored according to their homology. The specific name of the sub-type is written in brackets and numbers refer to protein names typically used for individual subunits of each subtype (e. g. subunit 3 of III-A (Csm) refers to Csm3). In type I-A, Cas3 nuclease and helicase domains have been suggested to be part of the complex, in type I-E the Cas3 HD domain is fused with the large subunits (highlighted by a dashed margin). Protein subunits with an RNA-recognition motif (RRM)-like fold are outlined boldly. The overview illustrates the essential role of Cas7 proteins in the complex. For type III-A Csm complex an additional small Cas7 homologue has been proposed (gray).
b. Comparison of selected crystal structures of of types I and III. Left: crystal structure of the type I-E complex from *E. coli*: two views of the crRNA-bound complex (120° rotation; PDB 1VY8), bound to the dsDNA target (EMDB accession 5314) and with additional Cas3 (EMDB 5929, 5930). Right: Cryo-EM structure of *Sulfolobus solfataricus* III-A Csm (EMDB 2420) and pseudo-atomic model of *Pyrococcus furiosus* III-B Cmr complex (Benda *et al.*, 2014, [69]). Adapted from ref. [10] and updated with the most recent available structures.

Interference and degradation mechanisms

The complementarity of the crRNA spacer to the invading nucleic acids provides the molecular basis for successful target capture and subsequent degradation. In addition, the immune response requires subtype-specific Cas proteins that facilitate sequence recognition, stabilize the hybridization and expose the target to nucleolytic cleavage.

Interference

In type I systems, two factors are responsible for target binding: First, PAM sequence recognition not only plays a role for the integration machinery, but the same motif ensures specificity of target recognition [71]. As the sequence is not complementary to crRNA, protein interactions mediate PAM recognition [72]. Chemical probing and EM studies have connected proteins of the crRNP complex such as *E. coli* Cse1 to PAM scanning [72]. In type III systems, which work PAM-independently, the role is attributed to the left-over 5′ repeat derived handle. Base-pairing of this with chromosomal host DNA inhibits the interference step and prevents self-destruction [33]. How type III complexes facilitate target binding beyond the general non-specific binding events and thermal diffusion remains elusive and will be better understood as soon as high resolution structural data with the target become available. The second factor that favors hybridization in type I and III-A complexes is the so-called seed sequence [51, 73]. This stands for the first 8 nucleotides (1–5, 7–8) of the spacer. Single mutations in the seed sequence escape the machinery, while mutations in other regions of the spacer will still bind with high affinity [51, 71].

Degradation

In Type I pathways, correct base pairing induces the recruitment of Cas3 and crRNA guided ATP-independent strand unwinding of the dsDNA, which results in R-loop formation [74]. Interestingly, biochemical data show that base pairing of crRNA and target DNA occur every 5–6 helical segments, with short non-helical segments up to three bases in between [51, 68]. This binding pattern is reminiscent of the DNA-DNA interaction mediated by the RecA protein during homologous recombination [75]. In addition, hybridization coincides with conformational changes within the complex, visualized by electron microscopic studies of apo and target-bound complexes [51, 76]. This conformational change is believed to induce the recruitment of a Cas3 supposedly in close proximity to the large subunit and Cas5 of the complex. In type I-E Cas3 a superfamily 2

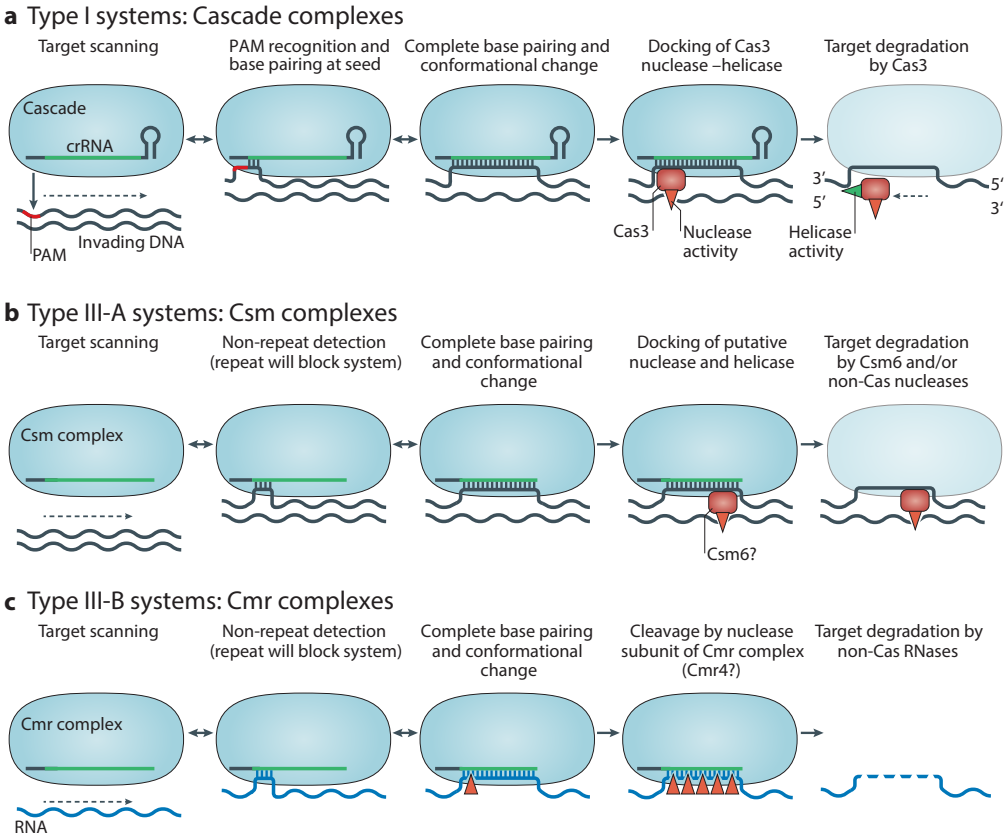


Figure 7: Target recognition and interference in type I and III crRNP interference complexes. The mechanism of target binding and degradation differs with regard to the substrate's (RNA in blue, DNA in gray) and system's protein machineries.
a. type I, Cascade. **b.** type III-A, Csm. **c.** type III-B, Cmr. Interference complexes are represented in blue; protospacer-adjacent motif (PAM) highlighted in red; arrows indicate directionality of cleavage. Adapted from ref. [10].

helicase domain is fused to a HD-nuclease [30]. The HD domain can be subtype specifically expressed as a single protein (I-A) or is fused with either the large subunit (I-D, III-B) or with Cas5 (I-B) protein [30]. For type III pathways, the understanding of target recognition and cleavage remains fragmentary. It is appreciated that type III-A *Sulfolobus solfataricus* Csm complexes target DNA *in vivo*, with potential involvement of Csm6, a Csx1 family protein [55]. Recent studies from V. Siksnys show that a Csm complex cleaves ssRNA *in vitro* via the catalytically active Cas7 family protein Csm3 (V. Siksnys, personal communication). Similarly in type III-B the Cas7 family protein Cmr4 has been shown

to possess ribonucleolytic activity [69]. Cas7 target cleavage explains the presence of a distinct cleavage pattern due to multiple active sites [22, 69].

The complex interplay of these proteins assures an effective immune response. The growing availability of structural data of single proteins and sub-complexes confirms bioinformatic classification. Combined efforts have contributed to the global appreciation of the CRISPR immune response. Despite the current general understanding of the pathways and detailed insights in functionalities of single proteins, molecular details and evolutionary connections remain elusive. The future of CRISPR research relies both on high-resolution atomic models of protein-nucleotide complexes and advanced knowledge of selective pressure within different ecological contexts.

1.6 Beyond adaptive immunity

1.6.1 Alternative CRISPR functions in prokaryotes

The CRISPR/Cas system developed functions outside the realm of innate immunity, some of which have evolved alongside its primary function. For instance, constitutive nuclease activity is prone to cause off-target cleavage effects. It is probably for this reason that Cas1 from the *E. coli* system I-E was found to associate and interact with known DNA repair factors and to be co-regulated with them after irradiation [46]. Off-target effects also generate an selective advantage for strains with increased selectivity for non-self (protospacer) vs. self (spacer) sequences and increase evolutionary fitness by accelerated local mutation rates [33]. Finally, the versatility of the system has led to it being repurposed, both by nature as a means of endogenous gene regulation, as well as by scientists as an extremely versatile tool for genome editing and manipulation (Box 3) [57, 77].

1.6.2 CRISPR versus RNAi

The programmable, sequence-directed, RNA-guided character of the CRISPR system has intriguing analogies to the eukaryotic RNA interference (RNAi) and related pathways [12]. In both cases, a protein machinery acts on nucleic acids, resulting in blockage

of gene function. Specificity is conferred not by the proteins, but solely by the sequence of an RNA guide that is processed from a longer precursor [78]. In many cases, the guide RNA as well as the target is originally derived from an invading genetic element, such as a virus or transposon, pointing at the evolutionary origin of both systems [79]. However, apart from these general principles, the prokaryotic and eukaryotic systems have distinct features: First and foremost, the protein machinery is entirely different and shows no similarity on sequence or structural levels [78, 80]. crRNA transcripts arise from genetic information that was incorporated into the host genome, whereas RNAi guides can be derived directly from the invasive nucleic acid [79]. Moreover, crRNA transcripts are single-stranded, with some systems requiring specific secondary structure elements. In contrast, RNAi requires long double-stranded RNA precursors, including stem-loop structures [80]. CRISPR/Cas systems act predominantly on invasive elements, whereas the RNAi machinery has been extensively repurposed for the regulation of endogenous transcripts, for instance by way of microRNAs; endogenous gene regulation by CRISPR/Cas is only beginning to emerge in some species [22, 77]. Finally, RNAi targets are exclusively RNAs, whereas CRISPR/Cas can act on both DNA and RNA. From a biotechnological point of view, this enabled scientists to exploit CRISPR/Cas for genome editing, whereas RNAi can be used only for post-transcriptional gene regulation.

Box 3 CRISPR-based technologies.

Given its simplicity, scientists have recognized the powerful potential of the type II pathway and rapidly exploited it for genetic engineering [57]. For this, the tracrRNA-crRNA duplex is substituted by a single guide RNA (sgRNA) – a fusion of the two RNA molecules. This eliminates the maturation steps required for Cas9 activation and allows the synthetic production of a vast variety of guide RNAs. The RNA-programmable Cas9 is directed towards site-specific DNA cleavage. Further genetic manipulation of Cas9 itself and its nucleic acid partners expand the repertoire of Cas9 genome editing from insertion, deletions, silencing, directed target of proteins to a DNA sequence in prokaryotes up to higher eukaryotes [81, 82].

1.7 Aim of the thesis

The heart of the CRISPR/Cas-mediated immune response pathway is the interference complex. In type I and III systems, multiple Cas7 proteins assemble around the guide crRNA, pre-ordering the RNA in a helical fashion. Depending on Cas7 position in the backbone, it contacts either Cas6, the large and small subunit. Thus Cas7 possesses multiple interaction surfaces for nucleic acid as well as protein interactions. I sought to investigate the structural basis that enables this versatile function and distinguishes Cas7 from the other major protein families. Due to poor sequence conservation, the structural biological approach was indispensable for subsequent biochemical analysis.

2 Results

2.1 Publication 1: Structure and RNA-binding properties of the type III-A CRISPR-associated protein Csm3

Hrle, A., Su, A. A., Ebert, J., Benda, C., Randau, L., Conti, E. Structure and RNA-binding properties of the type III-A CRISPR-associated protein Csm3. RNA Biol. (2013)

The manuscript 'Structure and RNA-binding properties of the type III-A CRISPR-associated protein Csm3' presents the first high-resolution protein structures of a Cas7 protein of the type III interference complex. The 2.37 Å crystal structure of *Mk* Csm3 shows that three domains are arranged around a central RRM-like fold. Structural superposition with representatives of the major Cas superfamililies 5, 6 and 7 confirm bioinformatic predictions that Csm3 is a Cas7 protein. *Mk* Csm3 binds RNA in a sequence-independent manner, but relies on a minimal length of 15 nucleotides. Based on surface potential and conservation we investigated the RNA-binding interface. Mutation analysis pinpointed a conserved surface area involved in RNA binding.

RNA Biology 10:11, 1670–1678; November 2013; © 2013 Landes Bioscience

Structure and RNA-binding properties of the Type III-A CRISPR-associated protein Csm3

Ajla Hrle[1,†], Andreas AH Su[2,†], Judith Ebert[1], Christian Benda[1], Lennart Randau[2,*], and Elena Conti[1,*]

[1]Structural Cell Biology Department; Max Planck Institute of Biochemistry; Munich/Martinsried, Germany; [2]Max Planck Institute for Terrestrial Microbiology; Karl-von-Frisch-Straße 10, Marburg, Germany

[†]These authors contributed equally to this work.

Keywords: RAMP, RRM domain, ferredoxin domain, Cas7, adaptive immunity

The prokaryotic adaptive immune system is based on the incorporation of genome fragments of invading viral genetic elements into clustered regularly interspaced short palindromic repeats (CRISPRs). The CRISPR loci are transcribed and processed into crRNAs, which are then used to target the invading nucleic acid for degradation. The large family of CRISPR-associated (Cas) proteins mediates this interference response. We have characterized *Methanopyrus kandleri* Csm3, a protein of the type III-A CRISPR-Cas complex. The 2.4 Å resolution crystal structure shows an elaborate four-domain fold organized around a core RRM-like domain. The overall architecture highlights the structural homology to Cas7, the Cas protein that forms the backbone of type I interference complexes. Csm3 binds unstructured RNAs in a sequence non-specific manner, suggesting that it interacts with the variable spacer sequence of the crRNA. The structural and biochemical data provide insights into the similarities and differences in this group of Cas proteins.

Introduction

For a long time, prokaryotic immune systems were believed to be restricted to "innate" immunity mechanisms (e.g., restriction modification systems)[1] and to defense mechanisms that result in cell death upon infection (e.g., toxin-antitoxin systems).[2] In the past decade, however, it has become clear that prokaryotes have evolved sophisticated and diverse adaptive immune systems that memorize previous attacks of foreign genetic elements. These systems consist of clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR-associated (Cas) proteins.[3-5] CRISPR-Cas is a nucleic acid-based defense system against mobile genetic elements such as viruses.[5] The CRISPR-Cas machinery distinguishes foreign (non-self) target DNA from (self) targets that are, for example, provided by a host CRISPR locus.[6,7]

The central element of CRISPR arrays is the arrangement of DNA sequences of variable length (spacers) derived from foreign genetic elements and separated by short 24–48 nt repeat sequences.[5] Upon infection, these clusters are transcribed into precursor crRNAs (pre-crRNA), which then are processed into mature CRISPR RNAs (crRNA).[8-10] The common features of mature crRNAs are the spacer, which identifies the matching target (protospacer) via base pairing, and the 5′-terminal 8 nt repeat tag (psi-tag), which is complementary to the self DNA but not to 2–4 nt short protospacer adjacent motif (PAM) sequences.[11] Adjacent to this array are the cas genes.[12,13] These encode proteins that are responsible for mediating the CRISPR response and that have a variety of functions, including nucleic acid binding and cleavage.[14]

CRISPR-Cas systems have been classified into three main types (I, II, and III) and 10 subtypes by bioinformatic analyses based on their cas gene organization, on the sequence and the structure (known or predicted) of the corresponding proteins.[15] The three CRISPR types also differ in the composition and mechanisms of their effector complexes.[16] Type I effector complexes are termed Cascade (CRISPR-associated complex for antiviral defense), type II effector complexes consist of a single Cas protein and two RNA molecules, and type III interference complexes are further divided into type III-A (Csm complex targeting DNA) and type III-B (Cmr complex targeting RNA).[11,17] In recent years, structural information on Cas proteins has started to provide insights into the molecular mechanisms of crRNA binding and target recognition. The combination of X-ray crystallography[8,18-22] and electron microscopic studies of the type I Cascade[23-25] and of the Type III-B Cmr-complex[26-28] has shown how some of the Cas proteins interact and bind crRNA.

Type I effector complexes are built around a central backbone composed of proteins of the Cas7 family.[23,29] The crystal structure of a Cas7 type I protein has revealed the presence of a central RRM/ferredoxin-like domain with several insertions and a C-terminal extension.[29] In type I systems, Cas7

*Correspondence to: Lennart Randau, Email: lennart.randau@mpi-marburg.mpg.de; Elena Conti, Email: conti@biochem.mpg.de
Submitted: 08/15/2013; Revised: 09/10/2013; Accepted: 09/16/2013
http://dx.doi.org/10.4161/rna.26500

Table 1. Data collection and structure refinement statstics of Csm3

Data collection		
	Native	**Native, Zn-SAD**
Resolution range (Å)	69.49 - 2.37 (2.49 - 2.37)	49.86 - 3.20 (3.00–3.08)
Unit cell(Å)[a]	a = 95.75 b = 101.02 c = 174.17	a = 70.51 b = 70.51 c = 193.36
Total reflections	308429	672581
Unique reflections	34685	38958
Multiplicity[a]	8.9 (7.1)	56.3 (52.4)
Completeness (%)[a]	99.5 (95.2)	99.9 (99.7)
Mean I/sigma(I) [a]	17.10 (2.72)	32.72(1.76)
Refinement		
B-factor	42.57	
R-factor(%)	18.09	
R-free(%)	21.34	
RMSD(bonds) (Å)	0.005	
RMSD(angles) (Å)	0.85	
Ramachandran favored (%)	95	
Ramachandran outliers(%)	0.15	

[a]Values in parentheses correspond to the highest resolution shell.

oligomerizes upon crRNA binding. In the best-characterized effector complex so far, the *Escherichia coli* Cascade complex, the crRNA binds within a super-helical grove formed by six copies of Cas7.[23,30] This helical arrangement has also been observed within other type I systems.[29,31,32] Despite the absence of significant sequence similarity, bioinformatic analysis has predicted that Cas7-like proteins also exist in type III systems.[33] Recently, it was shown that a Csm3 (CRISPR-Cas Subtype Mtube, protein 3) from *Staphylococcus epidermidis* binds RNA molecules at multiple sites.[34] Here, we present the crystal structure and RNA-binding properties of *Methanopyrus kandleri* Csm3. The structural and biochemical analysis of this type III-A Cas protein indicates that Csm3 is a Cas7-like protein capable of binding crRNA, suggesting it forms the backbone of the CRISPR-Cas Type III-A system effector complex.

Results and Discussion

Structure determination of Csm3

We expressed full-length *Methanopyrus kandleri* (*Mk*) Csm3 (351 residues) in *E. coli* and purified it to homogeneity (**Fig. S1A**). *Mk* Csm3 yielded crystals in an orthorhombic space group (*C*222) containing two molecules per asymmetric unit and diffracting beyond 2.4 Å resolution. An X-ray fluorescence scan on the crystals showed an unexpected peak at the Zinc excitation energy, suggesting the presence of intrinsically bound Zinc ions in the crystallized protein. We exploited the presence of this anomalous scatterer to solve the structure by single-wavelength anomalous dispersion method (SAD). The phases (obtained from a single bound Zinc ion) were of sufficient quality to build the polypeptide chain. The structure was refined at 2.37 Å resolution to an R_{free} of 21.0%/R_{work} of 18.0% and good stereochemistry (**Table 1**). The final model includes most of the protein, with the exception of a disordered region between residues 200 and 214. The two independent molecules in the asymmetric unit are very similar, superposing with a root mean square deviation (rmsd) of 0.22 Å for more than 95% of the Cα atoms. Static light scattering experiments of Csm3 in solution showed a mass of 33.3 kDa (**Fig. S1B**), consistent with the presence of a monomeric species. Thus, the interaction of the two molecules in the asymmetric unit reflects crystal packing contacts and not a physiological oligomer.

Csm3 is built of four domains organized around a central RRM-like fold

The crystal structure of *Mk* Csm3 reveals a compact architecture that can be described as composed of four domains: the core, the lid, the helical, and the C-terminal domains (**Fig. 1A**, in green, blue, red, and yellow, respectively). The core domain has a β1-α1-β2-β3-α2-β4 arrangement of secondary structure elements (**Fig. 1B**) with a topology typical of RRM-like and ferredoxin-like folds. Accordingly, the *Mk* Csm3 core domain folds into an antiparallel β-sheet, with two α-helices packed against the concave (back) surface. However, several features set the *Mk* Csm3 core domain apart from canonical RRM-like folds. In the β-sheet, strand β1 is long and highly bent, with a glycine

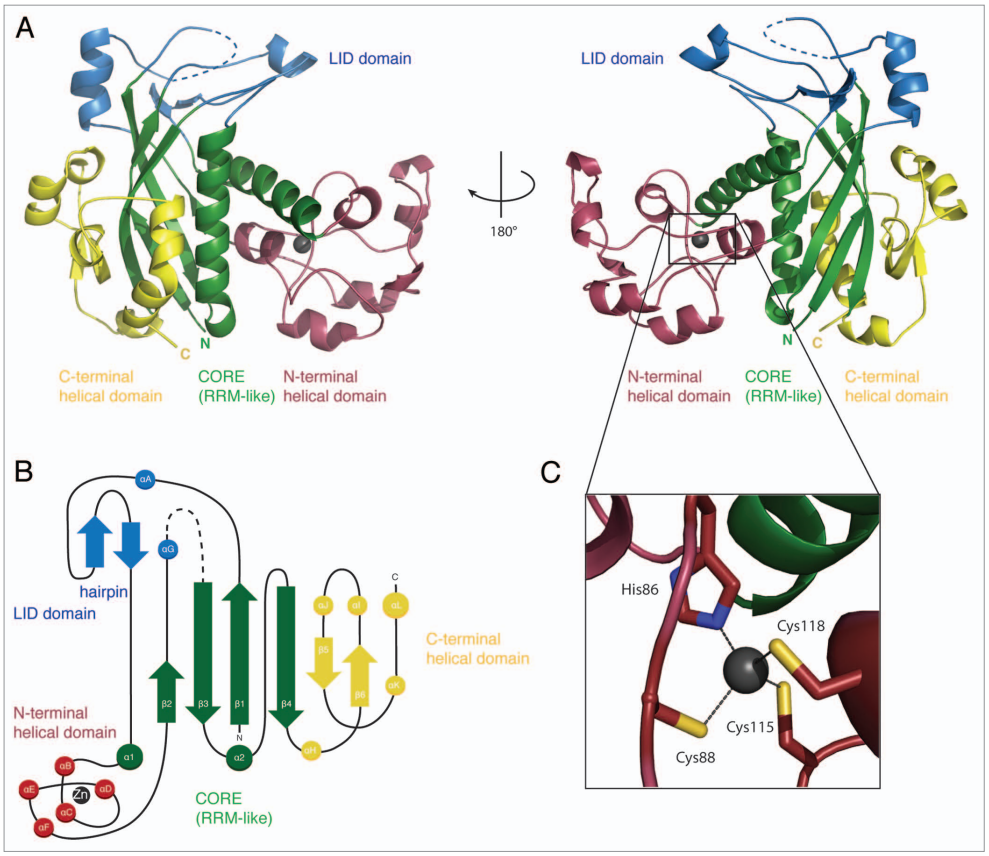


Figure 1. Structure of *Methanopyrus kandleri* Csm3. (**A**) The structure of *Mk* Csm3 can be divided into four distinct elements: the core (green) and lid domain (blue), a helical N-terminal (red), and a C-terminal domain (yellow). The structural elements of the core adopt a ferredoxin-like fold with β-α-β-β-α-β arrangement. The core is topologically interrupted by multiple insertions forming the lid and the helical N-terminal domain. The C-terminal domain packs against the core and is of mixed structural composition. The dashed blue line represents the missing disordered region between residues 200 and 214. The two views are related by a 180° rotation as indicated. (**B**) Topology diagram of *Mk* Csm3. Helices are represented as circles and β-strands as arrows. The secondary structure elements have been labeled numerically maintaining the nomenclature of RRM domains. The β-strands of the C-terminal domain extending the RRM β-sheet have also been labeled numerically. The additional α-helices have been labeled with letters (αA to αL). (**C**) A structural zinc ion present in the helical N-terminal domain is shown as a gray sphere, together with the coordinating residues (a cysteine and three histidine residues).

residue (Gly12) at the bending point effectively dividing it into two separate structural elements (strands $\beta1_A$ and $\beta1_B$, **Fig. 1B**). Strands β3 and β4, which sandwich β1, are also elongated (~12 residues), while strand β2 is very short (three residues).

The secondary structure elements of the core are connected by loop regions ranging from 2–10 amino-acid residues (between β3-α2 and between α2-β4, respectively) or by larger insertions (between β1-α1, β2-β3, and α1-β2) (**Fig. 1A and B**). The 35-residue long β1-α1 insertion contains a short β-hairpin and a one-turn α-helix (α_A). On one side, it packs against the 45-residue long β2-β3 insertion, which also contains an α-helix (α_G). On the other side, it packs against the α2-β4 loop. Overall, these interactions form the lid domain, which is positioned at the top of the β-sheet and is partially disordered (at a glycine-containing loop in the β2-β3 insertion).

The 100-residue long α1-β2 insertion contains five short α-helices (α_B to α_F) connected by extended segments (**Fig. 1A and B**). This insertion forms the α-helical domain and wedges between the two helices (α1 and α2) of the core domain, near the short edge of the β-sheet (i.e., near β2). The helical domain

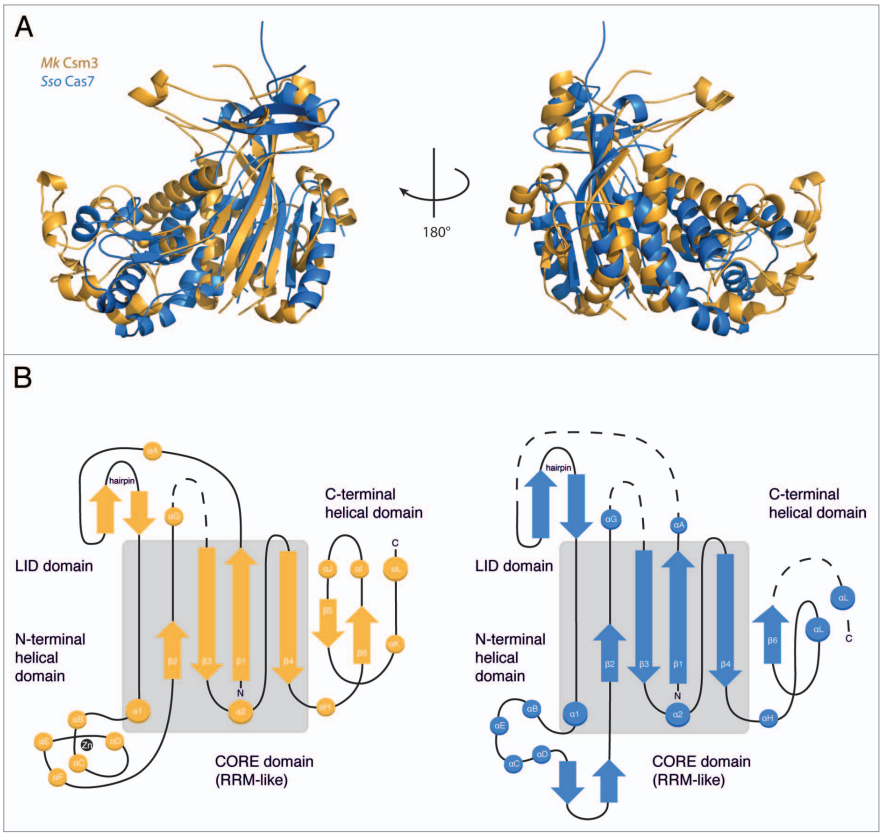


Figure 2. Structural similarity between Csm3 and Cas7. (**A**) *Sso* Cas7 (PDB ID: 3PS0, rmsd: 4.2Å, blue) shares the highest structural homology with *Mk* Csm3 (gold) beyond the core domain (gray). Both proteins have a similar arrangement of auxiliary domains surrounding the RRM-like fold, as well as a conserved architecture of the C-terminal domain. (**B**) Topology diagram of *Mk* Csm3 and *Sso* Cas7 showing the connectivity of the RRM fold relative to the other domains. The topological arrangement of the insertions is similar in both proteins. Similarities in secondary structure elements are highest within the core and low in the auxiliary domains.

binds a Zinc ion that is buried and is likely to have a structural role in stabilizing the fold of this domain (**Fig. 1C**). It connects helices (α_D to α_E) and is coordinated by His86, Cys88, Cys115, and Cys118. Only the latter two residues are well conserved among Csm3 orthologs (**Fig. S4A**). However, other cysteine and histidine residues are present in the α1-β2 insertion of Csm3 from other species (**Fig. S4A**). It is thus possible that other Csm3 proteins might have a Zinc-binding domain in the corresponding region of the structure, albeit with a different topology.

Finally, the RRM-like domain is followed by a C-terminal domain (αH-β6-αI-αJ-β5-αK-αL) (**Fig. 1A and B**). The C-terminal domain extends the core β-sheet by two antiparallel β-strands (β5 and β6), which flank the long edge of the core domain β-sheet (at β4). It also contains three short α-helices and a long C-terminal α-helix. The C-terminal α-helix packs against α2, at the convex surface of the β-sheet. The short helices (αI and αJ) contact the lid and partly occlude the front surface of the β-sheet. In canonical RRM domains, this front surface features hydrophobic residues that are part of the so-called RNP1 and RNP2 motifs and that bind RNA.[35] However, *Mk* Csm3 lacks the typical solvent-exposed hydrophobic residues that bind RNA in canonical RRM domains. Thus, the *Mk* Csm3 RRM seems to fulfill a structural purpose similar to other previously reported examples.[36,37]

Structural comparison of Csm3 with the Cas proteins of the RAMP superfamily

We compared the structure of *Mk* Csm3 with those of Cas5, Cas6, and Cas7, which represent the three major groups of

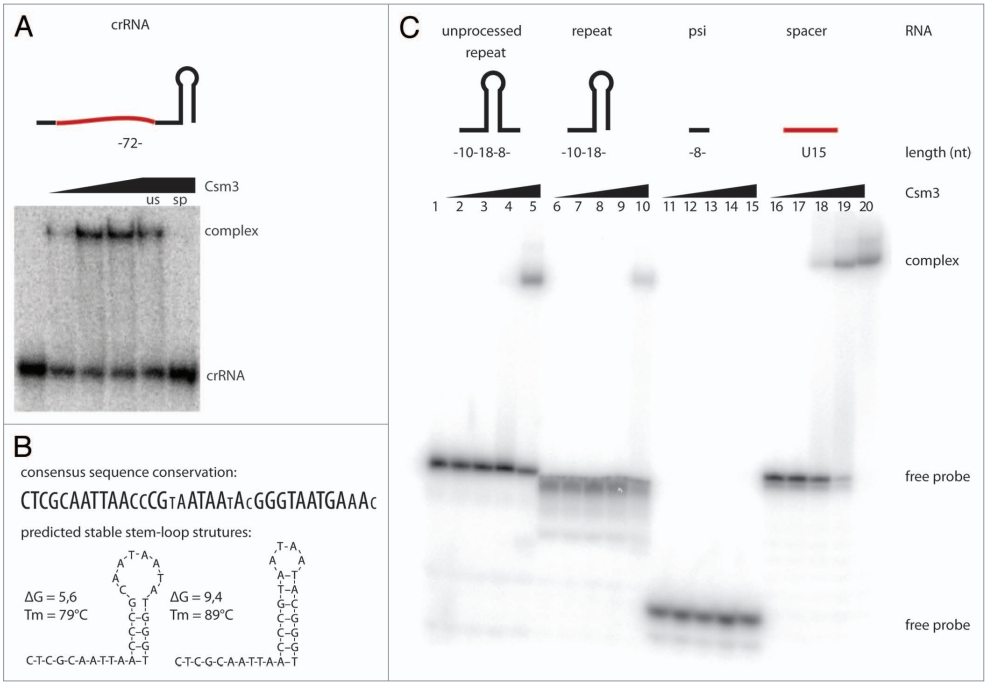


Figure 3. RNA-binding properties of Csm3. (**A**) *Mk* Csm3 binds to a physiological crRNA substrate (left panel). ^{32}P-labeled crRNA transcripts were incubated in the absence or presence of 5 μM, 10 μM, and 20 μM *Mk* Csm3. (**B**) Electrophoretic mobility shift assays were performed with the respective [^{32}P]-5′-end labeled RNAs and increasing concentrations of *Mk* Csm3 (0 μM, 1 μM, 30 μM, 100 μM). *Mk* Csm3 binds to single-stranded RNA substrates (lane 16–20) but not significantly to the repeat sequences (lanes 1–5 and 6–10). Binding to single-stranded RNA is dependent on length but not sequence (compare lanes 16–20 and 11–15). Weak binding of *Mk* Csm3 to processed and unprocessed repeat sequences (lanes 1–5 and 6–10, respectively) is likely attributed to the ssRNA overhangs. (**C**) *Methanopyrus kandleri* repeat sequence conservation and predicted RNA folding.

evolutionary distinct RRM-containing proteins in the RAMP (repeat-associated mysterious protein) superfamily[15] (**Fig. S2A**). *Bacillus halodurans* (*Bh*) Cas5d (PDB ID: 4F3M) has two RRM-like domains adjacent to each other and functions as an endoribonuclease in the pre-processing of crRNA transcripts. The N-terminal RRM-like domain contains the putative endoribonuclease site, which is centered at a histidine residue.[18,31] *Pyrococcus furiosus* (*Pf*) Cas6 (PDB ID: 3UFC) has an N-terminal RRM-like domain that packs against a twisted β-sheet domain.[21] This RRM-like domain contains an endoribonuclease site that is also centered at a histidine residue, although the exact position differs from that of *Bh* Cas5d. The similarity of *Mk* Csm3 with *Bh* Cas5d and *Pf* Cas6 is limited to the RRM-like domain (**Fig. S2A**). Using the structural alignment program SSM as implemented in Coot,[38] *Mk* Csm3 superposed with *Bh* Cas5d with an rmsd of 3.9 Å over 80 Cα atoms and with *Pf* Cas6 with an rmsd of 4.1 Å over 125 Cα atoms. No prominent histidine residue or possible catalytic triad is however apparent from these structural alignments of Csm3. Consistently, *Mk* Csm3 did not exhibit any prominent endonucleolytic activity with repeat RNA or precursor RNA substrates (data not shown).

Bioinformatic analyses have predicted that Csm3 belongs to the Cas7 family of RAMP proteins.[39] Superposition of *Mk* Csm3 with the *Sulfolobus solfataricus* (*Sso*) Cas7 (PDB ID: 3PS0) structure results in an rmsd of 4.2 Å over 110 Cα atoms. As with Cas5 and Cas6, the structural similarity with Cas7 is primarily at the RRM-like domain. However, *Mk* Csm3 shares significant overall architectural analogy with Cas7 (**Fig. 2A**). In particular, the two proteins have a similar arrangement of domains around the RRM-like fold. Cas7 contains a lid domain, a (mostly) helical domain, and a C-terminal domain at equivalent structural positions as described above for Csm3 (**Fig. 2B**). Although Cas7 is not a Zinc-binding protein and although the exact topological arrangement of secondary structure elements differs from *Mk* Csm3, the overall dimensions and shape of the two proteins is remarkably similar (**Fig. 2A**). As Cas7 is a scaffold RNA-binding protein, we assessed whether *Mk* Csm3 might have similar RNA-binding properties.

Csm3 binds single-stranded RNAs in a sequence non-specific manner

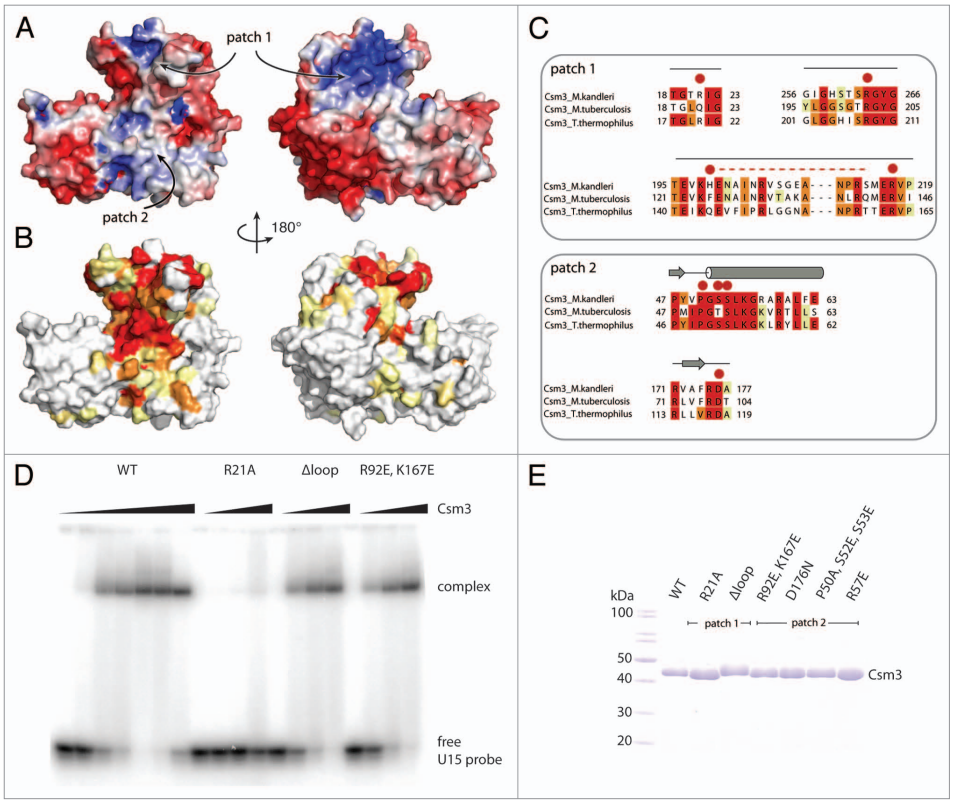


Figure 4. Identification of Csm3 RNA-binding residues. (**A**) The structure of Csm3 is shown in surface representations, in the same orientations as in **Figure 1A**, colored according to electrostatic potential. Charged patches (blue) are present at the back of the lid domain as well as at the interface between the core and N-terminal helical domain. Negatively charged surfaces (red) are located along the front of the N-terminal insertion and cover the C-terminal domain. Two surface patches discussed in the text (patch 1 and 2) are indicated. (**B**) Corresponding surface representations of Csm3 colored according to conservation with the Csm3 family. The conservation is based on a comprehensive alignment (**Fig. S4B**). Increase in conservation is shown in increasingly darker shades (from to red). No or low conservation (white and yellow) is found in the N-terminal insertion and the C-terminal domain. Highly conserved residues (orange and red) are located within the lid (patch 1) and core domains (patch 2) and coincide with positively charged surfaces (**A**). (**C**) Sequence alignments of Csm3 orthologs in regions corresponding to surface patches 1 and 2 (**A** and **B**). Residues selected for mutation analysis are highlighted with red dots. The unstructured loop (H199-S214) replaced by a $(GS)_3$ linker is represented as a dashed red line. (**D**) RNA binding of Csm3 mutants to a single-stranded RNA substrate U15. Wild-type (WT) protein and the double mutation within the core domain (patch2) bind with comparable affinity. Replacement of the unstructured loop (H199-S214) by a -$(GS)_3$- linker does not impair binding, while the single mutation R21A has completely lost RNA binding ability at this condition. (**E**) Coomassie-stained 12% SDS-PAGE gel of the purified protein samples used in the assays.

We performed electrophoretic mobility shift assays (EMSA) with crRNA substrates that were generated by in vitro transcription (**Fig. 3A**). These assays indicated that *Mk* Csm3 binds crRNAs (**Fig. 3A**). The *Mk* crRNAs contain a highly conserved repeat sequence of 36 nucleotides that includes a predicted stable stem-loop of 16–18 nucleotides (**Fig. 3B**) and a highly conserved eight nucleotide AATGAAA(C/G) motif at the 5′ end (psi-tag). They also contain variable spacer sequences ranging from 40–50 nucleotides.[40] We dissected which parts of the crRNA are recognized by *Mk* Csm3. In gel-shift assays, Csm3 showed weak binding to processed and unprocessed repeat sequences (**Fig. 3C**), but not to its stem-loop structure alone (**Fig. S3A**). We tested whether *Mk* Csm3 binds single-stranded RNA, which is present in part of the repeat sequence as well as in the variable spacer. In gel-shift assays, Csm3 bound a 15-mer polyU RNA or 15-mer polyA about 10 times stronger than the repeat sequence (**Fig. 3A**). Thus, the length of the single-stranded RNA might affect the strength of the interaction with *Mk* Csm3. *Mk*

Csm3 did not exhibit detectable RNA binding toward the 8 nt psi-tag in the gel-shift assays (**Fig. 3C**; **Fig. S3A**). We conclude that *Mk* Csm3 binds single-stranded RNAs from 15 nucleotides onwards in an apparently sequence non-specific manner and that RNA structures impair binding. This suggests that the variable sequence of the crRNA is bound by *Mk* Csm3, rather than the structured and conserved repeat.

To identify the RNA-binding interface, we examined the surface features of *Mk* Csm3 in terms of charge distribution (**Fig. 4A**) and evolutionary conservation (**Fig. 4B and C**). The lid domain contains a striking patch (1) of conserved and surface-exposed positively charged residues including Arg217, Arg263, and Arg267 (**Fig. S4A**). Another positively charged residue, Arg21, is located at the center of this patch and approaches the position of *Sso* Cas7 His160, a residue that has been shown to be important for RNA binding.[29] A single mutation of *Mk* Csm3 Arg21 to Ala abolished RNA binding in EMSA assays (**Fig. 4D**). In the lid domain, the positively charged surface patch is near the disordered glycine-containing loop (**Fig. 4C**). This loop is conserved (**Fig. 4C**) but does not appear to be involved in RNA binding, as its deletion did not show a significant change in the EMSA assay as compared with the wild-type (WT) protein (**Fig. 4D**). Another striking surface patch (2) of *Mk* Csm3 is located at the interface between the lid domain and the helical domain. In particular, helix α1 exposes several conserved residues, including Pro50, Ser52, Ser53, and Arg57 (**Fig. S4B**). Mutation of this conserved surface patch (2), however, did not significantly impair RNA binding (**Fig. S4C**). We concluded that Csm3 uses the lid domain to bind single-stranded RNA. It is possible to envisage that the other conserved surface patches on Csm3 mediate other types of macromolecular interactions, including protein–protein interactions that form in Csm3-containing effector complexes.

Conclusions

CASCADE/Cmr/Csm complexes share common functionalities, as reflected in their similar composition of proteins. Proteins of the Cas7 family are the backbone of the Type I effector complexes and are involved in interactions with both crRNA and other Cas proteins.[19,29,32] Computational analyses predicted that Csm3 might fulfill the role of the backbone protein Cas7 in type III interference assemblies.[39] Here, we show that *Mk* Csm3 has indeed a remarkably similar architecture as compared with *Sso* Cas7. We found that the structural similarity involves not only the central RRM-like domain, but also insertions at equivalent structural positions in the RRM fold. At the sequence level, however, the two proteins have almost completely diverged.

In line with the structural similarity to Cas7, *Mk* Csm3 recognizes crRNA. We found that Csm3 binds to a variable sequence of ssRNA via the flexible insertion that forms a lid on top of the RRM domain. The overall affinity toward RNA is significant yet not strong. It is abolished through mutation of an arginine residue (Arg21Ala) yet hardly reduced when mutating other conserved residues within the positively charged surfaces. It is possible however that this region contributes to RNA binding when in the context of a fully assembled Csm complex. Type III systems further process premature crRNA to mature crRNA; Csm3, together with Csm2 and Csm5, were reported to be required for crRNA 3′ termini maturation.[41] However, in our studies, we could not identify potential catalytic residues nor could we observe nucleolytic activity in biochemical assays. This is in agreement with *S. epidermidis* Csm3 studies indicating that crRNA maturation cleavage events are not performed by the Cas10/Csm complex.[34]

Cas7 proteins oligomerize with a helical arrangement around the crRNA and interact with other Cas proteins of the effector complex, such as the Cas5 in Type I-A.[42] The *S. solfataricus* Cas7 protein was shown to be monomeric and is thought to require Cas5 and crRNA for nucleation and stabilization of its assembly.[29] In agreement, *Mk* Csm3 also behaves as a monomer in solution and might only oligomerize in the context of the Csm complex. It is possible that the insertion domains that surround the RRM and/or the RRM itself might provide interfaces for protein–protein interactions.[35] We note, however, in contrast to observations with bacterial *S. epidermidis* Csm3, we did not observe binding of RNA molecules in six nucleotide increments for *Mk* Csm3.[34] Our structural observations provide a first step toward the structural elucidation of the Csm proteins and their respective role in the surveillance complex. Additionally, the structure will contribute to characterizing the evolutionary relationship within the Cas7 protein family. Further tentative type III members of this family (Cmr1, Cmr4, Cmr6, Csm5)[39] remain to be analyzed and classified.

Experimental Procedures

Protein expression and purification

Mk Csm3 wild-type and mutant proteins were expressed as recombinant His- and His-SUMO-tagged fusion protein using BL21-Gold (DE3) Star pRARE (Stratagene) in TB medium and induced overnight at 18 °C. The cells were lysed in buffer A (50 mM Tris pH 7.5, 200 mM NaCl, 10% Glycerol) supplemented with 10 mM Imidazole, DNase, protease inhibitors (Roche) by sonication. Proteins (wild-type and mutants) were purified using Nickel-based affinity chromatography. The His-SUMO tag was cleaved by adding SUMO protease overnight. Proteins were further purified by size-exclusion chromatography (Superdex 75, GE Healthcare) in gel-filtration buffer (buffer A supplemented with 2 mM DTT). Point mutations were introduced by Quick Change site directed mutagenesis according to the manufacturer's instruction (Stratagene).

Crystallization, data collection, structure determination, and analysis

Crystallization was performed at room temperature using hanging drop vapor diffusion method and equal volumes of the protein at 20 mg/ml (gel-filtration buffer) and of crystallization buffer (25% MPD and 50 mM MES 6.0). Crystals were both flash-frozen directly from the crystallization drop as well as subjected to further dehydration (increasing amounts of MPD up to 60%) and diffracted beyond 2.4 Å.

All diffraction data was collected at 100 K at the beamline PXII of the Swiss Light Source (SLS) synchrotron and processed

using XDS.[43] The structures were determined using the native data and Zn-SAD phases to build an initial model. This was then used as a search model for molecular replacement of higher resolution data using Phaser.[44] Model building was performed manually with the program Coot[38] and refined with PHENIX.[45] The data collection and refinement statistics are summarized in **Table 1**. Figures were prepared using PyMOL (http://www.pymol.org).

Biochemical assays

The RNA molecules U15, A10, A15, A20, A40 were synthesized (Purimex). The crRNA (locus 5, spacer 5) was produced by in vitro run-off transcription and purified by elution of the crRNA transcript from a polyacrylamide gel as described.[10] The RNA molecules were 5′-labeled with T4 polynucleotide kinase (New England Biolabs) and γ-[^{32}P] ATP (Perkin- Elmer).

For the gel-shift assays, 0.5 pmol labeled RNA was mixed with 1 µM, 10 µM, 30 µM, 100 µM protein in a 10 µL reaction containing 20 mM Hepes at pH 7.5, 100 mM KOAc, 4 mM $Mg(OAc)_2$, 0.1% (vol/vol) NP-40, and 2 mM DTT. Fifteen ng/µL (500 fmol/µl = 500x molar excess) yeast tRNA mix (Amicon) were used as non-specific and 15 ng/µl unlabeled crRNA transcripts were used as specific competitor molecules. The mixtures were incubated for 20 min at 55 °C before adding 2 µL 50% (vol/vol) glycerol containing 0.25% (wt/vol) xylene cyanole. Samples were run on a 8% (wt/vol) polyacrylamide gel at 4 °C and visualized by phospho-imaging (GE Healthcare).

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

The authors would like to thank Sutapa Chakrabarti and Marco Hein for critical reading of the manuscript and Jérôme Basquin, Karina Valer-Saldaña, and Sabine Pleyer at the MPI- Martinsried Crystallization Facility. The authors thank the staff of the PXII beamline at the Swiss Light Source for assistance during data collection. This study was supported by the Max Planck Gesellschaft, the DFG Research Group 1680 (FOR1680) to EC and LR, CIPSM to EC and the Schering Foundation fellowship to AH. Author contributions: AH, AS, and LR designed the experiments; AH and CB solved the structure; AS performed the experiment in **Figure 3A**; AH performed all other experiments. AH, EC, and LR wrote the manuscript.

Accession Number

The coordinates and the structure factors have been deposited to the Protein Data Bank with the accession code: 4N0L

Supplemental Materials

Supplemental materials may be found here:
www.landesbioscience.com/journals/rnabiology/article/26500

References

1. Labrie SJ, Samson JE, Moineau S. Bacteriophage resistance mechanisms. Nat Rev Microbiol 2010; 8:317-27; PMID:20348932; http://dx.doi.org/10.1038/nrmicro2315
2. Makarova KS, Wolf YI, Koonin EV. Comparative genomics of defense systems in archaea and bacteria. Nucleic Acids Res 2013; 41:4360-77; PMID:23470997; http://dx.doi.org/10.1093/nar/gkt157
3. Mojica FJ, Díez-Villaseñor C, García-Martínez J, Soria E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J Mol Evol 2005; 60:174-82; PMID:15791728; http://dx.doi.org/10.1007/s00239-004-0046-3
4. Mojica FJ, Díez-Villaseñor C, Soria E, Juez G. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. Mol Microbiol 2000; 36:244-6; PMID:10760181; http://dx.doi.org/10.1046/j.1365-2958.2000.01838.x
5. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. CRISPR provides acquired resistance against viruses in prokaryotes. Science 2007; 315:1709-12; PMID:17379808; http://dx.doi.org/10.1126/science.1138140
6. Marraffini LA, Sontheimer EJ. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nat Rev Genet 2010; 11:181-90; PMID:20125085; http://dx.doi.org/10.1038/nrg2749
7. Wiedenheft B, Sternberg SH, Doudna JA. RNA-guided genetic silencing systems in bacteria and archaea. Nature 2012; 482:331-8; PMID:22337052; http://dx.doi.org/10.1038/nature10886
8. Carte J, Wang R, Li H, Terns RM, Terns MP. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. Genes Dev 2008; 22:3489-96; PMID:19141480; http://dx.doi.org/10.1101/gad.1742908
9. Haurwitz RE, Jinek M, Wiedenheft B, Zhou K, Doudna JA. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. Science 2010; 329:1355-8; PMID:20829488; http://dx.doi.org/10.1126/science.1192272
10. Richter H, Zoephel J, Schermuly J, Maticzka D, Backofen R, Randau L. Characterization of CRISPR RNA processing in Clostridium thermocellum and Methanococcus maripaludis. Nucleic Acids Res 2012; 40:9887-96; PMID:22879377; http://dx.doi.org/10.1093/nar/gks737
11. Marraffini LA, Sontheimer EJ. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science 2008; 322:1843-5; PMID:19095942; http://dx.doi.org/10.1126/science.1165771
12. Grissa I, Vergnaud G, Pourcel C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res 2007; 35(Web Server issue):W52-7; PMID:17537822; http://dx.doi.org/10.1093/nar/gkm360
13. Jansen R, Embden JD, Gaastra W, Schouls LM. Identification of genes that are associated with DNA repeats in prokaryotes. Mol Microbiol 2002; 43:1565-75; PMID:11952905; http://dx.doi.org/10.1046/j.1365-2958.2002.02839.x
14. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV. 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 2006; 1:7; PMID:16545108; http://dx.doi.org/10.1186/1745-6150-1-7
15. Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, Horvath P, Moineau S, Mojica FJ, Wolf YI, Yakunin AF, et al. Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol 2011; 9:467-77; PMID:21552286; http://dx.doi.org/10.1038/nrmicro2577
16. Bhaya D, Davison M, Barrangou R. CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. Annu Rev Genet 2011; 45:273-97; PMID:22060043; http://dx.doi.org/10.1146/annurev-genet-110410-132430
17. Hale CR, Zhao P, Olson S, Duff MO, Graveley BR, Wells L, Terns RM, Terns MP. RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. Cell 2009; 139:945-56; PMID:19945378; http://dx.doi.org/10.1016/j.cell.2009.07.040
18. Garside EL, Schellenberg MJ, Gesner EM, Bonanno JB, Sauder JM, Burley SK, Almo SC, Mehta G, MacMillan AM. Cas5d processes pre-crRNA and is a member of a larger family of CRISPR RNA endonucleases. RNA 2012; 18:2020-8; PMID:23006625; http://dx.doi.org/10.1261/rna.033100.112
19. Nam KH, Haitjema C, Liu X, Ding F, Wang H, DeLisa MP, Ke A. Cas5d protein processes pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvulg CRISPR-Cas system. Structure 2012; 20:1574-84; PMID:22841292; http://dx.doi.org/10.1016/j.str.2012.06.016
20. Mulepati S, Bailey S. Structural and biochemical analysis of nuclease domain of clustered regularly interspaced short palindromic repeat (CRISPR)-associated protein 3 (Cas3). J Biol Chem 2011; 286:31896-903; PMID:21775431; http://dx.doi.org/10.1074/jbc.M111.270017
21. Wang R, Preamplume G, Terns MP, Terns RM, Li H. Interaction of the Cas6 riboendonuclease with CRISPR RNAs: recognition and cleavage. Structure 2011; 19:257-64; PMID:21300293; http://dx.doi.org/10.1016/j.str.2010.11.014

22. Reeks J, Naismith JH, White MF. CRISPR interference: a structural perspective. Biochem J 2013; 453:155-66; PMID:23805973; http://dx.doi.org/10.1042/BJ20130316
23. Wiedenheft B, Lander GC, Zhou K, Jore MM, Brouns SJ, van der Oost J, Doudna JA, Nogales E. Structures of the RNA-guided surveillance complex from a bacterial immune system. Nature 2011; 477:486-9; PMID:21938068; http://dx.doi.org/10.1038/nature10402
24. Wiedenheft B, Zhou K, Jinek M, Coyle SM, Ma W, Doudna JA. Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. Structure 2009; 17:904-12; PMID:19523907; http://dx.doi.org/10.1016/j.str.2009.03.019
25. Zhu X, Ye K. Crystal structure of Cmr2 suggests a nucleotide cyclase-related enzyme in type III CRISPR-Cas systems. FEBS Lett 2012; 586:939-45; PMID:22449983; http://dx.doi.org/10.1016/j.febslet.2012.02.036
26. Osawa T, Inanaga H, Numata T. Crystal Structure of the Cmr2-Cmr3 Subcomplex in the CRISPR-Cas RNA Silencing Effector Complex. J Mol Biol 2013; (Forthcoming); PMID:23583914; http://dx.doi.org/10.1016/j.jmb.2013.03.042
27. Cocozaki AI, Ramia NF, Shao Y, Hale CR, Terns RM, Terns MP, Li H. Structure of the Cmr2 subunit of the CRISPR-Cas RNA silencing complex. Structure 2012; 20:545-53; PMID:22405013; http://dx.doi.org/10.1016/j.str.2012.01.018
28. Zhang J, Rouillon C, Kerou M, Reeks J, Brugger K, Graham S, Reimann J, Cannone G, Liu H, Albers SV, et al. Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. Mol Cell 2012; 45:303-13; PMID:22227115; http://dx.doi.org/10.1016/j.molcel.2011.12.013
29. Lintner NG, Kerou M, Brumfield SK, Graham S, Liu H, Naismith JH, Sdano M, Peng N, She Q, Copié V, et al. Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE). J Biol Chem 2011; 286:21643-56; PMID:21507944; http://dx.doi.org/10.1074/jbc.M111.238485
30. van Duijn E, Barbu IM, Barendregt A, Jore MM, Wiedenheft B, Lundgren M, Westra ER, Brouns SJ, Doudna JA, van der Oost J, et al. 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 2012; 11:1430-41; PMID:22918228; http://dx.doi.org/10.1074/mcp.M112.020263
31. Nam KH, Haitjema C, Liu X, Ding F, Wang H, DeLisa MP, Ke A. Cas5d protein processes pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvulg CRISPR-Cas system. Structure 2012; 20:1574-84; PMID:22841292; http://dx.doi.org/10.1016/j.str.2012.06.016
32. Wiedenheft B, van Duijn E, Bultema JB, Waghmare SP, Zhou K, Barendregt A, Westphal W, Heck AJ, Boekema EJ, Dickman MJ, et al. RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. Proc Natl Acad Sci U S A 2011; 108:10092-7; PMID:21536913; http://dx.doi.org/10.1073/pnas.1102716108
33. Koonin EV, Makarova KS. CRISPR-Cas: evolution of an RNA-based adaptive immunity system in prokaryotes. RNA Biol 2013; 10:679-86; PMID:23439366; http://dx.doi.org/10.4161/rna.24022
34. Hatoum-Aslan A, Samai P, Maniv I, Jiang W, Marraffini LA. A ruler protein in a complex for antiviral defense determines the length of small interfering CRISPR RNAs. J Biol Chem 2013; (Forthcoming)
35. Maris C, Dominguez C, Allain FH. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. FEBS J 2005; 272:2118-31; PMID:15853797; http://dx.doi.org/10.1111/j.1742-4658.2005.04653.x
36. Fribourg S, Gatfield D, Izaurralde E, Conti E. A novel mode of RBD-protein recognition in the Y14-Mago complex. Nat Struct Biol 2003; 10:433-9; PMID:12730685; http://dx.doi.org/10.1038/nsb926
37. Kadlec J, Izaurralde E, Cusack S. The structural basis for the interaction between nonsense-mediated mRNA decay factors UPF2 and UPF3. Nat Struct Mol Biol 2004; 11:330-7; PMID:15004547; http://dx.doi.org/10.1038/nsmb741
38. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr 2004; 60:2126-32; PMID:15572765; http://dx.doi.org/10.1107/S0907444904019158
39. Makarova KS, Aravind L, Wolf YI, Koonin EV. Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. Biol Direct 2011; 6:38; PMID:21756346; http://dx.doi.org/10.1186/1745-6150-6-38
40. Su AA, Tripp V, Randau L. RNA-Seq analyses reveal the order of tRNA processing events and the maturation of C/D box and CRISPR RNAs in the hyperthermophile Methanopyrus kandleri. Nucleic Acids Res 2013; 41:6250-8; PMID:23620296; http://dx.doi.org/10.1093/nar/gkt317
41. Hatoum-Aslan A, Maniv I, Marraffini LA. 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 2011; 108:21218-22; PMID:22160698; http://dx.doi.org/10.1073/pnas.1112832108
42. Jore MM, Lundgren M, van Duijn E, Bultema JB, Westra ER, Waghmare SP, Wiedenheft B, Pul U, Wurm R, Wagner R, et al. Structural basis for CRISPR RNA-guided DNA recognition by Cascade. Nat Struct Mol Biol 2011; 18:529-36; PMID:21460843; http://dx.doi.org/10.1038/nsmb.2019
43. Kabsch W. XDS. Acta Crystallogr D Biol Crystallogr 2010; 66:125-32; PMID:20124692; http://dx.doi.org/10.1107/S0907444909047337
44. McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ. Phaser crystallographic software. J Appl Crystallogr 2007; 40:658-74; PMID:19461840; http://dx.doi.org/10.1107/S0021889807021206
45. Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung LW, Kapral GJ, Grosse-Kunstleve RW, et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr 2010; 66:213-21; PMID:20124702; http://dx.doi.org/10.1107/S0907444909052925

Supplementary Figure1

Analysis of the protein used in biochemical assays

(A) Coomassie-stained SDS-PAGE and size exclusion chromatography elution profile of purified MkCsm3.

(B) Static light scattering (SLS) chromatogram and values underline the monomeric behaviour in solution.

Supplementary Figure 2

Structure and sequence-based similarities of Csm3 with key Cas proteins

(A) Structure-based homology modeling of *Mk* Csm3 (gold) with known structures of CRISPR proteins (blue) structurally manually aligned based on the DALI server output and their common feature, the ferredoxin fold. The protein is shown in the same orientation to that used in Figure 1. The ferredoxin-like fold is the region of highest conservation. The homology of the N-terminal ferredoxin fold domains of *Bh* Cas5d (PDB ID: 3KG4, rmsd: 3.9) and *Pf* Cas6 (PDB ID: 3UFC, rmsd: 4.1Å) is restricted to the core fold of *Mk* Csm3. *Sso* Cas7 (PDB ID: 3PS0, rmsd: 4.2Å) shares the highest structural homology with *Mk* Csm3 beyond the core domain. Both proteins have a similar arrangement of insertions within the ferredoxin-like fold domain as well as the overall architecture of the C-terminal domain.

(B) Structure-based sequence alignment of *Mk* Csm3 (gold) and *Sso* Cas7 (blue). The alignment includes sequences from representative species of both families. Lighter letters denote residues identical in more than one third of the species considered; darker letters identify residues identical in more than two thirds of the species. The colors are based on the comprehensive alignment in *Supplementary Figure 4A*. Secondary structure elements are shown above the sequences with cylinders for α-helices and arrows for β-strands. Sequence conservation between *Mk* Csm3 and *Sso* Cas7 is mostly restricted to structural residues that define the core domain and is low within adjacent regions.

Supplementary Figure 3

Dissecting RNA-binding properties of Csm3

Electrophoretic mobility shift assays (EMSA) were carried out with the respective [^{32}P]-5′ end labeled RNAs and increasing concentrations of *Mk* Csm3. (A) *Mk* Csm3 does not specifically recognize the stem-loop (lanes 1-4) or psi-tag (lanes 13-16) of the conserved repeat sequence, yet shows sequence unspecific binding to U15 (lanes 5-8) and A15 (lanes 9-12).

Supplementary Figure 4

Mapping of Csm3 RNA-binding residues

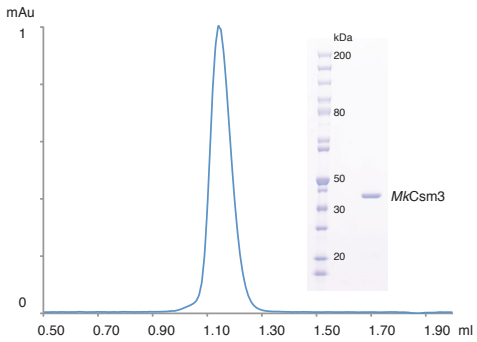
(A) Comprehensive sequence alignment of *Mk* Csm3 within the Csm-family. The alignment includes sequences from representative species. Lighter letters denote residues identical in more than one third of the species considered; darker letters identify residues identical in more than two thirds of the species.

(B) A close-up view of conserved residues (Pro50, Ser52, Ser53, Arg57 and D176) of the core domain, which were selected as representative mutations within the core (patch2).

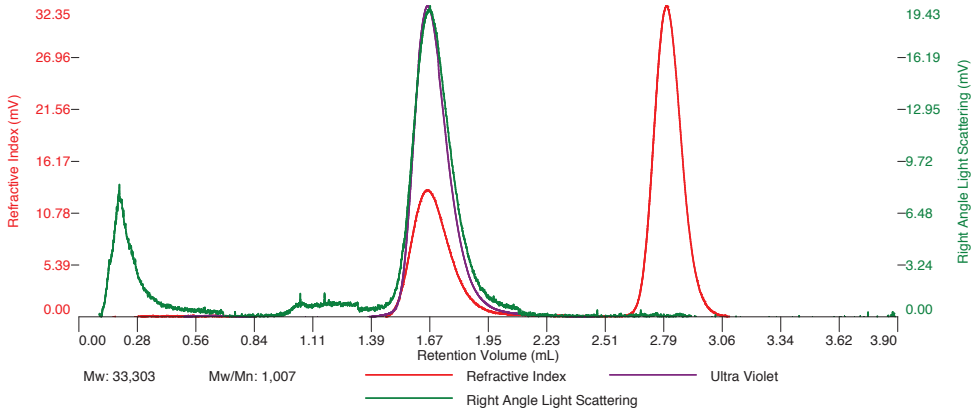
(C) Electrophoretic mobility shift assays (EMSA) were carried out with the respective [^{32}P]-5′ end labeled RNAs and increasing concentrations of *Mk* Csm3. Mutations located within the core (patch2) bind U15 ssRNA with comparable affinity to WT protein.

Supplementary Figure 1

A

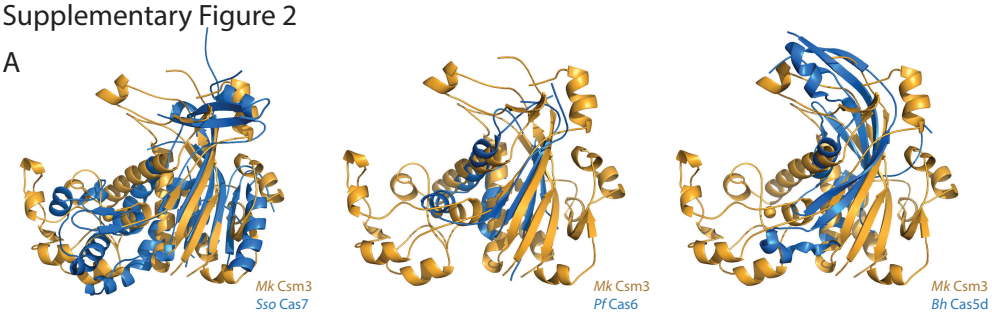


B

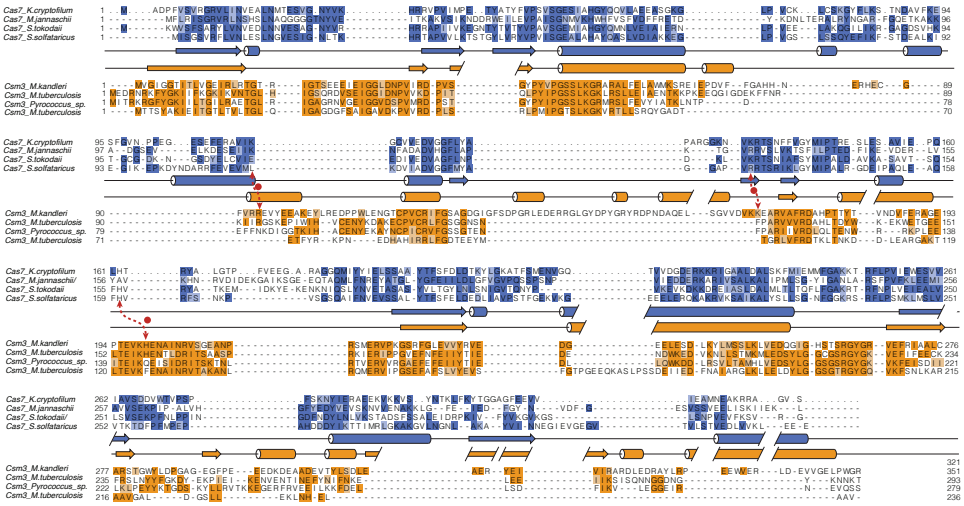


Supplementary Figure 2

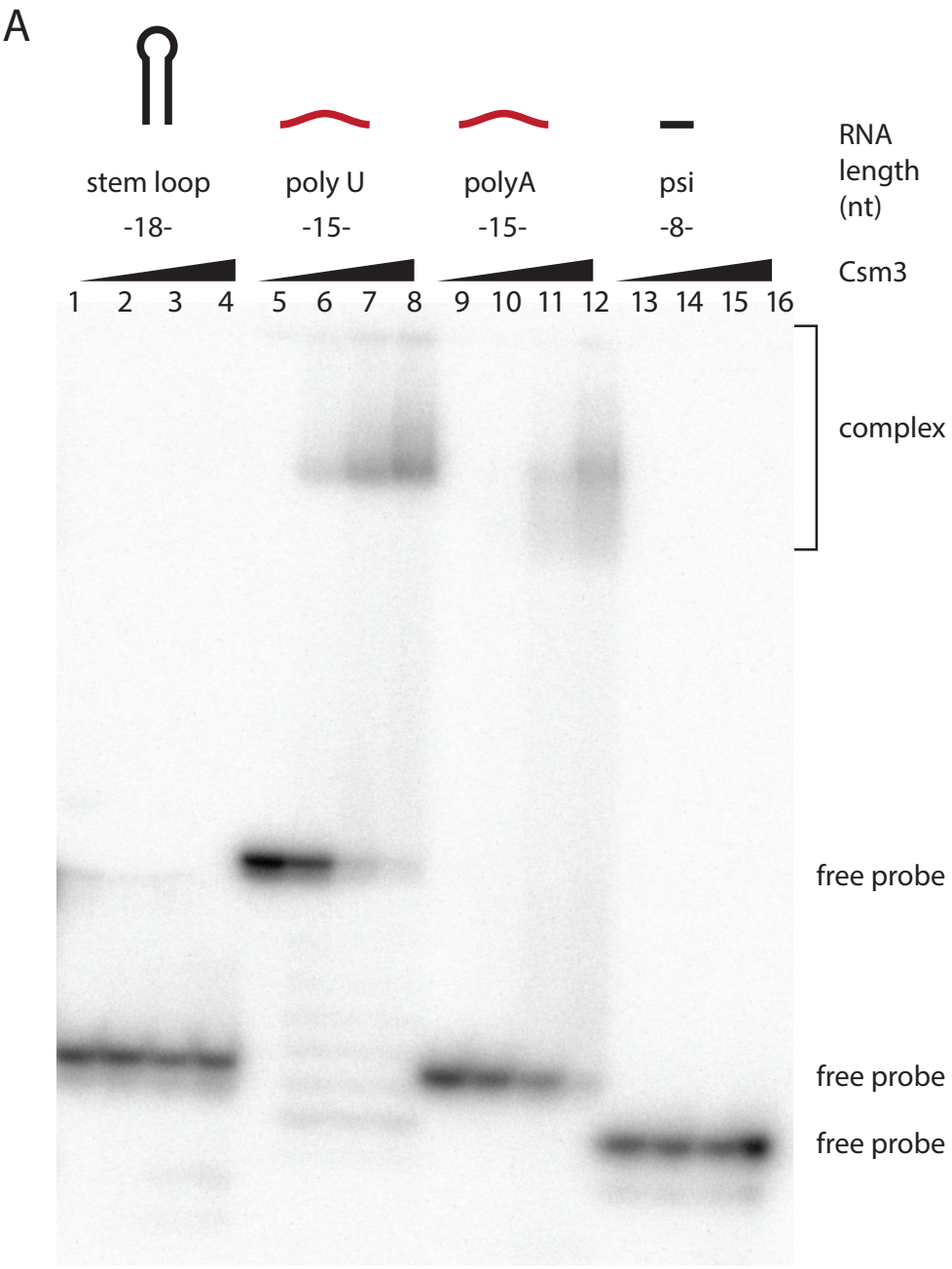
A



B

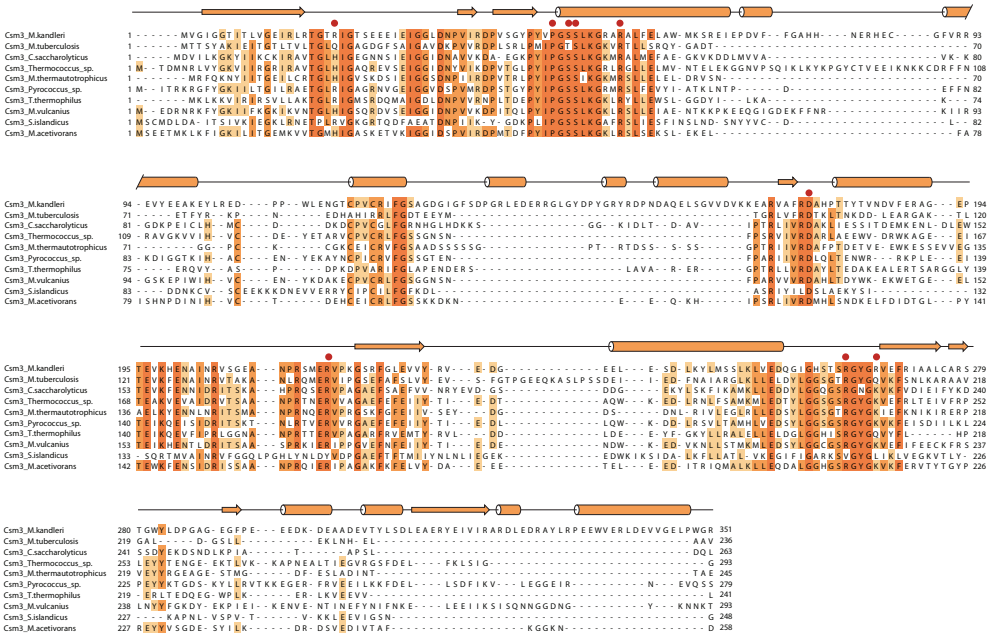


Supplementary Figure 3

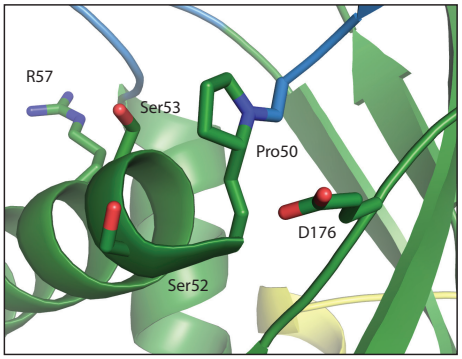


Supplementary Figure 4

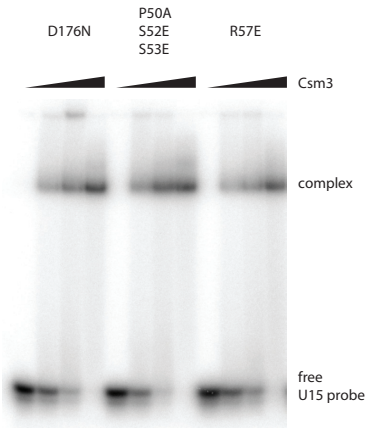
A



B



C



2.2 Publication 2: Structural analyses of the CRISPR protein Csc2 reveal the RNA-binding interface of the type I-D Cas7 family

Hrle, A., Maier, L.-K., Sharma, K., Ebert, J., Basquin, C., Urlaub, H., Marchfelder, A. & Conti, E. Structural analyses of the CRISPR protein Csc2 reveal the RNA-binding interface of the type I-D Cas7 family. RNA Biol. (2014)

The manuscript 'Structural analyses of the CRISPR protein Csc2 reveal the RNA-binding interface of the type I-D Cas7 family' presents a novel high-resolution protein structure at 1.8 Å of the Cas7 protein of *Thermophilum pendens*, the first representative of type I-D. The detailed structural homology comparison of *Tp* Csc2 with well-studied Cas proteins confirmed bioinformatic predictions and categorized it into the Cas7 family. Based on all available Cas7 structures, the study highlights denominators and differences within the Cas7 superfamily. Furthermore a combination of mutation analysis, protein–RNA cross-linking, mass spectrometry and fluorescence anisotropy and identified the conserved RNA binding surface and contributing amino acid residues.

RESEARCH PAPER

RNA Biology 11:8, 1072–1082; August 2014; © 2014 Taylor & Francis Group, LLC

Structural analyses of the CRISPR protein Csc2 reveal the RNA-binding interface of the type I-D Cas7 family

Ajla Hrle[1], Lisa-Katharina Maier[2], Kundan Sharma[3], Judith Ebert[1], Claire Basquin[1], Henning Urlaub[3,4], Anita Marchfelder[2,*], and Elena Conti[1,*]

[1]Structural Cell Biology Department; Max Planck Institute of Biochemistry; Martinsried, Germany; [2]Biology II; University of Ulm; Ulm, Germany; [3]Bioanalytical Mass Spectrometry Group; Max Planck Institute for Biophysical Chemistry; Göttingen, Germany; [4]Bioanalytics; Institute of Clinical Chemistry; University Medical Center Göttingen; Göttingen, Germany

Keywords: RRM domain, CRISPR, prokaryotic immune system, Cas7, RNA binding

Abbreviations: CRISPR, Clustered regulatory short interspaced palindromic repeats; Cas, CRISPR-associated; dCASCADE, interference complex subtype I-D; eCASCADE, interference complex subtype I-E; crRNA, CRISPR RNA; RAMP, Repeat associated mysterious protein; RRM, RNA recognition motif; RNAi, RNA interference; H1 and H2 and H1-2, β-hairpins of insertion domain 1 (or lid domain); Tp, Thermofilum pendens; Ss, Sulfolobus solfataricus; Mk, Methanopyrus kandleri; Rmsd, Root mean square deviation; SAD, Single-wavelength anomalous dispersion.

Upon pathogen invasion, bacteria and archaea activate an RNA-interference-like mechanism termed CRISPR (clustered regularly interspaced short palindromic repeats). A large family of Cas (CRISPR-associated) proteins mediates the different stages of this sophisticated immune response. Bioinformatic studies have classified the Cas proteins into families, according to their sequences and respective functions. These range from the insertion of the foreign genetic elements into the host genome to the activation of the interference machinery as well as target degradation upon attack. Cas7 family proteins are central to the type I and type III interference machineries as they constitute the backbone of the large interference complexes. Here we report the crystal structure of *Thermofilum pendens* Csc2, a Cas7 family protein of type I-D. We found that Csc2 forms a core RRM-like domain, flanked by three peripheral insertion domains: a lid domain, a Zinc-binding domain and a helical domain. Comparison with other Cas7 family proteins reveals a set of similar structural features both in the core and in the peripheral domains, despite the absence of significant sequence similarity. *T. pendens* Csc2 binds single-stranded RNA in vitro in a sequence-independent manner. Using a crosslinking - mass-spectrometry approach, we mapped the RNA-binding surface to a positively charged surface patch on *T. pendens* Csc2. Thus our analysis of the key structural and functional features of *T. pendens* Csc2 highlights recurring themes and evolutionary relationships in type I and type III Cas proteins.

Introduction

CRISPR (clustered regularly interspaced short palindromic repeats) confer an adaptive prokaryotic defense mechanism that recognizes and inactivates foreign genetic elements,[1] a mechanism that is functionally reminiscent of the eukaryotic RNA interference (RNAi) pathway.[2,3] In contrast to RNAi, CRISPR establishes a genetic memory of previously encountered pathogens that is accessed upon re-infection. Foreign nucleic acid sequences (spacers) derived from viruses or conjugative plasmids are integrated into the host genome.[4,5] The unique spacers are located within a CRISPR locus and interspersed by a series of identical host repeat sequences.[6] The CRISPR locus is transcribed into a precursor RNA that is subsequently processed to yield the mature functional crRNAs.[7] Adjacent to the CRISPR locus are genes encoding the protein machinery behind this response.[8] Upon infection, Cas (CRISPR-associated) proteins mediate spacer acquisition,[4,5] crRNA biogenesis,[9] target interference and degradation.[10]

The CRISPR-Cas systems have been classified into three major types (I, II and III), that can be further divided into at least 10 subtypes.[11] The classification of Cas proteins is hampered by the fact that even proteins with the same function have very little sequence similarity.[12,13] Therefore structural data are indispensible for accurate classification. The diversity within the CRISPR protein machinery is believed to have evolved out of the demand to respond to the specific nature of the pathogen as well as the environment of the host cell (such as thermophilic, mesophilic, halophilic, etc.). The majority of the Cas proteins contain RAMP (repeat-associated mysterious protein) domains.[14] These domains are based on a ferredoxin-like fold,[15] similar to that of an RRM (RNA-recognition motif), a ubiquitous RNA-binding

*Correspondence to: Elena Conti; Email: conti@biochem.mpg.de; Anita Marchfelder; Email: anita.marchfelder@uni-ulm.de
Submitted: 05/02/2014; Revised: 07/07/2014; Accepted: 07/11/2014
http://dx.doi.org/10.4161/rna.29893

domain.[16] However, the RRM-like domains in Cas proteins differ from those of canonical RRM domains.[16,17] First, they generally do not share the conserved consensus sequences that are involved in RNA binding in canonical RRM domains.[16-19] Second, they are structurally much more variable,[20] as they feature longer insertions[18,19,21] and extensions at the N- and C-termini.[22,23] This variation is reflected in their different functions, which range from having a structural role[18] to harboring catalytic activity.[21,24-26]

Insight into how Cas proteins assemble the functional interference complexes has been provided by electron-microscopy studies of type I-A[18]/I-E[27]/I-F[28] and III-A[29]/B[30,31] interference complexes and high resolution X-ray crystallography structures of single proteins.[20] Characteristic of type I and type III systems are members of the Cas7 family of proteins, which constitute the core subunit of the interference complexes. Multiple copies of Cas7 assemble in a helical fashion around the processed crRNA[18,27,28,32] and mediate interactions with further factors, which ultimately define complex length, activity and target recognition. To date, little information is available on the subtype I-D proteins and the associated *d*CASCADE interference complex. Here, we have studied *Thermofilum pendens* Csc2, a subtype I-D protein of the Cas7 family. Subtype I-D is commonly present in Archaea[33] and Cyanobacteria.[34] It harbors characteristic features of both subtypes I and III: a type I HD nuclease domain is fused to Cas10, the signature protein of type III. The general domain organization of CASCADE proteins is predicted to resemble type III proteins,[33] emphasizing the prominent role of this subtype as an evolutionary link between types I and III. We report the insights we obtained from the crystal structure and biochemical analysis of *Thermofilum pendens* (*Tp)* Csc2. The comparison of type I-D *Tp*Csc2 with type I-A *Sulfolobus solfataricus* (*Ss)* Csa2[18] and type III-A *Methanopyrus kandleri (Mk)* Csm3[19] allows building a comprehensive picture of the Cas7 protein family and its conserved RNA-binding properties.

Table 1. Data Collection and Structure Refinement Statstics of *Tp* Csc2

Data collection		
	Native *Tp* Csc2	SeMet *Tp* Csc2
Space group	P 2 21 21	P 2 21 21
Unit cell (Å)[a]	a = 60.47	a = 60.81
	b = 80.95	b = 81.24
	c = 112.60	c = 114.02
Resolution range (Å)[a]	46.22–1.82 (1.88–1.82)	48.68–2.37 (2.46–2.37)
Unique reflections[a]	50416 (7188)	23518 (2402)
I/σ (I)[a]	17.8 (1.6)	31.9 (6.4)
Multiplicity[a]	6.5 (6.0)	13.1 (12.6)
R_{merge} (%)[a]	6.7 (97.7)	7.3 (43.4)
CC(1/2) (%)[a]	99.9 (50.5)	99.9 (95.4)
Refinement		
Average B-factor	32.70	34.28
R_{work} (%)	18.15 (31.75)	20.85 (24.14)
R_{free} (%)	21.21 (34.64)	23.72 (25.13)
Rmsd bonds (Å)	0.017	0.004
Rmsd angles (°)	1.36	0.789
Ramachandran favored (%)	97.0	96.7
Ramachandran outliers (%)	0.0	0.0

[a]Values in parentheses correspond to the highest resolution shell; SeMet: Selenomethionine derivatized protein.

Structure Determination of Csc2

We expressed full-length *T. pendens* (*Tp*) Csc2 (374 residues) in *E. coli* and purified it to homogeneity. *Tp* Csc2 yielded crystals in an orthorhombic space group (*P*222), containing one molecule per asymmetric unit and diffracting beyond 1.8 Å resolution. X-ray fluorescence scans on the crystals showed a peak at the Zinc excitation, suggesting the presence of a bound zinc ion in the crystallized protein. We obtained phases by crystallizing the selenomethionine derivatized protein and solved the structure by single-wavelength anomalous dispersion method (SAD). The structure was refined at 1.8 Å resolution to an Rfree/Rwork of 21/18%. The final atomic model has good stereochemistry (**Table 1**) and includes most of the polypeptide. Disordered regions include a loop between residues Gln134 and Gly147, the four N-terminal residues and 21 C-terminal residues.

Csc2 has a Central RRM-Like Core Domain with Three Elaborate Insertion Domains

The overall architecture of *Tp* Csc2 can be described as composed of four domains (**Fig. 1A**). At the core of the molecule is a domain with β–α–β–β–α–β topology reminiscent of a RRM fold (**Fig. 1B**). The four β-strands form a twisted β-sheet, with two α–helices (α1, α2) resting against a concave groove. Strands β1 and β3 of the core domain lack residues of the so-called RNP2 and RNP1 motifs, which are required for RNA binding in canonical RRM domains. In addition, the canonical RNA-interacting interface of the RRM fold is obstructed from the solvent by an α helix (αE). Overall, a large part of the core domain is inaccessible to solvent. The most exposed structural element is helix α1. Helix α1 contains conserved residues and contacts a conserved glycine-rich loop between helix α2 and strand β4. The presence of a rather flexible glycine-rich loop at this structural position is a characteristic feature in the non-canonical RRM folds of the Cas superfamily, although its exact function remains elusive.

The core domain is flanked by three peripheral domains that are composed of elaborate insertions originating from the secondary structure elements of the core. The first insertion domain (insertion domain 1 or lid domain) is formed by three β-hairpins (H1, H2 and H1–2) that create a lid on top of the core (**Fig. 1A and 1B**). Hairpin H1 is formed between the β1–α1 elements of the RRM-like domain and contacts both strand β2 of the core and helix αE. Hairpin H2 is formed between the β2–β3 elements of the RRM-like domain. H2 features a sharp bend at the tip (where Gly205 is located) and a hinge at the bottom (where the residues Pro196 and Gly211 are located). The bottom of H2 packs against hairpin H1–2. Hairpin H1–2 constitutes the base of the lid domain and is formed by both the β1–α1 and β2–β3 segments of the RRM-like domain. This hairpin effectively

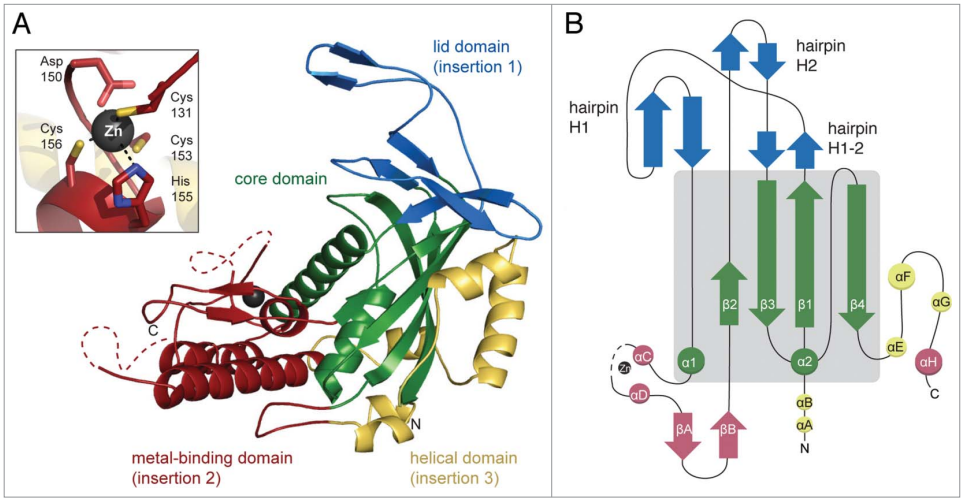


Figure 1. Crystal Structure of *Thermofilum pendens* Csc2. (**A**) Structure of *Tp* Csc2 can be divided into four distinct domains: a core domain (green), a lid domain (insertion 1, blue), a metal-binding domain (insertion 2, red) and a helical domain (insertion 3, yellow). Secondary structure elements of the core adopt a ferredoxin-like fold with β-α-β-β-α-β arrangement. Multiple insertions within the core define the accessory domains. Dashed lines indicate the disordered loops. The inset shows a detailed view of the zinc ion (gray sphere) with coordinating residues. (**B**) Topology diagram of *Tp*Csc2. α-Helices are represented as circles and β-strands arrows. The secondary structure elements have been labeled numerically maintaining the nomenclature of RRM domains. The hairpins of insertion domain 1 are labeled as described in the text (H1, H2 and H1–2). The α-helices of in the insertion domains are labeled with letters (αA to αH).

extends the β-strands β1 and β3 of the core, after a sharp bend created by the conserved residues Pro222 and Gly223 (**Fig. 2**).

The insertion domain 2 (or metal-binding domain) is defined by an 80 amino-acid long segment between α1–β2 and the very C-terminal helix αH (**Fig. 1B**). αH is an elongated helix, embedded within a predominantly hydrophobic cavity lined by the helices αC and αD. This domain coordinates a Zinc ion via the residues Cys131, Cys153, His155, Cys156 and in addition Asp150 (**Fig. 1A**). Metal binding appears to have a structural role, maintaining the close packing within the domain. The insertion domain 3 is a helical domain formed by three helices (αE, αF and αG) that are between the last β-strand of the core domain and the C-terminal helix αH. Insertion domain 3 contacts secondary structure elements of the core domain and, together with the small N-terminal helices αA and αB, it wraps around the convex surface of the β-sheets and helix α2 (**Fig. 1A**).

Csc2: a Cas7 Family Protein

Bioinformatic predictions categorized *Tp* Csc2 within the Cas7 protein family and suggested it as the Cas7 homolog in subtype I-D interference complexes.[11] We compared the structure of *Tp* Csc2 with those of *Sulfolobus solfataricus* (*Ss*) Csa2 (3PS0)[18] and *Methanopyrus kandleri (Mk)* Csm3 (4NOL)[19] (**Fig. 3A and 3B**). *Ss* Csa2 and *Mk* Csm3 are Cas7 homologs in the subtypes I-A and III-A and share a sequence identity of 9% and 20%, respectively, with *Tp* Csc2. The RRM-like fold of *Tp* Csc2 (76 amino acids) superposes with the respective domains with a root mean square deviation (rmsd) of 1.5 Å for *Mk* Csm3 and 3.0 Å for *Ss* Csa2. The main difference in the core domains is that *Tp* Csc2 lacks the fifth β-strand that is characteristic of the β-sheet of *Mk* Csm3 and *Ss* Csa2 (**Fig. 3A**).

Tp Csc2, *Ss* Csa2 and *Mk* Csm3 have insertions at equivalent positions within the core domain. Structure-based comparisons suggest that the peripheral insertion domains also share similarities. The lid domain (insertion 1) is in all cases the most flexible part of the molecule (**Fig. 3B**). A common structural feature of the lid domain is the β-hairpin corresponding to *Tp* Csc2 H1 (structural element 1 in **Figure 3B**), which protrudes toward the front of the RRM-like core. Insertion domain 2 is in all cases a predominantly α-helical domain. In both *Tp* Csc2 and *Mk* Csm3, this domain contains a structural Zinc ion (**Fig. 3A and 3B**). In the case of *Ss* Csa2, insertion domain 2 does not require a metal ion to be folded. A common structural feature of insertion domain 2 in the three structures is helix αD, which is buried in the heart of the proteins (structural element 2 in **Figure 3B**). Another conserved feature is a helix that connects this insertion domain to the core domain

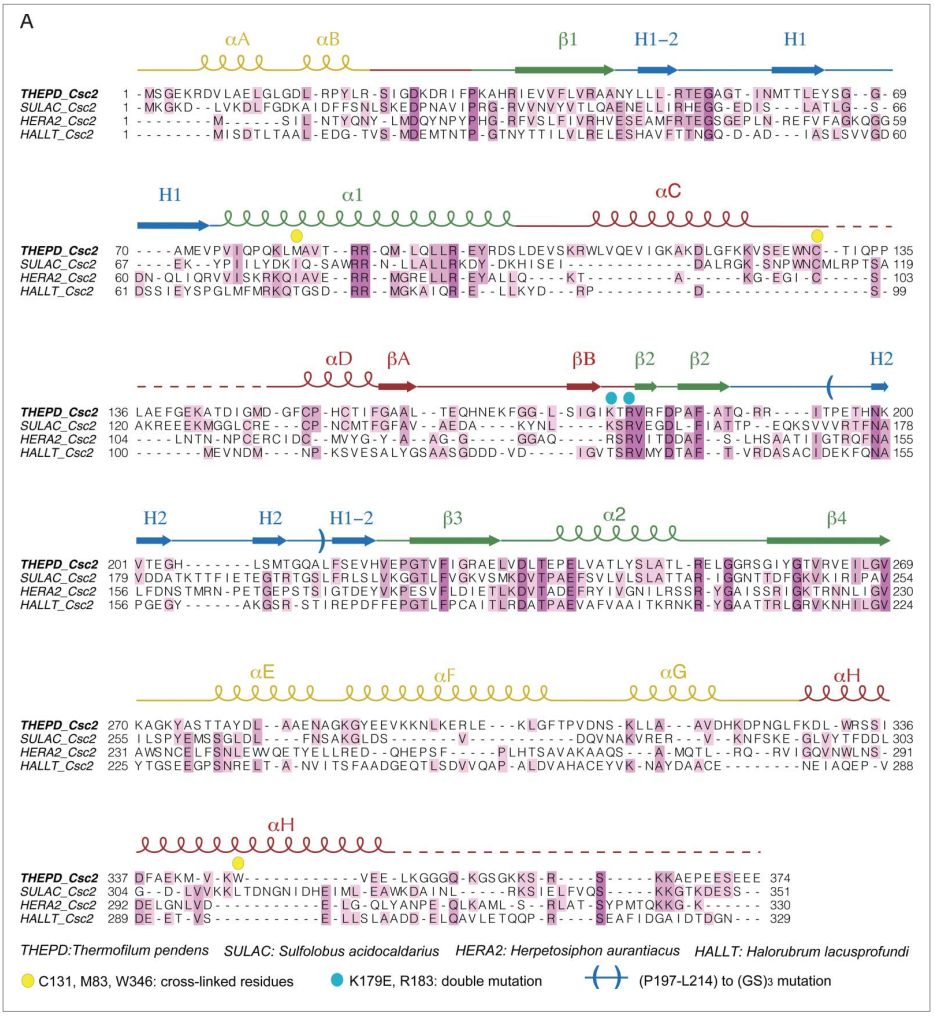


Figure 2. Structure-based sequence alignment of *Tp* Csc2. The alignment includes four sequences from representative species of the Csc2 family, based on a comprehensive alignment. Secondary structure elements are indicted by the cartoon above the sequences, color-coded and labeled according to **Figure 1A**. Colors represent the percentage of sequence identity (dark > 60%, light 60–30%). U15 cross-linked residues are highlighted with yellow dots. Blue dots above the K179 and R183 mark the mutated amino acids, brackets indicate the boundaries of the sequence spanning (P197-L214), which was replaced by $(GS)_3$ (Δloop mutant).

(structural element 3 in **Figure 3B**). In the case of *Tp* Csc2, this structural element corresponds to the C-terminal helix αH. In the case of *Mk* Csm3 and *Ss* Csa2, this structural element is derived from a topologically different region. Similarly, helices at the base of the core domain are present in all three proteins but are derived from different elements (structural element 4 in **Figure 3B**).

Generally, the structural conservation among the Cas7 proteins is not reflected in high sequence similarities. The exception is a solvent-exposed platform formed by the segment preceding β2 at the interface between the core domain and insertion domain 2 (structural element 5 in **Figure 3C**), which is not only conserved at the structural but also at the sequence level.

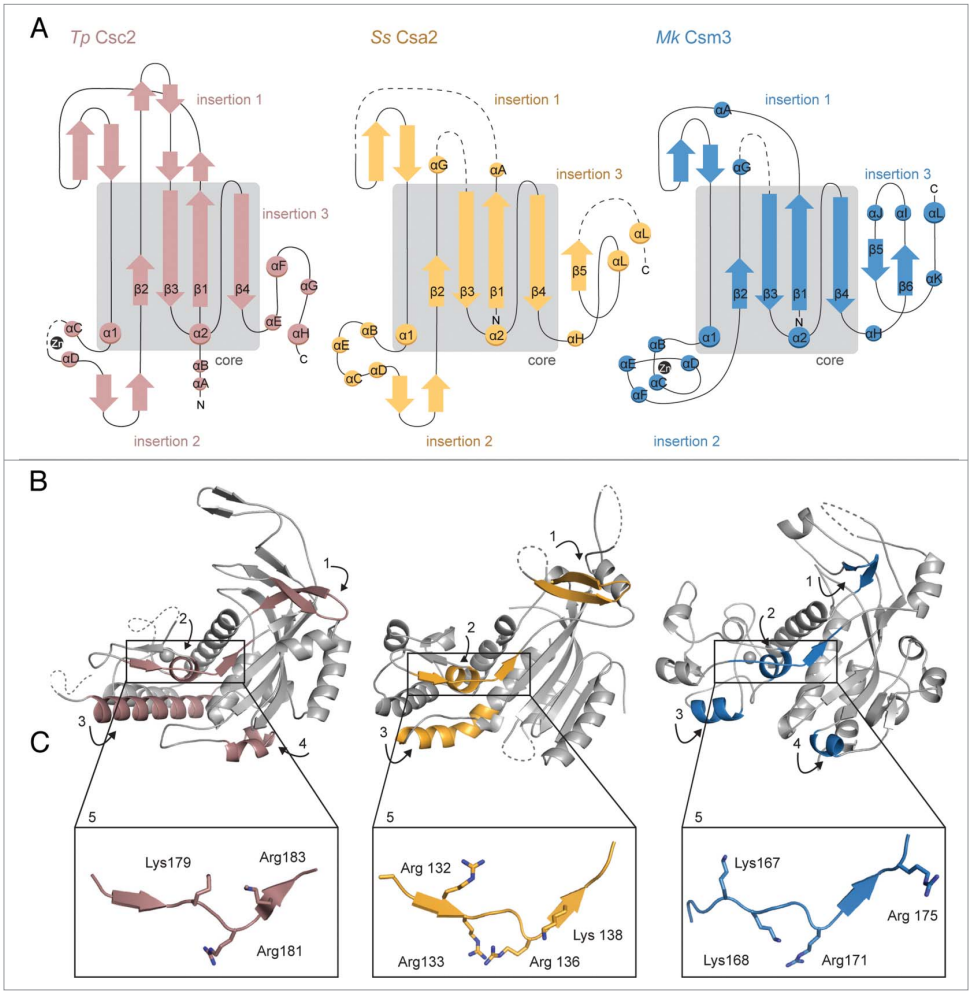


Figure 3. Structural comparison of Cas7 proteins. (**A**) Topology diagrams of *Tp* Csc2, *Ss* Cas7 and *Mk* Csm3 highlight the high structural conservation within the core RRM-like fold (boxed in gray) and show the connectivity of the insertion domains. The topological arrangement of the insertions 1–3 is similar in all proteins. Variations within secondary structure elements of the three proteins reflect subtype specificities. (**B**) Crystal structures of Cas7 orthologs, *Tp* Csc2, Ss Csa2, *Mk* Csm3, depicted according to the orientation in **Figure 1A** after optimal superposition of their RRM-like domains. The molecules are overall colored in gray. Significant structural similarities are colored according to the color-code of the respective proteins, *Tp* Csc2 (salmon), Ss Csa2 (orange), *Mk* Csm3 (blue). Numbers (1–5) refer to the significant structural elements discussed in the text. Dashed lines indicate the structurally unresolved loops. (**C**) Boxes highlight the structurally and sequence-conserved basic residues along β2 and the preceding insertion.

In order to confirm our classification and pervious bioinformatic analysis, we compared *Tp* Csc2 to well-studied Cas protein families (Cas6 and Cas5), based on the structural analysis by Reeks et al. Although Cas5 and Cas6 proteins also contain an RRM-like domain as a core structural element and a glycine-rich flexible region between α2 and β4, the peripheral domains diverge. First, the β hairpin between β2-β3 is part of the lid domain in *Tp* Csc2 and other Cas7 family members, while in Cas5d it can be seen as an extension of the RRM β strands, in Cas6 it is present in both RRM domains. Second, the insertion

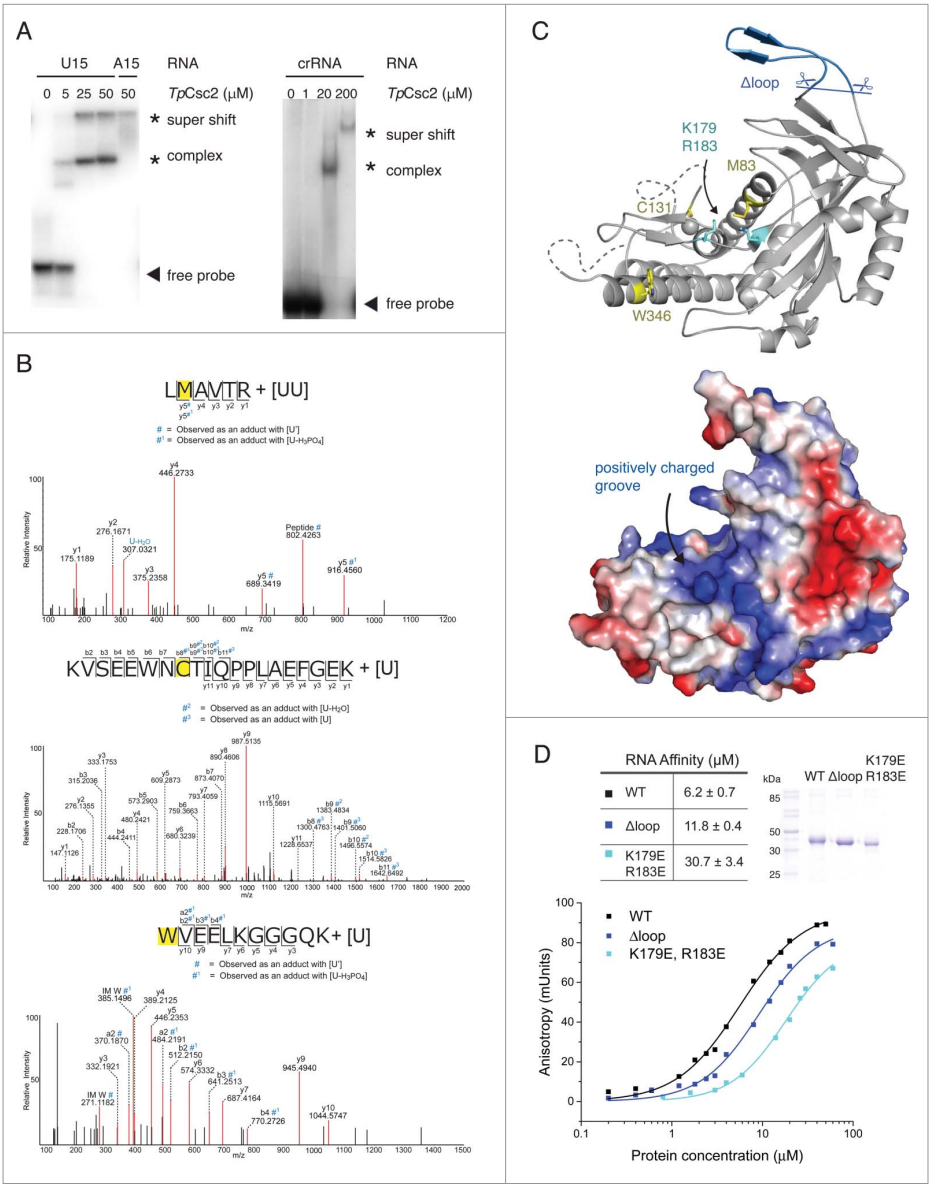


Figure 4. For figure legend, see page 1078.

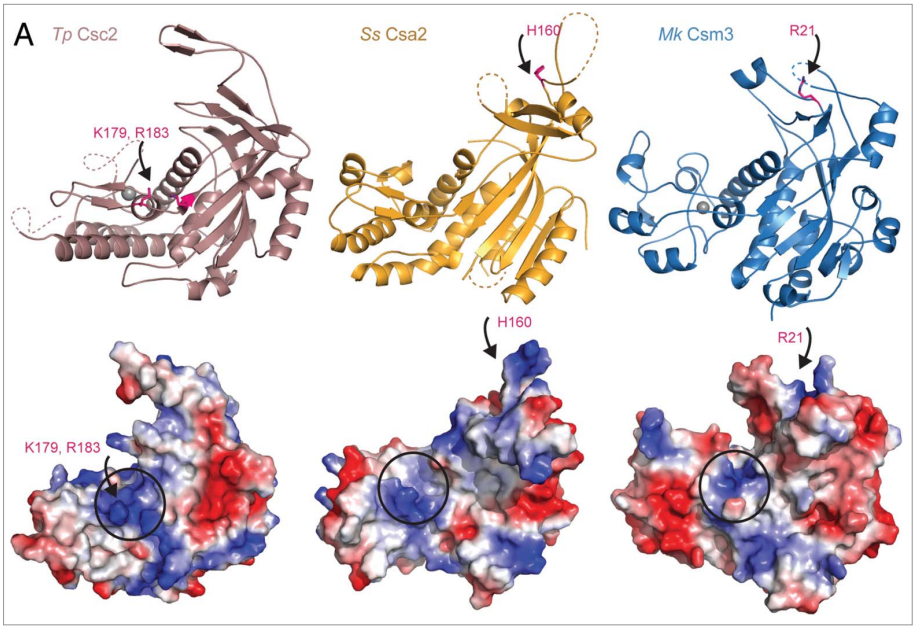


Figure 5. A structurally and functionally conserved surface groove in the Cas7 protein family. Upper panels: cartoon representation of the structures of the Cas7-like proteins *Tp* Csc2 (salmon), Ss Csa2 (orange), *Mk* Csm3 (blue) (in the same orientations as in **Figure 3B**). Lower panels: corresponding surface representations showing the electrostatic potential (red for electronegative and blue for electropositive). For all proteins positively charged patches are present at the interface between the core and insertion domain 2 (identified with a circle). Conserved lysines and arginines contribute to these patches (**Fig. 2**) and in Tp Csc2 (arrows point to pink residues) are involved in RNA binding (**Fig. 4**). Residues reported to have an effect on RNA binding in *Ss* Csa2 and *Mk* Csm3 (arrows point to pink residues) are located within positively charged surfaces of the respective lid domains.

domain 2 that is present between α1 and β2 in *Tp* Csc2 and other Cas7 proteins is absent in the Cas6 and Cas5 family. Third are the structurally variable C-terminal domains, which consist of a second RRM domain in the cases of most Cas6 proteins and an extended β-hairpin in Cas5d representatives. Despite the unifying RRM-like core, the structural variation of the peripheral domains might reflect the different RNA binding requirements.

Mapping the RNA-Binding Interface

Proteins of the Cas7-family assemble around processed crRNA and constitute the backbone of the interference complex.[27,29,31] Several Cas7 monomers are involved in binding to the variable pathogen derived spacer region, exposing it to the complementary target DNA.[27] We sought to determine the RNA-binding

Figure 4 (See previous page). Mapping the RNA-binding surface of *Thermofilum pendens* Csc2. (**A**) Electrophoretic mobility shift assays (EMSA) with wild-type *Tp* Csc2. Left panel: EMSA were performed with ^{32}P -5′-end labeled poly(U)$_{15}$ or poly(U)$_{15}$ RNAs and increasing concentrations of *Tp* Csc2 (0, 5, 25, 50μM). The positions of the free RNA probe (arrow head) and of the RNA-bound complexes (asterisks) are shown on the right. Right panel: EMSA assay with *Tp* Csc2 and ^{32}P -5′-end labeled crRNA. (**B**) MS/MS mass spectra of *Tp* Csc2 peptides, carrying an additional mass corresponding to one (panel two and three) or two (panel one) uracil nucleotides associated with the respective amino-acid. Peptide sequence and the fragment ions are indicated on top. The direcly crosslinked residues are colored yellow. The peptide fragmentation occurs with the cleavage of amide bonds resulting in b-ions and y-ions when the charge is retained by the N-terminal and C-terminal fragments, respectively. #, #,^{1} #^{2} and #^{3} indicate the b- and y-ions that were observed with a mass shift corresponding to U′, U-H_3PO_4, U-H2O and U, respectively. IM: Immonium ions. U′: U marker ion adduct of 112.0273 Da. (**C**) Mapping RNA-binding properties on the *Tp* Csc2 crystal structure. Upper panel: a cartoon representation of *Tp* Csc2 is shown in gray (in the same orientation as in **Figure 1A**) with the crosslinked residues colored in yellow (stick representation) and regions targeted for mutagenesis colored in blue (K178E/R183E and Δloop, indicated with scissors, stick representation). Lower panel: surface representation of *Tp* Csc2 (in the same orientation as in panel C) depicting the electrostatic potential (red for electronegative and blue for electropositive). (**D**) Quantitative measurements of RNA-binding affinities. Upper panel: 13% SDS-PAGE with the wild-type (WT) an mutant proteins used in the fluorescence anisotropy (FA) assay and a table with the K_d values obtained. The Δloop mutant was engineered by replacing the segment between Pro197 and Leu214 with a $(GS)_3$ sequence. Lower panel: FA measurements of WT and mutant *Tp* Csc2 with a 5′-6-carboxy-fluorescein-labeled poly(U)$_{15}$ -RNA.

interface in *Tp* Csc2. In electrophoretic mobility shift assays (EMSAs), *Tp* Csc2 bound a polyU RNA of 15 nt length (poly $(U)_{15}$ and poly$(A)_{15}$, (surrogates of the variable spacer sequence) in a comparable fashion as previously shown for *Mk* Csm3[19] (**Fig. 4A**, left panel). At increasing concentrations, *Tp* Csc2 binds likely in multiple copies on the 15-mer RNA oligonucleotides, as shown by the supershift in the gel (**Fig. 4A**, left panel, concentration 25 μM and 50 μM). A similar behavior was observed upon binding to the *Tp* crRNA (**Fig. 4A**, right panel).

Next, we sought to identify RNA binding interface of the protein using a crosslinking - mass spectrometry approach. We incubated *Tp* Csc2 with poly$(U)_{15}$ RNA, and cross-linked the complex by subjecting it to UV irradiation at 254 nm. We used LC-MS/MS mass spectrometry to detect and sequence peptides cross-linked to an RNA nucleotide as previously described.[35] UV crosslinks favorably occur with sulfur-containing or aromatic side chains that are in close proximity to the nucleic acid, although not necessarily in direct contact. The mass spectrometric analysis identified three modified peptides (**Fig. 4B**). The reactive residues that were directly conjugated to a uridine (Cys131, Met83 and Trp346, **Figure 4C**, upper panel) encircle a central positively-charged patch on the surface of the protein, suggesting that this region mediates RNA binding (**Fig. 4C**, lower panel). Moreover, this patch is adjacent to the conserved segment preceding β2 that is enriched in lysine and arginine residues (structural element 5 in **Figure 3C**).

We targeted surface exposed regions of *Tp* Csc2 for mutagenesis and used quantitative RNA-binding experiments to compare the mutants to the wild-type protein (**Fig. 4D**). In fluorescence anisotropy experiments, *Tp* Csc2 bound a poly$(U)_{15}$ RNA with an affinity in the low micro-molar range (**Fig. 4D**, black curve in lower panel). Reverse-charge mutations of two positively charged residues in structural element 5 (K179E/R183E) resulted in a 6-fold reduction of RNA-binding affinity as compared with the wild-type protein (**Fig. 4D**, light blue curve in the lower panel), consistently with the information form the structural and mass-spectrometry analyses. Importantly, a positively charged surface groove is present at the equivalent structural position (between the core RRM-like domain and insertion domain 2) in the structures of the Cas7 family proteins *Ss* Csa2 and *Mk* Csm3 (**Fig. 5**). This groove corresponds to the predicted site for crRNA binding on the Cas7 family protein CasC deduced from the interpretation of the cryo-EM structure of the *e*CASCADE[27] complex (**Fig. S1**).

In the case of *Ss* Csa2[18] and *Mk* Csm3,[19] the lid domain is involved in nucleic-acid binding (**Fig. 5**). For *Tp* Csc2 we did not identify direct RNA-binding residues from the lid by mass-spectrometry. Replacing the 18 residues that shape the tip of the lid domain (between Pro197 and Leu214) with a generic $(Gly\text{-}Ser)_3$ linker (**Fig. 4C**, upper panel) resulted in a modest reduction of poly$(U)_{15}$ RNA-binding affinity as compared with the wild-type protein (**Fig. 4D**, dark blue curve in the lower panel). We conclude that the influence of the lid domain in crRNA recognition or crRNA directed target DNA recognition depends on the specific Cas7 protein family, while the positively-charged surface groove between the core and the insertion 2 domain appears to be a conserved functional site.

Conclusions

A common structural feature of many Cas proteins is the central RRM-like fold and the presence of peripheral insertion domains. The structural diversity within these peripheral domains is thought to be responsible for the multitude of Cas protein functions.[20] Computational[12,17] and biochemical[18,19] studies have contributed to classifying the Cas7 proteins. Nevertheless weak sequence homology makes structural data indispensable to enable a complete understanding of their structure-to-function relationship. In this study, we solved the structure of *Tp* Csc2 and confirmed the classification of the Csc2 protein as a Cas7 protein of the subtype I-D. The structural similarity among *Tp* Csc2 and known Cas7 proteins, such as *Mk* Csm3 and *Ss* Csa2, encompasses the central RRM-like core domain as well as the arrangement of the insertion domains. *Tp* Csc2 and *Mk* Csm3 share higher sequence and structural similarities. This is in line with previous reports suggesting that Csc2 proteins resemble their type III Cas7 counterparts.[15] Despite lacking significant sequence similarity, the structural features as well as the charge distribution are strongly conserved within type I orthologs. Subtype specificities are reflected by the variations within the topology, structural composition and flexibility of the peripheral domains, such as the absence or presence of metal coordination and different secondary structure elements within the lid domain.

The basic physiological role of Cas7-like proteins of type I interference complexes[18] as well as its homologs, Csm3 in type III-A and Cmr4 in type III-B systems,[29,36] is to bind the variable crRNA spacer sequence and with it constitute a platform for stable binding of target DNA. In this study, we have defined the RNA binding interface of *Tp* Csc2. We show that *Tp* Csc2 binds variable sequences of ssRNA. The affinity toward the RNA is within the low micro-molar range, weak yet significant, and in line with the protein's physiological function. We further investigated the RNA-binding properties of *Tp* Csc2 using crosslinking mass spectrometry and structure-based mutation analyses. Here, we identified the RNA-protein interface and pinpointed functionally relevant residues. We show that *Tp* Csc2 possesses a critical, positively charged groove formed by conserved residues in the interface between the core and the second insertion domain. This feature is largely conserved among characterized protein family members. Our findings are in accordance to the predicted crRNA-binding interface of CasC, the *E. coli* Cas7 ortholog.[27] Therefore we speculate that upon *d*CASCADE assembly, multiple copies of *Tp* Csc2 may adopt a similar arrangement within the complex as CasC in the *e*CASCADE, defining a channel and exposing the central positively charged groove toward the solvent (**Fig. S1**).

Taken together our study highlights the evolutionary relationship within the Cas7 protein family and helps to better

understand the RNA-interacting features that are conserved among the Cas7 proteins. Further structural studies will identify the contribution of the insertion domains on protein interactions during *d*CASCADE assembly.

Accession Numbers

The coordinates and the structure factors of *Thermofilum pendens* Csc2 have been deposited in the protein Data Bank with the accession code 4TXD.

Experimental Procedures

Protein expression and purification

The gene for the Csc2 from *T. pendens* was ordered as a synthetic construct (GeneArt, Life technologies). His- and His-SUMO tagged *Tp* Csc2 proteins were expressed using *E. coli* BL21-Gold (DE3) Star pRARE cells (Stratagene) grown in TB medium and induced overnight at 18°C. The cells were lysed in buffer A (50 mM TRIS-HCl pH 7.5, 1 M NaCl, 10 mM imidazole, 10% glycerol) supplemented with protease inhibitors (Roche). The lysate was heated to a temperature of 55°C for 10 min and subsequently centrifuged at 25000 g. The protein was purified from the resulting supernatant at room-temperature by Ni^{2+}- affinity chromatography as an initial step and further purified over a HiTrap Heparin column (GE Healthcare) to remove minor contaminants. The His-tag was removed by treatment with SUMO protease. Size-exclusion chromatography (SEC) on a Superdex 75 column (GE Healthcare) was performed as a final step of purification using buffer B (20 mM HEPES pH 7.5, 150 mM NaCl and 5 mM DTT and 10% glycerol). Selenomethionine derivatized protein was purified as described above from *E. coli* grown in M9 media complemented with the essential amino acids and selenomethionine.[37]

Crystallization and structure determination

Native crystals of *Tp* Csc2 were grown at 20 °C by sitting-drop vapor diffusion from drops formed by equal volumes of protein (at 9.5 mg/ml) and of crystallization solution containing 0.05 M Mes pH 5.6, 0.2 M KCl, 0.01 M $MgCl_2$, 5% Peg 8000 and 17% glycerol. Crystals were cryoprotected with a final concentration of 20% glycerol prior to data collection. Selenomethionine derivatized crystals were obtained in similar conditions and cryo-protected as described above.

Native and SAD data were collected at the PXII and PXIII beamlines of the Swiss Light Source (SLS) (Villigen, Switzerland), respectively. Data were processed with XDS[38] and scaled using Aimless.[39] Selenium sites were located and experimental phases were calculated using the AutoSol pipeline in Phenix.[40] Model building and refinement were performed with COOT[41] and Phenix and the final model was validated using Molprobity.[42]

Biochemical assays

The RNA molecules poly$(U)_{15}$, poly$(A)_{15}$ and the crRNA (spacer 1 of locus 2, sequence: ACUAAGAGCC UCCUUUGCCC ACGGCAUCGG UAGGUCAGGU CCACGUUCAA AAUCAGCAAG)[43] were synthesized by Purimex. The poly$(U)_{15}$, and poly$(A)_{15}$ RNA were 5′ labeled with T4 polynucleotide kinase (New England Biolabs) and γ-^{32}P ATP (Perkin-Elmer), the cRNA was 3′ labeled with a-^{32}P-pCp (Hartmann Analytic) and T4 RNA ligase (Fermentas, Thermofisher Scientific). For the gel-shift assays using poly$(U)_{15}$ and poly$(A)_{15}$, 0.5 pmol labeled RNA was mixed with 5 μM, 25 μM, and 50 μM *Tp* Csc2 protein in 10 μL reaction volume containing 20 mM Hepes at pH 7.5, 100 mM KOAc, 4 mM $Mg(OAc)_2$, 0.1% (vol/vol) NP-40 and 2 mM DTT. For crRNA gel-shift assays, 0.14 pmol labeled RNA was mixed with 1 μM, 20 μM and 200μM *Tp* Csc2 protein.

The mixtures were incubated for 20 min at 55°C before adding 2 μL 50% (vol/vol) glycerol containing 0.25% (w/vol) xylene cyanole. Samples were separated on a 8% (w/vol) polyacrylamide gel at 4°C and visualized by phospho-imaging (GE Healthcare).

Fluorescence anisotropy

Fluorescence anisotropy measurements were performed with a 5′-6-carboxy-fluorescein-labeled poly$(U)_{15}$ RNA at 20°C in 50 μl reactions on a Genios Pro (Tecan). The RNA was dissolved to a concentration of 10 nM and incubated with *Tp* Csc2 for 20 min at 55°C before adding upon measurement. The excitation and emission wavelengths were 485 nm and 535 nm, respectively. Each titration was measured three times using ten reads with an integration time of 40 μs. The data were analyzed by nonlinear regression fitting using the BIOEQS software.[44]

Crosslinking–mass spectrometry analysis

Tp Csc2 – poly$(U)_{15}$ contacts sites were investigated with mass spectrometry after UV-induced protein–RNA crosslinking as described.[35,45] The purified crosslinks were analyzed using Top10HCD method on an Orbitrap Velos instrument and the data were analyzed using OpenMS and OMSSA as a search engine (see Supplementary Experimental Procedures).

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

The authors would like to thank O. Alkhnbasi and R. Backofen for bioinformatics advice; J. Basquin, K. Valer-Saldana, and S. Pleyer at the MPI-Martinsried Crystallization Facility for crystallization experiments; the staff of the PXII and PXIII beamline at the Swiss Light Source (Villigen, Switzerland) for assistance during data collection; M. Hein and members of our labs for critical reading of the manuscript.

Funding

This study was supported by the Max Planck Gesellschaft, the DFG Research Group 1680 (FOR1680) to E.C., H.U and A.M, and CIPSM to E.C.

Author Contributions

Author contributions: A.H. performed the structural and biochemical analyses; L.M. cloned WT *Tp* Csc2 and performed the crRNA experiment in **Figure 4A**; J.E. obtained native crystals;

C.B. performed fluorescence anisotropy experiments; K.S. performed the mass-spec experiment in **Figure 4B**; E.C, A.M and H.U. supervised research; A.H. and E.C wrote the manuscript.

Supplemental Materials

Supplemental data for this article can be accessed on the publisher's website.

References

1. Marraffini LA, Sontheimer EJ. Self versus non-self discrimination during CRISPR RNA-directed immunity. Nature 2010; 463:568-71; PMID:20072129; http://dx.doi.org/10.1038/nature08703
2. Makarova KS, Wolf YI, van der Oost J, Koonin EV. Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements. Biol Direct 2009; 4:29; PMID:19706170; http://dx.doi.org/10.1186/1745-6150-4-29
3. Carthew RW, Sontheimer EJ. Origins and Mechanisms of miRNAs and siRNAs. Cell 2009; 136:642-55; PMID:19239886; http://dx.doi.org/10.1016/j.cell.2009.01.035
4. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. CRISPR provides acquired resistance against viruses in prokaryotes. Science 2007; 315:1709-12; PMID:17379808; http://dx.doi.org/10.1126/science.1138140
5. Garneau JE, Dupuis MÈ, Villion M, Romero DA, Barrangou R, Boyaval P, Fremaux C, Horvath P, Magadán AH, Moineau S. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature 2010; 468:67-71; PMID:21048762; http://dx.doi.org/10.1038/nature09523
6. Al-Attar S, Westra ER, van der Oost J, Brouns SJ. Clustered regularly interspaced short palindromic repeats (CRISPRs): the hallmark of an ingenious antiviral defense mechanism in prokaryotes. Biol Chem 2011; 392:277-89; PMID:21294681; http://dx.doi.org/10.1515/bc.2011.042
7. Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, Snijders AP, Dickman MJ, Makarova KS, Koonin EV, van der Oost J. Small CRISPR RNAs guide antiviral defense in prokaryotes. Science 2008; 321:960-4; PMID:18703739; http://dx.doi.org/10.1126/science.1159689
8. Jansen R, Embden JD, Gaastra W, Schouls LM. Identification of genes that are associated with DNA repeats in prokaryotes. Mol Microbiol 2002; 43:1565-75; PMID:11952905; http://dx.doi.org/10.1046/j.1365-2958.2002.02839.x
9. Haurwitz RE, Jinek M, Wiedenheft B, Zhou K, Doudna JA. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. Science 2010; 329:1355-8; PMID:20829488; http://dx.doi.org/10.1126/science.1192272
10. Sinkunas T, Gasiunas G, Fremaux C, Barrangou R, Horvath P, Siksnys V. Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. EMBO J 2011; 30:1335-42; PMID:21343909; http://dx.doi.org/10.1038/emboj.2011.41
11. Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, Horvath P, Moineau S, Mojica FJ, Wolf YI, Yakunin AF, et al. Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol 2011; 9:467-77; PMID:21552286; http://dx.doi.org/10.1038/nrmicro2577
12. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV. 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 2006; 1:7; PMID:16545108; http://dx.doi.org/10.1186/1745-6150-1-7
13. Brendel J, Stoll B, Lange SJ, Sharma K, Lenz C, Stachler AE, Maier LK, Richter H, Nickel L, Schmitz RA, et al. A complex of Cas proteins 5, 6, and 7 is required for the biogenesis and stability of crRNAs in Haloferax volcanii. J Biol Chem 2014;In press; PMID:24459147; http://dx.doi.org/10.1074/jbc.M113.508184
14. Wang R, Li H. The mysterious RAMP proteins and their roles in small RNA-based immunity. Protein Sci 2012; 21:463-70; PMID:22323284; http://dx.doi.org/10.1002/pro.2044
15. Makarova KS, Aravind L, Wolf YI, Koonin EV. Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. Biol Direct 2011; 6:38; PMID:21756346; http://dx.doi.org/10.1186/1745-6150-6-38
16. Maris C, Dominguez C, Allain FH. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. FEBS J 2005; 272:2118-31; PMID:15853797; http://dx.doi.org/10.1111/j.1742-4658.2005.04653.x
17. Koonin EV, Makarova KS. CRISPR-Cas: evolution of an RNA-based adaptive immunity system in prokaryotes. RNA Biol 2013; 10:679-86; PMID:23439366; http://dx.doi.org/10.4161/rna.24022
18. Lintner NG, Kerou M, Brumfield SK, Graham S, Liu H, Naismith JH, Sdano M, Peng N, She Q, Copié V, et al. Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE). J Biol Chem 2011; 286:21643-56; PMID:21507944; http://dx.doi.org/10.1074/jbc.M111.238485
19. Hrle A, Su AA, Ebert J, Benda C, Randau L, Conti E. Structure and RNA-binding properties of the type III-A CRISPR-associated protein Csm3. RNA Biol 2013; 10:1670-8; PMID:24157656; http://dx.doi.org/10.4161/rna.26500
20. Reeks J, Naismith JH, White MF. CRISPR interference: a structural perspective. Biochem J 2013; 453:155-66; PMID:23805973; http://dx.doi.org/10.1042/BJ20130316
21. Nam KH, Haitjema C, Liu X, Ding F, Wang H, DeLisa MP, Ke A. Cas5d protein processes pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvulg CRISPR-Cas system. Structure 2012; 20:1574-84; PMID:22841292; http://dx.doi.org/10.1016/j.str.2012.06.016
22. Shao Y, Cocozaki AI, Ramia NF, Terns RM, Terns MP, Li H. Structure of the Cmr2-Cmr3 subcomplex of the Cmr RNA silencing complex. Structure 2013; 21:376-84; PMID:23395183; http://dx.doi.org/10.1016/j.str.2013.01.002
23. Carte J, Pfister NT, Compton MM, Terns RM, Terns MP. Binding and cleavage of CRISPR RNA by Cas6. RNA 2010; 16:2181-8; PMID:20884784; http://dx.doi.org/10.1261/rna.2230110
24. Garside EL, Schellenberg MJ, Gesner EM, Bonanno JB, Sauder JM, Burley SK, Almo SC, Mehta G, MacMillan AM. Cas5d processes pre-crRNA and is a member of a larger family of CRISPR RNA endonucleases. RNA 2012; 18:2020-8; PMID:23006625; http://dx.doi.org/10.1261/rna.033100.112
25. Haurwitz RE, Sternberg SH, Doudna JA. Csy4 relies on an unusual catalytic dyad to position and cleave CRISPR RNA. EMBO J 2012; 31:2824-32; PMID:22522703; http://dx.doi.org/10.1038/emboj.2012.107
26. Wang R, Preamplume G, Terns MP, Terns RM, Li H. Interaction of the Cas6 riboendonuclease with CRISPR RNAs: recognition and cleavage. Structure 2011; 19:257-64; PMID:21300293; http://dx.doi.org/10.1016/j.str.2010.11.014
27. Wiedenheft B, Lander GC, Zhou K, Jore MM, Brouns SJ, van der Oost J, Doudna JA, Nogales E. Structures of the RNA-guided surveillance complex from a bacterial immune system. Nature 2011; 477:486-9; PMID:21938068; http://dx.doi.org/10.1038/nature10402
28. Wiedenheft B, van Duijn E, Bultema JB, Waghmare SP, Zhou K, Barendregt A, Westphal W, Heck AJ, Boekema EJ, Dickman MJ, et al. RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. Proc Natl Acad Sci U S A 2011; 108:10092-7; PMID:21536913; http://dx.doi.org/10.1073/pnas.1102716108
29. Rouillon C, Zhou M, Zhang J, Politis A, Beilsten-Edmands V, Cannone G, Graham S, Robinson CV, Spagnolo L, White MF. Structure of the CRISPR interference complex CSM reveals key similarities with cascade. Mol Cell 2013; 52:124-34; PMID:24119402; http://dx.doi.org/10.1016/j.molcel.2013.08.020
30. Spilman M, Cocozaki A, Hale C, Shao Y, Ramia N, Terns R, Terns M, Li H, Stagg S. Structure of an RNA silencing complex of the CRISPR-Cas immune system. Mol Cell 2013; 52:146-52; PMID:24119404; http://dx.doi.org/10.1016/j.molcel.2013.09.008
31. Staals RH, Agari Y, Maki-Yonekura S, Zhu Y, Taylor DW, van Duijn E, Barendregt A, Vlot M, Koehorst JJ, Sakamoto K, et al. Structure and activity of the RNA-targeting Type III-B CRISPR-Cas complex of Thermus thermophilus. Mol Cell 2013; 52:135-45; PMID:24119403; http://dx.doi.org/10.1016/j.molcel.2013.09.013
32. van Duijn E, Barbu IM, Barendregt A, Jore MM, Wiedenheft B, Lundgren M, Westra ER, Brouns SJ, Doudna JA, van der Oost J, et al. 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 2012; 11:1430-41; PMID:22918228; http://dx.doi.org/10.1074/mcp.M112.020263
33. Staals RH, Brouns SJ. Distribution and Mechanism of the type I CRISPR-Cas systems, (Barrangou, R and van de Oost, J. eds.), pp115-144, Berlin-Heidelberg: Springer, 2013.
34. Cai F, Axen SD, Kerfeld CA. Evidence for the widespread distribution of CRISPR-Cas system in the Phylum Cyanobacteria. RNA Biol 2013; 10:687-93; PMID:23628889; http://dx.doi.org/10.4161/rna.24571
35. Sharif H, Ozgur S, Sharma K, Basquin C, Urlaub H, Conti E. Structural analysis of the yeast Dhh1-Pat1 complex reveals how Dhh1 engages Pat1, Edc3 and RNA in mutually exclusive interactions. Nucleic Acids Res 2013; 41:8377-90; PMID:23851565; http://dx.doi.org/10.1093/nar/gkt600
36. Zhang J, Rouillon C, Kerou M, Reeks J, Brugger K, Graham S, Reimann J, Cannone G, Liu H, Albers SV, et al. Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. Mol Cell 2012; 45:303-13; PMID:22227115; http://dx.doi.org/10.1016/j.molcel.2011.12.013
37. Bergfors T. Protein Crystallization: Second Edition. International University Line, 2009.
38. Kabsch W. XDS. Acta Crystallogr D Biol Crystallogr 2010; 66:125-32; PMID:20124692; http://dx.doi.org/10.1107/S0907444909047337
39. Evans PR, Murshudov GN. How good are my data and what is the resolution? Acta Crystallogr D Biol Crystallogr 2013; 69:1204-14; PMID:23793146; http://dx.doi.org/10.1107/S0907444913000061
40. Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung LW, Kapral GJ, Grosse-Kunstleve RW, et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr 2010; 66:213-21; PMID:20124702; http://dx.doi.org/10.1107/S0907444909052925

41. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr 2004; 60:2126-32; PMID:15572765; http://dx.doi.org/10.1107/S0907444904019158
42. Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall WB 3rd, Snoeyink J, Richardson JS, et al. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res 2007; 35: W375-83; PMID:17452350; http://dx.doi.org/10.1093/nar/gkm216
43. Lange SJ, Alkhnbashi OS, Rose D, Will S, Backofen R. CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems. Nucleic Acids Res 2013; 41:8034-44; PMID:23863837; http://dx.doi.org/10.1093/nar/gkt606
44. Royer CA. Improvements in the numerical analysis of thermodynamic data from biomolecular complexes. Anal Biochem 1993; 210:91-7; PMID:8489028; http://dx.doi.org/10.1006/abio.1993.1155
45. Schmidt C, Kramer K, Urlaub H. Investigation of protein-RNA interactions by mass spectrometry–Techniques and applications. J Proteomics 2012; 75:3478-94; PMID:22575267; http://dx.doi.org/10.1016/j.jprot.2012.04.030

2.3 Publication 3: Functional characterization of the type I-A Cas7 protein in *Thermoproteus tenax*

Plagens, A., Tripp, V., Daume, M., Sharma, K., Klingl, A., **Hrle, A.**, Conti, E., Urlaub, H., Randau, L. *In vitro* assembly and activity of an archaeal CRISPR-Cas type I-A Cascade interference complex. Nucleic Acids Res. (2014)

The study characterizes the archaeal type I-A interference complex of *Thermopreoteus tenax* (*Tt*). The complex was reconstituted from the individual subunits and used to identify interactions amongst the subunits as well as sequence motifs required for efficient DNA degradation. Mass spectrometry shows that the Cas3HD nuclease and Cas3 helicase proteins are an integral part of the assembly. Size exclusion chromatography (Fig. 8), RNA binding assays (Plagens *et al.*, Figure 2B) and TEM imaging (Fig. 8) demonstrate that *Tt* Cas7 (Csa2) binds unspecifically to crRNA and other small RNA contaminants, resulting in the oligomerization of Cas7. The oligomerization of Cas7 outside of the crRNP was also oberserved for the *S. solfataricus* homolog Csa2 and the Cas7-like protein Cmr4 [63, 69]. Interestingly, the proteins *Mk* Csm3 and *Tp* Csc2 did not exhibit this behavior at any stage.

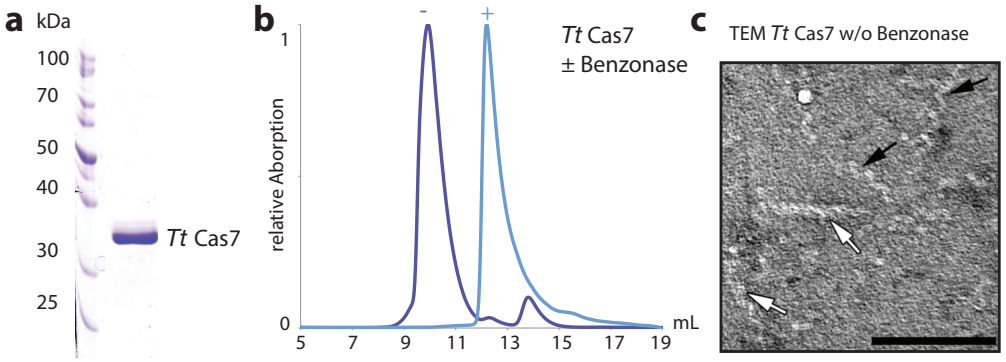


Figure 8: *Tt* Cas7. a. Coomassie stained gel and **b.** size exclusion profiles of *Tt* Cas7 with (+) and without (-) benzonase treatment shows monomeric (light blue) vs. oligomeric states (dark blue). **c.** Transmission electron micrograph of *Tt* Cas7 unspecifically bound to *E. coli* RNA (peak in panel b). Scale bar: 50 nm. See Plagens *et al.*, Fig. S6.

Published online 5 February 2014 *Nucleic Acids Research, 2014, Vol. 42, No. 8* ***5125–5138***
doi:10.1093/nar/gku120

In vitro assembly and activity of an archaeal CRISPR-Cas type I-A Cascade interference complex

André Plagens[1], Vanessa Tripp[1], Michael Daume[1], Kundan Sharma[2], Andreas Klingl[3], Ajla Hrle[4], Elena Conti[4], Henning Urlaub[2] and Lennart Randau[1,*]

[1]Prokaryotic Small RNA Biology Group, Max Planck Institute for Terrestrial Microbiology, D-35043 Marburg, Germany, [2]Bioanalytical Mass Spectrometry Group, Max Planck Institute for Biophysical Chemistry, D-37077 Göttingen, Germany, [3]Cell Biology and LOEWE Research Centre for Synthetic Microbiology, Philipps-Universität Marburg, D-35043 Marburg, Germany and [4]Department of Structural Cell Biology, Max Planck Institute of Biochemistry, D-82152 Martinsried, Germany

Received December 17, 2013; Revised and Accepted January 17, 2014

ABSTRACT

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-CRISPR-associated (Cas) systems of type I use a Cas ribonucleoprotein complex for antiviral defense (Cascade) to mediate the targeting and degradation of foreign DNA. To address molecular features of the archaeal type I-A Cascade interference mechanism, we established the *in vitro* assembly of the *Thermoproteus tenax* Cascade from six recombinant Cas proteins, synthetic CRISPR RNAs (crRNAs) and target DNA fragments. RNA-Seq analyses revealed the processing pattern of crRNAs from seven *T. tenax* CRISPR arrays. Synthetic crRNA transcripts were matured by hammerhead ribozyme cleavage. The assembly of type I-A Cascade indicates that Cas3′ and Cas3″ are an integral part of the complex, and the interference activity was shown to be dependent on the crRNA and the matching target DNA. The reconstituted Cascade was used to identify sequence motifs that are required for efficient DNA degradation and to investigate the role of the subunits Cas7 and Cas3″ in the interplay with other Cascade subunits.

INTRODUCTION

The coevolution of viruses with their prokaryotic hosts led to the development of specific and highly divergent antiviral prokaryotic immune systems. One complex group of adaptive immune systems that is widespread in bacterial and archaeal genomes is termed Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-CRISPR-associated (Cas). Cells that harbor these systems can be immunized against the attack of viruses by the integration of a virus-derived genome fragment into the host genome (1). The genetic memory of previous infections is mediated by CRISPR loci, which consist of a series of short repeat sequences (typically 24–37 bp) that are separated by spacer sequences (2–4). Cas proteins are often encoded in proximity to the CRISPR loci and are key players during all phases of immunization and protection of the cell (5,6). In the first phase, the adaptation, the injected viral DNA is recognized and a fragment is inserted into the host CRISPR array (7–9). This activity is often dependent on a short conserved sequence (2–5 bp) defined as the protospacer adjacent motif (PAM) that flanks the original spacer sequence (termed protospacer) in the viral genome (10,11). The genetic imprint is activated by the transcription of the CRISPR into a long precursor-crRNA (pre-crRNA), which is typically processed by the endoribonuclease Cas6 into short crRNAs that are characterized by an 8-nt 5′-hydroxyl repeat tag, a complete spacer sequence and a 2′–3′ cyclic phosphate repeat end (12–18). During a repeated viral attack, the mature crRNAs can be incorporated into a large Cas ribonucleoprotein interference complex to target the viral DNA for degradation (19–21).

These basic principles of CRISPR-Cas immunity are conserved, but careful computational and biochemical analyses of the differences among the executing interference machines, the composition of conserved Cas marker proteins and the nature of the targeted nucleic acids led to the identification of three distinct major types and several subtypes of CRISPR-Cas systems (5,22). The type I CRISPR-Cas systems can be further divided into six different subtypes (subtypes I-A to I-F), and the respective interference complex is termed Cascade (19). In type III systems, interference is executed by the Csm (subtype III-A, targeting DNA) or Cmr complex (subtype III-B, targeting RNA) (23–25). In contrast, bacterial type II systems are

*To whom correspondence should be addressed. Tel: +49 6421 178 600; Fax: +49 6421 178 599; Email: lennart.randau@mpi-marburg.mpg.de

5126 *Nucleic Acids Research, 2014, Vol. 42, No. 8*

characterized by the single large multifunctional protein Cas9, which is involved in both the maturation of crRNAs and the interference of DNA (26–28).

First details of the Cascade structure and the molecular mechanism were obtained for type I-E systems of *Escherichia coli*. The I-E Cascade shows a seahorse-shaped architecture with a size of 405 kDa and is composed of the conserved subunits Cas6e, Cas7, Cas5e and the subtype-specific nucleic acid-binding proteins Cse1 and Cse2 (19,29,30). The helical backbone of I-E Cascade is formed of six Cas7 copies that are tightly bound to a mature crRNA (31,32). The I-E Cascade facilitates the base pairing of the bound crRNA with the complementary DNA by screening for the short PAM sequence, resulting in strand invasion of the RNA and additional displacement of the non-complementary strand to form the R-loop structure, which recruits Cas3 to degrade the targeted viral DNA (18,33–36). The effector protein Cas3 contains a DExH-like helicase domain (Cas3′) and a HD phosphohydrolase domain (Cas3″), which are responsible for the unwinding of double-stranded DNA (dsDNA) and cleavage of single-stranded DNA (ssDNA) in dependence of adenosine triphosphate (ATP) and divalent metal ions, respectively (37–39).

The type I CRISPR-Cas subtypes differ in Cas protein content, which implies divergent Cascade assembly strategies and functional differences. Comparative studies of these different Cascade complexes will help to gain insight into the evolution and propagation of CRISPR-Cas systems, the integration of one or multiple immunity systems into the cellular protein network and the adaptation mechanisms to diverse prokaryotic environments (5,40). Analysis of the CRISPR-Cas system of the crenarchaeon *Thermoproteus tenax* identified a type I-A Cascade module (*csa5*, *cas7*, *cas5a*, *cas3′*, *cas3″*, *cas8a2*), essential genes for the adaptation of foreign DNA (*cas1*/*cas2*, *cas4*), a type III-A gene cluster, a second subset of a Cascade module and seven CRISPR loci spread throughout the genome (41,42). In *Sulfolobus solfataricus*, the type I-A Cascade sub-complex of Cas7 and Cas5a was identified and shown to bind crRNA and complementary ssDNA. The recombinant Cas7 proteins assembled into multimeric right-handed helical structures (43).

Here, we show the assembly of a complete type I-A Cascade from individual *in vitro*-produced Cas proteins, ribozyme-processed synthetic crRNAs and short protospacer DNA fragments. The strategy for the protein body assembly of the mature Cascade follows the co-refolding of insoluble recombinant I-A Cas proteins from bacterial inclusion bodies to recover the six protein complex. Synthetic crRNAs were created by *in vitro* transcription of crRNA constructs fused to *cis*-acting hammerhead ribozymes. The assembly of the Cascade ribonucleoprotein complex yielded active molecules that showed crRNA-specific DNA targeting and degradation. This *in vitro* assembly strategy allowed us to obtain insights into the Cascade assembly and DNA cleavage mechanism and to identify the PAM requirements for target degradation.

MATERIALS AND METHODS

Strains and growth conditions

Cells of *T. tenax* Kra1 (DSM 2078) grown heterotrophically in *Thermoproteus* medium (44) were a gift from R. Hensel (Essen). *E. coli* strains TOP10 (Invitrogen) and Rosetta2(DE3)pLysS (Stratagene) were cultured in LB medium at 37°C shaking at 200 rpm. For protein production, 1 mM isopropyl-β-D-1-thiogalactopyranoside (IPTG) was added to a growing culture (OD_{600}: 0.6) and incubated for 4 h.

Isolation of small RNAs, production of crRNAs and DNA substrates

For the preparation of *T. tenax* small RNAs (<200 nt), 0.1 g pelleted cells were lysed by homogenization and subsequently isolated according to the *mir*Vana™ miRNA Isolation Kit (Ambion). To generate synthetic crRNAs, forward and reverse complementary DNA oligonucleotides (cr5.2h or cr5.13h, respectively) were synthesized that contained a selected spacer sequence (spacer 5.2 or spacer 5.13) and were fused with the sequence of a minimal *cis*-acting hammerhead ribozyme at the 5′-end (Supplementary Table SI). The oligonucleotides were phosphorylated, hybridized and cloned under control of the T7 RNA polymerase promoter sequence (cr5.2h: BamHI/HindIII, cr5.13h: HindIII/EcoRI) into pUC19. The crRNA was prepared by run-off transcription in a reaction containing 40 mM HEPES/KOH, pH 8.0, 22 mM $MgCl_2$, 5 mM dithiothreitol, 1 mM spermidine, 4 mM of ATP, CTP, GTP and UTP, 20 U RNase inhibitor, 1 µg T7 RNA polymerase and template DNA [PCR products with sequence-specific primers (cr5.2PCRf/r or cr5.13PCRf/r, respectively)] at 37°C for 4 h. For the self-cleaving reaction of the hammerhead ribozyme, the transcription reaction was directly diluted with 4 volumes of 30 mM $MgCl_2$ in DEPC-H_2O and incubated for 1 h at 60°C. The cleaved crRNA was purified by phenol/chloroform extraction (pH 5.2), EtOH precipitated with the addition of glycogen (1:100, v/v), mixed with 2× formamide loading buffer (95% formamide, 5 mM EDTA, pH 8.0, 2.5 mg bromophenol blue, 2.5 mg xylene cyanol), heated for 5 min at 95°C, separated by a denaturing-PAGE (8 M urea, 1× TBE, 10% polyacrylamide) next to an RNA marker (low range ssRNA ladder, NEB) and visualized by toluidine blue staining. The gel bands were cut out and eluted overnight on ice in 500 µl elution buffer (20 mM Tris–HCl, pH 7.5, 250 mM sodium acetate, 1 mM EDTA, pH 8.0, 0.25% SDS) and EtOH precipitated. All DNA oligonucleotides used for cloning and as cleavage substrates were custom-synthesized (Eurofins MWG Operon, Supplementary Table SI).

RNA sequencing

The isolated small RNA was treated with T4 Polynucleotide Kinase (PNK) to ensure proper termini for adapter ligation (45). First, for the dephosphorylation of 2′–3′-cyclic phosphate termini, 10 µg of RNA was incubated at 37°C for 6 h with 20 U T4 PNK (NEB) in

1× T4 PNK buffer. Subsequently, 1 mM ATP was added, and the reaction mixture was incubated for 1 h at 37°C to generate monophosphorylated 5′-termini (46). RNA libraries were prepared with an Illumina TruSeq RNA Sample Prep Kit (Ambion), and sequencing on an Illumina HiSeq2000 sequencer was performed at the Max-Planck Genome Centre, Cologne, Germany. Reads were mapped to the *T. tenax* reference genome (FN869859) with CLC Genomics Workbench 6.0.

Purification of Cascade proteins

The gene constructs of *csa5*, *cas7*, *cas5a*, *cas3′*, *cas3″* and *cas8a2* in pET24a(+) (Novagen) were used as previously described (41). Cas3″ mutants were created using the QuikChange site-directed mutagenesis protocol (Stratagene) according to the manufacturer's instructions. Established mutations were confirmed by sequencing (MWG Eurofins). Soluble Csa5 could be purified, as cells were homogenized in buffer 1 (100 mM HEPES/KOH, pH 7, 10% glycerol, 10 mM ß-mercaptoethanol (ß-Me), 10 mM $CaCl_2$, 300 mM NaCl), lysed, cleared by centrifugation (45 000 × *g*, 1 h, 4°C) and heat precipitated (30 min, 90°C). The cleared supernatant (14 000 × *g*, 30 min, 4°C) was dialyzed overnight in salt-free lysis buffer 1 and purified via a Hi Screen Blue FF affinity column using a FPLC Äkta-Purification system (GE Healthcare) eluted with a linear salt gradient (0–2 M NaCl) at 420 mM NaCl. Fractions containing Csa5 were pooled, dialyzed overnight in salt-free buffer 1 and purified via an anion-exchange chromatography with a MonoQ 5/50 GL column (GE Healthcare) eluted with a linear salt gradient (0–1 M NaCl) at 100 mM NaCl. For pull-down assays of Cascade *in vivo*, Csa5 with a C-terminal 6× His-tag was used as a bait protein. Therefore, *csa5* was cloned into pET20b(+), protein expressed and cells lysed in buffer 1 without $CaCl_2$. The Csa5-His protein was purified from *E. coli* cell lysate by Ni-NTA affinity chromatography (HisTrap HP, GE Healthcare) and eluted with a linear imidazole gradient (0–500 mM) at 150 mM imidazole. Cas7 was purified by cell lysis in buffer 1, heat precipitation at 80°C and cation exchange chromatography with Heparin Sepharose (GE Healthcare) as described earlier (41). Additionally, the Cas7 full-length protein was expressed as a recombinant His-SUMO-tagged fusion protein using BL21-Gold(DE3)Star pRARE (Stratagene) overnight at 18°C. The cells were lysed by sonication in buffer 2 (100 mM potassium phosphate, pH 7.5, 500 mM NaCl, 10% glycerol, 1 mM ß-Me) supplemented with 20 mM imidazole, benzonase (NEB) and protease inhibitors (Roche), and Cas7-SUMO was further purified using Ni-NTA affinity chromatography. The salt concentration was lowered to 100 mM NaCl in 50 mM Tris, pH 7.5, and the His-SUMO-tag was cleaved off by adding SUMO protease overnight during dialysis. The protein was further purified over a HiTrap Heparin Sepharose HP column (GE Healthcare) to remove non-specifically bound nucleic acids. Size-exclusion chromatography on a Superdex 75 column (GE Healthcare) was preformed as a final step of purification in buffer 3 (100 mM NaCl, 50 mM Tris, pH 7.5, 10% glycerol, 5 mM dithiothreitol). The insoluble proteins Cas5a, Cas3′, Cas3″ and Cas8a2 were purified from inclusion bodies and solubilized in 4 M guanidine hydrochloride (GdmCl). The purity of all proteins was determined by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS–PAGE) and Coomassie blue staining alongside the protein marker (ColorPlus™ prestained protein ladder, broad range, NEB). The protein concentration was determined by the Bradford protein quantification method (BioRad).

Reconstitution assays

The reconstitution of the Cascade complex was previously described (41). Briefly, equal amounts (300 µg) of each solubilized protein Cas5a, Cas3′, Cas3″ and Cas8a2 were mixed with the purified proteins Csa5 and Cas7 in reconstitution buffer (3.5 M GdmCl, 100 mM HEPES/KOH, pH 7, 5% glycerol, 10 mM ß-Me, 10 mM $MgCl_2$, 300 mM NaCl) and the denaturing agent was removed via stepwise dialysis into GdmCl-free buffer at room temperature. Aggregated proteins were precipitated (14 000 × *g*, 30 min, 4°C), soluble proteins concentrated with centrifugal filter units (MWCO: 10 kDa) and analyzed by SDS–PAGE. The reconstitution of all 63 subunit combinations was tested via rapid dilution into 1 ml GdmCl-free reconstitution buffer of equal amounts (25 µg) of each subunit. To minimize aggregation, the protein solution was added drop-wise (5 µg total protein) to the reconstitution buffer. Aggregated proteins were precipitated (14 000 × *g*, 30 min, 4°C), soluble proteins were trichloroacetic acid (TCA)-precipitated and the identical amounts of supernatant and pellet analyzed by SDS–PAGE.

Pull-down assays and mass spectrometry analyses of *in vivo* Cascade proteins

To determine the protein–protein interaction *in vivo*, pull-down assays were performed with C-terminal His-tagged Csa5 used as a bait protein bound to a cobalt-chelate matrix and *T. tenax* cell extract as a prey. For the production of cell-free *T. tenax* extract, 1 g cells were resuspended in lysis buffer 3 (100 mM HEPES/KOH, pH 7.5, 300 mM ß-Me) and lysed as described elsewhere (47). To stabilize the interaction of proteins, a chemical cross-linking with formaldehyde was performed. One hundred micrograms purified Csa5, 1 mg cell extract and 1% formaldehyde (37% formaldehyde/10% MeOH) were mixed, incubated for 15 min on ice and the reaction stopped by the addition of 200 mM glycine for 5 min on ice. Subsequently, the pull-down assay was performed according to the ProFound Pull-Down PolyHis Protein:Protein Interaction Kit (Pierce). The cross-linked protein:cell extract was incubated for 30 min at 4°C on the cobalt resin, washed five times with 40 mM imidazole in buffer 1 without $CaCl_2$ and eluted at 290 mM imidazole. Proteins were TCA-precipitated, separated by SDS–PAGE and visualized with silver staining (Pierce Silver Stain Kit). For the mass spectrometry analysis, the protein:cell extract was pelleted by EtOH precipitation and digested

in solution in the presence of urea as described previously (48). The resultant peptides were desalted using STAGE tips (49). Further, the desalted peptides were analyzed by liquid chromatography tandem mass spectrometry on an LTQ Orbitrap Velos instrument (Thermo Fischer Scientific) under standard conditions. The protein identification was performed with MaxQuant (version 1.2.2.5) using the Andromeda search engine (50,51).

Protein–protein and protein–RNA interaction assays

The protein–protein interaction and the native molecular mass of Cascade or the Cas7 subunit was determined by size-exclusion chromatography with a Superdex 200 10/300 preparatory-grade column (GE Healthcare, 24 ml). Protein containing fractions were TCA-precipitated [1:4 of 100% (v/v)] and analyzed by SDS–PAGE. To verify the binding of Cascade to crRNA, 50 µg of synthetic crRNA was added to 500 µg of reconstituted Cascade complex incubated at 65°C for 30 min to facilitate RNA binding and separated by gel filtration. Fractions were split and protein extracted via TCA precipitation and analyzed during separation on SDS–PAGE, whereas RNA was isolated by Proteinase K treatment (10 µl, 6 U Proteinase K, 30 min at 37°C), phenol/chloroform extraction, EtOH precipitation and detection via denaturing-PAGE (10% polyacrylamide) and toluidine blue staining.

Electrophorectic mobility shift assay

The Cascade complex or the Cas7 subunit were tested for the ability of binding synthetic crRNA in electrophorectic mobility shift assays (EMSAs). As the binding of Cascade was metal-independent, the complex was reconstituted in the presence of 10 mM $CaCl_2$ instead of $MgCl_2$ to reduce RNA cleavage. A total of 5 pmol of the synthetic crRNA substrate was 5′-labeled with [γ-^{32}P]-ATP (5000 ci/mmol, Hartmann Analytic) and T4 PNK (Ambion) for 2 h at 37°C; the reaction stopped by the addition of formamide loading buffer and the separation by denaturing-PAGE (10% polyacrylamide). After autoradiographic exposure, the RNA band were cut out, gel eluted and EtOH precipitated. Up to 5 µM of reconstituted Cascade was incubated with 2–4 nM of 5′-labeled crRNA in binding buffer (100 mM HEPES/KOH, pH 7, 100 mM NaCl, 10 mM ß-Me, 50 ng yeast RNA) for 30 min at 70°C. The reaction was immediately stopped on ice, mixed with 1× loading buffer (Qiagen) and Cascade:crRNA complexes separated from free crRNA by non-denaturing TBE-PAGE (6% polyacrylamide, 1× TBE). The ability of monomeric Cas7 to bind to synthetic crRNA was tested by incubating protein with 2–4 nM of 5′-labeled crRNA in binding buffer without yeast RNA for 30 min at 80°C, ran on 2% TAE-agarose gels and visualized by phosphorimaging. Binding of the two Cas7 versions (Cas7, Cas7-SUMO) to contaminating nucleic acids during the purification process was verified by loading 100 µg protein on 1% TAE-agarose gels stained with ethidium bromide alongside the DNA marker (2-log DNA ladder, NEB). The specification of the contaminated nucleic acid was conducted by adding 2 U DNase I (RNase-free, NEB), 1 ng RNase A (biochemistry grade, Ambion) or 25 U benzonase (NEB) in the appropriate reaction buffer to the protein solution and incubated for 30 min at 37°C before loading on agarose gels.

Transmission electron microscopy (TEM)

To further analyze samples of Cas7 purifications, 5 µl of a concentrated protein solution was applied to carbon-coated 400 mesh copper grids. After blotting on a filter paper, the samples were washed twice on a drop of double distilled water and blotted each time. Finally, the proteins were negatively stained with 2% (w/v) uranyl acetate for 20 s, blotted again on filter paper and left for air drying (52,53). Electron microscopy and further analysis was carried out on a JEOL JEM-2100 transmission electron microscope (JEOL, Tokyo, Japan) operated at 120 kV. The microscope was equipped with a 2k × 2k CCD camera F214 in combination with the EM-Menu 4 software (TVIPS, Gauting, Germany).

Nuclease assays

The reconstituted Cascade was tested for the nucleolytical activity of the Cas3′′ subunit. A total of 5 pmol of the ssDNA oligonucleotide (int 5.2_CCT for) was 5′-labeled with [γ-^{32}P]-ATP (5000 ci/mmol, Hartmann Analytic) and purified as described earlier in the text. In all, 100 nM refolded Cascade was incubated with 2 nM of 5′-labelled ssDNA in nuclease buffer (100 mM HEPES/KOH, pH 7.0, 100 mM NaCl, 5 mM $MgCl_2$, 5 mM $MnCl_2$, 10 mM ß-Me, 20 ng yeast total RNA) at 70°C for 10 min, the reaction stopped via EtOH precipitation and resuspended in 10 mM Tris–HCl, pH 8.5, and formamide loading buffer. Each sample was boiled, measured via scintillation counting (Beckman LS6000) and samples with identical counts per minute after normalization were loaded on a 15% denaturing polyacrylamide gel alongside the low molecular weight marker (10–100 nt, Affymetrix) running in 1× TBE (8 watts, 2 hours). Cleavage products were visualized by phosphorimaging.

To test for the cleavage of dsDNA, short substrates were designed that mimic an invasive viral DNA. Therefore, the spacer sequence of crRNA 5.2 or 5.13 was used, the different PAM added upstream and flanked with random sequences up- and downstream (T7 promoter and T7 terminator, respectively) and each DNA oligonucleotide was synthesized either in forward or reverse complementary direction (for/rev: int 5.2_CCT, int 5.2_CCA, int 5.2_TCA, int 5.2_TCG, int 5.2_AAA, int 5.2_Rep or int 5.13_CCT, respectively). A total of 5 pmol of each forward and reverse oligonucleotide was 5′-labeled with [γ-^{32}P]-ATP (5000 ci/mmol, Hartmann Analytic), purified as mentioned earlier and hybridized with 1.5-fold molar excess of the respective cold complementary strand by heating at 95°C for 5 min and slowly cooling down to room temperature in hybridization buffer (10 mM Tris–HCl, pH 8, 1 mM EDTA, pH 8, 100 mM NaCl). For interference tests, 500 nM refolded Cascade and 500 nM crRNA (crRNA 5.2 or 5.13) were incubated in interference buffer 1 (100 mM HEPES/KOH, pH 7.0, 100 mM NaCl, 10 mM ß-Me, 20 ng yeast total RNA) at 70°C for 20 min to allow loading of crRNA into Cascade

and then the reaction started by adding 2 nM of 5′-labeled hybridized dsDNA and interference buffer 2 (5 mM $MgCl_2$, 5 mM $MnCl_2$, 2 mM ATP) at 70°C for 10 min. The reaction was stopped via EtOH precipitation and loaded on 20% denaturing TBE-polyacrylamide gels or on 10% sequencing gels (Model S2 sequencing gel electrophoresis apparatus, Life Technologies) alongside the low molecular weight marker or (10–100 nt) or seven specific labeled fragments (8–66 nt).

RESULTS

RNA-Seq analyses reveal the processing pattern of *T. tenax* crRNAs

One requirement for CRISPR immunity is the processing of crRNA precursors to yield mature crRNAs. To obtain a comprehensive view of the processing pattern of crRNAs in *T. tenax*, total small RNA was isolated and sequenced via Illumina HiSeq2000 RNA-Seq methodology. T4 PNK treatment of *T. tenax* small RNAs was required to facilitate adaptor ligation during RNA-Seq library construction, which confirms the presence of 5′-OH termini (45). Previously, seven CRISPR loci with a total of 142 spacer sequences were identified, and crRNA transcripts were verified for five clusters (TTX_1, 4, 5–7) via northern blot detection, whereas two clusters (TTX_2–3) showed no distinct processing pattern (41). In total, 13 357 720 individual sequence reads were mapped to the *T. tenax* reference genome (FN869859). Constitutive crRNA production and a gradual decline of mature crRNAs from the leader proximal to the leader distant region within each active cluster were identified (Figure 1). The read coverage indicates the relative abundance and the processed termini of crRNAs from the different CRISPR arrays. Individual crRNAs for the first spacers were represented by up to 1 334 585 reads (spacer 1.1, 10% of total reads), which demonstrates a massive accumulation of crRNAs *in vivo*. All identified crRNAs contained a clearly defined 5′-terminal 8 nt tag (TTX_1: 5′-UUUGAAGG-3′ or TTX_4–7: 5′-AUUGAAAG-3′). The 3′-termini are gradually shortened and mostly contain a minimal 2-nt tag (5′-GA-3′). Additionally, two spacers (spacer 4.34 and spacer 5.32) that were not listed in the CRISPR database could be identified, increasing the total number of 144 spacers in *T. tenax*. In contrast, the CRISPR loci (TTX_2–3) were inactive and showed no mature crRNAs with defined termini.

Cas3′ and Cas3″ are an integral part of the type I-A Cascade complex

Typical for archaeal subtype I-A CRISPR systems is the organization of Cascade genes in an operon structure and their transcription as a polycistronic unit, which includes individual genes for the helicase (Cas3′) and nuclease (Cas3″) domains of Cas3 (41). This organization suggests a different complex formation than Cascade:Cas3 complexes of the bacterial subtype I-E, in which Cas3 is recruited to an assembled Cascade bound to target DNA. To substantiate the protein–protein interactions of the Cascade complex *in vivo*, we performed pull-down assays of *T. tenax* cell extracts with the C-terminal His-tagged Csa5 as bait protein (Supplementary Figure S1A). The interacting proteins were subjected to an in solution trypsin digest followed by mass spectrometry analyses. In these experiments, we could identify Cas7, Cas5a, Cas3′, Cas3″ and Cas8a2 as interaction partners of Csa5, all six proteins whose genes are organized in an operon (Supplementary Table SII). This suggests an alternative mechanism of Cascade formation, in which the two Cas3 domains are not recruited but are an integral part of the Cascade complex in advance of the immune response. Noteworthy, none of the two copies of Cas6 in the genome of *T. tenax* were found in the data. Additionally, we could also detect a second Cas7 homolog (TTX_0235, 40% similarity), produced from a distant *cas* gene cluster without an encoded *csa5* gene, as a potential interaction partner of Csa5.

The *in vitro* assembled Cascade efficiently binds a synthetic mature crRNA

To obtain mature crRNAs with an 8 nt tag (5′-AUUGAA AG-3′) and a 5′-hydroxyl-terminus, we designed a ribozyme maturation strategy that overcomes the need for recombinant Cas6 to be available. A DNA construct was designed that contained a minimal *cis*-acting hammerhead ribozyme fused to the sequence of a mature crRNA under the control of a T7 RNA polymerase promoter (Supplementary Figure S2A). The synthetic crRNA was obtained by *in vitro* transcription and self-cleavage of the RNA transcript at 60°C. This methodology was tested for two different crRNA constructs [crRNA 5.2 (50 nt) and 5.13 (54 nt)], resulting in a cleavage efficiency of ∼60% (Supplementary Figure S2B) (54,55). These synthetic mature crRNAs were tested for their capability in complex binding to establish a fully *in vitro* assembly strategy of functional Cascade modules. In this procedure, the co-refolding of the insoluble subunits Cas5a, Cas3′, Cas3″ and Cas8a2, in combination with the purified proteins Csa5 and Cas7, yielded soluble Cascade (Supplementary Figure S1A–C). First attempts showed an increased refolding efficiency in the combination of all six Cascade subunits (41). Therefore, we investigated the co-refolding of all possible 63 compositions of Cascade subunits to scan for a minimal stable core of Cas proteins. Surprisingly, only the co-refolding of all six proteins resulted in high amounts of soluble protein recovery of ∼50%, whereas in contrast all other 62 combinations resulted in negligible amounts (up to 10%) of soluble protein (Supplementary Figure S3). An up-scaled protocol for the assembly of the six-protein Cascade revealed soluble proteins with concentrations of up to 4 mg/ml. The assembled Cascade was then incubated for 30 min at 65°C to remove impurities or misfolded heat-instable subunits before size-exclusion chromatography for screening of Cascade complex formation. The analysis of the elution profile (12–17 ml elution volume) showed that all six co-refolded proteins eluted in one fraction, followed by smaller sub-complexes and

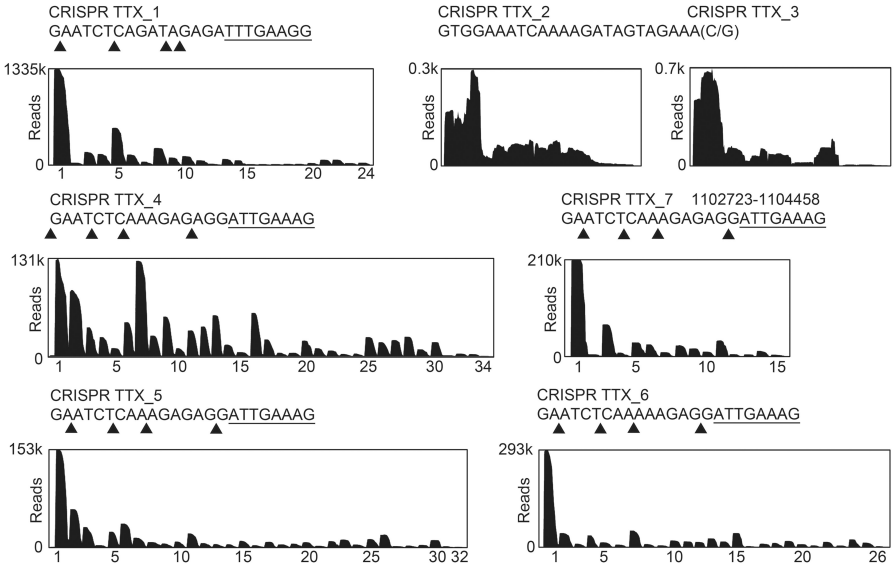


Figure 1. Processing of crRNAs in *T. tenax*. The abundance of crRNAs was verified by mapping the Illumina HiSeq2000 sequencing reads to the *T. tenax* Kra1 reference genome. Sequence coverage (reads) for the seven *T. tenax* CRISPR clusters (crRNAs are indicated by numbers on the *x*-axis) is visualized. The processing sites could be identified within the repeat elements, generating crRNAs with a 5′-terminal 8-nt tag [5′-(A/U)UUGAA(A/G)G-3′, underlined] and variable trimming of the 3′-ends with most of the crRNAs terminating with a 1–2-nt tag (5′-G/GA-3′). The trimming sites for each CRISPR locus are indicated by triangles.

monomeric protein subunits (Figure 2A and Supplementary Figure S4). The functionality of Cascade was first assayed by analysis of binding to the synthetic crRNA. The assembled complex was incubated with crRNA, applied to size-exclusion chromatography and identical fractions were checked for protein and RNA content. The first fractions of the Cascade profile co-eluted with the added crRNA. Visible on the denaturing gel are bands at the expected size of ∼50 nt (Figure 2A, crRNA). Two controls showed that the assembled Cascade was not cross-contaminated with small RNA from the *E. coli* expression host (Figure 2A, −RNA) and unbound crRNA was not eluting in the respective fractions (Figure 2A, −Cascade). Thus, mature crRNAs are supporting the Cascade formation, and the established ribozyme methodology can be used to prepare synthetic crRNA molecules to be loaded onto Cascade.

In type I-E systems, the subunit Cas7 constitutes the helical backbone of Cascade that is formed on binding of the crRNA (31). To verify the RNA-binding activity of Cas7 in I-A systems, we analyzed *T. tenax* Cas7 in more detail. The purification strategy for untagged soluble Cas7 resulted in multimeric proteins that eluted at the void volume from the used gel filtration column indicating protein complex sizes >600 kDa (Supplementary Figure S5A). The treatment of Cas7 with DNase I, RNase A or benzonase identified a cross-contamination of Cas7 with *E. coli* RNAs (<500 nt), which suggests that Cas7 multimerization is obtained by RNA binding independent of its sequence (Supplementary Figure S5B). Transmission electron microscopy (TEM) of this sample revealed long helical filaments of up to 50-nm length, similar to previously reported structural studies of *S. solfataricus* (43). Additionally, we could also observe the occurrence of interlaced filaments of two Cas7 helices illustrating the potential for complete coverage of RNA molecules with Cas7 subunits (Supplementary Figure S6).

To prevent polymeric Cas7 artifacts, an alternative purification strategy of a Cas7-SUMO fusion protein was established that resulted in monomeric and RNA-free Cas7 (Supplementary Figure S5A and C). This protein preparation was subsequently used for the *in vitro* assembly of Cascade (Figure 2A). A comparison of the crRNA binding behavior in EMSAs showed the recurrence of high-shifting multimeric Cas7 structures for the sole Cas7 protein (Figure 2B) and lower shifts and an increased affinity toward crRNA for Cas7 in a Cascade assembly. The addition of total yeast RNA or unlabeled crRNA resulted in similar Cas7 band shifts and underlines an unspecific affinity toward RNA for the individual Cas7 subunit (Supplementary Figure S7). In contrast, specific binding of Cascade to the synthetic crRNA 5.13 was seen in EMSAs, with protein concentrations ranging from 0.125–2 μM (Figure 2C). A second synthetic crRNA (crRNA 5.2) is bound by Cascade with a similar affinity (Supplementary Figure S8).

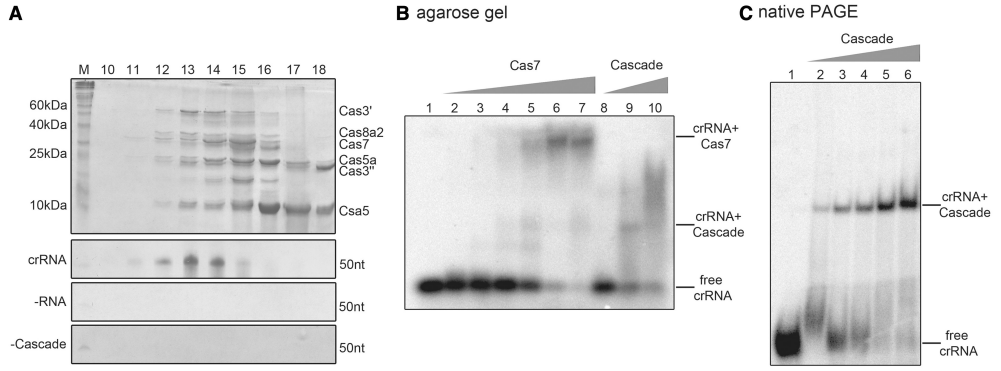


Figure 2. Cascade assembly and RNA binding. (**A**) The Cascade subunits (Csa5, Cas7, Cas5a, Cas3′, Cas3″ and Cas8a2) were assembled via a co-refolding procedure of insoluble recombinant proteins, incubated with synthetic crRNA 5.2 (crRNA) or no RNA (−RNA), and protein interaction was verified via size-exclusion chromatography. In all, 15% SDS–PAGE of individual fractions (fractions 10–18) shows distinct protein bands of assembled Cas protein subunits. The Cascade-containing fractions were additionally analyzed for bound RNA content by 10% Urea-PAGE. A protein-free sample served as a running control (−Cascade). (**B**) A comparison of the EMSAs for crRNA binding by the individual Cas7 subunit (lanes 1–7: 0, 0.2, 0.5, 1, 2, 4, 5 μM) or Cas7 assembled into Cascade (lanes 8–10: 0.5, 2, 5 μM) on 1% agarose gels demonstrates the formation of high shifts for Cas7 alone and lower, more diffuse shifts for Cascade. (**C**) EMSAs verified the binding of the 5′-[γ-^{32}P]-ATP labeled crRNA 5.13 by Cascade (with Cas7 purified via a SUMO-fused construct) in increasing concentrations (lanes 1–6: 0, 0.125, 0.25, 0.5, 1, 2 μM) and in the presence of yeast RNA via 6% native PAGE.

The exonucleolytically ssDNA cleavage activity of Cas3″ is inhibited by the addition of crRNAs

Previous analysis of the nuclease domain of a Cas3 enzyme indicated endo- and exonucleolytic ssDNA cleavage and revealed an HD motif within its active site, which is involved in the coordination of transition metal ions (38). Therefore, nuclease assays with a 93 nt linear ssDNA substrate were performed to biochemically characterize the *T. tenax* Cas3″ subunit within a coordinated Cascade structure. Nuclease assays with an assembled Cascade lacking crRNA showed increasing cleavage activity at Cascade concentrations of 0.05–0.5 μM. At the highest Cascade concentration, up to 75% ssDNA was degraded during 10 min incubation at 70°C. A time course cleavage assay with 0.1 μM Cascade demonstrated degradation of the substrate within 10 min (Figure 3A). The cleavage reaction showed a strict dependence on divalent metal ions, as increased cleavage of Cascade was only observed in the presence of Mn^{2+} ions, followed by Mg^{2+} ions, while Ca^{2+} ions inhibited the reaction. The highest cleavage rate was observed with the combination of Mg^{2+} and Mn^{2+} ions (Supplementary Figure S9). Next, we tested the influence of crRNA addition on the unspecific ssDNA nuclease activity. The used crRNA (crRNA 5.2) and the ssDNA fragment (int 5.2_CCT for) are not complementary to each other to prevent the formation of non-cleavable RNA:DNA hybrids. Cascade was first loaded with crRNA, followed by the addition of the ssDNA substrate to the reaction to mimic the *in vivo* situation of Cascade assembly. The comparison of Cascade ssDNA cleavage rate at 0.1 μM without crRNA and with 0.5 μM crRNA for the identical time points showed an inhibition of the nuclease activity (Figure 3A). The strongest inhibitory effect could be observed at 0.05 μM added crRNA after 10 min incubation, which resulted in an over twofold decrease in activity (65% versus 28.5% remaining ssDNA). The lowered cleavage rate can also be seen for higher concentrations of crRNA (0.5 μM: 55%, 2.5 μM: 45% uncleaved substrate), but is less significant, presumably due to hybrid formation of excess crRNAs (Figure 3B). To verify that Cas3″ is the sole ssDNA nuclease domain within the archaeal Cascade, we aligned the *T. tenax* Cas3″ with the previously characterized *Methanocaldococcus jannaschii* Cas3″ (MJ0384) sequence and introduced mutations of three of the four highly conserved residues of the HD domain motif (H19A, H55A and D56A) that are expected to inactivate exonuclease cleavage (37). The produced Cas3″ mutants were isolated and purified according to the wild-type Cas3″ purification protocol and assembled within Cascade via the co-refolding procedure (Supplementary Figure S10). Each Cas3″ mutant showed a deficiency in ssDNA nuclease activity, most dramatically observed for D56A (12% cleaved ssDNA), followed by H19A (18% cleaved substrate) (Figure 3C).

Cascade-mediated interference is dependent on crRNA and the protospacer DNA

Next, the assembled I-A Cascade was tested for cleavage of dsDNA in dependence of the spacer encoded crRNA to show *in vitro* type I-A Cascade-mediated interference. First, short dsDNA fragments were designed that mimic viral protospacer DNA. The spacer sequence 5.2 (40 bp) was flanked by random sequences on both sides (25 bp each). Directly upstream of the spacer the PAM sequence CCT was integrated. The PAM sequence CCN was identified in *S. solfataricus* on the basis of viral

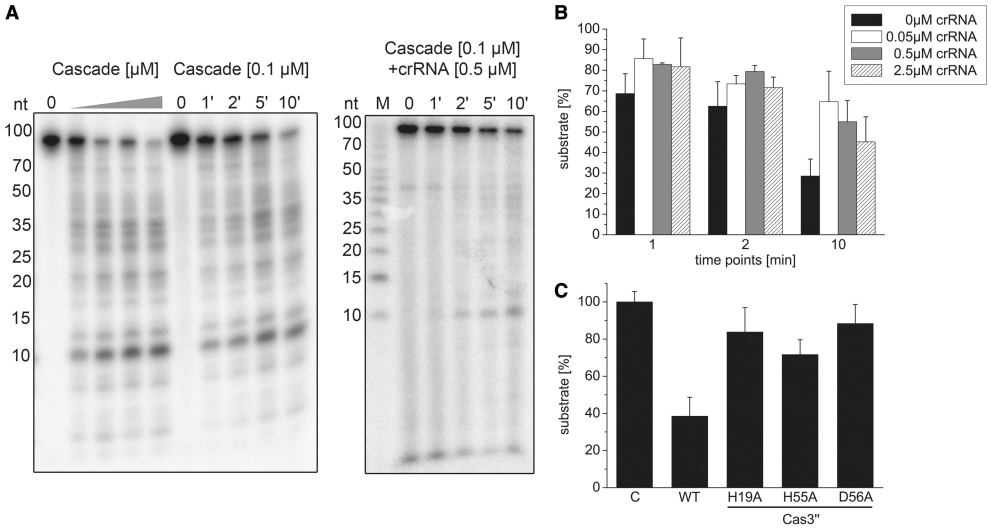


Figure 3. Type I-A Cascade cleaves ssDNA unspecifically and is inhibited by crRNA. (**A**) Cascade (0, 0.05, 0.125, 0.25, 0.5 µM) was incubated with a 5′-[γ-^{32}P]-ATP labeled ssDNA fragment (int_5.2 CCT for) in the presence of Mg^{2+} and Mn^{2+} ions and the nucleolytic cleavage reaction was resolved on 15% denaturing gels. After 1–10-min incubation at a fixed Cascade concentration of 0.1 µM, >70% of the substrate is cleaved. In the presence of 0.5 µM unlabeled crRNA, the reaction is inhibited. (**B**) This observation was tested in the presence of 0, 0.05, 0.5 and 2.5 µM crRNA, and the amount of remaining substrate estimated via line profile plots (Image J) was plotted for three different reaction times (1, 2, 10 min) and reactions performed in triplicate. (**C**) Cas3″ subunit with HD domain mutations (H19A, H55A, D56A) was assembled into Cascade and then tested in ssDNA cleavage assays.

BLAST hits of spacer targets (10). The non-target (int 5.2_CCT for, crRNA non-complementarity) or the target DNA strand (int 5.2_CCT rev, crRNA complementarity) was 5′-labeled and hybridized with 1.5-fold excess of the complementary cold strand to obtain two labeled dsDNA fragments. The synthetic mature crRNA construct 5.2 was supplied by hammerhead self-cleavage and first loaded into Cascade before the addition of the dsDNA substrates, the metal ions Mg^{2+} and Mn^{2+}, as well as ATP. The separation of cleavage products via 20% denaturing-PAGE identified a Cascade-dependent degradation pattern of the dsDNA substrates with differences in fragment length for the non-target and target strand (Figure 4A). As a control, the Cas3″ nuclease mutant D56A was assembled into Cascade, which showed no cleavage activity of the protospacer DNA. Cascade loaded with a synthetic crRNA 5.13 that is not complementary to the DNA target showed also no dsDNA cleavage activity demonstrating crRNA-mediated target guidance. In return, the reaction was performed using the spacer sequence 5.13 with identical flanks in the dsDNA substrate as used before (Figure 4B). Accordingly, only the matching crRNA 5.13 was capable to guide the Cascade-mediated cleavage of the dsDNA substrate, whereas neither the incorporated crRNA 5.2 nor the Cascade mutant Cas3″ D56A showed nuclease activity. The cleavage pattern of the non-target and target strand was similar for dsDNA 5.13 in comparison with dsDNA 5.2. Therefore, the cleavage products of the interference reaction with the labeled non-target and target strand of each dsDNA substrate 5.2 or 5.13 were separated on 10% sequencing gels to specify the fragment lengths (Supplementary Figure S11). Each reaction was loaded next to a marker with ssDNA fragments of 10–100 nt length and a mixture of seven ssDNA fragments of 8–66 nt length to pinpoint the dominant cleavage products. Cleavage sites were observed in the middle of the non-target strand within the protospacer (at position 43–47 nt), and smaller 5′-terminal fragments were generated with a minimal length of 10 nt (Figure 4C). The target strand is cleaved over its entire length, with smallest fragments of 9 nt. The comparison of both DNA substrates showed no significant differences in the fragmentation pattern, indicating a sequence unspecific cleavage of Cascade in dependence of the position within the R-loop structure (Figure 4D).

The Cascade reconstitution platform allows the identification of PAM-dependent target DNA cleavage

The co-refolded Cascade in combination with *in vitro* processed crRNA provides an active interference complex cleaving dsDNA in a RNA sequence-dependent manner. This set up allows the modulation of individual components and to test their influence on Cascade-mediated interference. First, the Cas3″ mutants H19A,

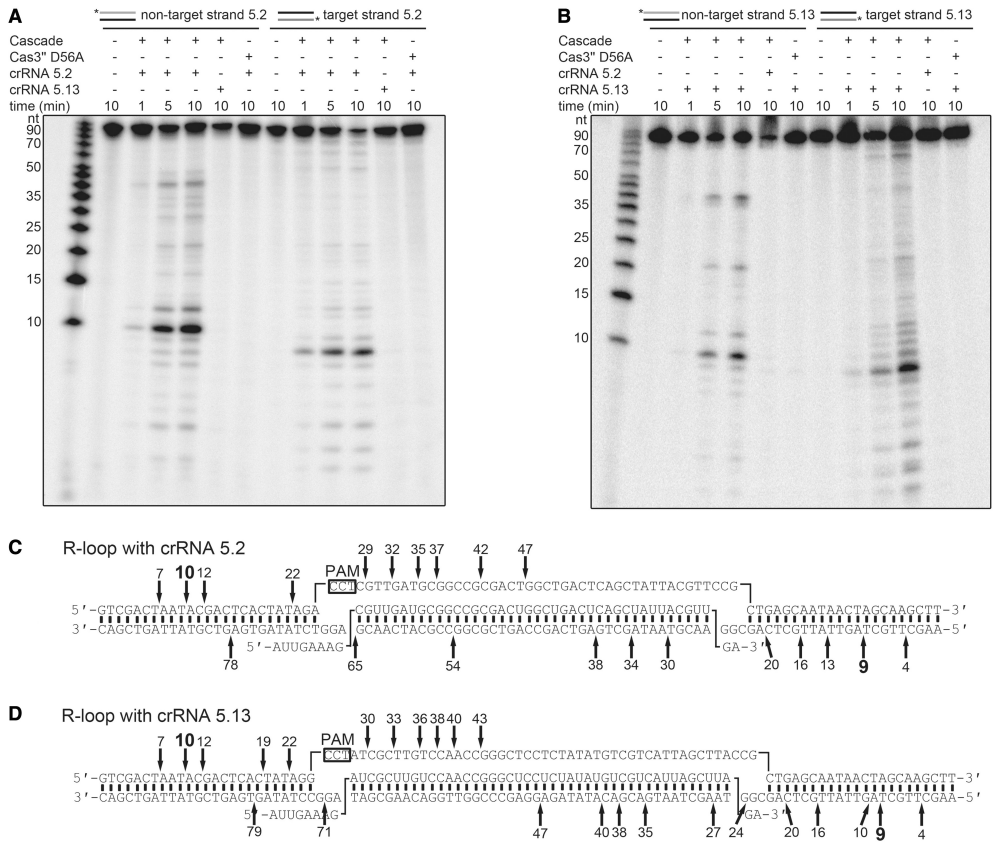


Figure 4. Interference activity of *in vitro* assembled type I-A Cascade. (**A**) The assembled Cascade complex is loaded with crRNA 5.2 for 20 min at 70°C, and the interference reaction is started with the addition of ATP, Mg^{2+}, Mn^{2+} and the dsDNA substrate (in_5.2 CCT), which is either 5′-[γ-^{32}P]-ATP labeled on the non-target (forward) or the crRNA target strand (reverse). Cleavage reactions were stopped at three different time points (1, 5, 10 min at 70°C). The reaction products of the cleaved dsDNA were separated on 20% denaturing gels. The non-matching crRNA 5.13 and the Cas3′′ D56A mutant are used as controls. (**B**) In parallel, the crRNA 5.13 is loaded into Cascade, and cleavage of the matching dsDNA substrate (in_5.13 CCT) is visualized. (**C** and **D**) The cleavage products are analyzed on 10% Urea-PAGE for each strand (in_5.2 CCT for/rev and in_5.13 CCT for/rev) with two different markers (Supplementary Figure S9). The cleavage sites are marked within the proposed R-loop structure that is formed during the interference reaction [(C) dsDNA 5.2, (D) dsDNA 5.13].

H55A and D56A were assembled into Cascade and tested for their ability to cleave dsDNA substrates (Figure 5A). All three mutants were impaired in dsDNA cleavage. The H19A and D56A mutants showed no activity for the non-target or target strand, while H55A is strongly reduced in dsDNA cleavage. Next, the effect of different PAM sequences on the *in vitro* interference reaction was assayed. It should be noted that Cascade recognizes PAM sequences as so-called target interference motifs that might slightly differ from the PAM sequence recognized during adaptation (56). Owing to missing viral targets of *T. tenax* spacers, the exact PAMs are not known. Detailed analysis of *S. solfataricus* CRISPR systems revealed CCN sequences located 5′ to the DNA strand corresponding to the crRNA (non-target strand) as functional PAMs (10,57), and computational studies identified the TCN motif as a potential PAM (56,58). Therefore, the designed dsDNA protospacer substrate 5.2 was modified upstream in the spacer-adjacent three base pair motif, while the spacer and the flanking sequences remained identical. The tested dsDNA substrates contained the PAM sequences CCT, CCA, TCA, TCG, AAA and a motif identical to the 5′-tag of the crRNA (mimicking self-targeting at the CRISPR locus) (Figure 5B). The interference reaction showed similar Cascade cleavage patterns of dsDNA for the PAM CCT and CCA,

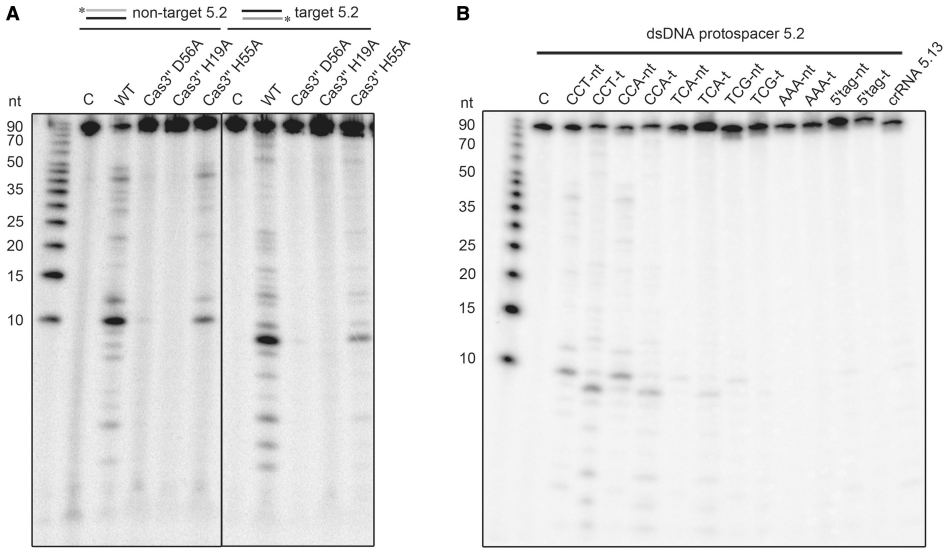


Figure 5. Analysis of Cas3′′ mutants and PAM recognition for *in vitro* assembled type I-A Cascade. (**A**) The Cas3′′-constructed mutants (H19A, H55A, D56A) were assembled into Cascade and tested for dsDNA cleavage. (**B**) The dsDNA substrate 5.2 was mutated to include the indicated 3-bp long PAM sequences (CCT to CCA, TCA, TCG, AAA or a PAM identical to the crRNA 8-nt tag). Cascade-mediated interference reactions were performed with either 5′-[γ-^{32}P]-ATP labeled non-target (forward) or the crRNA target strand (reverse) as a substrate, while Cascade was loaded with the spacer matching crRNA 5.2. The loaded non-matching crRNA 5.13 served as a negative control.

confirming that CCN motif is a PAM that allows DNA target cleavage. A reduced activity was observed for the TCA PAM, while substrates with the TCG, AAA or the 5′-tag motif were not cleaved by Cascade.

DISCUSSION

Type I CRISPR-encoded interference is mediated by a crRNA-guided Cascade complexed with Cas3 to degrade the foreign target DNA. Here, we described the production and assembly of an archaeal type I-A Cascade with co-refolded recombinant Cas proteins and synthetic crRNA transcripts, which resulted in an active complex capable of crRNA-guided degradation of dsDNA in dependence of PAM sequences.

The analysis of *T. tenax* RNA-Seq read mapping revealed the mature crRNA termini and confirmed previous northern blot analyses, which indicated that *T. tenax* harbors five highly active CRISPR loci and two inactive CRISPR loci (41). These inactive and active clusters contain conserved motifs for transcription initiation and differ mainly in their respective repeat sequences (seven base exchanges and one base length difference). Therefore, it is plausible that this repeat is not processed by the two encoded Cas6 proteins. Accordingly, in type III-A systems, crRNAs or their precursors cannot be detected in Cas6 deletion mutants suggesting degradation of the primary transcript (59). Alternatively, the transcription of the two inactive loci could be regulated by an unknown mechanism. To engineer functional synthetic crRNAs, the *in vivo* processing of mature crRNAs by Cas6 had to be mimicked, as soluble recombinant Cas6 was not available. The technique of fusing a *cis*-acting hammerhead ribozymes to RNA transcripts to obtain defined termini (54,55) was adapted for the production of crRNAs, which resulted in 8 nt tags with 5′-hydroxyl ends (Supplementary Figure S2A and B). This methodology allows the *in vitro* production of crRNAs that start without the guanosine base that is required for proper T7 RNA polymerase transcription initiation. The RNA-Seq data additionally revealed gradually degraded 3′-termini for all crRNAs. The used synthetic crRNAs contained a 2 nt tag (5′-GA-3′), which was observed for the majority of crRNA reads. The trimming of the crRNAs' 3′-end is typically observed in many different bacteria and archaea, but the exact mechanism is not known (60,61). It is likely that Cas6-processed crRNAs are protected by Cascade proteins binding to the conserved 5′-end, whereas the free 3′-ends are unspecifically trimmed by cellular RNases due to a missing secondary structure or that crRNAs are unstable at elevated temperatures (62,63). For organisms with multiple encoded CRISPR-Cas effector complexes, distinct crRNA species with varying 3′-ends might be sorted between complexes in dependence of their length (64). These potential parameters for crRNA loading into

recombinant Cascade could be tested by engineering multiple modified synthetic RNAs. Other possible applications for the use of synthetic crRNA transcripts are dye-labeled RNA constructs for smFRET measurements or crRNA-Cascade cross-linking with photoreactive nucleotide analogs that facilitates detailed information about the RNA structure and motion during complex formation and a precise mapping of the cross-link between crRNA and Cas protein subunit (65–67).

Previous research focused on the detailed analysis of the *E. coli* type I-E Cascade function and structure (19,31). One main feature of the type I-E interference mechanism is a separation into the Cascade module that is binding the crRNAs recognizing the matching DNA target and the Cas3 protein, which is recruited by Cascade to the crRNA:dsDNA R-loop structure. The *E. coli* Cascade is built up by the subtype specific subunits Cse1 and Cse2 that are proposed to interact with the DNA target, the conserved crRNA binding subunits Cas7 and Cas5e and the pre-crRNA cleaving endoribonuclease Cas6e (18,29,30). From a mechanistic point of view, this arrangement guarantees a streamlined processing of the pre-crRNA transcripts, immediate crRNA protection and the scanning for the complementary DNA target, localized within one protein complex. Additionally, the transcription of pre-crRNAs and *cascade* genes are strictly repressed by the global regulator H-NS and activated with the help of the transcription factor LeuO, ensuring a fast response to a viral attack (68–70).

In this study, we analyzed an archaeal type I-A CRISPR-Cas system, which showed remarkable differences in the assembly of Cascade and the crRNA production. A typical feature found in archaeal genomes is the splitting of the two Cas3 domains into separate genes (Cas3′: helicase domain, Cas3″: nuclease domain) and a gene organization into one *cascade* genes operon that can be regulated by changes of the environmental parameters (41). Similar to type I-E Cascade, the crRNA binding subunits *cas7* and *cas5a* and two subtype-specific subunits *csa5* and *cas8a2* are encoded within the operon. However, two *cas6* genes are encoded separately at distant locations in the genome. Pull-down assays supported that the genomic *cas* gene organization mirrors the Cas protein assembly, as both Cas3 subunits, Cas7, Cas5a and Cas8a2 were interacting with Csa5. The Cas6 protein could not be identified (Supplementary Table SII). In agreement, the co-refolding of all possible protein combinations showed the necessity of Cas3 for the formation of a stable Cascade complex, indicating that the two Cas3 subunits are not recruited but are rather an integral part of the Cascade I-A complex. In contrast to type I-E systems, pre-crRNA transcription and crRNA production appears to be constitutive, as evidenced by the high abundance of crRNAs detectable in RNA-Seq data. Therefore, regulation of Cascade immunity appears to rely on the activation of Cas protein production. The transcription of the Cascade genes might be activated by the encoded regulator protein Csx1 (Csa3), which is followed by assembly of Cascade:Cas3 and immediately loaded with crRNA to cleave the protospacer DNA target (71). A possible reason for the differences in Cascade assembly might be the thermophilic lifestyle of *T. tenax*, which would make a later recruitment of Cas3 to Cascade assembled in an R-loop structure under elevated temperatures challenging. Interestingly, type I-A systems are exclusively found in thermophilic organisms, supporting the evolutionary conservation of this alternative Cascade formation strategy (6).

The established assembly of the *T. tenax* Cascade allowed us to investigate the type I-A interference reaction and the role of individual subunits in more detail. In Cascade nuclease assays, Cas3″ was identified as the primary deoxyribonuclease subunit, cleaving linear ssDNA substrates exonucleolytically in the dependence of the divalent Mg^{2+} and Mn^{2+} metal ions. Mutations in the conserved HD domain motif (H19A, H55A and D56A) inhibited the cleavage activity. The large group of HD domain proteins comprises enzymes that are primarily involved in nucleic acid metabolism and signal transduction and react on a broad range of substrates, including ssDNA, RNA and R-loop structures (37,72). The degradation of ssDNA substrates by the Cas3″ subunit was previously characterized for different Cascade subtypes from *M. jannaschii*, *S. thermophilus* or *Thermus thermophilus*, and a crystal structure revealed the active site with two bound metal cations (37–39). The observed unspecific cleavage of ssDNA by Cas3″ *in vitro* would pose a problem within the cell. The addition of non-matching synthetic crRNA to the nuclease assays indicated an inhibition of this background cleavage activity. This suggests that Cascade might first bind crRNA, which inhibits the unspecific ssDNA cleavage until the correct dsDNA target is specifically located. Another example for the interplay between Cascade subunits was observed for the binding of crRNAs. The purification procedure, EMSA assays and TEM imaging confirmed that *T. tenax* Cas7 binds unspecifically to both crRNAs and other small RNA contaminants. This interaction resulted in the formation of long helical Cas7 multimer structures. The observation of helical Cas7 filaments was also made for a Cascade sub-complex of the subunits Cas7 and Cas5a of *S. solfataricus* (43). The cryo-electron microscopy structure of the I-E Cascade revealed a seahorse-like shape, in which six copies of the Cas7 subunit are forming the backbone, combined to a Cascade stoichiometry of 1× Cse1, Cas5e, Cas6e, 2× Cse2 and 6× Cas7 (18,31). The binding behavior of *T. tenax* Cas7 embedded in I-A Cascade changed significantly. The shifting of crRNA appeared lower during EMSA assays, the gel filtration showed no formation of extended Cas7 multimers and no helical structures were observed in TEM pictures. These observations indicate a coordinated assembly of I-A Cascade and a capping mechanism by other Cascade subunits that block Cas7 from forming extended multimeric filaments. The exact stoichiometry for I-A Cascade could not be determined as the amount of functional refolding is not known. Future work on the Cascade structure via TEM, crystallization or native mass spectrometry is required to address these questions.

In interference assays the assembled I-A Cascade exhibited crRNA-mediated cleavage of dsDNA molecules. In the targeting event, Cascade facilitates base pairing of the crRNA with the complementary target strand and

additional displacement of the non-target strand to produce the R-loop structure (33,35). An open question is the nature of the crRNA seed sequence specifically for the type I-A CRISPR-Cas system, defined as the minimal sequence complementarity of crRNA and target DNA for binding (32,73). The existence of an adjacent PAM was shown to be essential for directing DNA cleavage (57). Functional PAM sequences for the *T. tenax* I-A Cascade were CCA or CCT, whereas other flanking sequences yielded a reduced (TCA) or impaired activity (TCG). The origin of the *T. tenax* spacers is not known. Therefore, the established interference assay allowed us to identify this basic requirement for Cascade-mediated cleavage and can be used to complement plasmid-based *in vivo* assays that determined the PAM sequences *in vivo* for I-A and I-B systems (11,56,74,75). The main function of the PAM is the discrimination between host CRISPR loci DNA and protospacer/target sequences. Cascade first screens for the specific PAM followed by a helical destabilization and strand invasion of the crRNA, which leads to the R-loop formation (61,76). An outstanding question is the mechanism for the identification of the correct PAM, which triggers the crRNA-loaded Cascade for targeting. Two candidates Csa5 and Cas8a2 that are described as the small and large subunit in Cascade might interact with the DNA target and/or recognize the PAM (5,77). The Cascade-mediated interference assays reflect the stepwise degradation of the dsDNA, as Cas3″ first cleaves in the middle of the looped out non-target strand (at position 43–47 nt) followed by a gradual cleavage in 3′–5′ direction. In a second step, the target strand is then cleaved over the entire length also in 3′–5′ direction, which was reported previously for other members of the Cas3 family (37–39). The identified minimal cleavage products are similar to ones found in the type I-E system with 8–10-nt fragment length, which suggests a conserved cleavage mechanism (35).

In conclusion, the established *in vitro* assembly and interference activity protocols of a type I-A Cascade with synthetic crRNA molecules highlight similarities and differences between this archaeal interference complex and the type I-E Cascade. These studies are expected to aid in the assembly of other Cascade complexes to help us to understand commonalties and differences in the evolution of the diverse type I Cascade machineries.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors are grateful to Prof. Uwe-G. Maier for allocation of the electron microscopic facilities.

FUNDING

Deutsche Forschungsgemeinschaft [DFG, FOR1680] and the Max-Planck Society. Funding for open access charge: [DFG, FOR1680, Max-Planck Society].

Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. Grissa,I., Vergnaud,G. and Pourcel,C. (2009) Clustered regularly interspaced short palindromic repeats (CRISPRs) for the genotyping of bacterial pathogens. *Methods Mol. Biol.*, **551**, 105–116.
3. Mojica,F.J., 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.
4. 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.
5. Makarova,K.S., Haft,D.H., Barrangou,R., Brouns,S.J., Charpentier,E., Horvath,P., Moineau,S., Mojica,F.J., Wolf,Y.I., Yakunin,A.F. *et al.* (2011) Evolution and classification of the CRISPR-Cas systems. *Nat. Rev. Microbiol.*, **9**, 467–477.
6. 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**, e60.
7. Savitskaya,E., Semenova,E., Dedkov,V., Metlitskaya,A. and Severinov,K. (2013) High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in *E. coli*. *RNA Biol.*, **10**, 716–725.
8. Swarts,D.C., Mosterd,C., van Passel,M.W. and Brouns,S.J. (2012) CRISPR interference directs strand specific spacer acquisition. *PLoS One*, **7**, e35888.
9. Yosef,I., Goren,M.G. and Qimron,U. (2012) Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res.*, **40**, 5569–5576.
10. Mojica,F.J., 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.
11. Fischer,S., Maier,L.K., Stoll,B., Brendel,J., Fischer,E., Pfeiffer,F., Dyall-Smith,M. and Marchfelder,A. (2012) An archaeal immune system can detect multiple protospacer adjacent motifs (PAMs) to target invader DNA. *J. Biol. Chem.*, **287**, 33351–33363.
12. Carte,J., Pfister,N.T., Compton,M.M., Terns,R.M. and Terns,M.P. (2010) Binding and cleavage of CRISPR RNA by Cas6. *RNA*, **16**, 2181–2188.
13. Carte,J., Wang,R., 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.
14. 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.
15. Hale,C., Kleppe,K., Terns,R.M. and Terns,M.P. (2008) Prokaryotic silencing (psi)RNAs in *Pyrococcus furiosus*. *RNA*, **14**, 2572–2579.
16. Haurwitz,R.E., Jinek,M., Wiedenheft,B., Zhou,K. and Doudna,J.A. (2010) Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science*, **329**, 1355–1358.
17. 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.
18. 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.
19. Brouns,S.J., Jore,M.M., Lundgren,M., Westra,E.R., Slijkhuis,R.J., Snijders,A.P., 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.
20. Sapranauskas,R., Gasiunas,G., Fremaux,C., Barrangou,R., Horvath,P. and Siksnys,V. (2011) The *Streptococcus thermophilus* CRISPR/Cas system provides immunity in *Escherichia coli*. *Nucleic Acids Res.*, **39**, 9275–9282.
21. Rouillon,C., Zhou,M., Zhang,J., Politis,A., Beilsten-Edmands,V., Cannone,G., Graham,S., Robinson,C.V., Spagnolo,L. and White,M.F. (2013) Structure of the CRISPR interference complex CSM reveals key similarities with cascade. *Mol. Cell*, **52**, 124–134.
22. 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.
23. 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.
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. 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.
26. Deltcheva,E., Chylinski,K., Sharma,C.M., Gonzales,K., Chao,Y., 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.
27. 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.
28. 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. USA*, **109**, E2579–E2586.
29. 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.
30. 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.
31. Wiedenheft,B., Lander,G.C., Zhou,K., Jore,M.M., Brouns,S.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.
32. Wiedenheft,B., van Duijn,E., Bultema,J.B., Waghmare,S.P., Zhou,K., Barendregt,A., Westphal,W., Heck,A.J., Boekema,E.J., Dickman,M.J. *et al.* (2011) RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proc. Natl Acad. Sci. USA*, **108**, 10092–10097.
33. Westra,E.R., van Erp,P.B., Kunne,T., Wong,S.P., Staals,R.H., Seegers,C.L., 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.
34. Westra,E.R., Semenova,E., Datsenko,K.A., Jackson,R.N., Wiedenheft,B., Severinov,K. and Brouns,S.J. (2013) Type I-E CRISPR-cas systems discriminate target from non-target DNA through base pairing-independent PAM recognition. *PLoS Genet.*, **9**, e1003742.
35. 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.
36. Mulepati,S. and Bailey,S. (2013) *In vitro* reconstitution of an *Escherichia coli* RNA-guided immune system reveals unidirectional, ATP-dependent degradation of DNA target. *J. Biol. Chem.*, **288**, 22184–22192.
37. 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.
38. 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.
39. 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.
40. Makarova,K.S., Wolf,Y.I., Snir,S. and Koonin,E.V. (2011) Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. *J. Bacteriol.*, **193**, 6039–6056.
41. 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.
42. Siebers,B., Zaparty,M., Raddatz,G., Tjaden,B., Albers,S.V., Bell,S.D., Blombach,F., Kletzin,A., Kyrpides,N., Lanz,C. *et al.* (2011) The complete genome sequence of *Thermoproteus tenax*: a physiologically versatile member of the *crenarchaeota*. *PLoS One*, **6**, e24222.
43. Lintner,N.G., Kerou,M., Brumfield,S.K., Graham,S., Liu,H., Naismith,J.H., Sdano,M., Peng,N., She,Q., 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.
44. Brock,T.D., Brock,K.M., Belly,R.T. and Weiss,R.L. (1972) Sulfolobus-new genus of sulfur-oxidizing bacteria living at low ph and high-temperature. *Arch. Mikrobiol.*, **84**, 54–68.
45. Su,A.A., Tripp,V. and Randau,L. (2013) RNA-Seq analyses reveal the order of tRNA processing events and the maturation of C/D box and CRISPR RNAs in the hyperthermophile *Methanopyrus kandleri*. *Nucleic Acids Res.*, **41**, 6250–6258.
46. Schurer,H., Lang,K., Schuster,J. and Morl,M. (2002) A universal method to produce *in vitro* transcripts with homogeneous 3′ ends. *Nucleic Acids Res.*, **30**, e56.
47. Siebers,B., Wendisch,V.F. and Hensel,R. (1997) Carbohydrate metabolism in *Thermoproteus tenax*: *in vivo* utilization of the non-phosphorylative Entner-Doudoroff pathway and characterization of its first enzyme, glucose dehydrogenase. *Arch. Microbiol.*, **168**, 120–127.
48. Schmidt,C., Lenz,C., Grote,M., Luhrmann,R. and Urlaub,H. (2010) Determination of protein stoichiometry within protein complexes using absolute quantification and multiple reaction monitoring. *Anal. Chem.*, **82**, 2784–2796.
49. Rappsilber,J., Ishihama,Y. and Mann,M. (2003) Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.*, **75**, 663–670.
50. Cox,J. and Mann,M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.*, **26**, 1367–1372.
51. Cox,J., Neuhauser,N., Michalski,A., Scheltema,R.A., Olsen,J.V. and Mann,M. (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.*, **10**, 1794–1805.
52. Rachel,R., Meyer,C., Klingl,A., Gurster,S., Heimerl,T., Wasserburger,N., Burghardt,T., Kuper,U., Bellack,A., Schopf,S. *et al.* (2010) Analysis of the ultrastructure of archaea by electron microscopy. *Methods Cell. Biol.*, **96**, 47–69.
53. Klingl,A., Moissl-Eichinger,C., Wanner,G., Zweck,J., Huber,H., Thomm,M. and Rachel,R. (2011) Analysis of the surface proteins of Acidithiobacillus ferrooxidans strain SP5/1 and the new, pyrite-oxidizing Acidithiobacillus isolate HV2/2, and their possible involvement in pyrite oxidation. *Arch. Microbiol.*, **193**, 867–882.
54. Avis,J.M., Conn,G.L. and Walker,S.C. (2012) Cis-acting ribozymes for the production of RNA in vitro transcripts with defined 5′ and 3′ ends. *Methods Mol. Biol.*, **941**, 83–98.
55. Birikh,K.R., Heaton,P.A. and Eckstein,F. (1997) The structure, function and application of the hammerhead ribozyme. *Eur. J. Biochem.*, **245**, 1–16.
56. Shah,S.A., Erdmann,S., Mojica,F.J. and Garrett,R.A. (2013) Protospacer recognition motifs: mixed identities and functional diversity. *RNA Biol.*, **10**, 891–899.

57. Gudbergsdottir,S., Deng,L., Chen,Z., Jensen,J.V., Jensen,L.R., She,Q. and Garrett,R.A. (2011) Dynamic properties of the Sulfolobus CRISPR/Cas and CRISPR/Cmr systems when challenged with vector-borne viral and plasmid genes and protospacers. *Mol. Microbiol.*, **79**, 35–49.
58. Manica,A., Zebec,Z., Teichmann,D. and Schleper,C. (2011) *In vivo* activity of CRISPR-mediated virus defence in a hyperthermophilic archaeon. *Mol. Microbiol.*, **80**, 481–491.
59. 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. USA*, **108**, 21218–21222.
60. 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.
61. Marraffini,L.A. and Sontheimer,E.J. (2010) Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature*, **463**, 568–571.
62. Wang,R., 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.
63. Hatoum-Aslan,A., Samai,P., Maniv,I., Jiang,W. and Marraffini,L.A. (2013) A ruler protein in a complex for antiviral defense determines the length of small interfering CRISPR RNAs. *J. Biol. Chem.*, **288**, 27888–27897.
64. Hale,C.R., Majumdar,S., Elmore,J., Pfister,N., Compton,M., Olson,S., Resch,A.M., Glover,C.V. III, Graveley,B.R., Terns,R.M. *et al.* (2012) Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. *Mol. Cell*, **45**, 292–302.
65. Roy,R., Hohng,S. and Ha,T. (2008) A practical guide to single-molecule FRET. *Nat. Methods*, **5**, 507–516.
66. Weiss,S. (2000) Measuring conformational dynamics of biomolecules by single molecule fluorescence spectroscopy. *Nat. Struct. Biol.*, **7**, 724–729.
67. Schmidt,C., Kramer,K. and Urlaub,H. (2012) Investigation of protein-RNA interactions by mass spectrometry—techniques and applications. *J. Proteomics*, **75**, 3478–3494.
68. 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.
69. Westra,E.R., Pul,U., Heidrich,N., Jore,M.M., Lundgren,M., Stratmann,T., Wurm,R., Raine,A., Mescher,M., Van Heereveld,L. *et al.* (2010) H-NS-mediated repression of CRISPR-based immunity in *Escherichia coli* K12 can be relieved by the transcription activator LeuO. *Mol. Microbiol.*, **77**, 1380–1393.
70. Arslan,Z., Stratmann,T., Wurm,R., Wagner,R., Schnetz,K. and Pul,U. (2013) RcsB-BglJ-mediated activation of Cascade operon does not induce the maturation of CRISPR RNAs in *E. coli* K12. *RNA Biol.*, **10**, 708–715.
71. Lintner,N.G., Frankel,K.A., Tsutakawa,S.E., Alsbury,D.L., Copie,V., Young,M.J., Tainer,J.A. and Lawrence,C.M. (2011) The structure of the CRISPR-associated protein Csa3 provides insight into the regulation of the CRISPR/Cas system. *J. Mol. Biol.*, **405**, 939–955.
72. Yakunin,A.F., Proudfoot,M., Kuznetsova,E., Savchenko,A., Brown,G., Arrowsmith,C.H. and Edwards,A.M. (2004) The HD domain of the *Escherichia coli* tRNA nucleotidyltransferase has 2 $'$,3 $'$-cyclic phosphodiesterase, 2 $'$-nucleotidase, and phosphatase activities. *J. Biol. Chem.*, **279**, 36819–36827.
73. Semenova,E., Jore,M.M., Datsenko,K.A., Semenova,A., Westra,E.R., Wanner,B., van der Oost,J., Brouns,S.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. USA*, **108**, 10098–10103.
74. Manica,A., Zebec,Z., Steinkellner,J. and Schleper,C. (2013) Unexpectedly broad target recognition of the CRISPR-mediated virus defence system in the archaeon *Sulfolobus solfataricus*. *Nucleic Acids Res.*, **41**, 10509–10517.
75. Peng,W., Li,H., Hallstrom,S., Peng,N., Liang,Y.X. and She,Q. (2013) Genetic determinants of PAM-dependent DNA targeting and pre-crRNA processing in *Sulfolobus islandicus*. *RNA Biol.*, **10**, 738–748.
76. 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.
77. 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**, 762–769.

SUPPLEMENTARY DATA

SUPPLEMENTARY FIGURES

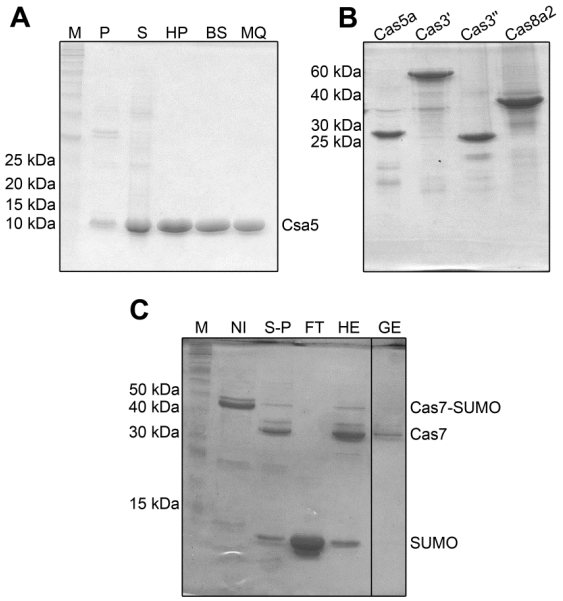


Fig. S1: Protein purification of all six Cascade subunits. (**A**) The soluble protein Csa5 was purified via a two-column strategy. Depicted are the pellet (P) and supernatant (S) fractions after cell lysis and ultracentrifugation, followed by the heat precipitated (HP, 90°C) Csa5 sample, further purified samples via Blue Sepharose affinity chromatography (BS) and MonoQ anion-exchange chromatography (MQ) on a 15% SDS gel next to the protein marker (M). (**B**) The insoluble subunits Cas5a, Cas3', Cas3'' and Cas8a2 were individually purified via inclusion body isolation and solubilization in 6 M GdmCl. The 15% SDS gel shows the subunits before usage in Cascade assembly. (**C**) The soluble subunit Cas7 was fused with a SUMO tag to omit cross-contamination with *E. coli* RNA and oligomerization. Depicted are the individual protein fractions alongside the protein marker (M) separated on a 15% SDS-PAGE. Cas7-SUMO was purified via Ni-NTA chromatography (NI), followed by a SUMO protease (S-P) treatment during dialysis. After concentration of native Cas7 (centrifugal filter units, MWCO: 30 kDa) the cleaved off SUMO tag was visible in the flow-through (FT). Cas7 was further purified with Heparin cation-exchange chromatography (HE) and gel filtration (GE) resulting in monomeric Cas7.

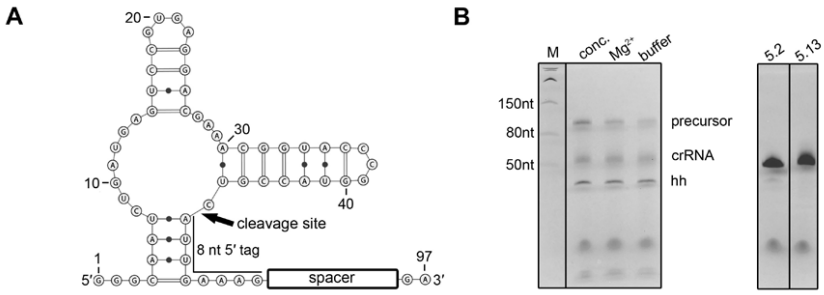


Fig. S2: Production of synthetic crRNAs. (**A**) Mature crRNAs 5.2 and 5.13 were produced via fusion of the crRNA sequence to the sequence of a minimal *cis*-acting hammerhead (hh) ribozyme, resulting in a self-cleavage (arrow marking the cleavage site) directly upstream of the 8 nt 5'-tag. (**B**) The self-cleavage of the transcript (precursor) into crRNA and hh was induced by a direct incubation at 60°C (conc.), dilution of the sample in 4 vol of 30 mM $MgCl_2$ in DEPC-H_2O (Mg^{2+}) or dilution in 4 vol of *in vitro* transcription buffer (buffer) followed by heat incubation, yielding the synthetic mature crRNAs 5.2 (50 nt) and 5.13 (54 nt).

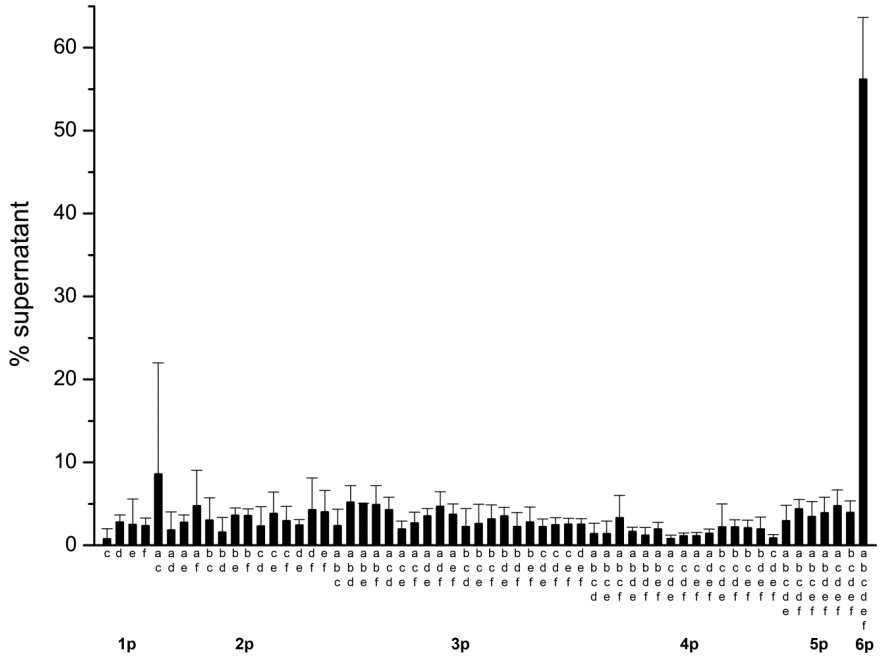


Fig. S3: Co-refolding combinations of Cascade subunits. All possible 63 combinations of Cascade subunits were reconstituted by the rapid dilution method and for each protein, the obtained bands on the SDS gel compared between soluble protein (supernatant) and aggregated protein (pellet). Plotted is the average of co-refolded supernatant (%) of 60 combinations that contained insoluble proteins (c: Cas5a, d: Cas3', e: Cas3'', f: Cas8a2), while the soluble proteins (a: Csa5, b: Cas7) are not taken into account. Only the combination that contained all six Cascade subunits (6p) showed a recovery of soluble protein above 50%, in contrast to combinations with missing subunits (1p-5p) that resulted in soluble protein of less than 10%.

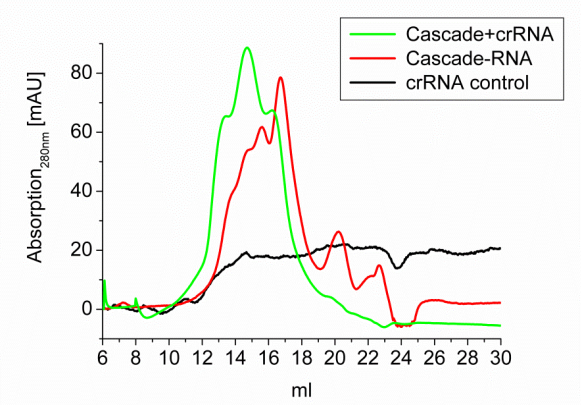


Fig. S4: Size-exclusion chromatograms of assembled Cascade. The Cascade subunits (Csa5, Cas7, Cas5a, Cas3', Cas3'' and Cas8a2) were assembled via the co-refolding procedure and incubated with synthetic crRNA 5.2 (crRNA) or no RNA (-RNA) and protein was followed via absorption at 280 nm. A protein-free sample of the crRNA served as a control of the elution profile (crRNA control). The chromatograms demonstrate that protein is eluting in the fractions 10-18 ml (Fig. 2A).

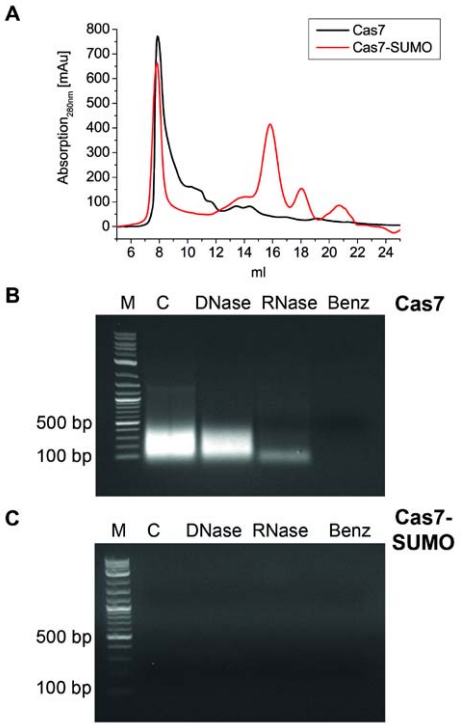


Fig. S5: Comparison of Cas7 and Cas7-SUMO purification procedures. (**A**) The two Cas7 versions (native purification or SUMO-tagged) exhibit different elution profiles on size-exclusion columns. The native purification of Cas7 results in completely assembled large multimeric complexes (black line, elution volume: 8.2 ml), while the Cas7-SUMO version elutes after cleavage of the SUMO-tag as a monomer (red line, elution volume: 15.5 ml). (**B**) 100 µg of Cas7 loaded on a 1% agarose gel and stained with EtBr next the DNA marker shows cross-contamination of *E. coli* RNA (C), which are cleaved by RNase A or Benzonase, but not by DNaseI. (**C**) The purification of Cas7-SUMO is free of any cross-contamination, which is shown by loading identical amounts of protein on a 1% agarose gel.

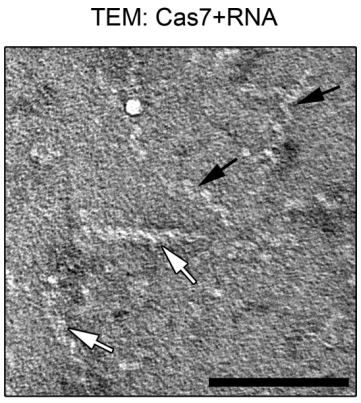


Fig. S6: Helical artifact formation of Cas7 with *E. coli* RNA. A typical TEM analysis of a Cas7 purification resulted in cross-contamination with *E. coli* RNA and the formation of helical (black arrows) and double-helical filaments (white arrows) of up to 50 nm length. Scale bar: 50 nm.

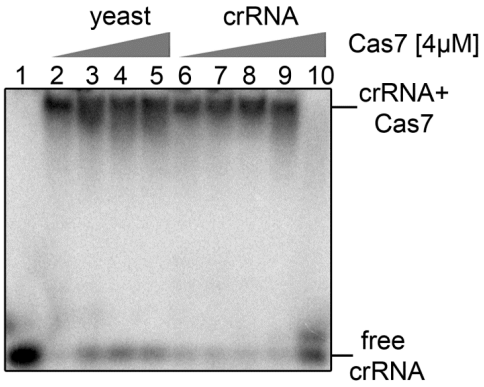


Fig. S7: Specificity of Cas7 binding to crRNA. At a fixed Cas7 concentration of 4 µM the specificity of crRNA binding (crRNA 5.2) was tested by competing amounts of yeast RNA (lane 2-6: 0, 10, 50, 250, 500 ng) or increasing amounts of unlabeled crRNA 5.2 (lane 7-10: 1:1, 1:10, 1:100, 1:1000). Only the highest concentration of unlabeled crRNA abolished the Cas7 shift indicating an unspecific binding of Cas7 to crRNA. Lane 1 serves as a crRNA loading control.

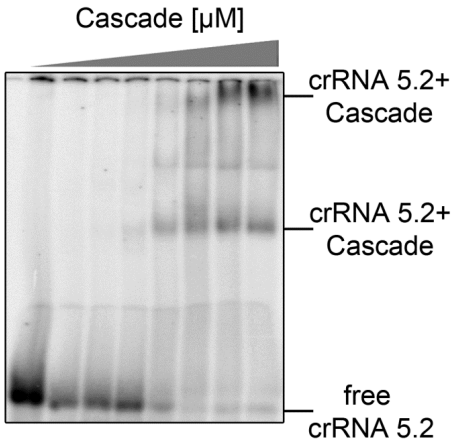


Fig. S8: Cascade binding to the synthetic crRNA 5.2. The assembled Cascade was tested in EMSAs for specific binding to the constructed crRNA 5.2. Increasing concentrations (0, 0.125, 0.25, 0.5, 1, 2, 4, 5 µM) of Cascade in the presence of yeast RNA showed a shifting of the 5'-labeled (with [γ-^{32}P]-ATP) crRNA 5.2 on 6% non-denaturing gels comparable to the synthetic crRNA 5.13.

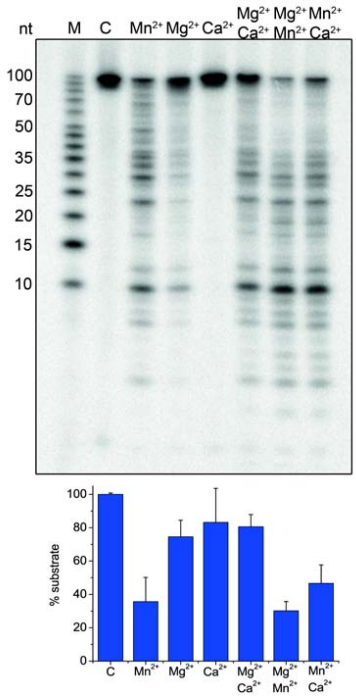


Fig. S9: The cleavage of ssDNA by Cascade is metal-dependent. Cascade was incubated with the 5'-[γ-^{32}P]-ATP labeled short ssDNA substrate (int 5.2_CCT for) and different metal combinations for 10 min at 70°C and the cleavage products were separated on a 15% denaturing gel next to the low molecular weight marker. The assay was repeated three times and in each reaction the remaining substrate estimated via line profile plots (Image J), which indicated that Cascade in the combination with Mg^{2+} and Mn^{2+} ions has the highest activity (~70% cleaved) on ssDNA.

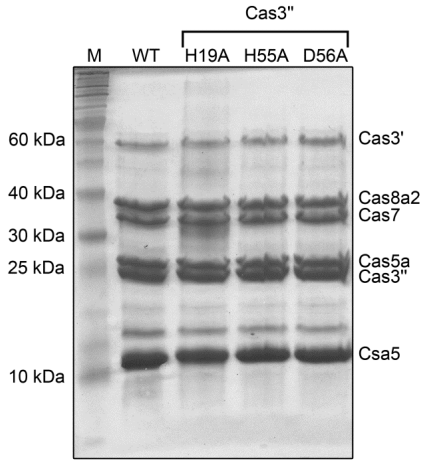


Fig. S10: Co-reconstitution of Cascade with wild-type Cas3'' and Cas3'' mutants. All Cascade complexes were assembled via the same co-refolding procedure of insoluble proteins including either Cas3'' WT, Cas3'' H19A, Cas3'' H55A or Cas3'' D56A. A 15 µl sample of the supernatant after Cascade co-refolding into native buffer, centrifugation and concentration was loaded on the 15% SDS gel alongside the protein marker (M), demonstrating the same reconstitution efficiency for all Cas3'' proteins within Cascade.

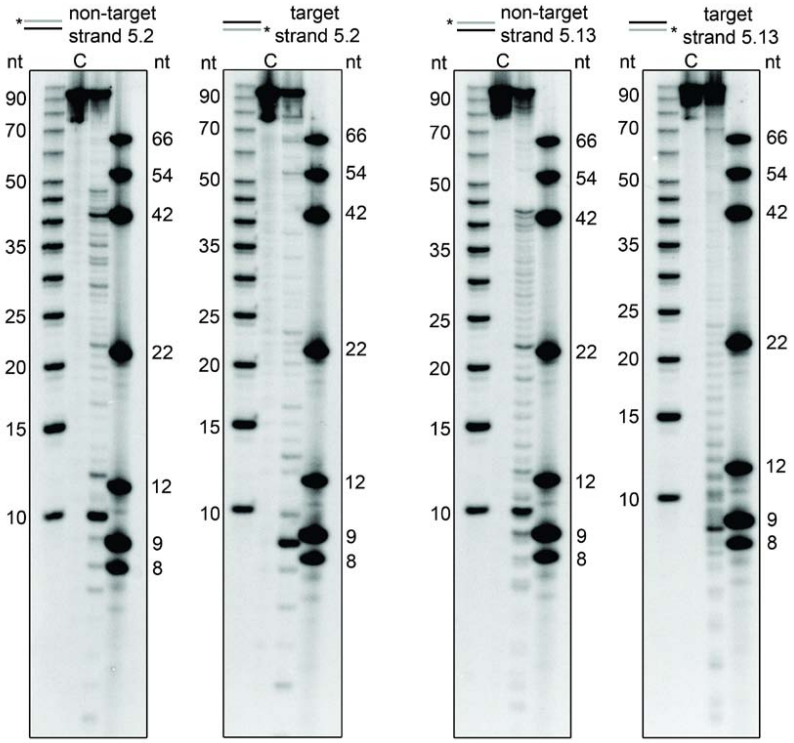


Fig. S11: Size determination of the cleavage products appearing in the Cascade interference reaction. The 5'-[γ-^{32}P]-ATP labeled non-target and target strand of each dsDNA substrate 5.2 or 5.13 were separated on 10% denaturing sequencing gels alongside the 5'-labeled low molecular weight marker (10-100 nt) or a mixture of seven ssDNA fragments (8-66 nt) to specify the fragment length of cleavage hot spots. Lanes marked with C are the loading controls of the dsDNA.

SUPPLEMENTARY TABLES

Table SI: Oligonucleotides and RNA sequences for cloning and substrate generation

name	**sequence**
cr5.2h (125bp)	BamHI – T7 promoter – hammerhead – crDNA, 5.2 – HindIII
cr5.2h for (119nt)	GATCCTAATACGACTCACTATAGGGCAATCTGATGAGTCCGTGAGGACGAAACGGTACC CGGTACCGTCATTGAAAGCGTTGATGCGGCCGCGACTGGCTGACTCAGCTATTACGTTG A
cr5.2h rev(119nt)	AGCTTCAACGTAATAGCTGAGTCAGCCAGTCGCGGCCGCATCAACGCTTTCAATGACGG TACCGGGTACCGTTTCGTCCTCACGGACTCATCAGATTGCCCTATAGTGAGTCGTATTA G
cr5.13h (126bp)	HindIII – T7 promoter – hammerhead – crDNA, 5.13 – EcoRI
cr5.13h for (120nt)	AGCTTAATACGACTCACTATAGGCAATCTGATGAGTCCGTGAGGACGAAACGGTACCCG GTACCGTCATTGAAAGATCGCTTGTCCAACCGGGCTCCTCTATATGTCGTCATTAGCTT AG
cr5.13h rev (120nt)	AATTCTAAGCTAATGACGACATATAGAGGAGCCCGGTTGGACAAGCGATCTTTCAATGA CGGTACCGGGTACCGTTTCGTCCTCACGGACTCATCAGATTGCCTATAGTGAGTCGTAT TA
cr5.2PCRf (19nt)	GGGGATCCTAATACGACTC
cr5.2PCRr (20nt)	TCAACGTAATAGCTGAGTCA
cr5.13PCRf (19nt)	GCCAAGCTTAATACGACTC
cr5.13PCRr (20nt)	TCTAAGCTAATGACGACATA
crRNA 5.2 (50nt)	AUUGAAAGCGUUGAUGCGGCCGCGACUGGCUGACUCAGCUAUUACGUUGA
crRNA 5.13 (54nt)	AUUGAAAGAUCGCUUGUCCAACCGGGCUCCUCUAUAUGUCGUCAUUAGCUUAGA
int 5.2 (93bp)	SalI – 5'-proto – PAM (NNN) – spacer 5.2 – 3'-proto – HindIII
int 5.2_CCT for (93nt)	GTCGACTAATACGACTCACTATAGACCTCGTTGATGCGGCCGCGACTGGCTGACTCAGC TATTACGTTCCGCTGAGCAATAACTAGCAAGCTT
int 5.2_CCT rev (93nt)	AAGCTTGCTAGTTATTGCTCAGCGGAACGTAATAGCTGAGTCAGCCAGTCGCGGCCGCA TCAACGAGGTCTATAGTGAGTCGTATTAGTCGAC
int 5.2_CCA for (93nt)	GTCGACTAATACGACTCACTCGCAGCCACGTTGATGCGGCCGCGACTGGCTGACTCAGC TATTACGTTCCGCTGAGCAATAACTAGCAAGCTT
int 5.2_CCA rev (93nt)	AAGCTTGCTAGTTATTGCTCAGCGGAACGTAATAGCTGAGTCAGCCAGTCGCGGCCGCA TCAACGTGGCTGCGAGTGAGTCGTATTAGTCGAC
int 5.2_TCA for (93nt)	GTCGACTAATACGACTCACTCGCAGTCACGTTGATGCGGCCGCGACTGGCTGACTCAGC TATTACGTTCCGCTGAGCAATAACTAGCAAGCTT
int 5.2_TCA rev (93nt)	AAGCTTGCTAGTTATTGCTCAGCGGAACGTAATAGCTGAGTCAGCCAGTCGCGGCCGCA TCAACGTGACTGCGAGTGAGTCGTATTAGTCGAC
int 5.2_TCG for (93nt)	GTCGACTAATACGACTCACTCGCAGTCGCGTTGATGCGGCCGCGACTGGCTGACTCAGC

	TATTACGTTCCGCTGAGCAATAACTAGCAAGCTT
int 5.2_TCG rev (93nt)	AAGCTTGCTAGTTATTGCTCAGCGGAACGTAATAGCTGAGTCAGCCAGTCGCGGCCGCA TCAACGCGACTGCGAGTGAGTCGTATTAGTCGAC
int 5.2_AAA for (93nt)	GTCGACTAATACGACTCACTCGCAGAAACGTTGATGCGGCCGCGACTGGCTGACTCAGC TATTACGTTCCGCTGAGCAATAACTAGCAAGCTT
int 5.2_AAA rev (93nt)	AAGCTTGCTAGTTATTGCTCAGCGGAACGTAATAGCTGAGTCAGCCAGTCGCGGCCGCA TCAACGTTTCTGCGAGTGAGTCGTATTAGTCGAC
int 5.2_Rep for (98nt)	GTCGACTAATACGACTCACTATAGAATTGAAAGCGTTGATGCGGCCGCGACTGGCTGAC TCAGCTATTACGTTCCGCTGAGCAATAACTAGCAAGCTT
int 5.2_Rep rev (98nt)	AAGCTTGCTAGTTATTGCTCAGCGGAACGTAATAGCTGAGTCAGCCAGTCGCGGCCGCA TCAACGCTTTCAATTCTATAGTGAGTCGTATTAGTCGAC
int 5.13 (97bp)	SalI – 5'-proto – PAM (CCT) – spacer 5.13 – 3'-proto – HindIII
int 5.13_CCT for (97nt)	GTCGACTAATACGACTCACTATAGGCCTATCGCTTGTCCAACCGGGCTCCTCTATATGT CGTCATTAGCTTACCGCTGAGCAATAACTAGCAAGCTT
int 5.13_CCT rev (97nt)	AAGCTTGCTAGTTATTGCTCAGCGGTAAGCTAATGACGACATATAGAGGAGCCCGGTTG GACAAGCGATAGGCCTATAGTGAGTCGTATTAGTCGAC
Csa5-His for (20nt)	TCCTAATACGACTCACTATA
Csa5-His rev (23nt)	GGAGCCACCCAAGCTTCCCCTTA
Cas3''-H19A for (34nt)	CCAGACCTACGAAGACGCCATCACGCAGGCTCTG
Cas3''-H19A rev (34nt)	CAGAGCCTGCGTGATGGCGTCTTCGTAGGTCTGG
Cas3''-H55A for (32nt)	CTAGCCGTGGAGTTCGCCGACCTAGGCAAGCT
Cas3''-H55A rev (32nt)	AGCTTGCCTAGGTCGGCGAACTCCACGGCTAG
Cas3''-D56A for (29nt)	CGTGGAGTTCCACGCCCTAGGCAAGCTCG
Cas3''-D56A rev (29nt)	CGAGCTTGCCTAGGGCGTGGAACTCCACG
MI_8 (8nt)	CATCAACG
MI_9 (9nt)	GCATCAACG
MI_12 (12nt)	GCCGCATCAACG
MI_22 (22nt)	GCCAGTCGCGGCCGCATCAACG
MI_42 (42nt)	GCATCTAATACGACTCACTATAGGGAGCGAATGAAACGAGCG
MI_54 (54nt)	GCAGCACTCGAGCAATTGTTACACGAAACCTTTACCCACACGTTCCACGGTGCC
MI_66 (66nt)	AGCTTTAATACGACTCACTATAGATTAATCCCATAATACTTTTCTAGGTCTGGGCGGAA TGGATCC

Table SII: MS analysis of *in vivo* pull-down Cascade experiments. Proteins co-purifying with Csa5 were identified by *in solution* trypsin digestion and followed by MS.

ORF	**protein**	**1) intensity**	**1) coverage [%]**	**2) intensity**	**2) coverage [%]**
TTX_1250	Csa5	8.46E+10	96.2	6.22E+10	96.2
TTX_1251	Cas7	-	-	836940	7.9
TTX_1252	Cas5a	-	-	-	3.5
TTX_1253	Cas3'	30243000	14	134950	1.2
TTX_1254	Cas3''	-	-	71119	4.9
TTX_1255	Cas8a2	12116000	14.4	-	-
TTX_0235	Cas7	27699000	17	138670	3.2

2.4 Follow-up on Publication 3: RNA-binding features of *Tt* Cas7

I followed up on the the collaborative study and investigated the RNA binding properties of *Tt* Csa2, the type I-A Cas7 homolog. As *Tt* Cas7 binds an unspecific poly-U_{20} sequence in comparable fashion and in the same affinity range as *Tp* Csc2 (Fig. 9), I chose to map the RNA surface using a similar approach. Mass spectrometric analysis identified six crosslinked peptides and 5 specific RNA-amino acid crosslinks on *Tt* Cas7 (Fig. 9).

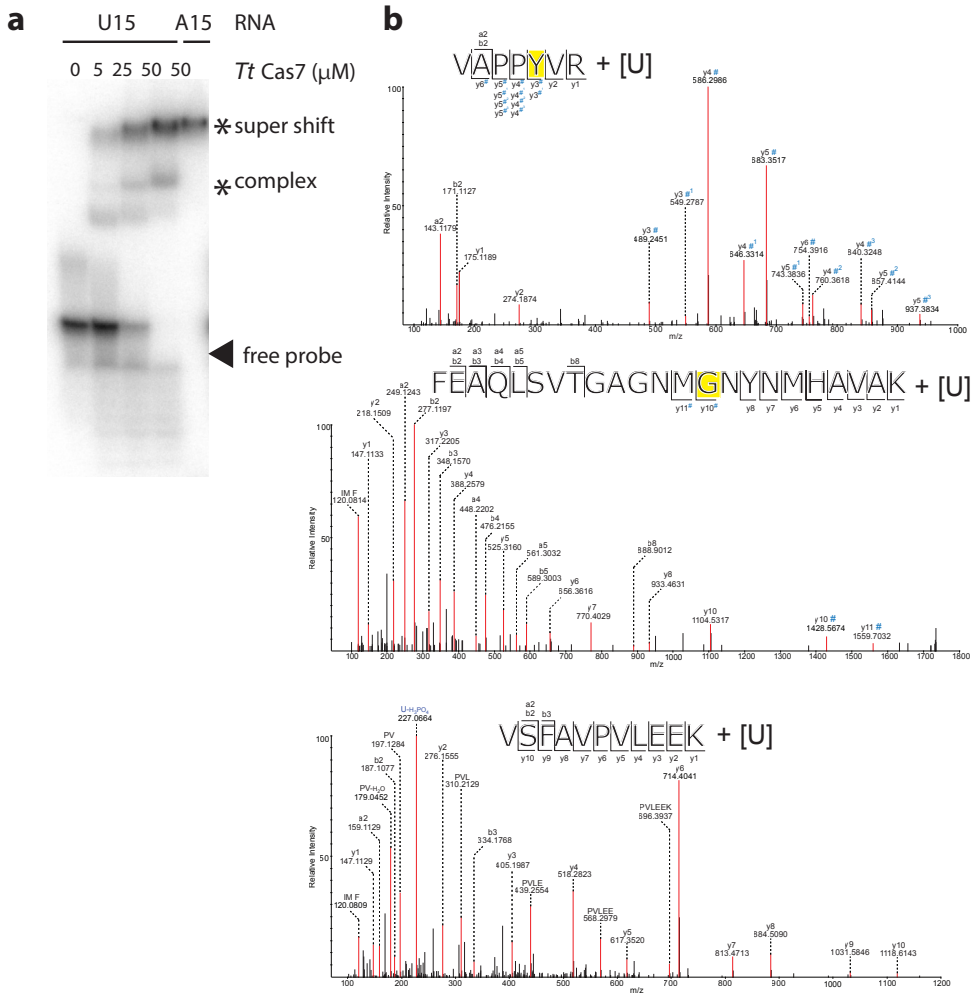


Figure 9. RNA-binding properties of *Tt* Cas7.

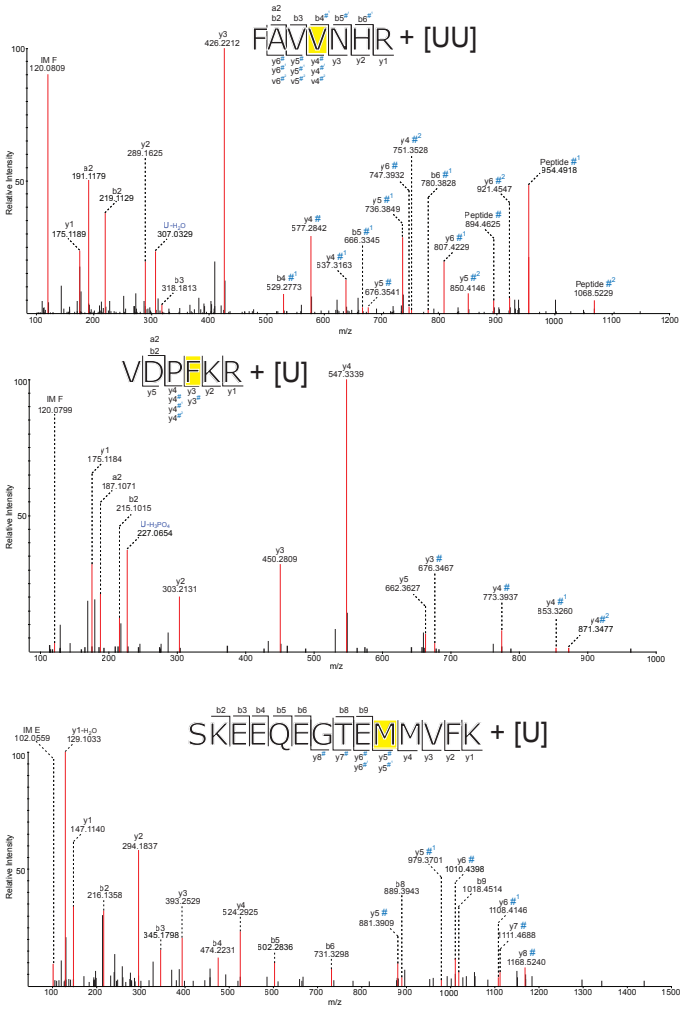


Figure 9: RNA-binding properties of *Tt* Cas7.
a. Electrophoretic mobility shift assays (EMSAs) with wild-type *Tp* Csa2. Left panel: EMSAs were carried out with ^{32}P-5′-end labeled poly-U_{15} or poly-A_{15} RNAs and increasing concentrations of *Tp* Csc2 (0, 5, 25, 50 μM). The positions of the free RNA probe (arrowhead) and of the RNA-bound complexes (asterisks) are shown on the right.
b. MS/MS mass spectra of *Tp* Csc2 peptides, carrying an additional mass corresponding to one (panel 2 and 3) or two (panel 1) uracil nucleotides associated with the respective amino acid. Peptide sequence and the observed fragment ions are indicated as follows: # adduct with [Fragment of U: -C_3O]. #^{1} adduct with [U′]. #^{2} adduct with [U-H_3PO_4]. #^{3} adduct with [U-H_2O].

The sequence conservation amongst homologous Cas proteins is weak, therefore I chose to compare the localization of the crosslinked residues on a structural level and investigate if the same features which I observed for other Cas7 proteins (Hrle *et al.*, 2014, Fig. 4) are present. As high-resolution structural data of *Tt* Csa2 were not available, I generated a structural model using the Phyre2 sever (Fig. 10). The mapped residues localize to two regions: the predicted lid domain and the central cleft defined by the interface of the RRM-like core and the α1-β2 insertion domain. Moreover the identified peptide lies adjacent to the conserved positively charged cleft (Fig. 10, bottom). This region harbors conserved arginine and lysine residues, which are found in other Cas7 family homologs (inset in Fig. 10). These results are in agreement with our previously presented findings (Hrle *et al.*, 2014, Fig. 5) and supported by high-resolution data of the well-studied type I-E interference complex from the Wiedenheft laboratory [68].

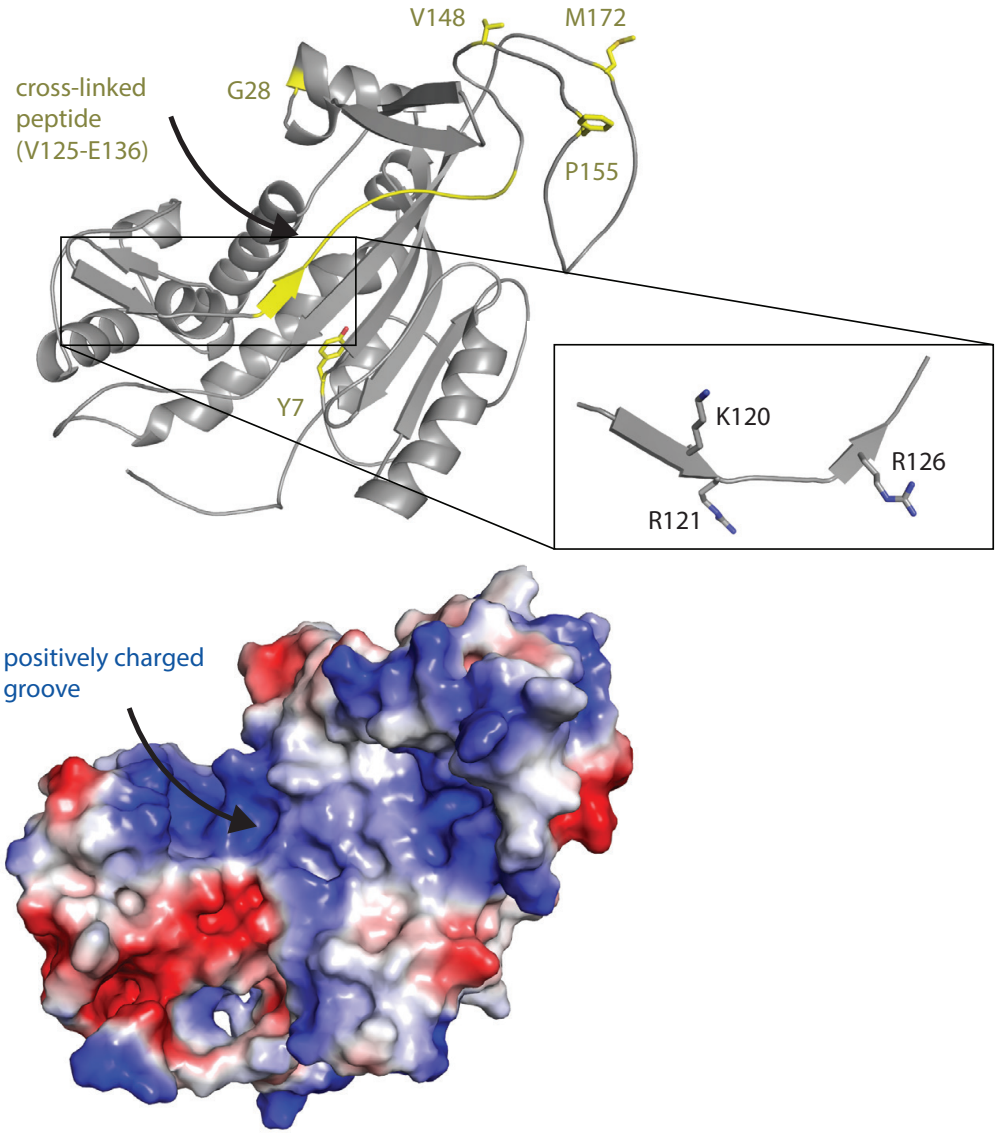


Figure 10:
Top: the predicted three dimensional model of *Tt* Cas7 (grey) generated using the Phyre 2 server. Direct U_{20}-crosslinked amino acid residues are represented as sticks (yellow) and an arrow points to the location of the identified peptide (highlighted in yellow). The conserved cleft (zoom in) harbors positively charged residues. A reoccurring feature in the structurally studied Cas7 proteins. Bottom: Surface potential distribution. Positive charge is shown in blue, negative in red. An arrow points to the prominent positively charged groove.

3 Discussion

The CRISRP field is expanding rapidly and structural studies are published at a fast pace. New insights from structural data of single proteins and interference complexes of all three types have become available in the past few months: The crystal structure of a type I interference complex [68], several structures of Cas9 [60, 61, 83], EM structures of type III-A [65] and an EM-aided pseudo-atomic model of the type III-B interference complex [69]. The following two sections discuss the two major achievements of the study. First, how our single Cas7 protein structures contribute to defining features that distinguish this protein family from others. Second, how the newly published structural data of entire complexes embed the knowledge into a functional context of the CRISPR/Cas adaptive immune response and places the results presented in this thesis in the wider perspective.

3.1 The Cas7 superfamily – a structural perspective

Cas protein families have been established on the basis of secondary structure predictions and sequence analyses [30]. These have been shown to correlate with the functional properties of the proteins. However, the high level of sequence divergence poses an obstacle. On the one hand, the degree of sequence diversity is a means of establishing lineages. On the other hand, it impairs the predictions of structural features and mechanistic details. This is why high-resolution structural data are paramount for our understanding of Cas proteins. The first published structure of a Cas7 representative was that of *Ss* Csa2, a type I-A homolog of *Tt* Cas7 (one of the subjects of this thesis) [63]. The *Ss* Csa2 structure laid the initial foundation for future structure-based classification and comparisons. The crystal structures of *Mk* Csm3 and *Tp* Csc2, the model of *Tt* Cas7 and recently published new structures of Cas7 family proteins of I-E and III-B now fully allow a comprehensive comparison of Cas protein families and the Cas7 sub-types alike [68, 69, 84, 85].

3.1.1 Common denominators and sub-type specificities of Cas7 proteins

The three proteins that were the center of investigation in this study – *Mk* Csm3, *Tp* Csc2 and *Tt* Cas7 – had been predicted to belong to the Cas7 protein family, yet no functional or structural data were available [70]. We compared our individual crystal structures of *Mk* Csm3, *Tp* Csc2 to representatives of three other major Cas protein families, namely Cas7, Cas5 and Cas6. All three protein families are representatives of the RAMP superfamily [29, 30]. Central to their architecture is at least one RRM-like domain and, despite limited sequence conservation, most of the proteins have a conserved glycine-rich region between α_2 and β_2 of the RRM domain [70]. The functional relevance of this reoccurring sequence motif remains elusive. However, one can speculate on a role in mediating RNA interaction by granting structural flexibility to a crucial and otherwise rather rigid structural element.

The features that characterize the different Cas protein families (5/6/7) within the RAMP superfamily are the absence or presence of additional peripheral domains surrounding the RRM-like core [29, 30]. These domains are defined by insertions between the secondary structure elements of the core; the distinct arrangements of secondary structure elements are specific to each protein family (Fig. 11). In case of the Cas7 family proteins, these peripheral domains give the proteins their crescent shape and unique RNA binding capabilities. Four major features distinguish Cas7 proteins from Cas6 and Cas5 representatives. Each feature is described below based on the comparison of our and all known structures.

RRM-like domain

All Cas7 protein structures, including *Mk* Csm3, *Tp* Csc2 and *Tt* Csa2, harbor one central RRM-like domain, with a β_1-α_2-β_2-β_3-α_2-β_4 topology. Four β-strands fold into a twisted antiparallel β-sheet, lined by two α-helices. As in all Cas7 proteins described to date, the strands β_1–β_3 lack the consensus sequence motifs RNP2 and RNP1 reported to be necessary and sufficient for binding RNA molecules in RNPs (ribonucleoproteins) [27, 30]. Thus Cas7 proteins mediate these interactions in a different manner. Structural super-positioning of the RRM-like domains with Cas7, Cas6 and Cas5 representatives highlights the identical arrangement of the secondary structure elements across the proteins families. The presence of either one or two RRM-like domains was long seen

as a discriminating feature between Cas protein families, as all Cas6 proteins contain two sequential RRM-like domains, linked by β_4 of the first one [70]. However, recent structures of the Cas7-like protein Cmr1 have extended the feature of a second RRM-like domain to the Cas7 family [69, 86]. Conversely, Cas5 proteins with two RRM-like domains have been predicted bioinformatically, but confirmation on structural level is pending [70].

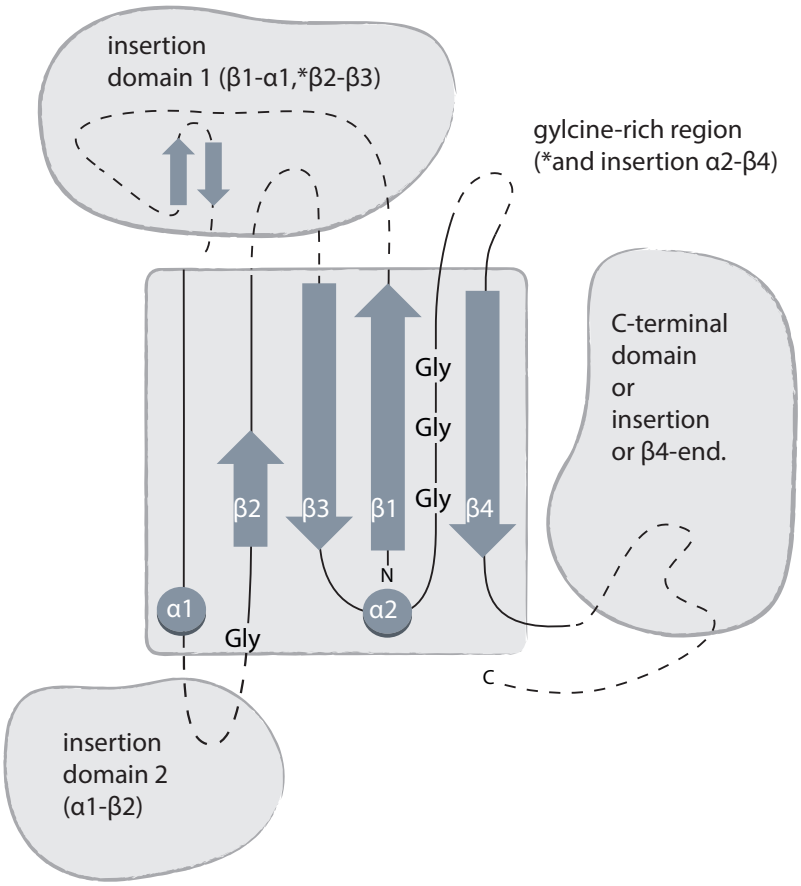


Figure 11: Domain organisation of RAMPs
Schematic representation of the domain organization of the RAMP superfamily. The core domain with a β_1-α_2-β_2-β_3-α_2-β_4 topology is highlighted by a gray square. Peripheral domains are constituted by insertions between the secondary structure elements of the core. A β-hairpin in domain 1 and a glycine-rich region between α_2-β_4 are common to Cas5 Cas7 and Cas6 protein families. Apart from that, Cas family proteins vary in the absence, presence of and different topological arrangements within the peripheral domains.

Domain 1

The domain 1, which we termed the lid (Jackson *et al.* refer to it as the thumb) is defined by secondary structure elements of two insertions between β_1 and α_1, as well as β_2 and β_3 (Fig. 11). The most prominent characteristic is its flexibility. It can change from partially unstructured to hairpin and helical structures, depending on the functional requirements (mediating protein–protein or protein–RNA interaction). One hairpin between β_1 and α_1 is present in Cas7, Cas6 and Cas5 protein families [29]. A second hairpin is formed by an elaborate insertion between β_2 and β_3 and appears to be a Cas7 family denominator. In all but *Tp* Csc2 and *Ec* CasC it was structurally not resolved (Figure 12, top panel). The crystal structure of the entire Cascade complex demonstrated that the hairpin observed in single-protein structures of *Tp* Csc2 is not a packing artifact, but functionally relevant: In *Tp* Csc2 and *Ec* CasC copies 1–5 the β_2–β_3 insertion included a hairpin, involved in intramolecular and RNA interaction (Fig. 12, CasC-2) [68]. In the first copy of *Ec* CasC however, the very same region folds into a short helix, which is essential for Cas6 interaction (Figure 12, CasC-1) [68]. The possibility to take up two different conformations underlines the flexibility and importance of this sub-type specific element in complex formation. Moreover, a hairpin formed by both the β_1–α_1 and β_2–β_3 segments constitutes the base of the lid domain. This hairpin extends the β-strands β_1 and β_3 of the RRM-like core – a feature that is also present in Cas5 and Cas6 families. However, there it is not part of an insertion, but formed by elongated beta sheets of the core itself [29].

Domain 2

The insertion domain 2 is located between α_1 and β_2. In all known cases, this domain contains helical elements and two anti-parallel β-strands (with the exception of the all-helical *Mk* Csm3) the topological order of which varies strongly. Together with the core domain this insertion defines a central cleft that harbors solvent-exposed, positively charged residues and a second glycine-rich cluster (Hrle *et al.*, 2014, Fig. 5). In *Tp* Csc2 and *Mk* Csm3, a zinc ion is structurally coordinated. Interestingly, in line with the divergent sequential arrangement of secondary structure elements, the metal coordination varies between subtypes, as residues required for coordination (histidines and cysteines) are not always present in this region. Superpositioning of the available Cas7 protein structures via the RRM showed that all have a different curvature. This can be attributed to the orientation of the insertion 2 domain relative to the core. Structural data from the

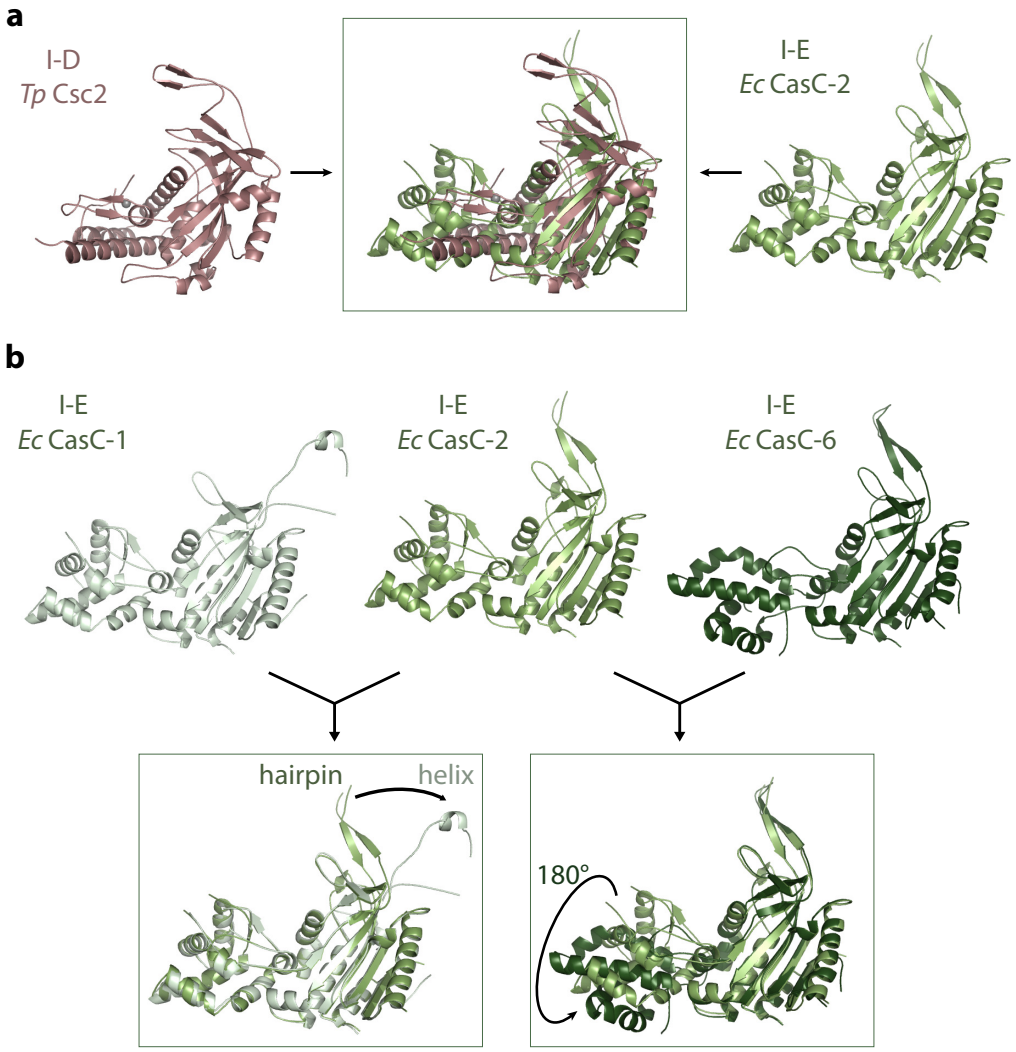


Figure 12: Structural flexibility of Cas7 proteins.
a. Crystal structures of I-D *Tp* Csc2 (purple) and I-E *Ec* CasC copy 2 (green) with resolved hairpin structure in domain 1. Superposition (middle box) highlights structural similarities beyond the RRM-like core.
b. Individual crystal structures of the three observed conformations of *Ec* CasC within the type I-E interference complex. The first copy (*Ec* CasC-1) features a small α-helix at the tip of insertion β_2-β_3 of domain 1. This structural element is rearranged to a β-hairpin in the copies 2 to 5 (displayed on the example of copy 2). Domain 2 of the sixth copy of *Ec* CasC is rotated by 180° compared to all other copies.

entire type I/E interference complex showed that domain 2 of the last copy of *Ec* CasC is rotated by 180° compared to the other copies (Fig. 12, CasC-6), triggered by its interaction with Cas5, the large subunit protein, and the 5′-end of the crRNA [68]. This indicates that there is an inherent structural flexibility of the junction between the core and insertion domain 2.

Domain 3 (C-terminal extension)

The third domain follows after β_4, packs against the RRM and (with exception of *Tp* Csc2), contributes a fifth β-strand to the β-sheet of the core domain. In *Tp* Csc2, three helices constitute the domain and the fourth C-terminal helix, re-joins with insertion domain 2. In Cmr4 this domain is defined by elements of the β_2-β_3 insertion [69].

Surface charge and conservation

Sequence diversity is the basis of all subtype-specific variation in topology and structural composition within the peripheral domains. Despite low sequence homology within and between subtypes, both positive surface charge and solvent-exposed conserved residues are located within the central groove and the lid domain. Both surfaces mediate RNA binding in all known structures [68, 84, 85].

3.1.2 Cas7-like proteins of type III-B

Additional valuable information on the Cas7 proteins has became available with the structures of proteins belonging to the RNA-targeting Cmr complex of *Pyrococcus furiosus (Pf)* [69]. The proteins Cmr4, Cmr6 and Cmr1 have been classified as Cas7-like proteins by Makarova *et al.* [30]. Previous EM studies of the same complex had shown that four copies of Cmr4, as well as one copy each of Cmr6 and Cmr1 assemble into a backbone-like superhelical arrangement (Fig. 13) [66, 67]. Biochemical analysis showed a distinct cleavage pattern that could just be explained by a catalytic activity that is part of Cmr4 itself [67, 69].

The comparison of the structures presented in this thesis with the above-mentioned individual structures of Cmr proteins from *Pyrococcus furiosus* beautifully shows how the peripheral domains of Cas7 have adapted to the requirements of the system.

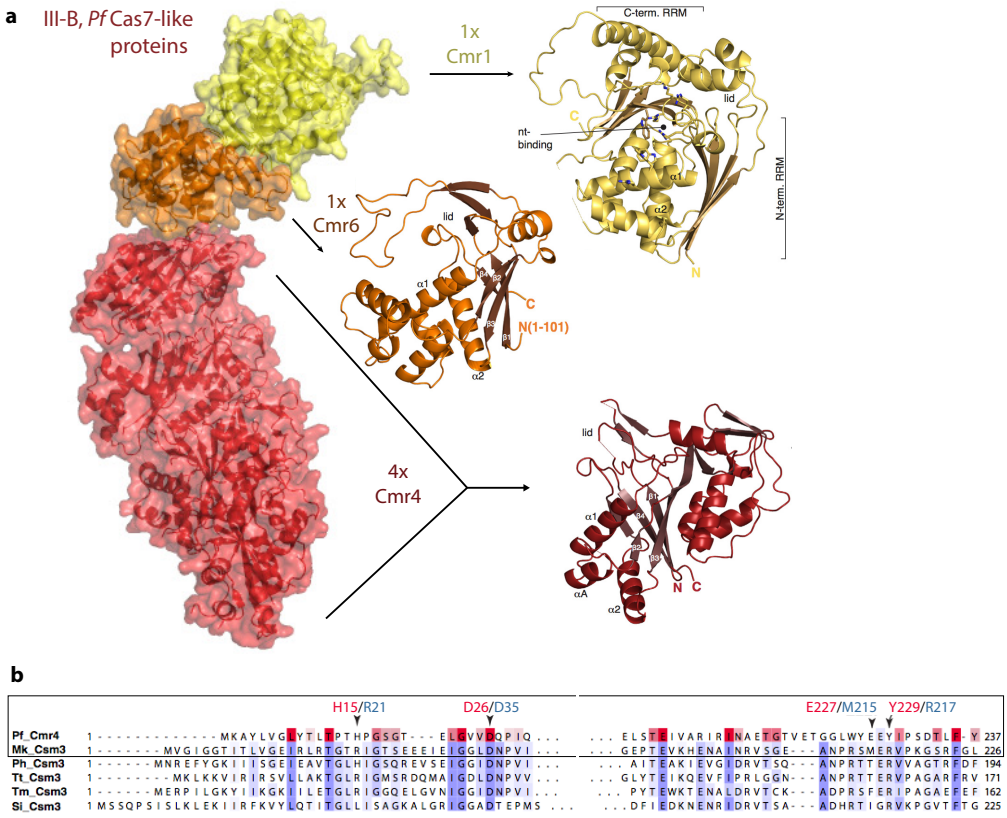


Figure 13: Cas7-like proteins of the Cmr4 complex of *Pyrococcus furiosus*.
a. Cas7-like proteins of the Cmr complex, assemble into a superhelical arrangement. Four copies of Cmr4 are followed by one copy of Cmr6 and one copy of Cmr1. Cmr4 is a unique Cas7-like protein with differently pronounced insertion domains and an active site within the domain 1 (lid). Cmr6 has similarities to both Cmr1 and Cmr4. Cmr1 harbors a second RRM after β_4. Similarly to Cas6 it is also positioned at the 3′-end of the crRNA and caps the complex.
b. Sequence alignment of Cmr4 and representatives of the Csm3 family. Both proteins have been shown to excite catalytic activity towards RNA targets. The residues that constitute the active site are conserved and highlighted (red: *Pf* Cmr4; blue: *Mk* Csm3).

Cmr4

Cmr4 is a Cas7-like protein with an RRM-like core featuring similar peripheral domain arrangements [69]. Four main features make Cmr4 unique (Fig. 13 a, Cmr4 in red). First, insertion 2 is reduced to a single helix between β_1 and α_1, which is a stark contrast to other described Cas7 proteins where this insertion is more pronounced and defines the nucleotide binding cleft. The modification of this structural element probably reflects the different nature of the substrate RNA that is bound. Second is the lid domain, which topologically formed by the same insertions. Here the insertion between β_2 and β_3 is very long (150 residues). It contributes a β-hairpin to the lid and also folds into a second elaborate domain of its own. Here helical elements pack against the front surface of the RRM β-sheet and a β-barrel flanks one edge of the RRM β-sheet, extending it after β_2. Compared to the other Cas7 proteins, this takes up the same position as the C-terminal extension domain. Third there is an additional insertion in β_3-β_4, forming a fifth blade that additionally extends the β-sheet. Fourth and finally, the most striking feature of the Cmr4 protein is that it is the first Cas7-like protein found to exhibit a catalytic activity in an *in vitro* reconstituted complex, generating a characteristic cleavage pattern of its RNA target [69]. The activity is located within the lid domain and residues from different unstructured elements come together to form an active site that is topologically distinct from nuclease sites in other family members. The same catalytic residues are conserved in the type III-A homologue Csm3 (Fig. 13 b), one of which corresponds to the arginine we found to be essential for RNA binding, drawing a functional connection from binding to catalysis (Hrle *et al.*, 2013, Fig. 4D). In our hands, the single *Mk* Csm3 did not show catalytic activity towards the bound RNA substrate. In context of the complete complex however, this may change. Recently, V. Siksnys reported to have measured an affinity towards an RNA target, expanding the function of the complex from targeting DNA to RNA (V. Siksnys, personal communication). This could provide an explanation for why it structurally resembles the type I DNA targeting complexes but possesses the catalytic site for RNA degradation.

Cmr6 and Cmr1

From 5′ to 3′ crRNA direction, Cmr6 and Cmr1 extend the superhelical arrangement of sequential Cmr4 RRM domains [69]. Cmr1 is a unique Cas7 family protein, as it has two RRM domains; just as in Cas6, the second RRM is located after β_4 (Fig. 13 a, Cmr1 in yellow). Two RRM domains of Cmr1 form a single structural unit by interacting

closely with each other via their insertion domains elements. This interaction creates a composite surface groove that is lined by conserved basic and hydrophobic residues. Mutation analysis and structural data suggest that Cmr1 employs this conserved cleft for nucleotide binding and of the 3′-end of the crRNA [69]. Despite lack of biochemical data and the fact that they belong to different families, one can speculate that Cmr1 may share functional analogies to Cas6 in crRNA processing. This is supported by its location and nucleotide binding of the 3′-end of the crRNA and the presence of two RRM domains. The second protein Cmr6 has an overall architecture that resembles the N-terminal RRM of Cmr1, with exception of the lid domain, which superposes well with Cmr4 (Fig. 13 a, Cmr6 in orange). The fully resolved Cmr4 β_2–β_3 hairpin connection hints towards what the currently unresolved equivalent insertion of Cmr4 would look like.

3.2 Form follows function: Cas7 proteins in the interference complex

Initially, our knowledge on Cas7 assembly and crRNA interaction was based on the interpretation of electron microscopy structures of type I and III interference complexes. These studies have confirmed that Csm3 is present in multiple copies in the type III-A Csm complex [65]. For *Tt* Cas7, we have shown its tendency to oligomerize around crRNA by EM, similar to the behavior of its *Sulfolobus sulfataricus* ortholog [63]. In all cases, multiple copies of Cas7 proteins were wrapped around the crRNA. They constitute the helical backbone of the interference complex and contribute to its elongated shape. Initial cryo-EM structures of crRNA and the target-bound type I-E interference complex gave a first higher resolution picture [51]. Now, the recently solved structure of the corresponding interference complex (containing a total of 11 proteins and a 61 nucleotide crRNA) allows a detailed view and understanding of the versatile protein interaction and RNA binding sites [68]. It shows that the endonuclease Cas6e, which is bound to the 3′ stem loop of the mature crRNA recruits the first copy of Cas7 via a hydrophobic cleft, the binding site for the short helix within the insertion β_2–β_3 of the lid domain of Cas7. In total, the six copies of Cas7 (*Ec* CasC) tightly intertwine with the crRNA and oligomerize, forming a helical backbone (Fig. 14 a, left panel). The structure shows that Cas7 has two distinct RNA contact sites, both of which are sequence-unspecific. Each Cas7 binds 5 nucleotides via a central cleft. A hairpin within the insertion β_2–β_3 of the next Cas7 copy folds over the top of the RNA and contacts an accessible positively charged binding

site within the cleft. This dents every sixth base and pins the RNA into the channel in a pseudo A configuration (Fig. 14 a, inset). The last Cas7-RNA interaction is locked by a similar hairpin structure of the terminal Cas5 protein. This positions and divides the RNA into six segments.

3.2.1 RNA binding properties of Type I-A/D and III-A Cas7 homologs

Our data on the Cas7–RNA interaction obtained from the functional analysis of three Cas7 homologs, *Mk* Csm3, *Tp* Csc2 and *Tt* Cas7, agree with the findings of type I-E Cas7 in context of its Cascade complex discussed above. First we showed that all proteins bind their natural substrate, the specific crRNA. In the case of *Tt* Csa2, we visualized the oligomerization along an RNA template by electron microscopy (Plagens *et al.*, 2014, Fig. S6). Oligomerization of *Tt* Csa2 also occurred with nucleic acid contaminants. The binding of Cas7 to an unspecific sequence did not come entirely unexpected, as it was assumed to contact mostly the variable spacer region. We demonstrated the RNA sequence-unspecific binding nature of Cas7 proteins by showing that *Mk* Csm3 prefers an unspecific, unstructured A_{20} and U_{20} substrate over a specific, structured repeat sequence. Later, we found A_{20} and U_{20} single-stranded RNA to bind to all three Cas7 proteins we studied, however with different affinities: Type I-A and I-D homologs bound with a higher affinity (μM) compared to the type III-A Csm3. This behavior can be due to the fact that additional subunits or Cas7-like proteins are needed to stabilize the interaction in the type III-A complex. In case of the type III-B homolog Cmr4, binding of RNA only occurs in context of complex assembly (C. Benda, personal communication). A similar yet less pronounced effect could account for the lower affinity of Csm3.

3.2.2 RNA binding surface of Type I-A/D and III-A Cas7 homologs

For all following experiments, we chose U_{20} as a single-stranded unstructured RNA sequence as it is a suitable substrate for protein-RNA cross-linking. By a combination of UV crosslinking and mutation analysis, we pinpointed two RNA interacting surfaces formed by residues within the central cleft and residues on the lid domain (Figure 14 b, inset). These findings turned out to be in line with the contact sites identified in the type I-E complex (Fig. 14 a, inset).

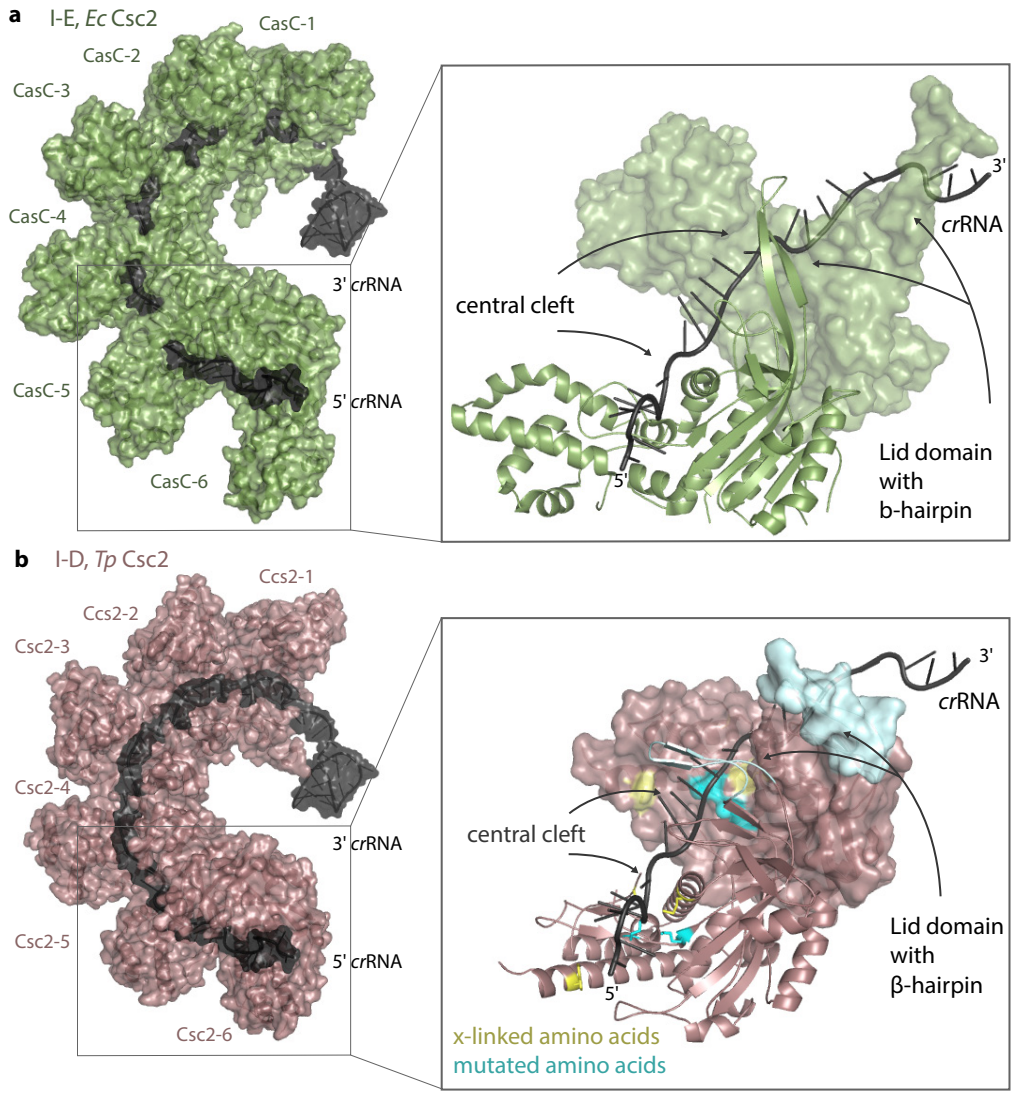


Figure 14: Cas7 backbone of interferences complexes.
a. Individual copies of the Cas7 protein *Ec* CasC are labeled 1 to 6 from 3′–5′ crRNA (black). Inset: Copies 5 (surface representation) and 6 (cartoon representation) and the first 22 nucleotides 3–5′ of the crRNA are shown. The RNA binds to *Ec* CasC via a central cleft and is pinned to this site by the insertion between β_2 and β_3, of the next copy. This insertion contains a hairpin as a structural element. **b.** The superposition of *Tp* Csc2 to *Ec* CasC 1–6 gives an initial picture of the type I-D backbone and underlines the conserved superhelical assembly of Cas7 proteins throughout the subtypes. The copies of *Tp* Csc2 are labeled 1 to 6 from 3′–5′ crRNA (black). Inset: Copies 5 (surface representation) and 6 (cartoon representation) and the first 22 nucleotides 3–5′ of the crRNA are shown in the same orientation as above for *Ec* CasC. Amino acids that we showed to interact with RNA (crosslinked: yellow; mutated: cyan) are in accordance with the RNA-binding interface of the type I-E complex [68].

For type III-A, we performed a mutational analysis, which demonstrated the involvement of a conserved arginine (R21) in the lid domain in RNA binding. Interestingly, mutations of conserved exposed residues along the α_1 helix of the core or a loop-out mutation of the insertion β_2–β_3 in the lid domain did not significantly perturb the binding (Hrle *et al.*, 2013, Figs. 4D, S4C). This can be due to the synergistic effect of the various binding sites: A combination of mutations within the two different binding sites might be necessary to show an effect. Aided by UV crosslinking we identified a significant number of contacts within the central positively charged cleft of *Tt* Csa2 and *Tp* Csc2 and identified a conserved consensus motif containing arginine and lysine residues. In the case of the *E. coli* Cas7, these residues are replaced by methionines that significantly contribute to the direct interaction with the phosphate backbone and the bases [68]. We observed that a minimal length of RNA is required for Cas7 interaction. This finding is in line with the model that Cas7 proteins work together to stabilize binding in a cooperative manner. Around 15 nucleotides are needed in order for one copy of Cas7 to stabilize the interaction of the subsequent one.

Next to the Cas6 family, the Cas7 family now belongs to the best understood amongst the Cas proteins from a structural and functional perspective. The proteins provide a comprehensive platform for protein-protein and protein-nucleic acid interactions. They interact with every single other subunit of the interference complex from the 3′ to 5′ of the crRNA [68, 87]. Moreover, the Cas7 homolog Cmr4 of the RNA-targeting Cmr complexes has evolved a catalytic site for RNA cleavage. Future structural studies of target-bound interference complexes will shed light on the structural rearrangements that occur within Cas7 proteins upon target recognition. Initial EM data suggest that the insertion domain between α_1 and β_2 of the last two Cas7 copies (closest to the 5′ crRNA end) undergoes a rotation and enables a distal lysine rich helix to contact the backbone of the target DNA [51, 68]. Interestingly one crosslink we obtained within *Tp* Csc2 was located in this region.

4 Outlook

The increasing availability of sequencing data of prokaryotic and phage genomes combined with comprehensive bioinformatics efforts has shed light on the expansive protein machinery that administrates the CRISPR immune response. Based the organization of the *cas* genes within the locus as well as secondary structure predictions of the respective proteins, these studies have classified the myriads of different Cas proteins [30]. The increasing number of crystal structures of individual Cas proteins has provided a wealth of information on the molecular mechanisms, such as for the assembly of Cas7 family proteins into the interference complex, the Cas6 mediated crRNA processing and Cas3-dependent target cleavage in type I systems [29]. In addition, a great challenge has been to obtain crystal structures of the entire interference complexes. However even though the novel crystal structures allow new insights into the diverse protein machinery and its underlying mechanisms, several aspects remain elusive and are subject to ongoing research. One point of interest will be to address the mechanisms of target DNA binding in a thermodynamically feasible fashion, given that the process is ATP-independent [9]. Jackson *et al.* suggest that the pre-ordering of crRNA-guide into a pseudo-A helical arrangement by the Cas7 subunits enhances target detection and may reduce the entropic penalty in order to provide a thermodynamic advantage for target binding. Structural comparisons with Cas9 and Argonaute proteins reveal that these proteins apply a similar mechanism: by kinking the RNA helix, the protein mediates pre-ordering of the RNA guide. In type I interference complexes, the Cas7 insertion that introduces the kink also covers this base and makes it inaccessible for the target DNA [68]. How full target hybridization is achieved remains elusive. One can speculate that a structural rearrangement, which has been shown by cryo EM of the type I-E, takes place to expose the RNA guide. In case of Cas9, the crystal structure of the target-bound complex suggests that the RNA–DNA hybrid forms a contiguous A-form duplex [61]. Another open question is the nucleic acid specificity of type III Csm and Cmr complexes towards RNA. Classically Csm3 is known to target DNA, whereas Cmr targets RNA [33, 47]. However, recent studies reported that Csm3 is capable of targeting RNA *in vitro* using the same catalytic site. These findings need to be validated *in vivo*, as they have important implications for archaeal biology: Is RNA targeting solely a response to RNA viruses, or is it used as a means of post-transcriptional regulation in these organisms?

Beyond the achievements of basic research, CRISPR has uncovered a powerful tool for laboratory and clinical applications alike: Cas9, a single nucleic acid-guided protein, which can perform a plethora of genome editing functions. Two discoveries have broadened its potential for genome engineering. First the discovery that one can disable one or both nuclease activities without interfering with target recognition [57]. Second the successful fusion of Cas9's two guide RNAs (tracrRNA and crRNA) to a single small guide RNA (sgRNA) [57]. These findings allowed the creation of programmable synthetic machinery, which unifies all three types of sequence-defined biological polymers. Proteins, RNA or DNA can be targeted to any dsDNA sequence by simply fusing them to either Cas9 or the sgRNA. Applications range from targeted genome editing, targeted genome regulation, programmable genome reorganization and visualization [81]. Thus the system holds tremendous potential for studying and engineering living systems.

CRISPR systems have stormed onto the scene. The speed at which they can be adapted to new targets, their high efficiency and versatility makes them poised to overtake more established technologies based on TALENs or zinc finger nucleases, both of which depend on engineering custom-made proteins for each DNA target [88, 89]. Within a year after their introduction, libraries of small guide RNAs have been synthesized that can be used to target 90 % of all genes in humans and model organisms [81]. Highlights of the application of CRISPR technologies include the generation of triple knockout mice in 'single shot' [90] and the excision of HIV from infected human cells [91]. Specificity and off-target effects still limit the applicability of CRISPR/Cas technology, in particular for the treatment of human diseases by gene therapy [92]. However, this also provides a strong driving force for basic research. For instance, biochemical analysis and crystal structures of Cas9 provide mechanistic details of target recognition and open up possibilities for engineering Cas9 towards higher target specificity [83, 93].

In summary, the emerging CRISPR/Cas fields provides ample opportunities for basic scientists trying to understand evolutionary relationships as well as molecular engineers aiming at benefiting human health.

References

[1] Clokie, M. R., Millard, A. D., Letarov, A. V., & Heaphy, S. Phages in nature. *Bacteriophage* **1**(1), 31–45 1 (2011). (↑ p. 1)

[2] Weinbauer, M. G. Ecology of prokaryotic viruses. *FEMS Microbiol Rev* **28**(2), 127–181 May (2004). (↑ p. 1)

[3] Stern, A. & Sorek, R. The phage-host arms race: shaping the evolution of microbes. *Bioessays* **33**(1), 43–51 Jan (2011). (↑ p. 1)

[4] Duckworth, D. History and basic properties of bacterial viruses. *Phage Ecology* , 1–44 (1987). (↑ p. 1)

[5] Bikard, D. & Marraffini, L. A. Innate and adaptive immunity in bacteria: mechanisms of programmed genetic variation to fight bacteriophages. *Curr Opin Immunol* **24**(1), 15–20 Feb (2012). (↑ p. 1)

[6] Heller, K. J. Molecular interaction between bacteriophage and the gram-negative cell envelope. *Arch Microbiol* **158**(4), 235–248 (1992). (↑ p. 1)

[7] Roberts, R. J., Belfort, M., Bestor, T., *et al.* A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res* **31**(7), 1805–1812 Apr (2003). (↑ p. 1)

[8] Molineux, I. J. Host-parasite interactions: recent developments in the genetics of abortive phage infections. *New Biol* **3**(3), 230–236 Mar (1991). (↑ p. 1)

[9] Westra, E. R., Swarts, D. C., Staals, R. H. J., *et al.* The CRISPRs, they are a-changin': how prokaryotes generate adaptive immunity. *Annu Rev Genet* **46**, 311–339 (2012). (↑ pp. 2 and 95)

[10] van der Oost, J., Westra, E. R., Jackson, R. N., & Wiedenheft, B. Unravelling the structural and mechanistic basis of CRISPR-Cas systems. *Nat Rev Microbiol* **12**(7), 479–492 Jul (2014). (↑ pp. 2, 3, 7, 8, 10, 15, and 17)

[11] Hall, A. R., Scanlan, P. D., Morgan, A. D., & Buckling, A. Host-parasite coevolutionary arms races give way to fluctuating selection. *Ecol Lett* **14**(7), 635–642 Jul (2011). (↑ p. 1)

[12] Makarova, K. S., Grishin, N. V., Shabalina, S. A., Wolf, Y. I., & Koonin, E. V. 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 (2006). (↑ pp. 1 and 18)

[13] Barrangou, R., Fremaux, C., Deveau, H., *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**(5819), 1709–1712 Mar (2007). (↑ pp. 1 and 9)

[14] Jansen, R., Embden, J. D. A. v., Gaastra, W., & Schouls, L. M. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* **43**(6), 1565–1575 Mar (2002). (↑ pp. 3 and 4)

[15] Grissa, I., Vergnaud, G., & Pourcel, C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8**, 172 (2007). (↑ p. 3)

[16] Hankin, E. L'action bactericide des eaux de la Jumna et du Gange sur le vibrion du cholera. *Ann. Inst. Pasteur* **10**, 511 (1896). (↑ p. 3)

[17] Mojica, F. J., Díez-Villaseñor, C., Soria, E., & Juez, G. Biological significance of a family of

regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Mol Microbiol* **36**(1), 244–246 Apr (2000). (↑ p. 4)

[18] Grissa, I., Vergnaud, G., & Pourcel, C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* **35**(Web Server issue), W52–W57 Jul (2007). (↑ p. 4)

[19] Kunin, V., Sorek, R., & Hugenholtz, P. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol* **8**(4), R61 (2007). (↑ p. 4)

[20] Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J., & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* **60**(2), 174–182 Feb (2005). (↑ p. 4)

[21] Bolotin, A., Quinquis, B., Sorokin, A., & Ehrlich, S. D. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**(Pt 8), 2551–2561 Aug (2005). (↑ p. 4)

[22] Hale, C. R., Majumdar, S., Elmore, J., *et al.* Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. *Mol Cell* **45**(3), 292–302 Feb (2012). (↑ pp. 4, 12, 18, and 19)

[23] Pougach, K., Semenova, E., Bogdanova, E., *et al.* Transcription, processing and function of CRISPR cassettes in Escherichia coli. *Mol Microbiol* **77**(6), 1367–1379 Sep (2010). (↑ p. 4)

[24] Makarova, K. S., Aravind, L., Grishin, N. V., Rogozin, I. B., & Koonin, E. V. A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res* **30**(2), 482–496 Jan (2002). (↑ pp. 4 and 6)

[25] Godde, J. S. & Bickerton, A. The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. *J Mol Evol* **62**(6), 718–729 Jun (2006). (↑ p. 4)

[26] Haft, D. H., Selengut, J., Mongodin, E. F., & Nelson, K. E. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput Biol* **1**(6), e60 Nov (2005). (↑ p. 4)

[27] Wang, R., Zheng, H., Preamplume, G., Shao, Y., & Li, H. The impact of CRISPR repeat sequence on structures of a Cas6 protein-RNA complex. *Protein Sci* **21**(3), 405–417 Mar (2012). (↑ pp. 6 and 84)

[28] Maris, C., Dominguez, C., & Allain, F. H.-T. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS J* **272**(9), 2118–2131 May (2005). (↑ p. 6)

[29] Reeks, J., Naismith, J. H., & White, M. F. CRISPR interference: a structural perspective. *Biochem J* **453**(2), 155–166 Jul (2013). (↑ pp. 6, 11, 12, 14, 84, 86, and 95)

[30] Makarova, K. S., Haft, D. H., Barrangou, R., *et al.* Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol* **9**(6), 467–477 Jun (2011). (↑ pp. 6, 13, 14, 17, 83, 84, 88, and 95)

[31] Deveau, H., Barrangou, R., Garneau, J. E., *et al.* Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *J Bacteriol* **190**(4), 1390–1400 Feb (2008). (↑ p. 9)

[32] Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J., & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**(Pt 3), 733–740 Mar (2009). (↑ p. 9)

[33] Marraffini, L. A. & Sontheimer, E. J. Self versus non-self discrimination during CRISPR RNA-

directed immunity. *Nature* **463**(7280), 568–571 Jan (2010). (↑ pp. 9, 12, 16, 18, and 95)

[34] Swarts, D. C., Mosterd, C., van Passel, M. W. J., & Brouns, S. J. J. CRISPR interference directs strand specific spacer acquisition. *PLoS One* **7**(4), e35888 (2012). (↑ p. 9)

[35] Yosef, I., Goren, M. G., & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic Acids Res* **40**(12), 5569–5576 Jul (2012). (↑ pp. 9 and 10)

[36] Datsenko, K. A., Pougach, K., Tikhonov, A., *et al.* Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat Commun* **3**, 945 (2012). (↑ p. 9)

[37] Stern, A., Keren, L., Wurtzel, O., Amitai, G., & Sorek, R. Self-targeting by CRISPR: gene regulation or autoimmunity? *Trends Genet* **26**(8), 335–340 Aug (2010). (↑ pp. 9 and 13)

[38] Vercoe, R. B., Chang, J. T., Dy, R. L., *et al.* Cytotoxic chromosomal targeting by CRISPR/Cas systems can reshape bacterial genomes and expel or remodel pathogenicity islands. *PLoS Genet* **9**(4), e1003454 Apr (2013). (↑ p. 9)

[39] Plagens, A., Tjaden, B., Hagemann, A., Randau, L., & Hensel, R. Characterization of the CRISPR/Cas subtype I-A system of the hyperthermophilic crenarchaeon Thermoproteus tenax. *J Bacteriol* **194**(10), 2491–2500 May (2012). (↑ pp. 10 and 11)

[40] Nuñez, J. K., Kranzusch, P. J., Noeske, J., *et al.* Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nat Struct Mol Biol* **21**(6), 528–534 Jun (2014). (↑ p. 10)

[41] Han, D., Lehmann, K., & Krauss, G. SSO1450–a CAS1 protein from Sulfolobus solfataricus P2 with high affinity for RNA and DNA. *FEBS Lett* **583**(12), 1928–1932 Jun (2009).

[42] Kim, T.-Y., Shin, M., Huynh Thi Yen, L., & Kim, J.-S. Crystal structure of Cas1 from Archaeoglobus fulgidus and characterization of its nucleolytic activity. *Biochem Biophys Res Commun* **441**(4), 720–725 Nov (2013).

[43] Samai, P., Smith, P., & Shuman, S. Structure of a CRISPR-associated protein Cas2 from Desulfovibrio vulgaris. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **66**(Pt 12), 1552–1556 Dec (2010).

[44] Nam, K. H., Ding, F., Haitjema, C., *et al.* Double-stranded endonuclease activity in Bacillus halodurans clustered regularly interspaced short palindromic repeats (CRISPR)-associated Cas2 protein. *J Biol Chem* **287**(43), 35943–35952 Oct (2012). (↑ p. 10)

[45] van der Oost, J., Jore, M. M., Westra, E. R., Lundgren, M., & Brouns, S. J. J. CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends Biochem Sci* **34**(8), 401–407 Aug (2009). (↑ p. 11)

[46] Babu, M., Beloglazova, N., Flick, R., *et al.* A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. *Mol Microbiol* **79**(2), 484–502 Jan (2011). (↑ pp. 11 and 18)

[47] Hale, C., Kleppe, K., Terns, R. M., & Terns, M. P. Prokaryotic silencing (psi)RNAs in Pyrococcus furiosus. *RNA* **14**(12), 2572–2579 Dec (2008). (↑ pp. 11 and 95)

[48] Niewoehner, O., Jinek, M., & Doudna, J. A. Evolution of CRISPR RNA recognition and processing by Cas6 endonucleases. *Nucleic Acids Res* **42**(2), 1341–1353 Jan (2014). (↑ pp. 11 and 12)

[49] Carte, J., Wang, R., Li, H., Terns, R. M., & Terns, M. P. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in

prokaryotes. *Genes Dev* **22**(24), 3489–3496 Dec (2008). (↑ p. 12)

[50] Haurwitz, R. E., Jinek, M., Wiedenheft, B., Zhou, K., & Doudna, J. A. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science* **329**(5997), 1355–1358 Sep (2010). (↑ p. 12)

[51] Wiedenheft, B., Lander, G. C., Zhou, K., *et al.* Structures of the RNA-guided surveillance complex from a bacterial immune system. *Nature* **477**(7365), 486–489 Sep (2011). (↑ pp. 12, 14, 16, 91, and 94)

[52] Garside, E. L., Schellenberg, M. J., Gesner, E. M., *et al.* Cas5d processes pre-crRNA and is a member of a larger family of CRISPR RNA endonucleases. *RNA* **18**(11), 2020–2028 Nov (2012).

[53] Nam, K. H., Haitjema, C., Liu, X., *et al.* Cas5d protein processes pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvulg CRISPR-Cas system. *Structure* **20**(9), 1574–1584 Sep (2012). (↑ p. 12)

[54] Hale, C. R., Zhao, P., Olson, S., *et al.* RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* **139**(5), 945–956 Nov (2009). (↑ p. 12)

[55] Hatoum-Aslan, A., Maniv, I., & Marraffini, L. A. 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**(52), 21218–21222 Dec (2011). (↑ pp. 12 and 17)

[56] Deltcheva, E., Chylinski, K., Sharma, C. M., *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**(7340), 602–607 Mar (2011). (↑ p. 13)

[57] Jinek, M., Chylinski, K., Fonfara, I., *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**(6096), 816–821 Aug (2012). (↑ pp. 13, 18, 19, and 96)

[58] Deveau, H., Garneau, J. E., & Moineau, S. CRISPR/Cas system and its role in phage-bacteria interactions. *Annu Rev Microbiol* **64**, 475–493 (2010). (↑ p. 13)

[59] Garneau, J. E., Dupuis, M.-È., Villion, M., *et al.* The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**(7320), 67–71 Nov (2010). (↑ p. 13)

[60] Jinek, M., Jiang, F., Taylor, D. W., *et al.* Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* **343**(6176), 1247997 Mar (2014). (↑ pp. 13 and 83)

[61] Nishimasu, H., Ran, F. A., Hsu, P. D., *et al.* Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**(5), 935–949 Feb (2014). (↑ pp. 13, 83, and 95)

[62] Jore, M. M., Lundgren, M., van Duijn, E., *et al.* Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nat Struct Mol Biol* **18**(5), 529–536 May (2011). (↑ p. 14)

[63] Lintner, N. G., Kerou, M., Brumfield, S. K., *et al.* Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE). *J Biol Chem* **286**(24), 21643–21656 Jun (2011). (↑ pp. 14, 49, 83, and 91)

[64] van Duijn, E., Barbu, I. M., Barendregt, A., *et al.* 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**(11), 1430–1441

Nov (2012).

[65] Rouillon, C., Zhou, M., Zhang, J., *et al.* Structure of the CRISPR interference complex CSM reveals key similarities with cascade. *Mol Cell* **52**(1), 124–134 Oct (2013). (↑ pp. 14, 83, and 91)

[66] Spilman, M., Cocozaki, A., Hale, C., *et al.* Structure of an RNA silencing complex of the CRISPR-Cas immune system. *Mol Cell* **52**(1), 146–152 Oct (2013). (↑ p. 88)

[67] Staals, R. H. J., Agari, Y., Maki-Yonekura, S., *et al.* Structure and activity of the RNA-targeting Type III-B CRISPR-Cas complex of Thermus thermophilus. *Mol Cell* **52**(1), 135–145 Oct (2013). (↑ pp. 14 and 88)

[68] Jackson, R. N., Golden, S. M., van Erp, P. B. G., *et al.* Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli. *Science* Aug (2014). (↑ pp. 14, 16, 80, 83, 86, 88, 91, 93, 94, and 95)

[69] Benda, C., Ebert, J., Scheltema, R. A., *et al.* Structural model of a CRISPR RNA-silencing complex reveals the RNA-target cleavage activity in Cmr4. *Mol Cell* **in press** (2014). (↑ pp. 14, 15, 18, 49, 83, 85, 88, 90, and 91)

[70] Makarova, K. S., Aravind, L., Wolf, Y. I., & Koonin, E. V. Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biol Direct* **6**, 38 (2011). (↑ pp. 14, 84, and 85)

[71] Almendros, C., Guzmán, N. M., Díez-Villaseñor, C., García-Martínez, J., & Mojica, F. J. M. Target motifs affecting natural immunity by a constitutive CRISPR-Cas system in Escherichia coli. *PLoS One* **7**(11), e50797 (2012). (↑ p. 16)

[72] Sashital, D. G., Wiedenheft, B., & Doudna, J. A. Mechanism of foreign DNA selection in a bacterial adaptive immune system. *Mol Cell* **46**(5), 606–615 Jun (2012). (↑ p. 16)

[73] Semenova, E., Jore, M. M., Datsenko, K. A., *et al.* Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc Natl Acad Sci U S A* **108**(25), 10098–10103 Jun (2011). (↑ p. 16)

[74] Westra, E. R., van Erp, P. B. G., Künne, T., *et al.* CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. *Mol Cell* **46**(5), 595–605 Jun (2012). (↑ p. 16)

[75] Chen, Z., Yang, H., & Pavletich, N. P. Mechanism of homologous recombination from the RecA–ssDNA/dsDNA structures. *Nature* **453**(7194), 489–494 (2008). (↑ p. 16)

[76] Hochstrasser, M. L., Taylor, D. W., Bhat, P., *et al.* CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. *Proc Natl Acad Sci U S A* **111**(18), 6618–6623 May (2014). (↑ p. 16)

[77] Sampson, T. R., Saroj, S. D., Llewellyn, A. C., Tzeng, Y.-L., & Weiss, D. S. A CRISPR/Cas system mediates bacterial innate immune evasion and virulence. *Nature* **497**(7448), 254–257 May (2013). (↑ pp. 18 and 19)

[78] Wiedenheft, B., Sternberg, S. H., & Doudna, J. A. RNA-guided genetic silencing systems in bacteria and archaea. *Nature* **482**(7385), 331–338 Feb (2012). (↑ p. 19)

[79] Farazi, T. A., Juranek, S. A., & Tuschl, T. The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members. *Development* **135**(7), 1201–1214 Apr (2008). (↑ p. 19)

[80] Jinek, M. & Doudna, J. A. A three-dimensional view of the molecular machinery of RNA interference. *Nature* **457**(7228), 405–412 Jan (2009). (↑ p. 19)

[81] Mali, P., Esvelt, K. M., & Church, G. M. Cas9 as a versatile tool for engineering biology. *Nat Methods* **10**(10), 957–963 Oct (2013). (↑ pp. 19 and 96)

[82] Mali, P., Yang, L., Esvelt, K. M., *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**(6121), 823–826 Feb (2013). (↑ p. 19)

[83] Anders, C., Niewoehner, O., Duerst, A., & Jinek, M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* Jul (2014). (↑ pp. 83 and 96)

[84] Hrle, A., Su, A. A. H., Ebert, J., *et al.* Structure and RNA-binding properties of the type III-A CRISPR-associated protein Csm3. *RNA Biol* **10**(11), 1670–1678 Nov (2013). (↑ pp. 83 and 88)

[85] Hrle, A., Maier, L.-K., Sharma, K., *et al.* Structural analyses of the CRISPR protein Csc2 reveal the RNA-binding interface of the type I-D Cas7 family. *RNA Biol* **11**(8) Aug (2014). (↑ pp. 83 and 88)

[86] Sun, J., Jeon, J.-H., Shin, M., *et al.* Crystal structure and CRISPR RNA-binding site of the Cmr1 subunit of the Cmr interference complex. *Acta Crystallogr D Biol Crystallogr* **70**(Pt 2), 535–543 Feb (2014). (↑ p. 85)

[87] Plagens, A., Tripp, V., Daume, M., *et al.* In vitro assembly and activity of an archaeal CRISPR-Cas type I-A Cascade interference complex. *Nucleic Acids Res* **42**(8), 5125–5138 Apr (2014). (↑ p. 94)

[88] Rebar, E. J. & Pabo, C. O. Zinc finger phage: affinity selection of fingers with new DNA-binding specificities. *Science* **263**(5147), 671–673 Feb (1994). (↑ p. 96)

[89] Sanjana, N. E., Cong, L., Zhou, Y., *et al.* A transcription activator-like effector toolbox for genome engineering. *Nat Protoc* **7**(1), 171–192 Jan (2012). (↑ p. 96)

[90] Wang, H., Yang, H., Shivalila, C. S., *et al.* One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153**(4), 910–918 May (2013). (↑ p. 96)

[91] Hu, W., Kaminski, R., Yang, F., *et al.* RNA-directed gene editing specifically eradicates latent and prevents new HIV-1 infection. *Proc Natl Acad Sci U S A* **111**(31), 11461–11466 Aug (2014). (↑ p. 96)

[92] Cho, S. W., Kim, S., Kim, Y., *et al.* Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res* **24**(1), 132–141 Jan (2014). (↑ p. 96)

[93] Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C., & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**(7490), 62–67 Mar (2014). (↑ p. 96)

Acknowledgements

Foremost I would like thank to my supervisor Elena Conti. Thank you for letting me contribute to the department scientifically and personally for the past years. Unconventionally and luckily I could dive into a vast range of biological fields and I appreciated this rare opportunity in structural biology. Alongside Elena I would like to thank Christian Benda for being my partner in crime and professionally and personally mentoring my graduate studies. My appreciation goes to all lab members, collaborators who shared their interests and entrusted me with their projects, especially Anita Marchfelder, Hennig Urlaub and Lennart Randau.

Thanks to Steffen, Walther, Peter, Michaela, Ulrike, Petra, Jérôme, Claire, Karina, Sabine, Ariane, Petra and Tatjana for always giving me a hand in everyday lab life.

My dear lab mates, Basti, Felix, Eva, Humayun, Varun, Rajan, Masami, Ingmar, Sevim, Gretel, Katharina, Marc – thank you for all the memorable dinners, celebrations, discussions, your scientific support and guidance, culinary delights, the good hearty laughs, and much much more – you are the good soul of my graduate studies.

My dearest Sutapa and Ben, I am more than grateful that our paths crossed and will continue to do so in the future. You have scientifically enlightened and personally supported me, and got me all stoked on surfing and salted caramel ;) Friends in need are friends indeed. Love you very much. Along this note I would like to thank all of my dear friends who have encouraged my ambition and been patient with me being so committed to my work.

My greatest gratitude and unconditional and eternal love goes to my family. My beloved grandparents, mama, tata, Dini, Marco (and the gremlins ;) I owe it all to you.