Philipps Universität Marburg

Characterization of a type I-B CRISPR-Cas system of *Clostridium thermocellum*

Dissertation

zur

Erlangung des Doktorgrades
der Naturwissenschaften
(Dr. rer. nat.)

Dem Fachbereich Biologie
der Philipps-Universität Marburg
vorgelegt von

Judith Zöphel
aus Hohenstein

Marburg/ Lahn, Juni 2015

Die Untersuchungen der vorliegenden Arbeit wurden von Oktober 2011 bis Juni 2015 unter Betreuung von Herrn Dr. Lennart Randau in Marburg am Max-Planck-Institut für terrestrische Mikrobiologie durchgeführt.

Vom Fachbereich Biologie

der Philipps-Universität Marburg als Dissertation

angenommen am: 14.07.2015

Erstgutachter: Dr. Lennart Randau

Zweitgutachter: Prof. Dr. Torsten Waldminghaus

Tag der mündlichen Prüfung: 23.07.2015

Teile dieser Arbeit wurden in folgenden Artikeln veröffentlicht:

Zoephel J, Dwarakanath S, Richter H, Plagens A, Randau L: *Substrate generation for endonucleases of CRISPR/cas systems*. J Vis Exp. 2012; **8**(67). pii: 4277.

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**(19): 9887-96.

Zoephel J, Randau L: *RNA-Seq analyses reveal CRISPR RNA processing and regulation patterns*. Biochem Soc Trans. 2013; **41**(6): 1459-63.

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich meine Dissertation mit dem Titel: "Characterization of a type I-B CRISPR-Cas system of *Clostridium thermocellum*" selbständig und ohne unerlaubte Hilfsmittel angefertigt und ich keine als die von mir ausdrücklich angegebenen Quellen verwendet habe.

Diese Dissertation wurde in der jetzigen oder ähnlichen Form bei keiner anderen Hochschule eingereicht und hat noch keinem sonstigen Prüfungszwecken gedient.

Marburg, den 10.08.2015

Judith Zöphel

„Das Ziel der Arbeit ist die Muße, die Muße ist die Schwester der Freiheit.“
—Aristoteles

I Summary

CRISPR-Cas systems are adaptive immune systems, found in bacteria and archaea that provide inheritable resistance against mobile genetic elements, e.g. viruses and plasmids. CRISPR-Cas systems comprise one or more CRISPR loci that contain virus-derived DNA sequences (spacers) that are interspaced by identical repeat sequences (repeats), and a set of *cas* genes. The degradation of nucleic acid targets is mediated by ribonucleoprotein (RNP) complexes, formed by Cas proteins, that are guided by small CRISPR RNA molecules (crRNAs). A Cas protein classification has been established which reflects the diversification of CRISPR systems during the co-evolution of phages and their respective hosts. In this study, the type I-B CRISPR-Cas system of the thermophilic bacterium *Clostridium thermocellum* was investigated.

CRISPR loci are transcribed into precursor-crRNAs and individual crRNAs are produced by Cas6 endonucleases. *C. thermocellum* contains two Cas6 proteins and the recombinant enzymes exclusively process their respective precursor transcripts *in vitro*. RNA-Seq analyses confirmed crRNA maturation and highlighted crRNA abundance differences *in vivo*. These analyses identified internal promotion of precursor-crRNA transcription and reverse crRNA transcripts (anti-crRNAs). Anti-crRNAs display a distinct processing pattern and the abundance of the complementary crRNA is often reduced *in vivo*. Cleavage assays with crRNAs and double-stranded crRNA/anti-crRNA hybrids identified RNase III to be capable of anti-crRNA processing. RNase III cleavage is mediated by recognition motifs within the repeat RNA duplexes.

In type I-B systems, CRISPR interference is mediated by a dsDNA targeting crRNP complex, termed Cascade, which consists of the Cas proteins Cas3, Cas5, Cas6, Cas7 and the subtype-specific Cas8b subunit. All five recombinant Cascade subunits were produced in *Escherichia coli*. Cascade assembly studies revealed that a stable core-complex is formed by Cas5, Cas7, Cas8b and crRNA *in vitro*. Cas3 does not assemble with the complex. Cas6 is a temporarily associated subunit. Mass-spectrometric analyses confirmed protein interactions between Cas5, Cas7 and Cas8b and determined an uneven complex stoichiometry of 1:1:6:2.5 for Cas5:Cas6:Cas7:Cas8b. The large subunit Cas8b forms an additional small C-terminal protein fragment that is also observed in *C. thermocellum* cell extracts and assembles with the complex.

This thesis provides details for the *in vitro* assembly of individual Cas proteins into type I-B Cascade. Furthermore, RNA-Seq analyses of the CRISPR arrays highlight the impact of individual spacer and repeat sequences on the functionality of CRISPR-Cas systems.

II Zusammenfassung

CRISPR-Cas Systeme stellen adaptive Abwehrsysteme dar, die in Bakterien und Archaeen gefunden werden und vererbbare Resistenz gegen mobile genetische Elemente wie z.B. Viren und Plasmide vermitteln. CRISPR-Cas Systeme bestehen aus einem oder mehreren CRISPR-Loci und assoziierten *cas* Genen. CRISPR-Loci enthalten virale DNA-Sequenzen (Spacer) die durch identische, repetitive Sequenzen (Repeats) getrennt werden. Cas-Proteine bilden Ribonukleoproteinkomplexe (RNP-Komplexe) mit kleinen CRISPR RNAs (crRNAs), die als Zielerkennungsmoleküle Nukleinsäuren detektieren und zum Verdau markieren. Die Einführung einer Nomenklatur zur Klassifizierung der Cas-Proteine verdeutlicht die Diversität von CRISPR-Cas Systemen, die durch die Co-Evolution von Phage und Wirt angetrieben wird. In der vorliegenden Arbeit wurde ein Typ I-B CRISPR-Cas System des thermophilen Bakteriums *Clostridium thermocellum* untersucht.

CRISPR-Loci werden in Precursor-crRNAs transkribiert und anschließend von Cas6 Endonukleasen zu individuellen crRNAs prozessiert. *C. thermocellum* enthält zwei Cas6 Proteine, die als rekombinante Enzyme *in vitro* ausschließlich die ihnen zugehörigen Precursor-Transkripte schneiden. RNA-Seq-Analysen bestätigten die Produktion von crRNAs und belegten variable Transkriptmengen individueller crRNAs *in vivo*. Es konnten Promotoren innerhalb von CRISPR-Loci beobachtet werden. Des Weiteren wurden entgegengesetzte crRNA-Transkripte (Anti-crRNAs) identifiziert, die ein deutliches Prozessierungsmuster aufweisen. Die Transkriptmenge der komplementären crRNAs ist häufig reduziert. Aktivitätsassays mit doppelsträngigen crRNA/Anti-crRNA-Hybriden zeigten, dass RNase III imstande ist, Anti-crRNAs zu schneiden. Diese Aktivität wird durch Erkennungsmotive in der Sequenz von Repeat-RNA-Hybriden vermittelt.

Typ I-B CRISPR Interferenz wird durch crRNP-Komplexe vermittelt, die doppelsträngige Fremd-DNA erkennen und als Cascade-Komplexe bezeichnet werden. Diese Cascade-Komplexe setzten sich aus den Cas-Proteinen Cas3, Cas5, Cas6, Cas7 und der Subtyp-spezifischen Untereinheit Cas8b zusammen. Alle fünf rekombinanten Cascade-Untereinheiten wurden in *Escherichia coli* hergestellt. *In vitro*-Studien zur Assemblierung von rekombinanten Cascade-Komplexen zeigten, dass Cas5, Cas7, Cas8b und crRNA stabile Untereinheiten sind. Cas3 ist kein fester Bestandteil des Komplexes und Cas6 interagiert nur zeitweise mit anderen Cas-Komponenten. Massenspektrometrische Analysen belegten die Interaktion zwischen Cas5, Cas7 und Cas8b, und ermittelten eine ungerade Komplexstöchiometrie von 1:1:6:2,5 für Cas5:Cas6:Cas7:Cas8b. Die große Untereinheit Cas8b weist ein zusätzliches C-terminales Proteinfragment auf, welches im Zellextrakt von *C. thermocellum* nachgewiesen wurde und ebenfalls mit dem Cascade-Komplex assembliert.

Die Ergebnisse dieser Studie ermöglichen einen Einblick in die Assemblierung der einzelnen Cas-Proteine zu einem Typ I-B Cascade *in vitro*. Des Weiteren verdeutlichen RNA-Seq Analysen der CRISPR-Arrays den Einfluss von individuellen Spacer-und-Repeat-Sequenzen auf die Funktionalität von CRISPR-Cas Systemen.

III Abbreviations

aa	amino acid
APS	ammonium persulfate
ATP	adenosine triphosphate
β-Me	β-mercaptoethanol
bp	basepair(s)
BSA	bovine serum albumin
C-terminal	carboxy-terminal
Cam	chloramphenicol
Cas	CRISPR-associated protein
Cascade	CRISPR-associated complex for antiviral defense
cpm	counts per minute
CRISPR	Clustered Regulary Interspaced Short Palindromic Repeats
crRNA	CRISPR-RNA
Da	Dalton
DEPC	diethylpyrocarbonate
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
kb	kilobases
kDa	kilo Dalton
M	molar (mol/L)
min	minutes
dsDNA	double-stranded DNA
dsRNA	double-stranded RNA
DTT	dithiothreitol
e.g.	for example
EDTA	ethylene-diamine-tetraacetic acid
EMSA	electrophoretic mobility shift assay
EtBr	ethidium bromide
et al.	*et alteri* = and others
Fig.	figure
FPLC	Fast Protein Liquid Chromatography
g	gram
x g	gravitational acceleration
h	hour(s)
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
(6x) His-tag	(hexa) histidine tag
IPTG	isopropyl β-D-1-thiogalactopyranoside
Kan	kanamycin
L	liter
LB	lysogeny broth
ssDNA	single-stranded-DNA
ssRNA	single-stranded-RNA

MOPS	3-(N-morpholino)propanesulfonic acid	SDS	sodium dodecyl sulfate
MW	molecular weight	T7 polymerase	RNA polymerase of the T7 bacteriophage
µ	micro (10^{-6})	tRNA	transfer-RNA
N-terminal	amino-terminal	TAE	trist-acetate EDTA-buffer
Ni-NTA	nickel-nitrilotriacetic acid	TCA	trichloroacedic acid
nt	nucleotides	TBE	tris-borate EDTA-buffer
NTP	nucleoside triphosphate	TEMED	N,N,N';N'-tetramethylethylenediamide
OD_{600}	optical density at 600 nm	Tris	tris-(hydroxymethyl)-aminomethane
ORF	open reading frame	U	unit (enzyme activity)
PAGE	polyacrylamide gel	Vol	volume
PCR	polymerase chain reaction	W	Watt
pH	negative logarithm of the hydrogen ion (H^+) concentration	% (v/v)	percent by volume
Phusion	*Pyrococcus*-like DNA polymerase fused to a processivity-enhancing domain	% (w/v)	percent by weight
RNA	ribonucleic acid	>	higher than
RNase	ribonuclease	<	lower than
rpm	rounds per minute		
RT	room temperature		

TABLE OF CONTENTS

1. Introduction

1.1 Prokaryotic defence mechanisms

Bacteriophages and their respective hosts are co-evolving in natural environments. Phages are the most abundant entities on our planet and it has been estimated that they outnumber their prokaryotic hosts by approximately tenfold [1]. Thus, bacteria have evolved or acquired protective mechanisms that can be classified as innate or adaptive anti-phage systems [2]. The innate immune systems prevent phage adsorption, cleave phage nucleic acids or abort phage infection [3]. One mechanism to escape phage adsorption is to modify the cell surface receptors [4]. Restriction-modification systems protect the cell by recognizing and degrading unmethylated phage DNA [5]. Finally, abortive infection systems lead to the death of a cell after recognizing its infection [3]. Adaptive immunity on the other hand is provided by CRISPR-Cas systems [6]. Phages have also evolved strategies to overcome bacterial protection resulting in a continuous arms race between the two populations.

1.2 The CRISPR-Cas immune system

CRISPR (**C**lustered_**r**egularly_**i**nterspaced_**s**hort_**p**alindromic_**r**epeats) - Cas (**C**RISPR-**as**sociated) are widespread adaptive and inheritable immune systems found in 45 % of bacterial and 84 % of archaeal genomes that have been sequenced (CRISPRdb status August 2014) [7]. CRISPR systems are formed by one or more CRISPR loci together with a set of *cas* genes that are often found in close proximity to the CRISPR locus (fig 1.1) [8].

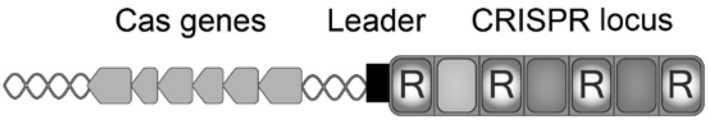


Figure 1.1: Common features of a CRISPR system. A CRISPR system consists of a CRISPR locus (repeat elements that are interspaced by spacer sequences) flanked by a leader sequence and a set of *cas* genes. (Modified from [9]).

A CRISPR locus consists of short identical repeat sequences (R) that were first discovered in *E. coli* by Ishino *et al*. in 1987 [10]. Based on sequence similarity, repeat sequences were initially organized into 12 clusters [11]. A recent bioinformatical classification identified 6 superclasses with 40 conserved sequence families and 33 potential structure motifs based on the repeat sequence and potential secondary structures within the repeats [12]. In the CRISPR locus, the repeat elements are interspaced by unique spacer sequences that were found to be mostly derived from extrachromosomal DNA, e.g. from viruses or plasmids [13-15]. CRISPR loci are flanked by 300-500 bp leader sequences that display the same orientation as the repeats [8]. CRISPR leaders are AT-rich sequences that contain promoter elements that direct CRISPR locus transcription [16, 17]. The leader-end is also the integration site of new spacers [18, 19]. CRISPR, together with the associated *cas* genes, provides resistance against mobile genetic elements based on sequence-complementarity between spacer sequences and invader DNA [6]. The Cas proteins that are associated with the CRISPR loci are fundamental components of the prokaryotic defence system [8, 20, 21]. CRISPR-Cas activity starts with the recognition of a short sequence of viral DNA termed protospacer by the adaptation complex which is integrated into the hosts CRISPR array, generating a new spacer [6, 22, 23]. The CRISPR array is transcribed into long precursor crRNA (pre-crRNA) and further processed into mature CRISPR RNA (crRNAs) [24-26]. The mature crRNAs are incorporation into CRISPR ribonucleoprotein complexes (crRNPs) and are then used as guide molecules to target and degrade the foreign DNA or RNA via base-complementarity in case of a repeated infection [27-29].

1.3 CRISPR-Cas diversification

CRISPR-Cas systems are highly diverse as the selective pressure of invaders drives the evolution of the prokaryotic immune systems [30]. Jansen *et al.* first discovered a link between CRISPR loci and four associated *cas* genes [8]. Comparative analyses of CRISPR systems revealed major differences in Cas protein sequences, repeat sequences and *cas* operon architecture [20, 21]. A Cas protein nomenclature has been developed by Makarova *et al.* based on these differences. CRISPR-Cas systems are classified into three distinct major types (I, II and III) and 11 subtypes (I A-F, II A-C, III A-B) [21, 31] (fig 1.2). Cas1 and Cas2 form the adaptation complex, which is responsible for spacer acquisition and are conserved in all active CRISPR-Cas systems, constituting the core of the major types. Each type as well as each subtype is defined by conserved signature genes (fig 1.2). Type I systems contain the *cas3* gene which encodes a DNase with helicase activity. The type I crRNPs are termed Cascade (**C**RISPR-**as**sociated-**c**omplex for **a**ntiviral **de**fence) and target dsDNA[27]. *cas9* is the signature gene of type II CRISPR-Cas systems, a stand-alone nuclease that is sufficient for dsDNA target cleavage [32]. Type III CRISPR systems contain the type- specific gene *cas10*. The type III crRNP complexes are known as Csm (type III-A), targeting DNA and Cmr (type III-B), targeting RNA [33, 34].

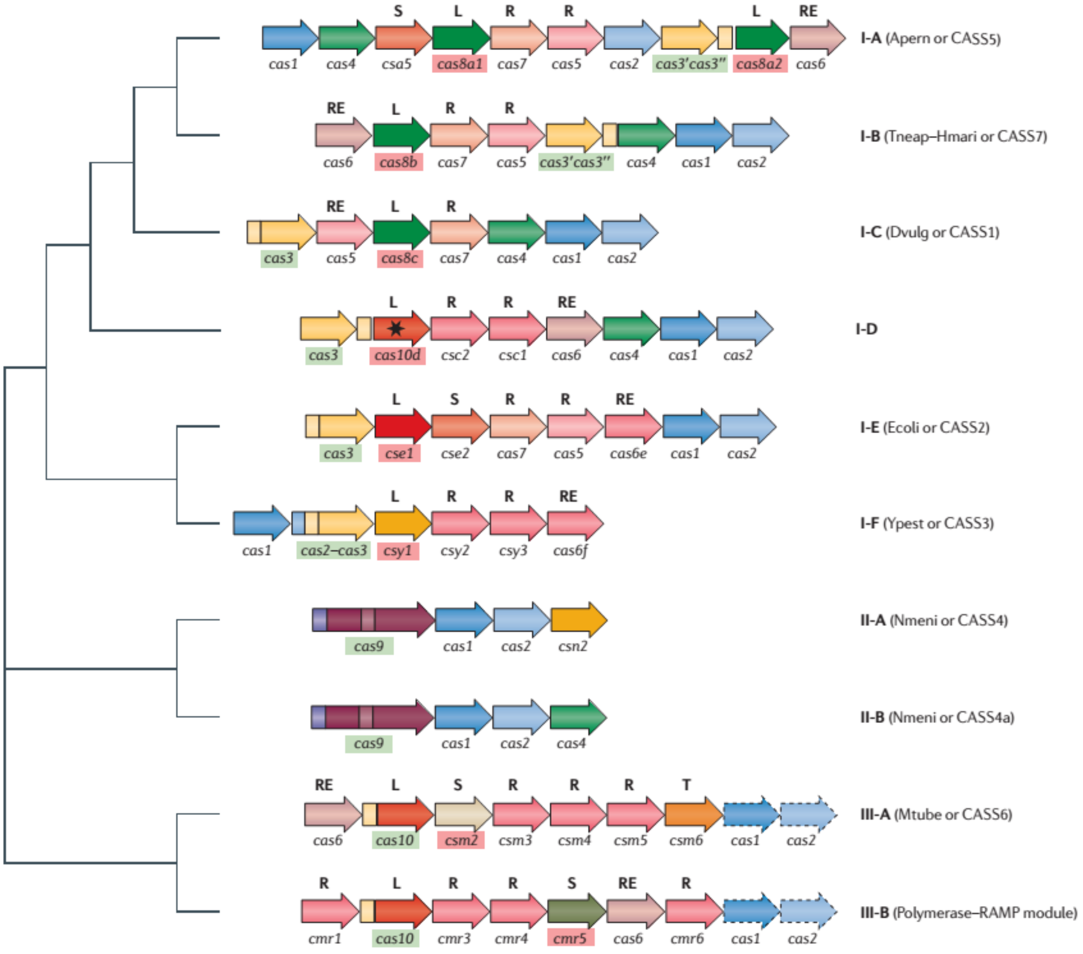

S L R R L RE
cas1 cas4 csa5 cas8a1 cas7 cas5 cas2 cas3'cas3'' cas8a2 cas6
I-A (Apern or CASS5)
RE L R R
cas6 cas8b cas7 cas5 cas3'cas3'' cas4 cas1 cas2
I-B (Tneap–Hmari or CASS7)
RE L R
cas3 cas5 cas8c cas7 cas4 cas1 cas2
I-C (Dvulg or CASS1)
L R R RE
cas3 cas10d csc2 csc1 cas6 cas4 cas1 cas2
I-D
L S R R RE
cas3 cse1 cse2 cas7 cas5 cas6e cas1 cas2
I-E (Ecoli or CASS2)
L R R RE
cas1 cas2–cas3 csy1 csy2 csy3 cas6f
I-F (Ypest or CASS3)
cas9 cas1 cas2 csn2
II-A (Nmeni or CASS4)
cas9 cas1 cas2 cas4
II-B (Nmeni or CASS4a)
RE L S R R R T
cas6 cas10 csm2 csm3 csm4 csm5 csm6 cas1 cas2
III-A (Mtube or CASS6)
R L R R S RE R
cmr1 cas10 cmr3 cmr4 cmr5 cas6 cmr6 cas1 cas2
III-B (Polymerase–RAMP module)

Figure 1.2: Classification of CRISPR-Cas systems by Makarova *et al.*: The typical architecture of *cas* operons is shown for each CRISPR-Cas subtype. The type-specific genes are boxed in green, the subtype-specific genes are shown in red boxes. Major Cas protein categories are depicted in letters above the genes: L=large crRNP complex subunit, S=small crRNP complex subunit, R=RAMP (repeat-associated mysterious protein) crRNP complex subunit. RE=RNases of the RAMP family that are associated with crRNA processing, T=transcriptional regulators [21].

In contrast to type I and type III CRISPR systems which are distributed among both, bacteria and archaea, type II CRISPR systems have been exclusively found in bacterial genomes [20, 35]. Type I and type III systems encode Cas proteins of a superfamily classified as RAMP (**r**epeat-**a**ssociated **m**ysterious **p**roteins) (fig 1.2, genes marked with “R”), including the proteins Cas5, Cas6 and Cas7. These proteins contain an RRM (RNA recognition motif) domain, a widespread RNA-binding domain, and interact with the crRNA. This suggests an evolutionary relationship between these Cas proteins and hints at a shared origin of type I and type III CRISPR systems [21, 31]. Often, one genome encodes CRISPR systems of more than one type and phylogenetically distant organisms can contain CRISPR systems of the same subtype. This highlights the mobility of CRISPR systems that are frequently exchanged via horizontal gene transfer. Thus, it is proposed that different CRISPR subtypes evolved independently (RAMPs and subtype specific proteins) in selected lineages, probably resulting from the selective pressure caused by phage predation [36-38].

1.4 Type I CRISPR-Cas action

CRISPR-Cas adaptive immunity comprises three stages of activity: acquisition, processing and interference (fig 1.3). The acquisition stage starts with a viral infection. A short stretch of viral DNA, termed protospacer, is recognized and integrated into the host’s CRISPR array. Thus, a newly acquired spacer is formed and spacer sequences can maintain a chronological record of recent invaders (fig 1.3 (1), (2)) [6, 22, 23, 39]. The new spacer is integrated directly downstream of the leader sequence by nicking the repeat element on opposite sites on both strands. The repeat sequence is subsequently duplicated [22, 23, 39-41].

Protospacers are selected for integration into the CRISPR-array via the detection of PAM (**p**rotospacer **a**djacent **m**otif) sequences. PAMs are 2-5 nt long sequences, located adjacent of the protospacer. They are found only in type I and type II systems and are located at the 3'-end of the protospacer on target strand complementary to the crRNA. The PAM sequences are subtype-specific [19, 22, 23, 42]. PAMs are crucial element for ‘self”- versus ‘non-self’ DNA discrimination in the host, as it does not base-pair with the respective positions of the repeat sequences next to corresponding spacer elements [13, 19, 37]. Hence, PAMs also play an important role in the Cascade interference reaction [43-45]. Cas1 and Cas2 are conserved in all systems and have been shown to mediate the acquisition of new spacers, but their exact function in the mechanism is still unknown. It is possible that additional proteins are involved in the adaptation process, as the *cas* gene architecture of type-I systems reveal a conserved gene clustering of *cas1* and *cas2* e.g. with *cas4* (fig 1.2) [21, 23, 39].

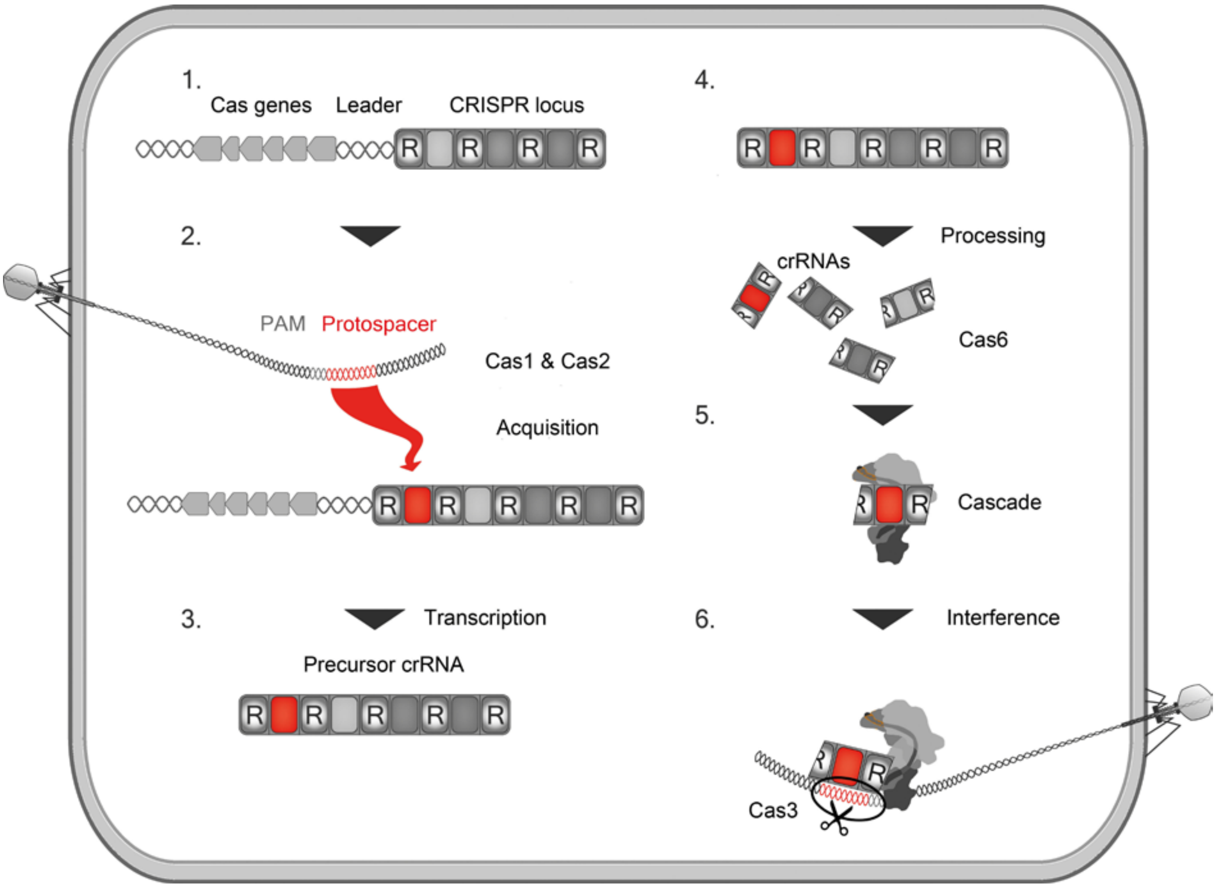


Figure 1.3: Schematic overview of CRISPR-Cas activity. A viral DNA sequence (protospacer, red) is inserted into the CRISPR locus (adaptation 1, 2). This is followed by the transcription of the array into pre-crRNA and subsequent processing into mature crRNAs (3, 4). Then, the crRNA is taken up into the Cascade complex (5) and, in case of a repeated infection, interference can be triggered by sequence complementarity between crRNA and protospacer. (Modified from [9]).

The next stage, termed CRISPR expression, covers transcription of the CRISPR array followed by pre-crRNA processing (fig 1.3 (3), (4)). The entire CRISPR locus is transcribed into a long pre-crRNA starting from the CRISPR leader [16, 18]. The pre-crRNA transcript is then processed at a single phosphodiester bond within the repeat sequences that is hydrolyzed by metal-independent Cas6 endonucleases. As a result, mature crRNAs are obtained that contain spacer sequences flanked by a repeat-derived 5'-terminal tag of 8 nucleotides and a longer repeat-derived 3'-tag [26, 27, 46-48]. Cas6 cleavage generates crRNAs that encompass a 5'-terminal hydroxyl and a 3'-terminal 2'-3' cyclic phosphate end [26, 49]. Type-I Cas6 enzymes usually comprise a catalytic triad with one invariant histidine residue, even though the relative positions of the catalytic amino acids are poorly conserved [26, 46, 47, 50-52]. The Cas6 homologues belong to the RAMP superfamily and share a common RRM motif, but their amino acid sequence was found to be very diverse and their protein structures vary. These divergences are thought to be responsible for the variability in recognition mechanisms of RNA substrates with different structures by different Cas6 homologues [26, 48, 51-53]. As an example, in type I-E and I-F systems, the respective Cas6 enzymes (Cas6f and Cas6e) bind repeat sequences that potentially form a hairpin structure. After Cas6 processing, the mature crRNAs are loaded into Cascade [45, 49] (fig 1.3 (5)). Cas6f and Cas6e are single-turnover enzymes that stay firmly associated with the repeat hairpin and form a stable Cascade subunit after crRNA delivery [11, 26, 51-56]. In contrast, other type-I variants are associated with repeat elements that are predicted to be unstructured and display mature crRNAs that harbor trimmed 3'-terminal ends. This suggests that the respective Cas6 variants are not a permanent subunit of Cascade, but function as stand-alone nucleases that only deliver the crRNAs to Cascade, which are then accessible for further enzymatic and/or chemical trimming [12, 43, 50]. The evolutionary basis for these mechanistic differences are not fully understood [57]. After associating with Cascade, mature crRNAs are used as guide molecules to target foreign DNA in case of a repeated infection via base complementaryty, which then results in the degradation of the viral DNA by a helicase/endonuclease called Cas3 (fig 1.3 (5), (6)). Experimental insights into type I-A, I-C, I-F and I-E Cascades have been published. These Cascades display similarities in the protein composition as they all share the three RAMP-containing Cas5, Cas6 and Cas7 variants, even though they differ in the subtype specific proteins that represent the large (and small) subunit of the complex (Cas8a and Csa5 for type I-A, Cas8c for type I-C, Cse1 and Cse2 for type I-E and Csy1 for type I-F, see fig 1.2). A major difference that can be observed between the Cascades is the composition of subunits that are permanently or temporarily associated components of the complexes (Cas6 and Cas3 variants) [21, 27, 43, 46, 48, 49, 55, 58, 59]. The type I-E Cascade was first described for *E. coli* and is the best studied Cascade in terms of structure and function [45, 54, 59, 60]. Its structure highlighted an uneven stoichiometry of: $(\text{Cse1})_1$-$(\text{Cse2})_2$-$(\text{Cas5})_1$-$(\text{Cas7})_6$-$(\text{Cas6})_1$ [49, 61]. Cryo-electron microscopy (cryo-EM) revealed an overall seahorse-shaped architecture [62]. Recently, the crystal structure of Cascade containing crRNA (and ssDNA) was solved by Jackson *et al*., Mulepati *et al*. and Zhao *et al*. [54, 60, 63] (fig 1.4). Cas6e and Cas5e anchor the crRNA repeat termini forming the head and tail of the complex. Cas6e is bound to the stem loop structure at the 3'-end of the crRNA and Cas5e reveals base-specific interactions at the 8 nt 5'-tag of the crRNA, kinking the nucleotide at position -1 with its thumb domain. Six subunits of Cas7 form the backbone along the spacer element. Each thumb domain of the Cas7 subunits kinks at every sixth position of the spacer sequence. Cas5e and Cas7, as well as the Cas7 subunits interact via palm and thumb domains. The thumb domain of Cas7.1 finally folds into a protein cleft of Cas6. The large subunit (Cse1) interacts with the Cas5 tail, Cas7 and Cse1. The small subunits (Cse2) display protein contact with the Cas7 backbone [54, 57, 60, 63].

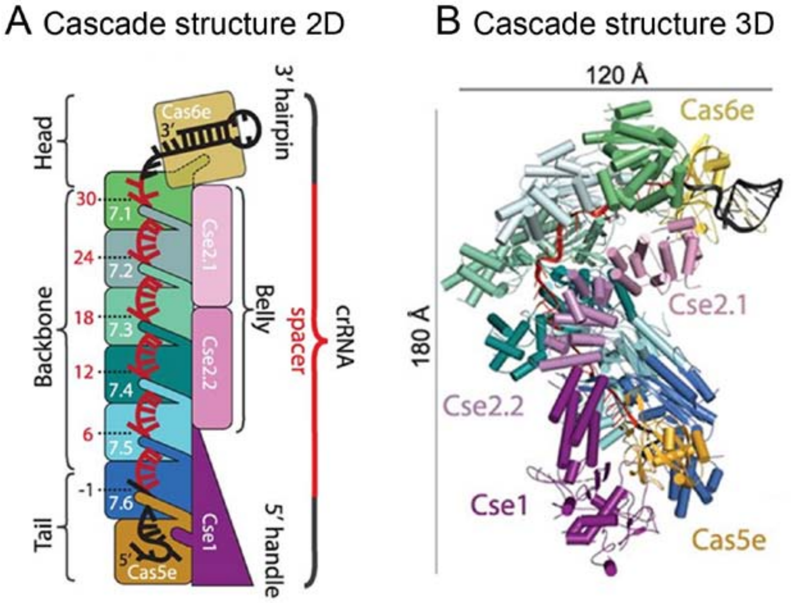


Figure 1.4: Type I-E Cascade structures in 2D and 3D. A) and B) Cas6e binds the hairpin structure at the 3'-terminal end of the crRNA (black) forming the head of the complex. The backbone consists of six Cas7 subunits that oligomerize along the spacer sequence (red). Cas5e caps the 5'-terminal repeat tag of the crRNA (black) as the tail. The large subunit Cse1 is associated with the tail and interacts with Cas5e, Cas7 and Cse2. The two Cse2 small subunits interact with Cas7 and form the belly of the complex (modified from [54]).

The fully assembled Cascade scans dsDNA for a potential PAM sequence. Cse1 is responsible for PAM identification [59, 64]. The PAMs that are crucial for target interference can differ from the PAMs that are required for spacer acquisition. Few motif variants of 2-3 bp are tolerated for interference [44, 65, 66]. Cse1 interacts with the PAM, which might destabilize the invader DNA and enable base-pairing between crRNA and target DNA [67, 68]. Base-pair complementarity between protospacer and the seed region of the crRNA (nucleotides 1-5 and 7-8 at the 5'-end of the spacer sequence) is required, whereas a limited number of mismatches in the remaining spacer sequence is tolerated [55, 69-71]. The non-target strand of the invader DNA is displaced, generating an R-loop [62]. The R-loop formation triggers a major conformational change of Cascade, which mediates the recruitment of the metal-dependent nuclease and ATP-dependent helicase Cas3 that interacts with the base of Cse1 [49, 59]. Cas3 nicks the displaced strand which induces a conformational change of the protein and activates the helicase domain [45, 72]. The target DNA is unwound and exonucleolytically degraded in 3' to 5' direction [45, 57, 59, 72, 73].

1.5 CRISPR-Cas subtype I-B

CRISPR-Cas systems of subtype I-B are distributed among archaea and bacteria. According to an older nomenclature by Haft *et al*., two I-B subtypes exist: I-B Tneap and I-B Hmari. These subtypes are named after a species that contains such a system (*Thermotoga neapolitana* and *Haloarcula marismortui*) [20].

Type I-B systems consist of Cas1, Cas2 and Cas4 that are proposed to be involved in the acquisition process, and the putative Cascade forming proteins Cas3, Cas5, Cas6, Cas7 and Cas8b. Cas3 represents the type I specific helicase, containing an N-terminal HD nuclease domain. The protein can also be found to be split into separated HD and helicase domains [21]. In *Methanococcus maripaludis, Thermococcus kodakarensis, Haloferax volcanii* and *Haloferax mediterranii* type I-B systems, the Cas6b endonuclease was shown to processes pre-crRNA transcripts into mature crRNAs that comprise an 8 nt 5'-terminal repeat tag and a spacer sequence followed by a 3'-terminal repeat handle [50, 74, 75]. In *Haloferax volcanii*, Cas7 and Cas5 were identified as type I-B Cascade subunits that interact with Cas6 and stabilize the mature crRNAs [75, 76]. A recent study identified a type I-B complex formed by Cas5 and Cas7 from *Pyrococcus furiosus* via immunopurification [77]. Cas8b is the subtype-specific protein and predicted to form the large Cascade subunit. A multiple alignment of representative sequences and predicted secondary structures of proposed type I large subunits by Makarova *et al.* is shown in figure 1.5. All large subunits are predicted to contain a finger, palm and thumb domain, and a zinc-finger domain is present in most proteins. These domains are shared by DNA polymerases and their shape resembles a right hand [78]. The large subunits of type I-B (Cas8b) and I-C (Cas8c) display a thumb domain with an extra alpha-helical region. In these subtypes, the small subunit, typically an alpha-helical protein (Cse2 in type I-E and Csa5 in type I-A), is not encoded in the Cas operons. It is possible that this C-terminal region is compatible with a small Cascade subunit according to size and structure [79]. However, structural and functional data on Cas8b proteins are not yet available.

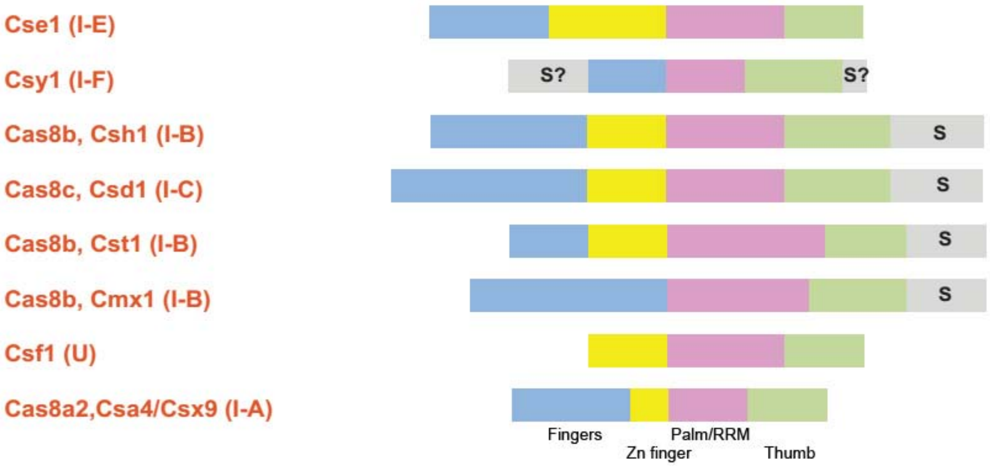


Figure 1.5: Alignment of large type-I Cascade subunits. Domain organization of different large type-I subunits is color-coded. S = regions that could be homologous to the small subunits encoded as separate genes in type I-A and I-E. Additional protein designation originates from an older nomenclature by Haft *et al.* 2004. (Modified from [79]).

1.6 Endoribonuclease III and its role in CRISPR-Cas

The bacterial endoribonuclease III (RNase III) is a double-stranded (ds) RNA-specific endonuclease that is characterized by a specialized endonuclease domain known as the RNase III domain. Many bacterial RNase III enzymes also display a dsRNA-binding domain (dsRBD) [80]. RNase III cleavage is metal-dependent (Mg^{2+}) and generates dsRNA products comprising 5'-phosphoryl and 3'-hydroxyl termini and a 2 nt 3'-overhang [81, 82]. The RNase III protein family includes bacterial RNase III enzymes as well as the eukaryotic proteins Dicer and Drosha that contain a more complex RNase III domain and are involved in the RNA interference mechanism (RNAi), mediating the production of small interfering RNAs (siRNAs) and the maturation of micro RNAs (miRNAs). Among the members of the endoribonuclease III family, the respective enzyme from *E. coli* is the most comprehensively studied protein [83-86]. Bacterial RNase III is involved in posttranscriptional gene expression, ribosomal RNA (rRNA) processing and the defence against viral infection [32 87, 88]. It is a global regulator and controls gene expression by processing mRNAs [89]. This includes dsRNA cleavage of RNA hybrids formed by mRNA and small regulatory RNAs, such as anti-sense RNAs or stem-loops within the mRNA [90, 91]. This cleavage can occur within the coding region, the 5' or 3' untranslated region (UTR) and can cause structural changes of an mRNA, resulting in promoted degradation or a more efficient translation. Gene expression might also be regulated by protein binding in the 3' and 5' UTRs [92]. RNase III forms a homo-dimer that comprises a catalytic valley which is formed by two RNase III domains and comprises two symmetric cleavage sites [93]. Nucleotides have been identified at specific positions in a dsRNA substrate that inhibit protein binding or allow binding, but prevent cleavage [94].

In type II CRISPR-Cas systems, RNase III is involved in crRNA maturation. Type II systems encode a *trans*-activating crRNA termed tracrRNA, in the vicinity of the CRISPR loci. The tracrRNA contains an anti-CRISPR repeat that pairs with the repeat sequences of the pre-crRNA transcripts [32]. The resulting RNA heteroduplexes are stabilized by Cas9 and subsequently cleaved by RNase III [95]. After additional processing of the crRNA, the mature crRNA-tracrRNA hybrid is used to direct the interference reaction, mediated by the Cas9 RNP complex [32, 95].

The RNase III domain, as well as the dsRBD are conserved and present in RNase III family proteins (RNase III, Dicer, Drosha). These endoribonuclease homologues are used in both, type II CRISPR-Cas systems as well as in the RNAi mechanism, to generate anti-invader guide RNAs [92].

1.7 *Clostridium thermocellum*

Clostridium thermocellum is a gram-positive bacterium with a rod-shaped cell body (fig 1.6). It is a strictly anaerobic and thermophilic (60°C) bacterium that is found in places with rotting biomass [96, 97]. *C. thermocellum* has gained biotechnological interest as it produces an active cellulase enzyme complex known as the "cellulosome" which is organized on the outer surface of the bacterium [98]. It enables the fermentation of cellulosic biomass to ethanol, acetic acid, lactic acid, CO_2 and H_2, suggesting future applications of this organism for biomass conversion [98, 99].

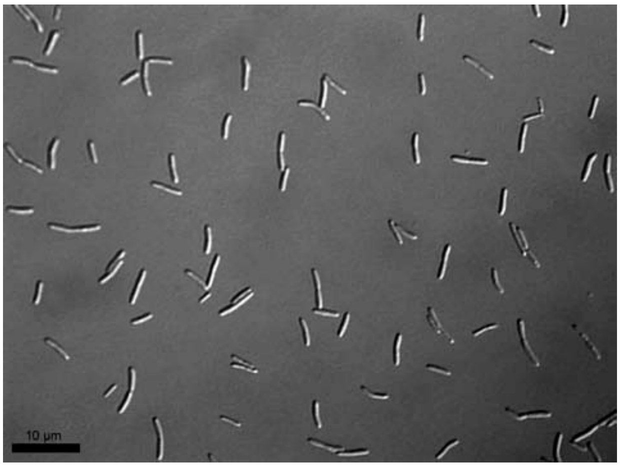


Figure 1.6: DIC picture of *Clostridium thermocellum*. Differential interference contrast microscopy of *C. thermocellum* reveals a rod-shaped cell body. Scale bar = 10µm. (Janine Haueisen, MPI Marburg).

The whole genome sequence of *Clostridium thermocellum* ATCC 27405 is available (GenBank no. CP000568; DOE Joint Genome Institute) and enabled the identification of two type I-B CRISPR-Cas systems. Computational analyses revealed that one of the CRISPR systems shares similarities with the archaeal type I-B CRISPR system of *Methanococcus maripaludis* which was previously analyzed in our laboratory, indicating a potential horizontal gene transfer event between Clostridia and methanogens. *C. thermocellum* can be used to study a bacterial type I-B CRISPR-Cas system and allows a comparison with the archaeal I-B system of *M. maripaludis*.

1.8 Aim of the study

The main focus of this study is placed on the *in vitro* characterization of a recombinant type I-B Cascade from *Clostridium thermocellum*. This necessitates the production of the Cascade forming proteins Cas3, Cas5, Cas6, Cas7 and Cas8b. Additionally, mature crRNAs are generated from pre-crRNA by the endonuclease Cas6. The assembly of the Cas proteins into Cascade is used to determine permanent or temporarily associated complex subunits. Furthermore, Cascade assemblies with modified crRNAs provide information about the Cascade loading mechanism. The protein stoichiometry of the complex, as well as the protein interaction sites are investigated via mass-spectrometric techniques in collaboration with Kundan Sharma (MPI Göttingen). RNA-Seq methodology is utilized to analyze small RNAs isolated from *C. thermocellum*. These data allow the characterization of the cellular crRNA pool. Together, these studies aim to provide information about crRNA processing and subsequent crRNP complex formation of a bacterial type I-B CRISPR system. The comparison of the respective mechanisms with other CRISPR-Cas subtypes will contribute to our understanding of CRISPR functionality and diversification.

2. Results

2.1 Genomic context of CRISPR-Cas subtype I-B in *Clostridium thermocellum*

In *Clostridium thermocellum* ATCC 27405, five CRISPR loci are annotated in the CRISPRdb database (fig 2.1) [7]. The loci 1-3 contain repeat sequences of 30 bp, whereas the repeat elements of locus 4 and 5 are 37 bp long. All five CRISPR arrays are flanked by transposase and integrase genes, which highlights the mobility of these systems. Two sets of *cas* genes were identified using the CRISPR interactive database "CRISPI" [100]. Each set is associated with CRISPR loci that comprise repeat elements of identical length. Several classification systems have been introduced based on the Cas protein composition independent from the repeats. According to the nomenclature by Makarova *et al.*, both sets of *cas* genes were classified as type I-B CRISPR systems that are defined by the subtype specific protein Cas8b and the presence of the type-specific *cas3* gene that is fused to a sequence encoding an additional HD nuclease domain. The subtype I-B Cas protein organization in *C. thermocellum* includes the adaptation proteins Cas1, Cas2 and Cas4, the crRNA processing endonuclease Cas6 and the proposed Cascade forming proteins Cas3, Cas5, Cas7 and Cas8b (Cthe_2296-2303 associated with loci 1-3, and Cthe_3201-3205 together with Cthe_3218-3220 associated with loci 4-5). However, according to an older nomenclature by Haft *et al.* of 2004, the two CRISPR systems would have been classified as type "I-B Tneap" (Cthe_2296-2303) and type "I-B Hmari" (Cthe_3201-3205, 3218-3220), which have been later summarized into subtype I-B. In this older nomenclature, each subtype is named after a species containing a genome with the single subtype. In general, the subtype Tneap (*Thermotoga neapolitana*) seems to be more distributed among bacteria, whereas subtype Hmari (*Haloarcula marismortui*) can rather be found in archaeal organisms. The distinction between these subtype I-B systems in *C. thermocellum* was based on a BLAST analysis of Cas8b, Cas7 and Cas5, known as either Csh1, Csh2 and Cas5/Hmari of subtype Hmari, or Cst1, Cst2 and Cas5/Tneap of subtype Tneap, as well as the analysis of organisms that contain homologous proteins.

Computational analyses also revealed that some of the closest relatives of the type I-B Hmari Cas proteins are also found in the archaeal model system *Methanococcus maripaludis* that is also analyzed in our laboratory. The high similarity between the *cas* genes of both organisms indicates a potential horizontal gene transfer event. The type I-B Hmari Cascade proteins (fig 2.1, black box) were chosen to be investigated in this study to allow the comparison of highly related CRISPR systems in the two prokaryotic domains of life.

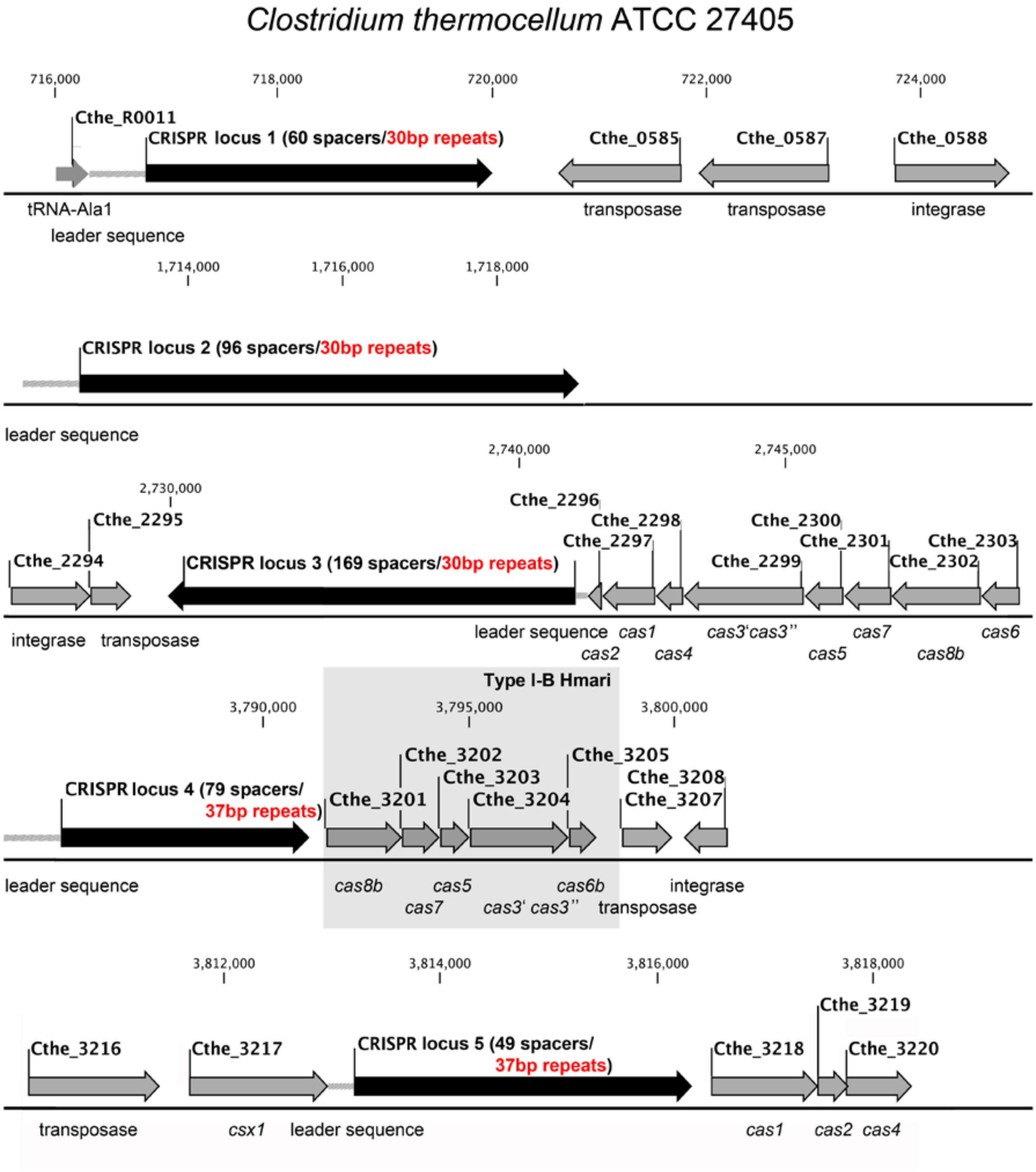


Figure2.1: Genomic context and type I-B *cas* gene organization. Two sets of *cas* genes are organized in operons, encoding the Cas proteins that are involved in CRISPR interference (Cas3, 5, 6, 7, 8b) and proteins that mediate CRISPR adaptation (Cas1, 2, 4). One set of *cas* genes is located adjacent to CRISPR locus 3 (30 bp repeats), whereas the other set is associated with the CRISPR loci 4 and 5 (37 bp repeats). (Modified from [50]).

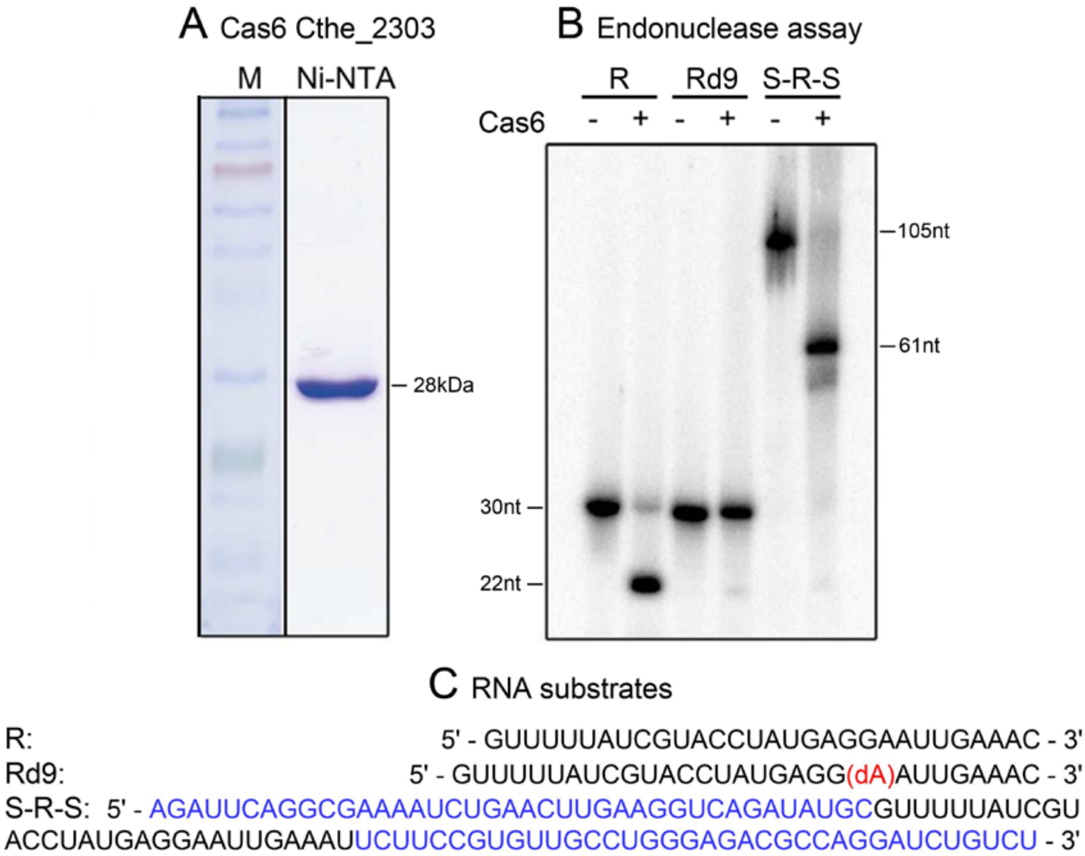
2.2 CRISPR RNA processing

CRISPR RNA processing was analyzed both *in vitro* and *in vivo* for the two CRISPR-Cas subtype I-B systems of *Clostridium thermocellum.*

2.2.1 CRISPR RNA processing *in vitro*

Clostridium thermocellum contains two *cas* genes that encode putative Cas6 proteins (fig 2.1). Cas6 Cthe_2303 is associated with the 30 nt repeat sequences and Cas6 Cthe_3205 with the 37 nt repeat sequences. Both genes were cloned into a pET20b vector and recombinant protein with a C-terminal 6-fold His tag was produced in *E. coli*. Both proteins were purified using a nickel-NTA column.

First, purified Cas6 Cthe_2303 (fig 2.2 A) was used for nuclease assays with 5'-terminal radiolabeled pre-crRNA and repeat RNA substrates and *in vitro* endonuclease activity could be shown (fig 2.2. B). Different RNA transcripts were analyzed (fig 2.2. C), consisting of a single 30 nt repeat element, a repeat sequence with a deoxyribonucleotide substitution at position -9 and a spacer145-repeat146-spacer146 sequence of CRISPR locus 3. The repeat, as well as the pre-crRNA substrates, were cleaved within the repeat sequence, whereas Cas6 endonuclease activity could not be detected for the modified RNA substrate. It has been shown that Cas6 creates mature crRNAs that contain a spacer unit and 8 nucleotides of the 5' repeat element as a 5' terminal tag [26, 27, 48]. This could be validated for Cas6 Cthe_2303 cleavage using the modified repeat RNA with a deoxyribonucleotide introduced at the proposed cleavage site (-9) which abolished Cas6 cleavage. Subsequently, the chimeric oligonucleotide could not only be used as a negative control, but also pinpointed and confirmed the Cas6 Cthe_2303 cleavage site within the repeat sequence in *C. thermocellum* [9]. In agreement with this, Cas6 Cthe_3205 nuclease assays with modified and unmodified 37 nt long repeat RNA substrates also revealed a cleavage site at position -9, generating crRNAs with an 8 nt 5'-terminal tag as well (data not shown).



R: 5' - GUUUUUAUCGUACCUAUGAGGAAUUGAAAC - 3'

Rd9: 5' - GUUUUUAUCGUACCUAUGAGG(dA)AUUGAAAC - 3'

S-R-S: 5' - AGAUUCAGGCGAAAAUCUGAACUUGAAGGUCAGAUAUGCGUUUUUAUCGU
ACCUAUGAGGAAUUGAAAUUCUUCCGUGUUGCCUGGGAGACGCCAGGAUCUGUCU - 3'

Figure 2.2: Cas6 Cthe_2303 purification and endonuclease assay. A) SDS-PAGE of a Cas6 Cthe_2303 (28 kDa) after nickel-NTA chromatography (Ni-NTA). B) Detection of endonucleolytic Cas6 activity for the 5'-terminal labeled repeat (R) and pre-crRNA (S-R-S) *in vitro* transcripts. The introduction of a dNTP at position -9 in the repeat sequence (Rd9) abolished Cas6 cleavage. The bands were separated on a denaturing 8 M urea 12 % polyacrylamide gel and visualized by autoradiography. C) The following RNA substrates were used: R = 30 nt repeat sequence, Rd9 = 30 nt repeat sequence including a dNTP at position -9 (red), S-R-S = spacer145 (blue) - repeat146 - spacer146 (blue) sequence. (Modified from [9]).

To obtain further information about the presence of two different types of repeat elements (30 nt and 37 nt) that are associated with different Cas6 homologous (Cthe_2303 and Cthe_3205) in *C. thermocellum* (fig 2.1), both purified Cas6 proteins (fig 2.2 A, 2.3 A) were used for nuclease assays with single 30 nt and 37 nt repeat elements (fig. 2.3 B). Both substrates were exclusively cleaved only by their corresponding Cas6 enzyme. The fact that both enzymes cleave two different repeat substrates at the same position (-9) and generate a 5'-terminal crRNA tag that is identical in sequence and length (fig B, red arrows) indicates a substrate recognition mechanism based on structure, repeat-length or the sequence upstream of the cleavage site. Computational analysis of both Cas6 enzymes revealed their very limited sequence homology. Cas6 Cthe_3205 showed high homology to the Cas6 protein found in the archaeon *M. maripaludis* (40 % AA identity and 62 % AA similarity). In contrast, BLAST search results of Cas6 Cthe_2303 identified it to be rather similar to Cas6 proteins found in bacteria (e.g. *Thermodesulfobacterium hveragerdense, Bacillus coagulans, Thermoanaerobacterium thermosaccharolyticum*). This might indicate that *C. thermocellum* contains two CRISPR-Cas systems that originate from the two prokaryotic domains of life (archaea and bacteria).

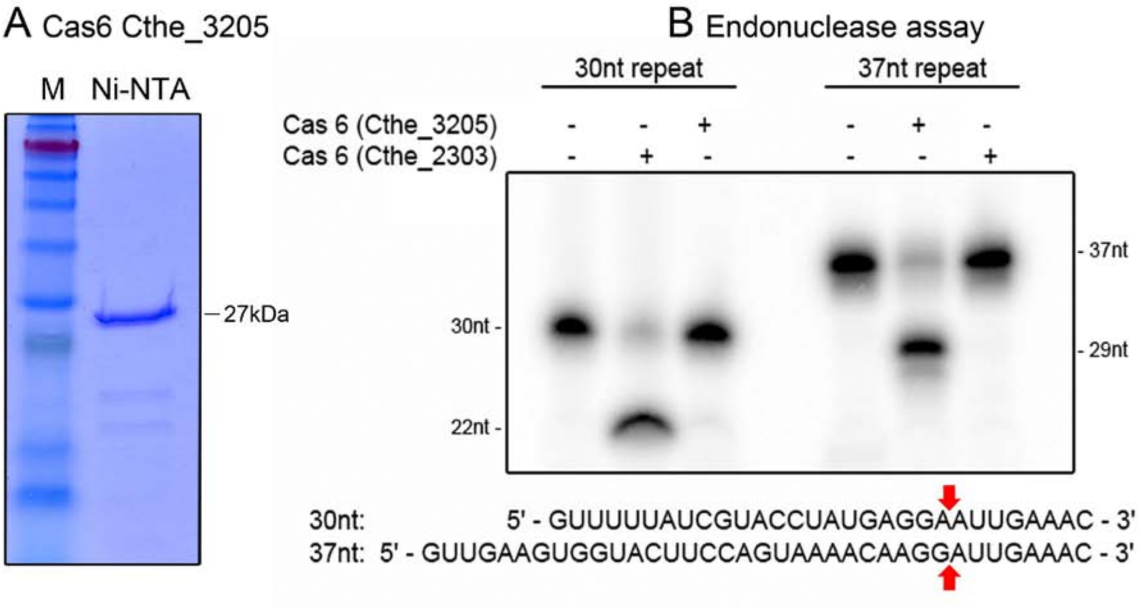

A Cas6 Cthe_3205
M
Ni-NTA
27kDa
B Endonuclease assay
30nt repeat
37nt repeat
Cas 6 (Cthe_3205) - - + - + -
Cas 6 (Cthe_2303) - + - - - +
30nt
22nt
37nt
29nt
30nt: 5' - GUUUUUAUCGUACCUAUGAGGAAUUGAAAC - 3'
37nt: 5' - GUUGAAGUGGUACUUCCAGUAAAACAAGGAUUGAAAC - 3'

Figure 2.3: Cas6 Cthe_3205 purification and endonuclease assay with repeat RNA. A) SDS-PAGE of a Cas6 Cthe_3205 (27 kDa) after nickel-NTA chromatography (Ni-NTA). B) Detection of endonucleolytic Cas6 (Cthe_2303 and Cthe_3205) activity for the 5'-terminal labeled 30 nt and 37 nt repeat RNA substrates. Both enzymes preferably cleave their respective substrate. The cleavage sites are marked within the sequence (red arrows). The bands were separated on a denaturing 8 M urea 12 % polyacrylamide gel and visualized by autoradiography.

2.2.2 CRISPR RNA processing *in vivo*

RNA-Seq methodology was used to analyze the cellular crRNA pool of *Clostridium thermocellum* to gain a more detailed insight into the transcription and the processing of crRNAs *in vivo*. Therefore, total small RNA (up to 200 nt) of *C. thermocellum* was isolated. Since mature crRNAs comprise 5'-hydroxy and 2'-3'-cyclic phosphate termini created by Cas6 cleavage, 5' phosphorylation and 3' dephosphorylation of the isolated RNA via T4 polynucleotid kinase (PNK) was performed to ensure proper adapter ligation.

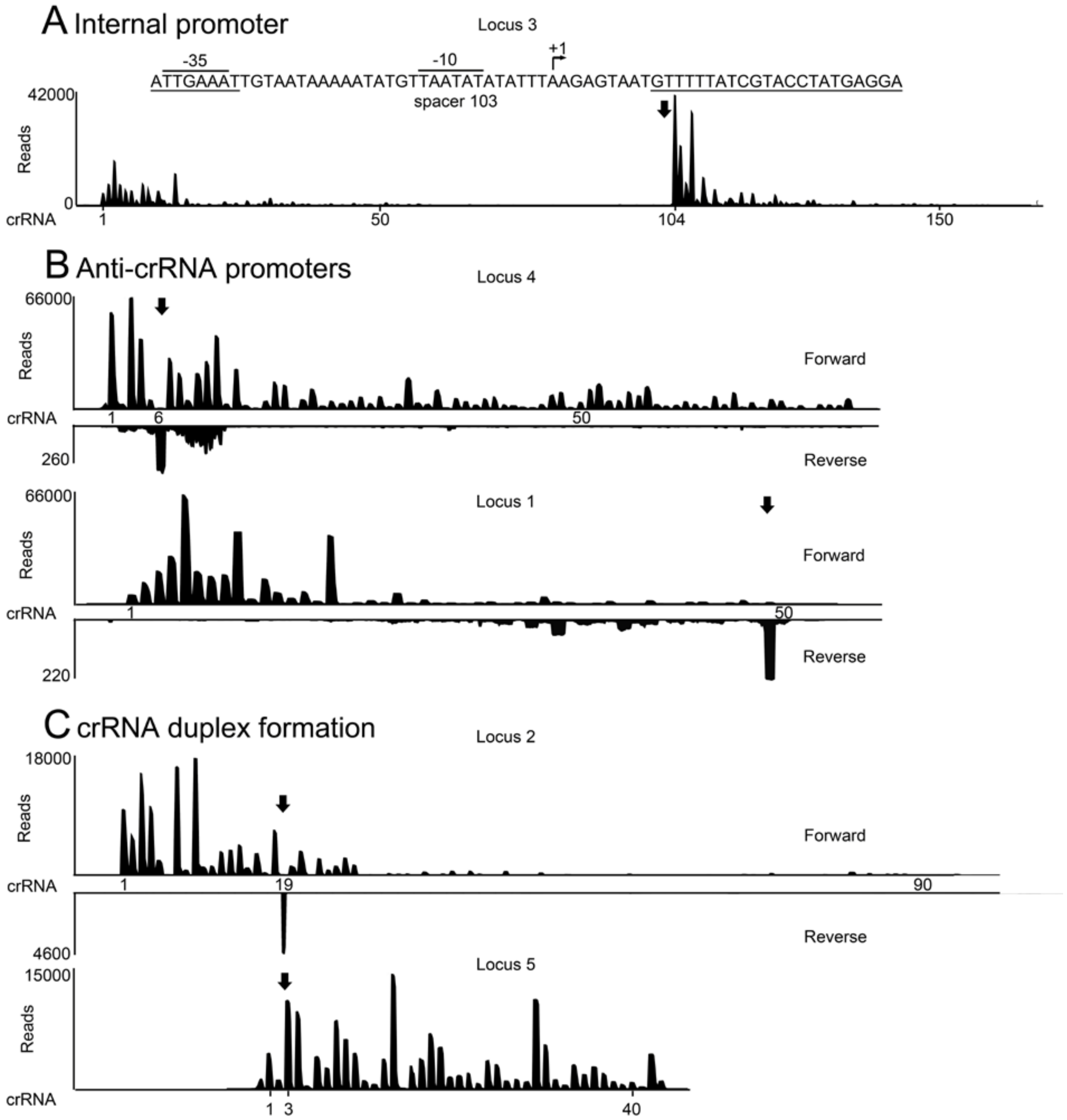


Figure 2.4: RNA-Seq data for *C. thermocellum* CRISPR loci. Illumina HiSeq2000 sequencing reads were mapped to the *C. thermocellum* ATCC 27405 reference genome to illustrate crRNA abundance and processing. A) CRISPR locus 3 reveals a proposed internal promoter located in spacer 103 (adjacent repeat sequences are underlined). B) Bidirectional transcription was exemplified for Loci 1, 2 and 4. Forward and reverse coverage was separated to highlight anti-crRNA promotion that can start within and at the leader-distal end of the locus (black arrows). The occurrence of anti-crRNAs is proposed to correlate with the abundance of crRNAs. C) The indicated crRNA of locus 2 is highly transcribed in reverse direction and complementary to a crRNA of locus 5 (black arrows). (Modified from [101]).

The small RNA libraries were sequenced via Illumina HiSeq2000 [102]. All sequence reads were trimmed to remove adapter sequences and individual sequence reads were mapped to the reference genome. The small RNA analysis revealed that all five CRISPR loci are constitutively transcribed and processed (fig 2.4). Nearly all crRNAs display the same 5'-terminal 8 nt tag (5'-AUUGAAAC-3') which shows that processing of both repeat elements (30 nt and 37 nt) yields the same crRNA 5'-tag *in vivo*. The 3'-termini are trimmed leaving tags of various length. The overall crRNA abundance follows the commonly observed trend which is a gradual decline in the abundance from the leader-proximal to the leader-distal region of the CRISPR loci. There are two plausible explanations for the variability in abundance i) the increase of the pre-crRNA transcript length provokes RNA polymerase stalling as well as premature termination of transcription or ii) the crRNAs that contain spacers from the leader-proximal region of the locus are needed in greater amounts to target the most recent invaders. An exception to commonly observed abundance pattern apparent in locus 3 (fig 2.4 A) where an internal promotion of transcription within the CRISPR array can be observed that is mediated by a spacer element. The overall crRNA abundance declines from crRNA 1 to crRNA 103 and then peaks again for crRNA 104 (fig 2.4 A). A closer analysis of this region revealed a potential -35 element (TTGAAA) encoded by the repeat elements and a putative pribnow box (-10), provided by spacer 103. Together, these two elements form a promoter that seems to be stronger than the promoter encoded in the leader region of CRISPR locus 3. One thing that should be noted is that the 8 nt 5'-terminal tags are not identical for all crRNAs of locus 3. Starting from repeat 115, the final U base changes to C which indicates that CRISPR loci might have recombined. By sequencing this particular genome region upstream of spacer 104, errors in the initial genome assembly during whole genome sequencing could be excluded.

The RNA-Seq data also provide evidence for bidirectional transcription of the CRISPR arrays (fig. 2.4 B), starting from regions within the clusters (CRISPR locus 4) or from the leader-distal ends (CRISPR loci 1, 2 and 5). Even though the number of anti-crRNAs is very small in comparison to the amount of crRNAs, in some cases the high abundance of anti-crRNAs correlates with the reduced amount of their crRNA counterparts, e.g. spacer 2 from locus 2 or spacer 6 from locus 4 (fig 2.4 B&C).

Another interesting aspect was found by analyzing the origin of these spacers from the CRISPR loci. CRISPR loci 2 and 5 harbor spacer sequences that are identical (locus 5, spacer 3 “crRNA 3”) and complementary (locus 2, spacer 19 “crRNA 19”) to parts of the anticodon arm, T arm and the acceptor stem of the single-host tRNA^{Pro} (GGG) isoacceptor RNA (fig 2.5).

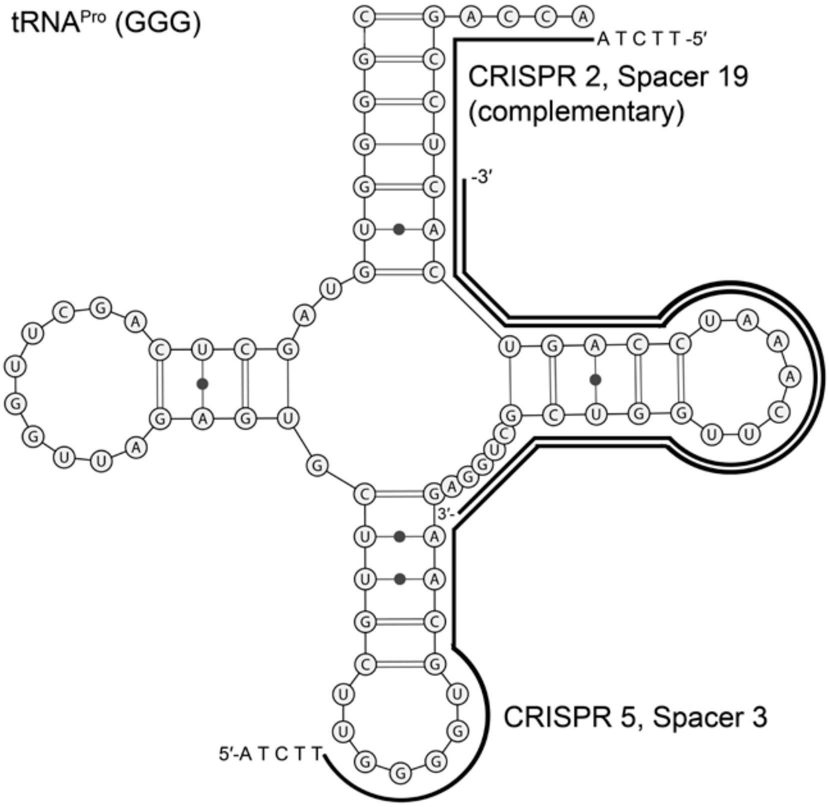


Figure 2.5: Host tRNA^{Pro} with matching crRNA spacers. CRISPR loci 2 and 5 comprise spacer sequences that are identical and complementary to the *C. thermocellum* ATCC 27405 tRNA^{Pro} (GGG) isoacceptor (outlined regions) [101].

These spacer sequences might either originate from the host tRNA^{Pro} gene or from viral attachment sites that target tRNA genes. The abundance of the complementary crRNA 19 might be problematic for the cell, as it potentially targets the essential host tRNA. This could inhibit precursor-tRNA folding and/or induce dsRNA cleavage. The RNA-Seq data (fig 2.4 C) show that crRNA 19 is nearly eliminated from the RNA pool, whereas crRNA 3 is highly abundant. This could indicate that crRNA 3 binds crRNA 19 and acts as an “RNA sponge” to prevent this tRNA^{Pro} targeting. In addition to this, antisense transcripts of spacer 19 could form hybrids with this problematic crRNA and induce dsRNA cleavage. In agreement with this, the RNA-Seq data reveal a great amount of anti-crRNA transcript for spacer 19 of locus 2 (fig 2.4 C).

2.2.3 Anti-crRNA processing *in vivo*

RNA-Seq data were used to analyze the reverse CRISPR transcripts *in vivo*. Individual anti-crRNAs show a conserved processing pattern within the repeat sequences that differs from the pattern of mature crRNAs (fig 2.6). The pattern of the anti-crRNAs from loci 1 and 2 (30 bp repeats) comprise an 18 nt 5′-terminal repeat fragment and a randomly processed 3′-end (fig 2.6 A), whereas the anti-crRNA processing within the 37 nt repeat elements yields a 22 nt 5′-tag and a randomly processed 3′-end (fig 2.6 B). It is plausible that CRISPR precursor RNAs in forward and reverse direction form double-stranded RNA (dsRNA) in the cell. These RNA duplexes could serve as substrates for dsRNA cleaving enzymes. The presence of distinct processed anti-crRNAs, together with the observation that complementary crRNAs are highly reduced in abundance (fig. 2.4 B, C), could indicate that anti-crRNAs neutralize crRNAs via base-pairing.

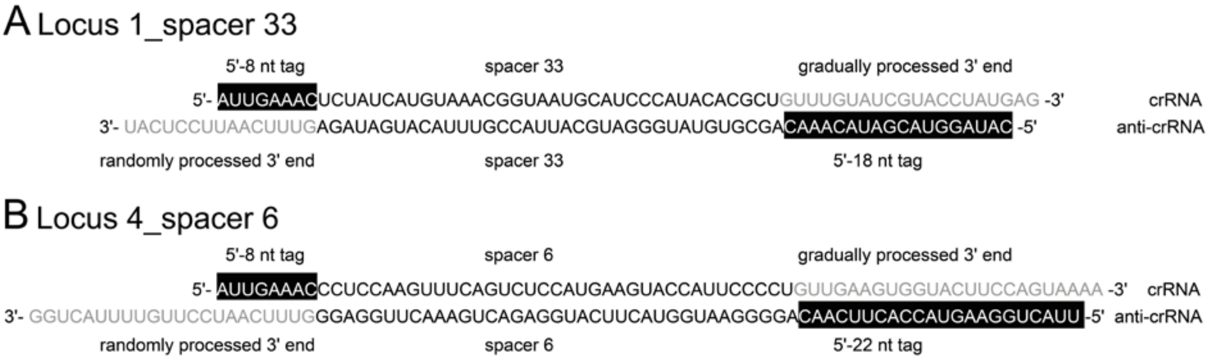

A Locus 1_spacer 33
5'-8 nt tag
spacer 33
gradually processed 3' end
5'- AUUGAAACUCUAUCAUGUAAACGGUAAUGCAUCCCAUACACGCUGUUUGUAUCGUACCUAUGAG -3'
crRNA
3'- UACUCCUUAACUUUGAGAUAGUACAUUUGCCAUUACGUAGGGUAUGUGCGACAAACAUAGCAUGGAUAC -5'
anti-crRNA
randomly processed 3' end
spacer 33
5'-18 nt tag
B Locus 4_spacer 6
5'-8 nt tag
spacer 6
gradually processed 3' end
5'- AUUGAAACCCUCCAAGUUUCAGUCUCCAUGAAGUACCAUUCCCCUGUUGAAGUGGUACUUCCAGUAAAA -3'
crRNA
3'- GGUCAUUUUGUUCCUAACUUUGGGAGGUUCAAAGUCAGAGGUACUUCAUGGUAAGGGGACAACUUCACCAUGAAGGUCAUU -5'
anti-crRNA
randomly processed 3' end
spacer 6
5'-22 nt tag

Figure 2.6: Processing pattern of anti-crRNAs *in vivo*. A) anti-crRNA 33 exemplifies processing within the 30 bp repeat elements leaving an 18 nt 5′-and a randomly processed 3′-end B) anti-crRNAs from loci 4 and 5 (37 bp repeats) display a 22 nt 5′-tag and a randomly processed 3′-end. (Modified from [50]).

2.2.4 RNase III cleaves anti/crRNA duplexes

To gain further insights into the processing of anti-crRNAs, double-stranded anti-/crRNA hybrids were produced and used for cleavage assays with recombinant RNase III from *C. thermocellum*.

RNase III was chosen as the candidate endonuclease, as it is a dsRNA specific nuclease and plays a major role in crRNA maturation in bacterial CRISPR-Cas type II systems. Type II crRNA maturation involves a *trans*-encoded small RNA (tracrRNA) that is partly base-complementary to the repeat regions of crRNA precursor transcripts [32]. The tracrRNA forms dsRNA with the precrRNA and is processed by RNase III, which serves as a host factor in type II crRNA maturation. Additionally, bacterial RNase III belongs to the same protein family as Dicer and Drosha, two well-studied type III nucleases that process short interfering (si) RNAs and micro (mi) RNAs which are involved in eukaryotic gene silencing[103-105]. Hence, the requirement of RNase III in a potential crRNA regulatory mechanism seemed reminiscent of the key roles of these related nucleases.

The RNase III gene from *C. thermocellum* was cloned into a pET20b vector with a C-terminal 6-fold His-tag. Recombinant RNase III was produced in *E. coli* and purified using nickel-NTA affinity chromatography (fig 2.7 A). An endonuclease assay using Cas6 Cthe_2303 and RNase III was performed on 5'-terminal labeled single-stranded (ss) and double-stranded (ds) RNA substrates in forward and reverse direction (fig 2.7 B, C). The RNA substrates consisted of a slightly truncated (25 nt instead of 30 nt) repeat 32 followed by spacer 33 and repeat 33 of CRISPR locus 1. RNase III endonuclease activity was observed on both sense- and antisense strands in a double-strand dependent manner. In contrast, Cas6 cleaves single-stranded repeat RNA as expected. An RNase T1 digest (RNase T1 cleaves upstream of every G residue) of the single-stranded sense-substrate as well as an alkaline ladder were used as size markers to pinpoint the cleavage sites. These sites were identified within the repeat sequences of the RNA substrates, showing the RNase III characteristic 2 nt 3'-overhang (fig 2.7 C). This cleavage pattern observed *in vitro*, confirms the 5'-terminal processing site of anti-crRNAs *in vivo* (fig 2.6). The cleavage products were sized in dependence on the Cas6 cleavage product and differed by 1 nt from the RNase T1 digest marker.

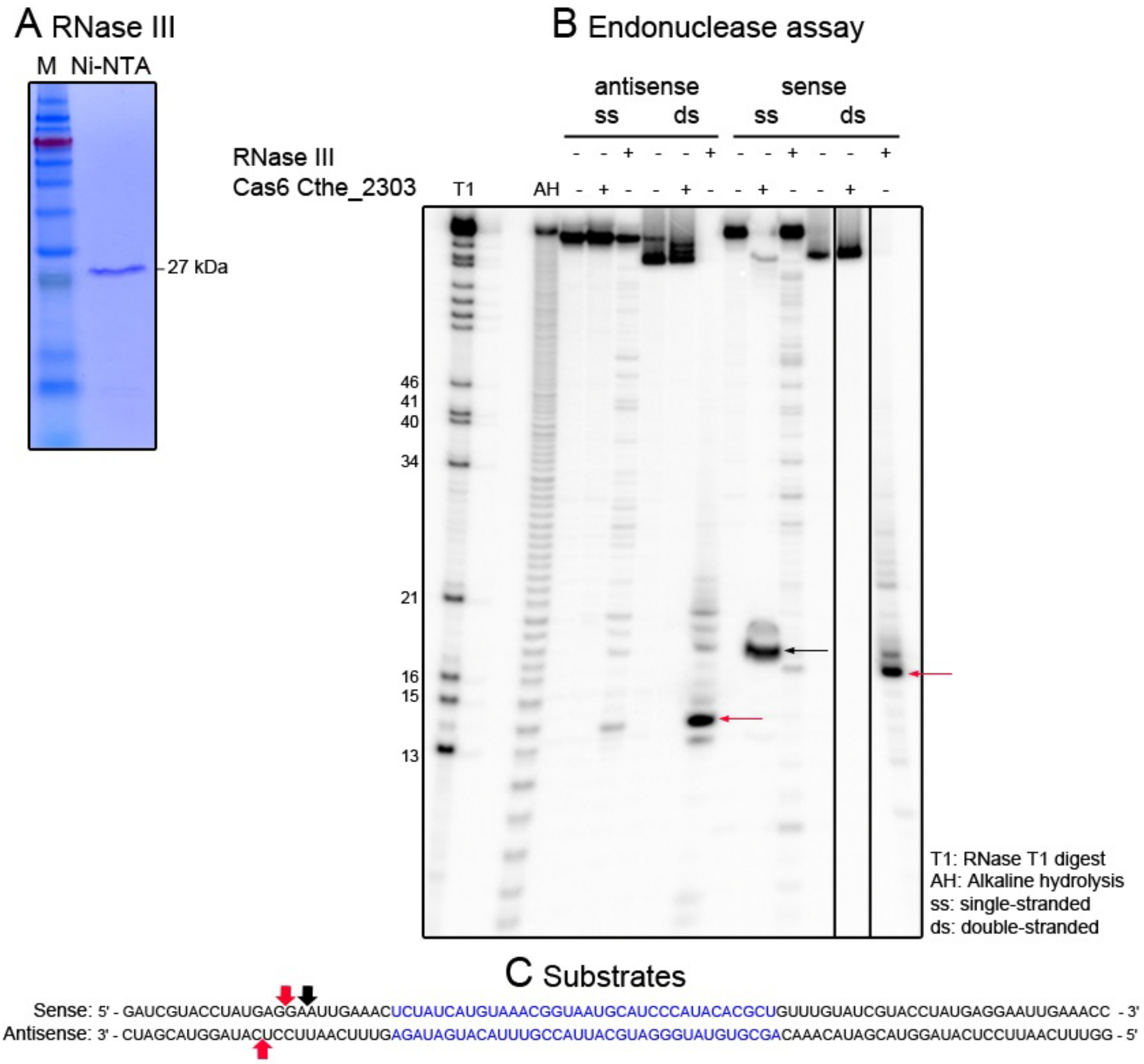

A RNase III
M Ni-NTA
27 kDa
B Endonuclease assay
antisense
sense
ss ds ss ds
RNase III - - + - - + - - + - - +
Cas6 Cthe_2303 T1 AH - + - - + - - + - - + -
46
41
40
34
21
16
15
13
T1: RNase T1 digest
AH: Alkaline hydrolysis
ss: single-stranded
ds: double-stranded
C Substrates
Sense: 5' - GAUCGUACCUAUGAGGAAUUGAAACUCUAUCAUGUAAACGGUAAUGCAUCCCAUACACGCUGUUUGUAUCGUACCUAUGAGGAAUUGAAACC - 3'
Antisense: 3' - CUAGCAUGGAUACUCCUUAACUUUGAGAUAGUACAUUUGCCAUUACGUAGGGUAUGUGCGACAAACAUAGCAUGGAUACUCCUUAACUUUGG - 5'

Figure 2.7: RNase III purification and endonuclease assay with anti-/crRNA hybrids. A) SDS-PAGE of RNase III after nickel-NTA purification. B) 5'-terminal labeled repeat-spacer-repeat substrates, both single-stranded (ss) and double-stranded (ds) in sense and reverse direction were used for an endonuclease assay with Cas6 (Cthe_2303) and RNase III. Endonucleolytic activity of Cas6 could be detected for the ss sense RNA substrate. RNase III cleavage was observed on both strands of the RNA hybrid. Bands were separated on a denaturing 8 M urea 12% PAA gel and visualized by autoradiography. C) RNA substrates in sense and antisense direction. Arrows denote specific enzymatic cleavage (red arrows = RNase III cleavage, black arrow = Cas6 cleavage).

As described in section 2.2.2, the CRISPR loci of *C. thermocellum* contain two spacer sequences that are complementary and identical with parts of the host tRNA^{Pro} (fig 2.5). The two crRNAs containing these spacers could form double-stranded RNA in the cell to prevent hybridization between the tRNA and the complementary crRNA, which would disturb translation in the cell. Either way, double-stranded RNA would be formed that could be a target for ds-RNA specific cleaving enzymes. To test whether RNase III could play a role in this scenario, an endonuclease assay was performed using RNA-hybrids that comprise the spacer sequence that is complementary to the tRNA (fig 2.5, spacer 19) in forward and reverse direction, resembling the tRNA-identical and tRNA-complementary parts (fig 2.8 A).

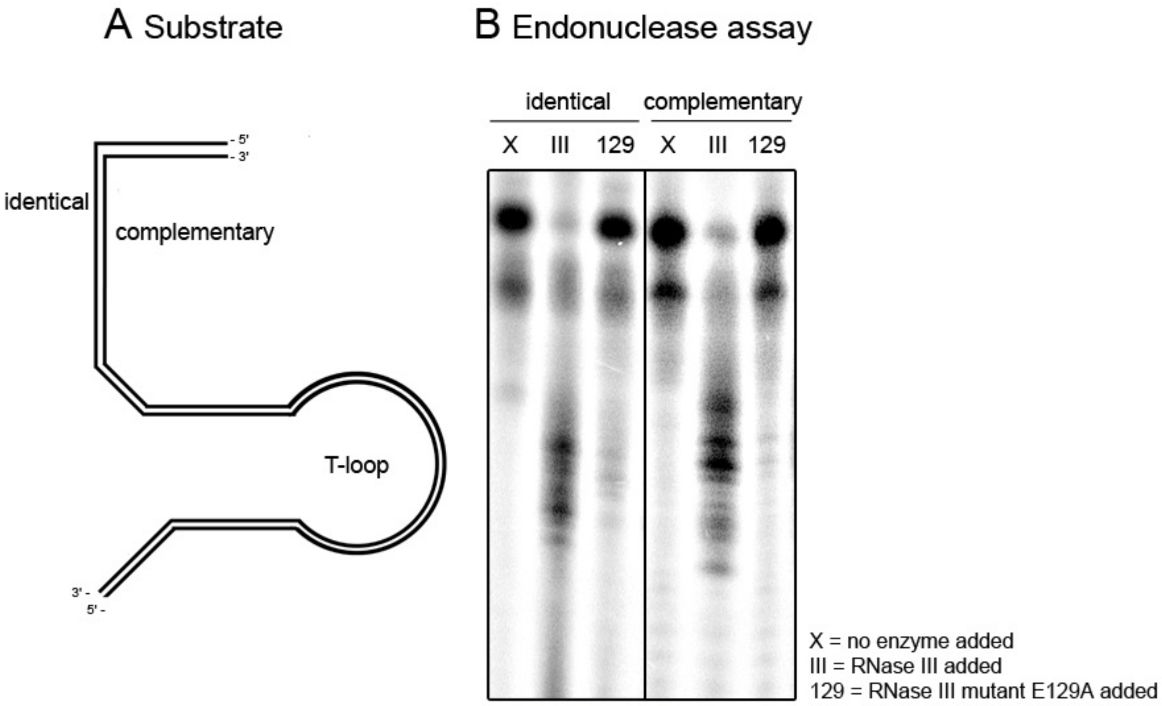


Figure 2.8: RNase III endonuclease assay with an RNA hybrid that resembles parts of the tRNA^{Pro}. A) A double-stranded RNA that comprises spacer 19 of CRISPR locus 2 in forward and reverse direction was used as a substrate. The RNA hybrid resembles a region that is identical to the tRNA^{Pro} isoacceptor in *C. thermocellum* and its complementary counterpart. B) Endonuclease assay with 5'-terminal labeled dsRNA using RNase III (III) and a cleavage-deficient RNase III mutant (E129A). Endonucleolytic cleavage by RNase III can be observed on both strands (identical and complementary, III), whereas cleavage is abolished by the site directed-mutagenesis E129 to A. Bands were separated on a denaturing 8 M urea 12% PAA gel and visualized by autoradiography.

It has been shown for the *E. coli* RNase III that the glutamic acid E117 is a highly conserved and a functionally essential residue of the active site. It is coordinated with divalent metal ions (Mg^{2+}) that participate in the hydrolysis of the RNA phosphodiester [106, 107]. Site-directed mutagenesis of this residue abolishes cleavage activity, probably due to the disruption of metal binding in the active site [108]. An alignment of the RNase III amino acid sequences from *E. coli* and *C. thermocellum* identified the glutamic acid E129 from *C. thermocellum* as the essential active-site residue. A mutation of E129 to alanine was introduced into RNase III to generate a catalytically inactive mutant (E129A). This mutant was also used in the cleavage assay and shown to be deficient of cleavage activity. RNase III cleavage activity could be observed on both strands, the tRNA-identical and complementary one (fig 2.8 B). This indicates that both i) the tRNA-crRNA hybrid as well as ii) the crRNA-crRNA hybrid could be targeted by RNase III.

2.3 RNase III substrate recognition of CRISPR repeat elements

Structure and sequence elements, so-called reactivity epitopes, determine the cleavage sites of bacterial RNase III [80, 109]. Two double-helical segments termed proximal-box (pb) and distal-box (db) that represent sites of enzyme-substrate contact participate in controlling substrate reactivity. It was proposed that in *E.coli*, RNase III substrate recognition is based on the absence of specific base pair elements (antideterminants) that inhibit substrate cleavage reactivity, limiting RNase III processing to a single target site [110]. In addition, specific base pair sequences have been identified that act as positive recognition determinants and cooperate with the antideterminants to control enzyme reactivity. The db is a 2 bp element positioned 11 bp away from the RNase III cleavage site. The db functions as a positive recognition determinant and base pair substitutions weakens the RNase III binding. Its sequence is not conserved. The pb is a 4 bp segment. Base pair substitutions at positions 1 and 3 reduce RNase III binding affinity. A strong inhibitory effect on binding can be observed for position 2 by a base pair substitution of AU, UA against GC or CG. Base pair substitution at position 4 causes a qualitatively different inhibition of RNase III reactivity. Here, the substitution of a GC or CG base pair suppresses cleavage without affecting substrate binding of the *E. coli* RNaseIII and therefore function as catalytic antideterminants [94].

As described in 2.2.4 *C. thermocellum* contains two types of repeat elements (30 bp and 37 bp) that can be processed by RNase III, once they form anti-/crRNA hybrids (*in vivo* fig 2.6. and *in vitr*o fig 2.7). It was possible to identify two potential db/pb sets within each double-stranded repeat sequences as they can be positioned relative to the RNase III cleavage site on both strands (fig. 2.9). Both potential pb of the 37 bp repeat duplex are very similar to an optimal *E. coli* pb with AU/UA base pairs at position 2 and 4. In contrast, both potential pb in the 30 bp hybrid comprise AU/UA base pairs at position 2 but display GC/CG at position 4 which strongly inhibits *E. coli* RNase III cleavage reactivity. This indicates sequence variations across species.

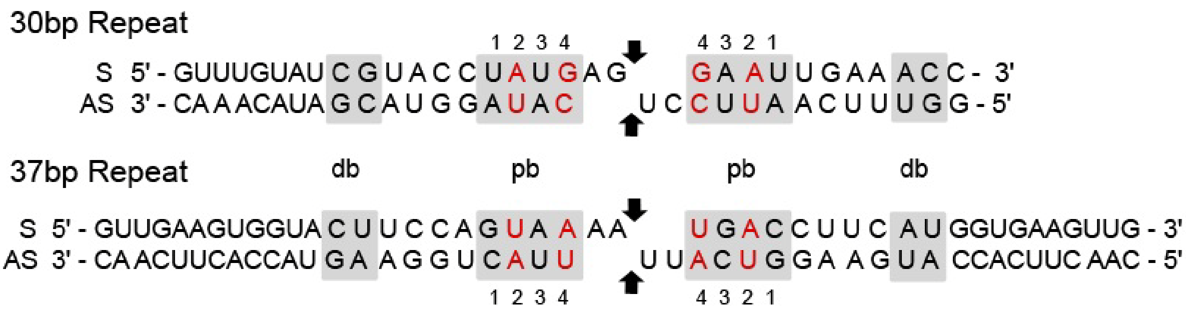

30bp Repeat
1 2 3 4
4 3 2 1
S 5' - GUUUGUAU C G U A C C U A U G A G
G A A U U G A A A C C - 3'
AS 3' - CAAACAUA G C A U G G A U A C
U C C U U A A C U U U G G - 5'
37bp Repeat
db
pb
pb
db
S 5' - GUUGAAGUGGUA C U U C C A G U A A A A
U G A C C U U C A U GGUGAAGUUG - 3'
AS 3' - CAACUUCACCAU G A A G G U C A U U
U U A C U G G A A G U A CCACUUC AAC - 5'
1 2 3 4
4 3 2 1

Figure 2.9: Potential RNase III substrate recognition motifs in double-stranded 30 bp and 37 bp repeat elements of *C. thermocellum*. Both ds-repeat sequences contain two potential proximal (pb) and distal (db) boxes that are depicted in grey, adjacent to the RNase III cleavage site indicated with black arrows. Within the proximal boxes, essential base pairs at position 2 and 4 that influence *E. coli* RNase III reactivity are marked in red.

In order to identify the essential set of distal and proximal boxes and investigate the sequence requirements for RNase III processing reactivity in *C. thermocellum*, RNA substrates were designed comprising the 30 nt repeat element in forward and reverse direction that differ in the nucleotide composition of the respective regions (fig 2.10 A). These included a wild-type substrate, a dsRNA substrate displaying a base pair substitution at positions 2 and 4 within the left pb and a repeat RNA duplex with substituted base pairs at position 2 and 4 of the right pb. An endonuclease assay was performed with 5′-terminal labeled RNA hybrids and RNase III (fig 2.10 B). With the use of an RNase T1 digest and an alkaline hydrolysis marker, the prominent cleavage products were sized and marked in the respective substrate sequences (fig 2.10 A). Cleavage products were sized according to figure 2.7. The RNase III processing pattern of the wild type substrate is consistent with the cleavage sites observed *in vivo* (fig 2.6). Additional cleavage products (fig 2.10 A) were also observed *in vitro* (fig 2.7 B), indicating variability of the RNase III cleavage site in the 30 bp repeat element. The substitution of the base pairs at position 2 and 4 in both pb causes a modified cleavage pattern. Extra cleavage products that are unique for the altered RNA substrates (fig 2.10 B) were marked in the substrate sequences (fig 2.10 A). The nucleotide exchange in the left pb results in the addition of one unique cleavage site in the forward and reverse strand (fig 2.10 A).Alterations in the nucleotide composition of the right pb cause two additional unique cleavage sites on both strands (fig 2.10 A). The wildtype cleavage products remain, which indicates that both sets of pb and db are recognized by RNase III in *C. thermocellum*.

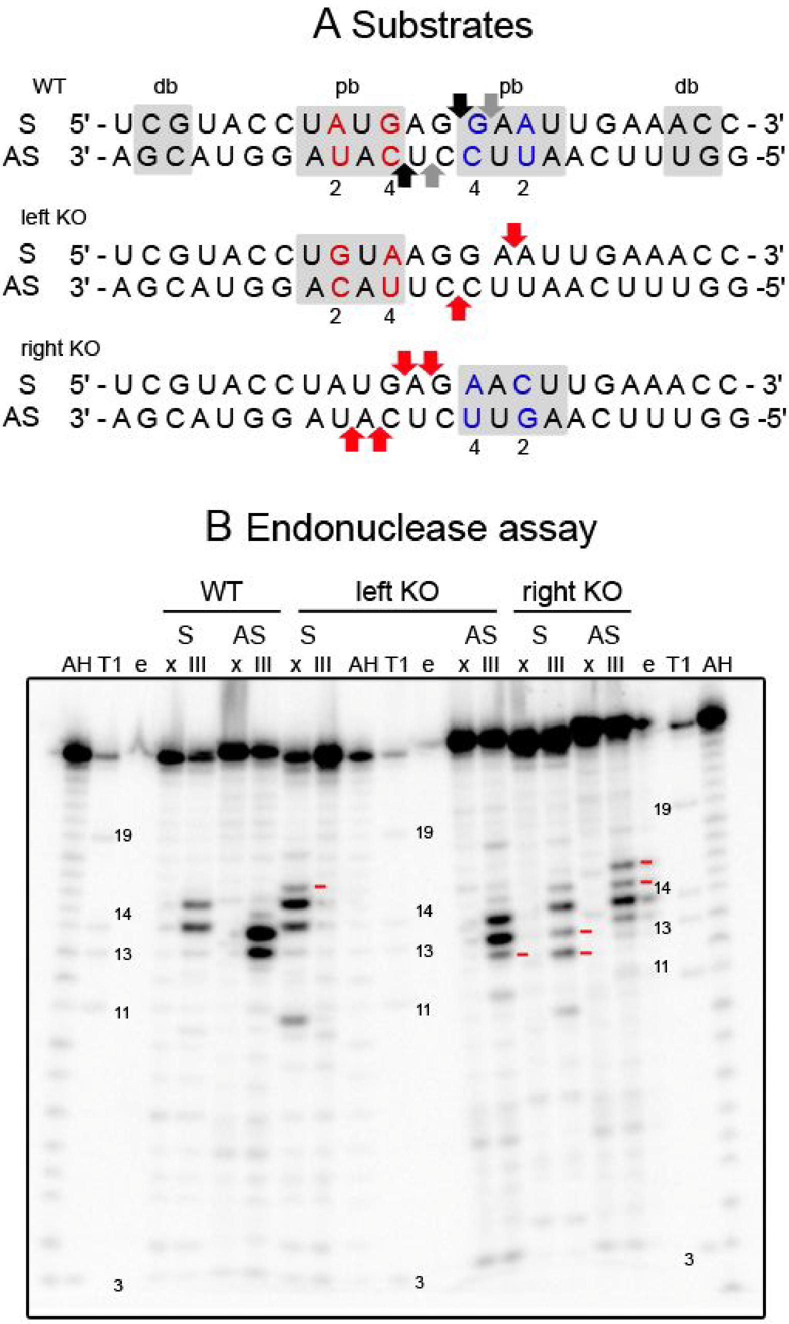

A Substrates
WT
db
pb
pb
db
S 5'-UCGUACCUAUGAG GAAUUGAAACC-3'
AS 3'-AGCAUGG AUACUCCUUAACUUUGG-5'
2 4 4 2
left KO
S 5'-UCGUACCUGUAAGG AAUUGAAACC-3'
AS 3'-AGCAUGG ACAUUCCUUAACUUUGG-5'
2 4
right KO
S 5'-UCGUACCUAUGAG AACUUGAAACC-3'
AS 3'-AGCAUGG AUACUCUUGAACUUUGG-5'
4 2
B Endonuclease assay
WT
left KO
right KO
S
AS
S
AS
S
AS
AH T1 e x III x III x III AH T1 e x III x III x III e T1 AH
19
14
13
11
3
19
14
13
11
3
19
14
13
11
3

Figure 2.10: Identification of essential RNase III recognition elements in double-stranded 30 nt repeat elements. A) Double-stranded RNA substrates comprising the 30 bp repeat (WT) with base pair substitutions at position 2 and 4 within the left proximal box (left KO) and the right proximal box (right KO) were used for RNase III assays. B) Endonuclease assay using the 5'-terminal labeled dsRNA substrates and RNase III from *C. thermocellum*. An RNase T1 digest (T1) and Alkaline hydrolysis (AH) with single-stranded WT substrate in sense direction (S) was used as RNA size markers. Empty lanes are marked with e. The WT substrate displays the predefined (A black arrows) as well as an additional processing sites (A grey arrows). Unique cleavage products were depicted (B red bars) and marked in the RNA sequences (A red arrows). Bands were separated on a denaturing 8 M urea 12% PAA gel and visualized by autoradiography.

2.4 Type I-B Cascade

In *Clostriudium thermocellum*, the five putative Cas proteins Cas3, Cas5, Cas6, Cas7 and Cas8b (Cthe_3201-3205) are proposed to form a type I-B Cascade complex that mediates the antiviral defense. However, the assembly of a functional type I-B complex has not been observed. In order to investigate the assembly and functionality of this CRISPR ribonucleoprotein complex (crRNP), the Cascade protein candidates were produced in *E. coli*. Since the expression of soluble Cas proteins in *E. coli* is often difficult due to insolubility issues, all five genes were cloned in their naturally occurring order into a single expression vector. The "Gateway cloning" technique was utilized for the generation of the expression construct. This method is based on homologous recombination of specific recombination sites that are introduced at the terminal ends of the genes during PCR amplification. This allows a simultaneous assembly of the genes of interest in the order of choice. However, this cloning strategy was not successful, and the "Golden Gate shuffling" cloning method was used to proceed. This technique also allows the simultaneous assembly of the genes of interest in the order of choice, but is based on a type II restriction enzyme digestion and a simultaneous ligation reaction which enables the generation of a construct lacking the original restriction sites [111]. For this approach, all five *cas* gene sequences were supplied with the respective terminal ends, codon-optimized for *E. coli* codon usage and commercially synthesized into pUC57 plasmids. This attempt did not yield any plasmid containing the five *cas* genes. Therefore, five individual expression constructs were generated containing the single *cas* genes (Bachelorthesis of Laura Penkert and Mastermodule of Franka Schreiner). For *cas* gene expression in *E. coli,* the codon-optimized versions of *cas3*, *cas5*, *cas7* and *cas8b* were utilized. A *cas6* codon-optimized expression construct could not be obtained and the native sequence was used for Cas6 protein expression. Toxicity of this endonuclease for *E. coli* cells could be a possible reason for the inability to clone a *cas6* codon-optimized variant into an expression vector and might explain complications during the *cas* gene assembly techniques.

2.4.1 Production of Cas proteins and Cascade assembly

For the recombinant production of all five Cas proteins in *E. coli*, optimal gene expression conditions were tested for each protein using two different expression vectors (Bachelorthesis of Laura Penkert and Mastermodule of Franka Schreiner).

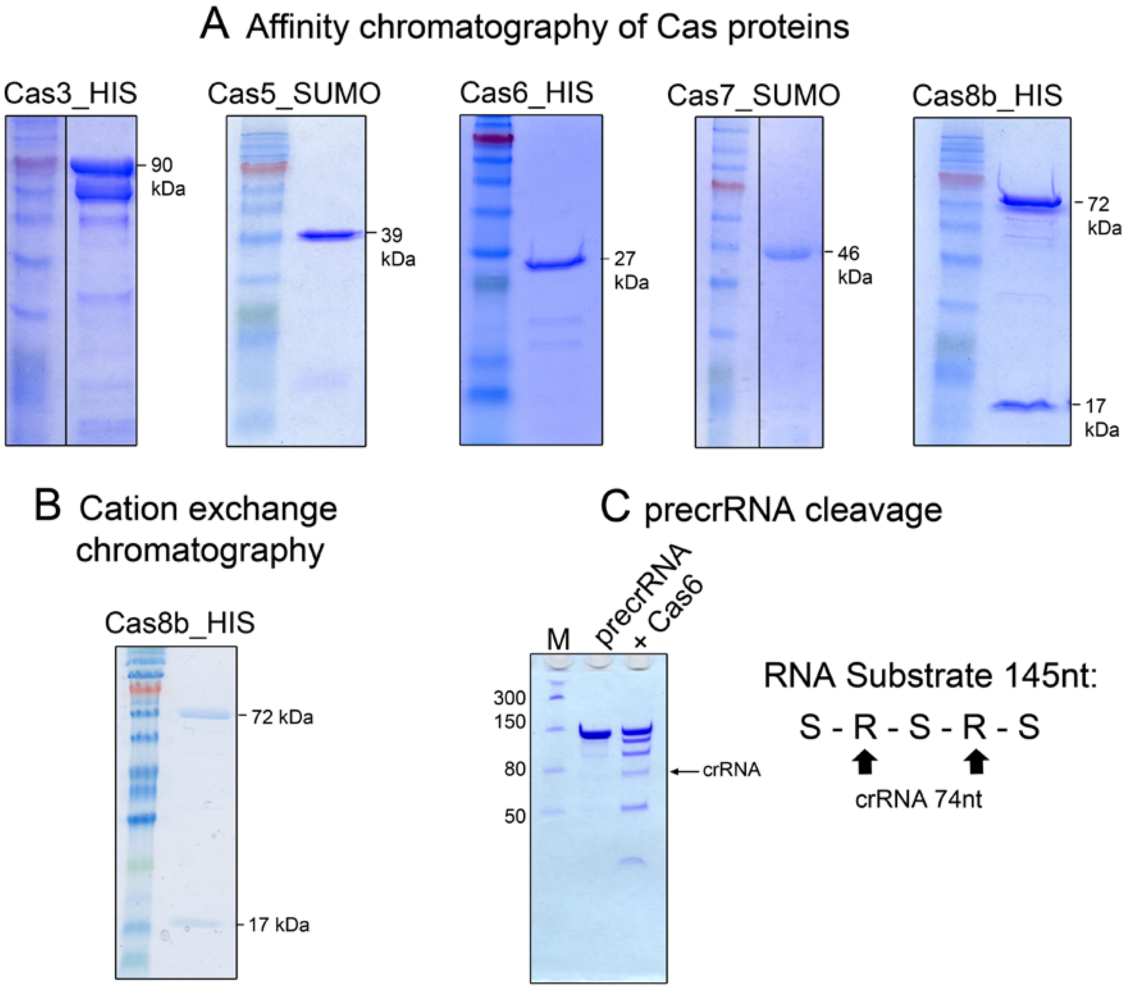


Figure 2.11: Cas protein purification and Cas6 pre-crRNA processing. A) SDS-PAGE of Cas3, Cas5, Cas6, Cas7 and Cas8b after nickel-NTA purification. Cas8b is purified with an additional 17 kDa protein fragment B) SDS-PAGE of Cas8b together with the protein fragment after cation exchange chromatography, used as an additional purification step to remove co-purified nucleic acids. C) Generation of a pre-crRNA comprising 3 spacers (S) that are interspaced by 37 nt repeat elements (R). Cas6 processing (black arrows) yields crRNAs and cleavage intermediates, respectively. Bands were separated on a denaturing 8 M urea 12 % PAA gel. (C) Bachelorthesis of Laura Penkert).

The pET20b vector generates protein with an N-terminal 6-fold His-tag, whereas the pEC-A-HI-SUMO vector provides an N-terminal His-tag and a SUMO protein fusion. This tag can be cleaved off using a SUMO protease which results in tag-free proteins. To yield sufficient amounts of Cas3, Cas6 and Cas8b, these proteins were produced from pET20b vectors. The individual SUMO protein fusions of these three proteins either increased the amount of protein impurities or rendered the proteins inaccessible for the SUMO protease. However, Cas5 and Cas7 expression was found to rely on the fusion to the SUMO protein, as it enhances protein stability. All five proteins were purified via affinity chromatography using a nickel-NTA column. Afterwards, the SUMO tags of Cas5 and Cas7 were removed (fig 2.11 A). Cas8b is purified with an additional 17 kDa protein fragment that comprises the C-terminal sequence of Cas8b which was verified by mass-spectrometry (Bachelorthesis of Laura Penkert). The Cas8b protein will be described in detail in section 2.4.4. Agarose gel-electrophoresis of all individual Cas proteins revealed that Cas8b is highly contaminated with nucleic acids. Hence, cation exchange chromatography with a Heparin column was used to remove the co-purified nucleic acids (fig 2.11 B). Finally, a pre-crRNA substrate containing three spacer sequences interspaced by 37 nt repeat elements was generated (fig 2.11 C). The RNA substrate was cleaved by Cas6 which resulted in the generation of one mature crRNA (74 nt) and processing intermediates.

In order to reconstitute the Cascade complex, the recombinant Cas proteins Cas5, Cas7, Cas8b and Cas6 with RNA cleavage products (fig 2.11) were mixed, incubated at 50°C and subjected to size-exclusion chromatography on a Superdex 200 column. Cas3 was not added to the mixture as it was shown that it is not tightly associated with Cascade during complex assembly (Mastermodule of Franka Schreiner), which was also observed for the type I-E Cascade [49]. 1 mL chromatography fractions from an elution volume of 6,6 to 13,6 mL (A7 to B12) were analyzed with respect to protein and RNA content via TCA-/ethanol precipitation and subsequent SDS-/urea-PAGE. Representative results are shown in fig 2.12 A. To generate a standard curve, a mix of protein standards (with known molecular weights) was subjected to gel-filtration chromatography under identical running conditions. With the logarithmic illustration of the molecular weights in relation to the elution volumes, an equation is obtained (e-function) to calculate the molecular weight of any molecule eluting from the gel-filtration column at a given volume. According to this, an elution volume of 8.6 mL to 10.6 mL corresponds to a molecular weight of 416 kDa to 194 kDa that is collected in the chromatography fractions A9 and A10. The assumed molecular weight of the type I-B Cascade (described in 2.4.6) lies within this range and a distinct protein peak can be observed eluting at around 9 mL (fig 2.12 A).

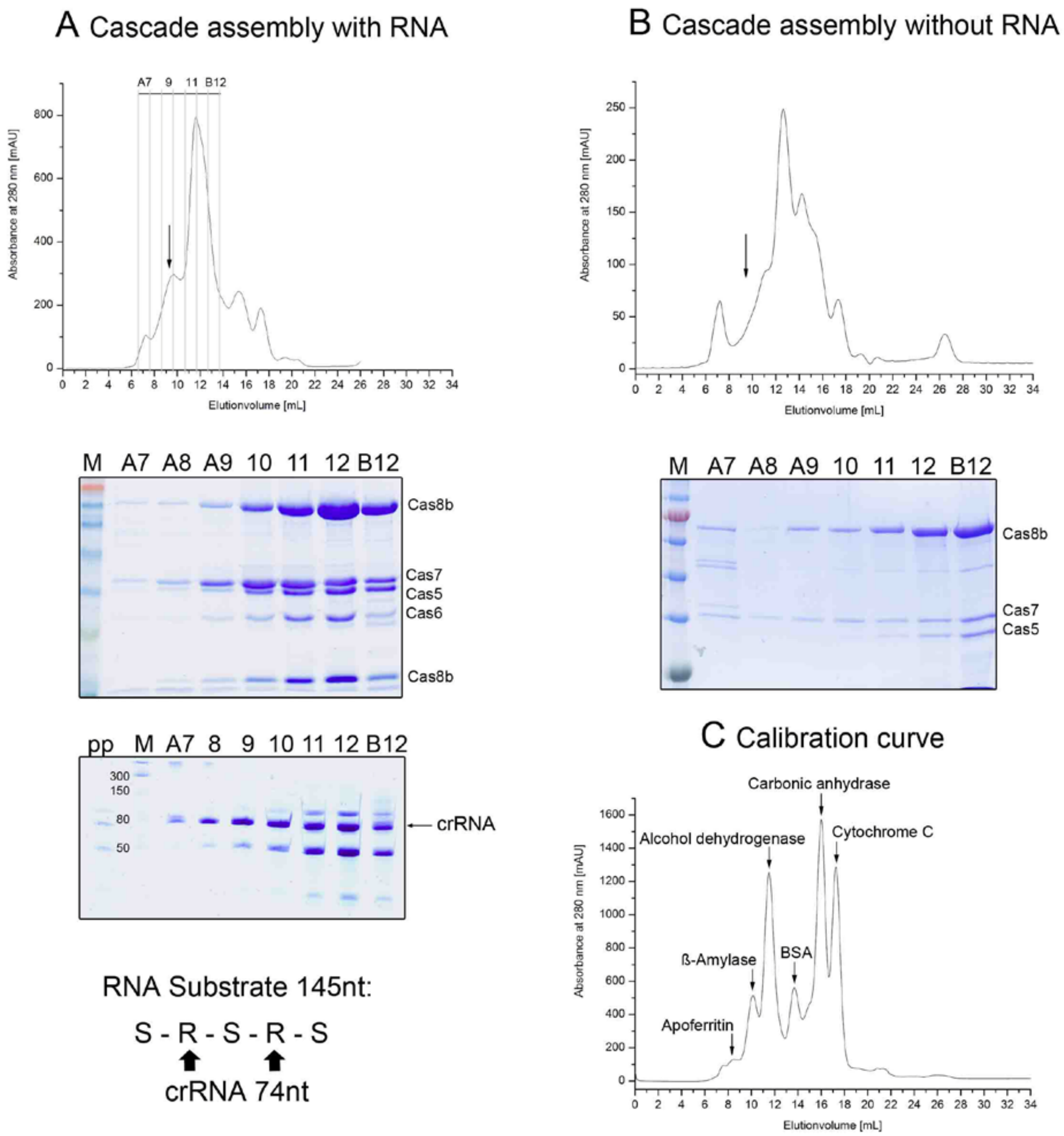


Figure 2.12: Cascade assembly with and without RNA. A) Cas proteins and processed pre-crRNA (pp) were mixed and subjected to gel-filtration chromatography. Fractions A7 to B12 (marked in the elution profile) were analyzed via TCA-/ethanol precipitation and subsequent SDS-/urea-PAGE. A protein peak around 9 mL that corresponds to the fractions A9 and A10 is marked with an arrow. The SDS- and urea-PAGE show a complex formed by Cas8b and the protein fragment, Cas7, Cas6, Cas5 and mature crRNA that elute in the respective fractions. B) Cas proteins were mixed without RNA and subjected to gel-filtration chromatography. Fractions A7 to B12 (6.6-13.6 mL) were analyzed via TCA-/ethanol precipitation and subsequent SDS-/urea-PAGE. The elution profile does not display a protein peak at an elution volume between 8.6-10.6 mL and the corresponding fractions A9 and A10. C) Gel-filtration chromatogram of a standard protein mixture that was used as a calibration curve for the calculation of a standard curve.

The corresponding fractions A9 and A10 contain the Cas proteins Cas8b, Cas7, Cas6 and Cas5 that co-elute with mature crRNA. An overrepresentation of the Cas7 protein can be observed, which is in agreement with the stoichiometry that was determined via mass-spectrometry described in section 2.4. A great protein peak can be observed eluting at around 11 to 13.5 mL that corresponds to the fractions A11 to B12. Gel-filtration chromatography of the individual Cas proteins showed that Cas8b elutes as a monomer at around 12 mL and that Cas7 forms multimers that elute between 11 and 13 mL. Thus, the peak represents unassembled Cas8b protein as well as Cas7 multimers that have been observed to interact with Cas5. Monomeric Cas7 protein elutes around 15 mL followed by monomeric Cas5 protein at 17 mL (Mastermodule of Franka Schreiner).

The next approach was to reconstitute the Cas proteins without adding RNA to the mixture to analyze whether RNA is required for proper complex formation. Representative results are shown in fig. 2.12 B. However, this approach did not yield any assembled complex. The protein peak resembling the complex is missing in the gel-filtration elution profile and SDS-PAGE of fractions A7 to B12 only revealed mainly unassembled protein. An increased amount of protein elutes at the void volume in fraction A7 that contains Cas7 multimers that interact with Cas8b. Compared to the assembly with crRNA, the amount of monomeric Cas7 protein eluting around 15 mL is increased.

In order to investigate whether Cas6 is required for Cascade formation, the recombinant proteins Cas8b, Cas7, Cas5 and pre-crRNA cleavage products were assembled without Cas6. A precursor-RNA was used that comprised a 3'-terminal repeat sequence that was shortened to 4 nt, as observed for the trimmed crRNAs *in vivo*. After Cas6 cleavage, the RNA products were phenol:chloroform extracted to efficiently remove the protein from the sample. After protein assembly, gel-filtration chromatography was performed and representative results are shown in figure 2.13. A distinct protein peak can be observed eluting around 10 mL and the corresponding fractions A9 and A10 show a stable complex formed by the subunits Cas5, Cas7, Cas8b and crRNA. Hence, Cas6 seems to be a temporarily associated Cascade subunit that dissociates after crRNA delivery.

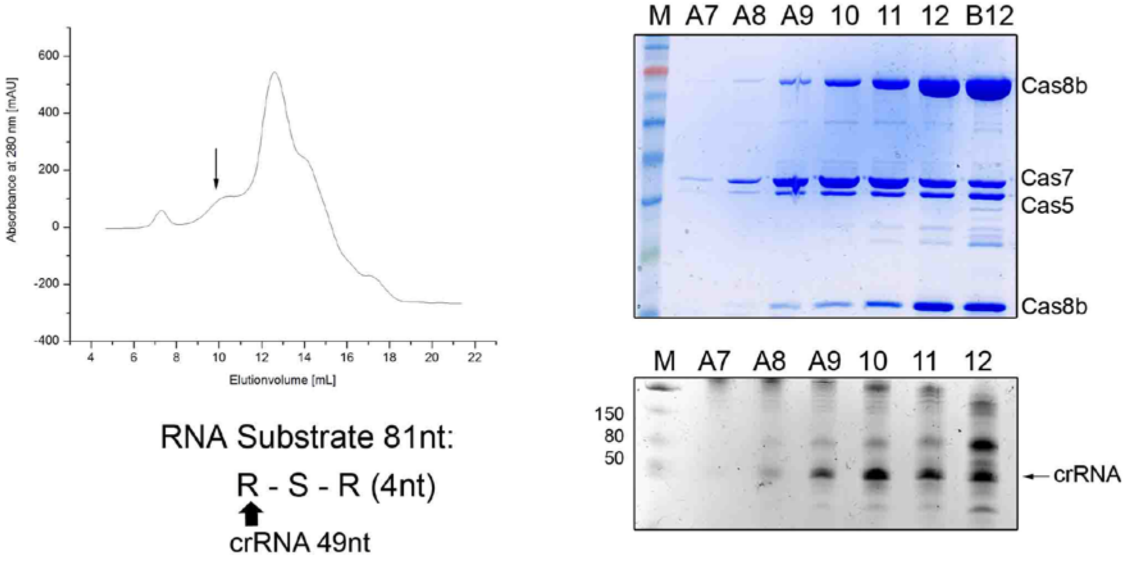


Figure 2.13: Cascade assembly without Cas6. Cas proteins and processed pre-crRNA with a shortened 3' end (4 nt) were mixed and subjected to gel-filtration chromatography. Fractions A7 to B12 that correspond to 6.8-13.8 mL elution volume were analyzed via TCA-/ethanol precipitation and subsequent SDS-/urea-PAGE. A protein peak around 10 mL (arrow) that is collected in the fractions A9 and A10 contains Cas8b, Cas7 and Cas5 that coelute with crRNA.

2.4.2 crRNA binding by Cascade

To prove that the crRNA is bound by Cascade, an Electrophoretic Mobility Shift Assay (EMSA) was performed (fig 2.14). Therefore, a 5′-terminal radiolabeled crRNA was incubated with different concentrations of the Cascade complex. The crRNA-protein complexes were separated from free crRNA on a native 6 % PAA gel (fig 2.14). The EMSA shows a crRNA mobility shift, representing an RNA:protein complex. The shift starts to appear at a protein concentration of 250 nM which is indicated by the reduction of free crRNA. A large excess of yeast RNA (~200 fold) was used in each reaction as a competitor, showing that the crRNA binding by Cascade is specific. Noteworthy, the radiolabeled crRNA is bound by a Cascade that was assembled in the presence of unlabeled crRNA as this is needed for proper complex formation. One plausible explanation is that the used Cascade sample contains a certain amount of unassembled Cascade subunits that form a crRNP complex with the added radiolabeled crRNA. It could also be an indication for “crRNA cycling” which would mean that the crRNA bound by Cascade can be replaced by a radiolabeled one.

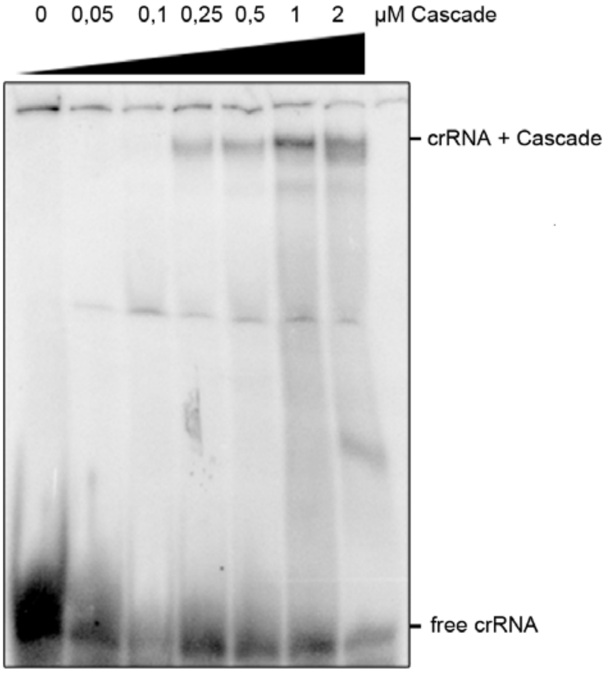


Figure 2.14: EMSA using crRNA and Cascade. Different concentrations (0-2 µM) of assembled I-B Cascade (without Cas6) were used for an EMSA with 5'-terminal radiolabeled crRNA (1,6 nM). A mobility shift can be observed starting at a 250 nM protein concentration which indicates RNA-protein complex formation. Bands were separated on a native 6 % PAA gel and visualized by autoradiography.

2.4.3 Cascade loading

Recently, the first crystal structure of a type I Cascade complex was obtained for the subtype I-E Cascade from *E. coli* [54, 60]. The backbone structure displays one copy of Cas5, capping the 5′-terminal end of the crRNA, six copies of Cas7 that oligomerize along the crRNA and form the backbone of the complex and one permanently associated Cas6 subunit sitting at the 3′-terminal end of the crRNA. In order to investigate crRNA loading of the type I-B complex, Cascade assemblies were performed using the purified Cas proteins Cas8b, Cas7, Cas5, Cas6 and pre-crRNA variants with modified repeat tags to analyze their role and identify their interaction partners in the complex. Two pre-crRNA variants were designed in a way that mature crRNAs were generated with modified repeat tags at both ends (fig 2.15).

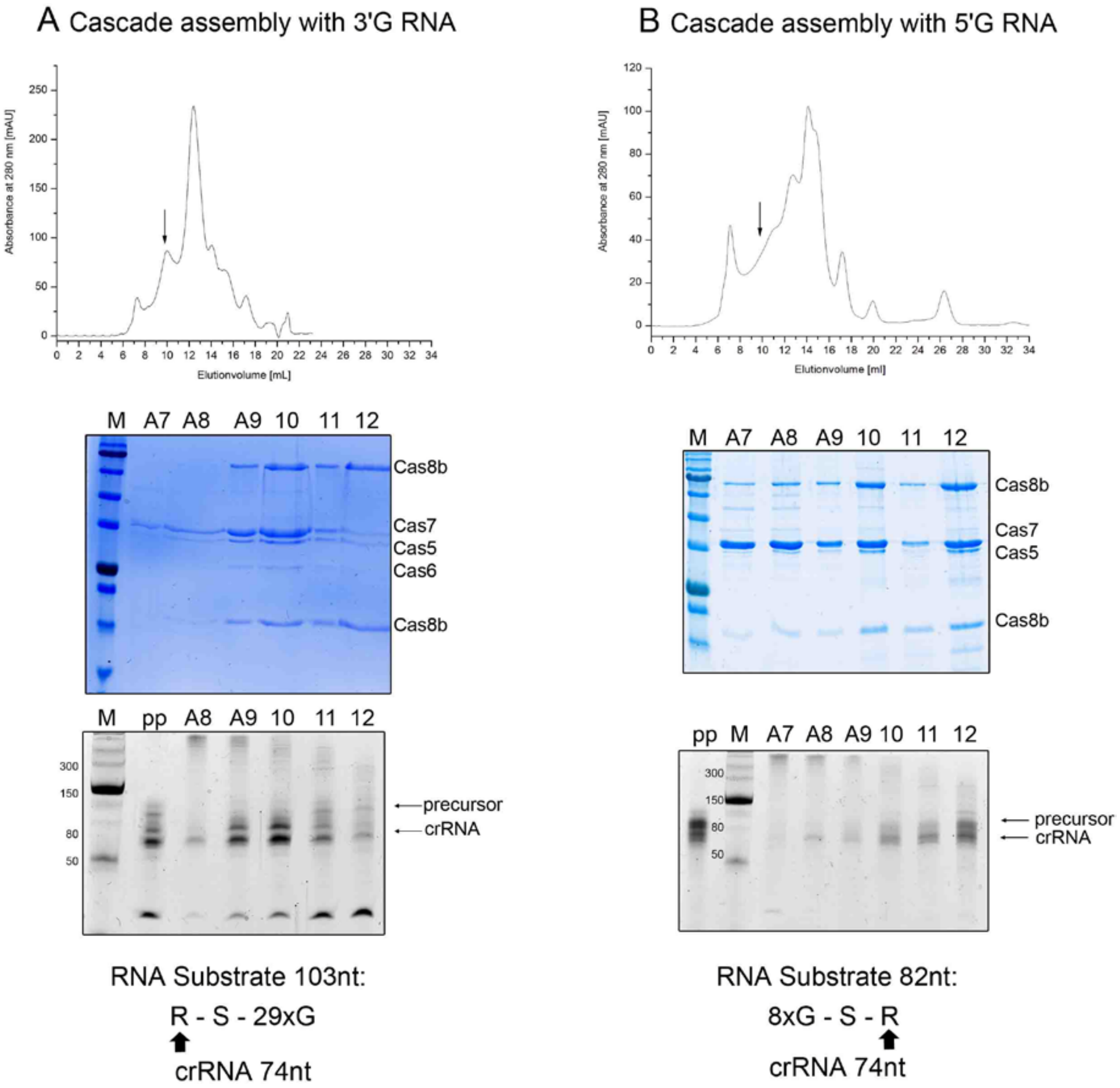


Figure 2.15: Cascade assembly with modified crRNAs. Cas proteins and processed pre-crRNA variants (pp) were mixed and subjected to gel-filtration chromatography. Fractions A7-A12 that correspond to 6.6-13.6 mL elution volume were analyzed by TCA/ethanol precipitation and SDS-/urea-PAGE. A) Gel-filtration elution profile of Cascade assembly with 3'-terminal modified crRNA (3'G RNA). The black arrow indicates the protein peak that corresponds to fractions A9 and A10 that contain assembled Cascade complex. B) The elution profile from gel-filtration purification of the Cascade assembly with 5'-terminal modified crRNA (5'G RNA). The black arrow indicates the region corresponding to fractions A9 and A10 where assembled complex would elute.

One pre-crRNA variant (fig 2.15 A) comprised a 37 nt repeat element and a spacer sequences followed by 29 G residues (R-S-29xG). Cas6 cleavage of this pre-crRNA produced a crRNA (3'G RNA) that contained an intact 5′-terminal tag and a modified 3′-end with 29 G residues instead of 29 repeat-derived nucleotides. In order to generate the pre-crRNA, *in vitro* run-off transcription was utilized that yielded transcripts of varying length. Analysis of the RNA production showed the formation of a double-band, presumably due to the 29 consecutive G residues that are difficult to transcribe for the T7 RNA polymerase. Hence, Cas6 processing of the precursor also generated crRNAs of different length in the form of a double band (fig. 2.15 A). The gel-filtration purification chromatogram of the Cascade assembly using the 3′G RNA displays a protein peak at an elution volume around 10 mL. The corresponding fractions A9 and A10 display all Cascade subunits. The derived pattern and the intensities of the individual bands are comparable to a Cascade assembly with wildtype crRNA (fig 2.12 A). The mature 3′G RNA is enriched in the respective fractions. Taken together, these results show the formation of a protein-RNA complex, indicating that the modification of the 3′-terminal repeat sequence of the crRNA does not interfere with crRNA loading or the assembly of Cas proteins.

The second pre-crRNA variant used for Cascade assembly contained 8 G residues followed by a spacer element and a 37 nt repeat sequence (8xG-S-R). Cas6 processing of the precursor generated a crRNA (5′G RNA) consisting of a modified 8 nt 5′-terminal tag and an unchanged 3′-end. It should be mentioned that the 5′G RNA contains a 5'-terminal triphosphate resulting from the generation via run-off transcription and therefore differs from the natural crRNAs that are processed by Cas6 and encompass a 5'-terminal hydroxyl group. After assembling the Cas proteins with this crRNA variant, gel-filtration purification was performed and representative results are shown in fig 2.15 B. The chromatogram displays a gradual increase instead of a distinct peak of protein between 8.6-10.6 mL elution volume. After TCA/ethanol precipitation and subsequent SDS-/UREA-PAGE, the respective fractions A9 and A10 contained Cas8b with the protein fragment, Cas7, Cas5 and mature 5′G RNA. The amount of Cas6 protein seems to be highly reduced indicating that the modification of the 5′-tag interferes with its complex association. The relative amount of Cas5 protein is slightly reduced, which suggests that the modification of the 5′-tag might reduce its ability to bind the crRNA. This would be in agreement with the observation made for the I-E Cascade, where Cas5 caps the 5′-terminal end of the crRNA. The fact that a significant amount of protein eluted within the void volume (elution volume around 7 mL, fractions A7 and A8) could indicate difficulties of the Cas proteins to assemble around the 5′G RNA.

As described in section 2.2.2, mature crRNAs display a gradually processed 3'-ends *in vivo*. To investigate whether the trimming of the 3'-terminal repeat sequence affects Cas6 loading ability, an assembly was performed with pre-crRNAs of different length (fig 2.16). One pre-crRNA substrate consisted of a 145 nt long spacer-repeat-spacer-repeat-spacer sequence that can be used to obtain an 74 nt long crRNA and cleavage intermediates via Cas6 processing and was also used for the wildtype assembly in fig 2.12. Another pre-crRNA substrate of 81 nt was used that consisted of the same spacer sequence, flanked by an intact 5'-terminal repeat and a 3'-terminal repeat shortened to 4 nt as observed for most of the crRNAs containing this spacer *in vivo*. Cas6 cleavage of this modified pre-crRNA generates 49 nt long crRNAs that were shown to assemble with Cascade (figure 2.13).

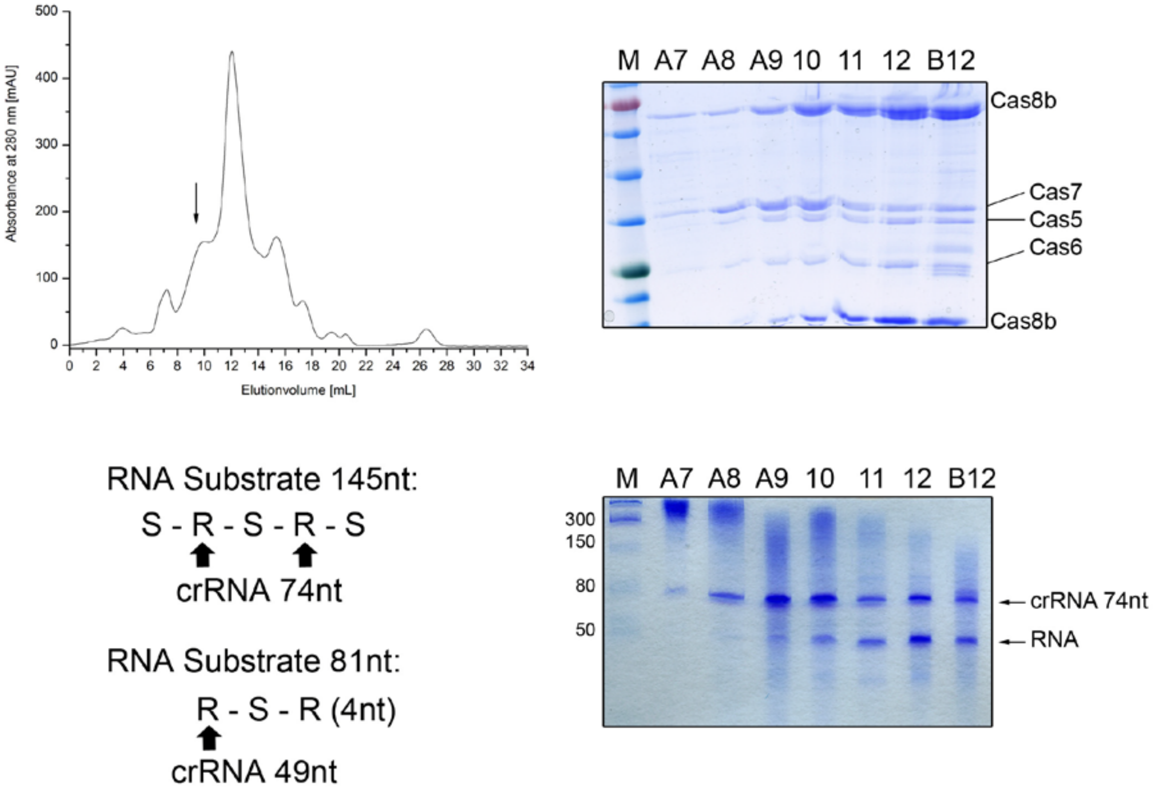


Figure 2.16: Cascade assembly with crRNAs of varying length. Equimolar ratios of two pre-crRNA variants (145 nt and 81 nt) were used to generate mature crRNAs via Cas6 cleavage that comprise either a full-length or shortened 4 nt 3'-terminal repeat ends. Processed crRNAs were mixed with Cas proteins and subjected to size-exclusion chromatography. The black arrow in the elution profile indicates the protein peak that corresponds to fractions A9 and A10 containing assembled complex. SDS- and urea-PAGE show the complex formed by Cas8b, Cas7, Cas6 and Cas5 in the respective fractions A9 and A10.

Equal molar amounts of both pre-crRNA transcripts were simultaneously processed using Cas6, assembled with Cascade and subsequently subjected to gel-filtration chromatography. The elution profile shows a protein peak at around 9 mL that corresponds to size of a fully assembled complex. SDS-/UREA PAGE of the respective fractions A9 and A10 displays all four Cas proteins and the crRNA of 74 nt. Additionally, RNA of around 50 nt can be observed that could correspond to either the 49 nt crRNA or an intermediate cleavage product (48 nt) that results from Cas6 cleavage of the 145 nt long pre-crRNA as observed in the wildtype assembly (fig 2.12). Nevertheless, the crRNA 49 that contains the shortened 3'-terminal repeat tag is less efficiently assembled with Cascade. This indicates that crRNAs comprising a full length 3'-terminal tag are preferably incorporated by the complex.

As described in section 2.1, *C. thermocellum* encodes two type I-B CRISPR systems (Hmari and Tneap) that are associated with CRISPR loci of 30 nt and 37 nt repeat elements. In section 2.2.1, it was shown that two Cas6 enzymes exclusively process their respective repeat substrates and generate crRNAs with identical 8 nt 5′-teminal tags. To investigate whether the studied Cascade (type I-B Hmari) could be loaded with a crRNA and associate with the Cas6 subunit of the other system (type I-B Tneap), recombinant Cascade assembly was performed using a pre-crRNA containing three spacer sequences that are interspaced by two 30 nt repeat elements that were processed by Cas6 Cthe_2303 (Tneap) shown in fig 2.17. The elution profile of the gel-filtration purification revealed a peak “shoulder” at the respective elution volume, indicating the potential formation of a complex. The SDS-/UREA PAGE analysis of the precipitated fraction A7 to A12 reveals the presence of all utilized Cas proteins and mature crRNA in fractions A9 and A10. The fact that Cas6_2303 can be observed in fractions A9 and A10 indicates that the Tneap Cas6 subunit interacts with the Hmari Cascade. The significant amount of protein in the void volume shown in fractions A7 and A8, as well as the elution profile that lacks the distinct complex protein peak and the small amount of crRNA that can be detected could be an indication of hindered crRNA loading conditions though.

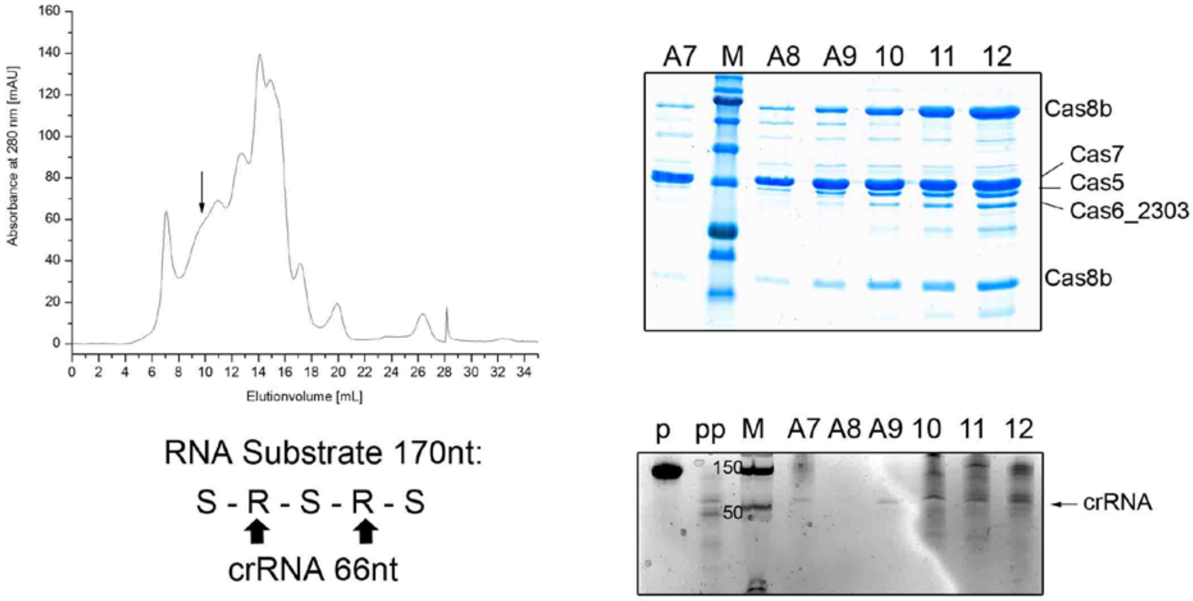


Figure 2.17: Cascade assembly using Cas6 of type I-B Tneap (Cthe_2303). Cas proteins Cas8b, Cas7 and Cas5 (type I-B Hmari) and pre-crRNA (p) processed by Cas6 of type I-B Tneap (pp) were mixed and subjected to gel-filtration chromatography. The black arrow in the chromatogram indicates the region, where assembled complex typically elutes. Fractions A7-A12 (elution volume 6.6-12.6 mL) were analyzed via TCA/ethanol precipitation and SDS-/urea-PAGE.

2.4.4 Characterization of Cas8b

The recombinant production of Cas8b and subsequent affinity chromatography using nickel-NTA revealed that soluble protein was co-purified with a small fragment of 17 kDa (section 2.4.1). This small protein fragment could not be removed via cation exchange chromatography or gel-filtration chromatography. When subjected to gel-filtration chromatography, both proteins co-elute at 12.5 mL elution volume which corresponds to a molecular weight of 90 kDa. A mass-spectrometric analysis identified the 72 kDa as the full-length Cas8b protein and the 17 kDa protein to be a C-terminal fragment of Cas8b (Kundan Sharma, MPI Göttingen).

Recombinant production and subsequent purification of the Cas8b protein (76 kDa) from *Methanococcus maripaludis* also yields the additional 17 kDa C-terminal fragment of Cas8b (fig 2.18 A), even though the proteins share only 17% amino acid homology. N-terminal Edman sequencing of 10 amino acids of the small fragment from *M. maripaludis* identified the amino acid sequence to start with a methionine residue at position 511 of the full-length protein (PhD thesis of Hagen Richter). An amino acid sequence alignment with

Cas8b of *C. thermocellum* identified this methionine residue at position 476 of the full-length protein as the potential first amino acid of the C-terminal fragment. Accordingly, the theoretical molecular weight of the C-teminal fragment would be 17.6 kDa. Mass-spectrometric analysis using MALDI-TOF (Matrix_assisted_laser_desoption_ionization–time_of_flight) of purified Cas8b revealed a molecular weight of 17.5 kDa that corresponds to the fragment (fig 2.18 B), confirming the start site of the amino acid sequences covering the C-terminal part of the full-length protein. Two masses of 53.1 kDa and 70.9 kDa could also be detected with very low relative intensities. These masses could correspond to the full-length protein (theoretical Mw: 72.6k Da) and a potential protein fragment covering the remaining N-terminal sequence (theoretical Mw: 54.9 kDa). The mass aberration of around 2 kDa is remarkable, but could be explained with the limited ability of the MALDI-TOF instrument to ionize large proteins that causes a significant decrease in the accuracy of mass determination of large molecules.

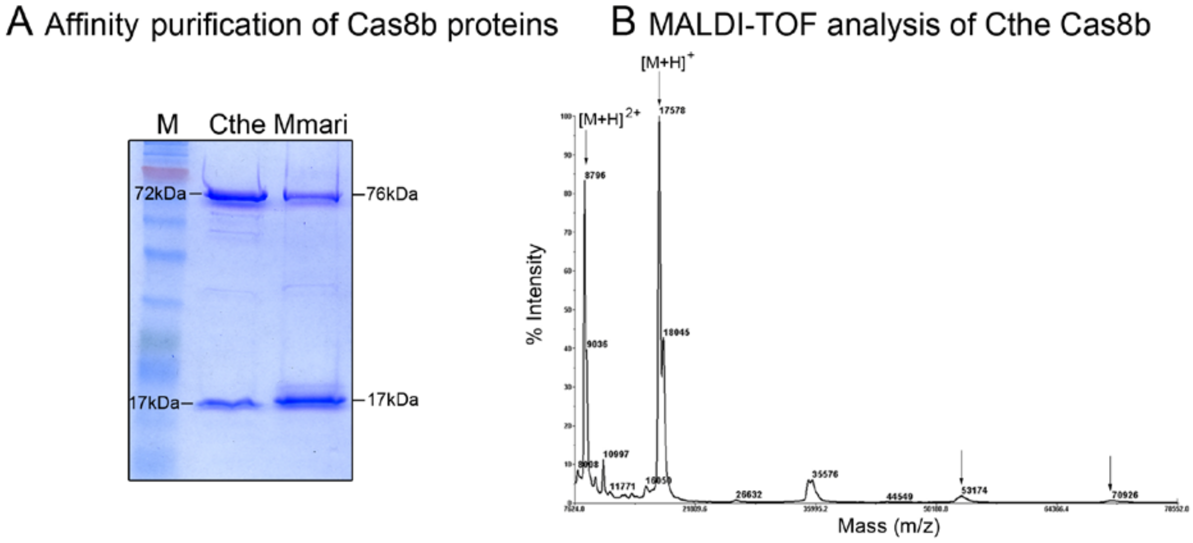


Figure 2.18: Recombinant production of Cas8b subunits and identification of the small protein fragment. A) Affinity purification of Cas8b proteins from *Clostridium thermocellum* (Cthe) and *Methanococcus maripaludis* (Mmari) that were expressed in *E. coli* yields similar protein patterns separated via SDS-PAGE. B) Mass-spectrometric analysis using MALDI-TOF of the Cas8b protein from *C. thermocellum* was used for identification and mass determination of the small protein fragment. (A) Bachelorthesis of Laura Penkert, B) Masterthesis of Kristina Rau, Jörg Kahnt, MPI Marburg).

To find out whether the appearance of the C-terminal Cas8b fragment is linked to the recombinant production in *E. coli*, the Cas8b protein of *C. thermocellum* was investigated *in vivo*. Therefore, a manufactured polyclonal antibody against the Cas8b protein was used. Western blot analysis using *C. thermocellum* cell extract together with this polyclonal antibody revealed that the C-terminal Cas8b fragment is also produced *in vivo* (Masterthesis of Kristina Rau). Together with the fact that the fragment also assembles with the Cascade *in vitro* (fig 2.12), this suggests a functional role of the 17 kDa C-terminal Cas8b protein in complex formation and/or interference.

2.4.5 Cascade interference

In order to perform *in vitro* DNA interference assays, 500 nM of purified Cascade complex (Cas8b, Cas7, Cas5, Cas6 and crRNA) was mixed with freshly prepared Cas3 protein in an equimolar ratio. A 5′-terminal radiolabeled double-stranded DNA target resembling viral DNA was provided. The DNA target contained both a protospacer and a PAM (protospacer adjacent motif), flanked by random DNA sequences in both directions (fig 2.19 A). The protospacer sequence on the DNA target strand is complementary to the crRNA assembled with the complex. It was shown that after PAM recognition by Cascade, the crRNA and the target strand base-pair, which leads to the displacement of the non-target strand and the formation of an R-loop [49]. The PAM sequence that is essential for target recognition by Cascade is not known for type I-B from *C. thermocellum*. Therefore, five different DNA substrates were tested that varied in the potential PAM sequence (table 1.1). AAG and GGT PAMs have been determined for type I-B systems from other organisms. In order to identify more potential PAM sequences, the CRISPRtarget tool was utilized [112]. This program aligns sequences with viral/phage genomes and all spacer sequences from loci 1-5 of *C. thermocellum* were analyzed. Three targets were detected that displayed at least 15 consecutive spacer-protospacer base-pairs. AAC, TTA and AGT were identified as potential PAM sequences downstream of the spacer-protospacer duplex on the DNA target strand.

Table 1.1: Potential type I-B PAM sequences. *

3′-PAM-5′	**Source**
AAG	Identified for type I-B of *Haloferax volcanii* [113]
GGT	Identified for type I-B of *Listeria monocytogenes* [114]
AAC	CRISPRtarget: complementarity of spacer 25 from locus 1 to *Clostridium kluyveri* prophage
TTA	CRISPRtarget: complementarity of spacer from locus 2 to *Clostridium periferengens* plasmid
AGT	CRISPRtarget: complementarity of spacer 34 from locus 4 to *Clostridium thermocellum* prophage

* status october, 2013.

A DNA target containing a CCC PAM was used as a substrate with a potential non-PAM to prove target specificity in case of an interference reaction. The divalent metal ions magnesium and manganese were added to the assay to induce Cas3 nuclease activity as described for the type I-A Cascade interference reactions [43]. Unfortunately, significant Cascade interference could not be detected for any of the five different DNA targets. Representative results are shown in figure 2.19 B for the substrates comprising the PAMs AAG, AGT, TTA and CCC.

The *in vitro* interference activity is determined by several parameters. It was shown for the type I-E CRISPR system of *Streptococcus thermophilus,* that after PAM recognition Cascade binds to the target DNA, generating an R-loop that is bound by Cas3 [45]. The active site of the Cas3 nuclease is responsible for divalent metal binding that is essential for ssDNA hydrolysis [72]. Therefore, metal-dependence of the Cas3 nuclease activity from *C. thermocellum* was tested using different combinations of magnesium, manganese, calcium, cobalt and copper ions with 5′-terminal radiolabeled ssDNA (fig 2.20 A). The substrate was most efficiently cleaved with the addition of magnesium ions together with manganese ions. Thus, these conditions were used in the interference assays.

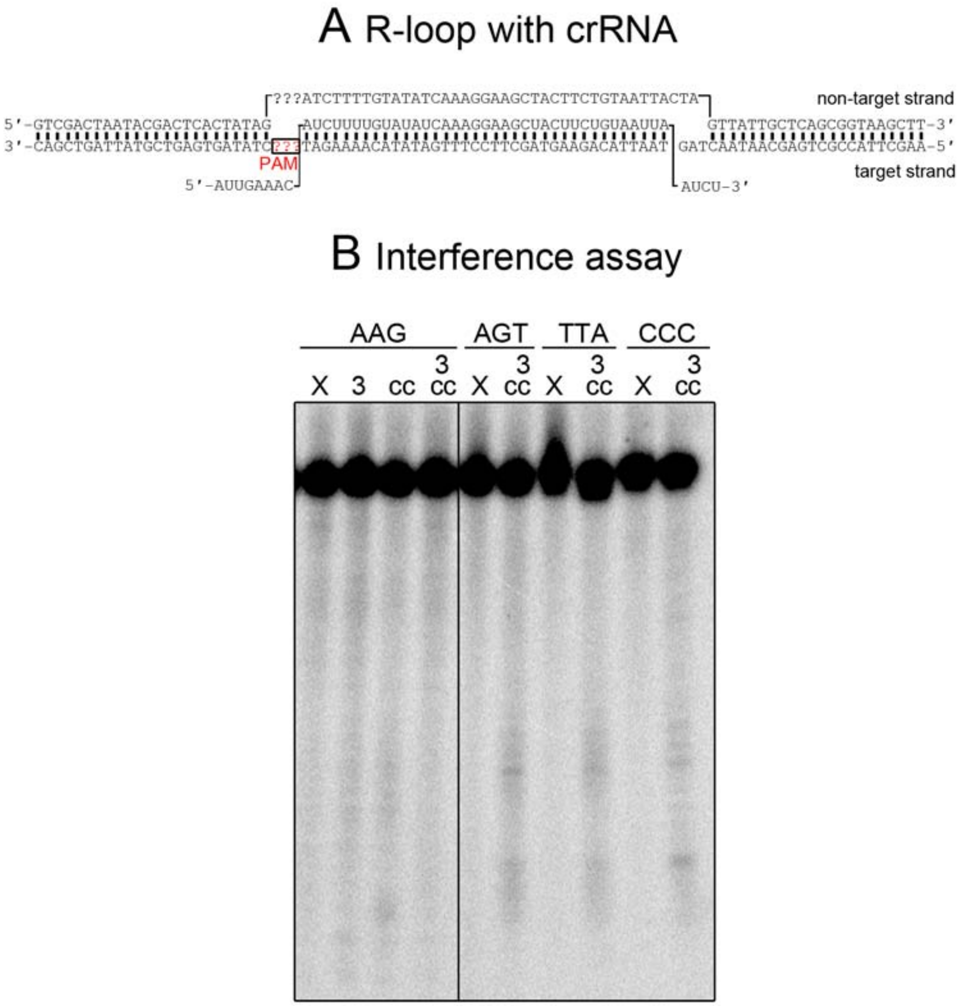

A R-loop with crRNA
???ATCTTTTGTATATCAAAGGAAGCTACTTCTGTAATTACTA
non-target strand
5'-GTCGACTAATACGACTCACTATAG
AUCUUUUGUAUAUCAAAGGAAGCUACUUCUGUAAUUA
GTTATTGCTCAGCGGTAAGCTT-3'
3'-CAGCTGATTATGCTGAGTGATATC???TAGAAAACATATAGTTTCCTTCGATGAAGACATTAAT
GATCAATAACGAGTCGCCATTCGAA-5'
PAM
target strand
5'-AUUGAAAC
AUCU-3'
B Interference assay
AAG
AGT
TTA
CCC
X
3
cc
3cc
X
3cc
X
3cc
X
3cc

Figure 2.19: Interference assay using type I-B Cascade with double-stranded target DNA. A) R-loop formation of crRNA and double-stranded DNA target. The spacer sequence of the crRNA assembled with Cascade basepairs with the protospacer sequence of the DNA target strand. The non-target DNA strand is displaced. B) Interference assay using 5'-terminal radiolabeled dsDNA substrates that vary in the PAM sequence (AAG, AGT, TTA, CCC). X =no protein added, 3 = Cas3 added, cc = Cascade (Cas8b, Cas7, Cas6, Cas5 and crRNA) added, 3cc = Cas3 together with Cascade added. Bands were separated on a denaturing 8 M urea 12 % PAA gel and visualized by autoradiography. (A) Modified from [43]).

Cas3 cleavage activity was also investigated for different protein concentrations on ssDNA after addition of magnesium and manganese (fig 2.20 B). However, only low cleavage activity can be observed using magnesium/manganese and a 4 µM Cas3 concentration. This indicates that recombinant Cas3 protein from *C. thermocellum* has little activity under these conditions and could explain unsuccessful interference assays. Recombinant Cas3 starts precipitating shortly after affinity-chromatography, which hints at problems of protein stability or folding.

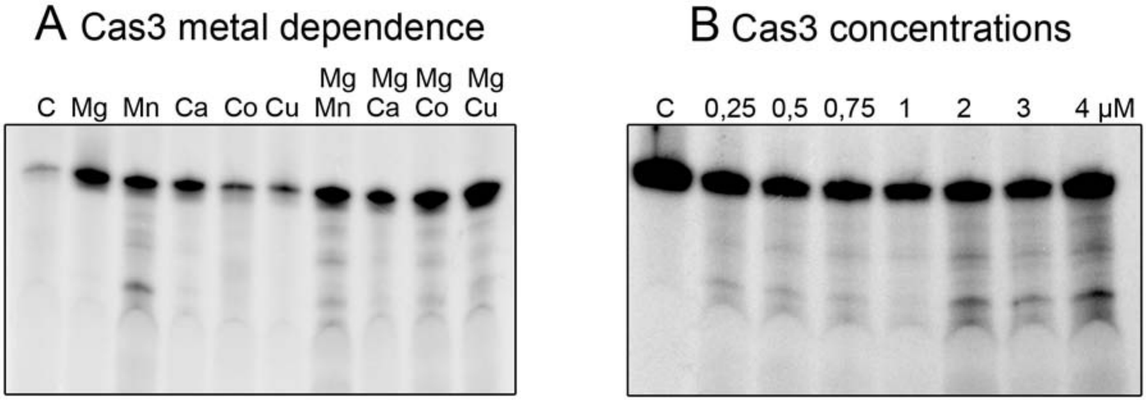


Figure 2.20: Cas3 nuclease activity on 5'-terminal radiolabeled ssDNA. A) Cas3 cleavage activity using different divalent metal ions (5mM). Weak nuclease activity revealed that the DNA substrate is most efficiently cleaved in the presence of Mg/Mn ions. B) Impact of Cas3 protein concentration from 0,25 – 4 µM on ssDNA cleavage. Bands were separated on a denaturing 8 M urea 12 % PAA gel and visualized by autoradiography.

2.4.6 Cascade stoichiometry and protein interaction sites

Assembled Cascade complex formed by Cas8b, Cas7, Cas6, Cas5 and crRNA was sent to Kundan Sharma (group of Prof. Urlaub, Max-Plank Institute Göttingen) to determine the stoichiometry of the complex and to analyze protein:protein interactions via mass-spectrometry.

Intensity-based absolute quantification (iBAQ) of proteins was used to identify the protein stoichiometry. This label-free method determines the abundance of a particular protein by calculating the sum of all peptide peak intensities obtained via mass-spectrometry matching to a specific protein divided by the number of theoretically observable peptides. This provides an accurate proxy for the number of protein copies [115, 116]. A ratio of 1 : 1 : 6 : 2.5 of the proteins Cas5 : Cas6 : Cas7 : Cas8b was obtained which correlates with the intensity of the protein bands of assembled Cascade (fig 2.12 A). Chemical cross-linking of the protein complex was obtained using a BS3 cross-linker which links lysine residues. After in-gel digestion of the cross-linked complex, the respective cross-linked peptides were analyzed via mass-spectrometry. It should be noted that cross-linked residues provide information about proximal locations, not direct contact sites [117].

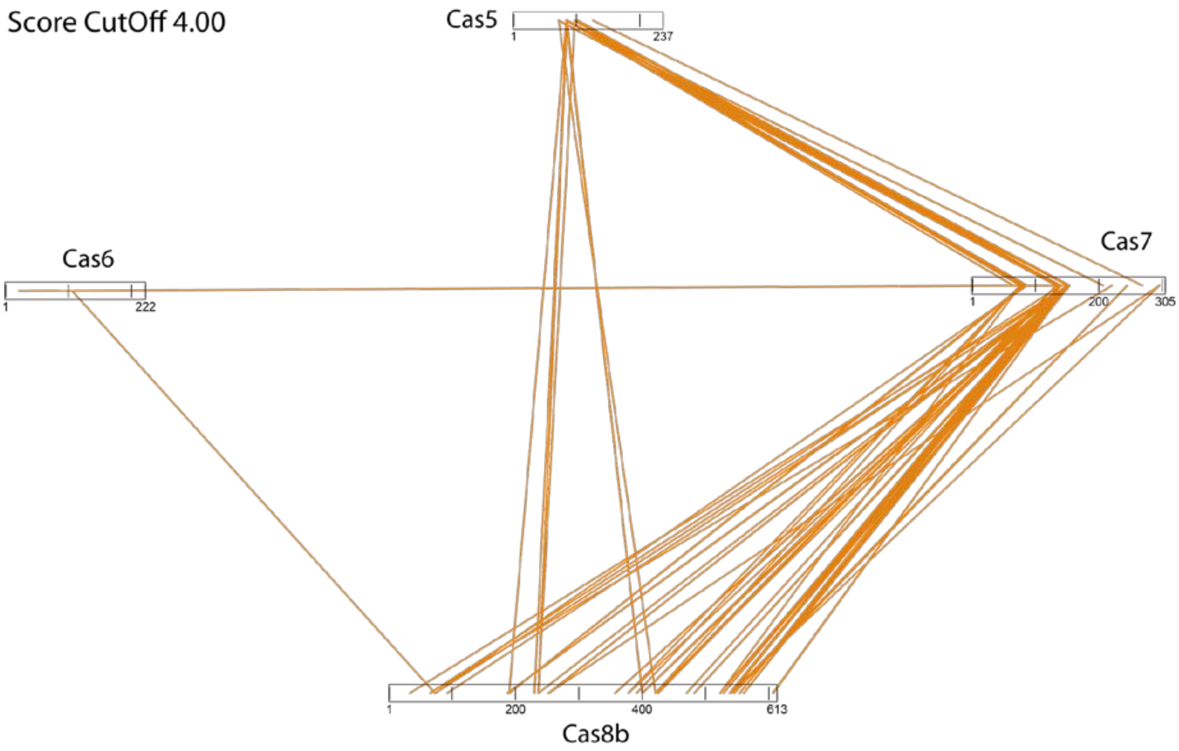


Figure 2.20: Inter-protein cross-links of the Cascade complex. A BS3 cross-linker was used to cross-link Cas protein lysine residues located in close proximity within Cascade. Inter-protein cross-links between Cas8b, Cas7, Cas6 and Cas5 are depicted as orange lines between respective amino acid residues represented by the bars. A score cut off of 4.0 was used that is based on the number of MS spectra counts that do not include low quality spectra. (Kundan Sharma, MPI Göttingen).

Inter-protein cross-links were identified for all four Cas proteins depicted in figure 2.20 with a relatively high score cutoff of 4.0. Most of the inter-protein cross-links between Cas7 and Cas8b are located within the N-terminal part of Cas7 forming two prominent clusters, whereas the corresponding cross-linked sites within Cas8b are widely distributed throughout the protein sequence. These two cross-link hot-spots within the Cas7 sequence are also the main cross-link sites between the lysine residues of Cas7 and Cas5. Cas5 inter-protein cross-links formed with Cas7 are centered in the middle of the amino acid sequence. Cas8b and Cas5 display cross-linked residues also in the middle of the sequence. Only two single residues of Cas6 could be observed that were cross-linked to Cas8b and Cas7, underlining the hypothesis that Cas6 is not a permanent Cascade subunit as described in 2.4.1 for the type I-B Cascade assembly.

2.4.7 Cas6-repeat RNA interaction

Purified Cas6_3205 protein and its corresponding 37 nt deoxy-repeat RNA was sent to Kundan Sharma (group of Prof. Urlaub, Max-Plank Institute Göttingen) to determine the protein:RNA interaction sites using UV-induced cross-linking. After UV irradiation of protein and RNA, cross-linked complexes were hydrolysed with trypsin. The resulting peptide:RNA conjugates were identified via mass-spectrometry and subsequent database search. Figure 2.21 A shows the spectrum of a peptide derived from Cas6 that was unambiguously identified as MIGFK by b- and y-type ions (resembling N- and C-terminal peptide fragments), with M (184) being cross-linked to the nucleotides UGA.

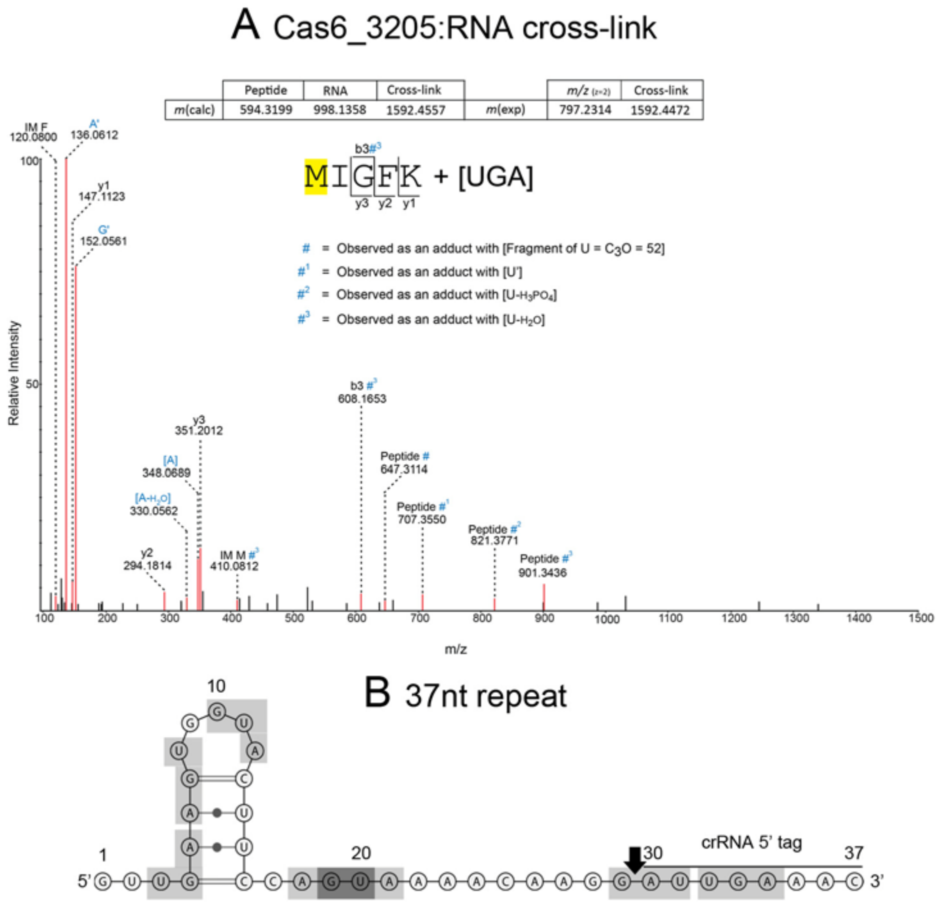


Figure 2.21: Cas6_3205 - repeat RNA interaction sites. A) Spectrum of the Cas6 peptide MFIGFK cross-linked to UGA RNA nucleotides. Identified fragment ions are indicated as y- and b-ions. Blue numbers indicate fragments with adducts. B) Predicted secondary structure of 37 nt repeat sequence. UGA protein interactionsites are indicated with grey boxes. (A)Kundan Sharma, MPI Göttingen).

An alignment of the Cas6 amino acid sequences from *C. thermocellum* and *M. maripaludis* identified an identical methionin residue at position 185 in the *M. maripaludis* Cas6 that was also determined as the interacting amino acid with the corresponding repeat RNA using the same method. It was shown that site-specific mutation of *M. maripaludis* Cas6 drastically decreases Cas6 binding of the corresponding repeat RNA (PhD thesis of Hagen

Richter). The 37 nt repeat element of *C. thermocellum* displays seven UGA sequences in different orders that could be potential interaction sites (fig 2.21 B).

2.4.8 Structure predictions of Cas5 and Cas7 and their Cascade interaction sites

The amino acid sequences of Cas5 and Cas7 were submitted to protein structure prediction tools. A model for the Cas5 protein was obtained using phyre^2 [118], whereas the I-TASSER software was used for the structure prediction of Cas7 [119]. These models display structural similarities to the *E. coli* crystal structures of Cas5e and Cas7e (fig 2.22) that were crystallized in the Cascade complex [54, 63, 68].

The *C. thermocellum* Cas5 model reveals a thumb domain as well as a palm domain with a modified RRM motif (RNA recognition motif) as observed in the *E. coli* Cas5e structure (fig 2.22 A, B). Cas7 from *C. thermocellum* displays thumb, palm and finger domains that are also present in the Cas7e structure (fig 2.22 C, D). In the *E.coli* Cascade model, Cas5e forms the tail of the complex and protects the 5′-terminal tag of the crRNA. The central bases of the 5′- terminal tag are positioned between the palm domains of Cas5e and the neighboring Cas7 protein. The thumb of Cas5e interacts with the finger domain of the adjacent Cas7 subunit. Cse1 is the large subunit of the *E. coli* Cascade and forms contact sites with the Cas5 RRM and the Cas7 thumb at the tail of the complex. The inter-protein lysine cross-links of Cas5, Cas7 and Cas8b (fig 2.20) of *C. thermocellum* were marked in the predicted structure models (fig 2.22 B, D). Within the Cas5 model, five regions were identified in the predicted thumb and palm domains that cross-linked with Cas7. Of these five regions, three were also found to cross-link with Cas8b that are mainly located in the thumb domain of Cas5. In the Cas7 model, nine regions all over the thumb, finger and palm domain were found to cross-link with Cas8b. Ten regions were found to cross-link with Cas5, partly overlapping with the cross-links observed with Cas8b. According to the *E.coli* Cascade structure, in *C. thermocellum,* the Cas5 thumb domain could interact with the Cas7 finger and the palm domains of both proteins could be located in cross-linkable proximity. The Cas5 thumb is also cross-linked to Cas8b, indicating protein contact. Cross-links between Cas8b and the Cas7 thumb and finger domain can be observed as well. As the Cascade of *C. thermocellum* reveals multiple copies of Cas7, it has to be noted that the distribution of cross-linked regions could spread over multiple protein copies. Therefore, the large Cas8b protein could interact with multiple Cas7 copies of the Cascade backbone or a Cas7 thumb and finger from different copies that are located in close proximity.

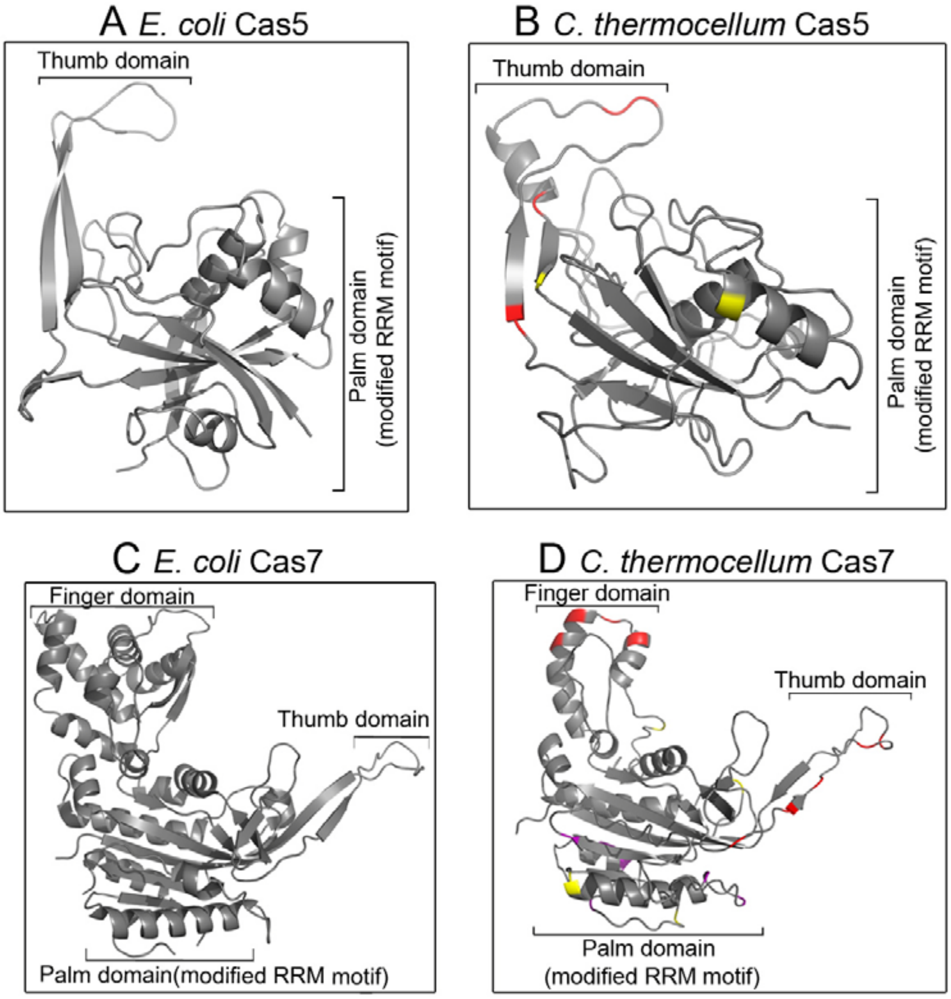


Figure 2.22: Predicted 3D structured of Cas5 and Cas7 from *C. thermocellum* and crystal structures of the homologous *E. coli* Cas proteins. A) and C) Cas5 and Cas7 from the Cascade crystal structure of *E. coli*. B) Phyre^2 model of *C. thermocellum* Cas5. Red = lysine residues cross-linked with Cas8b and Cas7, yellow = lysine residues cross-linked with Cas7. D) I-TASSER structure prediction for Cas7 from *C. thermocellum*. Red = lysine residues cross-linked with Cas8b and Cas5, yellow = lysine residues cross-linked with Cas5, purple = lysine residues cross-linked with Cas8b.

3. Material and Methods

3.1 Material and sources of supply

3.1.1 Chemicals and enzymes

The chemicals and enzymes used in this work were supplied by the companies AppliChem GmbH (Darmstadt), BioRad Laboratories GmbH (München), Biozym GmbH (Hessisch-Oldendorf), Difco Laboratories GmbH (Augsburg), Fermentas GmbH (St. Leon-Rot), Merck KGaA (Darmstadt), Roche GmbH (Mannheim), Roth GmbH (Karlsruhe), SERVA GmbH (Heidelberg) and Sigma-Aldrich Co. (Deisenhofen). Restriction endonucleases, T4-Ligase, Phusion DNA polymerase, murine RNase Inhibitor, dNTPs, 2-log DNA ladder, low range ssRNA ladder and Color-Plus protein marker were obtained from New England Biolabs GmbH (Frankfurt). Radioactive labeling of oligonucleotides was conducted using T4-Polynucleotide Kinase from Ambion/Life Technologies GmbH (Darmstadt). DNA and RNA oligonucleotides were obtained from Eurofins MWG GmbH (Ebersberg).

3.1.2 Kits

For the preparation of plasmid DNA from *E. coli* cells the Mini- and Maxi-Prep Kits from QIAGEN GmbH (Hilden) were used. The QIAquick Gel Extraction Kit from QIAGEN was utilized to extract DNA from agarose gels. Special kits are mentioned in the respective method chapters.

3.1.3 Buffers and solutions

Standard buffers and solutions were prepared according to Ausubel and Sambrook [120, 121]. Special buffers and solutions are mentioned in the chapter of the respective method. All media, solutions and buffers were, if necessary, autoclaved for 5 min at 121°C prior to usage. Heat sensitive solutions were sterilized using a sterile filter (Pore size 0.2 µm from Merck, Darmstadt) and a vacuum pump.

3.2 Culture conditions

3.2.1 *Escherichia coli*

Aerobic cultivation of *E. coli* cultures (3-800 mL) was performed in Erlenmeyer flasks by shaking in a rotatory shaker at 200 rpm and 37°C in lysogeny broth (LB) medium (1 % tryptone (w/v), 0.5 % yeast extract, 1 % NaCl (w/v), pH 7.2) or on solid medium plates (LB medium containing 1.5 % (w/v) agar-agar). Liquid LB medium containing the corresponding plasmid-encoded antibiotics (chloramphenicol 34 µg/mL), kanamycin 50 µg/mL, ampicillin 100 µg/mL and spectinomycin 50 µg/mL) was inoculated with a pre-culture (2 % (v/v)) and growth was monitored at 600 nm using a photometer NovaspecII device (Pharmacia Biotech/GE Life Science, München). Culture stocks were obtained by adding 40 % (v/v) glycerol and stored at -80°C. Protein expression was induced at OD_{600}=0.5-0.8 by the addition of 1 mM IPTG. After incubation of 3-4 hours at 37°C, cells were harvested via centrifugation (6,000 x g, 10 min, 4°) and stored at -80°C.

3.2.2 *Clostridium thermocellum*

Anaerobic cultivation of *C. thermocellum* cultures was performed at 50°C in a complex medium (modified from [122]) containing (per liter): 0.3 g NaH_2PO_4, 0.7 g K_2HPO_4, 10 g MOPS, 5 g yeast extract, 5 g cellobiose, 0.2 g $MgCl_2$ x $6H_2O$, 0.1 g $CaCl_2$ x $2H_2O$, 0.4 g cysteine HCl as a reducing agent, 0.01 g $FeSO_4$, 1.3 g $(NH_4)_2SO_4$ and 1.2 mL resazurin. The pH was adjusted to 7.0 using NaOH. 100 mL or 2 l bottles were filled with 20 or 500 mL of the medium, closed gastight and autoclaved. To obtain an anaerobic atmosphere, the bottles were flushed with N_2. The medium was incubated with 1 % of a primary culture and incubated at 50°C, agitating at 200 rpm. After incubating over-night, cells were centrifugated at 10,000 x g, 15 min at 4°C and stored at -80°C.

3.3 Strains and plasmids

3.3.1 Strains

The *E. coli* K12 DH5α strain [123] with the genotype *F*- Φ80*lac*ZΔM15 (Δ*lac*ZYA*arg*F)U169 *rec*A1 *end*A1 *hsd*R17(rK-,mK+) *pho*A *sup*E44λ-*thi*-1 *gyr*A96 *rel*A1, was used for cloning, storage and preparation of plasmid DNA.

Heterologous production of recombinant proteins from *C. thermocellum* was carried out in the *E. coli* strains Rosetta2(DE3)pLysS(Cam^r) with the genotype F-ompT hsdSB(rB- mB-) gal dcm (DE3) pLysSRARE2 (CamR) from Novagen and BL21(DE3)pLysS(Spec^r) with the genotype F-ompT hsdSB(rB- mB-) gal dcm (DE3) pLysS (Spec^r) from Invitrogen. In both strains, a T7 RNA polymerase under control of a *lacUV*-promoter is chromosomally encoded that can be induced via IPTG.

3.3.2 Plasmids and constructed recombinant vectors

Table 3.1: Sources of the used plasmids

Vector	Resistance	Application	Source
pUC19	Amp^r	*in vitro* transcription	
pRSF	Kan^r	*in vitro* transcription	
pET20b	Amp^r	heterologous gene expression	
Pec-A-HI-SUMO	Amp^r	heterologous gene expressio	Kind gift of Prof. Dr. Elena Conti (MPI Martinsried)

Table 3.2: Synthesized plasmids obtained from genscript with *C. thermocellum* genes modified for optimal codon usage in *E. coli*

Vector	Resistance	Description
pUC57+*cas8b*	Kan^r	*cas8b* gene Cthe_3201 in pUC57
pUC57+*cas7*	Kan^r	*cas7* gene Cthe_3202 in pUC57
pUC57+*cas5*	Kan^r	*cas5* gene Cthe_3203 in pUC57

pUC57+*cas3*	Kan^{r}	*cas3* gene Cthe_3204 in pUC57
pUC57+*cas6*	Kan^{r}	*cas6* gene Cthe_3205 in pUC57

Table 3.3: Constructed recombinant vectors for protein expression

Plasmid + Insert	**Description**
pET20b+*cas6* Tneap	*cas6* gene Cthe_2303 in pET20b, restriction sites NdeI / EcoRV
pET20b+*cas6* Hmari	*cas6* gene Cthe_3205 in pET20b, restriction sites NdeI / XhoI
pET20b+*cas8b*	*cas8b* gene Cthe_3201 codon optimized for *E.coli* in pET20b, restriction sites NdeI / XhoI
pET20b+*cas7*	*cas7* gene Cthe_3202 codon optimized for *E.coli* in pET20b, restriction sites NdeI / NcoI
pET20b+*cas5*	*cas5* gene Cthe_3203 codon optimized for *E.coli* in pET20b, restriction sites NdeI / XhoI
pET20b+*cas3*	*cas3* gene Cthe_3204 codon optimized for *E.coli* in pET20b, restriction sites NdeI / XhoI
pEC-A-HI-SUMO+*cas8b*	*cas8b* gene Cthe_3201 codon optimized for *E.coli* in pEC-A-HI-SUMO
pEC-A-HI-SUMO+*cas7*	*cas7* gene Cthe_3202 codon optimized for *E.coli* in pEC-A-HI-SUMO
pEC-A-HI-SUMO+*cas5*	*cas5* gene Cthe_3203 codon optimized for *E.coli* in pEC-A-HI-SUMO
pEC-A-HI-SUMO+*cas3*	*cas3* gene Cthe_3204 codon optimized for *E.coli* in pEC-A-HI-SUMO
pET20b+*RNase III*	*RNase III* gene Cthe_0931 in pET20b, restriction sites NdeI / XhoI
pET20b+*RNaseIII*E129A	*RNase III* E129A gene Cthe_0931 in pET20b, restriction sites NdeI / XhoI

Table 3.4 Constructed vectors for *in vitro* run-off transcription

Plasmid + Insert	Description
pUC19+*R-S-R* lo1	*C. thermocellum repeat33-spacer33-repeat34* of CRISPR locus 1 (30nt) with T7 promoter in pUC19, restiction sites BamHI / HindIII
pUC19+*R-S-R* lo1 reverse	*C. thermocellum repeat33-spacer33-repeat34* of CRISPR locus 1 (30nt) in reverse direction with T7 promoter in pUC19, restiction sites BamHI / HindIII
pUC19+*Spacer19* lo2	*C. thermocellum spacer19* of CRISPR locus 2 (30nt) with T7 promoter in pUC19, restiction sites BamHI / HindIII
pUC19+*Spacer19* lo2 reverse	*C. thermocellum spacer19* of CRISPR locus 2 (30nt) in reverse direction with T7 promoter in pUC19, restiction sites BamHI / HindIII
pUC19+*S-R-S* lo3	*C. thermocellum spacer145-repeat146-spacer146* of CRISPR locus 3 (30nt) with T7 promoter in pUC19, restriction sites BamHI / HindIII
pUC19+*S-R-S-R-S* lo3	*C. thermocellum spacer23-repeat24-spacer24-repeat25-spacer25* of CRISPR locus 3 (30nt) with T7 promoter in pUC19, restiction sites BamHI / HindIII
pRSF+*S-R-S-R-S* lo4	*C. thermocellum spacer2-repeat3-spacer3-repeat4-spacer4* of CRISPR locus 4 (37nt) with T7 promoter in pRSF, restriction sites PfoI / Eco0109
pUC19+*R-S-(37xG)* lo4	*C. thermocellum repeat3-spacer3-37xG-residues* of CRISPR locus 4 (37nt) with T7 promoter in pUC19, restiction sites BamHI / HindIII

3.3.3 Oligonucleotides

Table 3.5: Oligonucleotides uses for the amplification of genomic DNA from *C. thermocellum*, codon-optimized (co) gene sequences and site-directed mutagenesis of recombinant plasmid DNA. The exchanged base-pair is indicated in red. T7 promoter sequences are underlined.

Name	**Sequence 5'-3'**	**Genomic region /target vector**
Cas6_Tneap_fwd	GTAGGGCATATGAGGTTTAAAGTTGGTAT	*cas6* Cthe_2303,pET20b
Cas6_Tneap_rev	GCAGCAGATATCTATTAACTTAAAACACCCAAA	*cas6* Cthe_2303, pET20b
Cas6_Hmari_fwd	GTAGGGCATATGGATATTAAGATATTAACGGTAAG	*cas6* Cthe_3205, pET20b
Cas6_Hmari_rev	GCAGCACTCGAGTACAAAACCTTTACCAACTC	*cas6* Cthe_3205, pET20b
Cas8b_fwd	GTAGGGCATATGCTGGCTGGTGTGCTGCAA	*cas8b* Cthe_3201, co, pET20b
Cas8b_rev	GCAGCACTCGAGTTCTTCATCTTCACCGTCTTTCT	*cas8b* Cthe_3201, co, pET20b
Cas7_fwd	GTAGGGCCATGGGGATGATCAAAAACCGCCAAGAAATCC	*cas7* Cthe_3202, co, pET20b
Cas7_rev	GCAGCACTC GAG GAACGTTTTTCTTCGAAGCGGAT	*cas7* Cthe_3202, co, pET20b
Cas5_fwd	GTAGGGCATATG GTGGAAAAATACCTGGTGTTCG	*cas5* Cthe_3203, co, pET20b
Cas5_rev	GCAAAGCACCGGCCTCGTTAAATAATCAGAATGTTTTCATCAATACCTT	*cas5* Cthe_3203, co, pET20b
Cas3_fwd	GTAGGGCATATGCTGCTGTCTCACCCGGGC	*cas3* Cthe_3204, co, pET20b
Cas3_rev	GCAGCACTCGAGGAAGAAGTAATCTTTTTCCAGGTCAC	*cas3* Cthe_3204, co, pET20b
Cas8b_fwd	ACCAGGAACAAACCGGCGGCCGCTCGATGCTGGCTGGTGTGCTGCAA	*cas8b* Cthe_3201, co, SUMO
Cas8b_rev	GCAAAGCACCGGCCTCGTTATTCTTCATCTTCACCGTCTTTCT	*cas8b* Cthe_3201, co, SUMO
Cas7_fwd	ACCAGGAACAAACCGGCGGCCGCTCGATGATCAA	*cas7* Cthe_3202,

	AAACCGCCAAGAAATC	co, SUMO
Cas7_rev	GCAAAGCACCGGCCTCGTTAGAACGTTTTTTCTTC GAAGCGGAT	*cas7* Cthe_3202, co, SUMO
Cas5_fwd	ACCAGGAACAAACCGGCGGCCGCTCGATGGTGGAA AAATACCTGGTGTTCG	*cas5* Cthe_3203, co, SUMO
Cas5_rev	GCAAAGCACCGGCCTCGTTAAATAATCAGAATGTT TTCATCAATACCTT	*cas5* Cthe_3203, co, SUMO
Cas3_fwd	ACCAGGAACAAACCGGCGGCCGCTCGATGCTGCT GTCTCACCCGGGC	*cas3* Cthe_3204, co, SUMO
Cas3_rev	GCAAAGCACCGGCCTCGTTAGAAGAAGTAATCTTT TTCCAGGTCAC	*cas3* Cthe_3204, co, SUMO
Cas6_Hmari_fwd	ACCAGGAACAAACCGGCGGCCGCTCGATGATCAA AAACCGCCAAGAAATC	*cas6* Cthe_3205, co, SUMO
Cas6_Hmari_rev	GCAAAGCACCGGCCTCGTTACACGAAACCTTTACC CACACG	*cas6* Cthe_3205, co, SUMO
RNase III_fwd	GTAGGGCATATGTTTGATTTGGAAAGTTTTATGG	*rnase III* Cthe_0931, pET20b
RNase III_rev	GCAGCACTCGAGCGCATTAATATTCTCCAAAGAC	*rnase III* Cthe_0931, pET20b
RNase III_E129A_fwd	ACTTTCAGATGCCGTTGCAGCGTTAATAGGGGCTA	*rnase III* Cthe_0931 E129A, pET20b
RNase III_E129A_rev	TAGCCCCTATTAACGCTGCAACGGCATCTGAAAGT	*rnase III* Cthe_0931 E129A, pET20b
S/R 2-4_lo4_fwd	GTAGGGGCTCCGGGATAATACGACTCACTATAGCT GTATCCGCATGTCC	*Spacer/repeat 2-4*, CRISPR locus 4, pRSF
S/R 2-4_lo4_rev	AGCAGCAGGCCCCGATGTTTGCAGCACTGT	*Spacer/repeat 2-4*, CRISPR locus 4, pRSF
S/R 145/146_lo3_fwd	GCAGCAGGATCCTAATACGACTCACTATAGAGATT CAGGCGAAAATCTGA	*Spacer/repeat 145-146*, CRISPR locus 3, pUC19
S/R 145/146_lo3_rev	AGTCGGAAGCTTAGACAGATCCTGGCGTCTCC	*Spacer/repeat 145-146*, CRISPR locus 3, pUC19

Table 3.6: Oligonucleotides used for hybridization and subsequent ligation into pUC19 for *in vitro* run-off transcription. T7 promoter sequences are underlined.

Name	**Sequence 5'-3'**	**pUC19 insert**
S/R 23-25 lo3_fwd1	GATCCTAATACGACTCACTATAGAGATTCAGGCGAAAATCTGAACTTGAAGGTCAGATATGCGTTTTTATCGTACCTATGAGGAATTGAAACTCTTCCGTGTTGCCTGGGAGACGCCAGG	*S-R-S-R-S* lo3
S/R 23-25 lo3_fwd2	ATCTGTCTGTTTTTATCGTACCTATGAGGAATTGAAACATGATGTAAATAATAAAAGTCTCGATAAACTTAGGAACCAGGA	*S-R-S-R-S* lo3
S/R 23-25 lo3_rev1	AGCTTCCTGGTTCCTAAGTTTATCGAGACTTTTATTATTTACATCATGTTTCAATTCCTCATAGGTACGATAAAAACAGACAGATCCTGGCGTCTCCCAGGCAACACGGAAGAGTTTCAA	*S-R-S-R-S* lo3
S/R 23-25 lo3_rev2	TTCCTCATAGGTACGATAAAAACGCATATCTGACCTTCAAGTTCAGATTTTCGCCTGAATCTCTATAGTGAGTCGTATTAG	*S-R-S-R-S* lo3
R/S 33-34 lo1_fwd	GATCCTAATACGACTCACTATAGATCGTACCTATGAGGAATTGAAACTCTATCATGTAAACGGTAATGCATCCCATACACGCTGTTTGTATCGTACCTATGAGGAATTGAAACCAGGA	*R-S-R* lo1
R/S 33-34 lo1_rev	AGCTTCCTGGTTTCAATTCCTCATAGGTACGATACAAACAGCGTGTATGGGATGCATTACCGTTTACATGATAGAGTTTCAATTCCTCATAGGTACGATCTATAGTGAGTCGTATTAG	*R-S-R* lo1
Anti-R/S 33-34 lo1_fwd	GATCCTAATACGACTCACTATAGGTTTCAATTCCTCATAGGTACGATACAAACAGCGTGTATGGGATGCATTACCGTTTACATGATAGAGTTTCAATTCCTCATAGGTACGATCCAGGA	*R-S-R* lo1 reverse
Anti-R/S 33-34 lo1_rev	AGCTTCCTGGATCGTACCTATGAGGAATTGAAACTCTATCATGTAAACGGTAATGCATCCCATACACGCTGTTTGTATCGTACCTATGAGGAATTGAAACCTATAGTGAGTCGTATTAG	*R-S-R* lo1 reverse
S 19 lo1_fwd	GATCCTAATACGACTCACTATAGAGGTCGCTGGTTCAAATCCAGTCACTCCGACCAT C CAGGA	*Spacer19* lo2
S 19 lo1_rev	AGCTTCCTGGATGGTCGGAGTGACTGGATTTGAACCAGCGACCTCTATAGTGAGTCGTATTAG	*Spacer19* lo2
Anti- S19 lo1_fwd	GATCCTAATACGACTCACTATAGATGGTCGGAGTGACTGGATTTGAACCAGCGACCTCCCAGGA	*Spacer19* lo2 reverse

Anti- S19 lo1_rev	AGCTTCCTGGGAGGTCGCTGGTTCAAATCCAGTCACTCCGACCATCTATAGTGAGTCGTATTAG	*Spacer19* lo2 reverse

Table 3.7: Oligonucleotides used for hybridization and subsequent *in vitro* transcription. T7 promoter sequences are underlined.

Name	**Sequence 5'-3'**	**Description**
(8G)S/R 3 lo4_fwd	GTAGGGTAATACGACTCACTATAGGGGGGGGGATCTTTTGTATATCAAAGGAAGCTACTTCTGTAATTAGTTGAAGTGGTACTTCCAGTAAAACAAGGATTGAAAC	8xG-residues-spacer3-repeat3 of CRISPR locus 4
(8G)S/R 3 lo4_rev	GTTTCAATCCTTGTTTTACTGGAAGTACCACTTCAACTAATTACAGAAGTAGCTTCCTTTGATATACAAAAGATCCCCCCCCTATAGTGAGTCGTATTACCCTAC	8xG-residues-spacer3-repeat3 of CRISPR locus 4
R/S3(4nt) lo4_fwd	GTAGGGTAATACGACTCACTATAGGGAGTTGAAGTGGTACTTCCAGTAAAACAAGGATTGAAACATCTTTTGTATATCAAAGGAAGCTACTTCTGTAATTAGTTG	repeat3-spacer3-(4nt) of CRISPR locus 4
R/S3(4nt) lo4_rev	CAACTAATTACAGAAGTAGCTTCCTTTGATATACAAAAGATGTTTCAATCCTTGTTTTACTGGAAGTACCACTTCAACTCCCCTATAGTGAGTCGTATTACCCTAC	repeat3-spacer3-(4nt) of CRISPR locus 4

Table 3.8: DNA oligonucleotides used as Cascade interference targets. All sequences contain spacer9 of CRISPR locus 4 (underlined) with adjacent PAM sequences (red).

Name	**Sequence 5'-3'**
TTC_fwd	GTCGACTAATACGACTCACTATAGTTCATCTTTTGTATATCAAAGGAAGCTACTTCTGTAATTACTAGTTATTGCTCAGCGGTAAGCTT
TTC_rev	AAGCTTACCGCTGAGCAATAACTAGTAATTACAGAAGTAGCTTCCTTTGATATACAAAAGATGAACTATAGTGAGTCGTATTAGTCGAC
AAT_fwd	GTCGACTAATACGACTCACTATAGAATATCTTTTGTATATCAAAGGAAGCTACTTCTGTAATTACTAGTTATTGCTCAGCGGTAAGCTT
AAT_rev	AAGCTTACCGCTGAGCAATAACTAGTAATTACAGAAGTAGCTTCCTTTGATATACAAAAGATATTCTATAGTGAGTCGTATTAGTCGAC
GGG_fwd	GTCGACTAATACGACTCACTATAGGGGATCTTTTGTATATCAAAGGAAGCTACTTCTGTAATTACTAGTTATTGCTCAGCGGTAAGCTT
GGG_rev	AAGCTTACCGCTGAGCAATAACTAGTAATTACAGAAGTAGCTTCCTTTGATATACAAAAGATCCCCTATAGTGAGTCGTATTAGTCGAC

CCA_fwd	GTCGACTAATACGACTCACTATAGCCAATCTTTTGTATATCAAAGGAAGCTACTTCTGTAATTACTAGTTATTGCTCAGCGGTAAGCTT
CCA_rev	AAGCTTACCGCTGAGCAATAACTAGTAATTACAGAAGTAGCTTCCTTTGATATACAAAAGATTGGCTATAGTGAGTCGTATTAGTCGAC
TCA_fwd	GTCGACTAATACGACTCACTATAGTCAATCTTTTGTATATCAAAGGAAGCTACTTCTGTAATTACTAGTTATTGCTCAGCGGTAAGCTT
TCA_rev	AAGCTTACCGCTGAGCAATAACTAGTAATTACAGAAGTAGCTTCCTTTGATATACAAAAGATTGACTATAGTGAGTCGTATTAGTCGAC
TTG_fwd	GTCGACTAATACGACTCACTATAGTTGATCTTTTGTATATCAAAGGAAGCTACTTCTGTAATTACTAGTTATTGCTCAGCGGTAAGCTT
TTG_rev	AAGCTTACCGCTGAGCAATAACTAGTAATTACAGAAGTAGCTTCCTTTGATATACAAAAGATCAACTATAGTGAGTCGTATTAGTCGAC

Table 3.9: RNA oligonucleotides

Name	Sequence 5'-3'	Description
Repeat_30	GUUUUUAUCGUACCUAUGAGGAAUUGAAAC	30 nt repeat sequence
dRepeat_30	GUUUUUAUCGUACCUAUGAGGdAAUUGAAAC	30 nt repeat sequence with dNTP substitution
Repeat_37	GUUGAAGUGGUACUUCCAGUAAAACAAGGAUUGAAAC	37 nt repeat sequence
dRepeat_37	GUUGAAGUGGUACUUCCAGUAAAACAAGdGAUUGAAAC	37 nt repeat sequence with dNTP substitution
Repeat_WT_f	UCGUACCUAUGAGGAAUUGAAACC	Shortened 30 nt repeat
Repeat_WT_r	GGUUUCAAUUCCUCAUAGGUACGA	Shortened 30 nt repeat reverse
Repeat_left_f	UCGUACCUGUAAGGAAUUGAAACC	Shortened 30 nt repeat with GA substitution
Repeat_left_r	GGUUUCAAUUCCUUACAGGUACGA	Shortened 30 nt repeat with GA substitution reverse
Repeat_right_f	UCGUACCUAUGAGAACUUGAAACC	Shortened 30 nt repeat with AC substitution
Repeat_right_r	GGUUUCAAGUUCUCAUAGGUACGA	Shortened 30 nt repeat with AC substitution reverse

3.4 Working with DNA

3.4.1 Preparation of plasmid DNA from *E. coli*

Plasmid DNA was prepared from *E. coli* overnight cell cultures using the QIAGEN Plasmid Mini kit and the QIAGEN Plasmid Maxi-Plus kit (QIAGEN) according to the manufacturer's instructions.

3.4.2 Phenol/chloroform extraction of DNA

For DNA preparations from DNA/protein solution, the sample was mixed with 1 volume of phenol/chloroform (1:1) and vortexed for 1 min. After centrifugation for 1 min at 15,000 x g at RT, the upper aqueous phase was carefully transferred into a fresh tube. 1 voulme of chloroform was added to the sample, vortexed for 1 min and centrifuged again for 1 min at 15,000 x g at RT. Subsequently, the upper aqueous phase was transferred into a fresh tube and DNA precipitation was performed (3.4.3)

3.4.3 DNA precipitation

DNA preparations were concentrated via ethanol or isopropanol precipitation [120]. 2 volumes of 100 % ethanol (v/v) and 0.1 volume of 3 M Na-acetate were added to the sample and incubated for at least 30 min at -20°C. The sample was centrifuged at 12,000 x g for 10 min at RT. Afterwards, the supernatant was removed and the DNA pellet was washed with 70 % ethanol (v/v) and centrifuged again. This procedure was repeated, the supernatant was discarded, and the DNA pellet was dried for 5 min at RT. The pellet was resuspended in 30-300 µl of ddH_2O.

3.4.4 Quantitative and qualitative analysis of DNA

The concentration of DNA preparations in aqueous solutions was measured photometrically at a wavelength of λ=260 nm using a spectrophotometer (NanoDrop, Thermo Scientific, Wilmington, USA). The A_{260}/A_{280} ratio provides information on the purity oft the DNA. Sufficiently pure DNA preparations showed a ratio of ~ 1.8. Lower ratios indicated contamination with protein or phenol.

3.4.5 Agarose gel electrophoresis of DNA

Agarose gel electrophoresis of DNA molecules (plasmid DNA and PCR products) was carried out to determine their size and amount. Agarose gels with 1 % to 1.5 % (w/v) agarose in TAE buffer (40 mM Tris-acetate, 1 mM EDTA pH 8) and 0.5 µg/mL ethidium bromide were prepared depending on the size of the analyzed DNA fragments. Before the DNA samples were applied into the sample wells of the gels, they were mixed with loading dye (6x stock: 0.2 % bromphenol blue, 0.2 % xylene cyanol FF, 60 % (v/v) glycerol, 60 mM EDTA pH 8). 5 µl of 2-Log DNA ladder (New England Biolabs) were also applied on each gel. Electrophoresis was performed at 80-120 V at RT in TAE buffer. The DNA was visualized by UV irradiation at 254 nm (BioDocd-IT system, UVP).

3.4.6 Purification of DNA fragments

3.4.6.1 Gel extraction of agarose gels

For the extraction and purification of DNA fragments from agarose gels, DNA fragments were sliced out of the gel. Afterwards, the QIAquick gel extraction kit (QIAGEN) was used according to the instructions of the manufacturer. After dissolving the DNA from the gel matrix by incubation for 10 min at 50°C, the DNA bound to a silica gel column was washed several times and eluted in 25-50 µl ddH$_2$O.

3.4.6.2 Purification of PCR fragments

For purification of PCR fragment from the PCR assay, the QIAquick PCR purification kit (QIAGEN) was used according to the instructions of the manufacturer. The DNA bound to a silica gel column was eluted in 25-50 µl ddH$_2$O.

3.4.7 Polymerase chain reactions (PCR)

The PCR technique was utilized for the exponential amplification of DNA [124, 125]. Appropriate PCR primers for standard PCR were designed using the PCR primer design tool from MWG Operon (www.eurofinsdna.com). For manually designed primers the melting temperature was calculated using the following formula [126]:

$T_m = 64.9 + 41 \times (nG+nC-16.4)/(nA+nT+nG+nC)$

3.4.7.1 Amplification of genomic DNA and plasmid DNA

The PCR amplifications were performed using 50-100 ng of template DNA (genomic or plasmid DNA), 1 µM of each primer (forward and reverse), 200 µM dNTPs and 1 µl dimethyl sulfoxide (DMSO) in a 50 µl reaction volume. The reaction buffer of the Phusion polymerase (Finnzymes) supplied by the manufacturer was used. 1 U of the phusion polymerase was added directy to the sample prior to the PCR reaction that was carried out in a thermal cycler (BioRad). Standard cycling conditions are shown in table 3.10 and optimal cycling conditions were determined empirically for each DNA fragment.

Table 3.10: Standard PCR assay using Phusion polymerase

	Step 1	**Step 2-31**	**Step 32**
Denaturing	98°C, 5 min	98°C, 30 sec	
Annealing		T_m+3°C, 30 sec	
Elongation		72°C, 15-13 sec/kb	72°C, 10 min

3.4.7.2 PCR mutagenesis

Designed primer sets were used to introduce desired restriction sites at the 5' ends of DNA fragments. The restriction site overhang was considered in the calculation of the melting temperature of the primers.

3.4.7.3 Site-directed mutagenesis of plasmid DNA (Quik Change PCR)

Site directed mutagenesis was used to introduce mutations into plasmid DNA. Mutagenic primers were designed using the QuikChange Primer Design Program from Stratagene (www.stratagene.com) that contain the desired mutation, flanked by non-mutated sequences. 2 PCR reactions are carried out that contain 50 ng of the plasmid, 0.2 µM dNTP mix, 2 U of Phusion polymerase and 1 x of the corresponding polymerase buffer. Each reaction contains one of the mutagenic primers in a 0.8 µM concentration. After denaturation at 98°C for 1 min, 10 cycles followed of 98°C for 1 min; 1 min of the respective annealing temperature and 72°C for 3.5 min. Both samples were pooled and subjected to 18 cycles using the same conditions. The methylated template DNA was digested with 20 U of the methylation specific restriction endonuclease DpnI for 8 h at 37 °C. Afterwards, 2.5 µl, of the non-methylated intact PCR product was directly transformed into chemically competent *E. coli* DH5α cells.

3.4.8 Enzymatic modification of DNA

3.4.8.1 Restriction of DNA

Plasmid DNA and PCR products were designed using the respective restriction endonucleases (New England Biolabs) in appropriate buffers following the instructions of the manufacturer. The reaction mixture usually contained 5-20 U enzyme/µg DNA and were incubated at 37°C for 2 h.

3.4.8.2 5'-dephosphorylation of linearized vector-DNA

In order to avoid re-ligation of the digested plasmid DNA during the ligation reaction, Antarctic Phosphatase (New England Biolabs) treatment was used to remove the 5'-end phosphoryl groups. 1 U/ng DNA of Antarctic Phosphatase was added to the restriction mixture which was then further incubated at 37°C for 30 min. Afterwards, the phosphatase was inactivated by incubating of the sample for 5 min at 65°C.

3.4.8.3 Ligation

Ligation of restricted DNA fragments into vector DNA was performed using T4 DNA ligase [127]. 50 ng of digested, dephosphorylated plasmid DNA and ~150 ng of digested insert DNA (ration 1:3) were mixed. 200 cohesive end units of the T4 DNA ligase and 1 x supplied reaction buffer were added in a reaction volume of 20 µl and incubated overnight at 16°C. After inactivation of the T4 DNA ligase for 10 min at 70°C, the recombinant vector molecules were used for transformation.

For the generation of recombinant pec-A-HI-SUMO vectors, the respective inserts and linearized vector were processed with T4 DNA Polymerase LIC qualified (Novagen) according to the protocol of Prof. Dr. Elena Conti (MPI Munich) and subsequently ligated with a 2:1 insert:vector ratio.

3.4.8.4 Annealing of DNA oligonucleotides

1 or more sets of DNA oligonucleotides in forward and reverse direction were designed to contain the desired DNA sequence, flanked by terminal restriction sites that form sticky ends after annealing. 1 nmol of each oligonucleotide was 5'-phosphorylated in separate reactions that contained 50 U of T4-polynucleotide kinase (PNK), 1 x of the corresponding reaction buffer, and 10 mM ATP in 20 µl reaction volume. After incubation for 1 h at 37°C, 1 µl of the phophorylated forward primer was mixed with 1 µl of the corresponding phosphorylated reverse primer together with 1 U of T4 DNA ligase and 1 x of the respective reaction buffer in a 10 µl reaction volume. The samples were incubated for 5 min at 95°C on a heating block and then gradually cooled down over 2-3 h to RT. 4 µl of the hybridization mix was then ligated into appropriate vectors as described in 3.5.8.3 or directly used for *in vitro* run-off transcription (3.5.3).

3.4.9 Transformation

3.4.9.1 Preparation of chemical competent *E. coli* cells

Competent cells of all used *E. coli* strains were treated with $CaCl_2$ and RbCl (rubidium chloride). Therefore, 0.25 mL of an *E. coli* overnight culture was used for inoculation of 25 mL LB medium. Cells were incubated at 37°C and 200 rpm up to an OD_{600} of 0.3-0.5. Then, the culture was centrifuged at 2,300 x g for 10 min at 4°C and after discarding the supernatant, the cell pellet was gently resuspended in 10 mL of ice-cold solution A (10 mM MOPS, 50 mM RbCl, 50 mM $CaCl_2$) and subsequently pelleted again. The cell pellet was gently resuspended in 10 mL of ice-cold solution B (100 mM MOPS, 50 mM $CaCl_2$, 10 mM RbCl, pH 6.5) and incubated for 30 min on ice. The solution was centrifuged at 700 x g for 15 min at 4°C and the competent cells were carefully resuspended in 2 mL of ice cold solution B and 40 % glycerol and stored at -80°C for several months.

3.4.9.2 Transformation of competent *E. coli* cells

Plasmid DNA was gently mixed with 200-300 µl of competent *E. coli* cells (3.5.9.1). After incubation on ice for 1 h, the cells were subjected to a heat shock at 42°C for 45 sec and subsequently placed on ice. 1 mL of LB medium was added to the sample which was then incubated at 38°C for 1 h at 200 rpm. 100 µl of the freshly transformed cells were plated on LB agar plates that contained the respective antibiotics. The remaining cells were pelleted and carefully resuspended in 100 µl of LB medium prior to plating. The LB agar plates were incubated at 37°C overnight and colonies were screened for positive clones that contained the recombinant plasmid.

3.4.10 Sequencing

Automated sequencing of DNA [128] was performed by Eurofins MWG GmbH (Ebersberg). Primer sets that were used in addition to the standard sequencing primers from MWG operon for sequencing of *cas* genes as well as codon-optimized *cas* genes from *C. Thermocellum* over 1.5 kb are listed in table 3.11.

Table 3.11: Additional primers used for sequencing

Primer	Sequence 5'-3‘	Primer binding site
Cas3_seq_rev	TCTGTCAAGCCTCTCCTTC	Position 1653-1671 in Cas3 ORF
Cas8b_seq_fwd	ATCCTGTCTGTTCTTGTTGC	Position 694-713 in Cas8b ORF
Cas3_co_seq_fwd	ATTCGCATGGCGGAATGG	Position 760-778 in codon-optimized Cas3 ORF
Cas8b_co_seq_rev	CGAAATTGCCTTCGGTTTTATTGT	Position 1192-1216 in codon-optimezed Cas8b ORF

3.4.11 5'- terminal radioactive labeling of DNA

Single-stranded DNA oligonucleotides, synthesized by MWG Operon were used for radioactive labeling. 5 pmol of the respective template was mixed with 5 pmol of $\gamma[^{32}P]$-ATP (Hartmann Analytic), 10 U of T4 PNK and 1 x of the corresponding reaction buffer were mixed in a reaction volume of 10 µl and incubated for 1 h at 37°C.

3.4.12 Denaturing polyacrylamide gel electrophoresis of radiolabeled DNA

Denaturing 8 M urea, 12 % polyacrylamide TBE gels (mixed in a 40 mL volume with the addition of 0.1 % (v/v) APS and 0.01 / (v/v) TEMED. Gels were placed into a gel chamber (PROTEAN II Electrophoresis Chamber, BioRad). Prior to loading onto the gel, the reactions were mixed with 1 volume of formamide loading buffer (100 % formamide, 0.01 % bromphenol blue, 0.01 % xylene cyanol FF) and incubated at 95°C for 5 min. The separation was carried out for 1-2 h at 12 W, depending on the size oft the DNA fragments.

3.4.13 Detection of radiolabeled DNA by phosphorimaging

Gels were put into plastic bags and subsequently exposed to phosphor screens overnight at -20°C. The Storm 840 phosphorimager was used to visualize the bands on the phosphor screens.

3.4.14 Extraction of radiolabeled DNA from urea-polyacrylamide gels

After determining the location of the desired DNA fragments by phosphorimaging, the respective bands were cut out from the gel and dissolved in 500 µl of gel elution buffer (20 mM Tris/HCl pH 7.5, 250 mM sodium acetate, 1 mM EDTA, 0.25 % SDS) using a fresh reaction tube. The samples were incubated for 30 min at -20°C and then placed on ice on a shaker overnight. After pelleting the gel pieces (1 min, 13,000 rpm, RT), the supernatant was transferred into a fresh reaction tube. After EtOH precipitation, the activity of radiolabeled nucleid acid was measured in a scintillation counter (Beckmann LS 6500).

3.5 Working with RNA

3.5.1 Treatment of solutions, glassware and equipment

0.1 % (v/v) diethyl pyrocarbonate (DEPC) was mixed with all applied buffers and solutions, incubated overnight at RT and autoclaved to ensure protection against RNases. Glassware and tubes were sterilised at 210°C for 2 h before use. Equipment that is not heat-resistant was treated with 3 % RNase Exitus Plus (Applichem).

3.5.2 Isolation of small and total RNA from *C. thermocellum*

For the preparation of small RNAs, the *mir*Vana^{TM} miRNA Isolation Kit (Ambion) was used. 0.1 g of pelleted cells were resuspended in 0.9 mL lysis/binding buffer. After incubation with lysozyme for 20 min on ice, the cells were sonicated (Branson Sonifier 250, Danbury, CT, USA). The sample was centrifuged for 10 min at 4°C and 13.000 rpm and the supernatant was transferred to a fresh reaction tube for phenol/chloroform extraction. In order to separate the small RNAs (up to 200 nt) from the total RNA, the large RNAs were immobilized on the glass-fiber filter at a 25% ethanol concentration. Small RNAs were collected in the filtrate. Then, the ethanol concentration was increased to 55% in the filtrate and passed through a second glass-fiber filter to immobilize small RNAs. Both RNA fractions were washed and subsequently eluted using 100 µl elution solution.

3.5.3 *In vitro* run-off transcription

To generate RNA transcripts *in vitro*, run-off transcription was carried out using linearized plasmids (table 3.4 and 3.6) or hybridized oligonucleotides (table 3.7) that contained a T7 RNA promoter sequence as DNA templates. The reaction mixtures contained 40 mM Hepes/KOH, pH 8, 22 mM $MgCl_2$, 5 mM dithiothreitol (dTT), 1 mM spermidine, 4 mM of each NTP, 20 U RNase Inhibitor, 30 nM T7 RNA polymerase and 50 µg/mL linearized plasmid or 10-20 µg/mL hybridized oligonucleotides. The reaction volume varied from 20 µl up to 10 mL. The reaction mixture was incubated for 3–4 h at 37 °C. Afterwards, the samples were subjected to phenol/chloroform extraction and ethanol precipitation, followed by denaturing polyacrylamide gel electrophoresis, gel-extraction and ethanol precipitation before usage.

3.5.4 Denaturing polyacrylamide gel electrophoresis and gel extraction of RNA

Electrophoretic separation of RNA fragments obtained via *in vitro* transcription was performed using 8 M urea, 12 % polyacrylamide gels as described for radiolabeled DNA (3.4.12). The low range ssRNA ladder was used as a size standard. The gels were stained with a 0.1 % toluidine blue solution for 5 min and subsequently destained using water. The RNA fragments of choice were sliced out of the gel and transferred into a fresh reaction tube. Extraction of the RNA from the gel pieces was performed as described for radiolabeled DNA (3.4.14).

3.5.5 RNA precipitation and quantitative and qualitative analysis of RNA

Precipitation of small RNA transcripts was performed as described for DNA in 3.4.3 with the exception that glycerol was added to the sample mixed with 100 % ethanol in a 1/100 ratio. Quantitative analysis of the RNA was carried out as described for DNA in 3.4.15. Sufficiently pure RNA samples showed a A_{260}/A_{280} ratio higher than 1.8. A lower ratio indicates protein or phenol contaminations of the preparation.

3.5.6 5'- terminal radioactive labeling of RNA

Single-stranded RNA oligonucleotides (MWG Operon), as well as RNA transcripts generated via *in vitro* run-off transcription, were used for 5'-terminal radioactive labeling with $\gamma[^{32}P]$ in a T4 PNK reaction as described for DNA in 3.4.11. One exception to the protocol is that RNA transcripts were 5'-dephosphorylated as described in 3.5.8.2 prior to labeling.

3.5.7 RNA-sequencing

3 µg of small RNA from *C. thermocellum* was incubated for 3 h at 37°C with 20 U T4-PNK and 1 x of the respective reaction buffer to obtain dephosphorylation of 2'-3'-cyclic phosphate termini. Then, 2 mM ATP and another 10 U of T4 PNK were added to the mixture and incubated for 1 h at 37°C to generate monophosphorylated 5'-termini. The preparation of the RNA libraries was carried out using an Illumina TruSeq RNA Sample Prep Kit. Sequencing on an Illumina HiSeq2000 sequencer was performed at the Max Planck Genomecentre, Köln (Max Planck Institute for Plant Breeding Research).

3.5.8 Identification of crRNA abundance

Sequencing reads were trimmed by the removal of Illumina TruSeq linkers and poly-A tails as well as the removal of sequences using a quality score limit of 0.05. The trimmed reads were mapped to the reference genome of *C. thermocellum* (GenBank: CP000568) with CLC Genomics Workbench 5.0 (CLC Bio, Aarhus, Denmark). The following mapping parameters were used: mismatch cost: 2, insertion cost: 3, deletion cost: 3, length fraction: 0.5, similarity: 0.8. Additionally, reads <15 nt were removed. Gene annotations were obtained from GenBank.

3.6 Biochemical methods

3.6.1 Heterologous production of *C. thermocellum* proteins in *E. coli*

For the heterologous production of proteins from *C. thermocellum*, the respective genes that were cloned with the T7 RNA polymerase pET vector system (Novagen) and the pec-A-HI-SUMO vector were used. The heterologous expression was performed in *E. coli* BL21(DE)pLysS and Rosetta(DE)pLysS strains as described in 3.3.1. The pLys plasmid encodes T7 lysozyme which is a natural inhibitor of the T7 RNA polymerase. The transformed vector constructs are under the control of a T7 lac promoter which can be induced by the addition of IPTG. The used *E. coli* strains ensure minimal low-level expression of the genes before IPTG induction. Culture conditions are described in 3.2.1.

3.6.2 Enrichment and purification of recombinant enzymes

3.6.2.1 Enrichment of recombinant Cas8b protein

Recombinant *E. coli* Rosetta(DE)pLysS cells were resuspended in chilled buffer 1 (100 mM potassium phosphate pH 7.5, 500 mM NaCl and 1 mM ß-Me) and incubated with 1.5 mg lysozyme for 30 min on ice. Cells were sonicated 6 x 30 sec (40 % duty cycle, output control 4, Branson Sonifier 250) unbroken cells and cell debris were removed via ultracentrifugation (30,000 x g, 4°C, 30 min). The protein solution was injected into an FPLC system (ÄKTApurifier^{TM}, GE Healthcare) and purified via affinity chromatography using a nickel-NTA column (HisTrap HP, GE Healthcare). The column was equilibrated for 10 min with a flow-rate of 1 mL/min in buffer 1, and the protein was loaded onto the column (10 min; flow rate 1 mL/min). Afterwards, the protein solution was washed with buffer 1 for 10 min at 1 mL/min, buffer 2 (50 mM KCl, 1000 mM NaCl, 10 mM $MgSO_4$ and 2 mM ATP) for 5 min at 1 mL/min and again buffer 1 for 10 min. The protein was eluted from the column using buffer 3 (500 mM NaCl, 100 mM potassium phosphate pH 7.5, 1 mM ß-Me and 500 mM imidazole) with a linear imidazole gradient (0-500 mM, 20 min, 1mL/min). Fractions that contained protein were identified via absorption at 280 nm wavelength. The 1 mL peak fractions were collected and analyzed via SDS-PAGE (3.6.3). The fractions that contained Cas8b protein were pooled and dialysed in buffer 4 (50 mM Tris/HCl pH 7.5, 1mM dTT and 100 mM NaCl) overnight at 4°C. The protein solution was concentrated with Amicon filters units (Amicon Ultra-15, 30,000 MW, Millipore) to a volume of 10 mL and further purified using cation-exchange chromatography with a

Heparin column (HiTrap Heparin HP, GE healthcare). The solution was loaded onto the column (equilibration with purification buffer 4, 10 mL; 1 mL/min) and directly eluted with a linear salt gradient of 100-1000 mM NaCl for 20 min, 1mL/min using buffer 5 (50 mM Tris/HCl pH 7.5, 1 mM dTT and 1000 mM NaCl). The 1 mL peak fractions were collected and analyzed by SDS-PAGE (3.6.3).

3.6.2.2 Enrichment of recombinant Cas7 protein

Recombinant *E. coli* Rosetta(DE)pLysS cells were resuspended in chilled buffer 1 (100 mM potassium phosphate pH 7.5, 500 mM NaCl and 1 mM ß-Me) and incubated with 1.5 mg lysozyme for 30 min on ice. Cells were sonicated 6 x 30 sec (40 % duty cycle, output control 4, Branson Sonifier 250) unbroken cells and cell debris were removed via ultracentrifugation (30.000 x g, 4°C, 30 min). The protein solution was injected into an FPLC system (ÄKTApurifier^{TM}, GE Healthcare) and purified via affinity chromatography using a nickel-NTA column (HisTrap HP, GE Healthcare). The column was equilibrated for 10 min with a flow-rate of 1 mL/min in buffer 1, and protein was loaded onto the column (10 min; flow rate 1 mL/min). Afterwards, the protein solution was washed with buffer 1 for 10 min at 1 mL/min, buffer 2 (50 mM KCl, 1000 mM NaCl, 10 mM $MgSO_4$ and 2 mM ATP) for 5 min at 1 mL/min and again buffer 1 for 10 min. Afterwards the protein was eluted from the column using buffer 3 (500 mM NaCl, 100 mM potassium phosphate pH 7.5, 1 mM ß-Me and 500 mM imidazole) with a linear imidazole gradient (0-500 mM, 20 min, 1mL/min). Fractions that contained protein were identified via absorption at 280 nm wavelength. The 1 mL peak fractions were collected and analyzed via SDS-PAGE (3.6.3). The fractions that contained Cas7 protein were pooled.

3.6.2.3 Enrichment of recombinant Cas5 protein

Recombinant *E. coli* Rosetta(DE)pLysS cells were resuspended in chilled buffer 1 (100 mM potassium phosphate pH 7.5, 500 mM NaCl and 1 mM ß-Me) and incubated with 1.5 mg lysozyme for 30 min on ice. Cells were sonicated 6 x 30 sec (40 % duty cycle, output control 4, Branson Sonifier 250) unbroken cells and cell debris were removed via ultracentrifugation (30,000 x g, 4°C, 30 min). The protein solution was injected into an FPLC system (ÄKTApurifier^{TM}, GE Healthcare) and purified via affinity chromatography using a nickel-NTA column (HisTrap HP, GE Healthcare). The column was equilibrated for 10 min with a flow-rate of 1 mL/min in buffer 1, and the protein was loaded onto the column (10 min; flow rate 1 mL/min). Afterwards, the protein solution was washed with buffer 1 for 10 min at 1 mL/min, buffer 2 (50 mM KCl, 1000 mM NaCl, 10 mM $MgSO_4$ and 2 mM ATP) for 5 min at 1 mL/min and again buffer 1 for 10 min. Afterwards the protein was eluted from the column using buffer 3 (500 mM NaCl, 100 mM potassium phosphate pH 7.5, 1 mM ß-Me and 500 mM imidazol) with a linear imidazole gradient (0-500 mM, 20 min, 1mL/min). Fractions that contained protein were identified via absorption at 280 nm wavelength. The 1 mL peak fractions were collected and analyzed via SDS-PAGE (3.6.3). The fractions that contained Cas5 protein were pooled.

3.6.2.4 Enrichment of recombinant Cas3 protein

Recombinant *E. coli* Rosetta(DE)pLysS cells were resuspended in chilled buffer 1 (100 mM potassium phosphate pH 7.5, 500 mM NaCl and 1 mM ß-Me) and incubated with 1.5 mg lysozyme for 30 min on ice. Cells were sonicated 6 x 30 sec (40 % duty cycle, output control 4, Branson Sonifier 250) unbroken cells and cell debris were removed via ultracentrifugation (30,000 x g, 4°C, 30 min). The protein solution was injected into an FPLC system (ÄKTApurifier^{TM}, GE Healthcare) and purified via affinity chromatography using a nickel-NTA column (HisTrap HP, GE Healthcare). The column was equilibrated for 10 min with a flow-rate of 1 mL/min in buffer 1, then the protein was loaded onto the column (10 min; flow rate 1 mL/min). Afterwards, the protein solution was washed with buffer 1 for 10 min at 1 mL/min, buffer 2 (50 mM KCl, 1000 mM NaCl, 10 mM $MgSO_4$ and 2 mM ATP) for 5 min at 1 mL/min and again buffer 1 for 10 min. Afterwards the protein was eluted from the column using buffer 3 (500 mM NaCl, 100 mM potassium phosphate pH 7.5, 1 mM ß-Me and 500 mM imidazole) with a linear imidazole gradient (0-500 mM, 20 min, 1mL/min). Fractions that contained protein were identified via absorption at 280 nm wavelength. The 1 mL peak fractions were collected and analyzed via SDS-PAGE (3.6.3). The fractions that contained Cas3 protein were pooled and dialysed in 50 mM Hepes/KOH pH 7, 1mM dTT and 300 mM NaCl overnight at 4°C.

3.6.2.5 Enrichment of recombinant Cas6 (Cthe_2303 and Cthe3205) proteins

Recombinant *E. coli* BL21(DE)pLysS cells were resuspended in chilled buffer 1 (100 mM potassium phosphate pH 7.5, 500 mM NaCl and 1 mM ß-Me) and incubated with 1.5 mg lysozyme for 30 min on ice. Cells were sonicated 6 x 30 sec (40 % duty cycle, output control 4, Branson Sonifier 250) unbroken cells and cell debris were removed via ultracentrifugation (30,000 x g, 4°C, 30 min). The protein solution was injected into an FPLC system (ÄKTApurifier^{TM}, GE Healthcare) and purified via affinity chromatography using a nickel-NTA column (HisTrap HP, GE Healthcare). The column was equilibrated for 10 min with a flow-rate of 1 mL/min in buffer 1, then the protein was loaded onto the column (10 min; flow rate 1 mL/min). Afterwards, the protein solution was washed with buffer 1 for 10 min at 1 mL/min, buffer 2 (50 mM KCl, 1000 mM $NaCl_2$, 10 mM $MgSO_4$ and 2 mM ATP) for 5 min at 1 mL/min and again buffer 1 for 10 min. Afterwards the protein was eluted from the column using buffer 3 (500 mM NaCl, 100 mM potassium phosphate pH 7.5, 1 mM ß-Me and 500 mM imidazole) with a linear imidazole gradient (0-500 mM, 20 min, 1mL/min). Fractions that contained protein were identified via absorption at 280 nm wavelength. The 1 mL peak fractions were collected and analyzed via SDS-PAGE (3.6.3). The fractions that contained Cas6 protein were pooled and dialysed in 50 mM Hepes/KOH pH 7, 1mM dTT and 300 mM $NaCl_2$ overnight at 4°C.

3.6.2.6 Enrichment of recombinant wildtype and mutant RNase III proteins

Recombinant *E. coli* BL21(DE)pLysS cells were resuspended in chilled buffer 1 (50 mM Tris/HCl pH 7, 300 mM NaCl and 1 mM dTT) and incubated with 1.5 mg lysozyme for 30 min on ice. Cells were sonicated 6 x 30 sec (40 % duty cycle, output control 4, Branson Sonifier 250) unbroken cells and cell debris were removed via ultracentrifugation (30,000 x g, 4°C, 30 min). The protein solution was injected into an FPLC system (ÄKTApurifier^{TM}, GE Healthcare) and purified via affinity chromatography using a nickel-NTA column (HisTrap HP, GE Healthcare). The column was equilibrated for 10 min with a flow-rate of 1 mL/min in buffer 1, then the protein was loaded onto the column (10 min; flow rate 1 mL/min). Afterwards, the protein solution was washed with buffer 1 for 10 min at 1 mL/min and eluted from the column using buffer 2 (50 mM Tris pH 7, 300 mM NaCl, 1 mM dTT and 500 mM imidazole) with a linear imidazole gradient (0-500 mM, 20 min, 1mL/min). Fractions that contained protein were identified via absorption at 280 nm wavelength. The 1 mL peak fractions were collected and analyzed via SDS-PAGE (3.6.3). The fractions that contained wildtype or mutant RNase III protein were pooled and dialysed in 50 mM Tris pH 7, 1mM dTT and 300 mM NaCl overnight at 4°C. The protein solution was concentrated with Amicon filters units (Amicon Ultra-15, 30,000 MW, Millipore) to a volume of 10 mL and further purified using cation exchange chromatography with a Heparin column (HiTrap Heparin HP, GE healthcare) and eluted with a linear salt gradient (100-1000 mM). The 1 mL peak fractions were collected and analyzed by SDS-PAGE (3.6.3).

3.6.3 SDS-polyacrylamide gel electrophoresis (SDS-PAGE)

Denaturing sodium dodecylsulphate (SDS) polyacrylamide gel electrophoresis (PAGE) was used for protein analysis [129]. Polyacrylamide gels (10.5 x 11.5 x 1, Mini Protean system, BioRad) that comprise an upper stacking and a lower separation gel were used as support matrix in electrophoresis. The stacking gels (125 mM Tris/HCl pH 6.8, 0.1 % (v/v) SDS, 0.03 % (v/v) APS and 0.005 % TEMED) contained 6 % polyacrylamide, whereas separating gels (125 mM Tris/HCl pH 6.8, 0.1 % (v/v) SDS, 0.03 % (v/v) APS and 0.005 % TEMED) were prepared with 12 % polyacrylamide. The ingredients of the separating gels were mixed and subsequently poured into the gel casting chamber. The gel was covered with ddH_2O and polymerised. Afterwards, the stacking gel solution was poured on top and a 10-sample well was placed within the stacking gel. After polymerisation, the comb was removed and the freshly prepared gels were directly used for electrophoresis or stored at 4°C. The protein samples were mixed with 1 volume of 5 x loading buffer (100 mM Tris/HCl pH 6.8, 40 % (v/v) glycerin, 10 % (v/v) ß-Me, 3.2 % (w/v) SDS, 0.2 % (w/v) bromphenol blue) and incubated at 95°C for 5 min for denaturation prior to loading on the gel. Gel runs were performed in a Mini-Protein Tetra Cell (Biorad) that contained electrophoresis buffer (25 mM Tris/HCl pH8, 190 mM glycine and 0.1 % (v/v) SDS) at 200 V for 45 min. Proteins were visualized by gel staining (25 % (v/v) ethanol, 10 % (v/v) acetic acid, 0.25 % (v/v) Coomassie Brilliant Blue R-250) for 30 min at 37°C subsequently destained (5 % (v/v) methanol and 7.5 % (v/v) acetic acid) [130].

3.6.4 Protein quantitation

The Bio-Rad protein assay based on Bradford protein quantitation method [131] was used by following the instructions of the manufacturer to determine protein concentrations. Protein extinction was measured photometrically at 595 nm. Bovine serum albumin (BSA) concentrations of 2, 4, 6, 8 and 10 mg/mL served as a standard to create a calibration curve.

3.6.5 Molecular mass determination of proteins under native conditions

The approximate molecular mass of protein complexes was determined using gel-filtration chromatography. A gel-filtration Superdex column (Superdex 200 10/300GL, Separation

10.000 – 600.000 Da, volume: 24 mL, GE Healthcare) was used in a buffered solution (50 mM Hepes/KOH pH 7, 1 mM dTT and 300 mM NaCl) and equilibrated (3 h, 0.3 mL/min) before sample loading. Prior to the analysis of protein complexes, a mixture of calibration proteins (with known molecular weight) was applied to the column. The elution volumes V_e were monitored for each protein and together with the exclusion volume (void volume) V_0, a calibration curve was generated. The ratio of elution volume and exclusion volume V_e/V_o for each calibration protein is blotted against the size of the respective protein. The calibration curve is then used to calculate the molecular weight of the analyzed protein complexes. Appoferritin (used as V_o; MW 443,000 Da, 1 mg), ß-Amylase (MW 200,000 Da, 1.125 mg), Alcohol Dehydrogenase (MW 148,000 Da, 1 mg), BSA (MW 66,000 Da, 1 mg), Carbonic Anhydrase (MW 29,000 Da, 1.125 mg) and Cytochrome C (MW 12,400 Da, 1 mg) were used as calibration proteins.

3.6.6 Trichloroacetate precipitation of proteins

Protein samples were supplied with 20 % (v/v) TCA to precipitate the proteins. After incubation for 10 min on ice, precipitated proteins were pelleted at 14,000 rpm for 5 min. The supernatant was discarded and the pelleted proteins were washed two times with 200 µl of ice-cold acetone. Pellets were dried for 5 min at 95°C and subsequently solved in 50 mM Tris/HCl pH 7.

3.6.7 Assembly of recombinant Cas proteins

For the assembly of the Cascade complex, the recombinant proteins Cas6 (C-terminal 6 x His-tag), Cas5 (N-terminal SUMO tag), Cas7 (N-terminal SUMO tag) and Cas8b (C-terminal His-tag) were purified as described in 3.6.2. A 1:1:2 ratio of Cas5:Cas8b:Cas7 protein was used for the assembly, depending on the amount of protein that was obtained from the individual purified preparations. Cas5 and Cas7 proteins were mixed together with 100 µl of SUMO protease and dialysed overnight in 50 mM Hepes/KOH pH 7, 1 mM dTT and 300 mM NaCl. Cas8b was added and the mixture was concentrated to a volume of ~ 2 mL using Amicon filters units (Amicon Ultra-15, 30.000 MW, Millipore) and centrifuged for 10 min at 14.000 rpm. Meanwhile, the respective RNA precursors were processed using the Cas6 endonuclease as described in 3.6.9 for 2 h at 37°C. 5 µM of Cas6 was used in a total reaction volume of 2 mL to obtain 1.5 nmol of the respective crRNA variants. The Cas protein mixture and the endonuclease reaction were pooled and incubated for 30 min at 50°C. Afterwards, precipitates were pelleted and the sample was concentrated to a volume of ~ 500 µl and subjected to gel-filtration chromatography using an equilibrated Superdex 200 column. The protein complexes were separated at a flow rate of 0.4 mL/min and 1 mL fractions were collected. Afterwards, the 1 mL fractions A7-B12 (V_e: 6,6-12,6 mL) were further analyzed. From each fraction sample, 300 µl were used for TCA precipitation (3.6.6) and subsequent SDS-PAGE (3.6.3). 700µl of the fraction samples were used for phenol/chloroform extraction, ethanol precipitation (3.5.5) followed by urea-PAGE analysis (3.5.4).

3.6.8 Cascade interference assay

To test the cleavage of dsDNA, 5 pmol of ssDNA oligonucleotides (table 3.8) were 5'-terminal radiolabeled (3.5.6), purified and hybridized with a 1.5-fold molar excess of the respective unlabeled complementary strand. Therefore, both DNA strands were mixed, heated up to 95°C for 5 min and slowly cooled down to RT in hybridization buffer (10 mM Tris pH 8, 1 mM EDTA, 100 mM NaCl). For the interference reaction, 500 nM assembled Cascade complex and 500 nM of Cas3 (C-terminal 6 x His-tag) were mixed in 50 mM Hepes/KOH pH7, 1 mM dTT, 300 mM NaCl, 5 mM $MgCl_2$, 5 mM $MnCl_2$ and 2 mM ATP in a 12 µl reaction volume. The reaction was started by adding 20,000 cpm of 5'-labeled hybridized dsDNA to the mixture that was incubated for 20 min at 50°C. To stop the reaction, EtOH precipitation was performed. Subsequently, the samples were loaded onto 20 % denaturing polyacrylamide gels (3.5.4).

3.6.9 Cas6 endonuclease assay

1 µM of the respective Cas6 enzyme was incubated with 5'-terminal radiolabeled RNA (3.5.6) in 20 mM Hepes/KOH pH 8, 250 mM KCl, 2 mM $MgCl_2$, in 10 µl reaction volume and incubated at 37°C for 10 min. Afterwards, the samples were mixed with 2 x formamide buffer and incubated at 95°C for 5 min. The reactions were then loaded on 12 % denaturing polyacrylamide gels (3.5.4).

3.6.10 RNase III cleavage assay

5 pmol of ssRNA transcripts were 5'-terminal radiolabeled (3.5.6), purified and hybridized with 1.5-fold molar excess of the respective unlabeled complementary strand. Therefore, both strands were mixed, heated up to 95°C for 5 min and slowly cooled down to RT in hybridization buffer 10 mM Tris/HCl pH 7. For the cleavage reaction, 2 µM of the

respective enzyme was incubated with the 5'-labeled hybridized dsRNA in 5 mM $MgCl_2$, 1 mM dTT, 20 mM KCl, 50 mM Tris/HCl pH 7 in a 10 µl reaction volume and incubated for 5 min at 50 °C. The samples were mixed with 2 x formamide buffer and incubated for 5 min. The reactions were loaded on 12 % denaturing polyacrylamide gels (3.5.6) alongside an two markers using an RNase T1 digest (cleaves after each G residue) and an alkaline hydrolysis marker (fragments RNA substrate) that were obtained using the RNase T1 (Biochemistry Grade) Kit from Ambion according to the manufacturer’s instructions.

3.6.11 Electromobility shift assay

20,000 cpm of 5'-terminal radiolabeled crRNA substrate was used in each reaction with 50 mM Tris/HCl pH 7, 1 mM dTT, 75 ng yeast RNA and indicated concentrations of assembled Cascade complex in a volume of 10 µl. The samples were incubated at 50°C for 5 min, mixed with 2.5 µl GelPilot DNA loading dye (QIAGEN) and then loaded onto an 8 % non-denaturing TBE polyacrylamide gel.

4. Discussion

4.1 Cas6 processing and its association with Cascade

Previous studies have shown that Cas6 endonucleases of different CRISPR-Cas subtypes generate mature crRNAs that comprise an 8 nt 5'-terminal repeat tag [24, 26, 47, 50-52, 132]. *Clostridium thermocellum* contains two type I-B Cas6 endonucleases (Cthe_2303 and Cthe_3205). According to an older nomenclature, Cas6 Cthe_2303 is classified as type I-B “Tneap”, mainly found in bacteria, and Cas6 Cthe_3205 belongs to type I-B “Hmari” that is present in many archaea. Cas6 “Tneap” is associated with CRISPR clusters that contain 30 nt long repeat elements, whereas Cas6 “Hmari” is associated with CRISPR loci that comprise repeats of 37 nt length. Both Cas6 enzymes generate crRNAs that comprise an 8 nt 5'-terminal repeat tag, which is in agreement with the universal Cas6 processing pattern.

Cas6 proteins are a heterogeneous group of endonucleases with highly diverse sequences, structures and catalytic site compositions. In addition, the sequence and secondary structures of the CRISPR repeats differ between the subtypes. Thus, Cas6 diversity might result from the adaptation to their respective repeat sequences and/or structures [50]. Different mechanisms of Cas6 repeat RNA recognition have evolved. One general mechanism is exemplified by Cas6 from *Pyrococcus furiosus* (PfCas6) of type III-B. This enzyme is proposed to use a wrap-around mechanism, wherein unstructured repeats of long pre-crRNAs are bound in a bead chain-like manner and are subsequently processed [133]. In contrast, the Cas6 enzymes of type I-E, found e.g. in *Thermus thermophilus* and *E. coli*, as well as of type I-F present in *Pseudomonas aeruginosa*, specifically recognize repeats that form hairpin structures [26, 51, 52, 56]. As shown in section 2.2.1, Cas6 Cthe_2303 preferably cleaves the corresponding 30 nt repeat sequence, whereas Cas6 Cthe_3205 preferably processes the associated 37 nt repeat RNA. A hairpin structure was predicted to be formed in the 5'-terminal region of the 37 nt repeat using the “RNAfold” software (fig 2.21) [134]. However, *C. thermocellum* is a thermophilic organism that grows at an elevated temperature of 60 °C. Considering this parameter in the RNA folding prediction showed that this four bp stem is unlikely to be retained. According to the CRISPRmap tree, a CRISPR repeat classification based on sequence and structure similarities that represents evolutionary relationships, both repeat sequences of *C. thermocellum* are not predicted to contain a conserved structure motif [12], indicating that the Cas6 repeat recognition is likely not based on a conserved RNA structure. Furthermore, the two repeat sequences are classified into different superclasses. The 30 nt repeat belongs to superclass A, whereas the 37 nt repeat is grouped into superclass E, indicating that the two types of CRISPR repeats/clusters are phylogenetically distant [12]. Together with the finding that the two Cas6 enzymes are diverse in sequence and mainly distributed in different prokaryotic domains, this underlines the hypothesis that Cas6 enzymes and the respective CRISPR repeat elements co-evolved.

The archaeon *Methanococcus maripaludis* contains a single type I-B CRISPR-Cas system with a Cas6 endonuclease that is highly homologous to the Cas6 Cthe_3205 enzyme of *Clostridium thermocellum* (40 % aa identity and 62 % aa similarity). Furthermore, the CRISPR cluster of *M. maripaludis* contains 37 nt long, unstructured repeat elements that are classified into superclass E and are closely related to the 37 nt repeat elements of *C. thermocellum*. Together, these similarities in repeat sequences and Cas6 enzymes hint at a horizontal gene transfer event between archaea and bacteria and exemplify the mobility of CRISPR-Cas systems. In accordance with this, the Cas6 enzymes from *M. maripaludis* and *C. thermocellum* were shown to cleave both 37 nt repeat sequence (PhD thesis of Hagen Richter). The analysis of UV induced protein:RNA cross-links of Cas6 Cthe_3205 with the respective non-hydrolysable 37 nt repeat RNA and subsequent mass-spectrometry identified the methionine residue at position 184 of the protein to cross-link with a uridine base in a UGA sequence of the repeat as described in section 2.4.7. Using the same technique, the homologous methionine residue 185 of the Cas6 protein from *M. maripaludis* was identified to cross-link with UUGC of the repeat RNA and shown to be involved in RNA binding as the mutation of the respective residue resulted in a decreased binding affinity (PhD thesis of Hagen Richter). Recently, the co-crystal structure of Cas6 from *M. maripaludis* with non-hydrolysable repeat RNA was obtained (Richter *et al.*, unpublished). The data support a model wherein Cas6 dimerizes upon substrate binding and reshapes the RNA to form a 2-base pair stem-loop immediately upstream of the cleavage site. Similar base pairs that would also form this stem-loop can be observed in the *C. thermocellum* repeat. However, the RNA cross-link site UUGC at position 14-17 in the 37 nt repeat from *M. maripaludis* is not identical with any of the eight possible cross-linked UGA positions in the 37 nt repeat RNA of *C. thermocellum*. Therefore, the two Cas6 enzymes might display variations in repeat recognition.

After pre-crRNA processing, type I-E Cas6 endonucleases stay bound to the 3'-terminal end of mature crRNAs with a hairpin structure, forming the head of Cascade after complex assembly [52, 54]. Consequently, the crRNAs harbor complete 3'-tags. This was also shown for type I-F systems which contain similar structured repeat elements [26]. In contrast to this, all type III as well as type I-A and I-D systems comprise crRNAs with trimmed 3'-ends [43, 135, 136]. Here, the Cas6 homologues deliver the crRNAs to the crRNP complexes and are not associtaed as permanent subunits [43, 136]. It has been proposed that the presence or absence of stem-loop structures in the repeats correlates with the varying Cas6 affinity to the crRNA [53]. In *C. thermocellum,* the unstructured 3'-terminal repeat tags of mature crRNAs are trimmed as described in section 2.2.2. Together with the fact that the type I-B Cascade can be assembled in the absence of Cas6, this indicates that Cas6 is only a temporarily associated member of the complex. It was shown in section 2.4.3 that the Cascade assembly efficiency using a crRNA with a shortened 3'-terminal repeat tag is much lower than for a wildtype crRNA. Nevertheless, all crRNAs that were isolated from *C. thermocellum* and analyzed via RNA-Seq methodology harbored trimmed 3' ends. Hence, it is possible that *in vivo* crRNAs with intact 3'-termini are preferred for the initial Cascade assembly but once a stable complex is formed, Cas6 dissociates which results in the loss of 3'-end protection and crRNA trimming occurs. Additionally, crRNA cycling could occur, wherein the stable complexes would exchange the crRNAs with other trimmed crRNAs. This crRNA cycling ability of the type I-B Cascade could be shown by the electrophoretic mobility shift assays that were performed using fully assembled Cascade complexes that containing unlabeled crRNA and showed binding of the added radiolabeled crRNA.

To address the question, whether crRNAs that are associated with the different type I-B *C. thermocellum* CRISPR systems can be loaded into either of the two Cascade complexes, the investigated recombinant type I-B “Hmari” Cascade was assembled with Cas6 Cthe_2303 and crRNA of the type I-B “Tneap” system. Cas6 cleavage of the pre-crRNA substrate with 30 nt long repeats yields a mature crRNA that contains an 8 nt 5'-terminal repeat tag that is identical in sequence with the 8 nt 5'-terminal tag of the crRNAs associated with the “Hmari” CRISPR system. The experiment revealed an assembled type I-B “Hmari” complex including the Cas6 endonuclease of type I-B “Tneap” with low amounts of the respective crRNA. Compared to the wildtype assembly, a great amount of protein elutes at the void volume indicating protein complex sizes >600 kDa. The corresponding fractions A7 / A8 contain oligomerized Cas7 protein. Cas7 proteins have been shown to multimerize by RNA binding independent of the sequence, forming long helical filaments [43]. It is possible that Cas7 oligomerized along unspecific *E. coli* RNA that was co-purified with the recombinantly produced Cas proteins that is too long to be detected in the performed urea-PAGE. A relatively large amount of monomeric Cas7 protein can be observed and only a small amount of crRNA can be detected in the urea-PAGE. This indicates low affinity of the Cas proteins to the respective crRNA. Unbound crRNA was probably lost during concentration of the assembly mixture volume, prior to gel-filtration chromatography. The crRNA used for the assembly contains a spacer sequence of a CRISPR locus associated with the type I-B “Tneap” Cas6 endonuclease Cthe_2303 and therefore differs from the spacer sequence of the crRNAs used in all other assembly experiments. It has been proposed that the variations of crRNA abundance observed *in vivo* is influenced by varying processing and Cascade loading efficiency of crRNAs due to the spacer content [101, 137, 138]. Hence, the spacer sequence could have an impact on the efficiency of Cascade assembly. A complementary assembly experiment using the type I-B “Tneap” Cascade with the Cas6 endonuclease Cthe_3205 and a crRNA of type I-B “Hmari” could be performed to further investigate the interchangeability of crRNAs between the associated Cascade complexes in *C. thermocellum*.

4.2 Assembly of the type I-B Cascade

Type I Cascade complexes can differ in their composition with permanently and temporarily associated subunits. The type I-E and I-F Cascade complexes consist of the permanent subunits $(Cas6)_1$, $(Cas5)_1$, $(Cas7)_6$ and the respective large and small subunits of type I-E $(Cse1)_1$ and $(Cse2)_2$ or of type I-F $(Csy1)_1$ in an uneven stoichiometry. The Cas3 proteins of both subtypes are transiently recruited for target cleavage [27, 49, 55, 58, 59]. Similarly, the type I-C Cascade is composed of $(Cas7)_6$, the large subunit $(Cas8)_1$, and $(Cas5)_2$, wherein Cas5 also performs pre-crRNA cleavage in a Cas6-like manner [48]. In contrast, in type I-A CRISPR systems, Cas6 is not an integral part of the core complex, whereas Cas3 is a permanent subunit together with Cas7, Cas5 and the proposed small and large subunits Csa5 and Cas8 [43, 46]. The type I-B Cascade of *C. thermocellum* is formed by $(Cas5)_1$, $(Cas7)_6$ and $(Cas8b)_{2.5}$ and shown to bind crRNA. Cas3 was shown not to assemble with the complex (Mastermodule of Franka Schreiner). It has been reported that target DNA binding induces a protein rearrangement of the type I-E large and small subunits that enables Cas3 recruitment [60, 64]. Therefore, it is possible that the type I-B Cas3 protein is also recruited to Cascade after target DNA recognition. The $(Cas6)_1$ endonuclease of *C. thermocellum* seems to be temporarily associated with the complex, as a small amount of protein assembles with the Cascade forming Cas proteins, but is not required for proper complex formation. This is in agreement with the Cascade interprotein cross-link analysis which revealed only one interprotein contact between Cas6 and Cas7 as well as Cas8b. The analysis of the type I-B CRISPR system in *Haloferax volcanii* showed that the type I-B Cascade interference reaction is independent of the presence of Cas6 in the complex *in vivo* [139].

The assembly of recombinant Cascade was shown to strongly depend on the presence of crRNA. Without the addition of crRNA, the experiments only yielded unassembled protein and minor amounts of Cas7 protein multimers interacting with Cas8b, probably caused by the presence of unspecific *E. coli* RNA. So far, only the Cascade complexes from type I-A of *Thermoproteus tenax* and from type I-E of *E. coli* were shown to be produced without the presence of crRNA [140, 141].

In general, type I and type III crRNP complexes have been observed to display a common core architecture that is built up by the crRNA binding proteins that were classified as RAMPs (repeat associated mysterious proteins) [57, 79]. The crRNA-binding helical backbone of the crRNP complexes are composed of six Cas7 homologues that interact with one Cas5-like protein. This stoichiometry was confirmed for the type I-B Cascade complex of *C. thermocellum*. Furthermore, Cas7 and Cas5 were identified to be arranged in close proximity within the complex. The structure prediction of Cas5 revealed a thumb and a palm domain which are conserved among Cas5 family members [48, 54, 142, 143]. Cross-links between Cas5 and Cas7 were identified within Cas5 that are mostly located in the thumb domain. This is in agreement with the Cas subunit arrangement of the type I-E Cascade, wherein the thumb domain of Cas5 interacts with the 5'-terminal tag of the crRNA and contacts the adjacent Cas7 subunit at the finger domain. As multiple copies of Cas7 are part of the *C. thermocellum* Cascade, it is not possible to distinguish between cross-linked sites of different Cas7 copies. Nevertheless, cross-links with Cas5 were identified in the predicted Cas7 finger domain. Together, these results indicate a distinct core backbone architecture of the type I-B Cascade. In the type I-E Cascade, Cas6 and Cas5 are the most distal located subunits, forming the head and the tail of the complex [54, 60]. In *C. thermocellum,* one cross-link was identified for Cas6 with Cas8b and Cas7. Interestingly, no cross-link was observed between Cas6 and Cas5. This underlines the assumption that Cas6 is associated with the 3'-terminal end of the crRNA and therefore located most distant to Cas5.

In contrast to the conservation of the Cascade backbone, the large and small subunits of the type I and type III crRNP complexes differ in structure and their number of copies between the subtypes [21]. The respective Cascade subunits are proposed to mediate target interaction and probably adjusted to the requirements of specialized features for diverse target recognition events and the retention of the surveillance effectiveness (ssRNA versus dsDNA targets, various PAMs) [30, 57]. The crystal structure of the type I-E Cascade revealed that Cas5 interacts with the large subunit Cse1, and that both of these subunits are attached to the Cas7 backbone [54, 60]. Similar to this, cross-links were observed between Cas8b, Cas7 and Cas5 in *C. thermocellum*. Interestingly, 2.5 copies of Cas8b were identified in the type I-B Cascade. It has been proposed that in type I-B, I-C, I-D and I-F Cascades, the small subunit is fused to the large subunit [79]. The Cas8b protein of *C. thermocellum* co-purifies with a small protein fragment that was identified to represent the C-terminal part of Cas8b in addition to the full-length protein. This fragment was also detected *in vivo* (Masterthesis of Kristina Rau). Both, the full-length protein and the fragment assemble in the complex. It is possible that the C-terminal part of the protein has a similar function as the small subunit. Interestingly, cross-links between Cas5 and Cas8b can only be observed in the N-terminal part of the Cas8b protein, which could indicate that this part of the protein interacts with Cas5, whereas the C-terminal part of the full-length protein and the protein fragment acts as small subunits. However, bottom-up mass-spectrometric methods such as iBAQ (intensity-based absolute quantification) used for the analysis of the stoichiometry exhibit a limited accuracy in the determination of high protein copy numbers within complexes. This can influence the accuracy of the ratio of proteins that are present in more than one copy, such as Cas8b [115, 116]. Thus, to further investigate the exact copy number of full-length Cas8b protein and the additional Cas8b fragment, an absolute quantification approach (AQUA) is planned in collaboration with Prof. Dr. Henning Urlaub (MPI for biophysical chemistry, Göttingen). In this approach, peptides are chemically synthesized that are identical to N- and C-terminal parts of the protein sequence and contain stable isotopes. These are then used as internal standards in the Cascade complex mixture to estimate the absolute amount of proteins [144]. This approach will provide the exact number of full-length and small fragment Cas8b protein copies. Furthermore, the Cas8b protein was sent to our collaboration partner Dr. Scott Bailey (Johns Hopkins University, USA) for crystallization experiments (Masterthesis of Kristina Rau).

In order to investigate the interaction of Cas proteins with the repeat termini of the crRNA, recombinant Cascade was assembled using crRNAs with modified repeat tags (3'G RNA and 5'G RNA) as described in section 2.4.3. The substitution of the 3'-terminal tag with G residues does not seem to have an impact on Cascade formation as it looks very similar to the wildtype assembly. A distinct protein peak can be observed that represents the assembled Cascade consisting of all four Cas proteins and a significant amount of crRNA. In contrast, the analysis of the Cascade assembly using crRNA with a modified 5'-terminal repeat tag revealed some disparities. A distinct peak in the gel-filtration chromatogram corresponding to Cascade was not observed. Limited amounts of Cascade formed by Cas5, Cas7 and Cas8b but without Cas6 were detected. A large amount of Cas7 protein elutes in the void volume indicating that it likely oligomerized with *E. coli* RNA. The overall amount of extracted crRNA is very low which indicates that it did not assemble efficiently with the proteins and therefore most likely got lost during the concentration of the assembled protein mixture prior to gel-filtration chromatography. This also explains the relatively high amount of unassembled, monomeric Cas7 protein. Hence, the data suggest that the 5'-terminal repeat tag is essential for proper complex formation, whereas the sequence of the 3'-terminal end is dispensable for protein assembly after crRNA maturation *in vitro*. These results are in agreement with the *in vivo* analysis of the type I-B interference reaction of *Haloferax volcanii*, which showed that mutations of the 3'-terminal repeat end of the crRNA do not influence Cascade activity, whereas the manipulation of the 5'-tag causes the loss of interference activity [139]. Furthermore, in the type I-E Cascade, the 5'-terminal end of the crRNA is positioned between Cas5 and the adjacent Cas7 subunit and base-specifically interacts with Cas5 [54]. This verifies the crucial role of the 5'-terminal tag in Cascade backbone assembly.

In general, the *in vitro* analyses of Cas proteins assemblies with modified crRNAs exhibit methodic limitations. Four individual Cas proteins are recombinantly produced in *E. coli* and three of these proteins contain RRM motifs which indicates that they can bind RNA.

As a consequence, it is plausible that Cas6, Cas5 and Cas7 can be unspecifically complexed with *E. coli* RNA. Thus, individual assembly experiments have to be compared to the wildtype assembly. Defective Cascade formation around modified crRNAs is reflected in a relatively low amount of crRNA uptake. This suggests a low affinity to the respective crRNA which results in the observation of monomeric Cas7 protein and the formation of Cas7 oligomers with *E. coli* RNA. In addition to this, Cas6 is present in the assembly using the 3'G RNA, whereas it is absent from the assembly that was performed with the 5'G RNA. According to the data discussed in section 4.1, it would be assumed that the modification of the 3'-terminal tag hinders Cas6 interaction with the crRNA. The fact that Cas6 is not associated with the Cas protein assembly using the crRNA with a modified 5'-terminal tag indicates hindered complex formation. In contrast, the assembly experiment using a modified 3'-terminal tag showed proper Cascade formation, which indicates that Cas6 is rather associated on a protein:protein level to Cas7 than via crRNA contact in the *in vitro* studies. In order to get a better understanding of crRNA:Cas protein interactions, the use of modified crRNAs and subsequent *in vitro* assembly monitoring seems rather challenging as the crRNA binding Cas proteins interact unspecifically with RNA *in vitro*. As the crystallographic structures of the crRNA binding Cas protein homologues are available, it seems plausible to perform a single molecule Förster resonance energy transfer (smFRET) analysis, wherein dye molecules could be incorporated at predicted interaction sites between Cas proteins as well as crRNA, which could be used to monitor the order complex formation. Additionally, this would also give a better understanding on the localization of the crRNA within the Cascade complex [145].

4.3 Cascade interference

Prior to Cascade mediated DNA interference, the target DNA needs to be recognized. First, Cascade scans dsDNA with regard to a potential PAM sequence [59, 65]. Usually, a defined 2-3 bp PAM sequence is identified by Cascade [44, 65]. The type I-B CRISPR system of *Haloferax volcanii* forms an exception though, as six PAM sequences were identified to permit target recognition in the system *in vivo* using plasmid-invader assays [146]. For type I-E, the large subunit Cse1 was shown to mediate PAM recognition [60, 64, 67]. Second, target pairing between spacer and protospacer sequences triggers R-loop formation [147]. The small and large subunits stabilize the formed R-loop and undergo structural rearrangements which results in the recruitment of Cas3 [64, 71]. Two metal ions are required for the HD nuclease domain of Cas3 to nick the target DNA that activates the helicase which is responsible for ATP-dependent dsDNA hydrolysis [72, 148]. As described in section 2.4.5, a Cascade-mediated interference reaction could not be observed for the type I-B Cascade from *C. thermocellum in vitro*. Several factors could be responsible for this.

First, the PAM sequences used in the interference assays could differ from the one that mediates target recognition, as the PAM sequence specific for *C. thermocellum* has not been determined yet. The fact that the PAM sequences that were identified for the type I-B CRISPR systems of *Haloferax volcanii* and *Listeria monocytogenes* differ in sequence, indicates that the PAM is not subtype specific but has to be identified for each organism individually [113, 114]. The CRISPRtarget tool was utilized to identify potential DNA targets of the CRISPR system in *C. thermocellum.* Three potential PAM sequences were identified adjacent to protospacers from different prophages/plasmids of Clostridia species. Out of these, the 3'-AGT-5' PAM sequence represents the most promising hit as it was found in a *Clostridium thermocellum* prophage. However, one viral evasion strategy to escape CRISPR-Cas is to introduce variations in the PAM sequence in the own genome [44]. Thus, it is possible that the PAM sequences identified in the prophages/plasmid represents a modified PAM that can escape Cascade interference. Recently, EMSA experiments performed with Cas8 from *Methanothermobacter thermautotrophicus* showed that the protein responds to a 3'-GGG-5' PAM in dsDNA, which could also be tested for interference assays using recombinant Cascade from *C. thermocellum*.

Second, R-loop stabilization by the large subunit Cas8b of the complex could be inefficient. It was shown that Cas8b co-purifies together with a protein fragment that represents an additional C-terminal part of Cas8b. Consequently, the question arises whether this protein fragment plays a functional role or if it is artificially produced and hinders interference. To exclude internal translation and proteolytic self-cleavage of the protein, mutations were introduced within the region of a potential RBS adjacent to first amino acids of the Cas8b fragment and N-terminal protein truncations were generated. Analysis of the respective mutants revealed that production of the protein fragment is not prevented. Western blot analysis using a poly-clonal anti-Cas8b antibody and *C. thermocellum* cell extract confirmed the production of the additional protein fragment *in vivo* and excludes proteolytic cleavage of Cas8b during heterologous expression in *E. coli*. Interestingly, the expression of a truncated Cas8b protein, lacking the C-terminal part that represents the small fragment only yielded insoluble protein. This could indicate that the small fragment is produced by a split of the full-length Cas8b protein which causes subsequent precipitation of the remaining N-terminal part of the protein (Masterthesis of Kristina Rau). To investigate whether Cas8b is a functional large subunit of Cascade in the presence of the small protein fragment, EMSAs could be performed using Cascade and radiolabeled dsDNA targets to ensure target binding. Furthermore, an enzymatic or chemical footprint analysis specific for ssDNA using Cascade and the respective dsDNA target could confirm proper R-loop formation [49, 149]. Alternatively, once the crystal structure of the Cas8b protein is solved, smFRET could be used to monitor R-loop formation and investigate the interaction between the different copies of Cas8b within the complex [145].

A final explanation for the absence of Cascade DNA interference activity is the possibility that the Cas3 protein of *C. thermocellum* is not active *in vitro*. Recombinant Cas3 was produced with a 6-fold C-terminal His-tag and purified from *E. coli*. It is possible that the protein is misfolded due to the addition of the His-tag, but still remains soluble. Furthermore, the recombinant Cas3 protein could be loaded with unspecific nucleic acids from *E. coli*. The crystal structure of the Cas3 protein from *Thermobifida fusca* revealed ssDNA that was captured in the precleavage state. The DNA was of endogenous origin and copurified with the protein [150]. An additional purification step using ion-exchange chromatography could not be applied to Cas3 of *C. thermocellum*, as the protein started to precipitate shortly after affinity purification. Cas3 contains an N-terminal HD helicase domain which represents the proposed metal-dependent exo- and endonuclease [73]. The crystal structures of different Cas3 homologues revealed the interaction of iron (II), nickel or calcium ions with the HD residues [73, 150, 151]. In addition to this, Cas3 ssDNAse activity was observed with divalent magnesium, manganese, cobalt, copper and zinc ions [72, 73, 150, 152]. Hence, a wide range of cofactors could possibly be required for the activation of the Cas3 HD nuclease [57]. The nuclease assay using a ssDNA substrate with increasing concentrations of Cas3, magnesium and manganese ions showed that even at a 4 µM protein concentration, the substrate is not cleaved efficiently. This could indicate that magnesium and manganese are not the corresponding Cas3 cofactors. Therefore, the Cas3 cleavage assay was performed under varying metal conditions. Efficient DNA degradation could not be observed under any of the used metal conditions. Hence, further metal combinations have to be tested to determine optimal cleavage conditions for Cas3.

4.4 CRISPR RNA regulation patterns

The RNA-Seq analyses of the small RNome profiles of different archaea and bacteria e.g. *Methanococcus maripaludis* [50], *Methanopyrus kandleri* [102], *Methanosarcina acetivorans* [153], *Nanoarchaeum equitans* [154], *Pyrococcus furiosus* [137], *Sulfolobus solfataricus* [155], *Clostridium difficile* [156] and *Thermus thermophilus* [157] revealed that crRNA abundance follows a gradual decline from the leader-proximal to the leader-distal end of the locus. This is in agreement with the crRNA abundance pattern observed for *C. thermocellum*. In addition to the promoters that are located within the leader and drive CRISPR transcription, promoters have also been identified to be located within spacer and repeat sequences. In *Neisseria meningitidis,* repeat embedded promoters can be observed that enable the production of mature crRNAs independent of an RNA processing enzyme [158]. As described in section 2.2.2, CRISPR locus 3 of *C. thermocellum* also encodes a pair of promoter elements within a consecutive repeat and spacer element that initiates internal transcription of the locus. The final base of the 8 nt 5'-terminal tag of the crRNAs that are produced from the subsequent repeats is exchanged. Together, this indicates that the internal promoter might originate from the fusion of two CRISPR loci, wherein part of a leader was incorporated within this region. The production of high amounts of leader-distal crRNAs could facilitate a selective advantage for the host cell in case of an infection by a virus with complementary DNA [101].

The presence of CRISPR transcripts in reverse direction (anti-CRISPR transcripts) has been described for *P. furiosus* [137], *Sulfolobus acidocaldarius* [16] and *S. solfataricus* [155]. In *S. acidocaldarius,* potential BRE/TATA elements were identified in the leader-distal regions of the CRISPR loci that could be responsible for anti-CRISPR transcription. Internal promotion of anti-CRISPR transcription in *P. furiosus* was found to be initiated by a spacer sequence that contains promoter elements. It has been suggested that these anti-crRNAs could protect crRNAs from degradation [159]. In *P. furiosus*, the antisense RNA was shown to be cleaved *in vitro* by the respective type III-B crRNP complex CMR that targets RNA [34, 137]. *C. thermocellum* was also shown to contain antisense RNA transcripts starting either from the leader-distal end or within CRISPR clusters. The production of these antisense

RNAs is most likely driven by highly AT-rich leader-distal and internal spacer sequences. The analysis of the antisense RNAs revealed small anti-crRNAs that comprise a conserved processing pattern resulting in a defined 5'-terminal repeat tag. As described in section 2.2.4, the processing pattern is identical with the *in vitro* RNase III cleavage of anti/crRNA duplex transcripts. Thus, anti-crRNAs are proposed to play a regulatory role. These proposed contrasting roles of antisense RNA transcripts in bacteria and archaea could be explained by the fact that RNase III enzymes are found in bacteria, and were shown to be involve in type II crRNA maturation, whereas RNase III-like domain containing enzymes are not found in archaea [32]. However, antisense CRISPR transcripts appear to be absent in most organisms [101].

The general abundance of anti-crRNAs observed in the RNA-Seq data of *C. thermocellum* is significantly lower than that of their crRNA counterparts. However, it should be noted that anti-crRNAs are not protected by Cas7 oligomers in the cell and might exhibit a shorter half-life. Furthermore, unprocessed antisense transcripts longer than 200 bp are excluded from the RNA-Seq libraries due to the preparation protocol. In fact, crRNAs or precursor crRNAs have to hybridize with long antisense transcripts as RNase III cleavage is double-strand specific and Cas6 does not process single-stranded anti-crRNA precursors. Furthermore, the observed anti-crRNAs *in vivo* most likely result from RNase III cleavage of a duplex formed by untrimmed crRNAs and long antisense transcript, as the processed anti-crRNAs display randomly shortened 3'-ends which is probably caused by exonucleolytic trimming of the unpaired 3'-terminal region of the anti-crRNAs.

It was observed that the abundance of distinct processed anti-crRNAs correlates with the abundance of their crRNA counterparts. This could indicate that anti-crRNAs neutralize the matching crRNAs in the absence of invaders. The assumption results from the observation that in bacteria, antisense RNAs can manipulate the stability of target RNAs by promoting or blocking endo- or exoribonucleolytic cleavage [90]. One of the major endoribonucleases in bacteria that has been linked to mRNA cleavage induced by antisense RNA is the described double-strand specific RNase III [160]. The observation that RNase III cleaves repeat RNA duplexes in *C. thermocellum* indicates that the sequence elements (pb and db) that are required for RNase III reactivity are present. RNase III cleavage of a wildtype repeat RNA duplex yielded a prominent cleavage site and additional minor cleavage products, shifted by one nucleotide. This suggests the presence of different recognition motifs within the substrate.

In order to identify the db- and pb-elements that are recognized by RNaseIII and responsible for repeat duplex processing in *C. thermocellum*, two nucleotides that were shown to influence RNase III cleavage [94] were modified within the pb-elements on each side of the prominent cleavage site. Interestingly, none of the modifications caused complete loss of RNase III cleavage but induced additional cleavage sites, indicating that both sets of recognition motifs facilitate RNase III cleavage. The additional cleavage sites also display the characteristic 2 nt 3'-overhang. This suggests a certain flexibility in the architecture of the db/pb elements that are recognized by RNase III *in vitro*. *E. coli* RNase III cleavage action has been shown to be determined by positive as well as negative sequence elements that cooperate. This has been proposed to establish varying affinities among subatrates [94].

Apart from the potential impact of anti-crRNAs, other factors also have to be considered that could influence the abundance patterns of individual crRNAs in the cell. Cas6 cleavage activity of the repeats could be influenced by adjacent spacer sequences. It was shown in *M. maripaludis* that spacer sequences can influence Cas6 activity *in vitro*, using a set of consecutive spacer-repeat-spacer substrates [138]. In addition to secondary structures within and between spacer sequences of the CRISPR precursors that can influence Cas6 cleavage, loading efficiency as well as stability and turnover of crRNAs within the crRNP complexes could influence the crRNA abundance pattern [101].

The RNA-Seq analyses revealed the presence of a spacer sequence complementary to the host tRNA^{Pro}. Consequently, the question arises how this spacer was incorporated into the CRISPR array as it could mediate self-targeting. In the absence of crRNP complexes, chromosome-derived spacer acquisition has been described, but was observed to occur around 100-fold less frequently than the acquisition of foreign invader DNA [23]. It has been proposed that self-targeting is not a regulatory mechanism but mediates autoimmunity [161]. Another possible explanation is that the acquisition of the tRNA fragment originates from a viral attachment site. Temperate bacteriophages encode an identity block that comprises the crossover segment followed by tDNA which is used for integration into tRNA genes. After integration, tRNA gene function is retained [162]. As the tRNA targeting crRNA could inhibit precursor-tRNA folding, it is very likely that an RNA duplex is formed with the counter-crRNA, comprising the complementary spacer sequence that acts as an “RNA sponge”. As described in section 2.2.4, the spacer RNA hybrid displays a potential target for RNase III cleavage which would facilitate degradation of the harmful crRNA.

Overall, these results show that individual spacer sequences and repeat elements in forward and reverse direction can have an influence on the abundance of individual crRNAs and therefore impact CRISPR-Cas functionality [101].

5. References

1. Brussow, H., and Hendrix, R. W. (2002) Phage genomics: small is beautiful, *Cell 108*, 13-16.
2. Samson, J. E., Magadan, A. H., Sabri, M., and Moineau, S. (2013) Revenge of the phages: defeating bacterial defences, *Nature reviews. Microbiology 11*, 675-687.
3. Labrie, S. J., Samson, J. E., and Moineau, S. (2010) Bacteriophage resistance mechanisms, *Nature reviews. Microbiology 8*, 317-327.
4. Nordstrom, K., and Forsgren, A. (1974) Effect of protein A on adsorption of bacteriophages to Staphylococcus aureus, *Journal of virology 14*, 198-202.
5. Pingoud, A., Fuxreiter, M., Pingoud, V., and Wende, W. (2005) Type II restriction endonucleases: structure and mechanism, *Cell Mol Life Sci 62*, 685-707.
6. 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.
7. Grissa, I., Vergnaud, G., and Pourcel, C. (2007) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats, *BMC bioinformatics 8*, 172.
8. Jansen, R., Embden, J. D., Gaastra, W., and Schouls, L. M. (2002) Identification of genes that are associated with DNA repeats in prokaryotes, *Molecular microbiology 43*, 1565-1575.
9. Zoephel, J., Dwarakanath, S., Richter, H., Plagens, A., and Randau, L. (2012) Substrate generation for endonucleases of CRISPR/cas systems, *Journal of visualized experiments : JoVE*.
10. Ishino, Y., Shinagawa, H., Makino, K., Amemura, M., and Nakata, A. (1987) Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product, *Journal of bacteriology 169*, 5429-5433.
11. Kunin, V., Sorek, R., and Hugenholtz, P. (2007) Evolutionary conservation of sequence and secondary structures in CRISPR repeats, *Genome biology 8*, R61.
12. Lange, S. J., Alkhnbashi, O. S., Rose, D., Will, S., and Backofen, R. (2013) CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems, *Nucleic acids research 41*, 8034-8044.
13. 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, *Journal of molecular evolution 60*, 174-182.
14. 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.
15. Bolotin, A., Quinquis, B., Sorokin, A., and Ehrlich, S. D. (2005) Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin, *Microbiology 151*, 2551-2561.
16. 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, *Molecular microbiology 72*, 259-272.

17. 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, *Molecular microbiology 75*, 1495-1512.
18. Horvath, P., Romero, D. A., Coute-Monvoisin, A. C., Richards, M., Deveau, H., Moineau, S., Boyaval, P., Fremaux, C., and Barrangou, R. (2008) Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus, *Journal of bacteriology 190*, 1401-1412.
19. Deveau, H., Barrangou, R., Garneau, J. E., Labonte, J., Fremaux, C., Boyaval, P., Romero, D. A., Horvath, P., and Moineau, S. (2008) Phage response to CRISPR-encoded resistance in Streptococcus thermophilus, *Journal of bacteriology 190,* 1390-1400.
20. 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 computational biology 1*, e60.
21. 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., van der Oost, J., and Koonin, E. V. (2011) Evolution and classification of the CRISPR-Cas systems, *Nature reviews. Microbiology 9,* 467-477.
22. 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.
23. Yosef, I., Goren, M. G., and Qimron, U. (2012) Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli, *Nucleic acids research 40,* 5569-5576.
24. 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 & development 22*, 3489-3496.
25. Hale, C., Kleppe, K., Terns, R. M., and Terns, M. P. (2008) Prokaryotic silencing (psi)RNAs in Pyrococcus furiosus, *Rna 14*, 2572-2579.
26. 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.
27. 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.
28. 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 research 39*, 9275-9282.
29. 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, *Molecular cell 52*, 124-134.
30. van der Oost, J., Westra, E. R., Jackson, R. N., and Wiedenheft, B. (2014) Unravelling the structural and mechanistic basis of CRISPR-Cas systems, *Nature Reviews Microbiology 12*, 479-492.
31. Koonin, E. V., and Makarova, K. S. (2013) CRISPR-Cas: evolution of an RNA-based adaptive immunity system in prokaryotes, *RNA biology 10*, 679-686.

32. 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.
33. Marraffini, L. A., and Sontheimer, E. J. (2008) CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA, *Science 322*, 1843-1845.
34. 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.
35. Fonfara, I., Le Rhun, A., Chylinski, K., Makarova, K. S., Lecrivain, A. L., Bzdrenga, J., Koonin, E. V., and Charpentier, E. (2014) Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems, *Nucleic acids research 42*, 2577-2590.
36. Horvath, P., Coute-Monvoisin, A. C., Romero, D. A., Boyaval, P., Fremaux, C., and Barrangou, R. (2009) Comparative analysis of CRISPR loci in lactic acid bacteria genomes, *International journal of food microbiology 131*, 62-70.
37. Shah, S. A., and Garrett, R. A. (2011) CRISPR/Cas and Cmr modules, mobility and evolution of adaptive immune systems, *Research in microbiology 162*, 27-38.
38. Makarova, K. S., Grishin, N. V., Shabalina, S. A., Wolf, Y. I., and Koonin, E. V. (2006) A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action, *Biology direct 1*, 7.
39. Nunez, J. K., Lee, A. S., Engelman, A., and Doudna, J. A. (2015) Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity, *Nature 519*, 193-198.
40. Erdmann, S., and Garrett, R. A. (2012) Selective and hyperactive uptake of foreign DNA by adaptive immune systems of an archaeon via two distinct mechanisms, *Molecular microbiology 85*, 1044-1056.
41. Arslan, Z., Hermanns, V., Wurm, R., Wagner, R., and Pul, U. (2014) Detection and characterization of spacer integration intermediates in type I-E CRISPR-Cas system, *Nucleic acids research 42*, 7884-7893.
42. Datsenko, K. A., Pougach, K., Tikhonov, A., Wanner, B. L., Severinov, K., and Semenova, E. (2012) Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system, *Nature communications 3*, 945.
43. Plagens, A., Tripp, V., Daume, M., Sharma, K., Klingl, A., Hrle, A., Conti, E., Urlaub, H., and Randau, L. (2014) In vitro assembly and activity of an archaeal CRISPR-Cas type I-A Cascade interference complex, *Nucleic acids research 42*, 5125-5138.
44. 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 genetics 9*, e1003742.
45. 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, *The EMBO journal 32*, 385-394.
46. Lintner, N. G., Kerou, M., Brumfield, S. K., Graham, S., Liu, H., Naismith, J. H., Sdano, M., Peng, N., She, Q., Copie, V., Young, M. J., White, M. F., and Lawrence, C. M. (2011) Structural and functional characterization of an archaeal

clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE), *The Journal of biological chemistry 286*, 21643-21656.

47. Wang, R., Zheng, H., Preamplume, G., Shao, Y., and Li, H. (2012) The impact of CRISPR repeat sequence on structures of a Cas6 protein-RNA complex, *Protein science : a publication of the Protein Society 21*, 405-417.
48. Nam, K. H., Haitjema, C., Liu, X., Ding, F., Wang, H., DeLisa, M. P., and Ke, A. (2012) Cas5d protein processes pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvulg CRISPR-Cas system, *Structure 20*, 1574-1584.
49. 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., Beijer, M. R., Barendregt, A., Zhou, K., Snijders, A. P., Dickman, M. J., Doudna, J. A., Boekema, E. J., Heck, A. J., van der Oost, J., and Brouns, S. J. (2011) Structural basis for CRISPR RNA-guided DNA recognition by Cascade, *Nature structural & molecular biology 18*, 529-536.
50. 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 research 40*, 9887-9896.
51. 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, *Nature structural & molecular biology 18*, 688-692.
52. Sashital, D. G., Jinek, M., and Doudna, J. A. (2011) An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3, *Nature structural & molecular biology 18*, 680-687.
53. Niewoehner, O., Jinek, M., and Doudna, J. A. (2014) Evolution of CRISPR RNA recognition and processing by Cas6 endonucleases, *Nucleic acids research 42*, 1341-1353.
54. Jackson, R. N., Golden, S. M., van Erp, P. B., Carter, J., Westra, E. R., Brouns, S. J., van der Oost, J., Terwilliger, T. C., Read, R. J., and Wiedenheft, B. (2014) Structural biology. Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli, *Science 345*, 1473-1479.
55. 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., and Doudna, J. A. (2011) RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions, *Proceedings of the National Academy of Sciences of the United States of America 108*, 10092-10097.
56. Sternberg, S. H., Haurwitz, R. E., and Doudna, J. A. (2012) Mechanism of substrate selection by a highly specific CRISPR endoribonuclease, *Rna 18*, 661-672.
57. Plagens, A., Richter, H., Charpentier, E., and Randau, L. (2015) DNA and RNA interference mechanisms by CRISPR-Cas surveillance complexes, *FEMS microbiology reviews*.
58. Richter, C., Gristwood, T., Clulow, J. S., and Fineran, P. C. (2012) In vivo protein interactions and complex formation in the Pectobacterium atrosepticum subtype I-F CRISPR/Cas System, *PloS one 7*, e49549.
59. 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., de Vos, W. M., Dame, R. T., de Vries, R., Brouns, S. J., and van der Oost, J. (2012) CRISPR immunity relies on

the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3, *Molecular cell 46*, 595-605.

60. Mulepati, S., Heroux, A., and Bailey, S. (2014) Structural biology. Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target, *Science 345*, 1479-1484.

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

62. 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.

63. Zhao, H., Sheng, G., Wang, J., Wang, M., Bunkoczi, G., Gong, W., Wei, Z., and Wang, Y. (2014) Crystal structure of the RNA-guided immune surveillance Cascade complex in Escherichia coli, *Nature 515*, 147-150.

64. Hochstrasser, M. L., Taylor, D. W., Bhat, P., Guegler, C. K., Sternberg, S. H., Nogales, E., and Doudna, J. A. (2014) CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference, *Proceedings of the National Academy of Sciences of the United States of America 111*, 6618-6623.

65. Rollins, M. F., Schuman, J. T., Paulus, K., Bukhari, H. S., and Wiedenheft, B. (2015) Mechanism of foreign DNA recognition by a CRISPR RNA-guided surveillance complex from Pseudomonas aeruginosa, *Nucleic acids research 43*, 2216-2222.

66. Shah, S. A., Erdmann, S., Mojica, F. J., and Garrett, R. A. (2013) Protospacer recognition motifs: mixed identities and functional diversity, *RNA biology 10*, 891-899.

67. Sashital, D. G., Wiedenheft, B., and Doudna, J. A. (2012) Mechanism of foreign DNA selection in a bacterial adaptive immune system, *Molecular cell 46*, 606-615.

68. 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, *The Journal of biological chemistry 287*, 22445-22449.

69. 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, *Proceedings of the National Academy of Sciences of the United States of America 108*, 10098-10103.

70. Fineran, P. C., Gerritzen, M. J., Suarez-Diez, M., Kunne, T., Boekhorst, J., van Hijum, S. A., Staals, R. H., and Brouns, S. J. (2014) Degenerate target sites mediate rapid primed CRISPR adaptation, *Proceedings of the National Academy of Sciences of the United States of America 111*, E1629-1638.

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

72. 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, *The EMBO journal 30*, 1335-1342.

73. 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), *The Journal of biological chemistry 286*, 31896-31903.

74. Elmore, J. R., Yokooji, Y., Sato, T., Olson, S., Glover, C. V., 3rd, Graveley, B. R., Atomi, H., Terns, R. M., and Terns, M. P. (2013) Programmable plasmid interference by the CRISPR-Cas system in Thermococcus kodakarensis, *RNA biology 10*, 828-840.

75. Li, M., Liu, H., Han, J., Liu, J., Wang, R., Zhao, D., Zhou, J., and Xiang, H. (2013) Characterization of CRISPR RNA biogenesis and Cas6 cleavage-mediated inhibition of a provirus in the haloarchaeon Haloferax mediterranei, *Journal of bacteriology 195*, 867-875.

76. Brendel, J., Stoll, B., Lange, S. J., Sharma, K., Lenz, C., Stachler, A. E., Maier, L. K., Richter, H., Nickel, L., Schmitz, R. A., Randau, L., Allers, T., Urlaub, H., Backofen, R., and Marchfelder, A. (2014) A complex of Cas proteins 5, 6, and 7 is required for the biogenesis and stability of clustered regularly interspaced short palindromic repeats (crispr)-derived rnas (crrnas) in Haloferax volcanii, *The Journal of biological chemistry 289*, 7164-7177.

77. Majumdar, S., Zhao, P., Pfister, N. T., Compton, M., Olson, S., Glover, C. V., 3rd, Wells, L., Graveley, B. R., Terns, R. M., and Terns, M. P. (2015) Three CRISPR-Cas immune effector complexes coexist in Pyrococcus furiosus, *Rna*.

78. Steitz, T. A. (1999) DNA polymerases: structural diversity and common mechanisms, *The Journal of biological chemistry 274*, 17395-17398.

79. 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, *Biology direct 6*, 38.

80. Calin-Jageman, I., and Nicholson, A. W. (2003) RNA structure-dependent uncoupling of substrate recognition and cleavage by Escherichia coli ribonuclease III, *Nucleic acids research 31*, 2381-2392.

81. Nicholson, A. W. (1999) Function, mechanism and regulation of bacterial ribonucleases, *FEMS microbiology reviews 23*, 371-390.

82. Robertson, H. D., Webster, R. E., and Zinder, N. D. (1968) Purification and properties of ribonuclease III from Escherichia coli, *The Journal of biological chemistry 243*, 82-91.

83. Li, F., and Ding, S. W. (2006) Virus counterdefense: diverse strategies for evading the RNA-silencing immunity, *Annual review of microbiology 60*, 503-531.

84. Blaszczyk, J., Gan, J., Tropea, J. E., Court, D. L., Waugh, D. S., and Ji, X. (2004) Noncatalytic assembly of ribonuclease III with double-stranded RNA, *Structure 12*, 457-466.

85. Filippov, V., Solovyev, V., Filippova, M., and Gill, S. S. (2000) A novel type of RNase III family proteins in eukaryotes, *Gene 245*, 213-221.

86. Ji, X. (2008) The mechanism of RNase III action: how dicer dices, *Current topics in microbiology and immunology 320*, 99-116.

87. van Rij, R. P., and Andino, R. (2006) The silent treatment: RNAi as a defense against virus infection in mammals, *Trends in biotechnology 24*, 186-193.

88. Langenberg, W., Zhang, L., Court, D., Giunchedi, L., and Mitra, A. Transgenic tobacco plants expressing the bacterial mc gene resist virus infection.

89. Stead, M. B., Marshburn, S., Mohanty, B. K., Mitra, J., Pena Castillo, L., Ray, D., van Bakel, H., Hughes, T. R., and Kushner, S. R. (2011) Analysis of Escherichia

coli RNase E and RNase III activity in vivo using tiling microarrays, *Nucleic acids research 39*, 3188-3203.

90. Thomason, M. K., and Storz, G. (2010) Bacterial antisense RNAs: how many are there, and what are they doing?, *Annual review of genetics 44*, 167-188.
91. Grunberg-Manago, M. (1999) Messenger RNA stability and its role in control of gene expression in bacteria and phages, *Annual review of genetics 33*, 193-227.
92. Court, D. L., Gan, J., Liang, Y. H., Shaw, G. X., Tropea, J. E., Costantino, N., Waugh, D. S., and Ji, X. (2013) RNase III: Genetics and function; structure and mechanism, *Annual review of genetics 47*, 405-431.
93. Gan, J., Tropea, J. E., Austin, B. P., Court, D. L., Waugh, D. S., and Ji, X. (2006) Structural insight into the mechanism of double-stranded RNA processing by ribonuclease III, *Cell 124*, 355-366.
94. Pertzev, A. V., and Nicholson, A. W. (2006) Characterization of RNA sequence determinants and antideterminants of processing reactivity for a minimal substrate of Escherichia coli ribonuclease III, *Nucleic acids research 34*, 3708-3721.
95. Gasiunas, G., Barrangou, R., Horvath, P., and Siksnys, V. (2012) Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria, *Proceedings of the National Academy of Sciences of the United States of America 109*, E2579-2586.
96. Zverlov, V. V., Hiegl, W., Köck, D. E., Kellermann, J., Köllmeier, T., and Schwarz, W. H. Hydrolytic bacteria in mesophilic and thermophilic degradation of plant biomass.
97. Freier, D., Mothershed, C. P., and Wiegel, J. (1988) Characterization of Clostridium thermocellum JW20, *Applied and environmental microbiology 54*, 204-211.
98. Lynd, L. R., Weimer, P. J., van Zyl, W. H., and Pretorius, I. S. (2002) Microbial cellulose utilization: fundamentals and biotechnology, *Microbiology and molecular biology reviews : MMBR 66*, 506-577, table of contents.
99. Demain, A. L., Newcomb, M., and Wu, J. H. (2005) Cellulase, clostridia, and ethanol, *Microbiology and molecular biology reviews : MMBR 69*, 124-154.
100. Rousseau, C., Gonnet, M., Le Romancer, M., and Nicolas, J. (2009) CRISPI: a CRISPR interactive database, *Bioinformatics 25*, 3317-3318.
101. Zoephel, J., and Randau, L. (2013) RNA-Seq analyses reveal CRISPR RNA processing and regulation patterns, *Biochemical Society transactions 41*, 1459-1463.
102. 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 research 41*, 6250-6258.
103. Macrae, I. J., Zhou, K., Li, F., Repic, A., Brooks, A. N., Cande, W. Z., Adams, P. D., and Doudna, J. A. (2006) Structural basis for double-stranded RNA processing by Dicer, *Science 311*, 195-198.
104. Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S., and Kim, V. N. (2003) The nuclear RNase III Drosha initiates microRNA processing, *Nature 425*, 415-419.
105. Ketting, R. F., Fischer, S. E., Bernstein, E., Sijen, T., Hannon, G. J., and Plasterk, R. H. (2001) Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans, *Genes & development 15*, 2654-2659.

106. Gan, J., Shaw, G., Tropea, J. E., Waugh, D. S., Court, D. L., and Ji, X. (2008) A stepwise model for double-stranded RNA processing by ribonuclease III, *Molecular microbiology 67*, 143-154.
107. Sun, W., Pertzev, A., and Nicholson, A. W. (2005) Catalytic mechanism of Escherichia coli ribonuclease III: kinetic and inhibitor evidence for the involvement of two magnesium ions in RNA phosphodiester hydrolysis, *Nucleic acids research 33*, 807-815.
108. Sun, W., and Nicholson, A. W. (2001) Mechanism of action of Escherichia coli ribonuclease III. Stringent chemical requirement for the glutamic acid 117 side chain and Mn2+ rescue of the Glu117Asp mutant, *Biochemistry 40*, 5102-5110.
109. Franch, T., Thisted, T., and Gerdes, K. (1999) Ribonuclease III processing of coaxially stacked RNA helices, *The Journal of biological chemistry 274*, 26572-26578.
110. Zhang, K., and Nicholson, A. W. (1997) Regulation of ribonuclease III processing by double-helical sequence antideterminants, *Proceedings of the National Academy of Sciences of the United States of America 94*, 13437-13441.
111. Engler, C., Gruetzner, R., Kandzia, R., and Marillonnet, S. (2009) Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes, *PloS one 4*, e5553.
112. Biswas, A., Gagnon, J. N., Brouns, S. J., Fineran, P. C., and Brown, C. M. (2013) CRISPRTarget: bioinformatic prediction and analysis of crRNA targets, *RNA biology 10*, 817-827.
113. Stoll, B., Maier, L. K., Lange, S. J., Brendel, J., Fischer, S., Backofen, R., and Marchfelder, A. (2013) Requirements for a successful defence reaction by the CRISPR-Cas subtype I-B system, *Biochemical Society transactions 41*, 1444-1448.
114. Di, H., Ye, L., Yan, H., Meng, H., Yamasak, S., and Shi, L. (2014) Comparative analysis of CRISPR loci in different Listeria monocytogenes lineages, *Biochemical and biophysical research communications 454*, 399-403.
115. Wilhelm, B. G., Mandad, S., Truckenbrodt, S., Krohnert, K., Schafer, C., Rammner, B., Koo, S. J., Classen, G. A., Krauss, M., Haucke, V., Urlaub, H., and Rizzoli, S. O. (2014) Composition of isolated synaptic boutons reveals the amounts of vesicle trafficking proteins, *Science 344*, 1023-1028.
116. Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011) Global quantification of mammalian gene expression control, *Nature 473*, 337-342.
117. Christian, H., Hofele, R. V., Urlaub, H., and Ficner, R. (2014) Insights into the activation of the helicase Prp43 by biochemical studies and structural mass spectrometry, *Nucleic acids research 42*, 1162-1179.
118. Kelley, L. A., and Sternberg, M. J. (2009) Protein structure prediction on the Web: a case study using the Phyre server, *Nature protocols 4*, 363-371.
119. Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2015) The I-TASSER Suite: protein structure and function prediction, *Nature methods 12*, 7-8.
120. Sambrook J, F. E. a. M. T. (1989) Molecular Cloning - A Laboratory Manual, *2nd Ed.*
121. Ausubel, F. M. (1987) Current protocols in molecular biology, *Wiley*.
122. Lynd, L. R., Grethlein, H. E., and Wolkin, R. H. (1989) Fermentation of Cellulosic Substrates in Batch and Continuous Culture by Clostridium thermocellum, *Applied and environmental microbiology 55*, 3131-3139.

123. Hanahan, D. (1983) Studies on transformation of Escherichia coli with plasmids, *Journal of molecular biology 166*, 557-580.

124. Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., and Erlich, H. (1986) Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction, *Cold Spring Harbor symposia on quantitative biology 51 Pt 1*, 263-273.

125. Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B., and Erlich, H. A. (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase, *Science 239*, 487-491.

126. Wallace, R. B., Shaffer, J., Murphy, R. F., Bonner, J., Hirose, T., and Itakura, K. (1979) Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch, *Nucleic acids research 6*, 3543-3557.

127. Bankier, A. T., Weston, K. M., and Barrell, B. G. (1987) Random cloning and sequencing by the M13/dideoxynucleotide chain termination method, *Methods in enzymology 155*, 51-93.

128. Sanger, F., Nicklen, S., and Coulson, A. R. (1992) DNA sequencing with chain-terminating inhibitors. 1977, *Biotechnology 24*, 104-108.

129. Laemmli, U. K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4, *Nature 227*, 680-685.

130. Weber, K., and Osborn, M. (2006) SDS-PAGE to determine the molecular weight of proteins: The work of Klaus Weber and Mary Osborn - The reliability of molecular weight determinations by dodecyl sulfate-polyacrylamide gel electrophoresis (reprinted from J.Biol. Chem. vol. 244, pg. 4406-4412, 1969), *Journal of Biological Chemistry 281*.

131. Bradford, M. M. (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding, *Analytical biochemistry 72*, 248-254.

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

133. 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.

134. Zuker, M., and Stiegler, P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information, *Nucleic acids research 9*, 133-148.

135. Hein, S., Scholz, I., Voss, B., and Hess, W. R. (2013) Adaptation and modification of three CRISPR loci in two closely related cyanobacteria, *RNA biology 10*, 852-864.

136. 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, *Proceedings of the National Academy of Sciences of the United States of America 108*, 21218-21222.

137. Hale, C. R., Majumdar, S., Elmore, J., Pfister, N., Compton, M., Olson, S., Resch, A. M., Glover, C. V., 3rd, Graveley, B. R., Terns, R. M., and Terns, M. P. (2012) Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs, *Molecular cell 45*, 292-302.

138. Richter, H., Lange, S. J., Backofen, R., and Randau, L. (2013) Comparative analysis ofCas6b processing and CRISPR RNA stability, *RNA biology 10*, 700-707.

139. Maier, L. K., Stachler, A. E., Saunders, S. J., Backofen, R., and Marchfelder, A. (2015) An active immune defense with a minimal CRISPR (clustered regularly interspaced short palindromic repeats) RNA and without the Cas6 protein, *The Journal of biological chemistry 290*, 4192-4201.

140. Plagens, A., and Randau, L. (2015) In Vitro Co-reconstitution of Cas Protein Complexes, *Methods in molecular biology 1311*, 23-33.

141. Beloglazova, N., Kuznedelov, K., Flick, R., Datsenko, K. A., Brown, G., Popovic, A., Lemak, S., Semenova, E., Severinov, K., and Yakunin, A. F. (2015) CRISPR RNA binding and DNA target recognition by purified Cascade complexes from Escherichia coli, *Nucleic acids research 43*, 530-543.

142. Koo, Y., Ka, D., Kim, E. J., Suh, N., and Bae, E. (2013) Conservation and variability in the structure and function of the Cas5d endoribonuclease in the CRISPR-mediated microbial immune system, *Journal of molecular biology 425*, 3799-3810.

143. Shao, Y., Cocozaki, A. I., Ramia, N. F., Terns, R. M., Terns, M. P., and Li, H. (2013) Structure of the Cmr2-Cmr3 subcomplex of the Cmr RNA silencing complex, *Structure 21*, 376-384.

144. Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W., and Gygi, S. P. (2003) Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS, *Proceedings of the National Academy of Sciences of the United States of America 100*, 6940-6945.

145. Nagy, J., Grohmann, D., Cheung, A. C., Schulz, S., Smollett, K., Werner, F., and Michaelis, J. (2015) Complete architecture of the archaeal RNA polymerase open complex from single-molecule FRET and NPS, *Nature communications 6*, 6161.

146. 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, *The Journal of biological chemistry 287*, 33351-33363.

147. Szczelkun, M. D., Tikhomirova, M. S., Sinkunas, T., Gasiunas, G., Karvelis, T., Pschera, P., Siksnys, V., and Seidel, R. (2014) Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes, *Proceedings of the National Academy of Sciences of the United States of America 111*, 9798-9803.

148. 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, *The Journal of biological chemistry 288*, 22184-22192.

149. Raghavan, S. C., Tsai, A., Hsieh, C. L., and Lieber, M. R. (2006) Analysis of non-B DNA structure at chromosomal sites in the mammalian genome, *Methods in enzymology 409*, 301-316.

150. Huo, Y., Nam, K. H., Ding, F., Lee, H., Wu, L., Xiao, Y., Farchione, M. D., Jr., Zhou, S., Rajashankar, K., Kurinov, I., Zhang, R., and Ke, A. (2014) Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation, *Nature structural & molecular biology 21*, 771-777.

151. 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, *The EMBO journal 30*, 4616-4627.

152. Gong, B., Shin, M., Sun, J., Jung, C. H., Bolt, E. L., van der Oost, J., and Kim, J. S. (2014) Molecular insights into DNA interference by CRISPR-associated nuclease-

helicase Cas3, *Proceedings of the National Academy of Sciences of the United States of America 111*, 16359-16364.

153. Nickel, L., Weidenbach, K., Jager, D., Backofen, R., Lange, S. J., Heidrich, N., and Schmitz, R. A. (2013) Two CRISPR-Cas systems in Methanosarcina mazei strain Go1 display common processing features despite belonging to different types I and III, *RNA biology 10*, 779-791.
154. Randau, L. (2012) RNA processing in the minimal organism Nanoarchaeum equitans, *Genome biology 13*, R63.
155. Wurtzel, O., Sapra, R., Chen, F., Zhu, Y., Simmons, B. A., and Sorek, R. (2010) A single-base resolution map of an archaeal transcriptome, *Genome research 20*, 133-141.
156. Soutourina, O. A., Monot, M., Boudry, P., Saujet, L., Pichon, C., Sismeiro, O., Semenova, E., Severinov, K., Le Bouguenec, C., Coppee, J. Y., Dupuy, B., and Martin-Verstraete, I. (2013) Genome-wide identification of regulatory RNAs in the human pathogen Clostridium difficile, *PLoS genetics 9*, e1003493.
157. Juranek, S., Eban, T., Altuvia, Y., Brown, M., Morozov, P., Tuschl, T., and Margalit, H. (2012) A genome-wide view of the expression and processing patterns of Thermus thermophilus HB8 CRISPR RNAs, *Rna 18*, 783-794.
158. Zhang, Y., Heidrich, N., Ampattu, B. J., Gunderson, C. W., Seifert, H. S., Schoen, C., Vogel, J., and Sontheimer, E. J. (2013) Processing-independent CRISPR RNAs limit natural transformation in Neisseria meningitidis, *Molecular cell 50*, 488-503.
159. Garrett, R. A., Shah, S. A., Vestergaard, G., Deng, L., Gudbergsdottir, S., Kenchappa, C. S., Erdmann, S., and She, Q. (2011) CRISPR-based immune systems of the Sulfolobales: complexity and diversity, *Biochemical Society transactions 39*, 51-57.
160. Carpousis, A. J., Luisi, B. F., and McDowall, K. J. (2009) Endonucleolytic initiation of mRNA decay in Escherichia coli, *Progress in molecular biology and translational science 85*, 91-135.
161. Stern, A., Keren, L., Wurtzel, O., Amitai, G., and Sorek, R. (2010) Self-targeting by CRISPR: gene regulation or autoimmunity?, *Trends in genetics : TIG 26*, 335-340.
162. Williams, K. P. (2002) Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies, *Nucleic acids research 30*, 866-875.

7. Appendix

A1: Nucleotide sequences of *C. thermocellum cas* genes codon-optimized for *E. coli* usage

Sequence *cas*3

```
ATGCTGCTGTCTCACCCGGGCAAACTGCTGCTGGACCACCTGAAAAATGTGTTCCTGATCGGCGAC
TGTATCCTGATGCAGAAAAAGACCGAATTTGAAAGTTTCTCCGAACATGATATTCGTCAGCTGAAC
AAACTGAATCTGCTGACCCACGATCTGGGCAAAGCAACGAGCTATTTTCAGGACTACATTCGCAAC
CTGGATGACAATACCCAGAAAAACGATGAACGTAAACGCCATGGCCTGCTGTCTGGTGTTCTGAGT
TTTAAAATCGTGAACGCAGTTATGAAAAACGAAATTCTGGCTTTCCTGTCCTATATGGTGGTTTCA
AAACATCACGGCGAACTGGATGACTTTACGAATTTCATCAGCGTGATTTCTGGTGATGAGAAAAAC
AAAACCCTGCTGAAACTGCAGTTTGAAAGCATCGATAAAGGCAAACTGCAAGAAGTGATTACCCGT
CTGGGTATCGACTTTGATATTCTGTCCTATACGGTTGATGAATTTGAAAACGACATCGATTACATC
ACCTCACGTAAAGTCCGCAAAAAAGTGAAAGAACTGATGGGTCTGGAAATCTTCCTGCTGATTGAT
TACCTGTTTAGCCTGCTGATCTTCAGCGACAAACTGGAAGCAATTTACAACTCGGAAAACATGAAC
ATCGAAGAATTTATCGAGAAAAACACCAATCGTCCGCGCATTAGCTCTGATTGCGTGGACAAATTC
AAAAAATCTCTGGAAATCAAAAACATTCGCATGGCGGAATGGCGTAACCGCGCCTATCAGGATGTT
CTGGGCTCTGTCGAAAATCTGAGTCTGGAAGAAAAAATCCTGTCGATTAACCTGCCGACGGGCAGC
GGTAAAACCCTGACGGCTCTGAAAGCGGCCCTGCGTCTGAAAGAACGCCTGATTGAAGAAAAAGGT
TATAATCCGCGTATTATCTACGTGCTGCCGTTTACCTCCATTATCGAACAGAATTTTGATGTGTTC
CAAAAAGTTCTGGGCACCACGGAATCTAACGTTCTGCTGAAACATCACTATCTGAGTCAGCGCGTC
TACCAATGGGAAAAAGAAGGTGAAATTGAATCACTGTCGGATACCGTTAGCGAACATCTGGTCGAA
AGTTGGGACTCCGAAATCGTCGTGTCTACCTTTGTGCAGCTGCTGCATAGTATTTTCACGAACCGT
AACCGCAAACTGAAAAAATTCCACAACATCGTCAATTCGATCATCATTCTGGATGAAGTGCAGAGC
ATTCCGCATCGTTATTGGAATCTGGTTCGCGAAACCTTTTCTGCAATGGCTAAATACCTGAACTGC
CACTTTATTTTCATGACCGCAACGATGCCGCTGATCTTCTCGGAAGAAAACAAAGAAATCTACGAA
CTGGTGAAAGATAAACGTAAATACTTCGAAGAATTTAGCCGTATCACCATTGACGCTAAACGCCTG
ACGGATAAAACCACGCTGGACGAATATAAAACCCTGCTGCTGGATGACATTCTGCGTTACAAAAAA
GATGACTTTCTGATCGTTATGAACACCATTCGCACGTCCATCGAAATCTACTCATTCATCAAAGAA
GAACTGAAAGATGAAGCCGAAATCTACTACCTGTCTACGAACATCATCCCGAAAGAACGTCTGGAT
CGCATTAAGAAAATTAAAGAATCCAAAAACCGTAAAATCATTGTGTCAACCCAGATGATCGAAGCA
GGCGTGGACATCGATATTGACCGTGTTTATCGCGATTTTGGCCCGATGGACAGTATTAATCAAACG
GCTGGTCGTTGTAACCGCGAATGGGGCGATAAAAAAGGTCTGGTCACCCTGGTGAATCTGGTTAAC
GAAAATCATCACTCTCGCCCGTACGCGACCTATATCTACGATAACGTTCTGATTGAAGAAACGAAA
AAAGCCCTGAGTGGTCTGGAAATGATCGAAGAAAAAGAAATCTTCCATCTGGCGGAAAAATATTAC
GTGGGCCTGAAATCACACGGTTCGGATGAAAGCCAGAAACTGCTGGACTGCATCAAAGAACTGCGT
TATCGCGAAGCCTTTGAATGTGGCAAAGATGACAAAAATAGCGGTGTCTTCGAACTGATTCGTCAG
GATTTTAACACGGTTGACGTCTTCATCGAAATTGATGACGATGCGACCGAAGTGTGGCAGGAATAC
CAATCTATTAAGAAAATTAAAGACCGCTTCGAACGTAAACGCAAATTCAACCAGTTCAAAAAAGAC
CTGTACATGTACGTTCTGTCACTGCCGGAATTTGCCGTCCGTAAACAAGTGGATATTGACGAAAAA
GATATTACCTTCATCAACCGCGAAATGGTCTTCAACACCTACGATAAAGATACGGGCTTCATGCGT
GACCTGGAAAAAGATTACTTCTTCTAA
```

Sequence *cas*5

```
ATGGTGGAAAAATACCTGGTGTTCGACATTTCGGCAAGCTATGGCCACTTCAAAAAACCGTACACG
ACGACGAGCCCGCTGACCTACTCAATTCCGACCCGCACGGCAGTTTCGGGCATTATCGCGGCCGTC
CTGGGCTTTGGTAAAGAAGATTATCAGGAACATTTCACCAAACCGCAAGCAAAAATTGCTATCGGT
```

ATTCGCAACCCGGTCAAAAAAGTGCGTATCAGCGAAAACCTGATCAACACCAAAAAATCTATGAAC
ATCATCCACGAACGCACCCAGATCAAAATTGAATTTCTGAAAGATGCGTGCTATCGTATCTACTTC
ACCCATACGGACAAACAAATTTACGAACGCCTGAAAGAATCCCTGAAAGAACACCGTACCGTGTAT
ACGATCAGTATGGGCCTGTCCGAAAACCTGGCCAATTACACCTTTACGGGCGAATTTGATGGTCGC
GAAGTGAAAGGCAACAAAGAAATCGTTGAATTTAGCTCTGTGGTTCCGCTGGATATTCTGAAAAAA
GGTGACATTGAATTTGAAGATGACCGTGAATATTTCACCGAAACGATCCCGGTTGAAATGGACGCC
GAACGTAACATTAATGAATACCGCGAAGTCCTGTTCGAACGTAATGGTTGTAAAGTTCGTGCTCGT
GTGGACTCGTACATTAAAATTGAAGGTATTGATGAAAACATTCTGATTATTTAA

Sequence *cas*6

ATGGACATCAAAATCCTGACCGTGCGTCTGGAAGGCAACCGTGTGGAATCCCGTGACATCCCGAAA
ATCCGTGGCTACCTGGCTGGCCGTTTCCCGCAGTATCTGGAACTGCATAACCACCTGGGCGAAAAC
AAATTCAACTACGGTTACCCGGTTATCCAATACAAAAGTATTAACGGCGTCCCGAATATCCTGGCG
ATTAATGAAGCCTCCAAAATTCTGATCGATATTTTCTATGACGTGAAAGAAATCGATATGAAAGAC
AAAGTTATGTCAATTCTGGAAAAAGGTTACGTCCTGAAAACCCGTGAACTGAAAACCACGGAACGT
ATGATCGAATATCGCTTCCTGTCGCCGTGGCTGGCACTGAACCAGGAAAACTACGAAAAATTCATC
AACAGCGATTCTGGCGAACGCGTTGAAATCCTGAAGAAAATTCTGACCGGTAATATCCTGTCTATG
GCGAAAGGCCTGGGTTATTGCGTGGACAAACCGATTGAAGTGTTTGTTAACCTGCGTCCGGTCCAA
GTGAACTACAAAAATCGCAAAATGATCGGCTTCAAGGGTGAATTTATCACGAACTTCATCATCCCG
GATTATCTGGGCCTGGGTAAATCAGTCGCCCGTGGTTTCGGCACCGTGGAACGTGTGGGTAAAGGT
TTCGTGTAA

Sequence *cas*7

ATGATCAAAAACCGCCAAGAAATCCTGTTCCTGTATGACGTTACCGATGCTAATCCGAATGGTGAC
CCGCTGGACGAAAATAAACCGCGCATTGACGAAGAAACCGGTATTAACATCGTTACGGATGTCCGT
CTGAAACGCACCATCCGTGACTATCTGTACGATTATAAAGGCTTTGACGGTAGTAATGGCAAAGAT
ATTTTCGTGCGCGAAATCGAATCCGAAAAAGGCGGTATTAAAGATGGCAAAGCACGTGCTAAAGAC
TTCAACGAAAACGTCGATGAAATCCTGCAGAAAGCGATTGATATTCGCCTGTTCGGCGGTGTGATT
CCGCTGGATAAAGCCTCAATCACCTTTACGGGTCCGGTGCAATTCAACATGGGCCGTTCGCTGAAC
AAAGTTAACCTGAAACATATCAAAGGCACCGGTGCGTTTGCCAGCGGCGAGGGTAAAGCACAGAAA
ACGTTCCGCGAAGAATACATCGTTCCGTATTCTATTATCGCGTTTCACGGTATTATCAACGAAAAT
GCGGCCAAACGTACCGGCCTGACGGATGAAGACGTCGATCTGCTGGATGACGCGATGTGGAACGGT
ACCAAAAATCTGATTACGCGCTCCAAAATGGGCCATATGCCGCGCCTGATGCTGCGTGTGGTTTAC
AAACCGGGTGAAAACTTTTTCATTGGCGACCTGCAAAACCGCATCAGCCTGAACTTCGATGTGGAA
GAAGAAAAAATCCGTAGCATCAAAGATTTCTCTATTAAACTGGACGAACTGATCGATGAACTGGCA
AACTACGGTGACAAAATTGAAAAAGTCGTGTTTGTTGCTGATAAAAATCTGCGTCTGAGTTATAAA
GGCCGTGAAATCAACCTGAAAGACATCAAAGACATCCGCTTCGAAGAAAAAACGTTCTAA

Sequence *cas8b*

ATGCTGGCTGGTGTGCTGCAACTGGGTCAATACGCTCTGAATAAAAAATCCACGGATACCGAAGAA
TACCTGCAAGTCATCGAAAACCCGAACGATAAAGGCAACTACAACCATGTTCTGAAAATTGCGTTC
GAAAGTACCGAAGGCAATATCGTGTATCGTGGTGTTGAATACGAAGAATTTAGCAGCAAGAAAATT
AACAACTACGCGTACAAAAAAGGTTCCGCCCGCGGCGGTGATGTGACCCCGACGAGTAAATATACG
GACTCCAT

```
GAAAACCCTGAACAAAATCATGATCAGTTTCAACGATATTCTGAAATCCGCGAATCAGAACAATGA
CGAACAAAAAATCTTCAAAGGCATCTACGAATACATGGTGAGCAACCAGGAAAACATCGCCAACGA
TATTACGGAAAAAATCAAAAGTATCTCCCTGAAAAAAGGCGAATCGTTCATTATCACCGTGACGCT
GTTCGATAACAACACCGAAAAATACATGGGTAACTTCCAGCTGATCCGTAACCATCTGGCCCGCAT
CCTGAACGAACAATACTACAACAAATACGGCAAAACGAGCAAAGGCAAAGGTATCTGCTATTACTG
TAAAAACGAAGGCGAAGTCTTTGGTTTCGTGAATACGTATAACTCTTACACCGTGGATAAAATTGG
TTTCGTTACCGGCGGTTTTAAACAGGAAAATGCGTGGAAAAACTATCCGGTTTGCTCATGCTGTGC
CCAGAAACTGGAACAAGGCAAAAAATATATCCGTGAAAATCTGACGAGCAAATTCTCTGGTTTCGA
TTACTTCGTCATCCCGAAAGCAGTGATCAGCGATGAACACGACGAAGCTGAATTTATTGAAACCCT
GGAAGAATTTGAGAAAAACACCAACTTCTCAACGCAGGAATCGACCAAACAAAACCTGCTGGGCAG
CGAAAAAGATTTTCTGGAAATCATGAAAGACTCTAAAAACTACCTGAACTACAACATGCTGGTCTT
CAAAGAAGAACAGTCAGGTTCGGTGTTCCGTATTCTGCTGTACATCGAAGATATTGTTCCGAGCCG
CGTCAAAAACATCCTGCGTGTGAAAGATCGCGTTGACGAAACGGTCCTGTTTAAAAATCTGCCGGG
CAAAGATAACGCGACCTACGACCTGAAATTCGGTTTCGATAAAATCCGTACGTTTTTCCCGAACAA
TAAAACCGAAGGCAATTTCGACAAAAGTTTTCTGGAAATCCTGAACAACGTTTTCACCTACAAGAA
AATTAGCTACAAATTCCTGCTGGGTCGCATGATTAGCAAAATCCGTTCTGATTTCGCACGCGAAGA
ATATGTGAAAAACCTGGTTCTGCAGGCTCTGATGTGTATCATGTTCATCGACAAACTGAATCTGCT
GTCTGGCAAGGGTAAAGAAGTGCAAAAAATCATGATCGAAAAAACCGAGAAAAACAAAAAATATCT
GGATTTCTTTGAAAACGAATCATACAAAGATGTTTTCAACTCGGACTACAAACGTGCAGTTTTTCT
GACGGGCGTCCTGACCGAAAAACTGCTGAACATTCAGTATAAAGAACGTGGCTCAAAACCGTTTTA
CTCGCGCCTGAATGGTCTGAAACTGAACAAAAACATCGTGAAACGCATCTACACCGAAGCTATCAA
CAAACTGAACGAATACAACAAAAATTATTACAAAGAACTGGAATATCTGATCGGCATGTACATGCT
GAGTGAAGAATCCCAGAAAAACGTTAGCGATGACGAAATTTCTTTTTATTTCGTCCTGGGTATGAG
CCTGGCACGTTTCTTTAATGAAGAAAAGAAAGACGGTGAAGATGAAGAATAA
```

A2: Cross-linked peptides from Cas7, Cas8b and Cas5. The cross-linked lysine residue of each peptide is indicated in parenthesis. The lysine residues oft the peptides from Cas5 and Cas7 were marked in the aa sequence of the respective phyre^2 models.

Peptides Cas7	Peptides Cas8b	Spectra	Score
EIESEKGGIK(6)	KGSAR(1)	3	2,450996738
EIESEKGGIKDGK(6)	KGSAR(1)	1	2,007446482
EIESEKGGIKDGK(6)	YGKTSK(3)	1	2,454692884
EIESEKGGIKDGK(6)	FGFDKIR(5)	1	4,692503962
EIESEKGGIK(6)	KISYK(1)	1	4,30980392
EIESEKGGIKDGK(6)	KISYK(1)	2	6,815308569
EIESEKGGIKDGK(6)	LLNIQYKER(7)	1	4,166215625
AKDFNENVDEILQK(2)	KINNYAYK(1)	1	3,528708289
AKDFNENVDEILQK(2)	KGSAR(1)	1	4,690369833
AKDFNENVDEILQK(2)	VDETVLFKNLPGK(8)	1	3,064492734
SLNKVNLK(4)	KISYK(1)	27	11,27164622
SLNKVNLK(4)	GKEVQK(2)	7	9,139063379
LNGLKLNK(5)	VNLKHIK(4)	1	4,361510743
HIKGTGAFASGEGK(3)	KINNYAYK(1)	1	8,759450752

HIKGTGAFASGEGK(3)	KGSAR(1)	2	3,542118103
HIKGTGAFASGEGK(3)	YTDSMKTLNK(6)	1	2,966576245
HIKGTGAFASGEGK(3)	GSKPFYSR(3)	9	10,98296666
HIKGTGAFASGEGK(3)	LNGLKLNK(5)	26	11,4424928
HIKGTGAFASGEGK(3)	LNKNIVK(3)	1	4,203425667
HIKGTGAFASGEGK(3)	LNEYNKNYYK(6)	52	16,66554625
GTGAFASGEGKAQK(11)	GVEYEEFSSKK(10)	1	2,785156152
GTGAFASGEGKAQK(11)	KINNYAYK(1)	1	3,616184634
GTGAFASGEGKAQK(11)	KGSAR(1)	5	2,759450752
GTGAFASGEGKAQK(11)	YTDSMKTLNK(6)	2	4,643974143
GTGAFASGEGKAQK(11)	TSKGK(3)	1	3,077274542
GTGAFASGEGKAQK(11)	KYIR(1)	1	10,31247104
GTGAFASGEGKAQK(11)	VKNILR(2)	1	6,879426069
GTGAFASGEGKAQK(11)	KISYK(1)	1	7,133712661
GTGAFASGEGKAQK(11)	TEKNKK(3)	4	7,596879479
GTGAFASGEGKAQK(11)	TEKNKK(5)	2	3,468521083
GTGAFASGEGKAQK(11)	LLNIQYKER(7)	22	19,56224944
GTGAFASGEGKAQK(11)	GSKPFYSR(3)	10	10,37675071
GTGAFASGEGKAQK(11)	LNGLKLNK(5)	1	10,76700389
GTGAFASGEGKAQK(11)	NIVKR(4)	1	3,844663963
AQKTFR(3)	KGSAR(1)	4	5,514278574
AQKTFR(3)	LEQGKK(5)	2	6,634512015
AQKTFR(3)	VKNILR(2)	1	4,754487332
AQKTFR(3)	KISYK(1)	1	2,151195299
AQKTFR(3)	TEKNKK(3)	1	4,270025714
GVEYEEFSSKKINNYAYK(11)	VVYKPGENFFIGDLQNR(4)	2	10,84771166
SLNFDVEEEKIR(11)	LNKNIVK(3)	4	14,59859946
SLNFDVEEEKIR(11)	NIVKR(4)	4	4,761953897
SIKDFSIK(3)	KGSAR(1)	2	4,011441043
KINNYAYK(1)	LSYKGR(4)	1	3,074172425
YTEAINKLNEYNK(8)	LSYKGR(4)	1	2,431798276
EINLKDIK(5)	KGSAR(1)	2	2,387216143
DIKDIR(3)	KGSAR(1)	1	2,248720896
DIKDIR(3)	LEQGKK(5)	1	5,272458743

Peptides Cas7	Peptides Cas5	Spectra	Score
GFDGSNGKDIFVR(8)	ISENLINTKK(9)	5	13,00217692
GFDGSNGKDIFVR(8)	KSMNIIHER(1)	2	11,14387556
EIESEKGGIKDGK(6)	ISENLINTKK(9)	37	20,54060751
EIESEKGGIKDGK(6)	KSMNIIHER(1)	30	15,44129143
KSMNIIHER(1)	EIESEKGGIK(6)	25	12,75448733
EIESEKGGIKDGK(10)	ISENLINTKK(9)	2	4,540607512
KSMNIIHER(1)	GGIKDGK(4)	12	5,612610174

TQIKIEFLK(4)	GGIKDGK(4)	2	4,031517051
KSMNIIHER(1)	DGKAR(3)	19	2,707743929
AKDFNENVDEILQK(2)	ISENLINTKK(9)	10	12,76700389
AKDFNENVDEILQK(2)	KSMNIIHER(1)	10	13,05601112
SLNKVNLK(4)	NPVKK(4)	12	10,11804503
ISENLINTKK(9)	SLNKVNLK(4)	2	8,866461092
IEFLKDACYR(5)	SLNKVNLK(4)	6	12,60554832
VNLKHIK(4)	NPVKK(4)	7	5,35261703
ISENLINTKK(9)	VNLKHIK(4)	6	19,0428718
TQIKIEFLK(4)	VNLKHIK(4)	12	15,14448084
KGDIEFEDDREYFTETIPVEMDAER(1)	VNLKHIK(4)	1	3,354577731
HIKGTGAFASGEGK(3)	NPVKK(4)	5	6,119186408
HIKGTGAFASGEGK(3)	ISENLINTKK(9)	12	12,99139983
HIKGTGAFASGEGK(3)	KSMNIIHER(1)	12	15,326058
HIKGTGAFASGEGK(3)	TQIKIEFLK(4)	20	13,1580152
HIKGTGAFASGEGK(3)	IEFLKDACYR(5)	1	4,183096161
KGDIEFEDDREYFTETIPVEMDAER(1)	HIKGTGAFASGEGK(3)	1	3,420216403
GTGAFASGEGKAQK(11)	NPVKK(4)	4	5,552841969
GTGAFASGEGKAQK(11)	ISENLINTKK(9)	6	15,95467702
GTGAFASGEGKAQK(11)	KSMNIIHER(1)	15	9,987162775
GTGAFASGEGKAQK(11)	TQIKIEFLK(4)	14	12,22694531
ISENLINTKK(9)	AQKTFR(3)	9	9,057000407
GTGAFASGEGKAQKTFR(14)	KSMNIIHERTQIK(1)	1	2,913640169
KSMNIIHER(1)	AQKTFR(3)	14	7,26760624
KSMNIIHERTQIK(1)	AQKTFR(3)	1	3,244125144
TQIKIEFLK(4)	AQKTFR(3)	9	6,374687549
ISENLINTKK(9)	SKMGHMPR(2)	7	12,68824614
KSMNIIHER(1)	SKMGHMPR(2)	5	11,21112488
LDELIDELANYGDKIEK(14)	ESLKEHR(4)	4	8,279014256
VVFVADKNLR(7)	KSMNIIHER(1)	1	2,567030709

Peptides Cas5	**Peptides Cas8b**	**Spectra**	**Score**
KISYK(1)	NPVKK(4)	4	7,329754147
ISENLINTKK(9)	YGKTSK(3)	9	15,07262964
QENAWKNYPVCSCCAQK(6)	KSMNIIHER(1)	10	19,52287875
TFFPNNKTEGNFDK(7)	KSMNIIHER(1)	3	9,649751982
QENAWKNYPVCSCCAQK(6)	TQIKIEFLK(4)	11	11,06854213

Danksagung

Ganz besonders bedanke ich mich bei Dr. Lennart Randau für die Möglichkeit meine Doktorarbeit in seinem Labor durchführen zu dürfen, die exzellente Betreuung und stete Gesprächsbereitschaft.

Herrn Prof. Dr. Erhard Bremer, Frau Prof. Dr. Eva Stukenbrock und Herrn Prof. Dr. Torsten Waldminghaus danke ich für ihre Anregungen und Ratschläge als „IMPRS Thesis Advisory Committee“.

Bei Herrn Prof. Dr. Torsten Waldminghaus bedanke ich mich für die Übernahme des Zweitgutachtens. Prof. Dr. Stefan Bauer und Herrn Prof. Dr. Martin Thanbichler danke ich für die spontane Zusage meiner Prüfungskommission anzugehören.

Desweiteren bedanke ich mich bei allen Mitgliedern der AG Randau sowie der Abteilung B2 für die wunderbare Arbeitsatmosphäre und die schöne Zeit. Ein ganz besonderer Dank gilt Dr. André Plagens für seine Diskussionsbereitschaft und guten Ratschläge.

Ich bedanke mich bei Laura Penkert, Franka Schreiner und Kristina Rau, die im Rahmen ihrer Master- bzw. Bachelorarbeit viel zu dieser Arbeit beigetragen haben.

Ganz besonders herzlich danke ich meiner Familie und meinem Freund Andre für die große Unterstützung.

Abgrenzung der Eigenleistung

Die in dieser Arbeit präsentierten Ergebnisse wurden von mir selbständig ohne andere als die hier aufgeführte Hilfe durchgeführt. Im Folgenden werden weitere an dieser Arbeit beteiligten Personen sowie deren experimentellen Beiträge genannt:

Laura Penkert

Hat im Rahmen ihrer Bachelorarbeit die Konstrukte pET20b+*cas3*_co und pET20b+*cas8b*_co erstellt, die zur Anfertigung von Figure 2.11 A verwendet wurden. Figure 2.11 C und Figure 2.18 A stammen aus ihrer Bachelorarbeit.

Franka Schreiner

Hat im Rahmen ihres Mastermoduls die Konstrukte pEC-A-HI-SUMO+*cas5* und pEC-A-HI-SUMO+*cas7* erstellt, die zur Anfertiugung von Figure 2.11 A verwendet wurden.

Kristina Rau

Hat im Rahmen ihrer Masterarbeit das Protein Cas8b zur massenspektrometrischen Analyse von Jörg Kahnt (MPI Marburg) hergestellt. Figure 2.18 stammt aus ihrer Masterarbeit.

Kundan Sharma (MPI Göttingen)

Hat im Rahmen der DFG Forschergruppe FOR1680 als Kollaborationspartner u.a. die massenspektrometrischen Analysen der Protein:Protein und Protein:RNA Cross-link Experimente durchgeführt. Von ihm stammen Figure 2.20 und 2.21 A.