Characterization of the CRISPR-Cas subtype I-B proteins Cas6b and Cas8b of *Methanococcus maripaludis* C5

Dissertation

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

dem Fachbereich Biologie
der Philipps-Universität Marburg
vorgelegt von

Hagen Klaus Gunther Richter

aus Salzwedel

Marburg/Lahn
Dezember 2013

Characterization of the CRISPR-Cas subtype I-B proteins Cas6b and Cas8b of *Methanococcus maripaludis* C5

Dissertation

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

dem Fachbereich Biologie
der Philipps-Universität Marburg
vorgelegt von

Hagen Klaus Gunther Richter

aus Salzwedel

Marburg/Lahn
Dezember 2013

Die Untersuchungen zur vorliegenden Arbeit wurden von Dezember 2010 bis Dezember 2013 am Max-Planck-Institut für Terrestrische Mikrobiologie unter der Leitung von Herrn Dr. Lennart Randau durchgeführt.

Vom Fachbereich
der Philipps-Universität Marburg als Dissertation
angenommen am: 06.12.2013

Erstgutachter: Dr. Lennart Randau
Zweitgutachter: Prof. Dr. Kai Thormann

Tag der mündlichen Prüfung: 19.12.2013

Teile dieser Arbeit sind in folgenden Artikeln veröffentlicht:

Richter H, Lange SJ, Backofen R, Randau L (2013) Comparative analysis of Cas6b processing and CRISPR RNA stability. *RNA Biology* **10:** 700-707

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

Weitere Veröffentlichungen:

Richter H, Mohr S, Randau L (2013b) C/D box sRNA, CRISPR RNA and tRNA processing in an archaeon with a minimal fragmented genome. *Biochemical Society Transactions* **41:** 411-415

Richter H, Randau L, Plagens A (2013c) Exploiting CRISPR/Cas: interference mechanisms and applications. *International Journal of Molecular Sciences* **14:** 14518-14531

Zoephel J, Dwarakanath S, **Richter H**, Plagens A, Randau L (2012) Substrate Generation for Endonucleases of CRISPR/Cas Systems. *Journal of Visualized Experiments : JoVE*

Da steh ich nun, ich armer Tor,
und bin so klug als wie zuvor.

(Johann Wolfgang von Goethe)

List of contents

Chapter 0

Summary

Zusammenfassung

Mit dem CRISPR-Cas System wurde ein adaptives Immunsystem identifiziert, mit dem sich sowohl Archaen also auch Bakterien gegen fremde Nukleinsäuren zur Wehr setzen können. Dabei wird die Komplementarität einer kleinen RNA (crRNA) zur eindringenden Nukleinsäure ausgenutzt um diese abzubauen. Namensgebend für dieses System ist der CRISPR-Array oder Lokus, welcher sich aus sich wiederholenden DNA-Sequenzen (Repeats) und einzigartigen Elementen zwischen den Repeats (Spacer) zusammensetzt. Diese Spacersequenzen können aus vorausgegangenen Infektionen von Viren stammen und vermitteln, als Teil der kleinen crRNA, die nötige Komplementarität im Falle einer Neuinfektion.

Im Zuge der Co-Evolution von Prokaryoten und ihren Viren hat sich eine hohe Diversität von verschiedenen CRISPR-Cas Systemen entwickelt. Es wird zwischen drei Haupttypen unterschieden, welche in weitere Subtypen unterteilt werden können.

Die vorliegende Arbeit zeigt die erste Charakterisierung eines Subtyp I-B CRISPR-Cas Systems. Die *in vivo* Aktivität dieses CRISPR-Cas Systems wurde mittels RNA Sequenzierung in *Methanococcus maripaludis* C5 bestätigt. Die in den RNA-Seq Daten identifizierten crRNAs bestehen jeweils aus einer kompletten Spacersequenz und einer konservierten 8 Nukleotid Sequenz am 5' Terminus sowie einer 2 Nukleotid Sequenz am 3' Terminus.

Es wurden 8 *cas* Gene identifiziert. Von den Cas-Proteinen wurden mit Cas8b und einem zunächst nur mit hypothetischer Funktion annotierten Protein zwei subtypisch spezifische Vertreter genauer analysiert. Für das letzte Protein wurde eine Endoribonuclease-Aktivität identifiziert und das Protein als Cas6b bezeichnet. Cas6b ist ein “single-turnover” Enzym, welches die längeren Vorläufer crRNAs in kleine reife crRNAs prozessiert. Dabei wurde bewiesen, dass Cas6b für die 8 Nukleotid Repeatsequenz am 5' Terminus der reifen crRNAs verantwortlich ist. Trotz einer sehr geringen Sequenzidentität von nur 11%, zeigte eine Modellierung der Cas6b Struktur eine hohe Ähnlichkeit zur Kristallstruktur des in *Pyrococcus furiosus* identifizierten Cas6 Enzyms. Mit Hilfe von eingefügten Punktmutationen konnten vier Aminosäurereste des katalytischen Zentrums von Cas6b identifiziert werden: Lysin (K30), Histidin (H38), Histidin (H40) und Tyrosin (Y47). Analysen der RNA-Bindefähigkeit von Cas6b haben gezeigt, dass Cas6b nach erfolgter Substratbindung in der Lage ist Dimerstrukturen zu formen. Weitere Untersuchungen mittels „RNA-Crosslinking“ gefolgt von massenspektrometrischen Analysen haben ein Methionrest (M185) identifiziert, welcher eng mit einem Uridin (U15) der Repeatsequenz koordiniert ist. Cas6b Aktivitäts-Assays mit Repeatvarianten von *M. maripaludis* sowie der Repeatsequenz von *Clostridium*

thermocellum konnten belegen, dass die Prozessierung von Vorläufer crRNA durch Cas6b unabhängig von möglichen Sekundärstrukturen in der RNA stattfindet.

Neben dem Vorhandensein von reifen crRNAs geht aus den RNA-Seq Daten auch eine hoche Variabilität der crRNA Häufigtkeit hervor. Um die identifizierten crRNA Mengen zu bewerten, wurde ein experimentaler Ablauf entworfen, mit welchem der Einfluss jeder einzelnen Spacersequenz auf a) die Prozessierung durch Cas6b und b) die Stabilität der crRNAs analysiert werden konnte. Mit Hilfe dieses globalen Analyseansatzes wurden geringe Einflüsse der Spacerlänge sowie Spacersequenz auf die *in vitro* Prozessierung und Stabilität beobachtet. In diesem Zusammenhang werden zukünftige Experimente auch den Einfluss des Cas-Protein Interferenzkomplexes (Cascade) sowie mögliche regulatorische Effekte auf die Häufigkeit der crRNAs betrachten.

Die Charakterisierung des subtypspezifischen Cas8b ergab eine Spaltung des rekombinaten Proteins in zwei definierte Fragmente. Mittels Edman-Sequenzierung wurde die Schnittstelle innerhalb der Proteinsequenz identifiziert. In anderen CRISPR-Cas Subtypen sind zwei Proteine vorhanden, welche als große und kleine Untereinheit des Interferenzkomplexes Cascade beschrieben werden. Im Subtyp I-B hingegen ist Cas8b als einziges äquivalent zu den beiden Proteinen zu finden und es wurde postuliert das seine mögliche autokatalytische Spaltung die kleine und große Cascade-Untereinheit generiert. Die biochemische Charakterisierung von Cas8b und dessen mögliche Rolle innerhalb der Interferenzantwort ergab eine unspezifische Bindefähigkeit zu Nukleinsäuren und konnte keine nukleolytische Aktivität zeigen. Mögliche Funktionen von Cas8b werden diskutiert und in zukünftigen Studien sollen diese im Kontext des Interferenzkomplexes Cascade weiterführend untersucht werden.

Summary

The CRISPR-Cas system is an adaptive immune system found in archaea and bacteria to defend themselves against mobile genetic elements (e.g. phages). The system employs base complementarity of small RNA species (crRNAs) to target the foreign nucleic acids for degradation. The hallmark of the system is the CRISPR array or locus, which is composed of repetitive DNA sequences (repeats) that are interspersed by unique sequences (spacers). Spacer sequences can be derived from earlier encounters with viruses and, as part of the crRNAs, confer the base complementarity during a reoccurring attack. During the ongoing battle between prokaryotes and viruses diverse CRISPR-Cas systems evolved into three main types that are further subdivided.

This thesis shows the first characterization of a subtype I-B CRISPR-Cas system. RNA-Seq data proved the *in vivo* activity of this CRISPR-Cas system in *Methanococcus maripaludis* C5. The data further revealed that the crRNAs are always composed of a complete spacer sequence flanked by an 8 nt 5' repeat tag and a 2 nt 3' repeat tag.

Eigth *cas* genes were identified for *M. maripaludis*. Two Cas proteins, Cas8b and an annotated hypothetical protein were characterized in more detail. The hypothetical protein was shown to be the endoribonuclease responsible for the single-turnover catalysis of precursor crRNA into mature crRNA and was termed Cas6b. The reaction performed by Cas6b yields the 8 nt 5' terminal tag of the mature crRNAs. Despite sharing only low sequence identity of 11 %, the two Cas6 proteins of *M. maripaludis* and *Pyrococcus furiosus* could be well aligned using a structural model of Cas6b and the crystal structure of *P. furiosus* Cas6. Cas6b mutant analysis was used to determine four amino acid residues (lysine 30, histidine 38, histidine 40 and tyrosine 47) that comprise the catalytic site of Cas6b. The RNA binding properties of Cas6b were determined and showed a dimerization upon binding to a non-cleavable substrate. Further analyses including RNA crosslinking experiments followed by mass spectrometry identified a methionine residue (M185) that tightly coordinated to a uridine (U15) of the repeat sequence. Cas6b activity assays employing differently structured repeat variants of *M. maripaludis* and a 37 nt repeat sequence of *Clostridium thermocellum* could show, that the processing reaction performed by Cas6b does not recognize a secondary structure of the substrate.

In addition to the verification to the *in vivo* activity of the CRISPR-Cas system, the RNA-Seq data also revealed a varying abundance pattern of crRNAs. To assess the crRNA abundance a experimental procedure was designed, which was aimed to analyse the influence of spacer sequences on a) the processing by Cas6 and b) the stability of crRNAs. With the help of this global approach influences of the spacer length and spacer sequence on the crRNA maturation and *in vitro* stability were recognized. In this context, future experiments will also

determine further possible influences on crRNA abundance including i) crRNA loading into the Cas protein interference complex (Cascade) and ii) possible regulatory effects in terms of crRNA utilization dependent regulation.

The characterization of the subtype-specific protein Cas8b revealed a splitting of the recombinant protein into two defined fragments. The exact point of cleavage was determined by Edman sequencing and provides evidence for a proteolytic cleavage of the full-length protein (either autocatalytically or by a protease). Other CRISPR-Cas subtypes were reported to contain two proteins serving as small and big subunit of the interference complex Cascade. For subtype I-B on the other hand Cas8b was found to be the only equivalent to these two proteins and it was proposed that the identified cleavage generates the large and small Cascade subunit. A biochemical analysis of Cas8b with respect to its putative roles during CRISPR-Cas immunity showed an unspecific binding to nucleic acids while no nucleolytic cleavage was observed. Possible functions of Cas8b are discussed and future studies will focus on the analysis of the protein functions in the context of a complete Cascade.

Chapter I

Introduction

Introduction

In an evolutionary arms race between viruses on one side and Bacteria and Archaea on the other side, the latter ones created a remarkable arsenal of defense mechanisms. In the first line of defense prokaryotes developed various systems to block penetrators using extracellular polymeric substances (e.g. capsules) (Abedon, 2012; Hyman & Abedon, 2010; Labrie et al, 2010). Further defense systems include (i) the alteration or bare absence of receptors for phage adsorption (Labrie et al, 2010), (ii) the blocking of phage movement to prevent the penetration of the cytoplasm in case of successful adsorption or (iii) the recognition of phage particles via prophage endoced proteins during an infection event, which leads to repressed phage proliferation and prevents superinfection of the cell (Raivio, 2011). A different strategy is the abortive infect, in which the community is protected by the death of an individual infected cell to prevent phage proliferation (Hyman & Abedon, 2010; Labrie et al, 2010). Additional specificity in the response to invaders is achieved by restriction-modification systems that use enzymes targeting special sequence motifs to degrade the alien DNA (Arber & Linn, 1969; Enikeeva et al, 2010). Very recently, an adaptive immune system was discovered in Archaea and Bacteria. This CRISPR-Cas (clustered regularly interspaced short palindromic repeats – CRISPR associated) immune system acts via complementary base pairing of small RNA species, CRISPR RNA (crRNA), that target and degrade invading nucleic acids.

Originally thought to be a DNA repair system, the CRISPR-Cas system was later identified to constitute a prokaryotic immune system (Barrangou et al, 2007; Brouns et al, 2008). Hallmark of the system is the CRISPR array or locus, which contains repeated DNA elements (repeats) that are interspersed by unique DNA sequences (spacers) (Fig. 1). These sequences can be derived from an invading nucleic acid (e.g. virus DNA/RNA) and, as part of the small interfering RNA (crRNA), spacers confer the immunity during a reoccurring attack of a particular virus via base complementarity. Spacer sequences can be added or removed from a given CRISPR array and entire CRISPR clusters can be inherited. Hence, spacers define the adaptive character of the immune system. A leader sequence is located upstream of a CRISPR array, in which the regulatory elements and transcriptional start sites for the array are found. The leader also serves as the introduction site of new spacer sequences as newly adapted sequences are always found at the leader proximal end of the array. A set of Cas (CRISPR associated) genes, providing functionality of the system, is situated in close proximity of a CRISPR array, even though sometimes two or more CRISPR loci can share a set of *cas* genes (Barrangou & Horvath, 2011; Bolotin et al, 2005; Brouns et al, 2008; Deveau et al, 2010; Horvath & Barrangou, 2010; Marraffini & Sontheimer, 2010a; Sorek et al, 2008; Terns & Terns, 2011; van der Oost & Brouns, 2009).

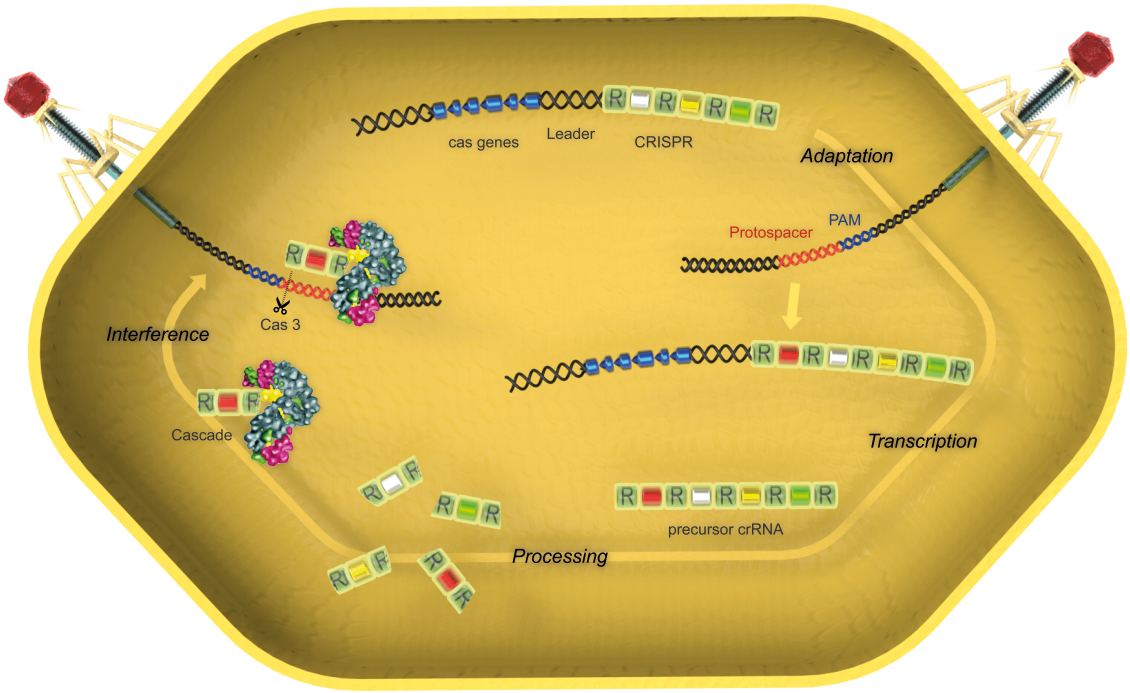


Figure1. Schematic view of the function of Type-I CRISPR-Cas systems. CRISPR-Cas systems comprise i) a **CRISPR** array, composed of repeats (R) and spacer sequences (coloured blocks), ii) a **leader** sequence and iii) a set of ***cas* genes** coding for proteins involved in the functionality of the system.

In clock-wise direction, the injection of phage DNA leads to recognition of the protospacer adjacent motif (PAM) (blue) by Cas proteins inluding Cas1 and Cas2 resulting in the **Adaptation** of a protospacer sequence (red) as a new spacer into the CRISPR array. **Transcription** of the array results in a long precursor crRNA, which is processed by a Cas6 enzyme producing mature crRNA. The crRNA (red) is bound by the CRISPR associated complex for antiviral defense (CASCADE) and subsequently binds the foreign DNA in case of a new infection guided by crRNA specific recognition of the target sequence. The recruitment of Cas3 leads to endonucleolytic decay of the invading nucleic acid.

The functionality of CRISPR-Cas systems can be described in three major stages: i) the adaptation of new spacer sequences by Cascis (CRISPR associated complex for integration of spacer) (Cady & O'Toole, 2011; Datsenko et al, 2012; Erdmann & Garrett, 2012; Plagens et al, 2012; Shah et al, 2013; Swarts et al, 2012; Yosef et al, 2012), ii) the transcription of the array into a precursor crRNA (pre-crRNA) which is further processed into mature crRNA (Carte et al, 2008; Gesner et al, 2011; Haurwitz et al, 2010; Richter et al, 2012; Sashital et al, 2011) and iii) the interference of a CRISPR associated complex for antiviral defense (Casade) or Cascade-like complexes (Jore et al, 2011b; Lintner et al, 2011a; Lintner et al, 2011b; Sashital et al, 2012; Sinkunas et al, 2011; Sinkunas et al, 2013; Wiedenheft et al, 2011a; Wiedenheft et al, 2011b) with the invading nucleic acid based on the base

complementarity of the crRNA (Fig. 1) (Barrangou & Horvath, 2011; Horvath & Barrangou, 2010; Sorek et al, 2008; van der Oost & Brouns, 2009; Westra et al, 2012b).

CRISPR-Cas systems share the same basic principles of functionality. However, while the adaptation process seems to be similar for different systems and two of the involved proteins (Cas1 and Cas2) are uniform, the detailed interference mechanisms with the foreign DNA/RNA show significant differences. Therefore, three major CRISPR-Cas types are described based on the presence of Cas3 (Type I), Cas9 (Type II) and Cas10 (Type III), which serve as the enzymes degrading the alien DNA or RNA species. The three types can be further classified into at least 10 additional subtypes based on subtype-specific proteins (e.g. Cas8b for subtype I-B) (Makarova et al, 2011b).

Shortly after the system was discovered, the idea of exploiting its key players for biotechnical and bioengineering purposes arose (Makarova et al, 2006). As of today, the usage of CRISPR-Cas systems ranges from a simple identification of different bacterial strains based on the spacers of a CRISPR array (spoligotyping) (Kamerbeek et al, 1997) to gene silencing (CRISPRi) (Qi et al, 2013) and genome editing (RGEN) (Cong et al, 2013; Mali et al, 2013) on the basis of the interfering crRNA and the type II protein Cas9. The development of genome-editing tool based on Cas9 and a guide RNA, led to an increased interest in CRISPR-Cas systems. Unlike established genome-editing tools (e.g. transcription activator like effector nucleases (TALEN), zinc-finger nucleases (ZFN)), RGEN does not require a redesign of the used protein and therefore is less money- and time - consuming (recently reviewed in (Charpentier & Doudna, 2013; Horvath & Barrangou, 2013; Richter et al, 2013b)). Future application proposals (Makarova et al, 2006) include the use of engineered CRISPR systems to protect bacteria utilized in industrial processes (e.g. wine and dairy industry).

While most uses of CRISPR-Cas in industrial processes refer to Bacteria, Archaea also possess some interesting aspects regarding their CRISPR-Cas systems. While only about 40% of the sequenced bacteria have a CRISPR-Cas system, roughly 90 % of all sequenced archaea use CRISPR-Cas (Grissa et al, 2007a; Grissa et al, 2007b; Kunin et al, 2007). In some archaea the CRISPR arrays can make up to 1 % of the genome and as many as 18 CRISPR loci have been identified in one species (*Methanocaldococcus sp.* FS406-22). Other archaea accumulated huge amounts of up to 47 *cas* genes (e.g. *Pyrococcus yayanosii*) (numbers taken from CRISPI data base). Given that most prokaryotes try to keep their genome as compact and small as possible it is remarkable that some organisms maintain a large variety of the CRISPR-Cas systems, while others do not possess any CRISPR loci.

Archaea were proposed to constitute a third domain of life in 1977 (Woese & Fox, 1977), next to the already established domains Eukarya and Bacteria, and were accepted as such in the 1990's (Woese et al, 1990).

Members of the Archaea are known to strive under extreme conditions (e.g. high/low pH, high temperature) (Whitman et al, 1999) but were later found to be as ubiquitous as Bacteria (Miller & Wolin, 1982). With the recent advances in the growth and maintenance of archaeal strains and the development of genetic tools for archaeal model organisms (Albers et al, 2006; Berkner et al, 2007; Berkner et al, 2010; Blank et al, 1995; Gardner & Whitman, 1999; Sandbeck & Leigh, 1991) it became possible to survey some fundamental problems of biology: How do the extremophiles abide the harsh conditions? What drives the co-evolution of organisms from the three domains? Especially the methanogenic archaea got into focus as reserves of fossil fuels are decreasing and these organisms could provide a new source of bio-methane (Weiland, 2010). *Methanococcus maripaludis* is one exemplary organism, which produces methane during methanogenesis. It is a coccoidal, methanogenic archeon growing under mesophilic and strictly anaerobic conditions with a hydrogen/carbondioxide atmosphere and a 2 bar over-pressure (Jones et al, 1987; Jones et al, 1983). Isolated from the marshlands in Georgia (US), several further strains (Keswani et al, 1996) were found and sequenced of which some do and some do not have a CRISPR-Cas system. *M. maripaludis* C5, as a representative strain possessing a single subtype I-B system, serves as a perfect model organism to not only characterize the CRISPR-Cas system of this archaeon but also to investigate the evolutionary aspects of strains with and without a CRISPR-Cas system. *M. maripaludis* became an archael model organism due to the ability to grow and maintain this organism in an artificial environment and due to the availability of a fully sequenced genome and genetic tools (Gardner & Whitman, 1999; Sandbeck & Leigh, 1991). The single subtype I-B CRISPR-Cas system (Fig. 2) of *M. maripaludis* C5 is composed of a CRISPR locus with 28 direct repeats of 37 nt length and 27 spacers of 34 – 40 nt length, which do not match to any known virus. Located adjacent to the CRISPR array are two operons of *cas* genes coding for the adaptation proteins Cas1, Cas2 and Cas4, for the interference proteins Cas3, Cas5, Cas7 and the subtype-specific protein Cas8b. In addition, the gene for the pre-crRNA processing enzyme Cas6b is located here (Richter et al, 2012).

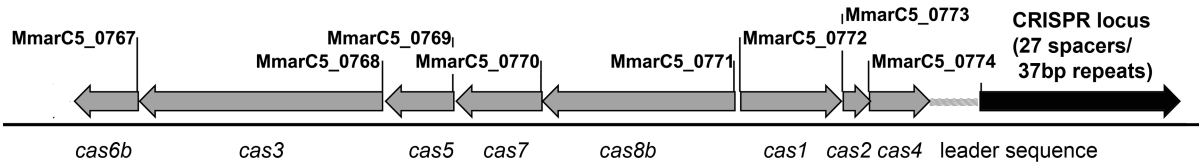


Figure 2. Genetic organization of the subtype I-B CRISPR-Cas system of *M. maripaludis* C5. The system is comprised of i) a CRISPR array (black arrow) with 28 direct repeats of 37 nt length and 27 unique spacers, ii) a leader sequence containing the transcriptional start site (grey shaded block) and iii) a set of 8 *cas* genes coded in two operons (grey arrows) (Richter et al, 2012).

The aim of the work presented in this thesis was to biochemically characterize the previously undescribed CRISPR-Cas subtype I-B based on the system identified in *M. maripaludis*. During the work all 8 *cas* genes were cloned and the proteins expressed. Of these, Cas2, Cas6b, Cas7 and Cas8b turned out to be soluble while the remaining four proteins yielded insoluble inclusion bodies. Therefore, the work focused on the biochemical characterization of the subtype-specific proteins Cas6b and Cas8b. Initial experiments included the verification of *in vivo* transcription and processing and the analysis of mature crRNA abundance and sequence of the CRISPR-Cas system of *M. maripaludis*. The pre-crRNA processing enzyme Cas6b was identified and characterized regarding its RNA binding and cleavage properties, including mutagenesis of important amino acid residues to determine a putative active site and binding domain. It was shown that Cas6b contains four amino acids: K30, the two interchangeable residues H38 and H40 and Y47 that play a major role for the pre-crRNA processing activity. This study also shows a possible dimerization of Cas6b during the processing step and reveals a methionine (M185) residue that is important for repeat recognition. A global *in vitro* analysis of crRNA processing and stability shows possible influences of spacer sequence and length on crRNA abundance. Analysis of the 37 nt repeat sequence indicated that the sequence seems to be more important for processing than the structure of the repeat sequence. The results further indicate an interchangeability of repeat sequences of different I-B subtypes.

The subtype-specific Cas8b protein was investigated considering its putative role within the interfering complex Cascade with focus on RNA and DNA binding and RNA and DNA cleavage properties. Cas8b of a bacterial and an archaeal subtype I-B system proved to co-purify as two fragments, which suggests a possible self-cleavage of the protein. Further studies of Cas8b regarding its possible role in Cascade using R-loop mimicking structures showed that Cas8b does not cleave DNA or RNA and revealed unspecific binding to nucleic acids. Possible functions within the full Cascade assemble will be tested in future studies.

Chapter II

Characterization of CRISPR RNA processing in *Clostridium thermocellum* and *Methanococcus maripaludis*

Hagen Richter[1], Judith Zoephel[1], Jeanette Schermuly[1], Daniel Maticzka[2], Rolf Backofen[2] and Lennart [1]Randau[1,*]

[1] Max-Planck-Institute for Terrestrial Microbiology, Karl-von-Frisch Straße 10, 35043 Marburg
[2] Institut für Informatik, Albert-Ludwigs-Universität, Georges-Koehler-Allee, Geb 106, 79110 Freibug
[*] Corresponding author

Abstract

The CRISPR arrays found in many bacteria and most archaea are transcribed into a long precursor RNA that is processed into small clustered regularly interspaced short palindromic repeats (CRISPR) RNAs (crRNAs). These RNA molecules can contain fragments of viral genomes and mediate, together with a set of CRISPR-associated (Cas) proteins, the prokaryotic immunity against viral attacks. CRISPR/Cas systems are diverse and the Cas6 enzymes that process crRNAs vary between different subtypes. We analysed CRISPR/Cas subtype I-B and present the identification of novel Cas6 enzymes from the bacterial and archaeal model organisms *Clostridium thermocellum* and *Methanococcus maripaludis* C5. *Methanococcus maripaludis* Cas6b *in vitro* activity and specificity was determined. Two complementary catalytic histidine residues were identified. RNA-Seq analyses revealed *in vivo* crRNA processing sites, crRNA abundance and orientation of CRISPR transcription within these two organisms. Individual spacer sequences were identified with strong effects on transcription and processing patterns of a CRISPR cluster. These effects will need to be considered for the application of CRISPR clusters that are designed to produce synthetic crRNAs.

Introduction

Clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR-associated (*cas*) genes define an anti-viral defence system in Archaea and Bacteria. CRISPR loci are composed of repeat sequences with an average length of 24 – 47nt, which alternate with unique spacer sequences derived from previous encounters with foreign nucleic acids (i.e. viruses, plasmids) (Barrangou et al, 2007; Bolotin et al, 2005; Sorek et al, 2008; van der Oost et al, 2009). CRISPR loci are transcribed and processed to generate the small interfering crRNAs. Diverse sets of cas genes are often found adjacent to a CRISPR locus and encode proteins that are involved in the three phases of CRISPR/Cas activity: acquisition of new spacers, processing of crRNAs and interference with foreign nucleic acid (Barrangou & Horvath, 2011; Cui et al, 2008; Horvath & Barrangou, 2010; Koonin & Makarova, 2009; Terns & Terns, 2011). Although there is little information available for the process of new spacer acquisition, recent progress has led to a better understanding of the other two phases. The maturation of precursor crRNA into small crRNAs is performed by diverse Cas endonucleases that belong to a protein family termed Cas6 (Carte et al, 2010; Carte et al, 2008; Gesner et al, 2011; Haurwitz et al, 2010; Sashital et al, 2011; Wang et al, 2011; Wang et al, 2012). In CRISPR/Cas Type-I the interference step is mediated by a complex of different Cas proteins (Cas complex for antiviral defence: Cascade) bound to crRNAs that

target the invading nucleic acid through base complementarity which ultimately results in the inactivation or degradation of foreign DNA by Cas3 (Howard et al, 2011; Jore et al, 2011b; Lintner et al, 2011b; Mulepati & Bailey, 2011; Plagens et al, 2012; Semenova et al, 2011; Sinkunas et al, 2011; Westra et al, 2012c). Type-II CRISPR/Cas systems use the single Cas9 protein for interference (Sapranauskas et al, 2011) and Type-III systems use a multi Cas protein complex that is distinct from Cascade (Cocozaki et al, 2012; Zhang et al, 2012).
Computational analyses of these defence systems identified a surprising diversity of different CRISPR/Cas types and subtypes, which are spread throughout archaeal and bacterial kingdoms.
This classification has defined three major types which can be further divided into at least 10 CRISPR/Cas subtypes (Makarova et al, 2011b). The subtype I-B, found, e.g. in Clostridia, methanogens and halophiles, is defined by the subtype- specific protein Cas8b. In *Clostridium thermocellum* and *Methanococcus maripaludis* the minimal subtype I-B Cas protein organization consists of the universal Cas1, Cas2 and Cas4 proteins that are proposed to mediate the integration of spacers as well as Cas3, Cas5, Cas7 and Cas8b, which are proposed to form the Cascade complex of this subtype. Finally, a Cas6 protein is required for the processing of crRNA (Carte et al, 2010; Carte et al, 2008; Gesner et al, 2011; Haurwitz et al, 2010; Sashital et al, 2011; Wang et al, 2011; Wang et al, 2012).
A Cas6 protein was first described for CRISPR/Cas subtype III-B in *Pyrococcus furiosus* as a metal-independent endonuclease involved in the processing of precursor crRNA into mature crRNA (Carte et al, 2010; Carte et al, 2008; Wang et al, 2011; Wang et al, 2012). Cas6 enzymes were also characterized for CRISPR/Cas subtype I-F in *Pseudomonas aeruginosa* (Cas6f, also termed Csy4) (Haurwitz et al, 2010) and CRISPR/Cas subtype I-E in *Thermus thermophilus* and *Escherichia coli* (Cas6e, also termed Cse3) (Gesner et al, 2011; Sashital et al, 2011). The amino acid sequence similarity of these Cas6 proteins is limited, yet they share ferredoxin-like folds and perform analogous reactions in the different CRISPR/Cas systems. These Cas6 proteins do not only differ in substrate specificity, but also in the composition of their active sites. For example *P. furiosus* Cas6 (Pf Cas6) interacts with single-stranded RNA while Cas6e and Cas6f seem to specifically bind to hairpin structures formed by the repeats (Carte et al, 2010; Carte et al, 2008; Gesner et al, 2011; Haurwitz et al, 2010; Sashital et al, 2011; Wang et al, 2011; Wang et al, 2012). Further differences can be found in the catalytic site of the Cas6 proteins. Pf Cas6 uses a catalytic triad composed of tyrosine, histidine and lysine residues (Carte et al, 2010; Wang et al, 2011), while in Cas6f a catalytic dyad of a histidine and a serine residue proved to be important for protein activity (Haurwitz et al, 2010; Haurwitz et al, 2012). Activity of Cas6e relies on a tyrosine and a histidine residue (Gesner et al, 2011; Sashital et al, 2011). Although there are variations in their active site composition and the recognition of RNA substrates, the different Cas6

cleavage reactions always generate crRNAs that consist of a spacer unit that is flanked by 8 nt of the repeat sequence as a 5'-terminal tag and a 3'-terminal repeat tag (Carte et al, 2008; Gesner et al, 2011; Haurwitz et al, 2010). Finally, Cas6 was shown to deliver the mature crRNA to the Cascade complex (Jore et al, 2011b; Wiedenheft et al, 2011b).

In this study, we provide the first analysis of crRNA processing for CRISPR/Cas subtype I-B for one bacterial model organism, *C. thermocellum* and one archaeal model organism, *M. maripaludis* (detailed information of CRISPR loci and gene organization can be found in Supplementary Figure S1). The abundance and processing of crRNAs were analysed *in vivo* by RNA-Seq methodology. In addition, the Cas6 enzymes of this CRISPR/Cas subtype (termed Cas6b) were identified and *M. maripaludis* Cas6b (Mm Cas6b) was analysed for crRNA processing *in vitro*.

Material and Methods

Growth of *E. coli*, *M. maripaludis* C5 and *C. thermocellum* cells

Methanococcus maripaludis C5 cells were a kind gift of W.B. Whitman (Georgia). *Clostridium thermocellum* (DSM1237) cells were obtained from DSMZ (German collection of micro-organisms and cell cultures). All *E. coli* cells were grown in LB-media with appropriate antibiotics at 37 °C and shaking at 200 rpm.
Methanococcus maripaludis C5 was grown at 37 °C in complex medium for methanococci (McC) (Jones et al, 1987) with H_2/CO_2 atmosphere (80%/20%) and one bar (15 psi) overpressure. *Clostridium thermocellum* cells were incubated in complex medium (Lynd & Grethlein, 1987) at 60 °C with an anaerobic atmosphere (N_2).

Production of Cas6 and mutants

The *cas6* genes MmarC5_0767, Cthe_3205 and Cthe_2303 were amplified from genomic DNA of *M. maripaludis* C5 or *C. thermocellum* ATCC 27405 and cloned into the vector pET-20b to facilitate protein expression with a C-terminal His-tag. Oligonucleotides for site-directed mutagenesis were designed using Agilents QuickChange Primer Design tool and *cas6* mutants were created using the QuickChange site-directed mutagenesis (Stratagene) according to the manufacturer's instructions. Mutations were confirmed by sequencing (MWG Eurofins).
All Cas6 variants were produced in *E. coli* (Rosetta2 DE3) cells. Induction of protein expression was performed by addition of isopropylthio-b-D-galactoside (IPTG) to a final concentration of 0.5 mM after growing the cells to an OD_{578} of 0.6. Four hours after induction the cells were harvested, the pelleted cells re-suspended in lysis buffer (10 mM Tris–HCl [pH8.0], 300 mM NaCl, 10 % glycerol and 0.5 mM DTT) and incubated on ice with lysozyme (1 mg/g cell pellet) for 30 min. Cell disruption was performed using sonication (8 x 30 s; Branson Sonifier 250). Clearing of the lysate was achieved by centrifugation (20000 rpm, 30 min, 4 °C) and the supernatant was applied to a Ni–NTA–Sepharose Column (GE-Healthcare) and purified using a FPLC Äkta- Purification system (GE-Healthcare). Elution of the proteins was performed by a linear imidazole gradient (0 - 500 mM). Purity of the proteins was determined by sodium dodecyl sulphate–polyacrylamide gel electrophoresis (SDS–PAGE) and Coomassie Blue staining. The protein was dialysed into lysis buffer and the protein concentration was determined by Bradford Assay (BioRad).

Generation of RNA substrates

The spacer2–repeat–spacer3 and repeat–spacer27–repeat RNA substrates were generated by *in vitro* run-off transcription using T7 RNA polymerase and internally labelled using [α-^{32}P] adenosine triphosphate (ATP) (5000 ci/mmol, Hartman Analytic) (Sampson & Uhlenbeck, 1988). The repeat RNAs and repeat RNAs with a substitution of the first unprocessed nucleotide against a dexoy nucleotide were synthesized by Eurofins MWG Operon. End labelling of these substrates was performed using T4 polynucleotide kinase (Ambion) and [γ-^{32}P] ATP (5000 ci/mmol) according to the manufacturer's instructions.

Templates for *in vitro* transcription were obtained by cloning of the pre-crRNA sequences with an upstream T7 RNA polymerase promotor sequence into pUC19 vector. After linearization of the plasmid with HindIII, *in vitro* transcription was performed in a final volume of 20 µl [40 mM HEPES–KOH (pH8.0); 22 mM $MgCl_2$; 5 mM DTT; 1 mM spermidine; 4 mM UTP, CTP, GTP and 2 mM ATP; 20 U RNase inhibitor; 1 µg T7 RNA polymerase; 1 µg linearized plasmid] at 37 °C for 1 h. End labelling of synthesized RNA was done in a 20 µl reaction volume: 10 µl of the RNA was labelled using 2 µl T4 Polynucleotide Kinase (PNK) buffer (New England Biolabs (NEB)) and 25 U T4 PNK (Ambion) at 37 °C for 30 min.

The RNAs were separated by denaturing PAGE (8 M urea; 1 x TBE; 10% polyacrylamide), and afterwards respective bands were cut out using sterile scalpels in reference to brief autoradiographic exposure. The RNA was eluted from the gel piece using 500 µl RNA elution buffer [250 mM NaOAc, 20 mM Tris–HCl (pH 7.5), 1 mM ethylenediaminetetraacetic acid (EDTA) (pH8.0), 0.25 % SDS] and overnight incubation on ice. Precipitation of RNA was performed by adding two volumes EtOH (100 %; ice cold) and 1/100 glycogen for 1 h at -20°C and subsequent washing with 70 % EtOH of pelleted RNA.

Endonuclease assay

Different indicated concentrations of purified Cas6 enzyme were incubated with radio labelled RNA substrates and buffer [250 mM KCl, 1.875 mM $MgCl_2$, 1 mM DTT, 20 mM HEPES–KOH (pH 8.0)]. The reaction mix was incubated for 10 min at 37 °C and then immediately mixed with 2x formamide buffer [95 % formamide; 5 mM EDTA (pH 8.0); 2.5 mg bromophenol blue; 2.5 mg xylene cyanol] and incubated at 95 °C for 5 min to stop the cleavage reaction. The reaction was applied to a denaturing 12–15 % polyacrylamide gel running in 1 x TBE with 12 W for 1.5 h. Visualization was achieved by phosphorimaging.

RNA-sequencing

RNA and DNA were extracted from cell lysates with phenol/chloroform (1:1; phenol pH 5 for RNA and pH 8 for DNA) (Randau et al, 2005b). A Proteinase K and 55 °C heat shock treatment preceded the phenol/chloroform step. Small RNAs (<200 nt) were purified from total RNA using the mirVana RNA extraction kit (Ambion). Three micrograms of isolated small RNA from either *M. maripaludis* C5 or *C. thermocellum* were treated with T4 PNK to ensure proper termini for ligation. A protocol for the dephosphorylation of 2'-, 3'-cyclic phosphate termini was modified from (Schurer et al, 2002): 1 µg of RNA was incubated at 37°C for 6 h with 10 U T4 PNK and 10 µl 5x T4PNK buffer (NEB) in a total volume of 50 µl. Subsequently, 1 mM ATP was added and the reaction mixture was incubated for 1 h at 37 °C to generate monophosphorylated 5'-termini. RNA libraries were prepared with an Illumina TruSeq RNA Sample Prep Kit and sequencing on an Illumina HiSeq2000 sequencer was performed at the Max-Planck Genomecentre Cologne.

Identification of crRNA abundance

Sequencing reads were trimmed [(i) removal of Illumina TruSeq linkers and poly-A tails and (ii) removal of sequences using a quality score limit of 0.05] and mapped to the reference genomes (GenBank: CP000568 and CP000609) 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). Reads <15 nt were removed. Initial crRNA identification was obtained from crisprdb (Grissa et al, 2007a) and gene annotations were obtained from Genbank.

Modelling of *M. maripaludis* Cas6b

A model of the Mm Cas6b (MmarC5_0767) protein structure was built with the I-TASSER platform (Roy et al, 2010). The program identified *P. furiosus* Cas6 (pdb ID 3PKM) as the top template for structure prediction. The protein model was compared with the Pf Cas6 crystal structure using the program DaliLite (Holm & Park, 2000) and their alignment revealed two homologous structures (Z-score 19.7, RMSD 2.5 Å). Cas6b sequences were aligned with ClustalW2 (Larkin et al, 2007).

Results

crRNA processing for CRISPR/Cas subtype I-B

The processing of crRNAs of the CRISPR/Cas subtype I-B was analysed by RNA-Seq for *M. maripaludis* C5 and for *C. thermocellum* ATCC 27 405. The isolated total small RNAs were modified with T4 polynucleotide kinase to allow proper adapter ligation and were sequenced through Illumina HiSeq2000 RNA-Seq methodology. Over 14 million individual sequence reads were mapped to the corresponding reference genomes and elucidated the abundance and processing patterns of the CRISPR arrays of these two organisms. *M. maripaludis* C5 possesses a single CRISPR cluster with 28 repeats of 37-nt length that are interspersed by 27 unique spacers. The CRISPR region is constitutively transcribed and processed into small crRNAs (Figure 1). The crRNAs contain a clearly defined 5'-terminal 8-nt tag with the sequence 5'-AUUGAAAC-3'. The 3'-termini are gradually shortened and most often contain a minimal 2-nt tag with the repeat nucleotides 5'-CU-3'. The abundance of crRNAs declines gradually from the leader proximal to the leader distant region with the crRNA containing the highly AT-rich (30 A or T out of 34 nt) spacer 3 being underrepresented.

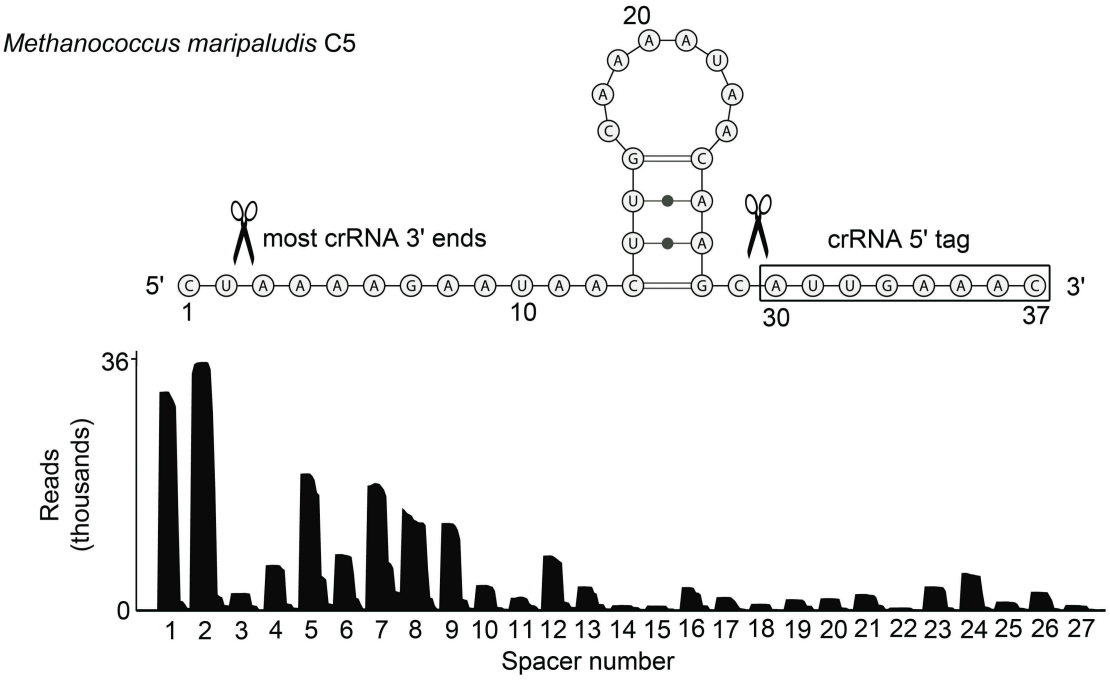
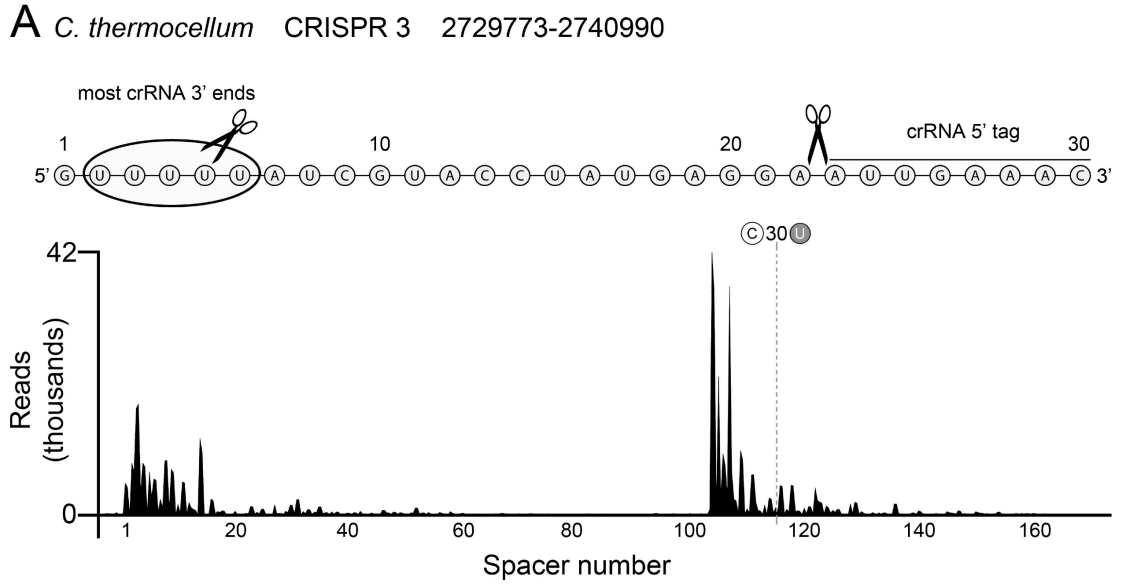


Figure 1. crRNA processing in *M. maripaludis*. Illumina HiSeq2000 sequencing reads mapped to the *M. maripaludis* C5 reference genome highlight the abundance of crRNAs. Processing occurs within the repeat elements, generating crRNAs with a 5'-terminal AUUGAAAC 8-nt tag (boxed) and more variably trimmed 3'-terminal tags. Cleavage sites are indicated and a possible hairpin structure was predicted by RNAfold (Zuker & Stiegler, 1981).

crRNA processing patterns in *C. thermocellum* reveal long-range influence of spacer sequences

RNA-Seq analysis of the small RNAs of *C. thermocellum* revealed five constitutively transcribed and processed CRISPR clusters. Two of these CRISPR/Cas subtype I-B systems are very similar to the one found in *M. maripaludis* and contain 37-nt repeat elements. The other three CRISPR clusters have 30-nt repeat sequences. Processing of both *C. thermocellum* CRISPR repeat sequences into mature crRNAs yields the same 5'-terminal 8-nt (5'- AUUGAAAC-3') tag that is also found for *M. maripaludis* crRNAs (Figure 2 and Supplementary Figure S2). The 3'-termini are trimmed leaving mostly short tags. The abundance of crRNAs follows the pattern found in *M. maripaludis* and described for other CRISPR/Cas subtypes with one notable exception. The CRISPR locus 3 contains an internal signal to promote crRNA transcription within the CRISPR array (Figure 2A and Supplementary Table S1). The overall crRNA abundance gradually declines from Spacer 1 to Spacer 103 before crRNA production peaks again starting with the crRNA containing spacer number 104. Interestingly, the 8-nt repeat tags are not identical for the crRNAs from this CRISPR locus 3 as at Spacer 116 the final U base of the 5'-terminal tag is changed to the commonly found base C (Figure 2A and Supplementary Table S1). Close analysis of this sudden spike of internal crRNA abundance revealed a transcription start site at the A residue at Position 29 within Spacer 103. Our data suggest that this spacer is sufficient to promote transcription within the CRISPR region and that the 28-nt upstream of the transcription start within the spacer provide the necessary promoter elements in the context of the flanking repeats. Although it is difficult to pinpoint the pribnow box, the extreme AT-richness of the spacer (26 out of 28 nt upstream of the transcription start site are A and T residues) suggests relaxed strand separation. DNA sequencing of the genomic region upstream of Spacer 104 excluded errors in the initial genome assembly during whole-genome sequencing. In addition to internal promotion, we observed several cases of bidirectional transcript production for the CRISPR arrays. Anti-crRNA transcripts can start at the region opposite of the leader (CRISPR loci 1,2,5) or internally (CRISPR locus 4) (Figure 2B and Supplementary Figure S2). Although the amount of these anti-crRNA transcripts is usually very small in comparison to the abundance of crRNAs, individual anti-crRNAs show a conserved processing pattern within the repeats that yields RNAs with complete reverse complementary spacer sequences. These anti-crRNAs usually contain 18-nt 5'-tags and 15-nt 3'-tags for CRISPR loci 1 and 2 and 22-nt 5'-tags for CRISPR loci 4 and 5 (Supplementary Figure S3). The presence of processed anti-crRNAs can correlate with the reduced abundance of the respective sense crRNA (Figure 2B and Supplementary Figure S2).


A C. thermocellum CRISPR 3 2729773-2740990
most crRNA 3' ends
crRNA 5' tag
5' G U U U U U A U C G U A C C U A U G A G G A A U U G A A A C 3'
1
10
20
30
C 30 U
42
0
Reads (thousands)
1
20
40
60
80
100
120
140
160
Spacer number

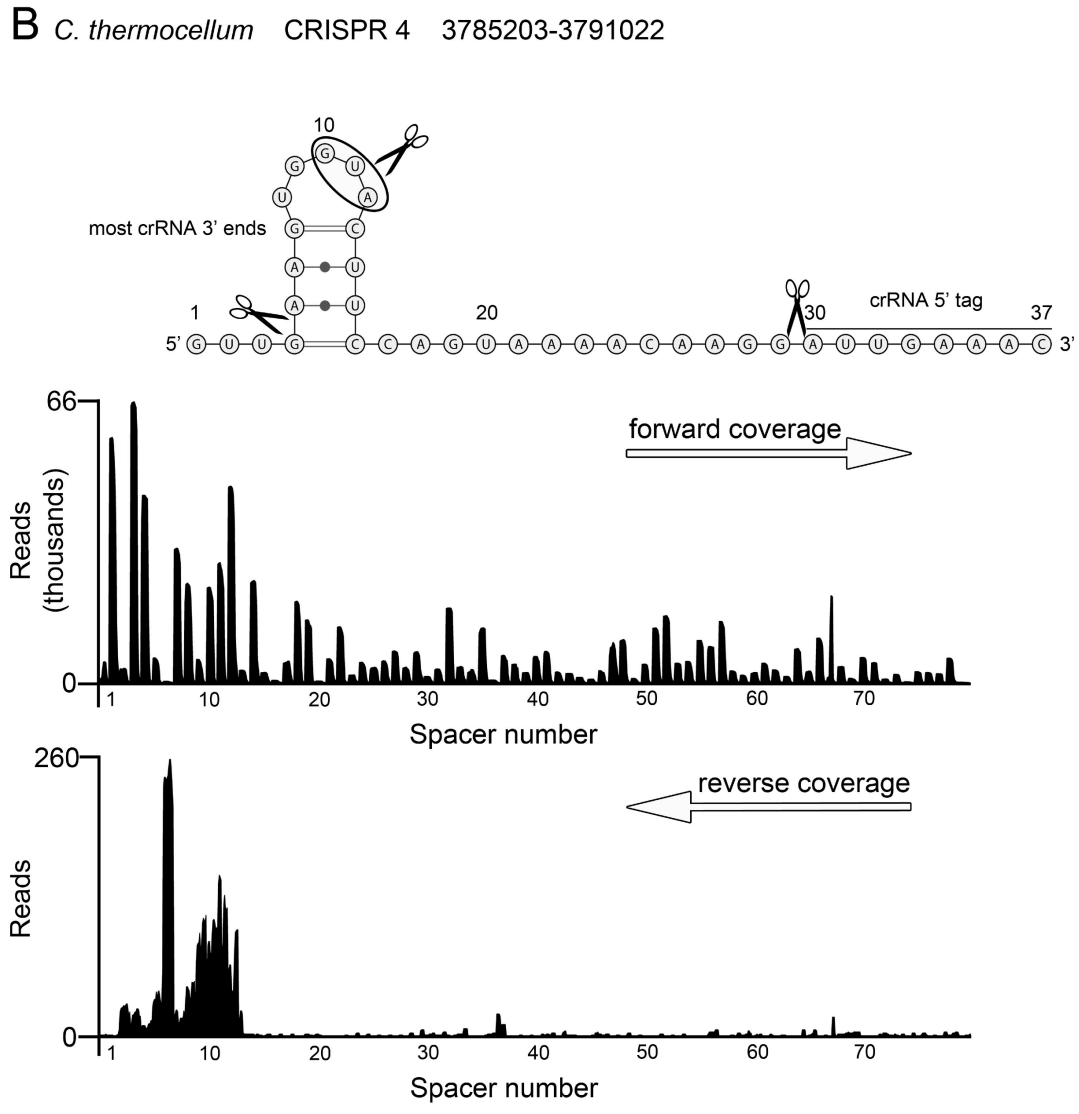

B C. thermocellum CRISPR 4 3785203-3791022
10
most crRNA 3' ends
crRNA 5' tag
5' G U U G C C A G U A A A A C A A G G A U U G A A A C 3'
1
20
30
37
66
0
Reads (thousands)
forward coverage
1
10
20
30
40
50
60
70
Spacer number
260
0
Reads
reverse coverage
1
10
20
30
40
50
60
70
Spacer number

Figure 2. crRNA processing in *C. thermocellum*. Illumina HiSeq2000 sequencing reads were mapped to the *C. thermocellum* ATCC 27405 reference genome and selected CRISPR regions are displayed. Conserved 5'-terminal crRNA cleavage sites and variably trimmed 3'-termini are indicated within the repeat sequence. A possible hairpin structure was predicted by RNAfold (Zuker & Stiegler, 1981). All *C. thermocellum* CRISPR mappings are found in Supplementary Figure S2. (**A**) CRISPR locus 3 reveals internal promotion of crRNA transcription at Spacer 104. (**B**) CRISPR locus 4 exemplifies bidirectional CRISPR transcription. Forward and reverse reads were separated to highlight the occurrence of processed anti-crRNAs that can correlate with reduced crRNA abundance.

Identification of Cas6 I-B enzymes that generate crRNAs

To identify the enzyme that generates crRNAs for CRISPR/Cas subtype I-B, we analysed the *cas* genes of *M. maripaludis* and *C. thermocellum*. A set of only eight *cas* genes was identified in the genome of *M. maripaludis* C5. One of these potential *cas* gene products (MmarC5_0767, Mm Cas6b) showed 12 % amino acid identity to Pf Cas6 which identified it as a Cas6b candidate for CRISPR/Cas subtype I-B. As the protein shares limited sequence identity with Cas6 proteins of other CRISPR/Cas subtypes (Carte et al, 2008; Gesner et al, 2011; Haurwitz et al, 2010), the structure of Mm Cas6b was modelled with I-Tasser (Roy et al, 2010; Zhang, 2008). Pf Cas6 was identified as the closest structural homologue and shares a very similar overall architecture [Dali-Lite Z-score 19.7, RMSD 2.5 Å, (Holm & Park, 2000)] (Figure 3). The structural alignment of the Mm Cas6b model and Pf Cas6 also reveals a conserved histidine residue of Mm Cas6b in close proximity to the catalytic histidine of Pf Cas6. The comparison of different Cas6b homologues of CRISPR subtype I-B (Figure 3) indicates high sequence similarity and conserved residues. *Clostridium thermocellum* contains one Cas6b homologue (Cthe_3205) associated with the 37-nt repeat sequences and a potential second Cas6 enzyme (Cthe_2303) associated with the 30-nt repeat sequences. Classification of Cthe_2303 is not unambiguously possible as the neighbouring *cas* genes do not clearly fit into the commonly used 10 CRISPR/Cas subtypes.

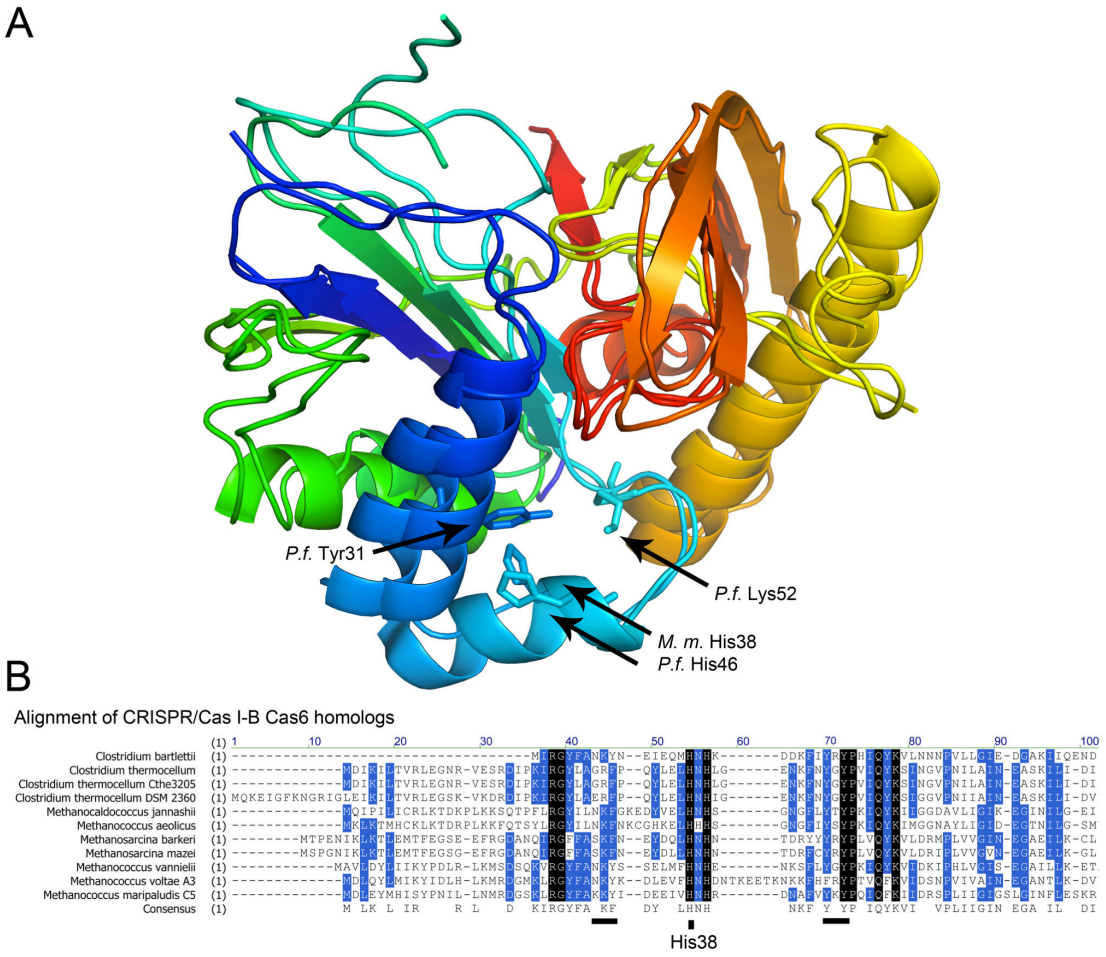


Figure 3. Structural model of Cas6b shows high similarity to Pf Cas6. (**A**) A structure model (I-Tasser) of *M. maripaludis* Cas6b was aligned to Pf Cas6 (DaliLite). The catalytic site of Pf Cas6 is indicated and a histdine at Position 38 in Cas6b is located in close proximity to the catalytic histdine in *P. furiosus*. (**B**) An alignment of Cas6b homologues (ClustalW) of subtype I-B shows conserved amino acid residues (black coloured residues) and reveals putative catalytic residues (black bars).

Repeat specific endonucleolytic activity of subtype I-B Cas6 homologues

The genes for Cas6 enzymes from *M. mariplaudis* and *C. thermocellum* were cloned and the recombinant enzymes were produced. *Clostridium thermocellum* Cas6b (Cthe_3205) production yielded only insoluble protein, but the two other Cas6 proteins were obtained in soluble form and allowed the analysis of their involvement in crRNA processing. Size exclusion separation of Mm Cas6b revealed a monomeric structure of the protein. The purified Cas6 candidates (Mm Cas6b and Cthe_2303) showed *in vitro* endonuclease activity and processed repeat RNA sequences from *M. maripaludis* and *C. thermocellum*, respectively (Figure 4). Therefore, the unclassified Cthe_2303 provided a good control for crRNA processing specificity of Mm Cas6b, as both enzymes specifically recognized only the repeats from their associated CRISPR clusters. Addition of bivalent metal ions did not

influence the activity of Mm Cas6b. To define the cleavage site, a repeat RNA was synthesized with a deoxy nucleotide substitution at the proposed processing position that generates the 8-nt crRNA 5'-tag observed *in vivo*. This substitution resulted in the loss of Mm Cas6b and Cthe_2303 processing.

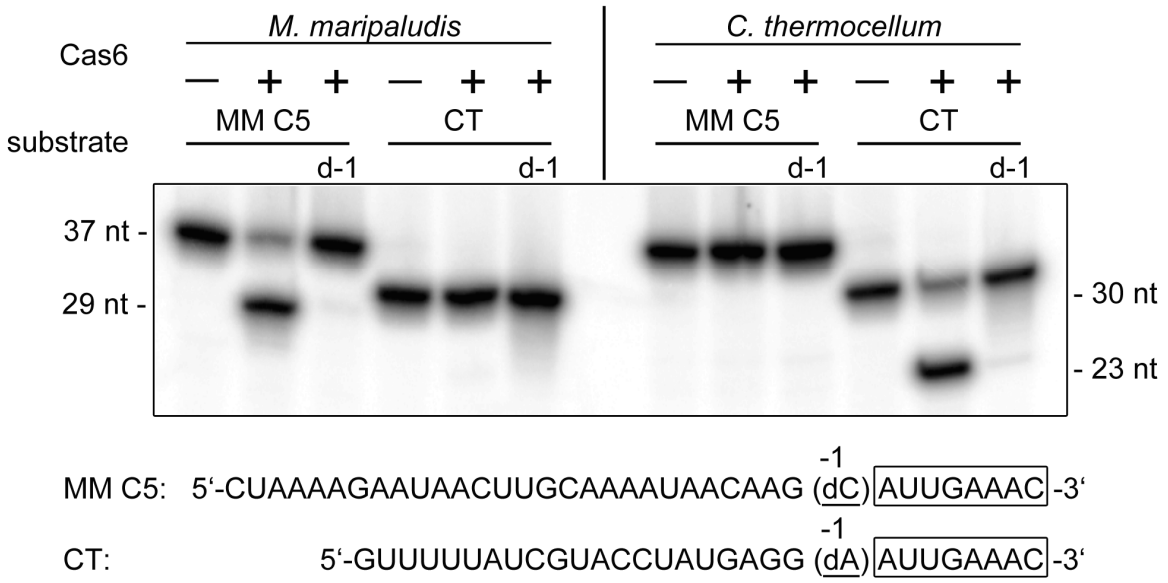


Figure 4. Cas6b of *M. maripaludis* (MM C5) and *C. thermocellum* Cthe_2303 (CT) cleave their specific repeat structure. Cas6b endonuclease assay were performed with 5'-terminal radioactively labelled repeat RNA and the respective deoxy variants (indicated at the bottom, -1 displaying the first base upstream of the 5'-tag) of *M. maripaludis* and *C. thermocellum*. Cas6b processes the 37-nt repeat into the smaller 29-nt fragment, while the deoxy variant (d-1) and the 30-nt repeat RNA of *C. thermocellum* are not cleaved. Cthe_2303 is specific for its 30-nt repeat RNA.

A single repeat structure is sufficient for Mm Cas6b *in vitro* processing

To test the influence of different RNA substrates (Supplementary Table S2) for Mm Cas6b activity, cleavage assays were performed with (i) repeat RNA, (ii) repeat - spacer - repeat RNA and (iii) spacer - repeat - spacer RNA using repeat and spacer sequences of the *M. maripaludis* CRISPR. For all three substrates, product formation was observed that corresponds with Mm Cas6b processing at the cleavage site determined by the deoxy nucleotide substitution within the repeat (Figure 4). This cleavage site determines that the conversion of the repeat (37 nt) results in a 29-nt fragment, while the repeat - spacer27 - repeat structure (131 nt) is processed into three fragments (74, 38 and 19 nt) and the spacer2 - repeat - spacer3 substrate (110 nt) is cleaved into two fragments (67 and 43nt) (Figure 5). Since all used substrates were cleaved in similar efficiency, the repeat RNA was used for further analysis of the catalytic site of Mm Cas6b. In order to test the

influence of the computationally predicted [RNAfold (Zuker & Stiegler, 1981)] short hairpin structure of the repeat (Figure 1), the mutation G16C was introduced that disrupts a G–C base pair within this hairpin. The mutated repeat was cleaved less effectively than wild-type repeats (Supplementary Figure S4).

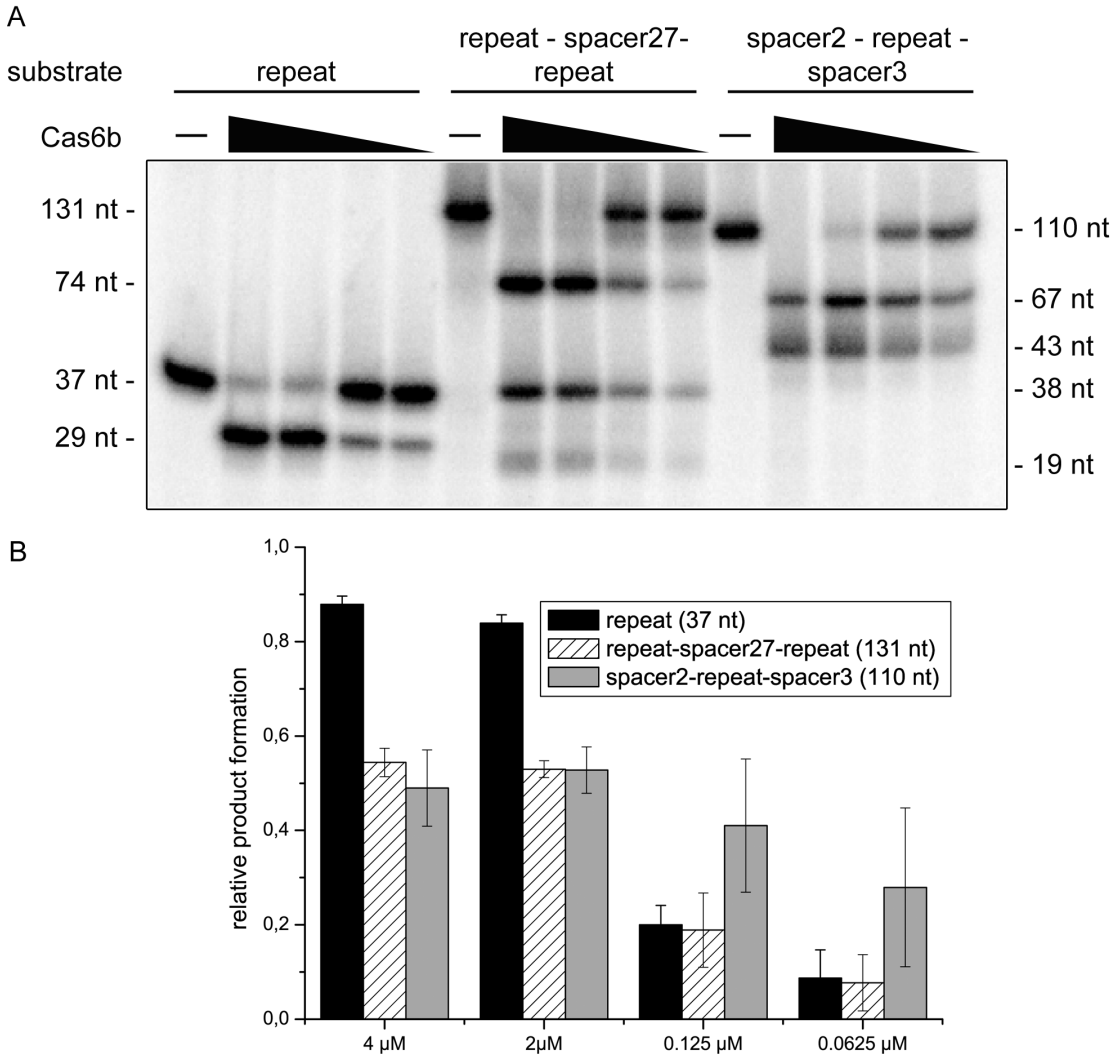


Figure 5. RNA substrates for Cas6b processing. The two 'repeat - spacer27 - repeat' and 'spacer2 - repeat - spacer3' substrates were internally labelled by *in vitro* transcription and repeat RNA molecules were 5'-end labelled. The substrates were used in three independent cleavage assays using different concentrations of Cas6b (4, 2, 0.125 and 0.0625 µM). (**A**) A representative assay shows the ability of Cas6b to process all used substrates in similar manner and efficiency. (**B**) Product formation for three independent reactions was quantified.

Mm Cas6b contains two catalytic histidine residues

To deduce catalytic residues, potentially important amino acids were identified based on the structural model of Mm Cas6b and the observed conservation of amino acids in the

alignment of Cas6b homologues (Figure 3). Cas6 proteins of other CRISPR/Cas subtypes contain a single catalytic histidine residue. In Cas6b, there are two conserved histidine residues (H38 and H40), separated only by a single amino acid that could potentially fulfil this role. A set of Mm Cas6b mutants was produced (Supplementary Figure S5) of which two mutants (Y49A and Y47A/Y49A) yielded insoluble proteins. The other mutants were used in endonucleolytic cleavage assays testing the processing of the repeat RNA substrate in comparison to wild-type Mm Cas6b (Figure 6). The single histidine mutant H38A and the tyrosine mutant Y47A showed reduced processing activity compared with wild-type Mm Cas6b. The H40A mutation reduced Mm Cas6b activity by >50 %. Surprisingly, both single histidine mutants retained considerable cleavage activity. However, the mutation of both histidine residues into alanine (H38A/H40A) resulted in a drastic loss of substrate processing. Mutation of the lysines at Position 29 or 30 did not show any notable effect on endonucleolytic activity.

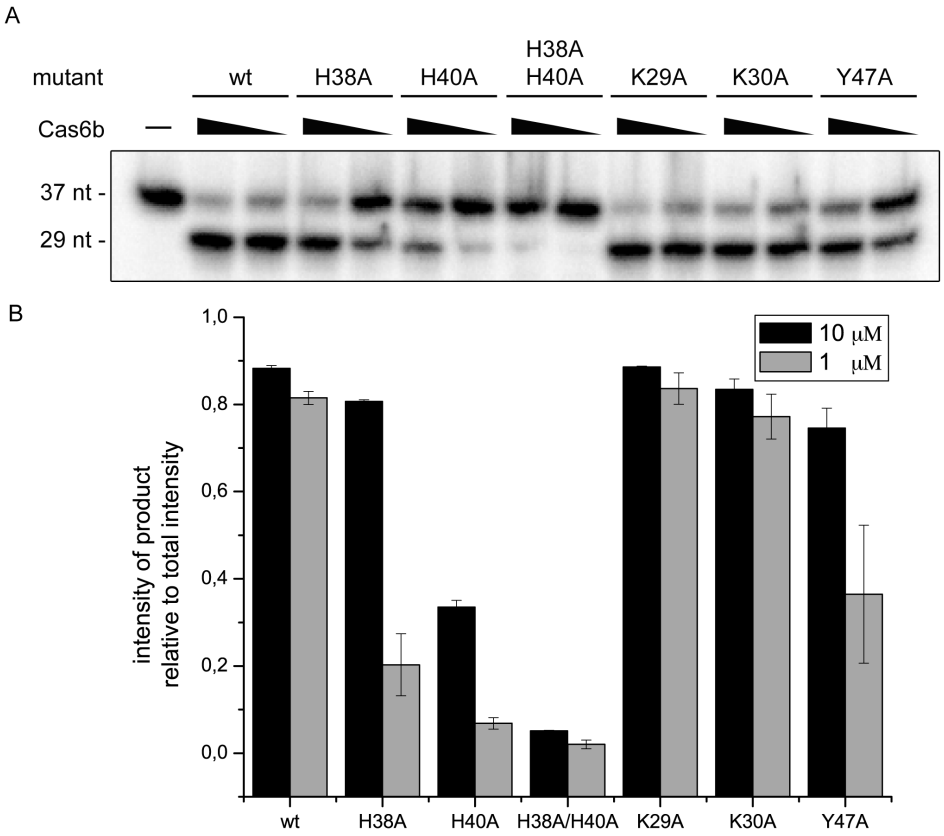


Figure 6. Two histidine residues play a critical role for Cas6b activity. Cas6b endonuclease assays were performed with the indicated Cas6b variants and 5'-end-labelled repeat RNA substrates. (**A**) A representative assay shows the activities of the Cas6b variants. While mutation of two lysines at Position 29 and 30 did not show any influence on activity in comparison to wild-type (wt), a mutation of two histidines at Positions 38 and 40 as well as a mutation of tyrosine at Position 47 show reduced processing activity. (**B**) Product formation for three independent reactions was quantified.

Discussion

The observed crRNA processing and abundance patterns of CRISPR/Cas subtype I-B are in good agreement with crRNA maturation previously analysed for other CRISPR/Cas subtypes and substantiate that 8-nt 5'-terminal and trimmed 3'-terminal crRNA tags are an universal feature of many CRISPR/Cas subtypes. The surprising observation of a spacer sequence that promotes crRNA production internally in *C. thermocellum* exemplifies the effect that an individual spacer can have for the abundance of mature crRNAs and subsequently the efficiency of entire CRISPR regions. The exchange of 1 nt of the otherwise universal 5'-terminal 8-nt tag of the crRNAs in the vicinity of the internal CRISPR transcription start site opens the possibility that two CRISPR elements might have been fused and subsequently portions of a leader element might have been incorporated into the repeat - spacer - repeat pattern. In addition, spacer elements were shown to promote transcripts of the reverse orientation. These anti-crRNAs were first described in Sulfolobus (Lillestol et al, 2009) and were also found in *P. furiosus* (Hale et al, 2012). However, they appear to be absent in most organisms. The occurrence of specific processing patterns for anti-crRNAs from different repeats of *C. thermocellum* CRISPR and the absence of anti-crRNAs in *M. maripaludis* indicate that this phenomenon might be specific for organisms with relaxed transcription start site definition rather than for the CRISPR/Cas subtype. Individual anti-crRNAs appear to be better suited for a conserved maturation process at their termini by a currently unknown mechanism. Reverse transcripts can form double-stranded RNA duplexes that might reduce the abundance and efficiency of crRNAs. Taken together these results highlight the effects that individual spacer sequences within a CRISPR region can have in both forward and reverse direction. These strong effects will need to be taken into consideration in the anticipated and proposed design of synthetic CRISPR regions for biotechnologically or medically important processes. We identified the Cas6b endonuclease responsible for crRNA maturation in the CRISPR/Cas subtype I-B. Cas6 enzymes are among the most diverse members in the sets of Cas protein of the different CRISPR/Cas subtypes and can be used to classify CRISPR/Cas systems. The similarity of repeat sequences and Cas6b enzymes between Bacteria (i.e. Clostridia) and Archaea (i.e. methanogens) hints at a horizontal transfer event for these CRISPR/Cas systems. Cas6 proteins might show this remarkable degree of divergence due to their individual adaptation to the given repeat sequence and/or structure. Evidence for this can also be found in the different principles of recognition for Pf Cas6, Cas6e and Cas6f. While Pf Cas6 binds to unstructured RNA, both Cas6e and Cas6f need a secondary structured RNA to bind and process *in vitro* (Haurwitz et al, 2010; Sashital et al, 2011; Wang et al, 2011). For Type II CRISPR systems, no Cas6 activity was reported. In these systems, the presence of a guide RNA (tracrRNA) recruits RNase III for the

processing of crRNAs (Deltcheva et al, 2011). These pathways exemplify the differences in crRNA maturation among organisms and CRISPR subtypes. The Cas6 enzymes Pf Cas6, Cas6e and Cas6f of different CRISPR/Cas subtypes were all shown to require a single conserved histidine residue for catalysis (Carte et al, 2010; Gesner et al, 2011; Haurwitz et al, 2010; Haurwitz et al, 2012; Sashital et al, 2011; Wang et al, 2011). In this study not one but two conserved histidine residues were identified for Mm Cas6b. Only the simultaneous mutation of both histidine residues resulted in a drastic loss of endonuclease activity. This implies that Cas6b exhibits the first example of a more flexible catalytic core in which both histidine residues are potentially representing the catalytic histidine and able to complement the loss of the other residue. Why did Cas6b evolve two catalytic histidine residues where this function can be fulfiled by a single histidine in other Cas6 enzymes? One possible explanation is the advantage such setup would have in coping with different substrates, e.g. with crRNA precursors that contain spacer of different length or structure. In Pf Cas6 a catalytic triad was described that provides a catalytic site for general acid/base catalysis (Carte et al, 2010; Wang et al, 2011). Mm Cas6b does not contain an identical catalytic triad but our observation of the importance of tyrosine 47 for Mm Cas6b activity and the occurrence of clustered amino acids that could provide general bases and acids might indicate more flexible active site architecture.

In conclusion, we provide the first description of crRNA processing *in vivo* and *in vitro* for CRISPR/Cas subtype I-B. These analyses of Cas6b in a bacterial and an archaeal model organism highlight the similarities between different CRISPR/Cas subtypes and the differences in crRNA processing. Two interchangeable catalytic histidine residues in Cas6b and internal promotion of crRNA production in *C. thermocellum* exemplify two new concepts that were found for CRISPR/Cas I-B systems.

Acknowledgements

The authors thank Andreas Su for technical help and André Plagens for advice and discussions. The authors are very grateful to Michael Rother, Rolf Thauer and William B. Whitman for their help with the handling of *Methanococcus maripaludis*.

Supplementary Material

Supplementary Figures

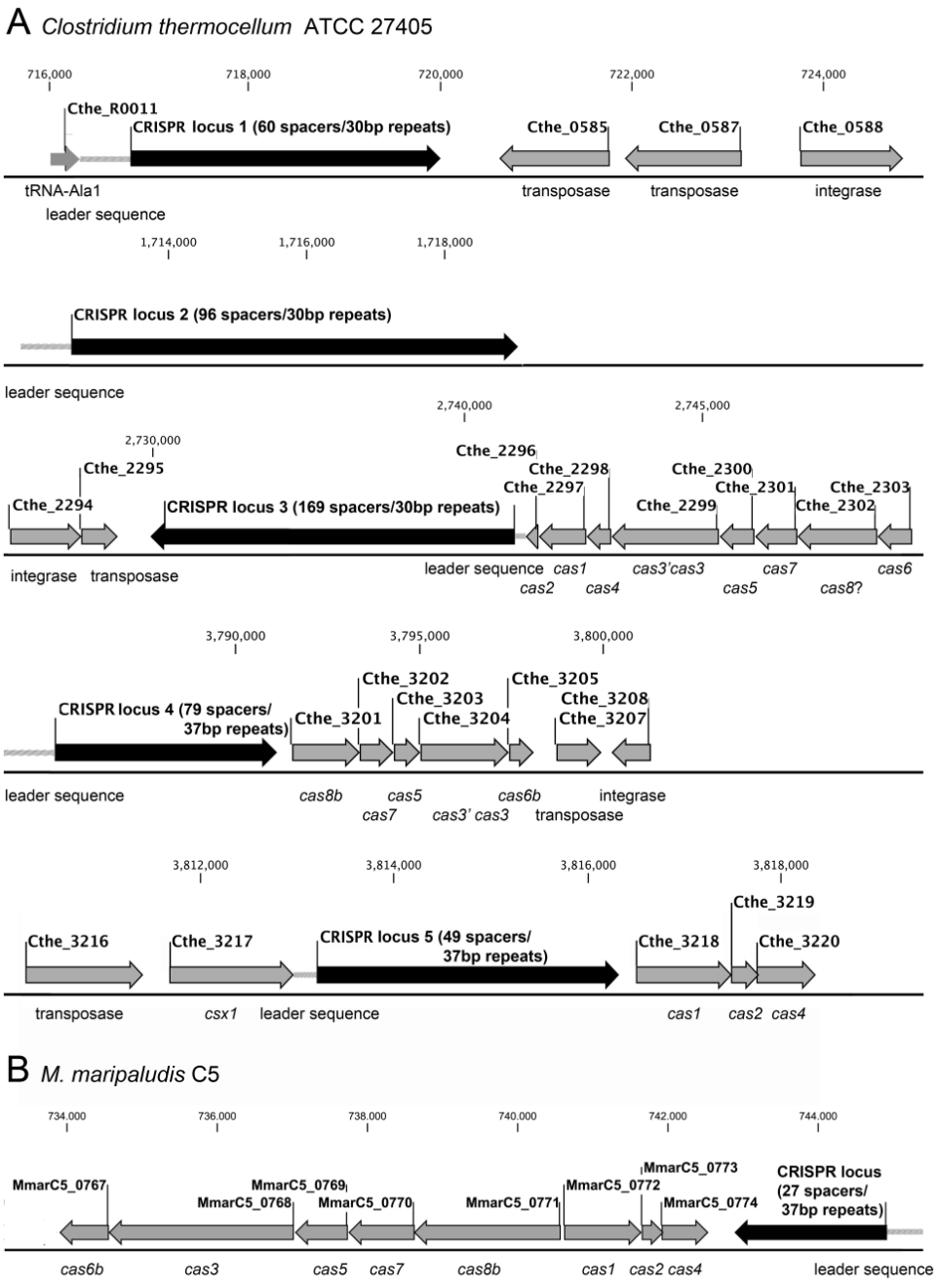


Figure S1. Genomic context and gene organization of the CRISPR loci in *C. thermocellum* and *M. maripaludis*. Schematic view of the CRISPR loci organization in a) *C. thermocellum* and b) *M. maripaludis*. Indicated are the number of spacers and the repeat length.

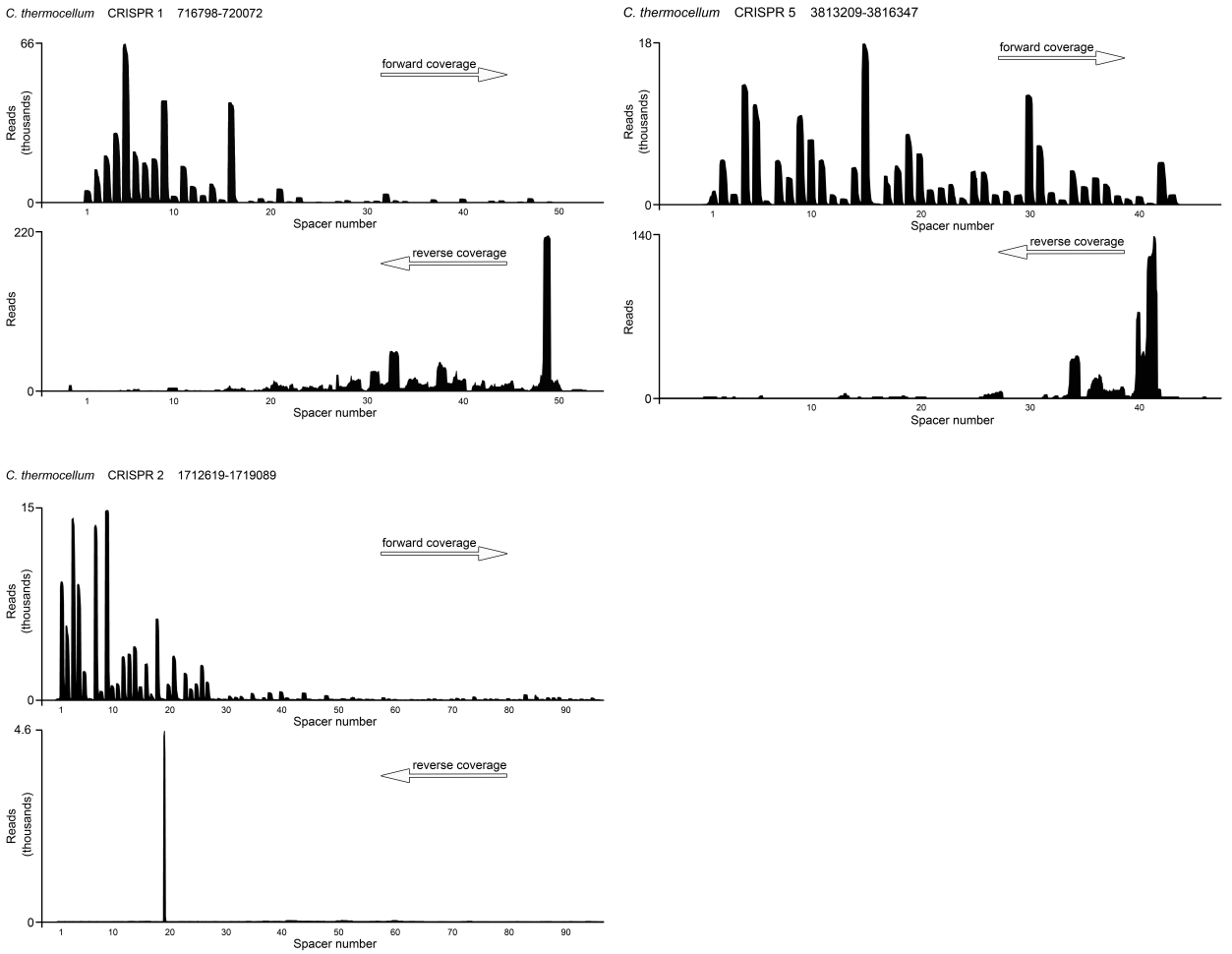


Figure S2. CRISPR RNA processing in *C. thermocellum*. Illumina HiSeq2000 sequencing reads were mapped to the *C. thermocellum* ATCC 27405 reference genome and CRISPR loci 1, 2 and 5 are displayed. Forward and reverse reads highlight the occurrence of processed anti-crRNAs.

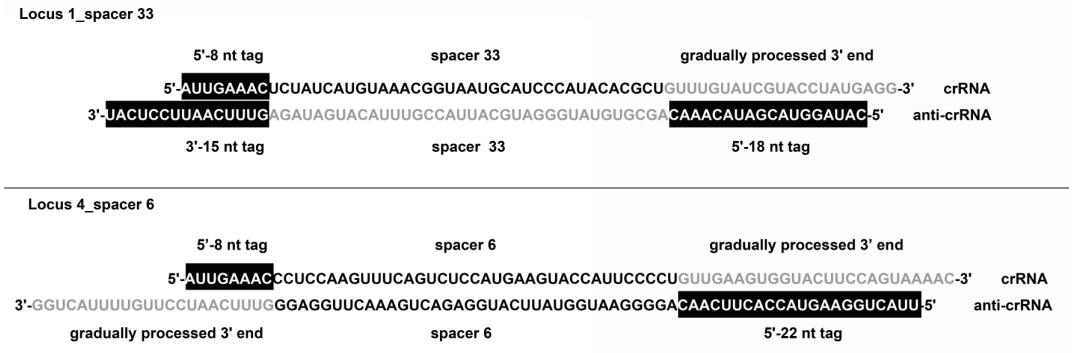


Figure S3. Processing patterns of crRNA and anti-crRNA transcript maturation in *C. thermocellum*. Examplary transcripts for CRISPR locus 1 (30 bp repeats) and CRISPR locus 4 (37 bp repeats) are shown. 5' and 3' terminal tags that display a conserved cleavage pattern are indicated by black boxes and gradually processed termini are indicated in grey.

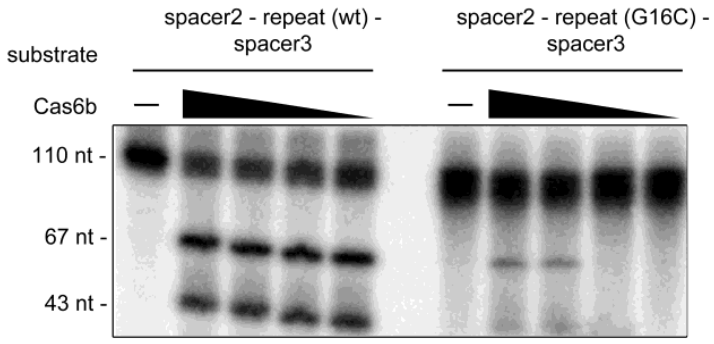


Figure S4. Repeat structure influence on Mm Cas6b activity. Mm Cas6b endonuclease assays were performed using internally labelled spacer2 - repeat - spacer3 substrates containing a wild type repeat sequence and a repeat variant with a mutation of G16 to C (G16C) that eliminates a G-C basepair and the computationally predicted hairpin formation (RNAfold). Mm Cas6b processing of the mutated repeat structure is strongly impaired.

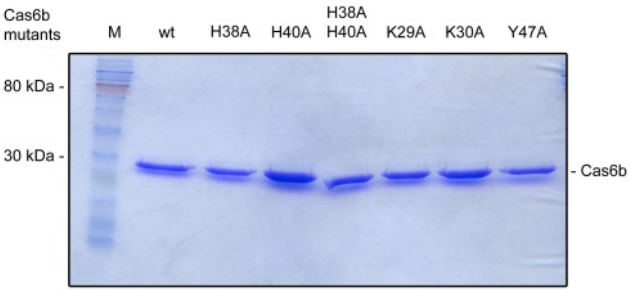


Figure S5. Purity of wild-type (wt) Cas6b and mutant variant preparations. 5 µg of the indicated purified protein samples were separated on a coomassie blue stained 15 % SDS-PAGE.

Supplementary tables

Table S1 Abundance patterns of *C. thermocellum* crRNAs. The crRNA production of all 5 *C. thermocellum* CRISPR clusters is detailed.

C. thermocellum CRISPR 1
genome position: 716798 - 720142
61 units
GTTTGTATCGTACCTATGAGGAATTG AAAC (30 bp)

spacer #	start position	end position	spacer sequence (sense direction)	# reads	5'tag	comments
1	716828	716864	UAAAACAAAGGAGAAUUGUAAAUGGAAG AGUAUUUGA	4366	AUUGAAAC	
2	716895	716930	UCGAGAUUUUGGAACUUUGUUUCGCCAU UCGUGUCU	11390	AUUGAAAC	
3	716961	716997	UCCGCAAAAGAGAGAAUCAGAGCUGCCG AACUUUUGG	16074	AUUGAAAC	
4	717028	717062	CGUUUCAUGGAUUUGAUAAGUUCAAGCU CUAUCUC	23737	AUUGAAAC	
5	717093	717128	GAUUCAGAAUUGAACGAGUACGUCAAAA GCGUUUUA	53779	AUUGAAAC	
6	717159	717195	UAGAUGAAUCAAAUAGUUCUACUGGACU UGGAGAUAC	17396	AUUGAAAC	
7	717226	717261	AAAUAAAUCUAAAGGGUAUGCCACGGAA ACGUUAUA	13672	AUUGAAAC	
8	717292	717327	CCCAGUAUGGAAUGGCGGUGGAACUGU CCGGCUGGC	15046	AUUGAAAC	
9	717358	717394	GCUCUAACAGUAAUUCAAUUAUUUUUGU AUUCAUUUG	34638	AUUGAAAC	
10	717425	717461	CGGGCCGCGCCUAUUUGCGUGGGAACC CUUAUAAUGG	2467	AUUGAAAC	
11	717492	717527	CGGUACUGCUGUGGUUAUCGACGAGAU UCAGACUUG	12496	AUUGAAAC	
12	717558	717593	ACGGAAGAGAAAAAACAGCAGAUAAUACA AAGCGUA	5756	AUUGAAAC	
13	717624	717660	UCAUUUUUUUCACGACUAUACAAUAUAC CAUUUGAAU	2599	AUUGAAAC	
14	717691	717725	UGGCUGAAAGGUUAUAGGUUGAAAGGAA AAGUCCC	6538	AUUGAAAC	
15	717756	717791	UCUCGCCUUUUGGCGGUUCGACUUUUG CCGCUUUCA	1286	AUUGAAAC	
16	717822	717859	GAAUUGCCAUUAAGUCUUUGGUUUCUUG UCUUGUCAUG	33989	AUUGAAAC	
17	717890	717928	UACAGCCUCCUUCAUGCUUUCUCCUUUG UCAGCUAACUC	201	AUUGAAAC	
18	717959	717994	AAGGCUGGGGUGAGUGGGUUGAAGUAA CUGAAGGUG	757	AUUGAAAC	
19	718025	718059	UUAUUUUUAGUACAUCAUUAUUUGUGAU UAUUCUA	1594	AUUGAAAC	
20	718090	718127	UUUGCCCUCUCCUUUCUUUUCUUUUGUU UGUUUUAGCG	675	AUUGAAAC	
21	718158	718193	UCUACCUGUUGAGCUUCUGCCACUUGUU CUAAUGUA	4869	AUUGAAAC	
22	718224	718260	UUGUUCAGCCUCCUUUCUAAGCAAUGUA AUAAUAAAA	566	AUUGAAAC	
23	718291	718328	CUUCAUAAUCUUUAGGGGUUCUUUUCUU CAUAUUGUCA	1925	AUUGAAAC	
24	718359	718396	CAACACUCCUUUAUUAUUAUUAUUUUUC CUGCCUUUGC	353	AUUGAAAC	
25	718427	718462	UAGGCCAGCUCGAAAAAGCCGGCGCUAA CACCGAAG	432	AUUGAAAC	
26	718493	718529	AUGUGCCUCCGCCACGUCUUGUUAUAAU UGCUUCUCC	200	AUUGAAAC	
27	718560	718595	CAUUGUUGCCUAGUUGUUUUUUUGGUU UUACUUCUC	671	AUUGAAAC	
28	718626	718663	AUUGCCGGCGGCAAGUCUGAAGCGCCAA GAAUUGACCG	942	AUUGAAAC	
29	718694	718729	CCGACUGUUCGACGUAGACGUCGACAG GAGGAAAGA	148	AUUGAAAC	
30	718760	718795	CCUCAAUUUCAUUGUAUCACCGGAGCAG UGGAGCGU	683	AUUGAAAC	identical with spacer #34

31	718826	718861	CAGAAAACUUGUGAACUACCGCCACUUA UAGAAGUG	758	AUUGAAAC	
32	718892	718928	UUGUUCUGCAACUUAUGUAAACAUGUAG UCGAAAUCA	3153	AUUGAAAC	
33	718959	718994	UCUAUCAUGUAAACGGUAAUGCAUCCCA UACACGCU	970	AUUGAAAC	
34	719025	719060	CCUCAAUUUCAUUGUAUCACCGGAGCAG UGGAGCGU	672	AUUGAAAC	identical with spacer # 30
35	719091	719126	CCCCAUUCAAGAAUAUUCACUGCUAUUC ACCUUCUU	50	AUUGAAAC	
36	719157	719192	AUAAUUCACAUAAUGCAAAUGUUGCACA ACUUGUAU	325	AUUGAAAC	
37	719223	719258	CCGAUAAGCUAAAUGUCAACGCUGAGAU GAUGGUAA	1261	AUUGAAAC	
38	719289	719325	AAUCCCGCCCAUUCGGCUGCCAAAUCAA UUCCGCCUA	48	AUUGAAAC	
39	719356	719392	UAUUGUGCCAAGUGUCAAUACUUUUGGU ACAAUAUUU	115	AUUGAAAC	
40	719423	719458	CAAAUAGGUCUGUAUACAUUCUAUUUUG AUACAAGU	1486	AUUGAAAC	
41	719489	719524	UUGUGGCGACAGAACGGCGGGAAUAGC AGUGUUGCA	257	AUUGAAAC	
42	719555	719591	AUGUAACCAGCCCAUGUCAUUUCUGCCU UUUGCUCCA	90	AUUGAAAC	
43	719622	719658	UGAGGAUAAAAUACAGCAGUACCUUGAA CUAGAUGAG	790	AUUGAAAC	
44	719689	719725	AUAAAACUAUUUAUUUUUUGUUCACAAU UGUGUACAA	878	AUUGAAAC	
45	719756	719791	AAAAAUGGCGAGCAGGCAGGGGGUUAG GACCCACCC	180	AUUGAAAU	
46	719822	719859	CGAUUUUGCAGACAGUCAAGAAGAUGCA GAACGUAAAC	496	AUUGAAAU	
47	719890	719927	AGCAAAUAUUUGAUUCAAAAAUAGAGCU AGUAAAGAAU	1587	AUUGAAAC	
48	719958	719996	UUUUUCAAACUUGCAGGGCUAGGCAGGA AUCGAACCUGC	250	AUUGAAAC	

C. thermocellum CRISPR 2
genome position: 1712609 - 1719059
97 units
GUUUUUAUCGUACCUAUGAGGAAU UGAAAC (30 bp)

spacer #	start position	end position	spacer sequence (sense direction)	# reads	5'tag	comments
1	1712639	1712672	CUACAUUGAAACUGGCAAUGCUACAGAA GCAGCA	8733	AUUGAAAC	
2	1712703	1712742	CUAUCUUUGUCAAACAGCAACCACAUGA AAGCGUCGGCAA	5358	AUUGAAAC	
3	1712773	1712811	AAUCUACUAUAAGUCAGUAUGAAACAUG CAAAAGAGAAC	13400	AUUGAAAC	
4	1712842	1712877	GGCGGUGUUGGUGCACAUACAAAACUUC CAAGCACA	8589	AUUGAAAC	
5	1712908	1712945	CACCUCAUAUGUUCAGGCAUACAUUUGU UUAUAUGGCA	2097	AUUGAAAC	
6	1712976	1713011	UCUUCUUAGCCUGCCACUGCUCCCAUCU GAAACCUU	97	AUUGAAAC	
7	1713042	1713078	UAAAAAUAUCUUCGGUAAUUGCAUCCGU UUUAAUUUC	12884	AUUGAAAC	
8	1713109	1713146	AUUUUAAAAAACCCAUAUUAUAUCCUCCU UUAGUUUUC	623	AUUGAAAC	
9	1713177	1713213	UACUUUGCUUAAGUUUUCAUAAUUAUCA GAUUUAUC	14114	AUUGAAAC	
10	1713244	1713279	UUGUAACCGAGAAUGUCCCUCCAAUGUU CUAACUUU	1050	AUUGAAAC	
11	1713310	1713346	GUGAUUGCAGACCCUGUUAUUAUUCAAG CCGCUGAAA	1205	AUUGAAAC	
12	1713377	1713414	AAGCCGGAUUUAGUGAGGCUUUUGGUG AGUCUAAUGUC	3164	AUUGAAAC	
13	1713445	1713483	AUAAACCUCAGACAUAUGUCACAGCAAAA GGCGUUUCCU	3405	AUUGAAAC	

14	1713514	1713549	GCCCAAAAGGUAGCUGUAGCUGUUGCU GGCGCAAUU	3955	AUUGAAAC	
15	1713580	1713618	UUUUCCCCAGGUGACUAAUUUCAGGAAC UUUUGUUAAUG	978	AUUGAAAC	
16	1713649	1713683	UCUUGAUAAAAUAAACAAUUUUUCUGUU GUAGAUA	2660	AUUGAAAC	
17	1713714	1713749	UGAUGUGACUGUUGUGGUUCUCUACUG CUGCGCCUU	431	AUUGAAAC	
18	1713780	1713816	AUUCAAGUCUCUUUUUUUUGGUGUUGU GCUUGCGAGC	6008	AUUGAAAC	
19	1713847	1713885	UUCUAUGGUCGGAGUGACUGGAUUUGA ACCAGCGACCUC	4149	AUUGAAAC	spacer sequence promotes anti-crRNA production
20	1713916	1713952	AAGGCGAAAUAGGUUAUAGAAAAAUAAG UAAUAUAAC	1051	AUUGAAAC	
21	1713983	1714020	AACAGGAGCAGGAAGACAAUGAAGCCAU GCGCCAGGGC	3205	AUUGAAAC	
22	1714051	1714087	CCAUAUCACCAAGUAAAUUUGUUUCAAC UCCGGCACC	91	AUUGAAAC	
23	1714118	1714154	AGAGAAGGGUUUGUAGAAGACGACUUUU UUACUGUCA	1972	AUUGAAAC	
24	1714185	1714220	CCGACGCUGACUUUGCCGAGUGUGAAG AAUGGCUGA	814	AUUGAAAC	
25	1714251	1714288	CUUCACGGGAAAGGUUCUUGCUGUAAUA UUCGGCCAUA	1184	AUUGAAAC	
26	1714319	1714354	AGGAACAGGACUUCAUACAACGUCACCG UUGAAGUG	2562	AUUGAAAC	
27	1714385	1714421	UAAGCCAUCGGGCAUGACGAACCAUUAU CGCAUAAAC	1308	AUUGAAAC	
28	1714452	1714487	AUCUCCAAUCCACCUCCGGUAUGCGGCU GCGUCUGC	62	AUUGAAAC	
29	1714518	1714557	UAAAGCUCUGGUCGGUAAAUGUGCGGAU UGGUCGCUCUGC	102	AUUGAAAC	
30	1714588	1714624	AACUCACCUUCUCCCCAUUUACCGGACA AACUAACCG	50	AUUGAAAC	
31	1714655	1714692	GGUCUAGCUGUGUACCAGAAAUACUGGC ACAGCUUUUC	271	AUUGAAAC	
32	1714723	1714758	GGACUGGCUGCAACCUUGAACAUAAACU UGAGUAUG	188	AUUGAAAC	
33	1714789	1714823	GAGAGGCCUUUUGGCCGGAGGAAGUUG UGGAGGUG	249	AUUGAAAC	
34	1714854	1714891	UCGCCAUAUCUACCACAUAACCAGCGGU UGUUGCCAAU	31	AUUGAAAC	
35	1714922	1714958	AUAGCAUUUCUUGAAAAAAACCGUAAAG AGUGGGGAU	470	AUUGAAAC	
36	1714989	1715027	UUUUUUUUGCAAUUCCUUGUAACCCUGC UUUGCUACUCA	35	AUUGAAAC	
37	1715058	1715097	UACCUUUAUUAGUCGAUACAACACUUAA AGUCUAUGACAA	165	AUUGAAAC	
38	1715128	1715163	CAUACAGCGAUGGUAGAAAGGUGAAUAU UGCUUCCU	524	AUUGAAAC	
39	1715194	1715232	UUGCCGCACUGCCCAAAACCAGCCCAAU UUUUGAACCCA	18	AUUGAAAC	
40	1715263	1715303	UUCGACGUAGAAAAUGCAAUAUCUAUCG GGGAAUGGUACUU	610	AUUGAAAC	
41	1715334	1715370	CCGAACAUUUUAUGCUCGUAGGGUACAU CUUCGCAGA	172	AUUGAAAC	
42	1715401	1715437	CACGGCAGCACAACCAGCCGCACGUCUG CUUUGGCUU	47	AUUGAAAC	
43	1715468	1715502	CCGUUUCCCACAUCUGAUUGAGUUGUUU GUUGAUA	67	AUUGAAAC	
44	1715533	1715569	UGCAAUGAUUUCGAAUUCUUUGUAACAU UGACUUUGG	512	AUUGAAAC	
45	1715600	1715635	AACCAUAAAAAAGUAAACUGAACCAUAAA UGAUACA	60	AUUGAAAC	
46	1715666	1715702	UGCCCCAUUCUUUUCUUCGCUCUCCUUC GCCCAUUAG	10	AUUGAAAC	
47	1715733	1715771	ACUUUUUGUCUUCACCUACCAUGUCCAU AUUCGUCAUAA	32	AUUGAAAC	
48	1715802	1715839	CUUUUAUAAAUUGAAAAGGAGGAUAAGU AUGUCAAACA	316	AUUGAAAC	
49	1715870	1715905	UACCAUCUGAGGGCUAUUUUGAAAGGUU CGUCAGUG	47	AUUGAAAC	
50	1715936	1715973	UCGGUAUGCACAAAGAAGUACAUACCGC AAUAUCUCUC	41	AUUGAAAC	
51	1716004	1716038	GCUGCCAUUUUGACAAUGGUGUUCUGC UGAUCUGC	109	AUUGAAAC	
52	1716069	1716103	UCUCGCUCGAAUAAAAUUGGUACAUCAU AUUUUUC	39	AUUGAAAC	
53	1716134	1716172	AAUAAAAAAUUUAGCAUACUUUAUAUUGA CCUUGAAGCG	183	AUUGAAAC	
54	1716203	1716242	UAGACAUGGAUUAUGUGAACGUUGUGGC AGACCGGGAAAG	65	AUUGAAAC	

55	1716273	1716308	GACUAAGUCCUGCAUCAGAUGGCAAUGUCGGUUCAG	63	AUUGAAAC	
56	1716339	1716374	AAGAAAACUGACAAUGCAGAAGCACAAAGAGAAAAG	55	AUUGAAAC	
57	1716405	1716440	UUCACCUCCUUUUUUCAUGAACUUUGGUCCUCCAAG	14	AUUGAAAC	
58	1716471	1716508	AUCAUAUAUGAUUCAUGCAAUGGAUGUCAAUAUGGAUG	174	AUUGAAAC	
59	1716539	1716575	GGUUUAUAGGCUUUUAGUUCUCCGCUGUGUAUCGCUC	94	AUUGAAAC	
60	1716606	1716643	CCGGCUUCACCUAUCAUCGCGAGAGUCGGAGCGUCUAC	69	AUUGAAAC	
61	1716674	1716709	CCUGAAAAUCUUCAUUCAAUGCGGCUAUGCCACUGC	16	AUUGAAAC	
62	1716740	1716775	AAUAAAUUUUCCUUUUCCAUUUAUUUACCUCCUUUU	12	AUUGAAAC	
63	1716806	1716844	UCAAGCAUUUCUUUCGGUAAUGGUUCGCCGAAAUCAAUC	44	AUUGAAAC	
64	1716875	1716910	AUUUUCCUCCUUUAUAUGUAUAAAAGACCUUUAUAA	20	AUUGAAAC	
65	1716941	1716980	GCCAUAUAUACACCCCCUACAGGUAGCUUGGGGCAAACUC	29	AUUGAAAC	
66	1717011	1717046	AUCUAAACGAAGAACAAUUGCAAAAGCUAAAUGACU	75	AUUGAAAC	
67	1717077	1717116	CAUUGAAACCGAAUGGGUUGAAGAAGGCGGUUUCUUUAGA	86	AUUGAAAC	
68	1717147	1717184	UACUAUCUGUAAAAUACUUGUAGCCUUGUCUUUUAAUU	36	AUUGAAAC	
69	1717215	1717251	CCAAUCCAUUCUUUCCGCAUCUUGUACUGCUGAUACC	5	AUUGAAAC	
70	1717282	1717317	UUUGUAUAAAGUGAGGUAAAUAGUAAAUAUAUAUGC	69	AUUGAAAC	identical with spacer # 78
71	1717348	1717386	AAAUCAAAGAAUCGGCAAAAGGAAAGACAUACGAAAUUG	128	AUUGAAAC	
72	1717417	1717453	UGCCUGCAAUGUUCUCGAUGGCAUUGUCGAAGAUGAG	91	AUUGAAAC	
73	1717484	1717520	UCCUCCCCUUGCCAUGUUAUGGUUUCGCCCACUCUUU	13	AUUGAAAC	
74	1717551	1717586	UCACCGUUAGGAAAGUUGAUAGUGAACACGCUGUUA	238	AUUGAAAC	
75	1717617	1717652	UACAAUAUACACAAGCAGAAUUCUUACAAAUAAUAA	19	AUUGAAAC	
76	1717683	1717720	CAAAAAGGAGGAGUAGUUCGUUUUGAGGAAAUAGACGC	79	AUUGAAAC	
77	1717751	1717787	AGUGUUCUAUAUGUAUUUGCUUUUCCUAAGCUUUUUG	82	AUUGAAAC	
78	1717818	1717853	UUUGUAUAAAGUGAGGUAAAUAGUAAAUAUAUAUGC	66	AUUGAAAC	identical with spacer # 70
79	1717884	1717923	CGCUUUUUGCUACUUUGAUUUACGCCAUUUAUUAUUCAAU	19	AUUGAAAC	
80	1717954	1717992	UGAACUGCACCUAUCCGCCAGUUUCCAGUCAUACUGGAU	108	AUUGAAAC	
81	1718023	1718060	UGAUUUUUAAUGUAUUGUAAUAUUUCCAUUUCCUUUUU	35	AUUGAAAC	identical with spacer # 82
82	1718091	1718128	UGAUUUUUAAUGUAUUGUAAUAUUUCCAUUUCCUUUUU	52	AUUGAAAC	identical with spacer # 81
83	1718159	1718194	GAGCCGGACGUGUGAAAAACUAUAUCCCGGAAGAUC	398	AUUGAAAC	
84	1718225	1718261	UAAUUAAUUAGCUUAAAUAAGCUAUAAUUAAUAUUGA	42	AUUGAAAC	
85	1718292	1718327	AAAUAAAAUUGUAAGAAUGAGUUAAGGGGUUCAAAA	355	AUUGAAAC	
86	1718358	1718396	UUUCCAGGCUCAAAAAUAAACCUAAUCAUUGUUAAAACC	55	AUUGAAAC	
87	1718427	1718462	UACGGACUUGCAACCGUCUAUAACUUGUUGUUAUAG	174	AUUGAAAC	
88	1718493	1718528	AAUCAAUUUUGCUUUUAAAUAGAGUUUAAUGUAUAC	164	AUUGAAAC	
89	1718559	1718595	GCAUGUUUUAAACUGUCAAGUAUUUUUACUUUACAGA	193	AUUGAAAC	
90	1718626	1718661	CGGUGGAUUAAAUGAGCUUUAUAAUCUGCUGGAUAA	24	AUUGAAAC	
91	1718692	1718727	AUCAAACCCUGAUAAAUAGACGGCUAGAUGCCAUUU	99	AUUGAAAC	
92	1718758	1718793	UACUUGAGUCUGUCCUUGAAGGCAUGGCUGAAUACU	31	AUUGAAAC	
93	1718824	1718859	UAUAGCAAAGGACAACGAAGUAAAAGAAAACUGGAA	69	AUUGAAAC	
94	1718890	1718925	UCAAAAAACACUGCGUAACUCAUUAUGCAUAACUCC	44	AUUGAAAC	

95	1718956	1718992	GCAUCUGUCAAUUUAUCACGGAUUGCAG UAUCGAUUU	154	AUUGAAAC	
96	1719023	1719059	GUGAAUUCCGAUGCUGUUUCCCUUUCG GGUAAGUUUU	27	AUUGAAAC	

C. thermocellum CRISPR 3
Genome position: 2740990-2729773
170 units
GUUUUUAUCGUACCUAUGAGGAAU UGAAAC (30 bp)

spacer #	start position	end position	spacer sequence (sense direction)	# reads	5'tag	comments
1	2741020	2741051	CAUCAUGCAUUGGCGCUGGCAGCAUCAA GUUUGUCUUCC	4677	AUUGAAAU	
2	2740953	2740982	UUCAAGCCGCUGAAAGAAAAAAAGUUAG ACCACUCUA	7875	AUUGAAAU	
3	2740886	2740915	AAGAGUUUGAAGCGUGCAAGCUUGAAUU UGCUGUUGA	17490	AUUGAAAU	
4	2740821	2740848	UUUGCAGCUUCGUUUUUGCAAGUUCUAC UGCUGUC	8069	AUUGAAAU	
5	2740753	2740783	AGGAAGCCUUUUGGCCGGAGGAAGUUG UGAAAGAUAAG	6601	AUUGAAAU	
6	2740687	2740715	CGAAAACAAUCUACCGGUUGAUAAUAAG CCGGACCC	5604	AUUGAAAU	
7	2740620	2740649	UUGGGGCAUUGGUAUUUGGCGGAAUCC CUGCGUUUGC	2390	AUUGAAAU	
8	2740552	2740582	UGCACAAUCCCAUAUUUUGCAUACAAGU CUAUUGCUUC	8496	AUUGAAAU	
9	2740486	2740514	UUCGGUUCAUUGGCACUGGGAAUCCCA GCAUAUUUG	7118	AUUGAAAU	
10	2740417	2740448	AACUAUAACGGCACAACUUAUGUUCCUC UUCGUGCUAUU	1559	AUUGAAAU	
11	2740350	2740379	AUAUAUUGGCUGAGCAAGAAAAGAAGUU GCUUUCGAU	5109	AUUGAAAU	
12	2740285	2740312	AUUACGAAGCAUUUAUAGAGAGUUAUGC UAAUACA	1837	AUUGAAAU	
13	2740216	2740247	AUCUCAGCUUCUCUUUUUGCUCUUUCGU AAGCGUCGCCA	819	AUUGAAAU	
14	2740150	2740178	UUUCUGAAUUCGCUGGAAAUUUGUCCGU CGUACGAA	11942	AUUGAAAU	
15	2740082	2740112	UGCCCCAUUCUUUUCUUCGCUCUCCUUC GCCCAUUAGC	27	AUUGAAAU	
16	2740015	2740044	ACUAAGAAAAUAGCAAGAAGACUCCCAC CUCUUUAGG	2368	AUUGAAAU	
17	2739949	2739977	GGAUCGUUGUUUGCUCCUCCGGAUUUU GUCUUCAGA	459	AUUGAAAU	
18	2739880	2739911	CUACAUUUGCAGGCAAUUUUUCAAUUCU CUAUCUAAUGG	562	AUUGAAAU	
19	2739813	2739842	UCUUCUGCUGUCUCAUCCCAUAUUUUUU GACCUUCUU	33	AUUGAAAU	
20	2739746	2739775	ACCUCUUGCUUCGUAACGAAUGUAACAC UAGAACCAG	522	AUUGAAAU	
21	2739679	2739708	CCCAAACUUUCAAUACUGGCUAAAAUAU CACCAAACC	19	AUUGAAAU	
22	2739612	2739641	UUUUGCAAUGUUCUGCUACAUAAUCAUA UACAAUUUC	202	AUUGAAAU	
23	2739546	2739574	GCAUAAUUCCUGUACCGGAAACGGUAGA AAUUAGCU	1261	AUUGAAAU	
24	2739481	2739508	CCAUAAAAAUCACAAGUGUAGUCAGGCC AGCGUCC	350	AUUGAAAU	
25	2739412	2739443	UCUACUAUUUGAGCUGGUAAUCCGCUAU GAAUCUUACGU	840	AUUGAAAU	
26	2739344	2739374	GCUGUUUUCUCGCUAUAACCAGCCCUUC UUGCUGCUUC	61	AUUGAAAU	
27	2739277	2739306	UUUCUGAAUUCGCUGGAAAUUCUUCCGU CGUACGAAA	1869	AUUGAAAU	
28	2739207	2739239	AAGGCUCACGCAGAUGCUAUAAGAACAG AAAUAGAACAGC	191	AUUGAAAU	
29	2739138	2739169	UCCAAAAUUAUUACGCAGAUGGACAAGC UUGAAAACGAA	495	AUUGAAAU	
30	2739069	2739100	UGAAACCUUUGUGUAAUCGUCCAGUGAU ACAGUUAGGAA	1496	AUUGAAAU	
31	2739003	2739031	UUGAUUUCAAUAUCAAACUGUAUAAACC UGCUUCGU	2325	AUUGAAAU	
32	2738937	2738965	UACUUGAGUCUGUAGAAGGUACUGCUGC	453	AUUGAAAU	

			CCAAAACC			
33	2738869	2738899	GGCAAGAAGUAGUAGAAGGUGUGAACCC ACCCUAUCGA	792	AUUGAAAU	
34	2738802	2738831	UGUUUUUCAAUGCUAUUUGUUCACCGAU UGAAUCCAC	367	AUUGAAAU	
35	2738735	2738764	CUGGAUUGCGUAGAGAUUUGUACUUUU GAAUUUUAAG	1223	AUUGAAAU	
36	2738669	2738697	UUCUUUUUGAAGUUCUUUAAGUUUGCUU UGCCAUUC	289	AUUGAAAU	
37	2738602	2738631	AGAAAUAGAAUUUGGUAUAGAGGAGACC CAUCAGAGC	384	AUUGAAAU	
38	2738535	2738564	CAUUUUGACCUAGCUUCUUUGGUUUUAC UUCUCUAGC	382	AUUGAAAU	
39	2738468	2738497	AUAGGUGAGUCAGAUGUAAAAAAACAGG GGAGGUAAU	121	AUUGAAAU	
40	2738402	2738430	UUGUCAACUAAGCAUGCGAACUGCUCCG GGCCAUUG	237	AUUGAAAU	
41	2738335	2738364	UGUUCCUUUGUCAAAUAAAGCUAUAUAC UAUCUAAAA	281	AUUGAAAU	
42	2738269	2738297	CGCCAAGUCUCCAAAGGAGCUACAGGAA AAAGUAAA	497	AUUGAAAU	
43	2738201	2738231	UGAAAACAACUUGCCAAUUGAUAAUAAG CCAGACCCAC	62	AUUGAAAU	
44	2738131	2738163	CAAAUAACAAAUAGUAAAUUGACCGUAG CAGAAUGGUUUG	243	AUUGAAAU	
45	2738064	2738093	CGUCCACGUCGAAUAGUCGGUAAUCCAU CAGCACCGA	109	AUUGAAAU	
46	2737998	2738026	AAUAAUUCUGCUAUUCCAGUAAUUUGUU UAAGUCGG	663	AUUGAAAU	
47	2737930	2737960	GACAGACAGAUGAAUAUGACGGAACCGU UUUUUGUGUG	247	AUUGAAAU	
48	2737863	2737892	AAGAUAAGGCUAACAAAAAGCUAUUGUG GGAAAUUUU	377	AUUGAAAU	
49	2737794	2737825	AUUGCUUCGUUGAUUCAGGAAGCAUAGC AUAUUCUUUUU	279	AUUGAAAU	
50	2737726	2737756	UACCGAAUCGCGCUUCUAUGAGUUGGAC CUUGUCGUCC	94	AUUGAAAU	
51	2737660	2737688	AUGCACUUGAUGAAGUAAUAUCCCAAUG GAUAGAUG	262	AUUGAAAU	
52	2737593	2737622	CGUUCGAUCACGAUGUAGUCUUCACCCA GAAGAAUCA	1039	AUUGAAAU	
53	2737527	2737555	UAUUGCCAGUGAUCACCAUAGUCGGCGG AAUGCUGG	252	AUUGAAAU	
54	2737459	2737489	ACGGGAACCUAUUCUGCACGAUCAGCAU UCGCGACAAA	322	AUUGAAAU	
55	2737393	2737421	CAGCCUGGAACAUUCCUGUUUCAGCGUC CUUCAUCA	25	AUUGAAAU	
56	2737327	2737355	CAUUCGGUACAAGAAAAAUCUUAUUGGC AUUACAAA	317	AUUGAAAU	
57	2737264	2737289	UUGUAGAUUUCAGCCUGCUCCUUCUUCA GGCGU	84	AUUGAAAU	
58	2737198	2737226	UCGCUGUGCGUGAAUUUCACUUUGUGU GUCAUGGUU	440	AUUGAAAU	
59	2737129	2737160	AGGAACCCUUCAAAGUAACCCUCAGACG GUAUAAAUGCU	57	AUUGAAAU	
60	2737063	2737091	AUCCUUGCAACCUCCAAGGCUUUUUCUU CAUCAGUC	191	AUUGAAAU	
61	2736995	2737025	UACCACCGUCUCUUACAUAACCGACAAG UCCACCUAUA	2	AUUGAAAU	
62	2736928	2736957	UACAUUUCAUCUCUGUCAAACCACCGUU CAUUUUCUU	15	AUUGAAAU	
63	2736862	2736890	UUUGUUGAUAGUCCAUGCAACUCUCCGG AAUCGUUA	92	AUUGAAAU	
64	2736796	2736824	CCUUGUUUUGACUUUGACGUAUCAAUGC UUGUUGCA	52	AUUGAAAU	
65	2736726	2736758	GAUCUUUCACUCCAAACUUUCGCUGUGU GAAAUAUAUCAC	46	AUUGAAAU	
66	2736661	2736688	CUAUGCGCCAAGUAAUUGGAUAUAUUGC AGGGUGU	95	AUUGAAAU	
67	2736595	2736623	GCAGGAAAGCCUAAGUUGUUCCAAACAG AAGAUGCA	210	AUUGAAAU	
68	2736528	2736557	ACUUAACAAUCGAAAAUGUUGAUUUUUG GAGUCCGAA	93	AUUGAAAU	
69	2736457	2736490	AUUUCAUCUCUUCAUACGGACUUUCAAA AUCUUCCCAAUCG	32	AUUGAAAU	
70	2736388	2736419	CUGAUAAUUACCCAAUAAUCGUAAACAU CGGAAGGAAGA	110	AUUGAAAU	
71	2736320	2736350	GCCAUCGGCAGCCACCUUCUCGCAAAUC AAUCUUGUCU	3	AUUGAAAU	
72	2736254	2736282	CUUGUUAGAUGAUUUCCGCUAACUUGUU CGCAGAUU	132	AUUGAAAU	
73	2736185	2736216	UACCAGAGAAACAUUCAGAUUGUAUAAA CUGGCAUAUGC	51	AUUGAAAU	
74	2736119	2736147	UUAGUUUUAUAGCCUUUAGGAAGGGCUU CGCUUUCA	66	AUUGAAAU	

75	2736051	2736081	CGGAUAGCUUGGGGUUUCUUACGCAAC CAAUUGAAAAC	95	AUUGAAAU	
76	2735983	2736013	AUGAUAUAAAGGAAGCUUUAGACGAGAA AUUUGCUUUU	26	AUUGAAAU	
77	2735914	2735945	UUUGAUAGUGGAAAUUAUUUUGGAUAUG AUAAAAGGUGC	39	AUUGAAAU	
78	2735844	2735876	UAAUAUGGACCCGCUAGAGUUUCGCGAG UAUGGUCUACAC	16	AUUGAAAU	
79	2735775	2735806	GCAAGAGCACGAACCUCAAAUCGCCUGU ACGGCUGUAUG	50	AUUGAAAU	
80	2735709	2735737	GUUGCUAAAGAAGAAAAUAAUGAUAAUC CUUCUCUA	56	AUUGAAAU	
81	2735643	2735671	AGAGAUAAACAACAAGACAAUGGACUGC AAUAUAUU	49	AUUGAAAU	
82	2735577	2735605	CAAACGGUUUGAAGAACGUUGGAGCGAG AAUCUCAA	111	AUUGAAAU	
83	2735508	2735539	AGGGGCUGUUAUGCCCCUCUUAUUAGG AACCUGAACCUG	3	AUUGAAAU	
84	2735442	2735470	ACAAAAGUAACACAGCCAAAUAAAGUGAA AGGUUCA	46	AUUGAAAU	
85	2735375	2735404	ACGGUGAACUGUAUGUAUCACCAAGUUC UAUAGGUCA	79	AUUGAAAU	
86	2735306	2735337	GAUUGUUGGAAAAAGGAAAUAGACGCAU UUUUUGAGUUG	44	AUUGAAAU	
87	2735238	2735268	GUUCCUCAAGCUGUGCAUCCCUAUCUGC AAUUUCUUUU	8	AUUGAAAU	
88	2735172	2735200	CAAAGCAAAAAACUGAAGCUAUGCUACU UUCAUCUG	22	AUUGAAAU	
89	2735104	2735134	UUUAGUUCUUCCGGCGCAUCGUUCUUG UAUUCAGGCAA	12	AUUGAAAU	
90	2735036	2735066	AGGCCAACAACUGAACUAGUUUUGUUUG UCAUUCUUAU	20	AUUGAAAU	
91	2734970	2734998	CCUUUGGAGAAUAUGCAAGCCCUGAUUU UGGCAGUA	53	AUUGAAAU	
92	2734904	2734932	UAGAUUCAACAUUCGAUUCUCCAAAGGC UUCGCUAA	73	AUUGAAAU	
93	2734837	2734866	UCCGCAAAAGAGAGAAUCAGAGCUGCCG AACUUUUGG	36	AUUGAAAU	
94	2734771	2734799	ACUAUAUCAAGUGUUCCAAAUCCAAUAU CAGAACUC	179	AUUGAAAU	
95	2734706	2734733	UGGCACAGAACAAAGCUAUAAUGGCAAA GUAUUCC	11	AUUGAAAU	
96	2734637	2734668	CUGUAUCUGAAUGCGUCUAAUUCAGGUA UAGCAUGUUCC	16	AUUGAAAU	
97	2734567	2734599	GAGAAUUAUUCUGGUUCAUCGGCUGCCA AAUAUAUUGUUG	132	AUUGAAAU	
98	2734502	2734529	AAGUCUUAUUAUAGACAACAAAAACACG GGUACCA	38	AUUGAAAU	
99	2734433	2734464	AUCAUUCAAGACCACUGCCUGAAUACAA GAACGAUGCGC	27	AUUGAAAU	
100	2734367	2734395	CCAAUUCAUAGUACCUAGCCGUAGAAGG AACAUAUA	153	AUUGAAAU	
101	2734301	2734329	CCGAACCAUUGACCAAUACUAUAUCCUA GAUUUUCA	49	AUUGAAAU	
102	2734234	2734263	AGCCAAGAAGAGGGGUCAAACCACUCAA AAUAGCCAU	119	AUUGAAAU	
103	2734167	2734196	UGUAAUAAAAAUAUGUUAAUAUAUAUUUA AGAGUAAU	775	AUUGAAAU	spacer with potential promoter sequence
104	2734099	2734129	UUGACAAUGAAAAUUACACUGUCAAAGA UGGAGAAGGA	41151	AUUGAAAU	drastic read # increase
105	2734033	2734061	AAAUAUACUCUAUGGUAUCUGCUGGACG CAAGUUGC	21552	AUUGAAAU	
106	2733967	2733995	AAGGUAUUGUCAUUGGCUGUUGCAUAUU UCAGUUUU	9585	AUUGAAAU	
107	2733899	2733929	UCGAUUGGCUCUUUUGUGAUUUGAUAUA UUUCUGCAGC	35845	AUUGAAAU	
108	2733831	2733861	CCGCGGUUCCUGCGGUCCAAUUUGCUU UCCCGACAAUA	2394	AUUGAAAU	
109	2733763	2733793	UAAUAAUAUUCUAUUUCUUCUUUUAUCC UGUUGAAUAU	10002	AUUGAAAU	
110	2733694	2733725	CAAUACCUCCUUUCAAUACCGGCGGUAU UUGUAAUCGCC	853	AUUGAAAU	
111	2733626	2733656	AUAAUUCAAUGUUACUGUUUAUUUUUUU AAAGUGCCAA	6202	AUUGAAAU	
112	2733558	2733588	UUAAAUAAUUCAAUUGUUCCGUUUAUUC CGCCCUUAAU	650	AUUGAAAU	
113	2733492	2733520	UCCUCAAUCUCUUUUUCGGUAAACAUCC ACCAUCAC	1514	AUUGAAAU	
114	2733426	2733454	UCAGGUGUAAGAGGUUUUACUUUCGCCC CUUGCGGC	2474	AUUGAAAU	
115	2733361	2733388	AGGUGGAAGGCUCCCUGAGGGGAGCAC CACCCUCC	792	AUUGAAAU	U to C change in

						5' tag
116	2733295	2733323	UGAUGUCGCGCCAGACAAAACGUAUGUA AUCAAAGA	4534	AUUGAAAC	
117	2733230	2733257	AAUUCUGCCAAGGCUCGCUCUCUGGCUU CUUGAAG	825	AUUGAAAC	
118	2733160	2733192	CUCAAGAACAGUUUGGUAAAUUGUUUGG UAUUGUUAAAUC	4582	AUUGAAAC	
119	2733094	2733122	CAAGGCUACCGAAUAUAUCAAUAAUGCC UAAAUCUU	600	AUUGAAAC	
120	2733025	2733056	UAUUCGUGAUGCUCCCAGCAAGCUGUG GGUCUCCGCCCA	1261	AUUGAAAC	
121	2732960	2732987	AAAUAAAACACGUUUCCAUCAAAUUCUCU UUUUUC	1292	AUUGAAAC	
122	2732893	2732922	ACAUAGUGAGUGCAUUGCCUUGGUUUG UUUUUUCUGC	4186	AUUGAAAC	
123	2732823	2732855	CACAAGCAGAGAGAGUGGGAGUACGUCA AGCAGUAGCGUG	1937	AUUGAAAC	
124	2732753	2732785	AGGCCGAACCAUAUGAUGUGAUUUCCAC CUUGUUGCAAAU	1328	AUUGAAAC	
125	2732687	2732715	UGAACUUUUUUCUUUUGUAAUAGAUUCC AUCAGUUA	628	AUUGAAAC	
126	2732620	2732649	CUUCAUCAAGUGCAUCAAAAGUAUCGCU UUUUCUUUC	819	AUUGAAAC	
127	2732551	2732582	UCUUCCAUCUAACCCCUCCCUAAUUCUA CAUAUUCUCCG	28	AUUGAAAC	
128	2732484	2732513	GUUUCAAAUUCACUGGACCAUUCACUGG ACCAGUAUA	1697	AUUGAAAC	
129	2732417	2732446	CUUGGAAGCGUAAUCGUUAGAAAAUCCU GUUCAUUCU	1919	AUUGAAAC	
130	2732348	2732379	ACAAUUCUGCAAAUUUUGAUUUACCGAG UAGCUUUUCUA	472	AUUGAAAC	
131	2732279	2732310	UACUCAACCGUAACUACUCUUUCAAAUU CUACUUCUCCA	37	AUUGAAAC	
132	2732212	2732241	CUGCUGACGAGGUGUUUUGAAGCCCAC CUAUUGCUAU	299	AUUGAAAC	
133	2732147	2732174	CCAAGCUGAAGGAUGGCAGGUACGUCAU CAUUGUU	267	AUUGAAAC	
134	2732081	2732109	CAGGCUGAGGAUGAGUAUAUAAAAUACA UUCUUAAA	258	AUUGAAAC	
135	2732013	2732043	AUAAUCCUUUCACCUUCAUAUCUCCCAU UCAAGUUCCU	10	AUUGAAAC	
136	2731944	2731975	GUUGUAUUCCUUUGAAUGAAAGAGCAAU GCUUAAAGAAA	1679	AUUGAAAC	
137	2731874	2731906	ACAAAAAUAUAGUUACCAUUGGCGAGUU UCUUUGCGACAC	65	AUUGAAAC	
138	2731807	2731836	CUCGGCACCGCUUGCCAUAUCACCAAGC AAAUUUGUU	22	AUUGAAAC	
139	2731739	2731769	ACAUCCACUGAGAAUCUCUUAGCGAAUA AAAAUGCAUC	249	AUUGAAAC	
140	2731672	2731701	UUGUUCAAAAUUGGGAUUCCAAGGCUAC GAUAUCGUG	572	AUUGAAAC	
141	2731605	2731634	CUUCUUUUUCGCUGUCGUAAUAUCCGAC AAUGAAAGA	87	AUUGAAAC	
142	2731538	2731567	ACUUAAUUGUUAACUUUGGUUGACGAAU AGGUUUACA	53	AUUGAAAC	
143	2731469	2731500	UUUGAGAAAUUCUGUUAUAUUAAAAUCA AAAUGUACAGC	181	AUUGAAAC	identical with spacer # 148
144	2731402	2731431	CUGAAGGGAGGGAGUGAUUUGAUAAAGA UAAACGAAU	28	AUUGAAAC	
145	2731333	2731364	AGAUUCAGGCGAAAAUCUGAACUUGAAG GUCAGAUAUGC	438	AUUGAAAC	
146	2731267	2731295	UCUUCCGUGUUGCCUGGGAGACGCCAG GAUCUGUCU	230	AUUGAAAC	
147	2731200	2731229	AUGAUGUAAAUAAUAAAAGUCUCGAUAAA CUUAGGAA	576	AUUGAAAC	
148	2731131	2731162	UUUGAGAAAUUCUGUUAUAUUAAAAUCA AAAUGUACAGC	167	AUUGAAAC	identical with spacer # 143
149	2731062	2731093	CCGUAUUGUAAACCCAAUUCUAAUAUUG AUAAGCCGCUA	86	AUUGAAAC	
150	2730995	2731024	AGGGCUUGGGUGCUAUAUCUGCAGAAG UAUCUGAUGC	436	AUUGAAAC	
151	2730924	2730957	AUGUUCCGGAUCAUGUUGUUGUCAUCAU CCGGGAAUUGGGC	222	AUUGAAAC	
152	2730856	2730886	UACCGAGCUUCUCAUGCAAUUCUCUUGC GCUUACGACU	122	AUUGAAAC	
153	2730790	2730818	AUCAUUUCCUGCAAAGCUAUGAACUGGU UCAUUCCG	91	AUUGAAAC	
154	2730722	2730752	UAGUAAGAACAAUAGUUCCGGAUGUAUC UGAAUAUAUC	375	AUUGAAAC	
155	2730653	2730684	AUCCUUAAAUUUGAGAUAUCUGAAUAAG AUAUCUUUUUC	60	AUUGAAAC	

156	2730587	2730615	UGCCCUGUGGAGCCGGGAAGACAGAAAC GGCAAUGG	62	AUUGAAAC	
157	2730521	2730549	CUAUGUACUCAAUCUUUCUAUCAGCUAU AGGUGUAA	121	AUUGAAAC	
158	2730451	2730483	UCUAUUGGAAUUCUGUUUGCAAUAUACA GUACCUUUAAGU	167	AUUGAAAC	
159	2730384	2730413	AAUCUAACUGUUGUUGUUUCUGUAUAAG UCAAUCCUU	144	AUUGAAAC	
160	2730318	2730346	AAACUCAAUUGCGUUAUGCCGUUUUCUU CUUUGUCA	184	AUUGAAAC	
161	2730250	2730280	AAAAUAAUAUAUCGAAUUUUCCCCUUAU UGGGUCCCAU	54	AUUGAAAC	
162	2730182	2730212	AUCCCUGGUUUUUGAGCCACCGUAUAGC AUUUGCAAUU	100	AUUGAAAC	
163	2730115	2730144	UGCAGAUCGUCCAUUGUCAAAAAGCGAG AAGGCGCAA	15	AUUGAAAC	
164	2730048	2730077	CUUCGCAAUAAAUGUAAUCUAUAUCGUA GUCGUCUUC	22	AUUGAAAC	
165	2729982	2730010	CUUUGGAAACCAAGGGUUUCGGCUGUUA UUCUCGCU	11	AUUGAAAC	
166	2729915	2729944	CACUUUAAUUCAAGUAAUAAUCUUAAUU GAUACUCAU	21	AUUGAAAC	
167	2729849	2729877	GAUGGUGCGGAUUUGAAGGAUGUGCCG UUCAAAUAC	0	AUUGAAAC	
168	2729781	2729811	UAAAUUUUCAGGCAACUGUUUUCAAAAG CUUUGAUAAA	0	AUUAAAAC	
169	2729714	2729743	CGUUUAUUCUUUCCCUCUUCGUCAACUU UCAAUAUCC	0	AUUGAACC	

C. thermocellum CRISPR 4
enome position: 3785203 - 3791022
80 units
GUUGAAGUGGUACUUCCAGUAAAA
CAAGGAUUGAAAC (37 bp)

spacer #	start position	end position	spacer sequence (sense direction)	# reads	5'tag	comments
1	3785240	3785275	AUGUAGAUGAAUGAUUACGAUGUUGGAG AAAUAUUU	56800	AUUGAAAC	
2	3785313	3785348	CUGUAUCCGCAUGUCCUACAGCAGAUGU CACAUCUG	3144	AUUGAAAC	
3	3785386	3785422	AUCUUUUGUAUAUCAAAGGAAGCUACUU CUGUAAUUA	64622	AUUGAAAC	
4	3785460	3785495	AGUGCUGCAAACAUCACAGCAGUAUUCA AUCAGGAA	43516	AUUGAAAC	
5	3785533	3785568	UCGCAUUUUCUACAUCAAACAAAGUUGC UCCGUCGU	5752	AUUGAAAC	
6	3785606	3785642	CCUCCAAGUUUCAGUCUCCAUGAAGUAC CAUUCCCCU	540	ACAAGGAU UGAAC	
7	3785680	3785716	GUGAAACUUGCCGGUUUUUAUGGCGUA UCUGUAGAUU	31097	AUUGAAAC	
8	3785754	3785789	UCAAAGUGCACAAAUAAAAAGUGCUUGU GUCAAUAC	22908	AUUGAAAC	
9	3785827	3785864	CACAAGUUGCAAAUGAUGUAUUGCCGCC GGUAAUUUCC	5477	AUUGAAAC	
10	3785902	3785936	GCCAUCAGUGGCCGAUAGAUAAGGCACA AUAUCUA	22137	AUUGAAAC	
11	3785974	3786009	GUUGCUAUUCGAGCAGAAAGAAACCCAA AAGUUAUC	27771	AUUGAAAC	
12	3786047	3786083	GUGAUAUUAUAUUGAAUCUGUUUAAAUG AGAUAUUCA	45648	AUUGAAAC	
13	3786121	3786157	CCACUGUCUUCUGUCAAUUGUUAUUAUA UCAUAUAUU	2905	AUUGAAAC	
14	3786195	3786229	AUGUAUGUUCAAAUACAUCAUAAAUAUG GUAAGGA	23146	AUUGAAAC	
15	3786267	3786305	UUGAUAAUGUUCUUUUCAAUAGUGCCAA GGUCUAAUCCU	2274	AUUGAAAC	
16	3786343	3786381	UCCCCGCCGUGUUAAUCCGAUUGUAAAC CUAUUCGUCAA	534	AUUGAAAC	
17	3786419	3786456	ACUAAUUUGAUUUAUAAAAAUGACUUAAU UACAAAUGA	4550	AUUGAAAC	
18	3786494	3786529	CUAUAGAUAUGUAUGGAGUACCGAAAGG CGACCCAU	18627	AUUGAAAC	
19	3786567	3786602	AAUACGGUGUCAGUGACGUUUUGAACCU	14523	AUUGAAAC	

			CAACCAAG			
20	3786640	3786674	ACUCUCCCAUGUUCAUUCCUCCUCUUUU CAAAUUU	368	AUUGAAAC	
21	3786712	3786747	AGAUAGACGUACAUAUGACAGGAGGUUG AAGUUAUU	5484	AUUGAAAC	
22	3786785	3786822	GAAGCUAUGCCUUAUGUGGGUGGACGC AGUUCUUCAGG	12913	AUUGAAAC	
23	3786860	3786897	CUAUUUUGUAAUAAUCACUUUUAAUUUC AUAACCUACA	1708	AUUGAAAC	
24	3786935	3786970	AAAUAGUACAUAGGUGGAGAGCAGCAAA CCCGCGAA	4685	AUUGAAAC	
25	3787008	3787044	UUUGCAACGUUUCCGUUGCUGUUAACUA CAUAGCAAG	3332	AUUGAAAC	
26	3787082	3787118	ACGAUGAGCUUAAAAACGAUUACUAAAG AGAGGAUGG	4869	AUUGAAAC	
27	3787156	3787191	AUUUUUGCUUCUCUAGCUUGUCUUUUCG UAUGUAUG	7242	AUUGAAAC	
28	3787229	3787264	ACUGCUUUAGCGUGAUCCUUGUGACCG UGCGAUACA	3394	AUUGAAAC	
29	3787302	3787337	UGGACUUUGCUCAGGUCAUUGGCCCGG AGAAACUGC	6830	AUUGAAAC	
30	3787375	3787409	UCAUUCUUGCAAGCCUGAAUCAGCCCUU UUAGAUA	1421	AUUGAAAC	
31	3787447	3787485	GGUUCUAUGAUUUCGAUAUACGACCCGC CUCUAUAUCCG	3131	AUUGAAAC	
32	3787523	3787558	GUUUUCUUAAUUUCAAAUACUCUUUUUG CCAUUUUU	17308	AUUGAAAC	
33	3787596	3787631	AAUCCCAGCGAAGUCCUUUAGAGAACAG AUUUUUGA	3571	AUUGAAAC	
34	3787669	3787706	UUAAUUUUAAAGACUUAUUCCGCAAUGA UAUAAAUAUG	2429	AUUGAAAC	
35	3787744	3787781	ACUUUGCCUAAGAAGAAUAAAAGUAUUC GUAGAAUAUU	11273	AUUGAAAC	
36	3787819	3787856	CUAACAUUAUUCACACCUCCCAUCAACG AACCUUUUUU	629	AUUGAAAC	
37	3787894	3787929	ACUUAGCCGUUGCAGUGGGCUUACCUU CUAAAAAUU	6222	AUUGAAAC	
38	3787967	3788002	ACUUUCCUUGAUAUGUUUACCUGUAAAG UAAAAAUA	4200	AUUGAAAC	
39	3788040	3788075	CUUAUAUAUAGAACAUAAAUUUUUCUAUA AUAUUAA	2283	AUUGAAAC	
40	3788113	3788147	CCCUUACAGUGGCGAUAUUUGUGGAAUG AGGCUGU	6055	AUUGAAAC	
41	3788185	3788221	AUCUCUUGUAGAAUUUUCUCUGCCUUUU CUAAGCUAC	7246	AUUGAAAC	
42	3788259	3788294	ACAGUUAUCAUAGGCUAUCCUCCUUAUA AAUGUUAA	2425	AUUGAAAC	
43	3788332	3788368	AAUAAUCCAAAAGCAUUUCUACUGCUUU UGGAAAUUU	2110	AUUGAAAC	
44	3788406	3788442	AUAUUUACUCCGGCAUCCCUCAGCCGGC GAAAAAGGU	1209	AUUGAAAC	
45	3788480	3788514	AUAAUUCUUCUCGUUUUUGUUUUAAUUC GUUUAUU	736	AUUGAAAC	
46	3788552	3788587	GUAAUAAGCUGUCCAAUCACCUCCUAGU UUUCCUGG	2758	AUUGAAAC	
47	3788625	3788662	UAGAACUUGAAAGAGAAUCCUUCAGGGA GAGUGGUACU	8239	AUUGAAAC	
48	3788700	3788736	UAUCUUCAAAGUAAUAGCUGUUAUAUGU GUCAAGUUC	9604	AUUGAAAC	
49	3788774	3788809	UUAAUUCAUCGCCAUUUUUGCCUAUAAG UCCAUAAA	953	AUUGAAAC	
50	3788847	3788882	CUUCUUCGAAGCUACUAAAGGUGGCCGU GCAGUAGU	4203	AUUGAAAC	
51	3788920	3788955	AUAGAUAAGGCUAUUGCAAGAGUAGCGA AUGUGCUA	12591	AUUGAAAC	
52	3788993	3789030	CAAAUAUUUCAGAAGCUAUUGCGGAUAU AUUUACUGUG	15333	AUUGAAAC	
53	3789068	3789106	UUGAAUUCCUGCAAUGUGGCAGGCUUUU CUCAACCUAGG	4407	AUUGAAAC	
54	3789144	3789181	GUCCCUGUGAUUUCAGCCAAUUUAUCGC AUUUUCAAUU	4918	AUUGAAAC	
55	3789219	3789257	ACAUACAGUUCGGUGUUUUCAUCUAGUG CUACAUUUAUU	9792	AUUGAAAC	
56	3789295	3789330	CCAGCAGGUUGAGCGUUAGGCAGGUUU CCCUGCCCG	8386	AUUGAAAC	
57	3789368	3789402	GUGGAAGAUGCGUUGUUUAUGUCAGCAA UUUCCGG	14203	AUUGAAAC	
58	3789440	3789475	CAAAAUGGCAGGUGGUAAAACCACAGCU AAUACAUC	2652	AUUGAAAC	
59	3789513	3789548	GUUUCAUCAACAUCCAUCUCGUCUACAU AGAACUCA	1719	AUUGAAAC	
60	3789586	3789622	AUAACAUAUUUUAAAGCAAAAAGCAGAAA AUCACUGG	2138	AUUGAAAC	
61	3789660	3789694	UGAUUGUAAAGGAAUAUGUAUCCCCAAG CCUUACA	4488	AUUGAAAC	

62	3789732	3789769	UGGAUAGUGCAAACAUAGAUGAAGCCUC CGUAGGCCUU	2940	AUUGAAAC	
63	3789807	3789843	UCUAUUUCAGCAACAAUAUAUUUGGCAG CCGAUGAAU	1431	AUUGAAAC	
64	3789881	3789917	AUCAUUGUAUAUAUCUUUGGAGUAACGU AUAAUACCC	7666	AUUGAAAC	
65	3789955	3789991	UUGAUAAUGUUCUUUUCAAUAGUGCCAA GGUCUAAUC	2434	AUUGAAAC	
66	3790029	3790065	ACAGGAACGUGCGUUCUCUGGACGAGG UAACUAUCGU	10260	AUUGAAAC	
67	3790103	3790137	UAGCACACAACAGCAAAAAAACAAAGAAA AAAAGA	20001	AUUGAAAC	
68	3790175	3790210	GCACCAACUUGUCUUUUUUGAUGAGGCU UGCAACAA	3007	AUUGAAAC	
69	3790248	3790286	AUAAUGUGCUGCUGGAUGCCCGUACAGU CUUCCACUGUA	980	AUUGAAAC	
70	3790324	3790362	UAGAUUGCGAAAAGCCUGCCACAUUGCA GGUCACAAGGU	5804	AUUGAAAC	
71	3790400	3790434	UUGUUUUUUCACUUAUAGACUUAAUUCA GUGUGGC	4535	AUUGAAAC	
72	3790472	3790512	CUUUCUCAAUGGCUGACUUCGAUGCGCC CUGGGCAAGCAUU	807	AUUGAAAC	
73	3790550	3790585	GUGCUGCUAUCUCGGCUAUUCUUUCAUC UAUACUCA	1911	AUUGAAAC	
74	3790623	3790657	UUUGCUCUUUAUCCGGUGCUCUUAUCUC AUUCACA	209	AUUGAAAC	
75	3790695	3790730	ACGAGAACGGGAAAGCAAUAUUGAGAAA ACCCUUUG	2543	AUUGAAAC	
76	3790768	3790802	UUGGCGGCAUAGCUUCUAUUAUAGAAAA UUCUACU	2304	AUUGAAAC	
77	3790840	3790875	UUAUAUCGAAGAAAAAAGAACAAAAGUGA GGCGAUU	2043	AUUGAAAC	
78	3790913	3790951	AUAAUUUCGGGAGUGUAGAAUUUCAACU CGGCAAUAAGA	5674	AUUGAAAC	
79	3790950	3790987	GAAUGAUUGAAUCAAUAACCAUUAUUGC CUGCAAUUA	98	AUUGAAAC	

C. thermocellum CRISPR 5
genome position 3813209 - 3816348
43 units
GUUGAAGAGGUACUUCCAGUAAAA
CAAGGAUUGAAAC (37 bp)

spacer #	start position	end position	spacer sequence (sense direction)	# reads	5'tag	comments
1	3813246	3813281	CUUUUUUUAUCUCGUUAUCUUCACCGUC UGUGUAUC	4791	AUUGAAAC	
2	3813319	3813355	AUUUCUCAUAUUCAAUUUUUACCUUUUC CAUUUUUUA	1136	AUUGAAAC	
3	3813393	3813432	AUCUUGGGGUGCAAGAGGUCGCUGGUU CAAAUCCAGUCAC	13100	AUUGAAAC	
4	3813470	3813506	AUGGGCAUGCUCAUAUGGUGUAUCCCUC AACGCUAUA	11003	AUUGAAAC	
5	3813544	3813580	CUAUCUGUUGCGAUUCUCCGCCACUGG GAUUAUCACC	454	AUUGAAAC	
6	3813618	3813654	AUUCACAAAAUCAGCACUCAAGAUGGGA AGUGACUUG	4750	AUUGAAAC	
7	3813692	3813727	AUCAUACUACCCAUAAAACCUUCUAAUU GCUUUAGC	3000	AUUGAAAC	
8	3813765	3813801	UUUUCCUUACAAUUGUUAAACAUCCAAA UAUCGUCAU	9504	AUUGAAAC	
9	3813839	3813874	CGCACAUUUCUAUCCCUAUACUGUUGCG GUUGGUAA	7097	AUUGAAAC	
10	3813912	3813949	UAUGUAUCAUAAUGCACGUGGCACUAAC UGUAAGACUU	4800	AUUGAAAC	
11	3813987	3814022	ACGACAACAUGAGAAGAAAAGGGGAUGA AGCACAGA	1111	AUUGAAAC	
12	3814060	3814095	UCUUCUCUUAAAGCACUACCAGUAUCUU CUACUAUU	651	AUUGAAAC	
13	3814133	3814168	UUGUAUGUAUUUUAUUAAUAUAGUGUGC UUGUAAUU	4003	AUUGAAAC	
14	3814206	3814242	GGAUUUGUCAAGUCUGCUAAAAUAGCGG GCGGCAAAA	17648	AUUGAAAC	
15	3814280	3814315	CCUUCUCCCCAUUUACCACCGGGCAAAU UAACUGUA	142	AUUGAAAC	
16	3814353	3814388	GUUAAUCCCAAAUUAUGCUCAAAAGAAAA AAACCAA	3147	AUUGAAAC	
17	3814426	3814462	AAUAUUGGAAUCCAAAAUAACCUUCAAUU UUCCACCA	4212	AUUGAAAC	

18	3814500	3814534	GUGACGUUUUGGCGUCUAUUGAAAAUCU AGGAUAU	7713	AUUGAAAC	
19	3814572	3814607	UUUCACAUGCAUAAUAUGUGAAAUAGCG ACCGAACA	5548	AUUGAAAC	
20	3814645	3814679	AACCUCCUUUUUAUAUUUCUGGAUCAUC AUCAUAA	1634	AUUGAAAC	
21	3814717	3814752	CAGUAUUGCAUUGUUAUUCCUCGCUUUC UUUUUGUU	1862	AUUGAAAC	
22	3814790	3814825	AUACACUUCAAGAGAAUAAUGUCCAGUU CGUGUCGG	2215	AUUGAAAC	
23	3814863	3814899	AUGGUGAUGAAGUGAGGUAUGAAGCUG GAAGCGGAGA	753	AUUGAAAC	
24	3814937	3814972	UUUUAUAGUAGAUUAAGUUUUCUGCGUU AAAAUAAU	3531	AUUGAAAC	
25	3815010	3815047	AUUGUCUUGAACUCAGGGUCUAUAUACC GCACCAAGCA	3584	AUUGAAAC	
26	3815085	3815122	CUUCGGAUAAUCCGUCUUCUCGUUUAAU UCUUAAGCCA	1117	AUUGAAAC	
27	3815160	3815196	ACAUAUACGAAGGUUUAAACAGGACAGC AGAGCAAAG	1491	AUUGAAAC	
28	3815234	3815271	GACUACAUUCUUUUAAUUUUUAAAAAAG AAAUAUUAAA	1009	AUUGAAAC	
29	3815309	3815347	AAUCUUCUGCGAGCAGAAGGAGUUCGUU UCUGUUGCAAC	12013	AUUGAAAC	
30	3815385	3815423	CCCAAAGAAAGCGCAUGAUAGGGUCUUU AGGUGGCUUCG	6504	AUUGAAAC	
31	3815461	3815495	AUAUUACAGUAACAACUAAAGCAGGACA UAAAAGU	1324	AUUGAAAC	
32	3815533	3815569	UACAUGAUAUUACCUCCUAUUUUUAUUU UUCUUCCAG	541	AUUGAAAC	
33	3815607	3815644	AGCAUCCCUACUUAAAGAGACUGGGGGC AUAUAACCCC	3765	AUUGAAAC	
34	3815682	3815722	UUCAAUCGUGGCCUCCAGGGAACGGGC GUAAAACAGUUCUU	2003	AUUGAAAC	
35	3815760	3815796	AUAGAAGGGGCUUAAGCCCUUUUUAUUC UGACCUGAA	2955	AUUGAAAC	
36	3815834	3815870	AGGAAGCAGAGGAUUACGUACGUCAGAA AACAGACUU	2279	AUUGAAAC	
37	3815908	3815942	UUUUGUCAAGCCAUUGAUACAUCCUUUA AUGUCUA	979	AUUGAAAC	
38	3815980	3816017	GUAUAGCACAGGCCCAAGUAAAAGAAAG GAGUUCAUCA	671	AUUGAAAC	
39	3816055	3816090	CUACAGCUAUAGCUUGCUUGAAUACAAA CUCAACGA	981	AUUGAAAC	
40	3816128	3816163	GUCCCUCCUUAUGCUUCAACAUUGAUUG CUACGCCU	305	AUUGAAAC	
41	3816201	3816238	AACCUUACUAUGAUGUUAUAACAUACAAA GGUUGUAAA	4603	AUUGAAAC	
42	3816276	3816311	CUCAAACUCAGUGAUACUAAGUAUUUUA AAAGAAAA	1090	AUUGAAAC	

Table S2 RNA substrates used for endonuclease cleavage assays. The repeat sequences are underlined and brackets indicate deoxy nucleotide substitutions.

RNA substrate	**Sequence 5' – 3'**
M. maripaludis repeat	CUAAAAGAAUAACUUGCAAAAUAACAAGCAUUGAAAC
M. maripaludis d-1 repeat	CUAAAAGAAUAACUUGCAAAAUAACAAG (dC) AUUGAAAC
C. thermocellum repeat	GUUUUUAUCGUACCUAUGAGGAAUUGAAAC
C. thermocellum d-1 repeat	GUUUUUAUCGUACCUAUGAGG (dA) AUUGAAAC
repeat - spacer27 - repeat	GAUUUCCGGCUAAAAGAAUAACUUGCAAAAUAACAAGCAUUGAA ACCUGAUGAAACAAGCGAAACAAACAAUAAAAAAUCAAGCUAAAA GAAUAACUUGCAAAAUAACAAGCAUUGAAACCGGGGCAAAGA
spacer2 - repeat - spacer3	AUUAAUCCCAUAAUACUUUUCUAGGUCUGGGCGGAAUCUAAAAG AAUAACUUGCAAAAUAACAAGCAUUGAAACUAAAAAAAGAAAAAA GUUAAAAAAGCUAGAAUAAA

Chapter III

Comparative analysis of Cas6b processing and CRISPR RNA stability

Hagen Richter,[1] Sita J. Lange,[3] Rolf Backofen[3,4,5,6] and Lennart Randau[1, 2,*]

[1] Max-Planck-Institute for Terrestrial Microbiology, Karl-von-Frisch Straße 10, 35043 Marburg
[2] LOEWE Center for Synthetic Microbiology (Synmikro), 35043 Marburg
[3] Institut für Informatik, Albert-Ludwigs-Universität, Georges-Koehler-Allee, Geb 106, 79110 Freibug
[4] Centre for Biological Systems Analysis (ZBSA), Albert-Ludwigs-Universität, 79110 Freibug
[5] Centre for Biological Singalling Studies (BIOSS), Cluster of Excellence, Albert-Ludwigs-Universität, 79110 Freibug
[6] Center for non-coding RNA in Technology and Health, University of Copenhagen, Frederiksberg C, Denmark
[*] Corresponding author

Abstract

The prokaryotic antiviral defense systems CRISPR (clustered regularly interspaced short palindromic repeats)/Cas (CRISPR-associated) employs short crRNAs (CRISPR RNAs) to target invading viral nucleic acids. A short spacer sequence of these crRNAs can be derived from a viral genome and recognizes a reoccurring attack of a virus via base complementarity. We analyzed the effect of spacer sequences on the maturation of crRNAs of the subtype I-B *Methanococcus maripaludis* C5 CRISPR cluster. The responsible endonuclease, termed Cas6b, bound non-hydrolyzable repeat RNA as a dimer and mature crRNA as a monomer. Comparative analysis of Cas6b processing of individual spacer - repeat - spacer RNA substrates and crRNA stability revealed the potential influence of spacer sequence and length on these parameters. Correlation of these observations with the variable abundance of crRNAs visualized by deep-sequencing analyses is discussed. Finally, insertion of spacer and repeat sequences with archaeal poly-T termination signals is suggested to be prevented in archaeal CRISPR/Cas systems.

Introduction

The adaptive immune system CRISPR/Cas consists of an array of clustered, regularly interspaced, short palindromic repeats (CRISPR) and a set of CRISPR-associated (Cas) proteins. Spacer sequences that are present between the repeats can be derived from the genomes of mobile genetic elements (e.g., viruses) and are utilized to protect host organisms against recurring attacks from these elements. The activity of the CRISPR/Cas system within the cell is divided into three phases: (1) the adaptation of new spacers, (2) the maturation of small CRISPR RNAs (crRNAs) that contain a single spacer sequence and (3) the interference with foreign nucleic acids using crRNAs bound to a Cas protein interference complex (Barrangou et al, 2007; Barrangou & Horvath, 2011; Bolotin et al, 2005; Cui et al, 2008; Horvath & Barrangou, 2010; Koonin & Makarova, 2009; Plagens et al, 2012; Sorek et al, 2008; Terns & Terns, 2011; van der Oost et al, 2009).

The diverse CRISPR/Cas systems are divided into three major types and at least 10 subtypes (Makarova et al, 2011b). In Clostridia, methanogens and halophiles, the subtype I-B is present, which is defined by the subtype-specific protein Cas8b. *Methanococcus maripaludis* C5 possesses a single subtype I-B CRISPR/Cas system with a minimal Cas protein set composed of the universal proteins Cas1, Cas2 and Cas4 (proposed to mediate spacer acquisition), as well as Cas3, Cas5, Cas7 and Cas8b, which are presumed to form the interference complex Cascade (Cas complex for antiviral defense). Cas6b completes the set of Cas proteins and was shown to process pre-crRNA in *M. maripaludis* (Richter et al,

2012). Cas6b belongs to a diversified family of Cas6 endonucleases that is responsible for crRNA processing. The first characterized Cas6 enzyme belongs to the subtype III-B system present in *Pyrococcus furiosus* (Pf Cas6). Pf Cas6 was described as a metal-independent endonuclease involved in the processing of pre-crRNAs into mature crRNAs (Carte et al, 2010; Carte et al, 2008; Wang et al, 2011; Wang et al, 2012). Different Cas6 enzymes were also characterized for subtype I-F (Cas6f, also known as Csy4 in *Pseudomonas* aerouginosa (Haurwitz et al, 2010; Haurwitz et al, 2012; Sternberg et al, 2012)*)* and subtype I-E (Cas6e or Cse3 in *Escherichia coli* and *Thermus* thermophilus (Gesner et al, 2011; Sashital et al, 2011)). For the subtype I-C that lacks a Cas6 protein, Cas5d appears to compensate for the functions of both Cas6e and Cas5e (Garside et al, 2012; Nam et al, 2012). Investigated Cas6 enzymes share a similar structure despite their low sequence similarities, feature a ferredoxin-like fold and operate in an analogous manner to create mature crRNA molecules. Nevertheless, significant differences in their catalytic site composition have been observed. In Pf Cas6, a catalytic triad formed by tyrosine, histidine and lysine residues can be found (Carte et al, 2010), while Cas6e utilizes a catalytic dyad of tyrosine and histidine residues (Haurwitz et al, 2012). Two adjacent conserved histidine residues play a major role for crRNA processing by Cas6b (Richter et al, 2012). Despite these differences in their active sites, all investigated Cas6 proteins generate crRNAs with a 5'-terminal 8 nt repeat tag (Carte et al, 2008; Gesner et al, 2011; Haurwitz et al, 2010; Richter et al, 2012; Sashital et al, 2011), whereas Cas5d processing yields an 11 nt 5'-terminal tag (Nam et al, 2012). The 3' end of a crRNA contains the remaining repeat nucleotides, but is often subsequently trimmed by a currently unknown mechanism (Hale et al, 2008; Hale et al, 2009; Richter et al, 2012). It was shown that Cas6 enzymes bind their respective substrates as monomers (Carte et al, 2010; Gesner et al, 2011; Haurwitz et al, 2010; Sashital et al, 2011; Sternberg et al, 2012; Wang et al, 2011) and are proposed to deliver the mature crRNA to the Cascade complex (Carte et al, 2010; Makarova et al, 2011b). However, a non-cleaving homolog of Pf Cas6 was observed to form dimers upon binding of unspecific RNA substrates (Wang et al, 2012). Monomeric Cas6 activity is in agreement with a proposed wrap-around mechanism for Pf Cas6, in which Cas6 was suggested to be bound to the pre-crRNA in a bead chain-like manner using the spacer sequences as guide elements for repeat binding (Wang et al, 2011). The proposed wrap-around mechanism correlates with the regularly interspaced arrangement of a pre-crRNA as determined by the CRISPR array. The repeats of a CRISPR are nearly uniform for one particular array and the spacer sequences are most often unique (Barrangou et al, 2007; Horvath & Barrangou, 2010; Koonin & Makarova, 2009; Sorek et al, 2008; Terns & Terns, 2011; van der Oost et al, 2009). The length of both spacers (26 - 72 nt) and repeats (24 - 47 nt) varies between different CRISPR systems (Grissa et al, 2007a; Sorek et al, 2008) and the arrays appear to follow restrictions regarding the preferred and

allowed lengths of repeat and spacer sequences. Spacer acquisition studies in *E. coli* indicated a favored length of 32 - 33 nt for newly integrated spacers, which is consistent with the length of previously existing spacers of the particular array (Yosef et al, 2012). A Gaussian-like distribution of spacer length was found for the three CRISPR loci in *Streptococcus thermophilus* (Horvath et al, 2008) also showing a particular range of spacer lengths that is favored. In *M. maripaludis,* the spacer length varies between 34 - 40 nt (Table S1**)**. Apart from their length, repeat sequences can differ in proposed secondary structures, which have been shown to be important for processing of some repeat sequences (Gesner et al, 2011; Haurwitz et al, 2010; Nam et al, 2012; Sashital et al, 2011). In other cases, the palindromic nature of repeats giving rise to small hairpin structures was shown to not be necessary for crRNA maturation (Carte et al, 2010; Carte et al, 2008; Wang et al, 2011). The position of the 5' processing site within the repeat was shown to influence the length of mature crRNA, as a ruler mechanism is applied to ensure a consistent length (Hatoum-Aslan et al, 2011). Based on their length, sequence and potential hairpin formation, 11 different repeat clusters were classified (Kunin et al, 2007). Deep sequencing analyses identified highly variable crRNA abundance patterns for different organisms. A general observed trend, e.g., for both CRISPR loci of *N.* equitans (Randau, 2012) and for the CRISPR locus of *M. maripaludis* (Table S1)(Richter et al, 2012), is a gradual decline in abundance of crRNAs from the first (closest to the promoter) to the last spacer. Newly acquired spacers are inserted closest to the promoter and this abundance pattern infers that the most recent interactions with viruses cause a Cascade-targeting system that might be most effective against these recently spotted viruses. However, it is puzzling why some crRNAs do not follow this trend and are found to be severely underrepresented *in vivo*. The analysis of the abundance of crRNAs bound to CMR complexes in *Sulfolobus solfataricus* identified a variable distribution of crRNAs that are loaded into the interfering complex (Zhang et al, 2012). Similar results were obtained by RNA sequencing of a total RNA isolation of *S. solfataricus,* where the crRNA abundances were highly variable (Deng et al, 2012; Wurtzel et al, 2010). Both repeat and spacer sequences could affect crRNA abundance in the cell. To assess the variable crRNA abundance pattern of the single CRISPR system of *M. maripaludis,* we generated 26 spacer - repeat - spacer substrates and analyzed the influence of these spacer sequences on Cas6b cleavage activity *in vitro*. Possible influences of crRNA stability were monitored via in-line probing of crRNAs of *M. maripaludis* in the presence and absence of Cas6b. In addition, we observed dimerization of Cas6b upon binding of non-cleavable substrates while mature crRNA is bound by Cas6b monomers.

Material and Methods

Production of Cas6b

Production and purification of recombinant Cas6b via Ni-NTA chromatography was performed as described earlier (Richter et al, 2012). Oligomeric states of Cas6b were determined by size-exclusion chromatography. Ni-NTA purified Cas6b in lysis buffer [10 mM TRIS-HCl (pH 8.0); 300 mM NaCl; 10 % glycerol; 0.5 mM DTT] or Cas6b incubated with repeat RNA harboring a 2'-deoxy modification (see below) was applied to an analytical Superdex column (200 10/300 GL, GE-Healthcare) and fractionated using a FPLC Äkta-Purification system (GE-Healthcare). Proteins of collected fractions were precipitated by addition of one quarter volume 100 % TCA (trichloracetic acid) and 15 min incubation on ice. The precipitated proteins were washed twice with 200 µl acetone (100 %, ice-cold) and dried at 95 °C for 5 min before resuspension in SDS loading buffer. Molecular weight of the fractions was determined by comparison to a calibration of the column with molecular weight markers, as described in the manufacturer manual (Kit for Molecular Weights, Sigma-Aldrich).

Generation of RNA substrates

PCR reactions with genomic DNA, isolated from *Methanococcus maripaludis* C5 and containing forward primers encoding the T7 RNA polymerase (RNAP) promoter, yielded spacer(n) - repeat - spacer(n+1) templates for *in vitro* run-off transcription (Fig. 2**)**. Transcriptions for unlabeled RNAs were performed for 1 h at 37 °C in a final volume of 20 µl [40 mM HEPES-KOH (pH 8.0); 22 mM $MgCl_2$; 5 mM DTT; 1 mM spermidine; 4 mM UTP, CTP, GTP, ATP; 20 U RNase inhibitor, 1 µg T7 RNAP, 1 µg PCR product]. Radiolabeled spacer - repeat - spacer substrates were generated by *in vitro* run-off transcription reducing the amount of ATP to 2 mM and adding 25 µCi (α-^{32}P) adenosine triphosphate (ATP) (5,000 Ci/mmol, Hartman Analytic). For the production of crRNA substrates, unlabeled spacer-repeat-spacer RNAs were processed with 15 µM Cas6b and purified by phenol-chloroform extraction. The obtained crRNAs were 5'-end labeled using (γ-^{32}P) ATP (5,000 Ci/mmol) and T4 polynucleotide kinase (PNK) in a volume of 20 µl [15 µl of purified RNA; 2 µl PNK buffer (New England Biolabs) and 25 U T4 PNK (Ambion)] for 1 h at 37 °C. Synthesis of repeat RNA with a deoxy substitution of the first unprocessed nucleotide was done by Eurofins MWG Operon. 5'-end labeling of the deoxy-modified repeat RNA was achieved as mentioned above. Labeled RNAs were separated by denaturing PAGE (8 M urea; 1 × TBE; 10% polyacrylamide) and visualized by autoradiography. Identified bands were excised from the gels and RNA species were eluted by overnight incubation on ice using 500 µl RNA elution buffer [250 mM NaOAc; 20 mM TRIS-HCl (pH 7.5); 1.5 mM

ehtylenediaminetetraacetic acid (EDTA) (pH 8.0); 0.25 % SDS]. After addition of two volumes, EtOH (100 %, ice cold) and 1/100 glycogen (Roche) the RNAs were precipitated for 1 h at -20 °C. Subsequent washing with 70 % EtOH was followed by pelleting of the RNA.

Endonuclease assay

In endonucleolytic cleavage assays the indicated Cas6b concentrations were incubated for 10 min at 37 °C together with the RNA substrates in cleavage buffer [250 mM KCl; 1.875 mM $MgCl_2$; 1 mM DTT; 20 mM HEPES- KOH (pH 8.0)]. The reaction was quenched by adding 2 x formamide buffer [95 % formamide; 5 mM EDTA (pH 8.0); 2.5 mg bromophenol blue; 2.5 mg xylene cyanol] and incubation at 95 °C for 5 min. Separation of RNA species was achieved by a denaturing 12 - 20 % polyacrylamide gel running at 12 W. Visualization of RNA was done by autoradiography. In order to obtain full processing of precursor RNA to generate mature crRNAs suitable for 5' labeling with (γ-^{32}P) ATP, the reaction time was increased to 1 h with 15 µM Cas6b.

Electrophoretic mobility shift assay (EMSA)

In EMSA reactions, indicated Cas6b concentrations were mixed with radiolabeled RNA substrates in binding buffer [10 mM TRIS-HCl (pH 8.0); 200 mM KCl; 5 % Glycerol; 0.5 mM DTT; 0.5 mM EDTA (pH 8.0); 1 µg BSA]. The reaction mix was incubated for 1 h at 37 °C and immediately mixed with 6 x DNA loading dye (4 g sucrose; 25 mg bromophenol blue; 25 mg xylene cyanol in 10 ml H_2O). Separation of the reaction was done with a 7 % native polyacrylamide gel running in 1 x TBE at 8 W for 2.5 h. Visualization was achieved by autoradiography.

In-line probing

Structural analysis of crRNAs was analyzed via in-line probing (Soukup & Breaker, 1999) incubating crRNA in probing buffer [5 mM TRIS-HCl (pH 8.5), 0.5 µg yeast RNA (Ambion), 100 mM KCl, 20 mM $MgCl_2$] for 16 h at room temperature. To test Cas6b influence on crRNA stability, 20 µM Cas6b was added to the probing reaction. The reaction was stopped by addition of 2 x formamide buffer and incubation at 95 °C for 5 min. Visualization was achieved by phosphorimaging after separation of RNA using 20 % denaturing polyacrylamide gels.

Results

Cas6b forms dimers upon binding to non-cleavable repeat substrates

Cas6b was previously identified to be the endonuclease responsible for the cleavage of repeat sequences of pre-crRNA molecules and for the generation of the crRNAs' 5'-terminal 8 nt tag (Richter et al, 2012). In this work, we analyzed the binding of the enzyme to a native repeat substrate. To prevent endonucleolytic cleavage of this repeat sequence during the performed binding assays, a non-cleavable variant of the repeat served as a substrate. This variant harbors a 2'-deoxy modification at the position of cleavage (Fig. 1A), which was shown to abolish nuclease activity (Haurwitz et al, 2010; Richter et al, 2012). Electrophoretic mobility shift assays (EMSA) were performed with varying concentrations of purified recombinant Cas6b. The EMSA assays revealed a band shift at low Cas6b concentrations that is fully converted to a distinct super shifted band with increasing Cas6b concentrations (Fig. 1B). This pattern of Cas6b repeat RNA binding is in accordance with the possible dimerization of the enzyme. Fractionation of a Cas6b-repeat complex via size-exclusion chromatography also supports Cas6b dimerization (Fig. S1). To prove that the observed Cas6b binding was specific, competition assays were performed. The addition of increasing concentrations of unlabeled 2'-deoxy-modified repeat RNA appears to preferentially diminish the occurrence of the super shift (Fig. 1C) and the total amount of unbound labeled RNA increased. The addition of up to 10 µg unlabeled yeast RNA did not influence the pattern of RNA shifts (Fig. 1C). Finally, we assayed the binding of a mature crRNA and observed a single band shift (Fig. 1D) that suggests the formation of a Cas6b monomer-crRNA complex after successful cleavage.

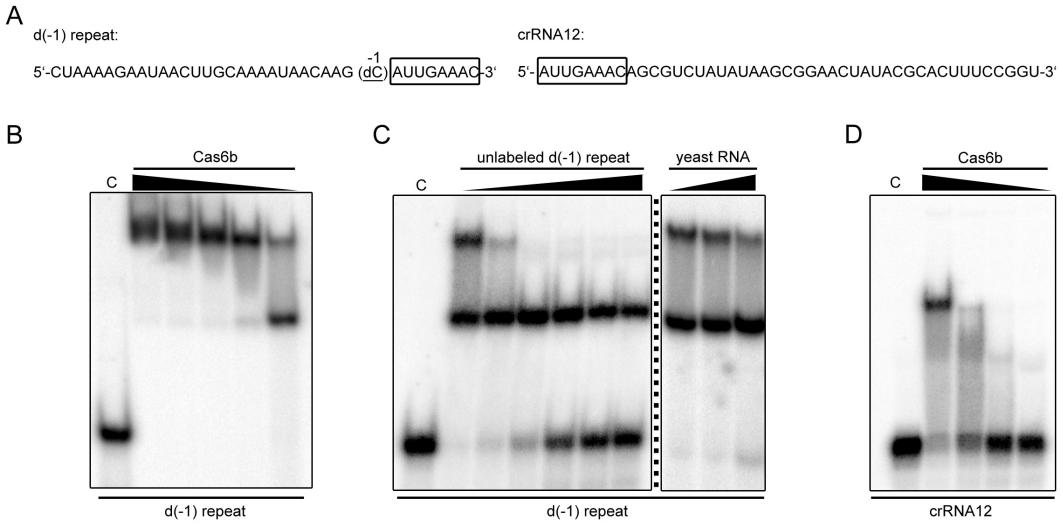


Figure 1. Cas6b-binding assays with non-cleavable native repeat RNA and mature crRNAs. (**A**) The employed RNA sequences are indicated. To prevent processing of the RNA repeat substrate, a deoxy substitution of the first base upstream of the 5' tag (boxed) was used in the binding assays. (**B**) Binding assays using decreasing concentrations of Cas6b (50 - 10 µM) and equal amounts of deoxy repeat RNA resulted in shift and super shift formation. The separation of samples was performed in native polyacrylamide gels (7 %). Bands were visualized by phosphorimaging. (**C**) Competition assays with unchanged Cas6b concentration (20 µM) and increasing concentrations of unlabeled deoxy repeat RNA (0 - 5 µM) and yeast RNA (10 - 1 µg) show specificity of Cas6b binding. (**D**) Only a single shift is observed for Cas6b binding (30 - 5 µM) with 5'-labeled crRNA substrate.

Establishment of an experimental approach to assess crRNA processing in *M. maripaludis*

To assess the previously reported highly variable abundance pattern of crRNAs isolated from *M. maripaludis* (Richter et al, 2012), we specifically designed an experimental setup to individually assay the influence of the two neighboring spacer sequences on Cas6b repeat RNA processing and crRNA stability *in vitro* (Fig. 2). Individual primer pairs were designed to amplify spacer-repeat-spacer sequences from *M. maripaludis* genomic DNA. One primer contained the T7 RNA polymerase promoter sequence. PCR products were generated for each spacer(n) - repeat - spacer(n+1) combination that were subsequently used as DNA templates for T7 RNA polymerase mediated *in vitro* run-off transcription. This approach allowed us to produce all individual RNA substrates. The addition of (α-^{32}P)-ATP in the transcription mixture yielded radioactively labeled RNA substrates that were used for Cas6b endonuclease assays. In an alternative approach, we generated unlabeled RNA molecules, processed these with an excess amount of Cas6b to obtain mature crRNA, which then were radioactively labeled at their 5'-termini and used in in-line probing assays (Soukup & Breaker, 1999) for structural analyses (Fig. 4). Even though the cleavage of spacer - repeat - spacer substrates yields two fragments, the T4 polynucleotide kinase labeling reaction strongly favored the crRNA product, which demonstrates the differences in the 5'-terminal phosphorylation state (5'-ppp vs. 5'-OH) of the two fragments.

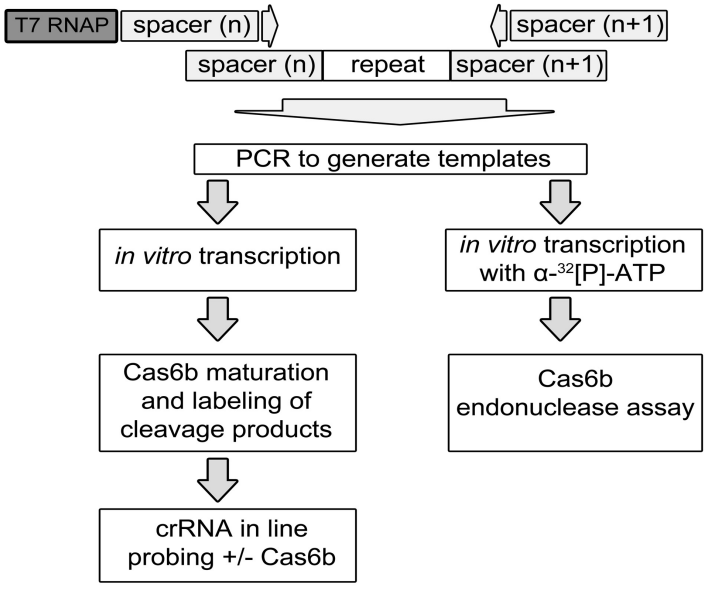


Figure 2. Experimental setup to assess the influence of adjacent spacers on repeat RNA processing by Cas6b. Individual primers for each spacer(n) - repeat - spacer(n+1) combination were designed and each forward primer contained a T7 RNA polymerase promoter sequence. The primers were used in PCR reactions with genomic DNA to generate individual spacer - repeat - spacer templates for *in vitro* run-off transcription. Adding (α-^{32}p)-ATP to the reaction yields radioactively labeled RNAs that were used for Cas6b endonuclease assays, whereas unlabeled RNAs were treated with Cas6b to generate mature crRNAs, which subsequently were labeled at the 5' tag and used for in-line probing reactions to evaluate crRNA stability.

Spacer sequences influence crRNA maturation *in vitro*

Three independent Cas6b endonuclease assays were performed with 26 consecutive spacer(n) - repeat - spacer(n+1) RNA substrates of the *M. maripaludis* CRISPR locus (spanning spacer 1 - 26) to screen for the influences of flanking spacer pairs on the efficiency of repeat RNA processing. Each RNA species was incubated with three different concentrations of Cas6b (10 µM, 1 µM and 0.1 µM) to evaluate and compare the cleavage efficiency for each substrate. All substrates displayed complete cleavage with 10 µM Cas6b. Some differences were observed for the other two concentrations that were quantified in triplicate (Fig. 3B). Cleavage assays with 1 µM Cas6b revealed that the two substrates containing spacer 8 (sp8) (60% and 70% substrate conversion, respectively) and the sp18 - repeat - sp19 (79% substrate conversion), as well as the sp25 - repeat - sp26 pre-crRNAs (72% substrate conversion), were less efficiently processed (Fig. 3). The addition of 0.1 µM Cas6b showed further differences with reduced crRNA maturation for some of the substrates (e.g., sp3 - repeat - sp4, sp18 - repeat - sp19, sp25 - repeat - sp26)

while other substrates displayed a comparatively increased product formation (e.g., sp2 - repeat - sp3 or sp9 - repeat - sp10).

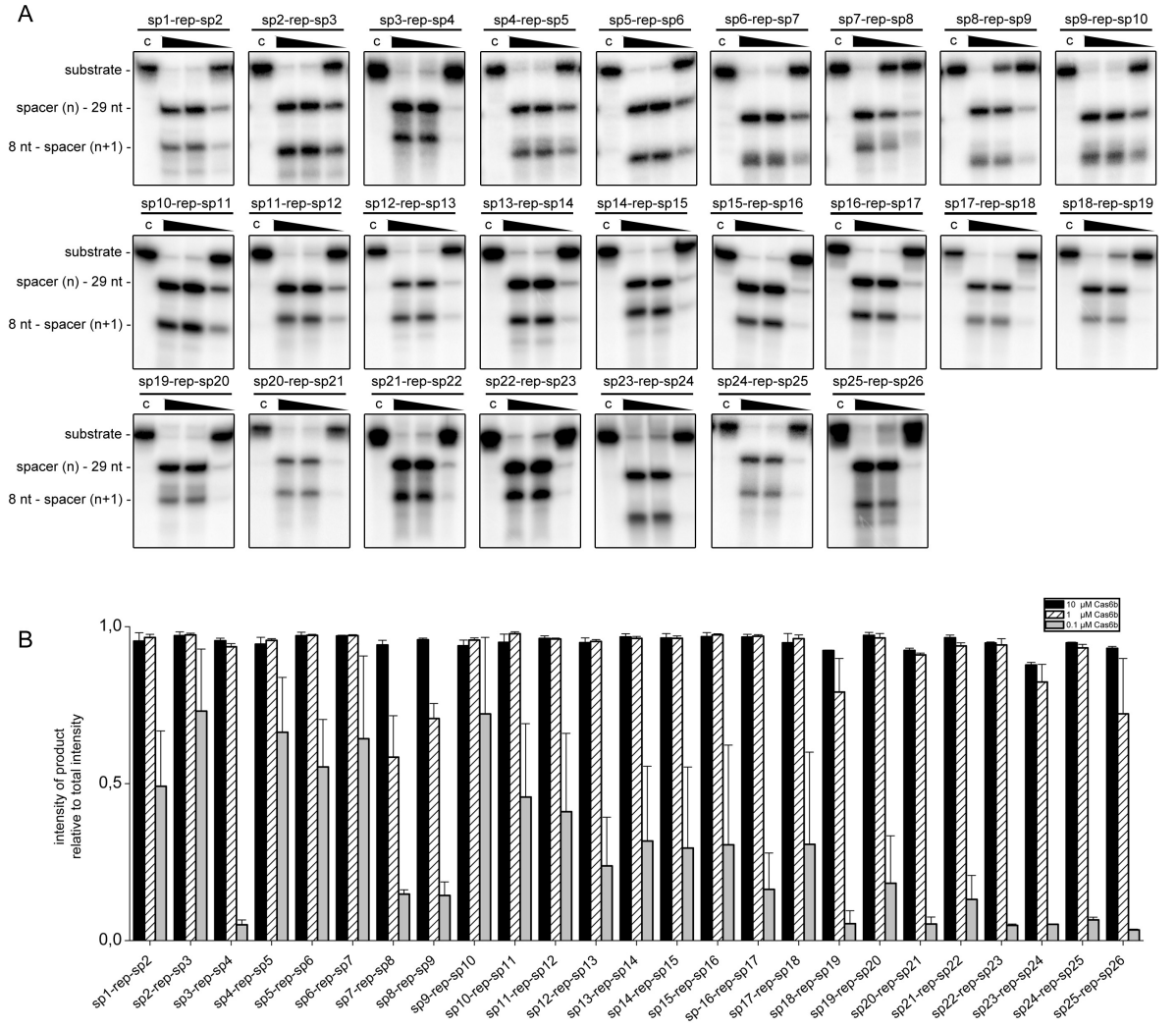


Figure 3. The influence of spacer sequences on crRNA maturation. (**A**) In three independent endonucleolytic cleavage assays, each internally labeled spacer(n) - repeat - spacer(n+1) substrate was processed using 10 µM, 1 µM and 0.1 µM Cas6b ("c" indicates controls). The RNAs were separated in denaturing polyacrylamide gels (12 %) running at 12 W and visualized by autoradiography. (**B**) Quantification of product formation. In three independent cleavage assays, the intensities of formed products in relation to the total intensity of observed bands were determined.

Minor differences of crRNA stability were observed *in vitro*

To analyze other possible factors that might influence the abundance of crRNA in *M. maripaludis in vivo*, the stability of the individual crRNAs was compared in in-line probing assays (Fig. 4). In these assays, the natural instability of crRNA is observed over time (16 h) to achieve partial digestion of the RNA, which can be slowed down or prevented by the formation of RNA secondary structures (Soukup & Breaker, 1999). In a second assay, Cas6b

was added to the reaction in order to investigate potential stabilizing or destabilizing effects of the endonuclease binding to mature crRNAs. The probing assays for most of the crRNAs of the *M. maripaludis* CRISPR locus showed no obvious differences in crRNA stability since all reactions resulted in uniform laddered degradation that is expected for unstructured RNAs (Fig. 4A, lane “I”). A potential small hairpin formed by inverted 5'-CUUG-3' sequences in the repeat was not apparent. Interestingly, the addition of Cas6b (Fig. 4A, lane “C”) to the probing assay causes, for most crRNAs, a change of the degradation pattern with the accumulation of smaller RNA fragments (most prominently e.g., for crRNAs containing spacer 10 or 23). Parallel assays with equal amounts of Cas6b storage buffer alone did not cause this destabilizing effect on the crRNAs. The crRNA24 was not degraded and the crRNA26 exhibits degradation that differs from the degradation pattern of the other crRNAs. In the probing reaction of this substrate with Cas6b, the accumulation of smaller RNA degradation products was not observed. However, the formation of a distinct larger RNA fragment can be seen. A time course experiment using this particular crRNA confirmed this observation (Fig. 4B). With longer incubation times, the amount of the large crRNA fragment increased while the substrate amount decreased. The degradation patterns for the probing reaction without Cas6b remained nearly identical during the investigated time period. RNase T1 digests were utilized to identify guanine bases in the RNA in order to pinpoint the sites of increased differential RNA stability found for probing with Cas6b (Fig. 4C). The RNase T1 digestion revealed that most of the accumulated 5'-labeled small RNA fragments have a length of 8 - 11 bases indicating the repeat tag of the crRNA. The accumulated crRNA26 fragment is cleaved between bases 22–23 of the crRNA. This is a very G-rich region of this spacer with the highest GC-content of all *M. maripaludis* spacers.

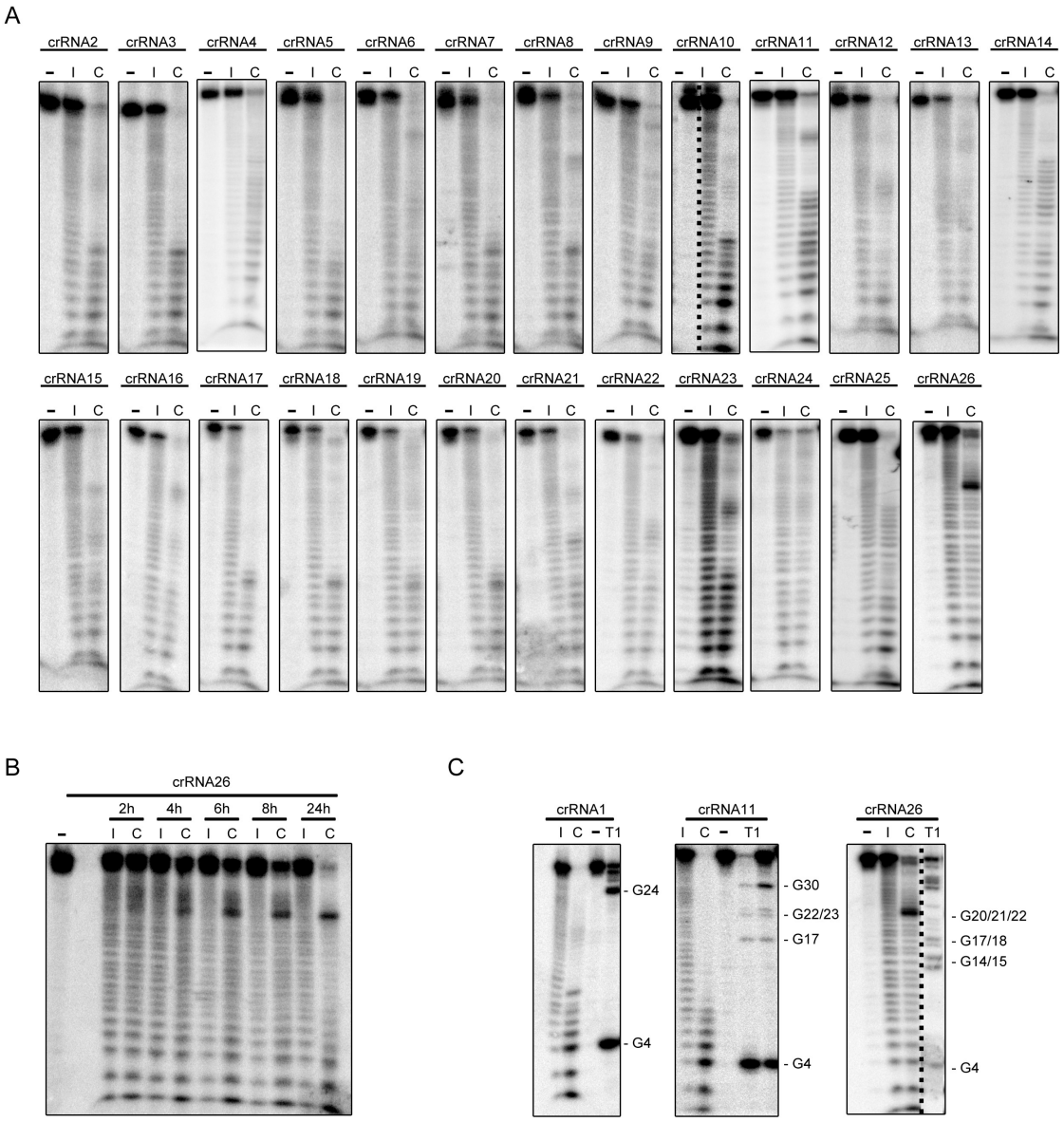


Figure 4. In-line crRNA probing assays. (**A**) Individual spacer(n) - repeat - spacer(n+1) RNAs were generated by *in vitro* run-off transcription, processed by Cas6b to yield mature crRNAs. The crRNAs were 5' labeled with (γ-^{32}p)-ATP and incubated for 16h at room temperature using a mild alkaline buffer to aim for partial digestion (lane "I"). Cas6b influence in the probing reaction was determined by addition of 20 µM Cas6b (lane "c"), untreated crRNA served as control ("-"). The probing reactions were separated by denaturing polyacrylamide gel-electrophoresis (20 %) and visualized by autoradiography. (**B**) A time-course experiment employing crRNA26 was performed taking samples at the indicated time spots. (**C**) Exemplary RNase T1 (Ambion) digests of the indicated crRNAs using 1 U of the enzyme were performed to identify specific cleavage of guanine bases and to determine the sizes of the RNA fragments. Two enzyme concentrations were tested for crRNA11 (1 U, 0.1 U)

Discussion

Different oligomerization states are reported for members of the Cas6 family of crRNA processing endonucleases. While most Cas6 proteins are reported to act as monomeric proteins, a non-catalytic Cas6 homolog from *P. horikoshii* was shown to feature RNA sequence-dependent dimerization (Wang et al, 2012). The active site of Cas6 can feature a catalytic triad that is also commonly found in tRNA splicing endonucleases (Carte et al, 2008). These splicing endonucleases always act as multimers with different families that form either heterotetrameric (Calvin et al, 2005; Randau et al, 2005a; Tocchini-Valentini et al, 2005; Yoshinari et al, 2005), homotetrameric (Li et al, 1998) or dimeric (Fujishima et al, 2011; Li & Abelson, 2000) assemblies. Our binding assays with *M. maripaludis* Cas6b and a non-hydrolysable repeat RNA clearly show two distinct shifts and, together with gel-filtration analysis, suggest dimerization of Cas6b upon substrate binding. However, binding assays with mature crRNA revealed only a single shift, which suggests monomeric crRNA binding of Cas6b after processing. Cas6b dimerization is observed in a state of stalled substrate processing, similar to the results obtained for *P. horikoshii.* Here, the non-catalytic active site of Cas6 hinders processing, while in the case of Cas6b, the 2'-deoxy modification of the repeat RNA prevents endonucleolytic cleavage. Pull-down assays with either Cas7 or Cas5 from *S. solfataricus* did not lead to co-purification of Cas6 proteins, suggesting that Cas6 does not interact (or interacts only transiently) with these Cascade proteins (Lintner et al, 2011b). In conclusion, we support a model that requires Cas6b dimerization during repeat RNA cleavage. A long CRISPR RNA precursor could potentially facilitate the binding of several Cas6b dimers to the various repeat sequences. Finally, Cas6 enzymes might act as a transporter to deliver the crRNA to the Cascade complex. Thus, commonly observed monomeric Cas6-crRNA complexes might act as a scaffold for Cascade formation around the mature crRNA. The processing of pre-crRNAs by Cas6 enzymes is highly dependent on the repeat sequence, but it is not fully understood which role spacer sequences play in influencing efficient crRNA processing and the total abundance of different crRNAs in the cell. Different factors determine which spacer is integrated into a CRISPR array. A crucial element is the protospacer adjacent motif (PAM), which is essential for recognition of protospacer sequences in the DNA of mobile genetic elements and, therefore, for the adaptation of new spacers into the growing CRISPR array (Deltcheva et al, 2011; Makarova et al, 2011b). PAM sequences were shown to be very small, only 2 nt (Gudbergsdottir et al, 2011) or 3 nt (Fischer et al, 2012) in length, and are also needed for the interference step to distinguish between self and non-self nucleic acids (Fischer et al, 2012; Gudbergsdottir et al, 2011; Semenova et al, 2011). In addition, it was shown that the length of the spacer is significant. Spacer acquisition assays in *E. coli* revealed that the majority of newly acquired spacers had a length between 32 - 33 nt with only few exceptions of longer or shorter

spacers (Yosef et al, 2012). A similar spacer distribution was identified for the three CRISPR loci of *S. thermophilus* (Horvath et al, 2008). Spacer length in *M. maripaludis* shows an analogous pattern with most of the spacers being 36 - 38 nt in length (Table S1**)**. The longest spacer is spacer 8, which is the only 40 nt spacer in the investigated CRISPR. Interestingly, both spacer - repeat - spacer RNA substrates that contained spacer 8 showed a reduced Cas6b processing efficiency, which suggests that spacer sequences or spacer length play a role in Cas6b activity. Our comparative analysis of all potential spacer - repeat - spacer combinations revealed additional reproducible differences in the efficiency of Cas6b processing of the enclosed repeat. However, these differences are often minor and, most importantly, do not clearly match with the highly variable crRNA abundance pattern observed *in vivo* (Table S1) (Richter et al, 2012). The crRNAs that are underrepresented in deep-sequencing data showed no reduced 5'-terminal maturation by Cas6b. However, potential sequencing biases in RNA-Seq studies should always be considered. Variable stability of crRNAs might also influence the crRNA abundance *in vivo*. However, our in-line probing experiments indicated a rather uniform pattern of crRNA degradation without observing stable secondary structure elements within the crRNAs. It was surprising to see that Cas6b destabilizes crRNAs, which could e.g., be explained by bending of the phosphate backbone upon Cas6b binding. Most retaining RNA degradation products show that the 5'-terminal repeat tag is protected from further degradation. In our experiment, the influence of spacer sequences on the stability of the RNA is obvious as the single spacer 26 with an unusually high GC-content forms a distinct RNA fragment that is protected from further degradation. However, crRNA26 appears not to be over-represented in the cell (Table S1**)**. In conclusion, these results show that while spacer sequences do have an effect on Cas6b processing and Cas6b-crRNA stability, these effects do not fully explain the *in vivo* crRNA abundance patterns. These patterns are suggested to be determined by several factors that include the ones we measured with the presented methodology. Additional parameters that could not be addressed with the current experimental setup are (1) the establishment of the interaction of crRNAs with Cas protein interference complexes in the cell, (2) differential turnover of crRNAs depending on their utilization in the cell and (3) different intrinsic stabilities of crRNAs after Cas protein interference complex binding. It could be shown for *S. solfataricus* that the abundance of crRNAs bound to the interference Cas protein complex CMR was also very diverse (Zhang et al, 2012). Therefore, the distance of the crRNA's spacer to the promoter and the success rate of the loading of crRNAs into a CMR or Cascade complex might have the strongest influences on crRNA abundance in the cell. Finally, the sequences of potential spacers have a huge impact on the global CRISPR transcription if they contain elements that might either serve as promoters or terminators. Recent studies revealed internal promotion of CRISPR array transcription (Deng et al, 2012; Richter et al, 2012) suggesting that spacer

sequences with promoter elements could be selected for during evolution if the increased transcription of crRNAs with older spacers was beneficial. A problematic scenario is the possible integration of spacer sequences with transcription termination sequences into a CRISPR. Many questions remain regarding the termination signals in archaeal genomes. However, one commonly observed feature, are poly-T stretches at termination sites (Santangelo & Reeve, 2006; Thomm et al, 1993). We observed that *M. maripaludis* and many archaea avoid the adaptation of spacer sequences with long poly-T stretches. In addition, several long poly-A stretches are found in the newly acquired spacers of the *M. maripaludis* CRISPR. These poly-A stretches specify poly-T stretches in the reverse direction and might provide beneficial termination signals for the prevention of anti-crRNA production. In addition, repeat sequences appear to have evolved to lack poly-T (more than 3 T bases in a row) motifs in Archaea while these are commonly found in bacterial repeats (Kunin et al, 2007). Careful manual inspection of archaeal repeats identified only two sequences with four consecutive T residues while the majority of them (over 300) contained poly-A sequences. Future research on the loading efficiency of Cascade complexes, internal promotion and termination signal will be required to fully assess the diverse influences that spacer sequences can have on crRNA abundance and their effectiveness against viral attacks.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We thank André Plagens for advice and discussions. This work was supported by grants from the Deutsche Forschungsgemeinschaft (DFG grants FOR1680 RA 2169/1-1 to L.R. and BA 2168/5-1 to R.B.) and the Max-Planck Society.

Supplementary Materials

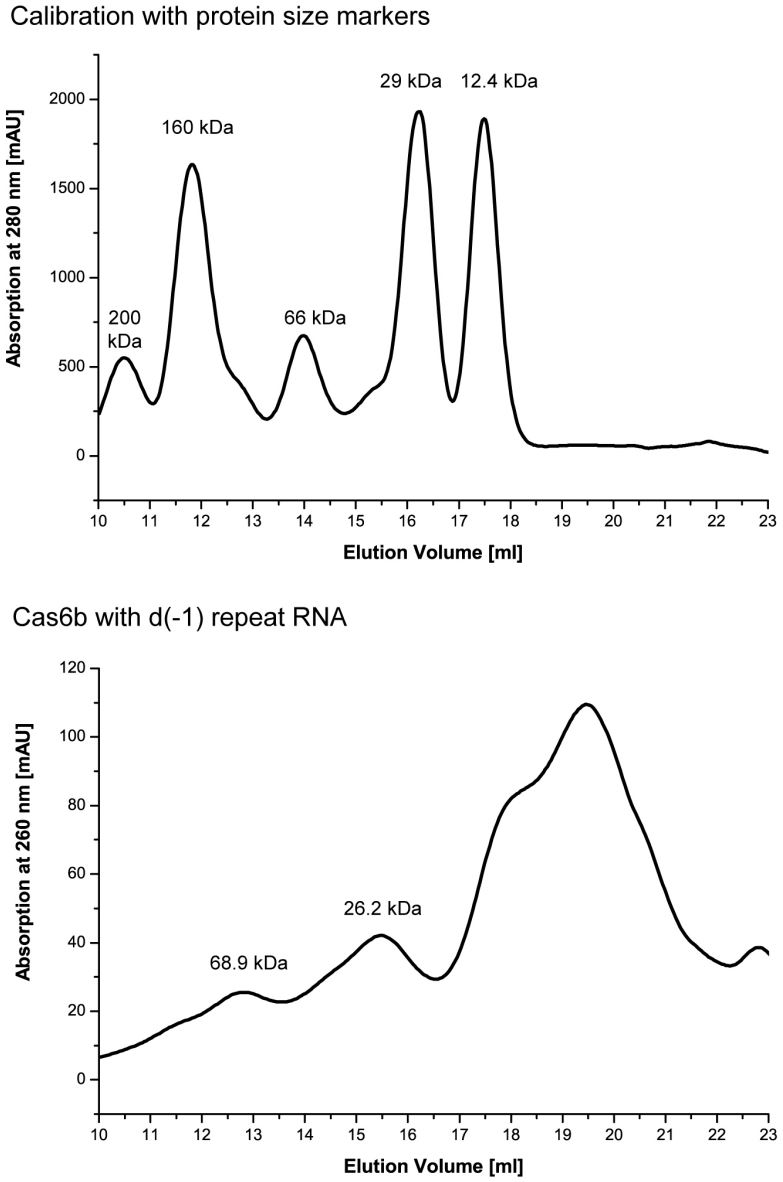


Figure S1. Size exclusion chromatography of Cas6b and d(-1) repeat RNA. SDS-PAGE analysis revealed the presence of Cas6b in fractions corresponding to two peaks: peak 1 (12 ml, 13 ml) and peak 2 (15 ml, 16 ml). Calibration of the molecular weight of fractionated proteins determined the mass of peak 1 at 69 kDa (Cas6b dimer + d(-1) repeat RNA has a mass of 69.2 kDa) and peak 2 at 26.2 kDa (Cas6b monomer has a mass of 28.3 kDa).

Table S1. *M. maripaludis* C5 spacers

Spacer #	Spacer position	Spacer length	Spacer sequence	Relative abundance[1]	Spacer GC (%)
1	742916	35	TTCTCCTTCCATCTTGAACAGTGTGAGGTAGCAGG	88,4	49
2	742988	37	ATTAATCCCATAATACTTTTCTAGGTCTGGGCGGAAT	100,0	38
3	743062	34	TAAAAAAAGAAAAAAGTTAAAAAAGCTAGAATAA	6,6	15
4	743133	37	AAAAAACAGATTGAAAAATTAACGGCAAAACAATGAA	18,0	24
5	743207	36	AAAACGAAAGAAGACTCAACAAGCTAGAAAAACAAT	55,2	31
6	743280	34	CAATTGCTAAGAGTGCACTAATCGCACTAATAAA	22,4	35
7	743351	35	ATTTTAAGTATGCGATGTACTCGGATTGATAAATT	50,6	29
8	743423	40	ACGTAGATACGCTTGTTGCTTCAATGGAAAATTCAGAAGG	40,3	40
9	743500	36	TTTGTGAATCCCATACAAGGGTTACAGCAGTTACAG	35,1	42
10	743573	37	ATTTATACGACTAGAAGACCGGCAAAAGTGTATACGG	10,0	41
11	743647	34	AAAATAAAGCAACGGTTATAAGTCCGGTTGCGAT	4,8	38
12	743718	37	AGCGTCTATATAAGCGGAACTATACGCACTTTCCGGT	21,9	46
13	743792	37	CGAATGAAACGAGCGAAACAAGTAATAAAAAATCAAG	9,4	32
14	743866	36	CAGTAAAAACATTTTGATTTATTATTTTGATCCATT	1,7	19
15	743939	39	CAGTAAAAAAAGTTAATGAAGAAATTACACCAAGTACTT	1,6	26
16	744015	38	TCGATATAAGCGGAACTATACGCACTTTCCGCTCTTAA	9,1	42
17	744090	38	TTTTTTTATACGTTTAGTAACAGGGTTAATTCTGTTTA	5,0	24
18	744165	37	TTTAATTCCCATTCATGGTTTGGATCTGGAATAAATG	2,3	32
19	744239	38	TTTTGTTCGTGTCTTTTCTTCTCAATAATTCTTTTAAA	4,2	24
20	744314	37	TTTTTTTATTGTAATCGGTTATACTGTTGTATTCTGT	4,6	24
21	744388	38	ACAATTTAGATGTTATAGAGTTTTTAAAACAGAATAAA	6,2	18
22	744463	36	TACAGAATTTACGTGATTTTAATTATGGTATATATT	0,8	19
23	744536	38	TATTGTCGATTTTATTTCGCTTAACTCTGTCGGATTAT	9,4	32
24	744611	36	TGATGGAGTCAAAAGGTAAGACTGTTAAATTTCCGG	14,8	39
25	744684	37	ACCTTGAAAATATAACATATGAGCAAAAAATTGAAAT	3,2	22
26	744758	38	CCACCGGCGGAGGGCTCTAAAACGAGGGTACTCGAATA	7,2	58
27	744833	37	CTGATGAAACGAGCGAAACAAACAATAAAAAATCAAG	1,8	32

[1] Relative abundance of individual *Methanococcus maripaludis* C5 crRNAs was calculated from RNA-Seq data taken from Richter et al, 2012. The most abundant spacer 2 was set as 100 and abundances were calculated accordingly.

Chapter IV

Recognition of crRNA precursor repeat substrates by the single turnover endoribonuclease Cas6b of *Methonococcus maripaludis* C5

Hagen Richter[1], Kundan Sharma[3], Henning Urlaub[3] and Lennart Randau[1, 2,*]

[1] Max-Planck-Institute for Terrestrial Microbiology, Karl-von-Frisch Straße 10, 35043 Marburg
[2] LOEWE Center for Synthetic Microbiology (Synmikro), 35043 Marburg
[3] Bioanalytical Mass Spectrometry Group, MPI for Biophysical Chemistry, Göttingen, Germany
[*] Corresponding author

Abstract

Bacteria and Archaea developed a variety of defense mechanisms against viruses including the adaptive immune system CRISPR-Cas. This system utilizes small interfering RNAs (crRNAs), which can contain unique sequences derived from viral genomes. Together with a complex of CRISPR associated proteins the crRNA base pairs with invading nucleic acids and mediates immunity (Barrangou et al, 2007; Brouns et al, 2008). CRISPR-Cas systems show a broad diversity (Makarova et al, 2011a; Makarova et al, 2011b), which is also exemplified in the Cas6 proteins, the enzymes generating mature crRNA (Carte et al, 2008; Gesner et al, 2011; Haurwitz et al, 2010; Reeks et al, 2013c; Richter et al, 2012; Sashital et al, 2011). Here, we detail a biochemical analysis of the Cas6 homolog Cas6b of subtype I-B in *M. maripaludis* C5. Cas6b is a single turnover enzyme with a binding affinity of 668 nM and an observed reaction rate of 0.296 min^{-1}. We identified an important amino acid residue (M185) for RNA binding and two additional residues that are involved in crRNA maturation (K30 and Y47). Our studies further reveal that the repeat sequence and not the repeat structure is important for precursor crRNA processing by Cas6b.

Introduction

CRISPR-Cas is an adaptive prokaryotic defense mechanism against invading nucleic acids that recently generated interest as a new genome editing tool (recently reviewed in (Charpentier & Doudna, 2013; Horvath & Barrangou, 2013; Richter et al, 2013b)). The system is composed of clustered regularly interspaced short palindromic repeats (CRISPR) adjoined to a set of CRISPR associated (Cas) genes. The CRISPR array comprises unique spacer sequences that intersperse a series of repeats. Spacer sequences can be derived from genomes of mobile genetic elements and, as part of the interfering small RNA (crRNA), these spacers help to counter a reoccurring attack of these elements. The basic mechanism of the CRISPR immunity can be divided into three major steps: i) adaptation of new spacers, ii) the maturation of crRNAs and iii) interference with the alien nucleic acid (Barrangou et al, 2007; Brouns et al, 2008; Horvath & Barrangou, 2010; Jore et al, 2011a; Sorek et al, 2008; Terns & Terns, 2011; van der Oost et al, 2009; Westra et al, 2012b). Although the details of the adaption process are still not completely understood, it was shown that Cas1 and Cas2 play a critical role in adaption and that the putative acquisition complex Cascis (CRISPR associated complex for integration of spacers) requires only the leader sequence and one repeat unit to adapt new spacer sequences (Datsenko et al, 2012; Erdmann & Garrett, 2012; Plagens et al, 2012; Swarts et al, 2012; Yosef et al, 2012; Yosef et al, 2013). In Type-I systems more information is present for the processing of precursor-crRNA (pre-crRNA),

mediated by Cas6 family proteins (Carte et al, 2008; Gesner et al, 2011; Haurwitz et al, 2010; Reeks et al, 2013c; Richter et al, 2012; Sashital et al, 2011) and the interference step in which a complex of different Cas proteins, called Cascade (CRISPR associated complex for antiviral defense) together with the mature crRNA marks the invader for degradation by Cas3 (Cady & O'Toole, 2011; Howard et al, 2011; Jore et al, 2011b; Lintner et al, 2011b; Mulepati & Bailey, 2011; Sashital et al, 2012; Sinkunas et al, 2011; Sinkunas et al, 2013; Westra et al, 2012a; Wiedenheft et al, 2011a; Wiedenheft et al, 2011b). While Cascade is a complex found in Type-I systems, Type-II systems utilize a single multifunctional protein called Cas9 to generate mature crRNAs and to interfere with the alien nucleic acid (Chylinski et al, 2013; Deltcheva et al, 2011; Jinek et al, 2012). Two Cascade-like complexes (CMR or CSM) are reported for Type-III systems (Hale et al, 2009; Marraffini & Sontheimer, 2010a; Osawa et al, 2013; Shao et al, 2013; Shao & Li, 2013; Zhang et al, 2012).

The three major types of CRISPR-Cas systems were defined by computational analyses and differ in their main interfering proteins: Cas3 (Type-I), Cas9 (Type-II) and Cas10 (Type-III). A further classification into at least 10 subtypes is based on subtype-specific proteins (e.g. Cas6b, Cas8b) (Makarova et al, 2011b). Subtype I-B with its specific proteins Cas8b and Cas6b can be found e.g. in Clostridia, methanogens and halophiles. Both in *Methanococcus maripaludis* and in *Clostridium thermocellum* an I-B system with 8 Cas proteins has been be identified. These include the universal Cas1, Cas2 and Cas4 proteins, proposed to be part of the spacer acquisition complex Cascis, the four interference complex proteins Cas3, Cas5, Cas7, Cas8b and the pre-crRNA processing enzyme Cas6b (Richter et al, 2012).

The first description of a crRNA maturation protein was published for the metal-independent subtype III-B Cas6 of *Pyrococcus furiosus* (Pf Cas6) (Carte et al, 2010; Carte et al, 2008; Wang et al, 2011; Wang et al, 2012). Further characterization of crRNA maturation by Cas6 family proteins was done for subtype I-E (Cas6e, formerly Cse3) in *Escherichia coli* and *Thermus thermophilus* (Gesner et al, 2011; Sashital et al, 2011), subtype I-F (Cas6f, also known as Csy4) of *Pseudomonas aeruginosa* (Haurwitz et al, 2010; Haurwitz et al, 2012; Sternberg et al, 2012) and subtype I-B (Cas6b) in *M. maripaludis* C5 and *Methanosarcina mazei* (Nickel et al, 2013; Richter et al, 2013a; Richter et al, 2012). Surprisingly, no Cas6 homolog is found in subtype I-C. Here, Cas5d appears to take over the functions of both Cas6e and Cas5e (Garside et al, 2012; Nam et al, 2012). Cas6 enzymes share similar features including a ferredoxin-like fold and the general acid/base mechanism of pre-crRNA processing. Considering that the overall sequence identities are very limited, the structural similarity of Cas6 enzymes is striking. The sequence differences of these enzymes are most obvious in the composition of the catalytic sites. While Pf Cas6 employs a triad of tyrosine, histidine and lysine residues (Carte et al, 2010; Wang et al, 2011), Cas6f utilizes a dyad of a serine and a histidine residue (Haurwitz et al, 2012). Cas6e activity also relies on the

canonical catalytic histidine and additionally a tyrosine has been reported to play a role in processing (Gesner et al, 2011; Sashital et al, 2011). Biochemical characterization of Cas6b reports the importance of two interchangeable histidine residues (Nickel et al, 2013; Richter et al, 2012). Despite their different active site composition and substrate recognition, the resulting mature crRNA always consists of a full-length spacer flanked by remaining repeat parts forming a distinct 8 nt 5' terminal tag and a gradually trimmed 3' terminal tag. Cas5d crRNA maturation on the other hand yields an 11 nt 5' terminal repeat tag (Carte et al, 2008; Gesner et al, 2011; Haurwitz et al, 2010; Nam et al, 2012; Richter et al, 2012; Sashital et al, 2011). The 5' repeat tags are important during the discrimination of self- and non-self DNA (Marraffini & Sontheimer, 2010b; Sashital et al, 2012; Westra et al, 2013).

Cas6 enzymes are often reported to bind their substrates as a monomer (Carte et al, 2010; Gesner et al, 2011; Haurwitz et al, 2010; Sashital et al, 2011; Wang et al, 2011). Exceptions are found in a non-catalytic Pf Cas6 homolog of *Pyrococcus horikoshii* (Wang et al, 2012), in a dimeric form of *Sulfolobus solfataricus* Cas6 (Sso Cas6) (Reeks et al, 2013c), in two Cas6e variants of *T. thermophilus* (Niewoehner et al, 2013) and Cas6b of *M. maripaludis* (Richter et al, 2013a), which were observed to bind RNA substrates in dimeric manner. While the reported monomeric binding correlates with the proposed wrap-around mechanism, in which Cas6 is bound to the long precursor in a bead chain like manner (Wang et al, 2011), the dimeric binding supports an individual processing of pre-crRNA by Cas6 dimers before being transported to Cascade. Repeat sequences play a critical role for Cas6 binding and also for Cas6 processing. Despite their different lengths, the repeats can differ in their secondary structures, which are specifically recognized by some Cas6 (Gesner et al, 2011; Haurwitz et al, 2010; Sashital et al, 2011; Sternberg et al, 2012). Other studies showed that structures of repeats, often formed due to their palindromic nature, do not influence the cleavage reaction (Carte et al, 2010; Carte et al, 2008; Wang et al, 2011). In addition to a possible structure of repeats, the sequence itself seems to be important not only for binding but also for processing by Cas6. The length and individual nucleotides of the sequence showed to be critical for binding and cleavage by Cas6 enzymes (Carte et al, 2010; Carte et al, 2008; Sternberg et al, 2012; Wang et al, 2011; Wang et al, 2012), while the deletion of some sequence stretches, even within the 5' terminal tag, does not influence processing by Cas6 (Wang et al, 2011; Wang et al, 2012). Some nucleotides seem to have only an influence on processing by affecting the binding behaviour of Cas6 (Carte et al, 2010; Sternberg et al, 2012), whereas other nucleotides were shown to inhibit processing, while binding was still possible (Wang et al, 2012).

In this study we show that Cas6b of *M. maripaludis* is a single turnover enzyme with an observed reaction rate of k_{obs} = 0.296 min^{-1} and a binding affinity to non-cleavable substrate of K_d= 668 nM. We further show that, besides the previously described catalytic histidines

(H38 and H40) (Richter et al, 2012), a lysine (K30) and a tyrosine (Y47) are critical for Cas6b processing activity. UV-crosslinking experiments with Cas6b that were followed by mass spectrometric analysis identified a methionine residue (M185) that has a major influence on repeat binding by Cas6b. Mutation of this residue resulted in a two times lower affinity to the respective RNA substrate (K_D=1.03 µM) compared to the wild-type enzyme. Analyses of different repeat mutants showed that the sequence of the *M. maripaludis* repeat is critical for Cas6b recognition and cleavage. Finally, we could show that repeats of an archaeal and a bacterial subtype I-B system are interchangeable and only the disruption of individual key nucleotides affects the processing reaction by Cas6b.

Material and Methods

Growth of *E. coli*

E. coli cells were grown in LB-media with appropriate antibiotics at 37 °C and shaking at 200 rpm.

Production of Cas6b and mutants

Using genomic DNA of *M. maripaludis* C5 the *cas6b* gene was amplified and cloned into the vector pET-20b to enable for a protein expression with a C-terminal His-tag. QuikChange site-directed mutagenesis was carried out according to the manufacturer's instructions (Stratagene mutations were confirmed by sequencing (MWG Eurofins)). The different Cas6b variants were produced in *E. coli* (Rosetta2 DE3) cells as described previously (Richter et al, 2012). After growth to an OD_{578} of 0.6 the protein expression was induced by addition of isopropylthio-β-D-galactoside (IPTG) to a final concentration of 0.5 mM. The cells were harvested 4 hours post induction and the cell pellet was resuspended in 5 ml lysis buffer (10 mM Tris-HCl [pH 8.0], 300 mM NaCl, 10 % glycerol and 0.5 mM DTT) per 1 g cells. Lysis was carried out on ice with lysozyme (1 mg / g cell pellet) for 30 min. Cell disruption was achieved by applying 8 cycles of 30 s sonication (Branson Sonifier 250) and 30 s rest on ice. The lysate was cleared by centrifugation (47800 g, 30 min, 4 °C) and subsequently the supernatant was loaded onto a Ni-NTA-Sepharose Column (GE-Healthcare) and purified using a FPLC Äkta-Purification system (GE-Healthcare). To elute the proteins, a linear gradient of imidazole was applied (0 – 500 mM) and the purity of eluted proteins was determined by SDS-PAGE and coomassie blue staining. Pure protein fractions were pooled and dialysed into lysis buffer following a protein concentration determination by Bradford Assay (BioRad).

Generation RNA substrates

M. maripaludis C5 wild-type repeat RNA, its deoxy substitution variant (d-1) and the wild-type repeat RNA of *C. thermocellum* were synthesized by MWG Eurofins. Synthesized RNA was 5' end labelled using T4 polynucleotide kinase (Ambion) and [γ-^{32}P] ATP (5000 ci/mmol, Hartman Analytic) according to the manufracturer's instructions. The labelling reaction was carried out for 1 hour at 37 °C in final volume of 10 µl (1 µl PNK buffer (NEB), 25 U T4 PNK (Ambion), 10 pmol RNA).

Generation of *M. maripaludis* repeat variants was achieved either by site-directed mutagenesis of a pUC19 vector containing a spacer2 - repeat - spacer3 construct (Richter et al, 2012) (G16C, G16C/C25G; Stratagene) using oligonucleotides designed by Agilent's Primer Design tool or by direct synthesis of oligonucleotide pairs with the desired mutation (U15A, U32A; MWG Eurofins). Substrates harbouring the repeat variants G16C and G16C/C25G were yielded by *in vitro* run-off transcription after linearization of the vector using HindIII. The variants U15A and U32A were also yielded by *in vitro* run-off transcription, but here a hybridization (heating of the mixed oligonucleotide pair for 5 min at 95 °C and slow decrease of temperature for 2 hours) of the synthesized oligonucleotides preceded the *in vitro* transcription. *In vitro* run-off transcription was carried out for 3 hours at 37 °C in a final volume of 20 µl using the linear DNA templates described above (40 mM HEPES-KOH [pH 8.0]; 22 mM $MgCl_2$; 5 mM DTT; 1 mM spermidine; 4 mM UTP, CTP, GTP, ATP; 25 µCi [α-^{32}P] adenosine triphosphate (ATP) (5000 Ci/mmol, Hartman Analytic) 20 U RNase inhibitor; 1 µg T7 RNA polymerase; 1 µg linear template).

RNA labelling reactions were separated by a denaturing polyacrylamide gel electrophoresis (8 M Urea; 1 x TBE; 10 % polyacrylamide). Using sterile scalpels the respective bands were cut out of the gel in reference to a brief autoradiographic exposure. Elution of the RNA took place in 500 µl RNA elution buffer (250 mM NaOAc; 20 mM Tris-HCL [pH 7.5]; 1 mM EDTA [pH 8.0]; 0.25 % SDS) over-night shaking on ice. Precipitation of RNA was achieved by addition of 2 Vol EtOH (100 %, ice cold) and 1/100 glycogen (Roche) and incubation at – 20 °C for 1 h. Subsequent to pelleting, the RNA was washed with 70 % EtOH and the dried pellet was resuspended in H_2O_{DEPC}.

Endonuclease assay

Indicated concentrations of the different Cas6b variants were incubated with the radiolabelled RNA substrates in a final volume of 10 µl (250 mM KCl; 1.875 mM $MgCl_2$; 1 mM DTT; 20 mM HEPES-KOH [pH 8.0]). Reactions were incubated at 37 °C for the indicated period of time and immediately mixed with 2 x formamide buffer (95 % formamide; 5 mM EDTA [pH 8.0]; 2.5 mg bromophenolblue; 2.5 mg xylene cyanol) and heated at 95 °C for 5 min to stop the cleavage reaction. For turnover and kinetic studies of Cas6b a master mix of 5 (50 µl) and 24 (240 µl) reactions, respectively, was prepared to assure constant reaction conditions. After the indicated time period a 10 µl sample was removed and mixed with 2 x formamide buffer and heated to 95 °C for 5 min.

The reactions were loaded on a denaturing polyacrylamide gel (12 %) and run in 1 x TBE at 12 W for 1.5 hours. Visualization was achieved by phosphorimaging. Substrate processing was quantified using ImageQuant software. Mean values of relative processed substrate out

of three independent turnover assays were plotted using Origin and standard deviation was calculated using Excel. Kinetic values were plotted using Origin and fitted according to first-order reactions. Linearization of the data was calculated with Excel and plotted with Origin; the slope of the equation equals k_{obs}, the observed reaction rate.

Electrophoretic mobility shift assay (EMSA)

In binding assays the indicated concentrations of either Cas6b wild-type or Cas6b M185A were mixed with the radiolabelled (d-1) repeat variant in a final volume of 10 µl (10 mM Tris-HCl [pH8.0]; 200 mM KCl; 5 % glycerol; 0.5 mM DTT; 0.5 mM EDTA [pH 8.0]; 1 µg BSA). The mixture was incubated at 37 °C for 1 h before addition of 6 x DNA loading dye (4 g sucrose; 2.5 mg bromophenolblue; 2.5 mg xylene cyanol in 10 ml H_2O). A 7 % native polyacrylamide gel running in 1 x TBE at 8 W for 2.5 hours was used to separate the reaction products. Visualization was achieved by autoradiography and quantification of EMSAs for K_D value determination was done with ImageQuant. Relative amounts of bound RNA were plotted against the used protein concentration and fitted with a hyperbolic equation using Origin. Linearization of the data was achieved by plotting the data in a Scatchard-plot in which the negative reciprocal of the slope equals the K_d value.

UV induced protein-RNA crosslinking and enrichment of cross-linked peptides

UV crosslinking and enrichment of cross-linked species was achieved as previously described (Schmidt et al, 2012). Cas6b and (d-1) repeat RNA were mixed in equimolar ratios (1 nmol each) in a total volume of 100 µl of lysis buffer. Formation of complex was achieved by incubation on ice for 30 min before being transferred in black polypropylene microplates (Greiner Bio-One) and irradiated at 254 nm for 10 min. Following a RNA precipitation (described above) the samples were denatured (4 M Urea, 50 mM Tris-HCl [pH 7.9]) and digested with Benzonase (Ambion) for 1 h at 37 °C. Proteolysis was performed over-night at 37 °C using trypsin (Promega). After desalting on a C18 column (Dr. Maisch GmbH, Göttingen) the cross-linked peptides were enriched using a TiO_2 column (GL sciences). Finally, the samples were dried and resuspended in 10 µl solvent (5 % acetonitrile (ACN), 1 % formic acid (FA)).

Nano-liquid chromatography and MS analysis

5 µl of the cross-linked sample (above) was subjected to a nano-liquid chromatography system (Agilent 1100 series, Agilent Technologies) including a C18 trapping column (length:

2 cm; inner diameter: 150 µm; flow rate: 10 µl/min; buffer A: 0.1 % FA) in line with a C18 analytical column (length: 15 cm; inner diameter: 75 µm) (both C18 AQ 120 Å 5 µm, Dr. Maisch GmbH). After elution from the trapping column the analytes were separated in the analytical column using a gradient of 7 - 38 % buffer B (95 % ACN, 0.1 % FA) with an elution time of 33 min (0.87 %/min) and a flow rate of 300 nl/min.
Online ESI-MS was performed using an LTQ-Orbitrap Velos (Thermo Scientific), operated in data-dependent mode using a TOP10 method. The MS scans were recorded in a m/z range of 350 – 1600 and for subsequent MS/MS the top 10 most intense ions were selected. Fragments, generated by HCD activation (higher energy collision dissociation, normalized collision energy = 40), and precursor ions were scanned in the Orbitrap. The resulting spectra were measured with high accuracy (< 5 ppm) both at MS and MSMS level.

MS data analysis

MS *.raw files were converted into *.mzML format using msconvert (Kessner et al, 2008). Open MS (Bertsch et al, 2011; Sturm et al, 2008) and OMSSA (Geer et al, 2004) search engines were employed to analyze the protein-RNA cross-links. High scoring cross-linked peptides were manually annotated for confirmation.

Repeat Analysis

RNA structures were predicted using RNAfold (McCaskill, 1990; Schuster et al, 1994; Zuker & Stiegler, 1981) at standard settings. Repeat sequences were aligned using VectorNTI (Invitrogen) with a ClustalW2 algorithm (Larkin et al, 2007)and a sequence logo of different subtype I-B repeat sequences was created with WebLogo using the basic settings (Crooks et al, 2004; Schneider & Stephens, 1990).

Results

Cas6b is a single turnover enzyme with a fast reaction rate

Endonucleolytic cleavage assays using recombinant Cas6b were employed to determine the enzymatic turnover of Cas6b (Fig. 1a). In three independent assays with three different RNA to protein ratios and different incubation times, only in an 1:1 ratio, the RNA substrate was cleaved with high conversion rates of about 80 %. Decreasing protein concentrations and maintaining constant RNA levels (10 nM) showed a significant reduction in substrate conversion rate to about 40 % and 15 % for 1:5 and 1:10 ratios, respectively (Fig. 1b), indicating a single turnover mechanism of Cas6b.

Determination of the observed reaction rate (k_{obs}) was performed in time course experiments in which the substrate conversion was monitored over 570 s (samples taken every 30 s). The assay shows maximum conversion already after 200 s and indicates a very fast reaction rate (Fig. 1c). Quantification and plotting of the data shows (Fig. 1d) that 70 % conversion is achieved after approximately 200 seconds and stays constant during the remaining incubation time. Linearization (small graph inserted in Fig. 1d) of the data was for the data obtained for the first 200 seconds after which the plateau of conversion was reached. Three independent assays show a substrate conversion rate k_{obs} = 0.296 min^{-1} ± 0.01 min^{-1}.

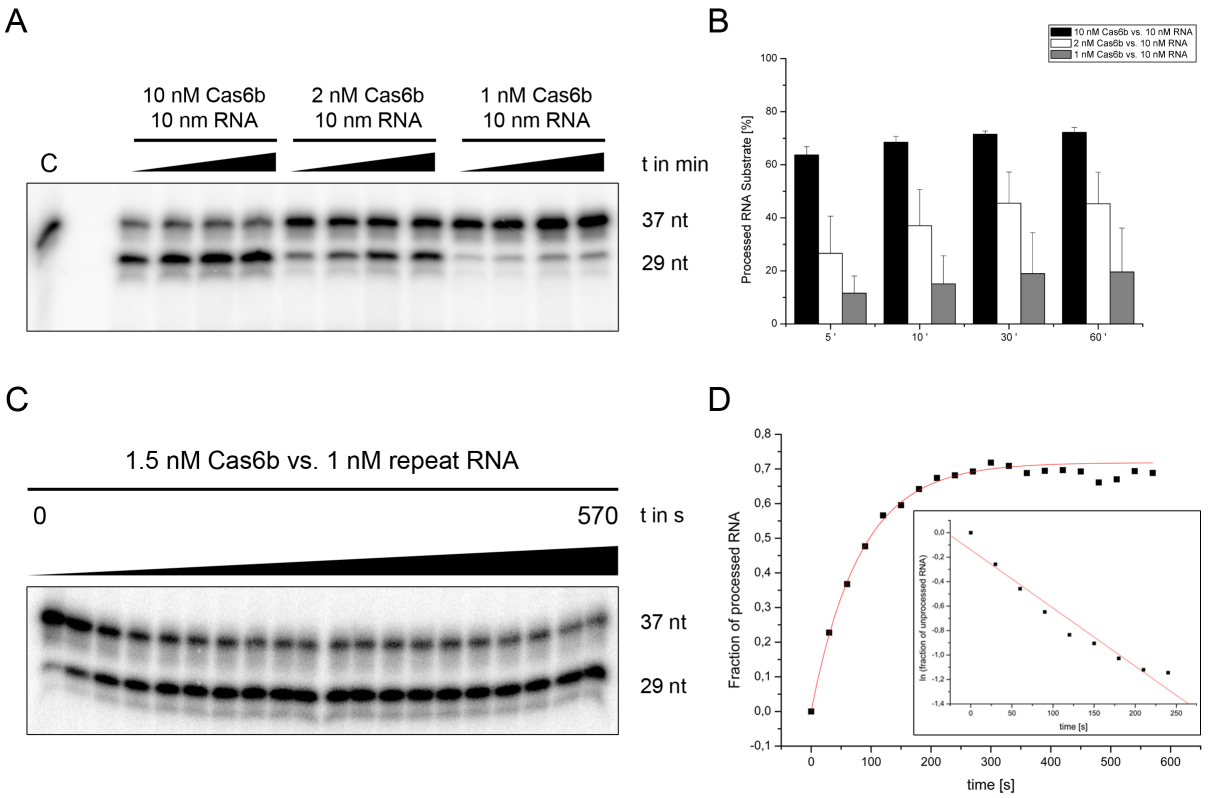


Figure 1. Cas6b is a single turnover endonuclease. To determine the turnover and reaction rate of Cas6b, endonuclease assays were performed with 5' end labelled repeat RNA. (**A**) Determining the turnover, a representative assay using 3 different Cas6b to RNA ratios (1:1, 1:5 and 1:10) incubated for 4 different time intervals (5 min, 10 min, 30 min and 60 mins) is shown. Processing occurs in all reactions, yet only in the 1:1 ratio a significant conversion rate is observed, indicative for a single turnover enzyme. (**B**) Substrate processing of three independent assays was quantified. (**C**) One representative assay in which 1.5 nM Cas6b with 1 nM repeat RNA were incubated over a time course of 570 seconds taking a sample every 30 seconds is displayed. (**D**) Quantified data was plotted and fitted by the first-order rate law. The data obtained for the first 200 seconds of the reactions were linearized and in three independent assays the reaction rate k_{obs} = 0.296 min^{-1} ± 0.02 min^{-1} was determined.

The active site of Cas6b is composed of four active residues

Previous studies identified two interchangeable histidine residues as part of the Cas6b catalytic site (Nickel et al, 2013; Richter et al, 2012). A structural model and alignment (Richter et al, 2012) was used to deduce further amino acid residues that might be involved in the processing reaction of Cas6b. Single mutations of conserved residues in Cas6b did not seem to have a significant impact on activity (Tab. 1). Therefore, double mutants were introduced by site-directed mutatgenesis. These mutants were analysed using endonuclease assays employing *M. maripaludis* C5 repeat RNA as substrate (Fig. 2a). Of the created mutants (Tab. 1), only a mutation of both lysine at position 30 together with a tyrosin at position 47 resulted in a significant reduction of conversion to 20 % (Fig. 2b). The mutation of two conserverd lysines at position 29 and 30 and the double mutant of K29 and Y47 resulted in a Cas6b with wild-type like conversion rates. Together four active residues were identified: lysine 29, histidine 38, histidine 40 and tyrosine 47.

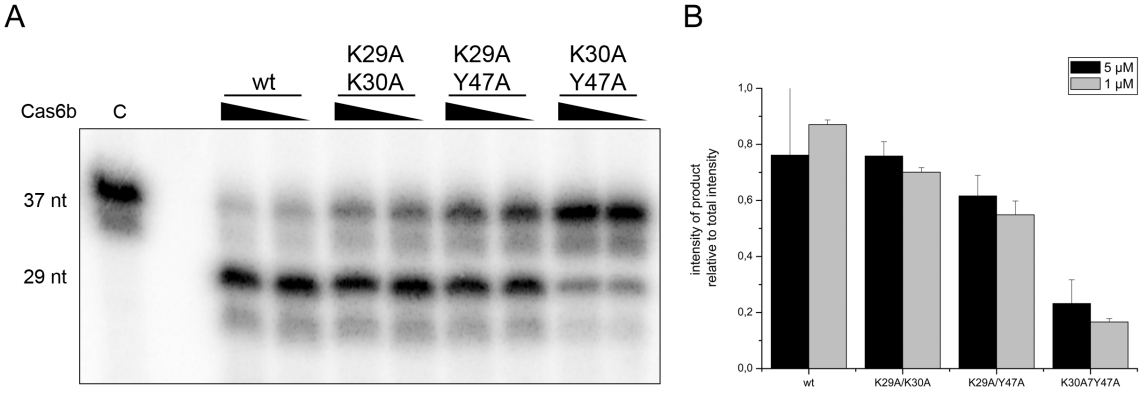


Figure 2. A lysine and tyrosine residue are part of the catalytic site of Cas6b. Indicated Cas6b mutants were used in endonuclease assays with 5' end labelled repeat RNA. (**A**) The activities of the Cas6b variants are shown in a representative assay. While the two double mutants of lysine 29/lysine 30 and lysine 29/tyrosine 47 did not show significant differences compared to wild-type (wt) Cas6b,

the double mutant lysine 30/tyrosine 47 displays a highly reduced processing activity. (**B**) Product formation by the indicated Cas6b variants was quantified in three independent assays.

It should be noted that several mutants did not yield soluble protein (Tab. 1) and it cannot be excluded that these residues might also be involved in the cleavage reaction.

Table 1. Set of different Cas6b mutants.

active site mutants			binding site mutants		
mutated residue	soluble protein	active protein	mutated residue	soluble protein	binding protein
R24A	yes	yes	M185A	yes	(yes)
Y26A	yes	yes	G187A	no	-
K29A	yes	yes	G190A	yes	yes
K30A	yes	yes	G203A	no	-
Y31A	yes	yes	G205A	no	-
H38A	yes	(yes)	G203A/G205A	no	-
H38F	yes	(yes)	G211A/G213A	no	-
H40A	yes	(yes)			
H40F	yes	(yes)			
Y47A	yes	(yes)			
Y49A	no	-			
Y49F	no	-			
K55A	yes	yes			
Y26A/K55A	no	-			
K30A/Y47A	yes	no			
K29A/K30A	yes	yes			
K29A/Y47A	yes	yes			
K29A/K30A/ Y47A	yes	no			
Y31/K55	no	-			
H38A/H40A	yes	no			
Y47A/Y49A	no	-			
Y47A/Y49F	no	-			

A methionine residue at position 185 binds to a UUGC motif of a non-hydrolysable repeat substrate

It was shown previously, that Cas6b binds to a non-cleavable repeat version with a deoxynucleotide substitution (d-1 repeat) and forms dimers (Richter et al, 2013a). UV-crosslinking of the d-1 repeat substrate to wild-type Cas6b followed by mass spectrometric analysis identified a methionine residue at position 185 that interacts with an UUGC sequence motif of the repeat RNA. Interestingly, analysis of the previously introduced Cas6b structural model (Richter et al, 2012), shows that the methionine residue is at the opposite site of the already identified catalytic residues, indicating two distinct domains for cleavage and substrate binding.

An approach to generate binding deficient Cas6b mutants by site-directed mutagenesis often resulted in insoluble proteins (Tab. 1), especially for mutants in the C-terminal glycine rich loop proposed to be important for substrate binding (Haurwitz et al, 2010). However, the mutation of the methionine 185 residue yielded soluble protein. Determination of K_d values for the wild-type and the M185A mutant of Cas6b were performed by standard EMSAs (Fig. 3). Binding assays for both Cas6b versions show a gradually inclining amount of shifted substrate with increasing amounts of Cas6b, yet the shifts for wild-type Cas6b appear already at concentrations as low as 50 nM (K_d of 668 nM ± 29 nM) (Fig. 3a, b) while over 250 nM are needed to cause a significant RNA substrate shift with the M185A mutant (K_d = 1.03 µM ± 0.3 µM) (Fig. 3c, d). Scatchard linearization of the quantified data (small graph inserted in Fig. 3b, c) was employed to yield K_D values of both proteins. An approximately two times higher affinity of Cas6b wild-type to its RNA substrate is displayed compared to the mutant, confirming the importance of this residue for substrate binding.

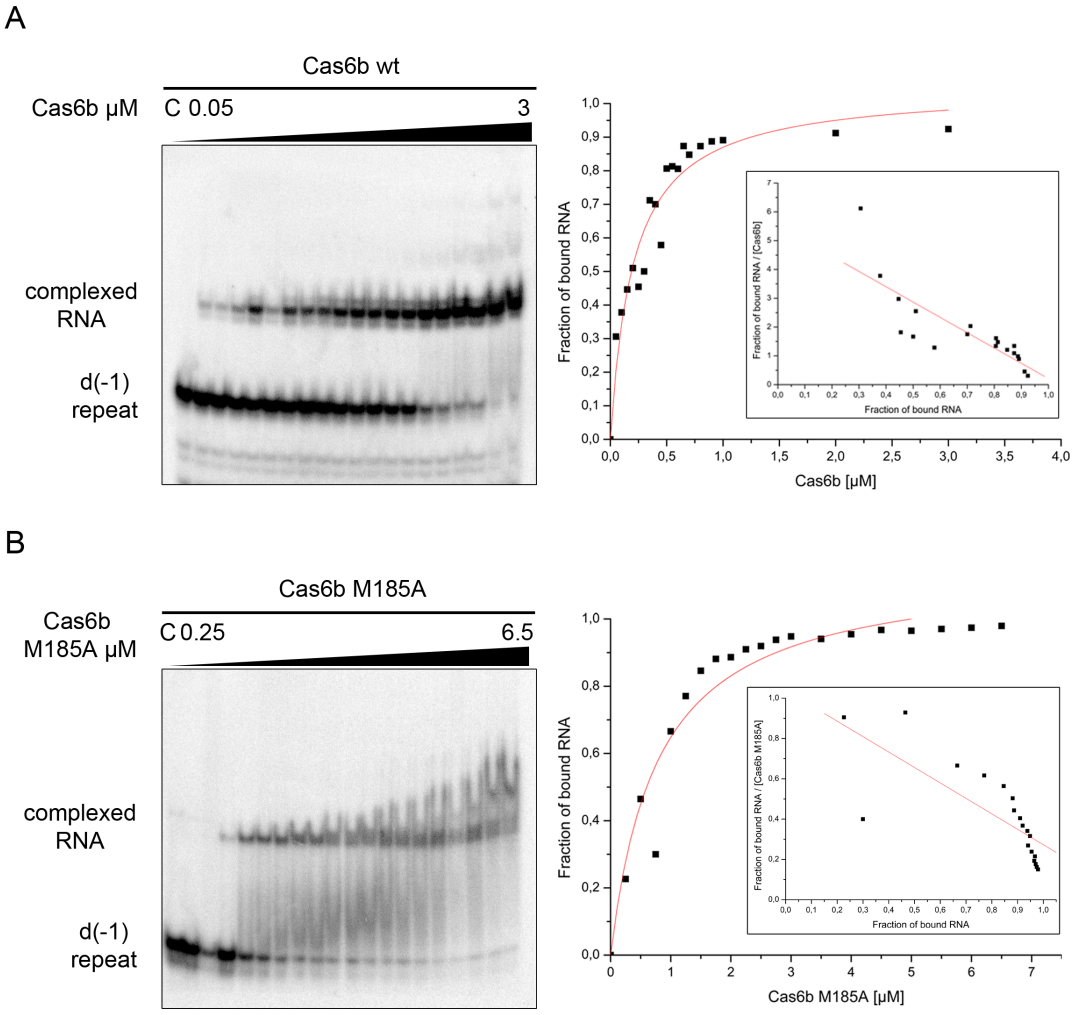


Figure 3. A methionine residue plays an important role in substrate binding of Cas6b. EMSAs using 5' radiolabelled deoxy repeat RNA incubated with indicated Cas6b variants and concentrations were used to determine the substrate binding affinity. (**A**) Shifting of the substrate starts at Cas6b

concentrations of 50 nM and a complete shift of the substrate into a Cas6b (wt):RNA complex is reached at a concentration of 2.5 µM. At higher Cas6b concentrations the reported Cas6b dimerization (Richter et al, 2013a) was observed to begin. (**B, D**) Quantified data was plotted and fitted with an hyperbolic curve. Linearization was achieved with a Scatchard plot and the K_d, as the negative reciprocal of the slope, was determined in three independent assays with $K_{d,wt}$ = 668 nM ± 29 nM and $K_{d,M185A}$ = 1.03 µM ± 0.3 µM. (**C**) Higher concentrations of the M185A Cas6b mutant are needed to initiate the shift of the RNA substrate (250 nM). Full substrate shift into Cas6b (M185A):RNA complex is achieved with comparably high concentrations of 6.5 µM.

The maturation of crRNA by Cas6b tolerates changes in repeat sequence and structures

Bioinformatical analysis of different 37 nt long subtype I-B repeats using sequence alignments (ClustalW2, (Larkin et al, 2007)) and sequence logo generation (WebLogo, (Crooks et al, 2004; Schneider & Stephens, 1990)) revealed several conserved nucleotides, predominantly located in the processing region that yields the AUUGAAAC 8 nt 5' terminal tag of the mature crRNA (Fig. 4a). The 5' region of the repeat shows a lower conversion suggesting that it is less important for Cas6b binding or cleavage.

Several single nucleotide mutations were introduced into the *M. maripaludis* C5 repeat sequence to evaluate the importance of the repeat sequence and structure for Cas6b processing (Fig. 4b,c). A mutation that is expected to resolve the hairpin structure (G16C) showed a lower substrate conversion rate compared to the wild-type, indicating a possible influence of repeat structure on Cas6b recognition. However, restoring the hairpin structure by repairing the Watson-Crick base pairs (G16C/C25G) resulted in a complete loss of processing by Cas6b. Two UUGC motifs are found within the repeat sequence. One of these motifs is found to be bound by the methionine residue 185 mentioned above. Individually changing each of the two UUGC motifs into UAGC, to analyse their effect on processing, resulted in cleavable substrates. Furthermore the mutation of uridine 15 to adenine resolves the hairpin structure and confirms the lack of structural influences on the processing by Cas6b. Interestingly, *M. maripaludis* Cas6b was able to cleave a wild-type repeat RNA of *C. thermocellum* with slightly reduced activity even though this repeat displays 12 nt changes and an altered structure (Fig. 4 c/d). The CRISPRmap tool (Lange et al, 2013) places both repeat sequences into superfamily E, however, both repeats have a completely divergent predicted structure (Fig. 4c) indicating an influence of the sequence during processing rather than an influence of the structure.

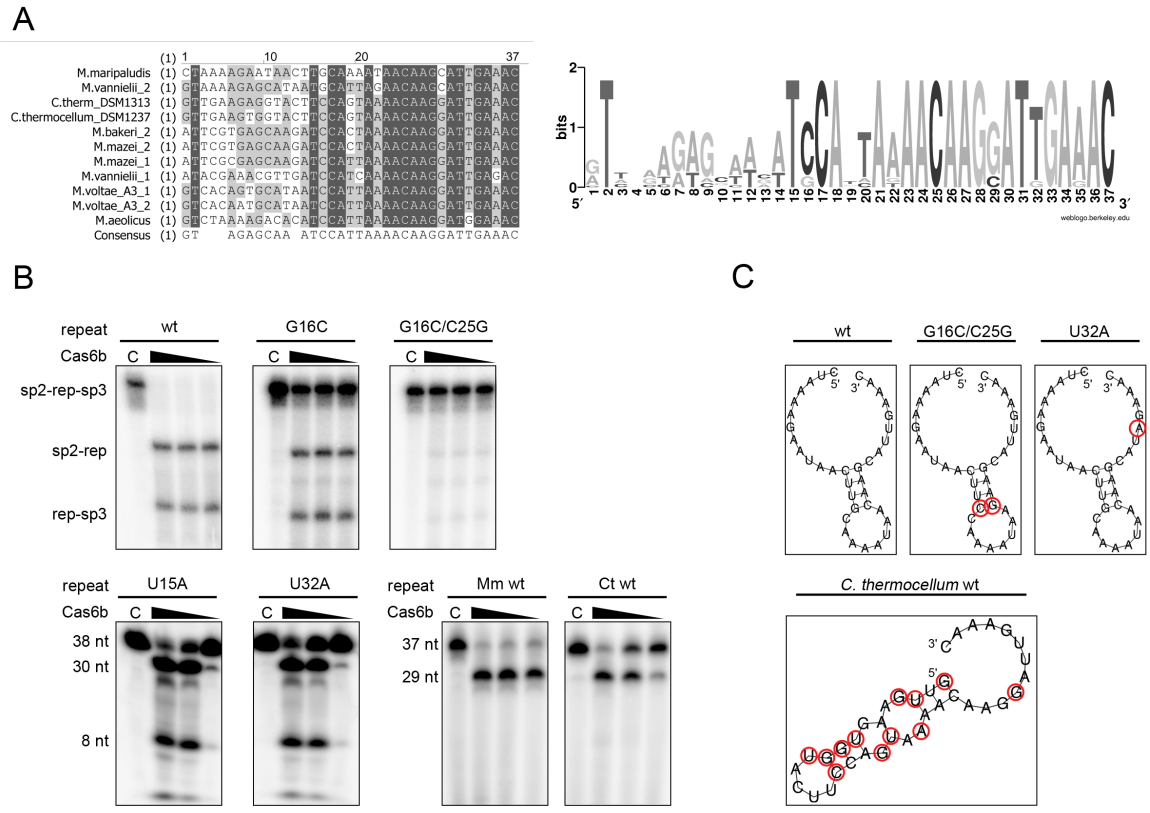


Figure 4. The influence of the repeat sequence on crRNA maturation by Cas6b. (**A**) A sequence alignment (ClustalW2) (Larkin et al, 2007) and a sequence logo (WebLogo) (Crooks et al, 2004; Schneider & Stephens, 1990) of different 37 nt long subtype I-B repeats show a highly conserved 3' region while the 5' regions are less conserved between these repeats. (**B**) In endonuclease assays incubating different internally labelled RNA substrates representing several repeat variants (G16C, G16C/C25G, U15A and U32A) as well as 5' end labelled *M. maripaludis* and *C. thermocellum* wild-type (Mm wt and Ct wt) repeat RNAs were incubated with different concentrations of *M. maripaludis* Cas6b (10 µM, 1 µM and 0.1 µM). While the G16C/C25G repeat substrate is not processed anymore, all other used substrates are still cleaved by Cas6b indicating a sequence influence rather than a structural influence during processing. (**C**) Structure predictions for the different repeat variants were created with RNAfold (McCaskill, 1990; Schuster et al, 1994; Zuker & Stiegler, 1981). (**D**) A list of the different used repeat sequences. Altered nucleotides compared to *M. maripaludis* wild-type (Mmar wt) repeat are marked in red.

Discussion

Structural and biochemical characterization studies are available for several members of the diverse family of Cas6 endonucleases present in the different CRISPR-Cas subtypes (Carte et al, 2010; Carte et al, 2008; Gesner et al, 2011; Haurwitz et al, 2010; Haurwitz et al, 2012; Nickel et al, 2013; Reeks et al, 2013c; Richter et al, 2013a; Richter et al, 2012; Sashital et al, 2011; Sternberg et al, 2012; Wang & Li, 2012; Wang et al, 2011; Wang et al, 2012). Canonical histidines (e.g. found in Cas6f, Cas6e and Pf Cas6) are a key feature of Cas6 catalytic sites, as the proposed general acid/base mechanism requires protonation of the leaving group by the histidine (Carte et al, 2008). Previous studies identified two histidine residues in the catalytic site of Cas6b (Nickel et al, 2013; Richter et al, 2012). The only known exception is a Cas6 homolog of *S. solfataricus* (Sso Cas6) that lacks the canonical histidine residiue (Reeks et al, 2013c). However, to stabilize the conformation during the reaction, more residues are required. In Pf Cas6 a tyrosine and lysine were reported to complete the active center of the enzyme (Carte et al, 2010; Wang et al, 2011). In this study, these residues were also identified for Cas6b. Although the composition of the Pf Cas6 and Cas6b catalytic sites appear similar, the order of the residues is reversed in Cas6b and two histidines seem to complement for each other. Therefore this is indicative for a distinct catalytic site and strengthens the hypothesis that Cas6 enzymes can be used to classify their respective subtypes (Richter et al, 2012). Until now, all the characterized Cas6 enzymes of different subtypes are reported to have different catalytic sites, which might also correlate with the respective mode of processing, e.g. the wrap-around mechanism (Wang et al, 2011), dimerization upon processing (Richter et al, 2013a) or the recognition of repeat hairpin structures (Gesner et al, 2011; Haurwitz et al, 2010; Sashital et al, 2011; Sternberg et al, 2012).

We showed that Cas6b of *M. maripaludis* is a single-turnover enzyme, which is in good agreement with other studies, reporting single-turnover reactions for Cas6f (Haurwitz et al, 2010), Cas6e (Sashital et al, 2011) and Sso Cas6 (Reeks et al, 2013c). However, the catalytic activity of Cas6b (0.296 min^{-1}) is about 20-fold lower compared to observed reaction rates of Cas6e (4.9 min^{-1}) (Sashital et al, 2011) or Cas6f (3.8 min^{-1}) (Haurwitz et al, 2010; Sternberg et al, 2012). Similar reaction rates to Cas6b were reported for the dimeric version of *S. solfataricus* (Sso) Cas6 (0.8 min^{-1}), while the monomeric version shows about 10-fold lower conversion rates (0.034 min^{-1}) (Reeks et al, 2013c). Cas6b of *M. maripaludis* has also been reported to form dimers at higher concentrations when supplied with non-hydrolysable pre-crRNA substrate (Richter et al, 2013a).

To be in a measurable window of conversion rate, only very low amounts of Cas6b could be used for the k_{obs} determination. These low concentrations, however, might not have

promoted the previously reported Cas6b dimerization (Richter et al, 2013a). Assuming that the monomeric form of Cas6b is also active, this is a possible explanation for the observed lower conversion rate.

With a substrate affinity of approximately 668 nM, the binding affinity of Cas6b to its substrate is significantly lower compared to other Cas6 homologs as the K_d values of Cas6f and Cas6e are reported with 50 pM and 3.6 nM, respectively (Sashital et al, 2011; Sternberg et al, 2012). One explanation could be a different mode of binding, as both Cas6e and Cas6f specifically recognize a RNA hairpin structure of their respective repeats followed by fast reaction rates (Gesner et al, 2011; Haurwitz et al, 2010; Sashital et al, 2011; Sternberg et al, 2012). Unfortunately, no kinetic data exists for Pf Cas6 and other Cas6 members close to Cas6b that do not rely on the recognition of a secondary repeat structure. The recognition of a hairpin structure of a repeat and the associated fast coordination of the cleavage site into the active site could be faster compared to the proposed wrap-around mechanism for unstructured pre-crRNAs. This mechanism (Wang et al, 2011) based on sequence rather then structure of the repeat may need more time for scanning of putative binding sites as well as for coordination of the cleavage site into the catalytic center of the Cas6 enzyme. The wrap-around theory is also supported by the distal loci for binding and cleavage within the protein, which was reported for Pf Cas6 (Carte et al, 2010; Wang et al, 2011). In agreement with this, a methionine residue at position 185, important for binding, is at the opposite site of the active center of Cas6b and wrapping of the repeat cleavage site into the active site might require additional fine-tuning. As Cas6f, for example, specifically recognizes the structure of the repeat and cleaves at the stem of the hairpin (Haurwitz et al, 2010; Sternberg et al, 2012), the wrap-around hypothesis also gives reasoning for the lower conversion rates observed for Cas6b.

While binding of the substrate always includes several amino acids (e.g. arginine, glycine) it seems that binding of a uridine base is important during recognition and the involved amino acids differ in the various Cas6 enzymes (Gesner et al, 2011; Wang et al, 2011). Cas6e of *T. thermophilus* for example employs an arginine residue for binding (Gesner et al, 2011), while in Cas6b the methionine residue binds the uridine. Mutational studies of repeats associated with Cas6e and Pf Cas6 revealed that the loss of this uridine base severely affects binding by Cas6 indicating a stronger influence of the repeat sequence (Carte et al, 2010; Sashital et al, 2011; Wang et al, 2011). The importance of individual nucleotides within the repeat were also shown for Cas6f binding (Sternberg et al, 2012). Here, the 5th nucleotide is crucial for binding as it marks the beginning of the hairpin structure. Further analysis, including mutations within the hairpin, proved the importance of the hairpin for Cas6f processing. In addition, single nucleotides within the stem-loop region that do not affect the repeat structure were important for recognition (Sternberg et al, 2012). Mutation of single nucleotides showed

similar effects on Cas6b cleavage of repeats, however the crucial influence of the structure for both binding and cleavage was not identified. Cas6b was able to cleave a *C. thermocellum* repeat substrate from a bacterial subtype I-B with a distinct predicted structure. Cas6 enzymes evolved two different binding sites for recognition of their substrates, one site specifically binding to the hairpin and the second site recognizing a 5' terminal segment preceeding the hairpin (Niewoehner et al, 2013). While some Cas6 variants seem to have both binding sites, studies of Pf Cas6 show that repeat binding occurs 12 nt upstream of the processing site indicating that only one of the two binding motifs is apparent in this Cas6 (Carte et al, 2010; Wang et al, 2011). The performed Cas6b studies confirm this result, as the uridine residue identified by UV–crosslinking experiments is located in this proposed binding area for Cas6 and no other binding was identified. This residue is highly conserved among subtype I-B repeats, which indicates its importance. A specific length of the RNA sequence is needed to direct the repeat into the catalytic site of the enzyme. In agreement, the distance of the identified binding site to the point of processing within the repeat matches the proposed wrap-around mechanism (Wang et al, 2011).

In conclusion, the data provides evidence for the apparent similarity of Pf Cas6 and subtype I-B archaeal Cas6b. The repeat recognition and cleavage activities are proposed to employ very similar mechanisms and both active centers feature similar yet reversed catalytic triads.

Chapter V

Proteolytic cleavage and nucleic acid binding properties of the CRISPR-Cas I-B subtype-specific protein Cas8b

Hagen Richter[1], Sabine Mohr[1], Judith Zöphel[1], Laura Penkert[1] and Lennart Randau[1, 2, *]

[1] Max-Planck-Institute for Terrestrial Microbiology, Karl-von-Frisch Straße 10, 35043 Marburg
[2] LOEWE Center for Synthetic Microbiology (Synmikro), 35043 Marburg
[*] Corresponding author

Abstract

CRISPR arrays are transcribed into long precursor RNA species, which are further processed into mature, interfering RNAs (crRNAs). Cas proteins utilize these RNAs, which can contain spacer sequences derived from viral genomes, to mediate immunity during a reoccurring viral attack (Barrangou et al, 2007; Brouns et al, 2008). Different CRISPR-Cas types are defined by the presence of different interference complexes in which a large subunit plays a major role for target binding mediated by crRNA base complementarity (Makarova et al, 2011a; Makarova et al, 2011b). We biochemically analysed the proposed large Cascade subunit Cas8b of the subtype I-B CRISPR-Cas system found in *Methanococcus maripaludis* C5 and *Clostridium thermocellum*. Evaluation of the two Cas8b proteins revealed a putative proteolytic cleavage site within the protein resulting in the creation of two protein fragments that are proposed to represent the large and small Cascade subunit of subtype I-B. Finally, we demonstrate unspecific nucleic acid binding by Cas8b and the absence of endonucleolytic cleavage activity.

Introduction

The prokaryotic immune system CRISPR-Cas, composed of arrays of clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR associated (Cas) genes, utilizes base complementarity between small interfering RNAs (crRNAs) and target nucleic acids to identify and degrade foreign DNA or RNA. Repetitive DNA sequences (repeat) of CRISPR arrays are interspersed by unique spacer sequences that can be derived from genomes of infecting viruses. These spacers are a part of the crRNAs and confer the necessary base complementarity with the target (Barrangou et al, 2007; Bolotin et al, 2005; Brouns et al, 2008; Horvath & Barrangou, 2010; Mojica et al, 2009; Mojica et al, 2005; Sorek et al, 2008; Terns & Terns, 2011; van der Oost et al, 2009). Mechanistically, the CRISPR-Cas immunity can be described in three phases: i) acquisition of new spacer sequences (Datsenko et al, 2012; Erdmann & Garrett, 2012; Plagens et al, 2012; Swarts et al, 2012; Yosef et al, 2012), ii) maturation of crRNAs (Carte et al, 2008; Gesner et al, 2011; Haurwitz et al, 2010; Richter et al, 2012; Sashital et al, 2011) and iii) interference with the foreign nucleic acids during a reoccurring attack by a ribonucleoprotein complex of Cas proteins and mature crRNA (Cady & O'Toole, 2011; Hale et al, 2009; Jore et al, 2011b; Lintner et al, 2011b; Sashital et al, 2012; Shao et al, 2013; Wiedenheft et al, 2011a; Wiedenheft et al, 2011b; Zhang et al, 2012).

Based on the three different proteins Cas3, Cas9 and Cas10, responsible for the degradation of the target nucleic acid, the CRISPR-Cas systems were classified into three major types.

Grouping of individual Cas proteins led to a classification into at least 10 additional subtypes (Makarova et al, 2011b). The Type-I subsystem I-B can e.g. be found in methanogens, halophiles and Clostridia. In the archaeon *Methanococcus maripaludis* C5 and in the bacterial organism *Clostridium thermocellum* a minimalistic I-B system is present, with two operons coding for the putative adaption machinery Cascis (CRISPR associated complex for integration of spacer) (Plagens et al, 2012) and the interference complex Cascade (CRISPR associated complex for antiviral defense) (Richter et al, 2013a; Richter et al, 2012). This system comprises the universal Cas proteins Cas1, Cas2 and Cas4, the interference proteins Cas3, Cas5, Cas7 and the subtype-specific proteins Cas8b and Cas6b (Richter et al, 2012). The large Cascade subunits, represented by Cas8 proteins are some of the most diverse components of CRISPR-Cas systems and show only low similarities in their proposed Zn-finger and polymerase like domains (e.g. thumb and palm). Cas8 proteins are predicted to play a major role during the interference step (Makarova et al, 2011a). This stage is the major difference between the three types and reflects the plethora of different CRISPR-Cas systems best (Makarova et al, 2011b). A reasonable explanation for this diversification is the co-evolution of viruses and their hosts, which led to the development of a variety of measures by viruses to penetrate their hosts, e.g. anti-CRISPR proteins (Bondy-Denomy et al, 2013).

Type-I systems are found in both, bacteria and archaea, and employ an interference system in which Cas3 is the key player. Cas3 contains an HD phosphohydrolase domain and a DExH-like helicase domain and subsequent to the recruitment by Cascade unwinds dsDNA and degrades the resulting ssDNA in an ATP and Mg^{2+} dependent manner (Beloglazova et al, 2011; Makarova et al, 2011a; Makarova et al, 2011b; Mulepati & Bailey, 2011; Sinkunas et al, 2011). Target recognition is mediated by base pairing of the crRNA with the complementary strand and results in a displacement of the non-complementary strand yielding a Cas3 accessible structure called R-loop (Jore et al, 2011b; Lintner et al, 2011b). To avoid self-targeting of genomic DNA, Cascade is able to discriminate between self- and non-self targets using two specific sequence motifs. The first motif is the 5' terminal repeat tag of mature crRNA, generated by Cas6 family enzymes (Carte et al, 2008; Gesner et al, 2011; Haurwitz et al, 2010; Richter et al, 2012; Sashital et al, 2011) and the second motif is represented by the protospacer adjacent motif (PAM), which is short sequence motif that can be located upstream or downstream of a protospacer sequence. Both crRNA and PAM sequences must not base pair to ensure discrimination between self- and non-self. Additionally, the 2 – 5 nt long PAM sequence serves as motif for Cascade binding (Marraffini & Sontheimer, 2010b; Sashital et al, 2012; Westra et al, 2013). Subsequently, a helical destabilization enables the invasion of the matching seed sequence of the crRNA. Seed sequences are described as the first 8 -10 nt of a spacer sequence that are necessary for

successful recognition of the foreign DNA (Semenova et al, 2011). Type I-E Cascade of *E. coli* is the best studied CRISPR interference complex (Brouns et al, 2008; Jore et al, 2011b; Wiedenheft et al, 2011a; Wiedenheft et al, 2011b), has a size of 405 kDa and is composed of the subunits, Cas6e, Cse1, Cse2, Cas7 and Cas5. While Cas6e processes precursor crRNA, Cas7 and Cas5 are tightly bound to the crRNA to protect its degradation. The conservation of Cas5, Cas7 and the helicase/nuclease Cas3 suggests a general mechanism of interference in Type-I systems, which is supported by studies of the related I-A system with a similar Cas7/Cas5 crRNA binding platform (Lintner et al, 2011b; Plagens et al, 2012). The large Cascade subunit Cse1 of subtype I-E and the small subunit Cse2 tightly bind their DNA target (Jore et al, 2011b; Mulepati et al, 2012; Sinkunas et al, 2013). However, the large subunit Cse1 is dispensable *in vitro* and the resulting sub-complex is able to specifically bind ssDNA and dsDNA. The addition of Cse1 enhances target localization, yet the binding becomes unspecific (Jore et al, 2011b). Strutural studies of Cse1 of *Thermus thermophilus* indicate an involvement during PAM recognition. In binding studies performed with this Cse1 variant a loop structure displayed a high specificity towards PAM sequences. (Sashital et al, 2012). A study of type I-F Cascade of *Pseudomonas aeruginosa* showed a similar assembly of the interference complex compared to the one described for I-E (Wiedenheft et al, 2011a). However, I-F systems lack a small Cascade subunit and only a large subunit is reported (Makarova et al, 2011a). Similarly to I-E, Csy1 (the I-F large subunit) and Csy2 (Cas5) form a heterodimer, which seems to be involved in target recognition and is also located at the periphery of the Cascade (Wiedenheft et al, 2011a). Type I systems display a broad variety regarding their large Cascade subunit. The subtypes I-A and I-E employ two proteins designated as small and large subunit, while the subtypes I-B, I-C and I-F have one gene coding for a protein that would be required to combine the functions of the small and large Cascade subunit. An exception is apparent in subtype I-D, where a Cas10 homolog is found, which additionally to the Cas8 features has a HD nuclease domain that seems to be reminiscent to the Type-III systems (Makarova et al, 2011a). Although Type-III systems also employ a large ribonucleoprotein complex for target interference (Hale et al, 2009; Osawa et al, 2013; Shao et al, 2013; Zhang et al, 2012), no genes coding for a small and large subunit are found in Type-III systems, which are thought to be the progenitor of Type-I systems (Makarova et al, 2011a; Makarova et al, 2011b). As Type-II systems only rely on one large multifunctional protein (Cas9) for interference and no large or small subunit of an interference complex are found (Chylinski et al, 2013; Deltcheva et al, 2011; Jinek et al, 2012), it seems that the presence of a small subunit is special for Type-I systems. The different large subunits of Type-I systems (Cse1, Csy1, Cas10d and Cas8) do not share high sequence similarity, but with an RRM fold and the putative Palm and Thumb domains, they have some features in common, which seem to correlate with their function in target binding and

recognition (Makarova et al, 2011a). The subtype I-B large subunit of *Methanothermobacter thermautotrophicus* (Nar71) has been analysed biochemically (Guy et al, 2004). Nar71 was proposed to be a novel nuclease-ATPase of a putative thermophilic DNA repair system, as the CRISPR-Cas immune system was not yet described. The protein was shown to bind Holliday junctions, flayed and flapped DNA structures. Furthermore, a strand specific nuclease activity towards 3' flapped structures was observed in the presence of ATP. The addition of ATP also resulted in DNA unwinding activities (Guy et al, 2004). Additional information is needed to evaluate the possible conserved functions of the different large Cascade subunits (e.g. Cas8b, Cas8a, Cas8c).

In this study we present the recombinant production of the subtype I-B specific protein Cas8b of *M. maripaludis* C5. We demonstrate that the purification of two different Cas8b proteins of the bacterial representative *C. thermocellum* and the archaeon *M. maripaludis* C5 results in two fragments. The similar sizes of the fragments would correlate with a proteolytic cleavage into the large and small Cascade subunit comparable to Cse1 and Cse2 of *E. coli* Cascade. Edman sequencing of the smaller fragment identified the cleavage site within the full length Cas8b between an asparagine and a methionine residue. The initial biochemical characterization of Cas8b of *M. thermautotrophicus*, Nar71, showed nuclease and helicase functions of the protein (Guy et al, 2004). Experiments using the purified Cas8b of *M. maripaludis* C5 and CRISPR-Cas adapted flayed and flapped substrates (DNA/RNA) as well as ssDNA, dsDNA and ssRNA substrates, did not display the reported cleavage activity. However, the protein exhibits nucleic acid binding activity, which is in agreement with the putative role within Cascade.

Material and Methods

Growth of *E. coli*

All *E. coli* cells used for cloning and expression were grown in LB–medium with the appropriate antibiotics at 37 °C shaking at 200 rpm.

Production of Cas8b and mutants

The *cas8b* genes MmarC5_0771 (Cas8b) and Cthe_3201 (Ct Cas8b) were amplified from genomic DNA isolated from *Methanococcus maripaludis* C5 or *Clostridium thermocellum* (ATCC 27405), respectively. To yield a C-terminal His-tag the genes were cloned into pET-20b vector, while MmarC5_0771 additionally was cloned in to pET-Duet1 to facilitate expression with an N-terminal His-tag. A sequence alignment of various Cas8b of different organisms was prepared with VectorNTI (Invitrogen) using the ClustalW2 algorithm (Larkin et al, 2007). Site-directed mutagenesis to generate *cas8b* mutants was performed according to the manufacturer's protocol of QuikChange site-directed mutagenesis (Stratagene). QuikChange primers were designed with the QuikChange Primer Design tool (Agilent) and synthesized by MWG Eurofins. Mutations were confirmed by sequencing (MWG).
Expression of the Cas8b variants was facilitated in *E. coli* (Rosetta2 DE3) cells by induction of protein expression using a final concentration of 0.5 mM isopropylthio-β-D-galactoside (IPTG) after the cells reached an OD_{578} of 0.6. Cells were harvested after 4 hours and the pelleted cells were resuspended in 5 ml lysis buffer (10 mM Tris-HCl [pH 8.0]; 300 mM NaCl; 10 % glycerol and 0.5 mM DTT) per 1 g cells. The cells were lysed with lysozyme (1 mg/g cells) incubated on ice for 30 min before being disrupted by sonication (8 x 30 sec, Branson Sonifier 250). The lysate was cleared by centrifugation (47800 g, 30 min, 4 °C) and supernatant was applied to a Ni-NTA-Sepharose column (GE-Healthcare) and purified using a FPLC Äkta-Purification system (GE-Healthcare). Elution of the column bound protein was achieved by a linear imidazole gradient (0 – 500 mM). SDS-PAGE and coomassie blue staining determined the purity of proteins. Pure protein samples were pooled and dialyzed into lysis buffer containing only 100 mM NaCl for further purification steps. In additional steps Cas8b of *M. maripaludis* C5 was purified using a cation exchange column (Heparin-Sepharose, GE-Healthcare) and a gradient of NaCl (0 – 1000 mM). SDS-PAGE and coomassie blue staining verified the purity of the eluted protein before pooled samples were subjected to dialysis into lysis buffer. Final purification was performed by size-exclusion chromatography applying the protein sample to an analytical Superdex column (200 10/300 GL, GE-Healthcare). The molecular weight of the collected fractions was determined by

comparison to a calibration curve of the column with molecular weight markers (Kit for Molecular Weights, Sigma-Aldrich). The final protein concentration was determined by Bradford assay (BioRad).

Generation of DNA and RNA substrates

A repeat variant with a deoxynucleotide substitution at the 29th nucleotide to prevent processing by Cas6b was synthesized by MWG. DNA oligonucleotides used to generate flapped, flayed and R-Loop like structures were manually designed to have complementary sequences and were synthesized by MWG. Cas6b processing of unlabeled spacer1 - repeat - spacer2 precursor crRNA yielded mature crRNA2 as described previously (Richter et al, 2013a).

The indicated substrates were 5' end labelled using T4 polynucleotide kinase (Ambion) and [γ-^{32}P] ATP (5000 Ci/mmol) in a total volume of 10 µl for synthesized oligonucleotides (10 pmol RNA/DNA, 1 µl PNK buffer (NEB), 25 U T4 PNK (Ambion)) and 20 µl for crRNA2 (15 µl RNA, 2 µl PNK buffer, 25 U T4 PNK). Labelled oligonucleotides and crRNA2 were separated using a 10 % denaturing polyacrylamide gel (8 M Urea, 1 x TBE) running in 1 x TBE at 10 W for 1.5 hours. After visualization by autoradiography bands were cut out accordingly and DNA and RNA species were eluted in 500 µl elution buffer (250 mM NaOAc; 20 mM Tris-HCl [pH 7.5]; 1 mM EDTA [pH 8.0]; 0.25 % SDS) over night shaking on ice. After addition of 2 Vol 100 % ice cold EtOH and 1/100 glycogen (Roche) the DNA and RNA species were precipitated for 1 h at -20 °C, pelleted and subsequently washed with 70 % EtOH.

Flapped, flayed and R-Loop structures were generated by hybridization of radiolabelled oligonucleotides with their respective unlabelled counterparts. The corresponding oligonucleotides were mixed with a slight excess of unlabelled oligonucleotides to assure complete hybridization of the labelled species. The final reaction volume of 10 µl was heated to 95 °C for 5 min followed by a 2 hour cooling step to facilitate proper hybridization of complementary sequences.

Endonuclease Assay

Endonucleolytic cleavage assays employed varying indicated Cas8b concentrations, which were mixed with the specified radiolabelled substrates. Reactions without ATP (20 mM HEPES-KOH [pH 8.0]; 1 mM DTT; 5 mM $MgCl_2$) and with ATP (20 mM HEPES-KOH [pH 8.0]; 1 mM DTT; 5 mM $MgCl_2$; 10 mM ATP) were brought to 10 µl final volume and incubated at 37 °C for 30 min. To define a possible cleavage of a single strand within the flapped or flayed substrates, only one strand that was used in the hybridization reaction was 5'

endlabelled. To analyse each individual strand for potential cleavage, the labelled strand was changed in different hybridization reactions. Separation of DNA species was achieved by a denaturing polyacrylamide gel (8 M Urea, 1 x TBE, 15 % acrylamide) followed by visualization with autoradiography.
In order to yield mature crRNA for 5' end labelling, an unlabelled spacer1 - repeat - spacer2 precursor RNA was processed by Cas6b. Full processing of the substrate was achieved by a reaction time of 1 h with 15 µM Cas6b (250 mM KCl; 20 mM HEPES-KOH [pH 8.0]; 1.875 mM $MgCl_2$; 1 mM DTT).

Electrophoretic mobility shift assay (EMSA)

The indicated concentrations of Cas8b were mixed in binding buffer (10 mM Tris-HCl [pH 8.0]; 200 mM KCl; 5 % glycerol; 0.5 mM DTT; 0.5 mM EDTA [pH 8.0]; 1 µg BSA) with the particular substrate and incubated at 37 °C for 1 h to facilitate binding. The reactions were immediately mixed with 6 x DNA loading dye (4 g sucrose; 2.5 mg bromophenol blue; 2.5 mg xylene cyanol in 10 ml H_2O) and applied to a native polyacrylamide gel (1 x TBE, 7 % acrylamide) running in 1 x TBE at 7 W for 3 hours. Visualization was achieved via autoradiography.

Results

Recombinant Cas8b is purified as two fragments

The genes of *M. maripaludis* C5 Cas8b (MmarC5_0771, Mm Cas8b) and *C. thermocellum* Cas8b (Cthe_3201, Ct Cas8b) were cloned into pET-20b vector to yield a C-terminal His-tag. Addtionally, Mm Cas8b was cloned into pET-Duet1 vector to obtain a N-terminal His-tag. The respective Cas8b variants were produced and subsequently purified via Ni-NTA affinity chromatography (Fig. 1a), which always resulted in elution fractions containing two protein fragments of approximately 60 kDa and 15 kDa size for the C-terminal tagged Cas8b. Further purification steps, including cation exchange chromatography and size-exclusion chromatography (Fig. 1c) generated cleaner protein fractions. The elution profile of the size-exclusion chromatography run showed one peak with an estimated size corresponding to the calculated size of full-length Cas8b (theoretical MW: 76.6 kDa, Fig. 1c). However, two bands were observed by SDS-PAGE (Fig. 1a, right panel). To prevent proteolytic cleavage, expression tests (e.g. using minimal medium or protein purification with supplemented protease inhibitors) were performed but did not change the protein fragmentation pattern (data not shown). In addition, Ct Cas8b expression and purification was analysed as well and indicated a nearly identical cleavage pattern during SDS-PAGE, with two bands of approximately 60 kDa and 15 kDa. Taken together this fits with the calculated full-length Ct Cas8b (theoretical MW: 71.5 kDa, Fig. 1b). To exclude, that the small band observed during SDS-PAGE is enriched due to the C-terminal His-tag, the tag was cloned to the N-terminal end, which would then result in a loss of the small band. Again two fragments were observed, but the smaller band was reduced in size likely due to the missing His-tag (Fig. 1b). Mass spectrometric analysis of the small fragment of the C-terminal tagged Mm Cas8b purification identified peptides of the C-terminal end of Cas8b, further indicating that the obtained results for Cas8b purification are not displaying artifacts of the purification procedure. An alignment (Fig. 1d) of different Cas8b variants of various organisms (ClustalW2, (Larkin et al, 2007)) reveals the predicted Zn-finger domain (Makarova et al, 2011a). Cas8b shows only very low sequence conservation. However, the two CxxC-motifs (Fig. 1d, upper part) are highly conserved and the entire C-terminal end of the alignment (Fig. 1d, lower part) shows a higher sequence similarity indicative for an important function of this protein region. The calculated size of this C-terminal part (16.4 kDa) of the protein correlates with the examined size of the small fragment of the Cas8b purifications (Fig. 1a,b). An N-terminal Edman-Sequencing (in collaboration with AG Lochnit, Gießen) of 10 amino acid residues of the small fragment (boxed sequence in the bottom alignment of Fig. 1d)

additionally determined the putative cleavage site within the full-length Cas8b between an asparagine residue and a methionine residue.

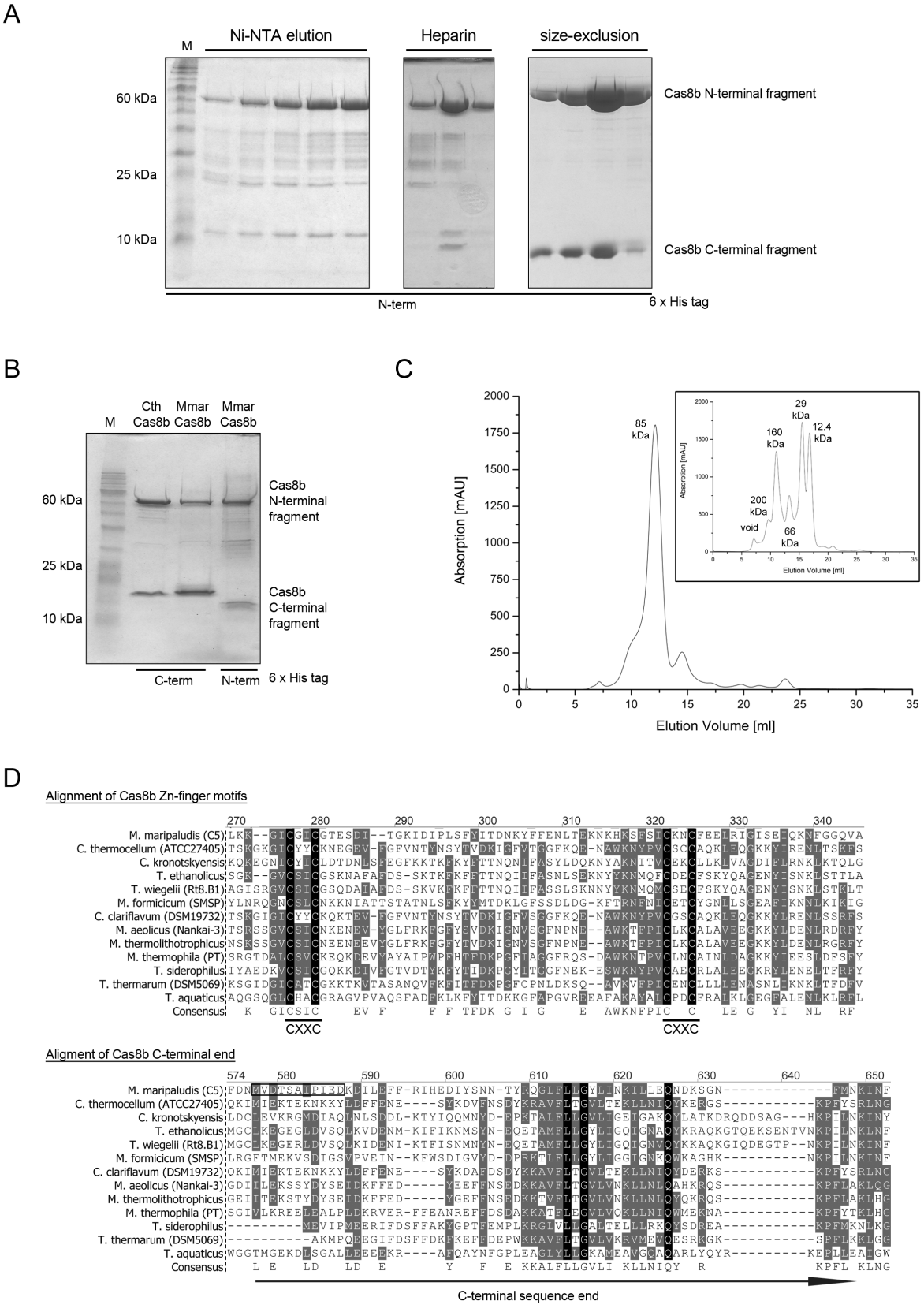


Figure 1. Purification of *M. maripaludis* (Mm) and *C. thermocellum* (Ct) Cas8b. (**A**) SDS-PAGE of Cas8b elution fractions with the different used chromatography procedures (Ni-NTA affinity, cation exchange (Heparin) and size-exclusion) indicated. Cas8b with a N-terminal tag His-tag always purifies in two fragments of roughly 60 kDa and 12 kDa (theoretical MW: 76.6 kDa) (**B**) SDS-PAGE of protein purifications of different recombinant Mm Cas8b (either with N- or C-terminal His-tag) and a recombinant Ct Cas8b with C-terminal His-tag. The three variants show purification of two fragments of 60 kDa and approximately 15 kDa for the C-terminal version and 60 kDa and roughly 12 kDa fragments for the N-terminal tagged version. (**C**) Size-exclusion elution profile of N-terminal tagged Mm Cas8b (main graph) and calibration of the size-exclusion column (small boxed graph). (**D**) ClustalW2 sequence alignment of Cas8b representatives (Larkin et al, 2007). The top part shows the putative Zn-finger domain with two CXXC motifs and the lower part depicts the C-terminal part showing a higher similarity compared to the overall sequence identity. Amino acids that were identified via Edman-Sequencing of the Mm Cas8b C-terminal fragment are displayed by a box within the alignment. Conserved residues are shown with black background while grey background highlights similar amino acid residues.

Cas8b shows no endonucleolytic cleavage activity

Studies with another Cas8b homolog (Nar71) of *M. thermautotrophicus* revealed a structure specific cleavage activity of the protein correlating with a postulated function in a novel kind of DNA repair system (Guy et al, 2004). With the knowledge of the association of Nar71 with a CRISPR-Cas system, this putative function was re-investigated using Mm Cas8b and substrates designed to mimic structures that are predicted to appear during CRISPR-Cas immunity (flapped and flayed DNA structures). Based on the experimental setup of the Nar71 studies (Guy et al, 2004), we were not able to confirm the reported cleavage activity for Cas8b in reactions that were carried out either with supplemented or without supplemented ATP. The DNA strands (Fig. 2a) that were used to generate the structures, were not processed (Fig. 2b) by Cas8b. Further assays with single stranded RNA and DNA also did not show any cleavage activity of Cas8b under the tested conditions (data not shown).

A

I) flay_D_sp2 (60 nt)	5' -	AGCTGAGCGACTAGCGGAACTTC*ATTCCGCCCAGACCTAGAAAAGTATTATGGGATTAAT* - 3'
II) flay_D_ran (40 nt)	5' -	CTCTCCGCGAGCTGCTAGCTGTTCCGCTAGTCGCTCAGCT - 3'
III) flap_sp2 (37 nt)	5' -	*ATTAATCCCATAATACTTTTCTAGGTCTGGGCGGAAT* - 3'

B

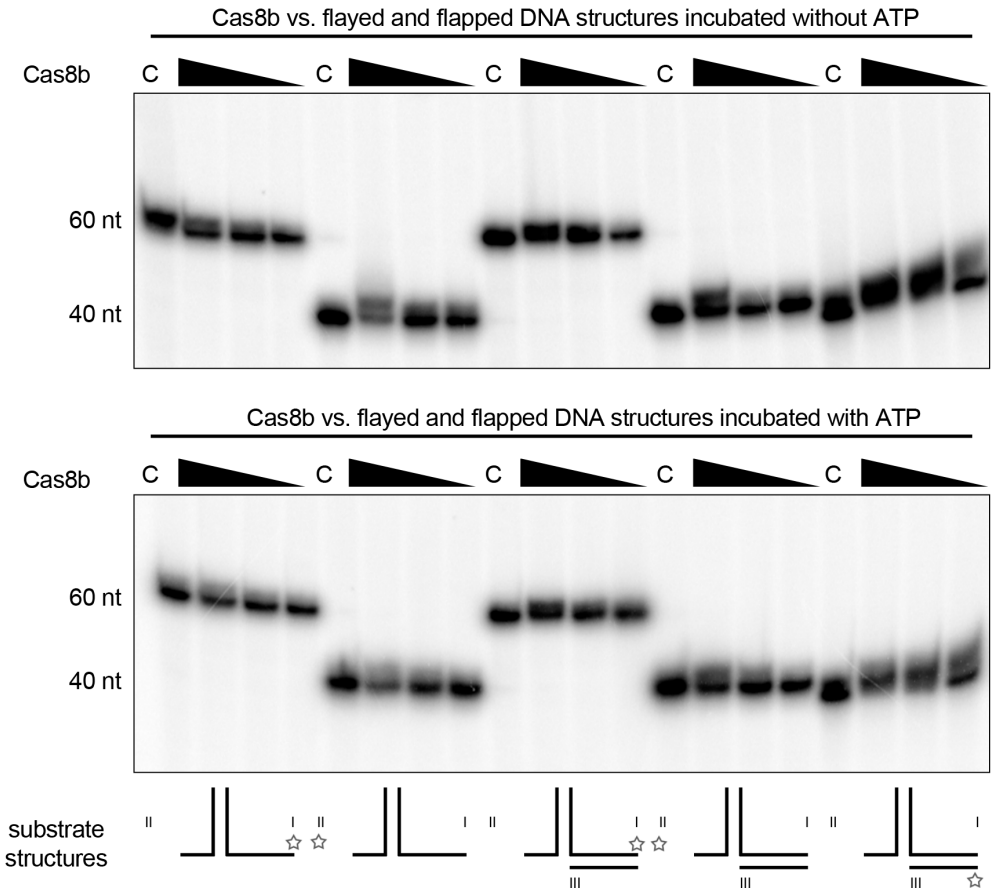


Figure 2. ***M. maripaludis*** **Cas8b does not exhibit endonucleolytic cleavage activity.** (**A**) Oligonucleotides designed for hybridization are displayed. Complementary sequence parts that hybridize with each other are shown by underlined and *italic* sequences. (**B**) Structures formed by hybridization are indicated below. Cas8b cleavage activity was determined with endonuclease assays using flapped and flayed DNA structures with alteration of the 5' labelled strand (marked by an asterisk). The upper panel shows a denaturing polyacrylamide gel of an assay without supplemented ATP and three decreasing concentrations of Cas8b (6 µM; 3 µM and 1 µM). In the lower panel the same reactions are shown with supplementation of ATP (10 mM), separation of the reaction was achieved by denaturing polyacrylamide gel electrophoresis. Cleavage activity was not observed under the tested conditions.

Cas8b binds nucleic acids unspecifically

Binding of Cas8b to different nucleic acid substrates was assayed by EMSA studies. Cas8b was shown to bind ssRNA, ssDNA, dsDNA and R-loop mimicking structures in an unspecific manner (Fig 3). Size-exclusion chromatography of a Cas8b:RNA complex shows a monomeric binding behaviour in assays performed with d-1 repeat RNA (data not shown). Competitive assays using either unlabelled d-1 repeat RNA or unlabelled and unspecific yeast RNA showed non-specificity of the binding reaction as the labelled RNA substrate is not shifted when higher yeast RNA concentrations were used (Fig. 3c). Cas8b also bound R-loop structures that were designed to should mimic structures generated during an interference event in which Cas8b would act as big Cascade subunit, without showing major preferences to any of the used substrates (Fig. 3d). In additional experiments we aimed to elucidate possible functions in PAM recognition that are reported for the large Cascade subunit Cse1 (Sashital et al, 2012) (substrate sequences are shown in Fig. 3a). However, Cas8b did not display a favoured binding to a TTC PAM containing substrate (PAM reported in (Fischer et al, 2012)) when compared to a non-PAM sequence (data not shown).

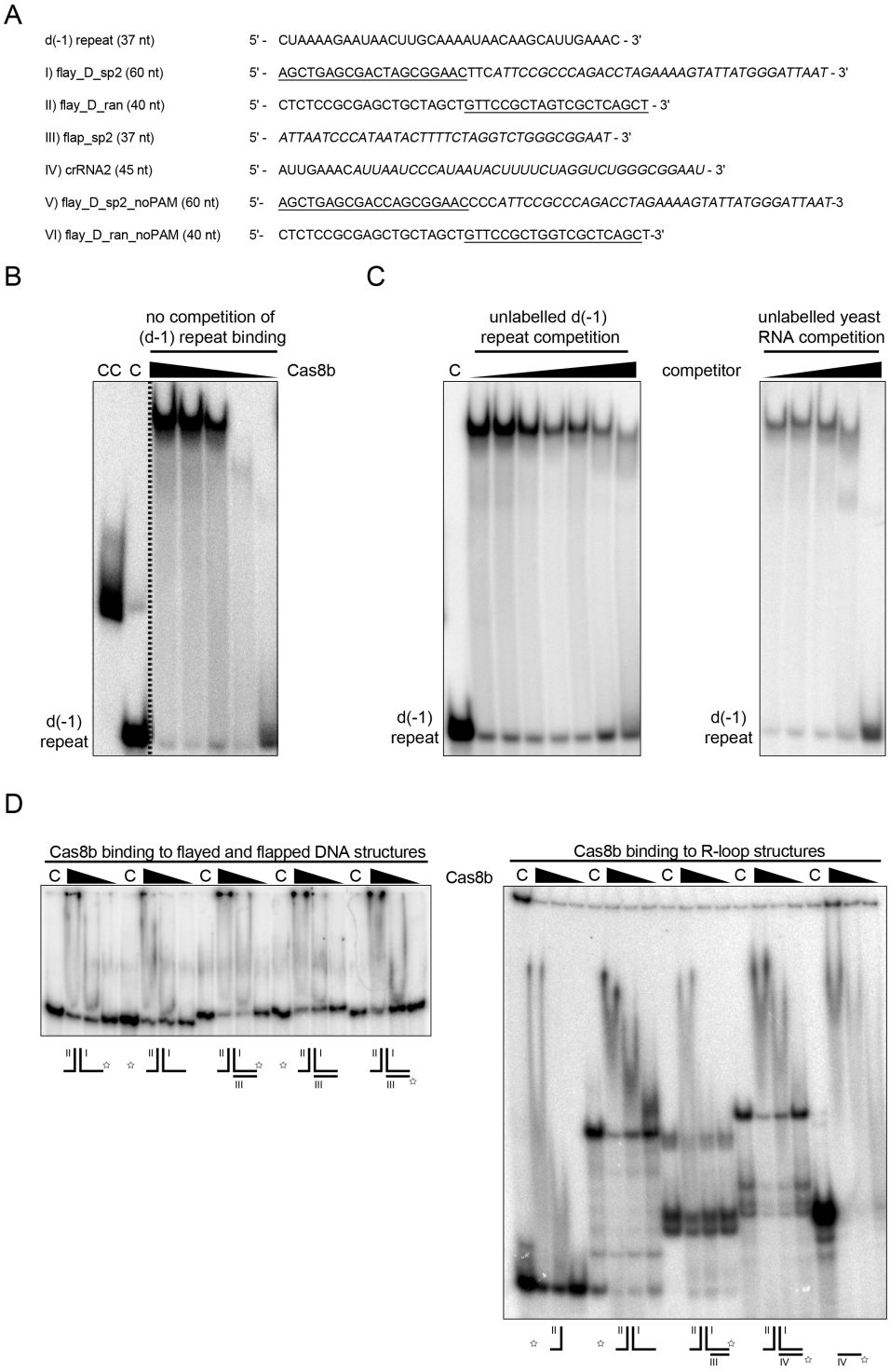


Figure 3. Unspecific nucleic acid binding by *M. maripaludis* Cas8b. (**A**) Substrates for EMSAs and oligonucleotides designed for hybridization are shown. Complementary sequence parts are depicted by underlined and *italic* sequences. (**B**) A native polyacrylamide gel of an EMSA using 5' end labelled d-1 repeat RNA and decreasing Cas8b concentrations (17.5 µM – 2.5 µM) shows a high shifting RNA:Cas8b complex. A reaction of RNA incubated in binding buffer (C) served as negative control and a reaction of Cas6b incubated with RNA conduced a positive binding control (CC). (**C**) Competition of the Cas8b:RNA shift with unlabelled d-1 repeat RNA (1 µM – 10 nM) and unlabelled yeast RNA (1 µg – 10 ng) shows decreasing amounts of shifted labelled RNA indicative for an

unspecific binding reaction. (**D**) Binding assays of different R-loop mimicking structures. The left panel shows binding to flapped and flayed DNA arrangements and the right panel depicts binding to flapped and flayed RNA structures. Hybrids formed by base complementarity are shown below the assays with small roman numerals marking the used oligonucleotide and asterisks highlighting labelled strands. Cas8b was able to bind to all tested structures.

Discussion

CRISPR-Cas interference complexes have been described for subtype I-E in *E. coli and T. themophilus*, subtype I-F in *P. aeruginosa* (Jore et al, 2011b; Sashital et al, 2012; Westra et al, 2012a; Wiedenheft et al, 2011a; Wiedenheft et al, 2011b) and also for Type II systems (Chylinski et al, 2013; Deltcheva et al, 2011; Jinek et al, 2012). Only little information was obtained for Type III systems (Hale et al, 2009; Osawa et al, 2013; Shao et al, 2013; Zhang et al, 2012) since a complete complex to investigate the underlying details is still not available. Similarly, the complete assembly of different Cascade-like complexes is lacking e.g. for subtype I-A systems. However, sub-complexes have been reported for an archaeal I-A system of *Sulfolobus solftaricus* (Lintner et al, 2011b; Reeks et al, 2013a; Reeks et al, 2013b). Studies of the *E. coli* Cascade showed that the complex is composed of Cas7, Cas5 and Cas6, proteins that can also be found in *M. maripaludis*. In addition to these three proteins, the I-E Cascade also possesses one large subunit (Cse1) and a homo-dimer of a small subunit (Cse2) (Jore et al, 2011b). Analysis of the genetic organization of the *cas* genes of subtype I-B reveals that there is no gene coding for a small Cascade subunit. Amongst the Type-I systems, only subtypes I-E (Cse1 and Cse2) and I-A (Cas8a and Csa5) have a small and large subunit, while in the other systems only one protein is found and thought to combine the functions of both subunits (Jore et al, 2011b; Makarova et al, 2011a; Reeks et al, 2013a). No biochemical data exists for the interference complexes of the two subtypes I-C and I-D. Subtype I-C, however, appears to be similar to the I-B systems as they share a resembling genetic organization and the Cas8c gene arrangement (putative Zn-finger, Palm and Thumb domains) suggest an evolutionary relationship (Makarova et al, 2011a). The large subunit Cas10d of the I-D systems is very different to all other Type-I systems and is reminiscent of Type-III systems, as Cas10d, in addition to the conserved Cas8 domains (Zn-finger, Palm and Thumb), has a N-terminal HD domain (Makarova et al, 2011a). Here, we present data for an archaeal Cas8b protein of *M. maripaludis*, the putative large subunit of a subtype I-B Cascade. The observed purification pattern of Cas8b into two fragments seems to be a feature of Cas8b proteins as two representative enzymes of different organisms and a variant with a different tag loacation of Mm Cas8b resulted in the same cleavage pattern. Bioinformatic analyses of Type-I large subunits show that the small subunit appears to be C-terminally fused to the large subunit resulting in the Cas8b protein (Makarova et al, 2011a). With respect to these analyses and to our results, we propose that in subtype I-B a proteolytic cleavage yields the small and the large subunit of the I-B Cascade. It seems that this cleavage so far is restricted to I-B systems, since no such cleavage is reported for the large subunit Csy1 of the I-F Cascade investigated in *P. aeruginosa*, which is also lacking a small subunit protein (Wiedenheft et al, 2011b). Our

results pinpoint the cleavage site within Cas8b to a region prone for proteolytic cleavage by aspartic proteases. Features of these proteases (e.g. Pepsin, Cathepsin) are a general two domain structure with two highly conserved aspartate residues in the N-terminal domain (Szecsi, 1992). Although our alignment does not show strongly conserved aspartate residues it is apparent that there is an accumulation of aspartate residues in the N-terminal protein region. Furthermore it is not a necessity to have two completely conserved aspartate residues as e.g. a retroviral protease of HIV is a heterodimer with each monomeric unit supplying one of the two aspartic protease domains (Graves, 1991; Toh et al, 1985). Next to the cleavage site, within Cas8b, a conserved region of the C-terminal end is found, which might contribute a recognition site for a protease. Proteolytic cleavage of proenzymes to yield active forms of enzymes has been described for many proteins. For example Protein Kinase C is activated by proteolytic cleavage of Ca^{2+} dependent Protease I (Kajikawa et al, 1983; Kishimoto et al, 1983; Takai et al, 1979). Also various other human systems use activation of proteins that are similar, for instance trypsin activation is achieved by proteolytic cleavage of trypsinogen by an enteropeptidase (Desnuelle & Gabeloteau, 1957; Liener, 1960; Rovery & Desnuelle, 1954). However, activation does not necessarily have to be performed *in trans*. Autocatalytic cleavage of proteins has been reported for several cases in which e.g. an N-terminal nucleophile is responsible for the processing. This would also be one putative mechanism for Cas8b. In this reaction, e.g. in the case of the glutamine PRPP amidotransferase (GAT), a cysteine residue takes the role of the active nucleophile for self-processing (Brannigan et al, 1995). In Cas8b four cysteine residues are conserved in two CxxC motifs and e.g. might take part in an autocatalytic cleavage. However, the CxxC motifs can also be part of a cysteine switch (Springman et al, 1990; Van Wart & Birkedal-Hansen, 1990) in which conserved cysteine residues can interact with Zn^{2+} ions. If zinc is bound to the cysteine residues the protein is in an inactive state. However, an active site with the motif HExxHxxGxxH that is a feature of cysteine metalloproteases (Springman et al, 1990; Van Wart & Birkedal-Hansen, 1990) cannot be identified in the Cas8b alignment.

The presented proteolytic mechanisms display only a few instances of different processing reactions to yield active proteins and show a possibility of regulating downstream processes after transcription and translation. These kinds of feedback regulations help to fine-tune active protein species to not disturb the cellular protein equilibrium. Although Cas8b might not necessarily follow one of the described processing pathways, our results suggest a potential for Cas8b to undergo post-translational cleavage to yield the large and small Cascade subunit.

In a first biochemical characterization of a Cas8b homolog of *M. thermautotrophicus* (Nar71), a nucleic acid cleavage activity and strand displacement for flapped and flayed DNA structures was reported. The authors suggested a novel DNA repair system in thermophiles

(Guy et al, 2004). However, in experiments using similar substrates based on the CRISPR-Cas immune system, Cas8b of *M. maripaludis* did not show any of those activities. The substrates were similar to the flayed and flapped DNA structures used in the Nar71 studies (Guy et al, 2004), but the designed oligonucleotides contained spacer sequences and putative PAM sequences instead of random sequences. Considering their size, a sequence identity of only 14.5 % of both proteins is very low and surprisingly the two CxxC motifs that are usually conserved between Cas8b sequences are not found in Nar71. This suggests two different forms of Cas8b since it is also not reported that Nar71 purifies into two fragments (Guy et al, 2004). Therefore, Nar71 might exhibit different functions. The binding activity that both proteins share might be needed for R-loop formation and stabilization. The lack of specificity of nucleic acid binding is in agreement with studies in which a subtype I-E Cascade sub-complex missing the big subunit Cse1 showed more specific binding compared to a full Cascade that exhibited unspecific DNA binding behaviour. However, Cse1 of *E. coli* alone was not able to bind DNA and only Cascade combinations with the small subunit Cse2 enabled binding (Jore et al, 2011b). However, in another study, Cse1 of *T. thermophilus* was suggested to be involved in PAM recognition (Sashital et al, 2012). Therefore, the chosen PAM for our Cas8b analysis, based on another subtype I-B study (Fischer et al, 2012), may not be correct and a future screening for the proper PAM sequence should be used to analyse the PAM-specificity of Cas8b.

The nucleic acid binding shown by both, Nar71 and Mm Cas8b, lacked any specificity, which may be only apparent in the context of a complete Cascade. Therefore, it might be that the assembly of Cas8b into the interference complex might be required to reveal the native function of the protein.

Chapter VI

Conclusion

Conclusion

CRISPR-Cas systems show a broad diversity with different described types and subtypes (Makarova et al, 2011b). Although the general aspects and mechanisms of the three major types are very similar, the underlying details can vary significantly between the subtypes.

The adaption step is proposed to be conserved between the different types, as all systems seem to employ a similar complex (Cascis) formed by the universal proteins Cas1 and Cas2 (Erdmann & Garrett, 2012; Fineran & Charpentier, 2012; Plagens et al, 2012; Swarts et al, 2012; Yosef et al, 2012).

The processing step, which yields mature crRNA, is similar amongst Type-I systems. All subtypes use a Cas6 homolog to process pre-crRNA yielding a mature form of the interfering RNA (Gesner et al, 2011; Haurwitz et al, 2010; Reeks et al, 2013c; Richter et al, 2012; Sashital et al, 2011). One exception is found within subtype I-C in which no Cas6 homolog is present and Cas5d is responsible for pre-crRNA processing (Garside et al, 2012; Nam et al, 2012). The acid/base chemistry that is applied by the Cas6 enzymes appears to be conserved. However, the details of substrate recognition and processing by Cas6 enzymes differ between subtypes. This study provides the identification of the subtype I-B endoribonuclease Cas6b of *M. maripaludis*, which processes pre-crRNA into mature crRNA. With the support of the results obtained in this work, a new classification of Cas6 enzymes based on their mode of substrate recognition is proposed. Furthermore it is proposed, that the differences of Cas6 enzymes can be used to distinguish different subtypes.

The interference step is the most diverse stage of CRISPR-Cas immunity with regard to participating complexes and modes of target recognition. Type-I systems employ an interference complex called Cascade (CRISPR associated complex for antiviral defense), which recognizes the invading nucleic acid via crRNA base pairing and prepares it for degradation by Cas3. Subtype I-E of *E. coli* is the best studied Type-I system and the Cascade is reported to be composed of Cas5, Cas7, Cas6 and the large and small Cascade subunits, Cse1 and Cse2, respectively (Jore et al, 2011b; Westra et al, 2012c; Wiedenheft et al, 2011a). Amongst the different Type-I systems only the subtypes I-E and I-A possess genes coding for a small and large Cascade subunit, while in the other systems one protein appears to combine the functions of both subunits. In this study we report the identification of proteolytic cleavage of the large Cascade subunit Cas8b of the subtype I-B system present in *M. maripaludis* and *C. thermocellum* and propose a novel mechanism to yield the small and large subunits of Cascade.

An evolutionary link between Cas6 endoribonucleases of different CRISPR-Cas subtypes

All studied Cas6 enzymes share some basic features, e.g. a ferredoxin-like fold or the metal independent processing of pre-crRNA. In addition, the cleavage reaction always yields mature crRNA with a 5' terminal repeat tag of 8 nucleotides (Carte et al, 2008; Gesner et al, 2011; Haurwitz et al, 2010; Richter et al, 2012; Sashital et al, 2011). Even though the Cas6 activities are similar, the sequence identity of these proteins is limited and mechanistic differences became apparent between the studied homologs. While Cas6e and Cas6f show a very high affinity towards their substrate and specifically recognize a hairpin structure within the repeat (Gesner et al, 2011; Haurwitz et al, 2010; Sashital et al, 2011; Sternberg et al, 2012), Pf Cas6 and Cas6b do not seem to rely on a structural recognition motif of their substrates. Instead a sequence motif located 12 nt upstream of the processing site appears to influence binding (Carte et al, 2010; Carte et al, 2008; Wang et al, 2011; Wang et al, 2012). Additionally, Cas6b showed a comparable low affinity towards its substrate, which might correlate with the wrap-around mechanism proposed for Pf Cas6 (see Fig. 1) (Wang et al, 2011). The wrap-around mechanism describes a model in which several Cas6 enzymes are bound along the pre-crRNA in a bead chain like manner, wrapping the RNA around the enzyme. As this mechanism is not based on secondary structures, the missing structural elements that highlight the binding sites could delay tight binding, as scanning for the recognition motifs and coordination of the repeat cleavage site into the catalytic center might take more time and result in less affine binding. The lower processing rate measured for Cas6b in comparison to the observed reaction rates of Cas6e and Cas6f (Gesner et al, 2011; Sternberg et al, 2012) is difficiult to explain by this mechanism since a chain of multiple Cas6 units could enable multiple cleavage events along a long pre-crRNA transcript.

Cas6 enzymes also differ in their catalytic site composition. The first identified and characterized Cas6 enzyme, Pf Cas6, harbors a catalytic triad composed of a tyrosine, an histidine and a lysine residue (Carte et al, 2010; Carte et al, 2008; Wang et al, 2011; Wang et al, 2012). Studies to identify the catalytic site of Cas6e show that only two residues (tyrosine and histidine) are involved in crRNA maturation (Gesner et al, 2011; Sashital et al, 2011). A catalytic dyad (histidine and serine) is also described for Cas6f (Haurwitz et al, 2010). The residues involved in the catalysis performed by Cas6b are the same as in Pf Cas6, but their order is reversed to lysine, histidine and tyrosine. Additionally, not one but two conserved histidine residues play a major role in Cas6b pre-crRNA processing (Richter et al, 2012). Whereas the composition of the active site shows variations, one constant is the canonical histidine, which is found in all of these reported enzymes and is required for target protonation (Carte et al, 2008). One exception to prove the rule is a Cas6 homolog of *S.*

solfataricus, which does not rely on the canonical histidine. Here the protonation of the target is performed by a catalytic site that is composed of three lysine residues and one arginine (Reeks et al, 2013c).

Cas6 sequence similarities are very low but a structural model of Cas6b (Richter et al, 2012) aligns very well with the crystal structure of Pf Cas6 (Wang et al, 2011). This structural conservation and the similarities in crRNA processing indicate that both Cas6 proteins might be evolutionary linked. Bioinformatical analysis of the three CRISPR-Cas types suggest that Type-III systems are the progenitors of the Type-I and –II systems. Evidential for this proposal are the presence of a Cas10 homolog in Type-I systems (Cas10d in I-D) and the high similarities between Cas8 proteins of Type-I and Cmx1 proteins of Type-III (Makarova et al, 2011a). Studies in *Methanosarcina mazei* further report that Cas6 enzymes of a Type I-B and a Type III-B system independently show identical processing patterns with a consensus repeat substrate (Nickel et al, 2013). Additionally, it was shown that Type–III systems lack a Cas1 gene and therefore usually pair with at least one subtype I-A or I-B system (Makarova et al, 2011a; Makarova et al, 2011b). This additionally implies a possible functional connection between Type-I and Type-III systems. Further evidence for this link of the two systems is provided by a non-catalytic Type I-A Pf Cas6 homolog of *Pyrococcus horikoshii,* which shares a sequence identity of 75 % with the III-B Cas6. This enzyme is also able to form dimers (Wang & Li, 2012) similar to Cas6b, which can form dimers with an uncleavable repeat substrate (Richter et al, 2013a). Furthermore, the binding site within the protein is found to be distal to the catalytic site in both Pf Cas6 and Cas6b (Wang et al, 2011; Wang et al, 2012), which is indicative for the proposed wrap-around mechanism and would separate both proteins from the mechanism found e.g. in Cas6e and Cas6f. Therefore, we propose two classes of Cas6 enzymes (Fig.1), one that specifically recognizes repeat structures (Cas6e, Cas6f) and a second class that is based on a sequence dependence (Pf Cas6, Cas6b). However, to confirm the similarities between Cas6b and Pf Cas6, a crystal structure of Cas6b is needed.

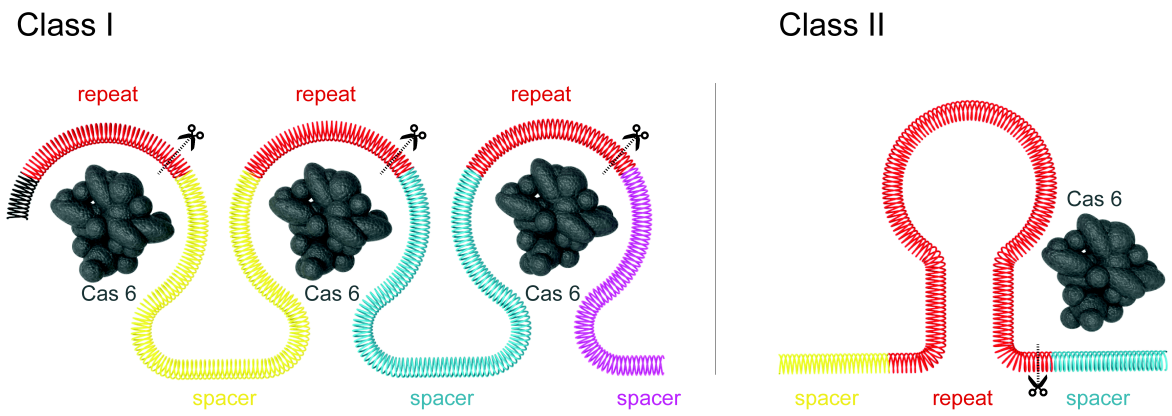


Figure 1. A proposed classification of Cas6 enzymes. Based on the mode of substrate recognition two classes of Cas6 enzymes are described. Class I Cas6 enzymes (e.g. Pf Cas6 and Cas6b) do not rely on secondary structures of their substrates and employ the proposed wrap-around mechanism. The spacer sequences help to guide the repeat sequences around multiple Cas6 units to mediate a bead chain like binding. Enzymes belonging to Class II (e.g. Cas6e and Cas6f) specifically recognize a stem loop structure of the repeat and facilitate cleavage at the base of the hairpin.

The large Cascade subunit Cas8b – two for the price of one?

The basic assembly of Type-I interference complexes appears to be conserved. Studies performed with Cascade complexes of the subtypes I-E and I-F suggest the universal presence of Cas5, Cas6 and Cas7 proteins (Jore et al, 2011b; Wiedenheft et al, 2011a; Wiedenheft et al, 2011b). However, differences can be found regarding the presence of a large and small Cascade subunit. While Cascade of subtype I-F does not employ a small subunit and only a large subunit is present (Wiedenheft et al, 2011b), the I-E Cascade is composed of a large and a small subunit (Jore et al, 2011b; Wiedenheft et al, 2011a). The occurrence of small and large subunits within Type-I systems differs, subtypes I-A and I-E possess genes coding for both subunits, whereas the subtypes I-B, I-C and I-F only have one gene coding for a large protein, which is likely required to combine the function of both, the small and the large subunit. Subtype I-D represents an exception with Cas10d, a large subunit with an HD domain that is reminiscent of the Type-III Cas10 enzymes (Makarova et al, 2011a; Makarova et al, 2011b).

Analysing Cas8b, the large subunit of the subtype I-B system of *M. maripaludis*, we observed a distinct proteolytic cleavage event of the protein. This cleavage within the C-terminal end of the protein is conserved in another Cas8b of *C. thermocellum* and we speculate that the proteolytic activity could create protein fragments that might function as the large and small

subunit. Additional evidence for this proposal is found in bioinformatic analysis of Cas8 proteins, which suggest that the small subunit is C-terminally fused to the large Cascade subunit (Makarova et al, 2011a). Large and small Cascade subunits are thought to be involved in PAM recognition (Sashital et al, 2012), target binding and R-Loop stabilization (Jore et al, 2011b; Makarova et al, 2011a; Makarova et al, 2011b). Biochemical analyses of Nar71, a Cas8b homolog of *M. thermautothrophicus,* showed that Cas8b is a nuclease with ATPase properties. Furthermore, this protein was able to bind Holliday junctions, as well as flayed and flapped DNA structures. However, without the knowledge about CRISPR-Cas the protein was proposed to be part of a novel DNA repair system of thermophiles (Guy et al, 2004). Analyses of Cas8b of *M. maripaludis* with similar substrates that may occur during CRISPR-Cas interference did not show nucleic acid cleavage activity and revealed unspecific DNA binding properties. The large subunit Cse1 of *E. coli* was not able to shift DNA substrates and only within Cascade and in combination with the small subunit Cse2 a non-specific shift of DNA was observed (Jore et al, 2011b). In contrast, studies with Cse1 of *T. thermophilus* using DNA substrates with and without PAM sequences revealed substrate binding based on PAM recognition (Sashital et al, 2012). These results confirm the predicted role of the large subunits during PAM recognition. The isolated Cas8b was tested for PAM recognition but did not show specificity towards any of the used substrates. Future screening of additional PAM sequence candidaters might be necessary. Additionally, future studies will focus on a completely assembled I-B Cascade to study if the functions of large and small Cas proteins are revealed in the context of the protein complex. In this study we established a Cas6b approach to create various crRNAs *in vitro*, which will be used to load the studied Cascade complexes.

Appendix

References

List of figures

List of tables

List of abbreviations

Curriculum vitae

Danksagung

Eigenständigkeitserklärung

References

Abedon ST (2012) Bacterial 'immunity' against bacteriophages. *Bacteriophage* **2:** 50-54

Albers SV, Jonuscheit M, Dinkelaker S, Urich T, Kletzin A, Tampe R, Driessen AJ, Schleper C (2006) Production of recombinant and tagged proteins in the hyperthermophilic archaeon Sulfolobus solfataricus. *Applied and environmental microbiology* **72:** 102-111

Arber W, Linn S (1969) DNA modification and restriction. *Annual review of biochemistry* **38:** 467-500

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

Barrangou R, Horvath P (2011) CRISPR: New Horizons in Phage Resistance and Strain Identification. *Annu Rev Food Sci Technol*

Beloglazova N, Petit P, Flick R, Brown G, Savchenko A, Yakunin AF (2011) Structure and activity of the Cas3 HD nuclease MJ0384, an effector enzyme of the CRISPR interference. *The EMBO journal* **30:** 4616-4627

Berkner S, Grogan D, Albers SV, Lipps G (2007) Small multicopy, non-integrative shuttle vectors based on the plasmid pRN1 for Sulfolobus acidocaldarius and Sulfolobus solfataricus, model organisms of the (cren-)archaea. *Nucleic acids research* **35:** e88

Berkner S, Wlodkowski A, Albers SV, Lipps G (2010) Inducible and constitutive promoters for genetic systems in Sulfolobus acidocaldarius. *Extremophiles : life under extreme conditions* **14:** 249-259

Bertsch A, Gropl C, Reinert K, Kohlbacher O (2011) OpenMS and TOPP: open source software for LC-MS data analysis. *Methods in molecular biology* **696:** 353-367

Blank CE, Kessler PS, Leigh JA (1995) Genetics in methanogens: transposon insertion mutagenesis of a Methanococcus maripaludis nifH gene. *Journal of bacteriology* **177:** 5773-5777

Bolotin A, Quinquis B, Sorokin A, Ehrlich SD (2005) Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151:** 2551-2561

Bondy-Denomy J, Pawluk A, Maxwell KL, Davidson AR (2013) Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature* **493:** 429-432

Brannigan JA, Dodson G, Duggleby HJ, Moody PC, Smith JL, Tomchick DR, Murzin AG (1995) A protein catalytic framework with an N-terminal nucleophile is capable of self-activation. *Nature* **378:** 416-419

Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, Snijders AP, Dickman MJ, Makarova KS, Koonin EV, van der Oost J (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321:** 960-964

Cady KC, O'Toole GA (2011) Non-identity-mediated CRISPR-bacteriophage interaction mediated via the Csy and Cas3 proteins. *Journal of bacteriology* **193:** 3433-3445

Calvin K, Hall MD, Xu F, Xue S, Li H (2005) Structural characterization of the catalytic subunit of a novel RNA splicing endonuclease. *Journal of molecular biology* **353:** 952-960

Carte J, Pfister NT, Compton MM, Terns RM, Terns MP (2010) Binding and cleavage of CRISPR RNA by Cas6. *Rna* **16:** 2181-2188

Carte J, Wang R, Li H, Terns RM, Terns MP (2008) Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes Dev* **22:** 3489-3496

Charpentier E, Doudna JA (2013) Biotechnology: Rewriting a genome. *Nature* **495:** 50-51

Chylinski K, Le Rhun A, Charpentier E (2013) The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems. *RNA biology* **10**

Cocozaki AI, Ramia NF, Shao Y, Hale CR, Terns RM, Terns MP, Li H (2012) Structure of the Cmr2 subunit of the CRISPR-Cas RNA silencing complex. *Structure* **20:** 545-553

Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, Zhang F (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science* **339:** 819-823

Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. *Genome research* **14:** 1188-1190

Cui Y, Li Y, Gorge O, Platonov ME, Yan Y, Guo Z, Pourcel C, Dentovskaya SV, Balakhonov SV, Wang X, Song Y, Anisimov AP, Vergnaud G, Yang R (2008) Insight into microevolution of Yersinia pestis by clustered regularly interspaced short palindromic repeats. *PloS one* **3:** e2652

Datsenko KA, Pougach K, Tikhonov A, Wanner BL, Severinov K, Semenova E (2012) Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nature communications* **3:** 945

Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, Eckert MR, Vogel J, Charpentier E (2011) CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471:** 602-607

Deng L, Kenchappa CS, Peng X, She Q, Garrett RA (2012) Modulation of CRISPR locus transcription by the repeat-binding protein Cbp1 in Sulfolobus. *Nucleic acids research* **40:** 2470-2480

Desnuelle P, Gabeloteau C (1957) [Role of calcium during the activation of trypsinogen by trypsin]. *Archives of biochemistry and biophysics* **69:** 475-485

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

Enikeeva FN, Severinov KV, Gelfand MS (2010) Restriction-modification systems and bacteriophage invasion: who wins? *Journal of theoretical biology* **266:** 550-559

Erdmann S, Garrett RA (2012) Selective and hyperactive uptake of foreign DNA by adaptive immune systems of an archaeon via two distinct mechanisms. *Molecular microbiology* **85:** 1044-1056

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

Fischer S, Maier LK, Stoll B, Brendel J, Fischer E, Pfeiffer F, Dyall-Smith M, 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

Fujishima K, Sugahara J, Miller CS, Baker BJ, Di Giulio M, Takesue K, Sato A, Tomita M, Banfield JF, Kanai A (2011) A novel three-unit tRNA splicing endonuclease found in ultrasmall Archaea possesses broad substrate specificity. *Nucleic acids research* **39:** 9695-9704

Gardner WL, Whitman WB (1999) Expression vectors for Methanococcus maripaludis: overexpression of acetohydroxyacid synthase and beta-galactosidase. *Genetics* **152:** 1439-1447

Garside EL, Schellenberg MJ, Gesner EM, Bonanno JB, Sauder JM, Burley SK, Almo SC, Mehta G, Macmillan AM (2012) Cas5d processes pre-crRNA and is a member of a larger family of CRISPR RNA endonucleases. *Rna* **18:** 2020-2028

Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH (2004) Open mass spectrometry search algorithm. *Journal of proteome research* **3:** 958-964

Gesner EM, Schellenberg MJ, Garside EL, George MM, Macmillan AM (2011) Recognition and maturation of effector RNAs in a CRISPR interference pathway. *Nat Struct Mol Biol* **18:** 688-692

Graves MC (1991) Human immunodeficiency virus proteinase: now, then, what's next? *Advances in experimental medicine and biology* **306:** 395-405

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

Grissa I, Vergnaud G, Pourcel C (2007b) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic acids research* **35:** W52-57

Gudbergsdottir S, Deng L, Chen Z, Jensen JV, Jensen LR, She Q, Garrett RA (2011) Dynamic properties of the Sulfolobus CRISPR/Cas and CRISPR/Cmr systems when challenged with vector-borne viral and plasmid genes and protospacers. *Molecular microbiology* **79:** 35-49

Guy CP, Majernik AI, Chong JP, Bolt EL (2004) A novel nuclease-ATPase (Nar71) from archaea is part of a proposed thermophilic DNA repair system. *Nucleic acids research* **32:** 6176-6186

Hale C, Kleppe K, Terns RM, Terns MP (2008) Prokaryotic silencing (psi)RNAs in Pyrococcus furiosus. *Rna* **14:** 2572-2579

Hale CR, Majumdar S, Elmore J, Pfister N, Compton M, Olson S, Resch AM, Glover CV, 3rd, Graveley BR, Terns RM, Terns MP (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

Hale CR, Zhao P, Olson S, Duff MO, Graveley BR, Wells L, Terns RM, Terns MP (2009) RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* **139:** 945-956

Hatoum-Aslan A, Maniv I, Marraffini LA (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

Haurwitz RE, Jinek M, Wiedenheft B, Zhou K, Doudna JA (2010) Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science* **329:** 1355-1358

Haurwitz RE, Sternberg SH, Doudna JA (2012) Csy4 relies on an unusual catalytic dyad to position and cleave CRISPR RNA. *The EMBO journal*

Holm L, Park J (2000) DaliLite workbench for protein structure comparison. *Bioinformatics* **16:** 566-567

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

Horvath P, Barrangou R (2013) RNA-guided genome editing a la carte. *Cell research*

Horvath P, Romero DA, Coute-Monvoisin AC, Richards M, Deveau H, Moineau S, Boyaval P, Fremaux C, Barrangou R (2008) Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. *Journal of bacteriology* **190:** 1401-1412

Howard JA, Delmas S, Ivancic-Bace I, Bolt EL (2011) Helicase dissociation and annealing of RNA-DNA hybrids by Escherichia coli Cas3 protein. *The Biochemical journal* **439:** 85-95

Hyman P, Abedon ST (2010) Bacteriophage host range and bacterial resistance. *Adv Appl Microbiol* **70:** 217-248

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

Jones WJ, Nagle DP, Jr., Whitman WB (1987) Methanogens and the diversity of archaebacteria. *Microbiological reviews* **51:** 135-177

Jones WJ, Whitman WB, Fields RD, Wolfe RS (1983) Growth and plating efficiency of methanococci on agar media. *Applied and environmental microbiology* **46:** 220-226

Jore MM, Brouns SJ, van der Oost J (2011a) RNA in Defense: CRISPRs Protect Prokaryotes against Mobile Genetic Elements. *Cold Spring Harb Perspect Biol*

Jore MM, Lundgren M, van Duijn E, Bultema JB, Westra ER, Waghmare SP, Wiedenheft B, Pul U, Wurm R, Wagner R, Beijer MR, Barendregt A, Zhou K, Snijders AP, Dickman MJ, Doudna JA, Boekema EJ, Heck AJ, van der Oost J, Brouns SJ (2011b) Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nat Struct Mol Biol* **18:** 529-536

Kajikawa N, Kishimoto A, Shiota M, Nishizuka Y (1983) Ca2+ -dependent neutral protease and proteolytic activation of Ca2+-activated, phospholipid-dependent protein kinase. *Methods in enzymology* **102:** 279-290

Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, Bunschoten A, Molhuizen H, Shaw R, Goyal M, van Embden J (1997) Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. *Journal of clinical microbiology* **35:** 907-914

Kessner D, Chambers M, Burke R, Agus D, Mallick P (2008) ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **24:** 2534-2536

Keswani J, Orkand S, Premachandran U, Mandelco L, Franklin MJ, Whitman WB (1996) Phylogeny and taxonomy of mesophilic Methanococcus spp. and comparison of rRNA, DNA hybridization, and phenotypic methods. *International journal of systematic bacteriology* **46:** 727-735

Kishimoto A, Kajikawa N, Shiota M, Nishizuka Y (1983) Proteolytic activation of calcium-activated, phospholipid-dependent protein kinase by calcium-dependent neutral protease. *The Journal of biological chemistry* **258:** 1156-1164

Koonin EV, Makarova KS (2009) CRISPR-Cas: an adaptive immunity system in prokaryotes. *F1000 Biol Rep* **1:** 95

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

Labrie SJ, Samson JE, Moineau S (2010) Bacteriophage resistance mechanisms. *Nature reviews Microbiology* **8:** 317-327

Lange SJ, Alkhnbashi OS, Rose D, Will S, Backofen R (2013) CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems. *Nucleic acids research* **41:** 8034-8044

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23:** 2947-2948

Li H, Abelson J (2000) Crystal structure of a dimeric archaeal splicing endonuclease. *Journal of molecular biology* **302:** 639-648

Li H, Trotta CR, Abelson J (1998) Crystal structure and evolution of a transfer RNA splicing enzyme. *Science* **280:** 279-284

Liener IE (1960) Chromatographic studies on trypsin, trypsinogen and the activation process. *Archives of biochemistry and biophysics* **88:** 216-221

Lillestol RK, Shah SA, Brugger K, Redder P, Phan H, Christiansen J, Garrett RA (2009) CRISPR families of the crenarchaeal genus Sulfolobus: bidirectional transcription and dynamic properties. *Molecular microbiology* **72:** 259-272

Lintner NG, Frankel KA, Tsutakawa SE, Alsbury DL, Copie V, Young MJ, Tainer JA, Lawrence CM (2011a) The structure of the CRISPR-associated protein Csa3 provides insight into the regulation of the CRISPR/Cas system. *Journal of molecular biology* **405:** 939-955

Lintner NG, Kerou M, Brumfield SK, Graham S, Liu H, Naismith JH, Sdano M, Peng N, She Q, Copie V, Young MJ, White MF, Lawrence CM (2011b) 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

Lynd LR, Grethlein HE (1987) Hydrolysis of dilute acid pretreated mixed hardwood and purified microcrystalline cellulose by cell-free broth from Clostridium thermocellum. *Biotechnology and bioengineering* **29:** 92-100

Makarova KS, Aravind L, Wolf YI, Koonin EV (2011a) Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biol Direct* **6:** 38

Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV (2006) A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct* **1:** 7

Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, Horvath P, Moineau S, Mojica FJ, Wolf YI, Yakunin AF, van der Oost J, Koonin EV (2011b) Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol* **9:** 467-477

Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM (2013) RNA-guided human genome engineering via Cas9. *Science* **339:** 823-826

Marraffini LA, Sontheimer EJ (2010a) CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat Rev Genet* **11:** 181-190

Marraffini LA, Sontheimer EJ (2010b) Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature* **463:** 568-571

McCaskill JS (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29:** 1105-1119

Miller TL, Wolin MJ (1982) Enumeration of Methanobrevibacter smithii in human feces. *Archives of microbiology* **131:** 14-18

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

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

Mulepati S, 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

Mulepati S, Orr A, 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

Nam KH, Haitjema C, Liu X, Ding F, Wang H, DeLisa MP, 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

Nickel L, Weidenbach K, Jager D, Backofen R, Lange SJ, Heidrich N, Schmitz RA (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

Niewoehner O, Jinek M, Doudna JA (2013) Evolution of CRISPR RNA recognition and processing by Cas6 endonucleases. *Nucleic acids research*

Osawa T, Inanaga H, Numata T (2013) Crystal Structure of the Cmr2-Cmr3 Subcomplex in the CRISPR-Cas RNA Silencing Effector Complex. *Journal of molecular biology*

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

Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, Lim WA (2013) Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152:** 1173-1183

Raivio T (2011) Identifying your enemies--could envelope stress trigger microbial immunity? *Molecular microbiology* **79:** 557-561

Randau L (2012) RNA processing in the minimal organism Nanoarchaeum equitans. *Genome biology* **13:** R63

Randau L, Calvin K, Hall M, Yuan J, Podar M, Li H, Söll D (2005a) The heteromeric Nanoarchaeum equitans splicing endonuclease cleaves noncanonical bulge-helix-bulge motifs of joined tRNA halves. *Proceedings of the National Academy of Sciences of the United States of America* **102:** 17934-17939

Randau L, Münch R, Hohn MJ, Jahn D, Söll D (2005b) Nanoarchaeum equitans creates functional tRNAs from separate genes for their 5'- and 3'-halves. *Nature* **433:** 537-541

Reeks J, Graham S, Anderson L, Liu H, White MF, Naismith JH (2013a) Structure of the archaeal Cascade subunit Csa5: relating the small subunits of CRISPR effector complexes. *RNA biology* **10:** 762-769

Reeks J, Naismith JH, White MF (2013b) CRISPR interference: a structural perspective. *The Biochemical journal* **453:** 155-166

Reeks J, Sokolowski RD, Graham S, Liu H, Naismith JH, White MF (2013c) Structure of a dimeric crenarchaeal Cas6 enzyme with an atypical active site for CRISPR RNA processing. *The Biochemical journal*

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

Richter H, Randau L, Plagens A (2013b) Exploiting CRISPR/Cas: interference mechanisms and applications. *International journal of molecular sciences* **14:** 14518-14531

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

Rovery M, Desnuelle P (1954) [The chemical mechanism of activation of beef chymotrypsinogen by trypsin]. *Comptes rendus des seances de la Societe de biologie et de ses filiales* **148:** 1437-1439

Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nature protocols* **5:** 725-738

Sampson JR, Uhlenbeck OC (1988) Biochemical and physical characterization of an unmodified yeast phenylalanine transfer RNA transcribed in vitro. *Proceedings of the National Academy of Sciences of the United States of America* **85:** 1033-1037

Sandbeck KA, Leigh JA (1991) Recovery of an integration shuttle vector from tandem repeats in Methanococcus maripaludis. *Applied and environmental microbiology* **57:** 2762-2763

Santangelo TJ, Reeve JN (2006) Archaeal RNA polymerase is sensitive to intrinsic termination directed by transcribed and remote sequences. *Journal of molecular biology* **355:** 196-210

Sapranauskas R, Gasiunas G, Fremaux C, Barrangou R, Horvath P, Siksnys V (2011) The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. *Nucleic acids research* **39:** 9275-9282

Sashital DG, Jinek M, Doudna JA (2011) An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3. *Nat Struct Mol Biol* **18:** 680-687

Sashital DG, Wiedenheft B, Doudna JA (2012) Mechanism of foreign DNA selection in a bacterial adaptive immune system. *Molecular cell* **46:** 606-615

Schmidt C, Kramer K, Urlaub H (2012) Investigation of protein-RNA interactions by mass spectrometry--Techniques and applications. *Journal of proteomics* **75:** 3478-3494

Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. *Nucleic acids research* **18:** 6097-6100

Schurer H, Lang K, Schuster J, Morl M (2002) A universal method to produce in vitro transcripts with homogeneous 3' ends. *Nucleic acids research* **30:** e56

Schuster P, Fontana W, Stadler PF, Hofacker IL (1994) From sequences to shapes and back: a case study in RNA secondary structures. *Proceedings Biological sciences / The Royal Society* **255:** 279-284

Semenova E, Jore MM, Datsenko KA, Semenova A, Westra ER, Wanner B, van der Oost J, Brouns SJ, Severinov K (2011) Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc Natl Acad Sci USA* **108:** 10098-10103

Shah SA, Erdmann S, Mojica FJ, Garrett RA (2013) Protospacer recognition motifs: mixed identities and functional diversity. *RNA biology* **10:** 891-899

Shao Y, Cocozaki AI, Ramia NF, Terns RM, Terns MP, Li H (2013) Structure of the Cmr2-Cmr3 subcomplex of the Cmr RNA silencing complex. *Structure* **21:** 376-384

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

Sinkunas T, Gasiunas G, Fremaux C, Barrangou R, Horvath P, Siksnys V (2011) Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *The EMBO journal* **30:** 1335-1342

Sinkunas T, Gasiunas G, Waghmare SP, Dickman MJ, Barrangou R, Horvath P, Siksnys V (2013) In vitro reconstitution of Cascade-mediated CRISPR immunity in Streptococcus thermophilus. *The EMBO journal* **32:** 385-394

Sorek R, Kunin V, Hugenholtz P (2008) CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol* **6:** 181-186

Soukup GA, Breaker RR (1999) Relationship between internucleotide linkage geometry and the stability of RNA. *RNA* **5:** 1308-1325

Springman EB, Angleton EL, Birkedal-Hansen H, Van Wart HE (1990) Multiple modes of activation of latent human fibroblast collagenase: evidence for the role of a Cys73 active-site zinc complex in latency and a "cysteine switch" mechanism for activation. *Proceedings of the National Academy of Sciences of the United States of America* **87:** 364-368

Sternberg SH, Haurwitz RE, Doudna JA (2012) Mechanism of substrate selection by a highly specific CRISPR endoribonuclease. *Rna* **18:** 661-672

Sturm M, Bertsch A, Gropl C, Hildebrandt A, Hussong R, Lange E, Pfeifer N, Schulz-Trieglaff O, Zerck A, Reinert K, Kohlbacher O (2008) OpenMS - an open-source software framework for mass spectrometry. *BMC bioinformatics* **9:** 163

Swarts DC, Mosterd C, van Passel MW, Brouns SJ (2012) CRISPR interference directs strand specific spacer acquisition. *PloS one* **7:** e35888

Szecsi PB (1992) The aspartic proteases. *Scandinavian journal of clinical and laboratory investigation Supplementum* **210:** 5-22

Takai Y, Kishimoto A, Iwasa Y, Kawahara Y, Mori T, Nishizuka Y (1979) Calcium-dependent activation of a multifunctional protein kinase by membrane phospholipids. *The Journal of biological chemistry* **254:** 3692-3695

Terns MP, Terns RM (2011) CRISPR-based adaptive immune systems. *Curr Opin Microbiol* **14:** 321-327

Thomm M, Hauser W, Hethke C (1993) Transcription Factors and Termination of Transcription in Methanococcus. *Systematic and Applied Microbiology* **16:** 648-655

Tocchini-Valentini GD, Fruscoloni P, Tocchini-Valentini GP (2005) Structure, function, and evolution of the tRNA endonucleases of Archaea: an example of subfunctionalization. *Proceedings of the National Academy of Sciences of the United States of America* **102:** 8933-8938

Toh H, Ono M, Miyata T (1985) Retroviral gag and DNA endonuclease coding sequences in IgE-binding factor gene. *Nature* **318:** 388-389

van der Oost J, Brouns SJ (2009) RNAi: prokaryotes get in on the act. *Cell* **139:** 863-865

van der Oost J, Jore MM, Westra ER, Lundgren M, Brouns SJ (2009) CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends Biochem Sci* **34:** 401-407

Van Wart HE, Birkedal-Hansen H (1990) The cysteine switch: a principle of regulation of metalloproteinase activity with potential applicability to the entire matrix metalloproteinase gene family. *Proceedings of the National Academy of Sciences of the United States of America* **87:** 5578-5582

Wang R, Li H (2012) The mysterious RAMP proteins and their roles in small RNA-based immunity. *Protein science : a publication of the Protein Society* **21:** 463-470

Wang R, Preamplume G, Terns MP, Terns RM, Li H (2011) Interaction of the Cas6 riboendonuclease with CRISPR RNAs: recognition and cleavage. *Structure* **19:** 257-264

Wang R, Zheng H, Preamplume G, Shao Y, 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*

Weiland P (2010) Biogas production: current state and perspectives. *Applied microbiology and biotechnology* **85:** 849-860

Westra ER, Nilges B, van Erp PB, van der Oost J, Dame RT, Brouns SJ (2012a) Cascade-mediated binding and bending of negatively supercoiled DNA. *RNA biology* **9:** 1134-1138

Westra ER, Semenova E, Datsenko KA, Jackson RN, Wiedenheft B, Severinov K, Brouns SJ (2013) Type I-E CRISPR-Cas Systems Discriminate Target from Non-Target DNA through Base Pairing-Independent PAM Recognition. *PLoS genetics* **9:** e1003742

Westra ER, Swarts DC, Staals RH, Jore MM, Brouns SJ, van der Oost J (2012b) The CRISPRs, they are a-changin': how prokaryotes generate adaptive immunity. *Annual review of genetics* **46:** 311-339

Westra ER, van Erp PB, Kunne T, Wong SP, Staals RH, Seegers CL, Bollen S, Jore MM, Semenova E, Severinov K, de Vos WM, Dame RT, de Vries R, Brouns SJ, van der Oost J (2012c) CRISPR Immunity Relies on the Consecutive Binding and Degradation of Negatively Supercoiled Invader DNA by Cascade and Cas3. *Molecular cell* **46:** 595-605

Whitman WB, Pfeifer F, Blum P, Klein A (1999) What archaea have to tell biologists. *Genetics* **152:** 1245-1248

Wiedenheft B, Lander GC, Zhou K, Jore MM, Brouns SJ, van der Oost J, Doudna JA, Nogales E (2011a) Structures of the RNA-guided surveillance complex from a bacterial immune system. *Nature* **477:** 486-489

Wiedenheft B, van Duijn E, Bultema JB, Waghmare SP, Zhou K, Barendregt A, Westphal W, Heck AJ, Boekema EJ, Dickman MJ, Doudna JA (2011b) RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proc Natl Acad Sci USA* **108:** 10092-10097

Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America* **74:** 5088-5090

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences of the United States of America* **87:** 4576-4579

Wurtzel O, Sapra R, Chen F, Zhu Y, Simmons BA, Sorek R (2010) A single-base resolution map of an archaeal transcriptome. *Genome research* **20:** 133-141

Yosef I, Goren MG, Qimron U (2012) Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic acids research* **40:** 5569-5576

Yosef I, Shitrit D, Goren MG, Burstein D, Pupko T, Qimron U (2013) DNA motifs determining the efficiency of adaptation into the Escherichia coli CRISPR array. *Proceedings of the National Academy of Sciences of the United States of America* **110:** 14396-14401

Yoshinari S, Fujita S, Masui R, Kuramitsu S, Yokobori S, Kita K, Watanabe Y (2005) Functional reconstitution of a crenarchaeal splicing endonuclease in vitro. *Biochemical and biophysical research communications* **334:** 1254-1259

Zhang J, Rouillon C, Kerou M, Reeks J, Brugger K, Graham S, Reimann J, Cannone G, Liu H, Albers SV, Naismith JH, Spagnolo L, White MF (2012) Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. *Molecular cell* **45:** 303-313

Zhang Y (2008) I-TASSER server for protein 3D structure prediction. *BMC bioinformatics* **9:** 40

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

List of figures

Chapter I

Chapter II

Chapter III

Chapter IV

Chapter V

Chapter VI

List of Tables

Chapter II

Chapter III

Chapter IV

List of abbreviations

CRISPR-Cas

Cas	CRISPR associated
Cascade	CRISPR associated complex for antiviral defense
Cascis	CRISPR associated complex for integration of spacers
CRISPR	clustered regularly interspaced short palindromic repeats
crRNA	CRISPR RNA
Cse	Cas subtype *E. coli*
Csy	Cas subtype *Yersinia*
Nar	Nuclease-ATPase in Repair
PAM	protospacer adjacent motif
pre-crRNA	precursor CRISPR RNA
tracrRNA	trans activating CRISPR RNA

organisms

Ct, Cthe	*Clostridium thermocellum*
Mm, Mmar	*Methanococcus maripaludis*
Pf	*Pyrococcous furiosus*
Sso	*Sulfolobus solfataricus*

amino acids

A	Alanine
C	Cysteine
F	Phenylalanine
G	Glycine
H	Histidine
K	Lysine
M	Methionine
R	Arginine
Y	Tyrosine

nucleobases

A	Adenine
C	Cytidine
G	Guanine
T	Thymidine
U	Uridine

nucleotides

ATP	adenosine triphosphate
CTP	cytidine triphosphate
GTP	guanosine triphosphate
TTP	thymidine triphosphate
UTP	uridine triphosphate

chemicals and enzymes

ACN	acrylonitrile
DEPC	diethylpyrocarbonate
dsDNA	double strand DNA
DTT	Dithiothreitol
EDTA	ethylenediaminetetraacetic acid
EtOH	ethanol
FA	formic acid
GAT	glutamine PRPP amidotransferase
HD nuclease	histidine aspartate nuclease
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
IPTG	Isopropyl-ß-D-thiogalactosid
McC	*Methanococcus* complex media
NTA	nitrilotriacetic acid
PNK	polynucleotide kinase
PRPP	phosphoribosyl pyrophosphate
RGEN	RNA guided endonuclease
RNAP	RNA polymerase
SDS	sodium dodecyl sulfat
ssDNA	single strand DNA
ssRNA	single strand RNA
TALEN	transcription activator like effector nuclease
TBE	TRIS/Borate/EDTA

TCA	trichloracetic acid
tRNA	transfer RNA
ZFN	zinc finger nucleases

units

bp	base pair
ci	curie
kDA	kilo dalton
M, mM, µM, nM	Molar, millimolar, micromolar, nanomolar
MW	molecular weight
nt	nucleotides
OD_{xnm}	optical density at x nm excitation wave lenght
ppm	parts per million
rpm	revolutions per minute
T	temperature
UV	ultra violet

miscellaneous

ATCC	American Type Culture Collection
DSMZ	Deutsche Sammlung von Mikroorganismen und Zellkulturen
EMSA	electrophoretic mobility shift assay
ESI	electrospray ionisation
FPLC	fast protein liquid chromatography
HCD	higher energy collision dissociation
K_d	dissociation constant
k_{obs}	observed reaction rate
LB	Luria Bertani
LTQ	linear trap quadrupole
MS	mass spectrometry
PAGE	polyacrylamide gel electrophoresis
pdb	protein data base
RMSD	root mean square deviation
Seq.	sequencing
wt	wild-type

Curriculum Vitae

Personal Information

Name:	**Hagen Richter**
Date of Birth:	26 / 06 / 1985
Address:	Naunhofer Str. 10 D-04821 Waldsteinberg
E-mail:	hagen.richter.1985@gmail.com

School

1992 – 1996	Elementary school, Beucha, Saxony
1996 – 2004	Grammar school, Brandis, Saxony (A-levels: 1.8)

Military Service

2004 – 2005	Basic military service, Rotenburg a. d. Fulda and Hammelburg

Academics

2005 – 2008	Bachelor of Science in Biotechnology at the Technical University of Braunschweig (1.9) Bachelor thesis in the laboratory of Prof. Dr. Petra Dersch: "Molecular analyses of the role of regulatory RNAs in the regulation of the transcriptional factor RovA in *Yersinia pseudotuberculosis*" (1.3)
2008 – 2010	Master of Science in Biotechnology at the Technical University of Braunschweig (1.3)
2009 – 2010	Master thesis in the laboratory of Dr. B. Duncker at the University of Waterloo (Ontario, Canada): "The Consequences of perturbing pre-RC subunit levels on DNA replication and genome integrity." (1.1)
2010 ongoing	Ph.D. student in the laboratory of Dr. L. Randau at the Max-Planck-Institute for Terrestrial Microbiology "Cas6b and Cas8b, the subtype-specific proteins of CRISPR-Cas subtype I-B in *Methanococcus maripaludis*"
2010 ongoing	Associate member of the International Max Planck Research School for Environmental, Cellular and Molecular Microbiology (IMPRS-MIC)

Scientific Engagement

08/2008 - 10/2008	Student assistant in the labs of Prof. Dr. Petra Dersch (infection biology, Braunschweig)
06/2009 - 07/2009	Student assistant in the labs of Prof. Dr. R. Hehl (plant, yeast genetics, Braunschweig)
10/2011 - 10/2013	IMPRS student representative (Marburg)

Soft Skills

Language:	German (native) English (fluent in spoken and written language) French (school knowledge, 4 years)
IT:	MS Office, Apple Keynote Adobe Photoshop Vector NTI, EndNote

Other Interests

Photography
Sports (Soccer, Squash, Cycling)
Travelling

Danksagung

Die letzten drei Jahre in denen ich meine Doktorarbeit angefertigt habe, war eine besondere Zeit, die meinen zukünftigen Werdegang sicherlich stark beeinflussen wird. Während dieser Zeit habe ich sehr viel Unterstützung erhalten für die ich mich an dieser Stelle bedanken möchte.

Zunächst möchte ich mich bei Dr. Lennart Randau dafür bedanken, dass er mir, ohne mich zuvor persönlich gesehen zu haben, die Möglichkeit gegeben hat, meine Arbeit in seiner Arbeitsgruppe anzufertigen. Ich bin auch sehr dankbar, dass er für sämtliche Fragen immer eine offene Tür hatte und mit Rat und Tat zur Seite stand.

Ein weiterer Dank geht an Prof. Dr. Kai Thormann, welcher nicht nur 3 Jahre lang Teil meines Thesis Advisory Committee's war, sondern sich auch dazu bereit erklärt hat, das Zweitgutachten für meine Arbeit zu übernehmen.

In diesem Zusammenhang möchte ich mich auch bei Prof. Dr. Martin Thanbichler und Prof. Dr. Stefan Bauer dafür bedanken, dass Sie zugestimmt haben, meiner Prüfungskommission anzugehören.

Für die drei jährige Unterstützung während meiner Doktorarbeit möchte ich mich auch bei meinem Thesis Advisory Committee, bestehend aus PD Dr. Sonja-Verena Albers, Prof. Dr. Kai Thormann und PD Dr. Werner Liesack, bedanken.

Besonders bedanken möchte ich mich auch bei Dr. André Plagens, welcher mir stehts mit seinem Tipps und Tricks zur Seite stand und immer ein offenes Ohr für Fragen sowie anregende Diskussionen hatte.

Ein großer Dank gilt auch den restlichen Mitgliedern der AG Randau ebenso wie der AG Stukenbrock, für eine angenehme Arbeitsatmosphäre sowie einer schönen Zeit auch außerhalb des Labors, auf dem Fußballfeld, auf dem Squashcourt oder in den Bars bei dem einen oder anderen Bier.

Zusätzlicher Dank gebührt natürlich auch den Freunden außerhalb Marburgs, welche für den einen oder anderen Tapetenwechsel während dieser Zeit gesorgt haben. Dabei möchte ich besonders Eike, Janina, Martin, Meusel, Timon, Ronny, Markus, Alex und Katja danken. Norman möchte ich an dieser Stelle insbesondere für die riesige Unterstützung bei der Anfertigung einiger Abbildungen danken, welche sicherlich niemals so professionell ausgesehen hätten ohne seine Hilfe.

Nicht zuletzt möchte ich mich auch bei meiner Freudin Michi bedanken, welche mich während dieser Zeit unterstützt hat und auch meine Macken, besonders das frühe Aufstehen, über sich hat ergehen lassen.

Der größte Dank gilt meiner Familie, welche mich während meiner ganzen Ausbildung unterstützt hat und auch während der letzten drei Jahre ständig für mich da war und immer wieder aufmunternde Worte parat hatte.

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich meine Dissertation mit dem Titel: "Characterization of the CRISPR-Cas subtype I-B proteins Cas6b and Cas8b of *Methanococcus maripaludis* C5" selbständig, ohne unerlaubte Hilfsmittel angefertigt und mich keiner anderen als der von mir ausdrücklich bezeichneten Quellen und Hilfen bedient habe.

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

Marburg, den 05.11.2013 Hagen Richter

Erklärung des Eigenanteils

Richter H, Lange SJ, Backofen R, Randau L (2013) Comparative analysis of Cas6b processing and CRISPR RNA stability. *RNA Biology* **10:** 700-707

Die Versuche wurden von Hagen Richter durchgeführt. Die bioinformatische Auswertung verschiedener Repeatsequenzen wurde von Sita Lange und Rolf Backofen ausgeführt. Die Veröffentlichung wurde von Hagen Richter und Lennart Randau geschrieben.

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

Versuche am I-B system von *M. maripaludis* welche zu den Abbildungen 1 und 3-6 geführt haben wurden von Hagen Richter durchgeführt. Die Analysen des Systems von *C. thermocellum* waren Experimente von Judith Zöphel und haben zu Abbildung 2 und den Supplements beigetragen. Bioinformatische Auswertungen der Daten wurde von Daniel Maticzka und Rolf Backofen übernommen. Die Veröffentlichung wurde von Hagen Richter und Lennart Randau geschrieben.

Marburg, den 05.11.2013 Hagen Richter