Investigation of prokaryotic immune defense system with quantitative and structural mass spectrometry

Dissertation

for the award of the degree

"Doctor rerum naturalium" (Dr.rer.nat.)

of the Georg-August-Universität Göttingen

within the doctoral program *Molecular Biology*

of the Georg-August University School of Science (GAUSS)

submitted by

Kundan Sharma

from New Delhi, India

Göttingen, 2015

Thesis Committee

Prof. Henning Urlaub	Bioanalytical Mass Spectrometry Group Max Planck Institute for Biophysical Chemistry, Göttingen Bioanalytics, Department of Clinical Chemistry, University Medical Centre Georg-August-Universität, Göttingen
Prof. Jörg Stülke	Department of General Microbiology Institute for Microbiology and Genetics Georg-August-Universität, Göttingen
Prof. Peter Rehling	Department of Cellular Biochemistry University Medical Centre Georg-August-Universität, Göttingen

Members of the Examination Board

Prof. Henning Urlaub (Reviewer)	Bioanalytical Mass Spectrometry Group Max Planck Institute for Biophysical Chemistry, Göttingen Bioanalytics, Department of Clinical Chemistry, University Medical Centre Georg-August-Universität, Göttingen
Prof. Jörg Stülke (Second Reviewer)	Department of General Microbiology Institute for Microbiology and Genetics Georg-August-Universität, Göttingen

Further members of the Examination Board

Prof. Peter Rehling	Department of Cellular Biochemistry University Medical Centre Georg-August-Universität, Göttingen
Prof. Patrick Cramer	Department of Molecular Biology Max Planck Institute for Biophysical Chemistry, Göttingen
Prof. Uwe Groß	Department of Medical Microbiology University Medical Centre Georg-August-Universität, Göttingen
Prof. Rolf Daniel	Department of Genomics and Applied Microbiology Institute for Microbiology and Genetics Georg-August-Universität, Göttingen

Date of the oral examination: 29th April, 2015

Affidavit

I hereby declare that the presented thesis entitled “Investigation of prokaryotic immune defense system with quantitative and structural mass spectrometry” has been written independently and with no other sources and aids than quoted.

Gottingen, 24th March 2015

Kundan Sharma

Abstract

Bacteria and archaea have evolved an adaptive and heritable immune defense system comprising a CRISPR (clustered regularly interspaced short palindromic repeats) locus and Cas (CRISPR-associated) proteins that targets mobile genetic elements such as phages and plasmids. The molecular memory of previous infections is transcribed into crRNAs (CRISPR RNAs) which serve as a template to guide the hydrolysis of incoming foreign genetic material. The CRISPR-Cas system is divided into three Types (I, II and III) on the basis of signature *cas* genes and further subtypes defined by the protein machinery and target specificity. During my PhD work, I investigated the Type I and Type III CRISPR-Cas systems using quantitative and structural mass spectrometry approaches.

The first part of this work focused on using relative quantitative approach to study the effect of a *cas* gene deletion on the expression levels of other Cas proteins in *Haloferax volcanii*. A comparison between the proteomes of *H. volcanii* wild type and deletion strains using stable-isotope dimethyl labeling showed that the removal of a *cas* gene is compensated by an overall increase in the expression of *cas* gene cluster. In addition, the absolute amounts of Cas protein components in multi-subunit Cascade complexes from *H. volcanii* and *Clostridium thermocellum* were determined using intensity based absolute quantification. The results were used to determine the stoichiometry of Cas proteins in these multi-subunit protein complexes which is valuable for the further investigation of molecular interactions within these complexes.

Further in this work, the UV induced protein-RNA cross-liking approach was utilized to investigate RNA binding regions in single (recombinant) Cas proteins such as the archaeal and bacterial Cas6b proteins and the Cas7 family proteins from four different organisms. These structural studies were also extended to multi-subunit crRNP complexes such as the Type I-E Cascade complex from *E. coli* and Type III-A Csm complex from *Thermus thermophilus*. The information derived from the cross-linking studies could validate several protein-RNA interactions reported earlier in X-ray crystallography studies. In addition to identifying new RNA binding regions in the Cas proteins, the identified cross-links could also be mapped to conserved regions of these proteins within modified RNA binding domains. The results provided unbiased evidence of direct protein-RNA interaction in *in vitro* and *in vivo* purified crRNP complexes. Lastly, a lysine directed chemical cross-linking approach is presented for the investigation of protein-protein interactions between different Cas proteins in the *C. thermocellum* Cascade complex where more than 126 inter-protein interactions were identified. These results constitute the first step towards MS based structural modeling of crRNP complexes.

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 The immune system of bacteria and archaea

The viruses that infect bacteria (bacteriophages) and archaea are the most abundant forms of life on this planet, even so that they have outnumbered their hosts in various orders of magnitude [1]. To survive the predation from these viruses both bacteria and archaea have evolved various defense mechanisms, such as modification of surface receptors to prevent virus adsorption, restriction enzymes (Restriction-Modification systems, RM systems) for nucleolytic cleavage of non-self DNA and abortive infection by undergoing lysis and sacrificing the infected cell [2]. These are referred to as the innate immune responses in prokaryotes. However, it is not just the viruses that represent a threat to these organisms. They are constantly exposed to foreign genetic material that is exchanged among related or unrelated species by various mechanisms of horizontal gene transfer (HGT) including transformation, conjugation and transduction [3, 4]. The acquisition of this foreign genetic material through viruses or HGT might not always be beneficial for the host and may lead to host cell lysis and death; therefore bacteria and archaea have developed an adaptive immune response for protection against these mobile genetic elements such as viruses and plasmids. Recently, an adaptive, heritable immune response has been identified in microbes, the CRISPR-Cas (clustered regularly interspaced short palindromic repeats – CRISPR associated) system [5]. Throughout this thesis, both conserved and unique features of the CRISPR based adaptive immune response found in prokaryotes have been investigated.

1.1.1 CRISPR-Cas system

Computational analyses have revealed that the genomes of around 90% of archaea and 40% bacteria comprise a family of DNA repeats known as CRISPR [6, 7]. These repeats are interspaced with non - repetitive spacer units which are acquired during the invasion of foreign genetic material. The spacer units in the CRISPR loci serve as a genetic memory for the acquired immune response, because they reflect the number of different phages and plasmids that were encountered by the host during past infections. These loci express small CRISPR RNAs that would target the invading DNA or RNA with complementary sequence during a subsequent infection. The mechanism of recognition and degradation of foreign genetic material is analogous to the RNA interference mechanism in eukaryotes. It also reflects the survival ability

of a bacterium or an archaeon because an expanded CRISPR locus would render the host with an efficient defense mechanism against mobile genetic elements [8].

1.1.1.1 CRISPR locus

A hallmark feature of CRISPR-Cas system is the CRISPR locus. The first CRISPR array was described in *E.coli* in 1987 with 14 repeats of 29 base pairs that are interspaced with 32-33 bp spacer sequences [9, 10]. It was based on the particular structure of this loci that led Jansen and co-workers coin the term CRISPR in 2002 [11]. As the name suggests, a CRISPR locus consists of palindromic repetitive sequences (the repeats) that are separated by similar sized spacer sequences which are identical to the fragments of plasmids and viral genomes and therefore they specify the targets of CRISPR interference (Figure 1.1).

The length of a CRISPR locus can vary in different microbial species due to several reasons. The size of both repeat and spacer units can vary between 25 – 40 bp [12], the number of repeats and spacers can range between a few to several hundred with an average around 66 [8] and the number of CRISPR loci per genome can also vary, mostly there is a single CRISPR locus but there are exceptions with 18 clusters in *Methanocaldococcus jannaschii* [13]. These sequences are preceded by a leader sequence that is AT rich, several hundred bp in length, but not conserved between different species [11]. The new spacer elements are inserted near the leader sequence which also comprises of binding sites for regulatory proteins that control spacer acquisition. Preceding or following the repeats are a set of CRISPR associated genes (*cas* genes) which encode the Cas protein machinery responsible for CRISPR activity. These *cas* genes also form the basis of CRISPR classification.

1.1.1.2 Cas (CRISPR-associated) proteins

In addition to the CRISPR locus, the Cas proteins encoded by *cas* genes are key players in the immune defense, they are responsible for mediating the adaptive immune response. Cas proteins are highly diverse in terms of functionality as they can act as a single protein catalyzing endonucleolytic cleavage of the target DNA, processing of CRISPR RNAs (crRNAs) and can even come together in the form of multi-subunit CRISPR ribonucleoprotein (crRNP) complexes along with the crRNAs for the processing of CRISPR loci transcripts as well as targeting and cleavage of invading DNA. The Cas proteins have been observed to comprise of several nuclease domains, distinct helicase domains and also certain domains that are characteristic of RNA

binding proteins [11]. The most conserved domains across various classes of RNA binding proteins are the RNA Recognition Motifs (RRMs) [14], which are also found in a special category of Cas proteins known as Repeat associated mysterious proteins (RAMPs) [15]. Recently solved crystal structures of RAMPs indicate the presence of one or two domains similar to the RRM (also called the ferredoxin fold) [16-19]. An overview of the major Cas protein families and their characteristics and distinct functions are summarized in Table 1.1.

Table 1.1 Overview of major Cas protein families, the core component of CRISPR-Cas systems. Based on [20-24].

Cas Protein Family	Nomenclature in CRISPR subtypes	Characteristics and functions
Cas1	Type I, II, III: Cas1	Metal dependent endonuclease, targets dsDNA, ssDNA or branched DNA in sequence independent manner. Involved in spacer acquisition.
Cas2	Type I, II, III: Cas2	Metal dependent nuclease, RAMP-like fold with $\beta_1\alpha_1\beta_2\beta_3\alpha_2\beta_4$ arrangement. Involved in spacer acquisition.
Csn2	Type II-A: Csn2	Accessory role during spacer acquisition (lacks catalytic activity).
Cas3	Type I: Cas3 Cas3' Helicase domain Cas3'' Nuclease HD domain	ssDNA nuclease and ATP dependent helicase, both characteristics significant during interference, for strand separation and target DNA cleavage.
Cas4	Type I-A-D and II-B: Cas4	Stabilization of Cas1-Cas2 complex during spacer acquisition.
Cas5	Type I: Cas5, III-A: Csm4, III-B: Cmr3	Consists of RRM domains, belongs to RAMP superfamily. Part of the crRNP complex, important during Interference. When catalytically active, can substitute Cas6 in crRNA processing.
Cas6	Type I and III: Cas6	Metal dependent endoribonuclease, consists of RRM domains, belongs to RAMP superfamily, involved in crRNA processing
Cas7	Type I-A: Csa2, I-B/C/E: Cas7, I-D: Csc2, I-F: Csy3, III-A: Csm3, III-B: Cmr4	Consists of RRM domains, belongs to RAMP superfamily. Multiple copies that form the backbone of crRNP complex. Important during the interference step. Cmr4: catalytic Cas7 protein responsible for target RNA cleavage.
Large subunit	Type I: Cas8 homologs I A-C: Cas8a-c, ID: Cas10 I-E: Cse1, I-F: Csy1, Type III: Cas10	Consists of RRM domains. Interacts with the Cas7 and Cas5 in the multi-subunit crRNP complex, capping of the 5'end of crRNA. Important during interference step.

	homologs III-A: Csm1, III-B Cmr2	
Small subunit	Type I-A: Csa5, I-E: Cse2, III-A: Csm2, III-B: Cmr5	Present in multi-subunit crRNP complexes, Weakly interacts with other proteins such as Cas7. Important in interference step.
Cas9	Type II	Involved in both crRNA processing and cleavage of the target DNA

1.1.2 The three steps of CRISPR-Cas action

The CRISPR-Cas based immunity is an acquired form of immune response where the mode of action involves three distinct stages: adaptation, expression and interference (Figure 1.1).

1.1.2.1 Adaptation by spacer acquisition

CRISPR loci acquire the fragments of invading DNA and therefore these newly acquired spacers would result in a sequence specific resistance mechanism to the corresponding phage. The new spacers are integrated into the CRISPR locus in a polarized manner, starting from the leader end [5, 25]. The spacers with specific protospacer adjacent motif (PAM) are selected by Cas proteins from the invading DNA and integrated into the CRISPR locus in a PAM dependent orientation [26-28]. The most highly conserved Cas proteins Cas1 and Cas2 have been reported to have a key role in this spacer acquisition process [5, 29]. This is how the cell is able to adapt to the mobile genetic elements in the environment, hence this stage is referred to as “Adaptation” phase [8].

1.1.2.2 Expression of CRISPR transcripts

The CRISPR locus is transcribed into precursor crRNA or pre-crRNA. The pre-crRNA is then processed into smaller crRNA with the endonucleases such as Cas6 homologs (which might be present as a part of multi-subunit crRNP complex) or housekeeping endonucleases such as RNase III. At this step there are differences in the different CRISPR systems in terms of key candidates involved in crRNA processing (these distinct features are discussed under section 1.1.3). The crRNA generated, comprises a part of repeat sequences flanking both ends of a complete spacer. The spacer sequence in this crRNA is then responsible for guiding the crRNP complex towards a complementary target sequence in any invading mobile genetic element such as a viral DNA or a plasmid [12, 22].

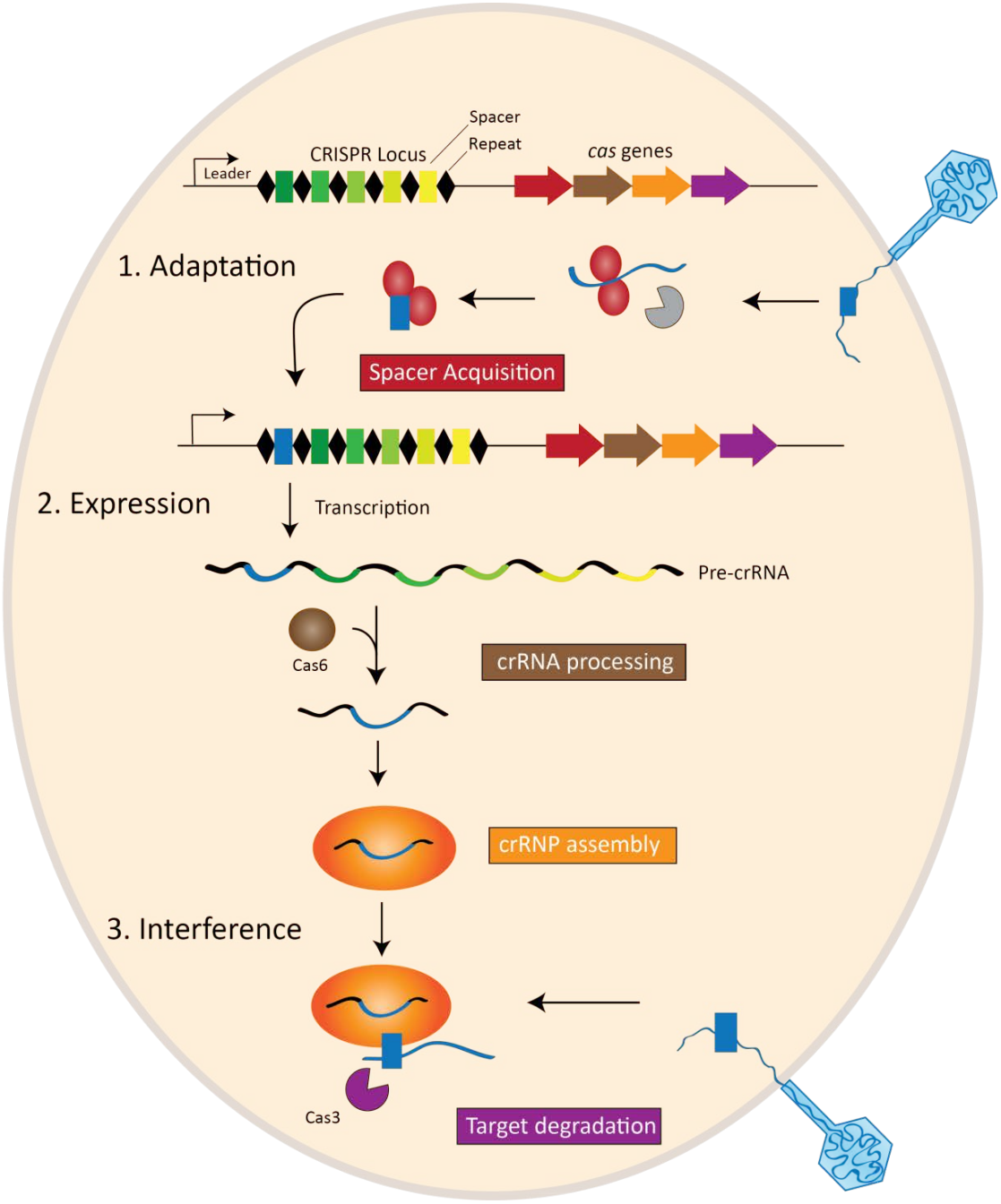


Figure 1.1 Overview of the CRISPR-Cas Type I-E system in *E. coli*.

The CRISPR locus showing the palindromic repeats (black) separated by spacers (green) that are acquired from the invading genetic elements such as bacteriophage DNA. The locus is preceded by a leader sequence (arrow) and a followed by set of *cas* genes. The *cas* genes are indicated with different colors (red, brown, orange and violet) with respect to their role in the CRSIPR system (indicated in colored boxes). The CRISPR immunity is mediated by crRNAs and Cas proteins that together form the crRNP complex (orange). The three stages of the CRISPR based adaptive immune response: (i) Adaptation: Acquisition of new spacer units (blue), (ii) Expression: Transcription of CRISPR locus and processing of the crRNAs by Cas6 (brown) (iii) Interference: Target surveillance by the crRNP complex and degradation of invading DNA by endonucleases e.g. Cas3 (violet). Adapted from [22] with permission from the publisher.

1.1.2.3 RNA directed CRISPR interference

The mature crRNAs associate with Cas proteins to form crRNP complexes, which could comprise of multiple Cas protein subunits like the Type I and Type III systems or a single large protein (Cas9) like in Type II system. The assembled crRNP complex then scans the invading genetic element for a sequence similarity between the crRNA and a protospacer sequence. Hybridization of the crRNA and target strand, results in a conformational change in the crRNP complex which acts as a signal for the activation of endonucleases (Type I and III-B) or intrinsic nuclease activity of the crRNP complex (Type II and Type III-A) for the degradation of target DNA [30]. (Further details of the activity of different crRNP complexes and candidate proteins that assemble together to form these complexes is discussed under section 1.1.3).

1.1.3 Three types of CRISPR-Cas systems

The recent classification of CRISPR-Cas systems by Kira S. Makarova divides them into three distinct Types I, II and III [21, 23, 31]. Two universal *cas* genes *cas1* and *cas2* are present in all CRISPR subtypes, and they play a significant role in the spacer acquisition process [32-34]. The adaptation is therefore very similar in all the three types of CRISPR-Cas systems. Substantial difference between the three types lies in their sets of constituent genes and signature *cas* genes. These include: In Type I, *cas3* gene (comprising of both helicase and nuclease domain) [35]; in Type II, *cas9* gene (a large protein that singularly controls the process of crRNA processing and interference) and in Type III, *cas10* gene (the large subunit, important during interference). Further the three types are divided into various subtypes along the phylogeny of universal *cas1* gene [21]. The characteristics of different Cas protein components of the three systems are also described in Table 1.1.

1.1.3.1 Type I CRISPR-Cas systems

In Type I system the spacer acquisition, like all CRSPR-Cas subtypes, is mediated by Cas1 and Cas2 proteins. The distinct characteristics of Type I system that makes them different from the rest of CRISPR-Cas types are: the Cas6 endonuclease responsible for crRNA processing, the CRIPSR-associated complex for antiviral defense (Cascade) that is formed by assembly of multiple Cas proteins and the crRNA, the Cas3 endonuclease responsible for the degradation of target DNA [36-38], also illustrated in Figure 1.2.

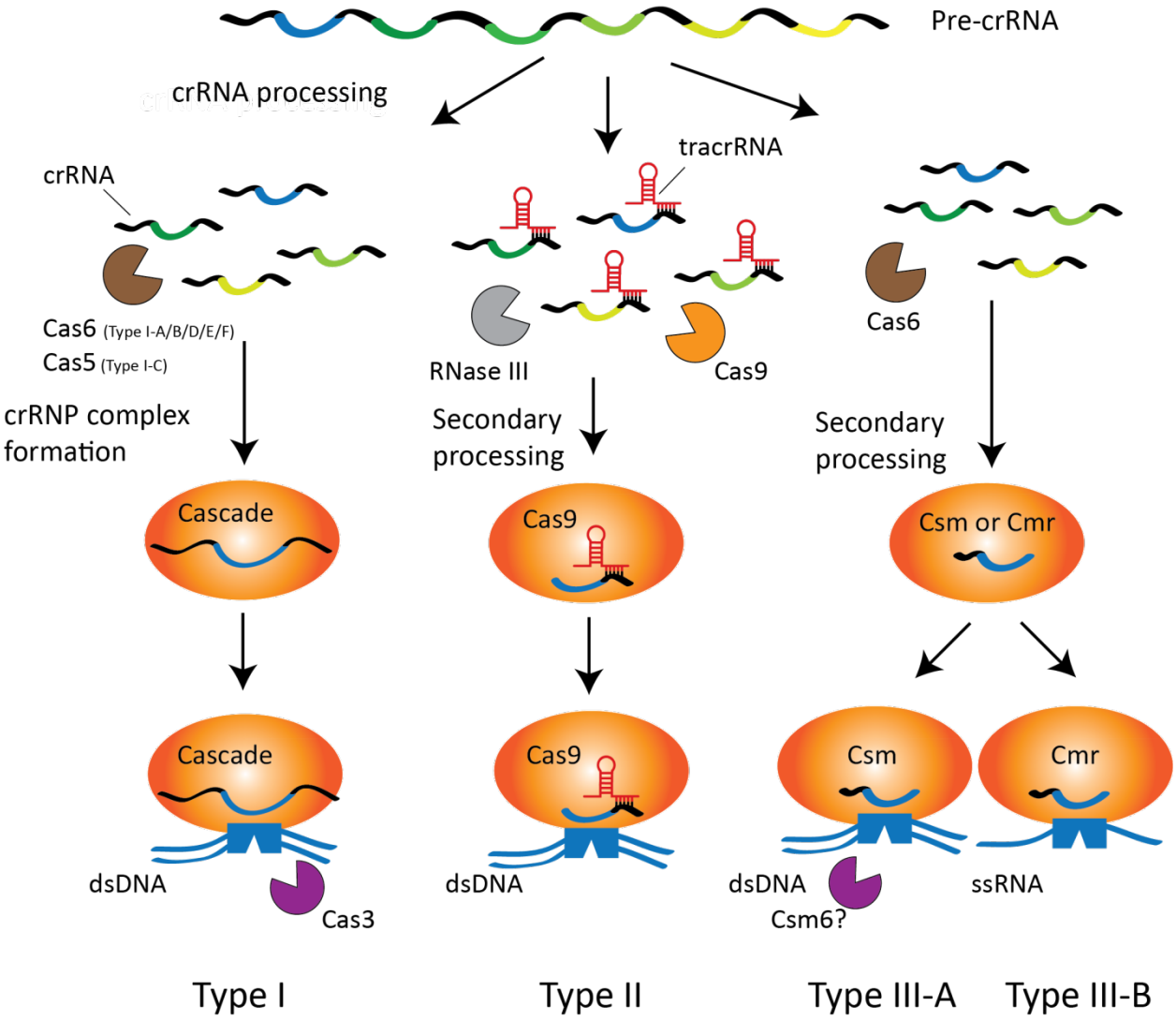


Figure 1.2 RNA directed CRISPR interference in the three types of CRISPR-Cas systems.

Adaptation phase of the CRISPR immunity is common in all the three types of CRSIPR-Cas systems (see Figure 1.1). However the key Cas protein candidates involved in processing of crRNA, formation of multi-subunit or single protein crRNP complex and target surveillance and degradation are the major criteria of difference between the subtypes. The three types are therefore characterized by the distinct features of their Cas proteins. In Type I, the crRNA are processed mostly by Cas6 (Figure 1.3) and other Cas proteins form multi-subunit Cascade complex that targets dsDNA. In Type II Cas9 is the sole player mediating these roles, whereas in Type III the assembled multi-subunit Csm or Cmr complex is similar to the Cascade complex, with DNA or RNA as target. Adapted from [12] with permission from the publisher.

The Cas6 endonucleases are responsible for the processing of pre-crRNA, resulting in a processed mature crRNA [24, 39]. This crRNA has three components (i) the complete spacer (ii) upstream of the spacer, 8 nucleotides (nt) derived from the repeat and (iii) downstream of the spacer a sequence of invariable size derived from the downstream repeat, comprising of a palindromic repeat that tends to form a stem-loop structure [30, 36, 40] (Figure1.3).

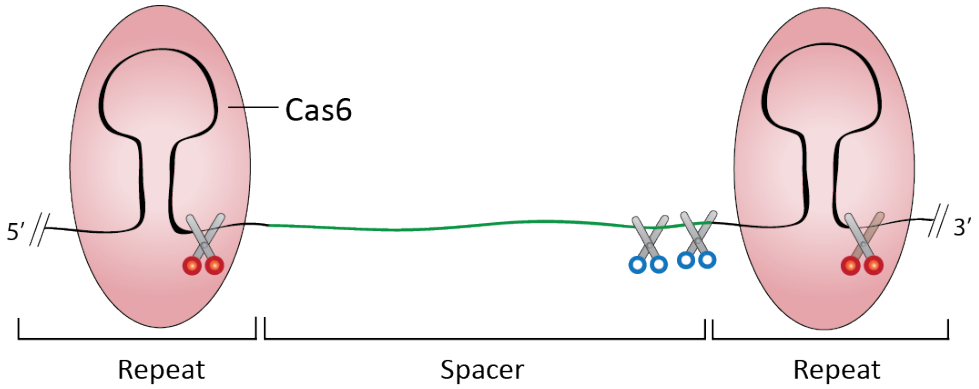


Figure 1.3 Generation of processed crRNA in Type I systems.

Primary processing of pre-crRNA is performed by Cas6 by cleavage (red scissors) within the repeat sequences resulting in a crRNA with 5' handle of 8 nt, a central spacer sequence and a longer 3' handle. In some subtype the 3' handle forms a stem-loop structure and in some CRISPR-Cas systems the 3' handle is further processed by unknown ribonucleases (blue scissors). Adapted from [22] with permission from the publisher.

In the Type I-E system from *E. coli*, the mature crRNA and Cas proteins together form the Cascade complex. The complete complex has a stoichiometry of $Cse1_1Cse2_2Cas5_1Cas7_6Cas6_1$ which is a typical feature in Type I and Type III complexes [30, 41]. The crRNP complexes over different CRISPR subtypes share a helical backbone formed of multiple copies of Cas7 protein, Cas5 and Cas8 proteins assembled around the crRNA with an addition of a small subunit in some cases (Figure 1.4). Due to these structural similarities the crRNP complexes in other subtypes of Type I are also referred to as Cascade complex [24, 42].

The fully assembled Cascade complex then scans the target dsDNA for a PAM (protospacer adjacent motif). On recognition of PAM by Cse1, the dsDNA destabilizes allowing the hybridization between the spacer region of crRNA and the complementary sequence on target DNA [37]. This base pairing results in an R-loop formation further triggering a conformational change in the Cascade complex [30, 43]. This structural change acts as a signal for the activation of Cas3. The Cas3 using both its helicase and nuclease activity, unwinds and then degrades the target DNA [36, 44].

1.1.3.2 Type II CRISPR-Cas systems

The Type II system is most distinct from all other CRISPR subtypes. The only similarity with other types is in terms of spacer acquisition by Cas1 and Cas2. The signature protein of Type II system is Cas9, a large protein which acts as an endonuclease. It works as single protein machinery for the generation of mature crRNAs as well as the cleavage of target DNA. Recent reports of the high resolution crystal structures of Cas9 from *Streptococcus pyogenes* and

Actinomyces naeslundii have been a major achievement in the understanding of this system [45, 46].

The CRISPR locus of Type II system comprises of a gene (in addition to the *cas* genes) for the synthesis of trans-activating crRNA (tracrRNA). The tracrRNA has a sequence complementarity to the repeat region of the pre-crRNA and a duplex formation between the two results in processing of the crRNA-tracrRNA (dsRNA) hybrid by RNase III in presence of Cas9 [47]. This mature crRNA-tracrRNA hybrid is then responsible for the target recognition in a PAM dependent manner followed by cleavage [48, 49].

In the crystal structures it was shown that Cas9 has two distinct nuclease domains. The HNH domain responsible for the cleavage of target DNA (the one complementary to the guide RNA sequence) and a RuvC nuclease domain that cleaves the non-target strand (non-complementary strand), leading to double strand breaks in the target DNA [48, 50]. This ability of Cas9 for creating dsDNA breaks at specific sites defined by a guide RNA has led to its use as versatile tool in genome engineering [51].

1.1.3.3 Type III CRISPR-Cas systems

The Type III systems are characterized with the presence of a signature gene *cas10*, that encodes the large subunit Cas protein, homologous to palm-domain polymerases. Also there are multiple genes encoding for RAMPs. Further the Type III systems are classified into subtypes based on *cas1* gene phylogeny. There are two major subtypes, Type III-A systems have a signature *csm2* gene and Type III-B systems have a signature *cmr5* gene [21].

The crRNA biogenesis in Type III system is very similar to the Type I system where the sole player, Cas6 endonuclease, mediates the processing or pre-crRNAs into mature crRNAs. Also, the architecture of crRNP complexes in both Type I and Type III complexes have a lot of structural similarities [52-54] also depicted in Figure 1.4. The Type III-A Csm complex has a helical backbone of multiple copies of Csm3 and in Type III-B this backbone comprises of Cmr4 proteins, in a similar morphology as Cas7 proteins in Type I-E Cascade complex. Also the crystal structure of Csm3 and Cmr4 revealed structural homology with Cas7 protein [55, 56].

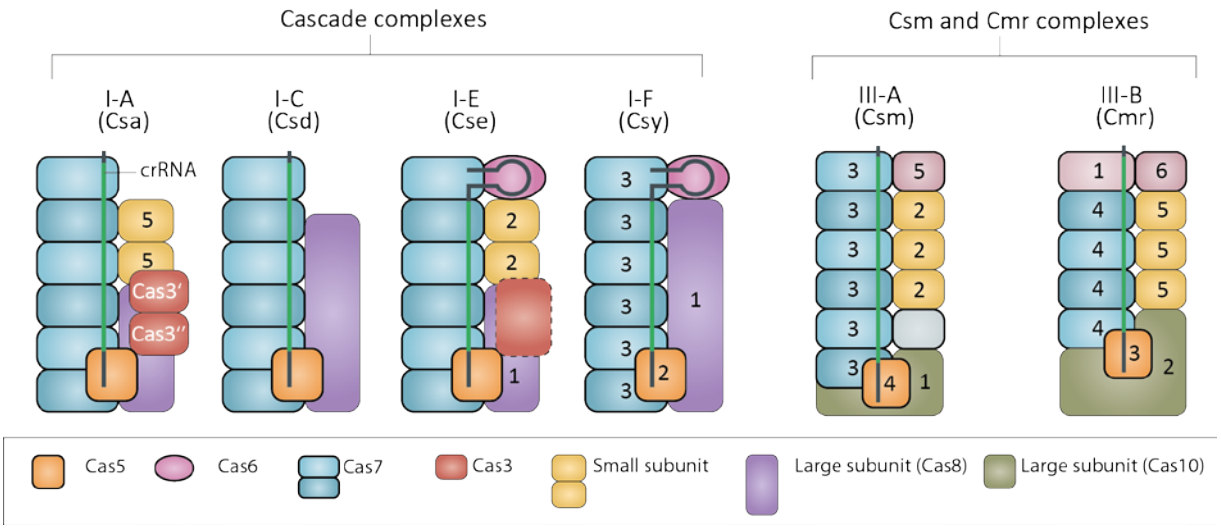


Figure 1.4 Schematic representations of assembled crRNP complexes from Type I and Type III CRISPR-Cas systems.

The Cas proteins are assembled around the crRNA with a spacer component (green) and flanking repeats (black). The colored boxes/circles represent different Cas proteins components of the crRNP complexes, as indicated in the box below. Same color across different subtypes represents the homology of conserved Cas proteins. Numbers represent the nomenclature of a particular Cas protein in a given subtype, for e.g. The Cas 7 protein (blue) referred to as Csy3 in subtype I-F, Csm3 in subtype III-A and Cmr4 in subtype III-B. The two truncated domains of Cas3 are shown as Cas3' and Cas3'' in Type I-A Csa complex and as a fused component of the Type I-E Cascade complex. Adapted from [22] with permission from the publisher.

In the interference stage, there is a major difference in the two Type III subtypes:

- The Type III-A Csm complex targets dsDNA in a PAM dependent manner, with Cas10 also playing a significant role [57]. Also a helicase/nuclease Csm6 has been speculated to be involved in target degradation [58]. In addition, recent reports for the Type III-A Csm complex from *T. thermophilus* suggest that the Type III-A system targets RNA in a flexible manner without relying on a PAM [59].
- The Type III-B Cmr complex has been shown to be the most unique out of all CRISPR-Cas systems in terms of targeting RNA and not DNA. The Cmr4 protein that forms the helical backbone of the Cmr complex was suggested as the catalytic subunit due to a multiple catalytic sites observed along the backbone of Cmr complex in *in vitro* experiments [60].

1.2 Mass spectrometry

Mass spectrometry (MS) is an analytical technique for the identification of compounds based on their elemental composition and charged state [61]. In a MS analysis, the chemical compounds are ionized in the gas-phase to generate charged molecules which are then measured on the basis of their mass-to-charge ratio and abundance. It is very widely used for the analysis of simple and complex biological samples such as proteins, nucleic acids, lipids and macromolecular complexes.

A mass spectrometer consists mainly of three components: i) Ion source - to produce multiply charged ion droplets from the sample, ii) mass analyzer - to separate the ions based on their mass-to-charge (m/z) ratio and iii) detector - to count the number of ions at each m/z value emerging from the analyzer (Figure 1.5). Once the gas-phase ion droplets enter the instrument, they are inside a vacuum system comprising a mass analyzer and a detector. Before entering the mass analyzer the solvent in the droplets evaporates under high temperature. In the mass analyzer, ions are guided and separated by electric or magnetic fields according to their m/z ratio. After separation in the mass analyzer, ions are measured by a detector which transforms them into usable signals comprising the information of their m/z ratio and abundance. The detector is coupled to a data processing system which uses specialized software to produce a suitable form of mass spectrum for data analysis [61].

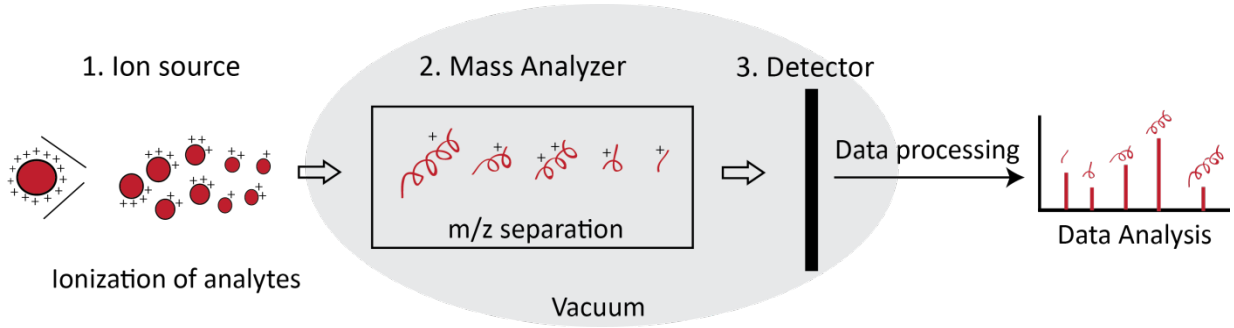


Figure 1.5 Basic components of a mass spectrometer.

The mass spectrometer consists of three main components: Ion source for producing ions from the sample, mass analyzer for m/z based separation and a detector for measuring these ions and their abundance. The mass analyzer and detector are present inside the vacuum chamber of the instrument. The signals from the detector are then transferred to dedicated software to process the data for data analysis purposes.

MS has become very popular in the last two decades due to the developments in soft ionization techniques such as electrospray ionization (ESI) and matrix-assisted laser desorption ionization

(MALDI) and it has become a method of choice for analyzing biological macromolecules, especially proteins. In addition, in 2002, the Nobel Prize in Chemistry was awarded to John Fenn and Koichi Tanka for the development of ESI and MALDI, respectively [62, 63].

1.2.1 Electrospray Ionization

Electrospray ionization (ESI) is used to ionize the analytes out of a solution using high voltage. Due to the rapid transition of ions from liquid to gas phase, the ESI ion source is mostly coupled to a liquid chromatography (LC) system. The LC column is a narrow capillary that tapers into a fine needle tip at the end made of glass or metal, which makes sure that the outgoing liquid is sprayed in form of droplets. These droplets undergo extensive evaporation and solvent loss, facilitated by the presence of volatile organic solvents such as acetonitrile in the solution. In order to enable the ionization in positive mode, a proton rich environment is needed for which the pH of solution is kept very low using volatile acidic components such as formic acid. The spray needle is held at high potential difference (several kV) from the inlet of the mass spectrometer, to make sure these droplets undergo intense electrostatic dispersion due to repulsion of like charges resulting in smaller droplets [64, 65]. This is also referred to as 'Coulombic fission' where the original droplet bursts creating more stable and smaller droplets [66]. These droplets further vaporize as they reach closer to the heated inlet of the mass spectrometer, becoming smaller as the like-charge repulsion increases, leading to further dispersion of these droplets. This phenomena is currently supported by two coexisting theories: i) Charge residue model – the cycle of coulombic fission and evaporation repeats until there is only one analyte ion left per droplet and ii) Ion evaporation model – the big and highly charged droplets burst to produce free ions [67]. The final charge that is present on the naked ion generated after this process allows the mass spectrometer to accelerate these ions through the remaining system.

1.2.2 Mass spectrometry instruments

All mass spectrometers comprise three basic components as shown in Figure 1.6, however the nature of these components varies with respect to the type of data to be generated and the kind of sample to be analyzed. One of the key distinguishing features in these mass spectrometers is the mass analyzer.

A mass analyzer is the core of a mass spectrometer that separates the ions based on their m/z ratio. Based on the principle of how m/z separation is achieved there is a wide variety of mass analyzers available. The three most common examples include:

1. Quadrupole analyzer - It consists of four cylindrical metal rods held parallel to each other. Two opposite rods carry a positive charge while the other pair of opposite rods carry a negative charge. In addition to this direct voltage a high radio frequency (RF) voltage is applied to all the four rods, resulting in an oscillating electric field. The analyte ions are separated based on the stability of their trajectories as they fly through these electrodes. Direct and RF voltages are changed so that only the ions with very narrow interval of m/z values successfully pass through the rods to the detector [68].

2. Time-Of-Flight (Tof) analyzers - They are based on the basic principle that when same amount of force is applied to different ions, the resulting acceleration on the different ions is inversely proportional to their mass. Heavier ions will have a slower acceleration and thus will take longer to reach the detector whereas the lighter ions will move faster and reach the detector in less time. The ions are thus separated based on their time of flight to reach the detector [61, 68].

3. Ion traps - They are also referred to as a 3D quadrupole and have an edge over canonical quadrupoles as they are able to perform tandem MS analysis. For tandem analysis a precursor ion is selected based on its m/z value and isolated in the trap followed by its collision with an inert gas (helium), resulting in dissociation. The ions that are generated after dissociation of the precursor are then scanned to produce MS/MS spectrum of the precursor. The most commonly used ion traps include: i) linear ion trap which is similar to the quadrupole with slight modifications that enables higher capacity to store more ions and ii) Orbitrap which is the latest advancement among mass analyzers, here the ions travel in a circular motion along a spindle shaped electrode as shown in Figure 1.6. The ions can be trapped inside the Orbitrap and it also serves as a detector generating the mass spectrum using fourier transformation [69]. The

Orbitraps also provide a very high resolution up to 280000 at 400 m/z and high mass accuracy of <5 ppm [70].

The performance of a mass spectrometer is evaluated based on two important characteristics: i) Resolution - ion separation with very small difference in their m/z values and ii) Mass accuracy - precision in determining the m/z value. However the instruments currently flourishing are hybrid instruments, which are comprised of two analyzers in order to overcome the limitations of a single mass analyzer and combine the strengths of different mass analyzers for better performance. Here, the schematic layout of one such hybrid instrument is explained, the LTQ (Linear trap quadrupole) Orbitrap mass spectrometer from Thermo Fischer Scientific (Schwerte, DE), a prototype for the current generation of mass spectrometers, which were routinely used in the progress of this thesis (Figure 1.6).

1.2.2.1 LTQ Orbitrap mass spectrometer

The LTQ Orbitrap XL mass spectrometer is a hybrid Fourier-Transform mass spectrometer (FTMS) which combines a linear ion trap (LTQ) and an Orbitrap mass analyzer (Figure 1.6). Ions generated from the ion source are collected in the LTQ followed by ejection into the C-shaped storage trap which is used to store the ions before injection into the Orbitrap. In the Orbitrap, a very high resolution precursor ion scan is performed to generate the MS1 spectrum and at the same time, the ions are separated in the ion trap and fragmented by low-energy collision induced dissociation (CID) to record the product ion scan i.e., the MS2 spectrum. The LTQ ion traps have very high sequencing speed as compared to the Orbitrap, therefore various product ion scans (MS2) can be performed in the LTQ while the Orbitrap is performing precursor ion scan (MS1). The Orbitrap can also perform CID with very high mass accuracy and resolution but owing to its low acquisition speed the MS2 scans are performed in LTQ. This is the advantage of combining the strengths of two mass analyzers in one hybrid instrument [71].

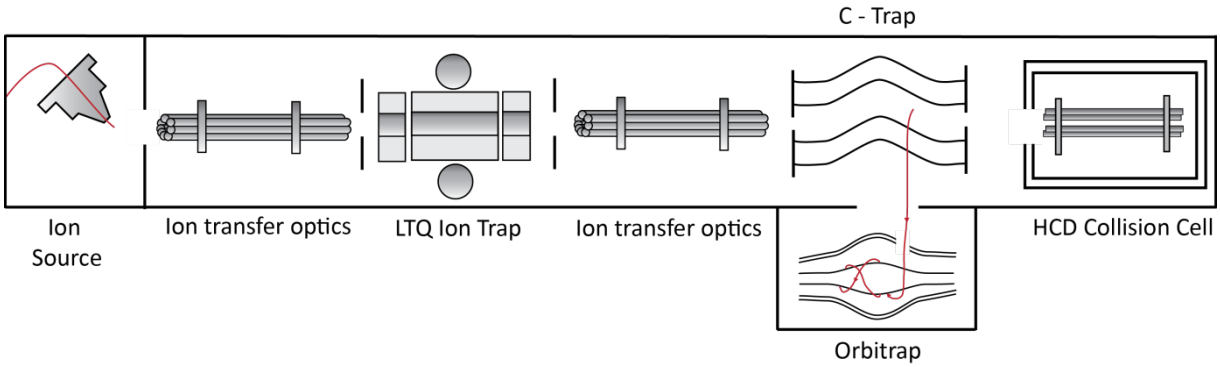


Figure 1.6 Schematic layout of an LTQ-Orbitrap mass spectrometer.

A hybrid mass spectrometer comprising an ion source through which the ions are sprayed into the mass spectrometer, ion transfer optics (multipole) for enabling the movement of ions through different parts of the mass spectrometer. It also comprises a C-shaped trap for ion storage and an HCD collision cell for fragmentation, although fragmentation can also occur in the ion trap.

Additionally, this hybrid instrument is also equipped with an HCD collision cell, to perform high-energy collision dissociation (HCD) for the fragmentation of selected precursors. For HCD fragmentation, the ions are collected in the LTQ, and the selected ions depending upon the MS1 are then passed into the HCD cell where they are fragmented under normalized collision energy. The product ions are then transferred into C-trap which further transfers them into the Orbitrap. In the Orbitrap the fragment spectrum or the MS2 scan is recorded [71]. The HCD collision cell provides the ability to perform a broad range of fragmentation experiments from advanced level of proteomics to even small molecule research but due to its slow speed it is used only for certain samples or questions.

1.2.3 MS based fragmentation of peptides and RNA

Mass spectrometry has been routinely used to gain deeper insight into the complexity of biological samples. Tandem MS plays a very important role in studying different biomolecules such as proteins and nucleic acids. Both CID and HCD based fragmentation can be used for the investigation of proteins and DNA/RNA in the samples. As most of the biological samples used during the course of this thesis comprised of proteins and RNA, here I discuss the basic principle of fragmentation of peptides and RNA in the MS experiments.

For proteomics studies, the peptide fragmentation is carried out from acidic solutions in positive ion mode which can give rise to different ion species (Figure 1.7 A). The fragment ions that are generated are named according to Roepstorff-Fohlmann-Biemann nomenclature [72, 73]. Under the low energy dissociation (CID) the most common fragment ions are generated

upon the cleavage of a peptide bond between two amino acids. Depending on whether the charge is retained at the amino-terminal or carboxy-terminal fragment of the peptide they are referred to as b- ions or y- ions, respectively. Another common observation is the pair of a- and b- ions, separated by a mass of 27.9949 Da (corresponding to loss of a C=O group). In the MS2 scan from the quadrupole instruments the y- ions predominate whereas in the ion trap instruments both b- and y- ions are observed [64].

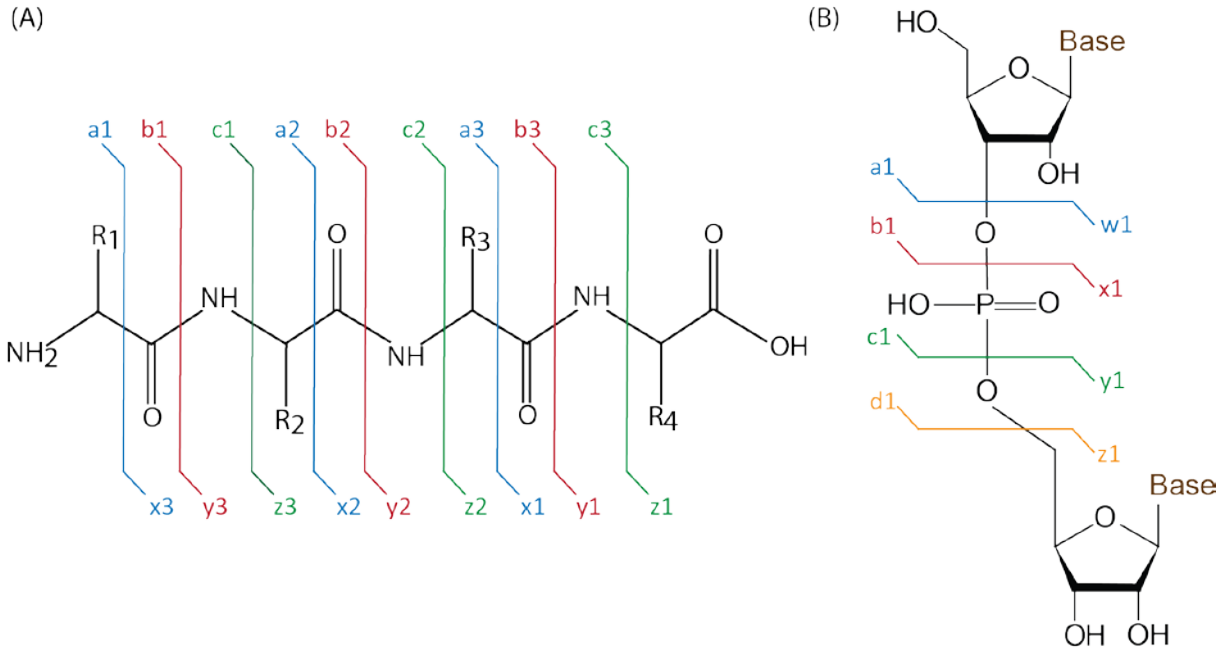


Figure 1.7 Schematic representations of peptide and RNA fragmentation and their nomenclature.

(A) Nomenclature of different fragment ions generated upon fragmentation of peptide in low collision energy. The ions from carboxyl- terminal are named as x-, y- and z- ions whereas the amino- terminal ions are called a-, b- and c- ions. The cleavage of peptide bonds results in two sets of ion species: b- ions, if the charge resided N- terminal to the cleavage site and y- ions if the charge is C- terminal. (B) Nomenclature of fragment ions generated upon fragmentation of RNA from the cleavage of phosphate backbone. The fragment ions containing the 5'end are named as a-, b-, c- and d- ions and the ones containing 3' end are called w-, x-, y- and z- ions.

Although MS has gained immense popularity in proteomics field, it is also used as a viable tool for structural studies of nucleic acids. With the recent developments in ionization techniques, it has become easier to generate gaseous ions from heavy biomolecules. The fragmentation of nucleic acids is carried out from basic solutions in negative ion mode. The fragment ions generated from RNA are named according to the nomenclature proposed by McLuckey *et al* [74]. The nomenclature is analogous to the one widely used for peptides. There are four possible sites for cleavage along the phosphodiester chain (Figure 1.7 B) and they are named based on the retention of charge at 5' or 3' end. Under CID conditions it has also been noticed

that there is high tendency of the cleavage of N-glycosyl bond between different base and the sugar moiety, leading to the release of nucleobase as a separate ion or as a neutral loss [75].

1.2.4 MS based proteomics and data analysis

The term 'proteome' refers to the entire collection of proteins expressed by a genome, cell, tissue or a whole organism at a given point of time under different conditions and the study of proteome is referred to as 'Proteomics'. It deals with study of different aspects of molecular and cellular biology at the protein level. Proteomics studies can be very challenging owing to the complexity of the protein populations extracted from cells and tissues and sometimes the protein of interest might be very low abundant in such a complex mixture. Therefore a sensitive and advanced analytical approach such as MS can be useful to deal with complex protein samples. In the past MS has become a method of choice for identification of proteins, post-translational modifications and protein-protein interactions when applied to smaller protein datasets [76]. However with increasing developments in new experimental approaches, the MS-based proteomics is now also used for analysis of very large protein systems such as the analysis of entire human proteome [77, 78].

For typical proteomics experiments, the proteins are extracted and isolated from cells or tissues by different fractionation and affinity purification strategies. The isolated proteins are then separated using one dimensional gel electrophoresis (1D-PAGE) [79]. For higher sensitivity of the MS analysis the proteins are digested into peptides enzymatically using endoproteinases such as trypsin and the protein identification is carried out by peptide sequencing in the MS analysis (Figure 1.8). Upon digestion of proteins multiple peptides that are generated might add to the sample complexity, therefore depending on the experimental requirements or scientific question being addressed the sample complexity can be further reduced by carrying out a separation at the peptide level. The complex peptide mixtures can be separated based on peptide characteristics such as isoelectric point using peptide isoelectric focusing (pIEF) [80].

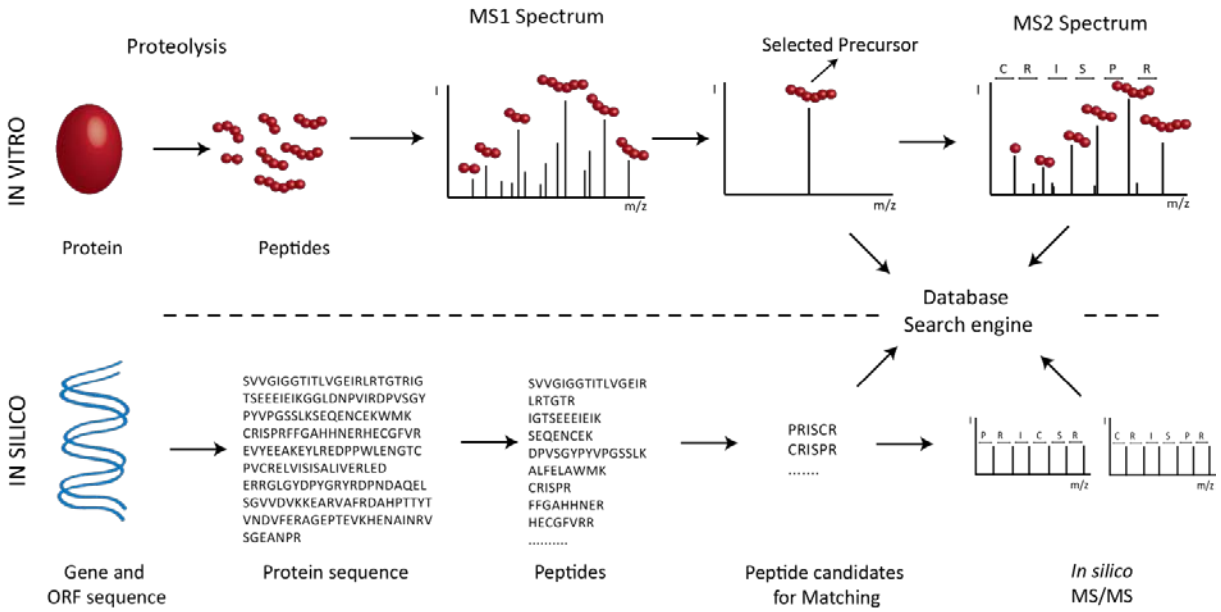


Figure 1.8 Schematic representation of the workflow for proteomics data analysis.

For the protein identification by MS, the protein is hydrolyzed with endoproteinases under *in vitro* conditions and the peptides generated are scanned in the mass spectrometer for their m/z values. After the MS1 analysis the precursor ions are selected based upon their m/z intensity, for the MS/MS fragmentation. From the MS2 spectrum the amino acid sequence of the peptide can be determined. For the identification of the protein, the experimentally determined mass of the precursor and the fragments are compared with the theoretically determined masses obtained from the database search performed *in silico*.

Nonetheless before the MS analysis the peptides are separated by a step of high-pressure liquid chromatography (HPLC) and eluted into an electrospray. This is followed by the MS analysis where the MS1 scan is acquired and a selected precursor is fragmented to generate the MS2 scan as shown in Figure 1.8. This mode of data collection in tandem MS analysis where a fixed number of precursor ions whose m/s values are recorded in a survey scan are selected using predetermined rules and subjected to a second stage of fragmentation in MS2 or MS/MS analysis is also referred to as data dependent acquisition (DDA) [81]. The MS and MS/MS data is used for matching against protein sequence databases to identify the peptides and therefore the proteins. The entire workflow of using liquid chromatography separation, followed by electron spray ionization and mass spectrometry is also called LC-ESI-MS/MS or in short LC-MS/MS.

1.2.5 Quantitative proteomics

One of the most challenging aspects of proteomics is to quantify the differences between different physiological states in a biological system. The MS based quantitation have gained immense popularity over the past decade making use of differential stable isotope labeling to create specific isotopic mass tags that can provide a basis for quantification. The isotopic tags can be introduced at the level of proteins or peptides as shown in Figure 1.9. Broadly, the quantitative MS approaches are classified into two categories i) Relative quantification - the comparison between amount of proteins or entire proteomes between two or more samples in order to yield a quantitative ratio and ii) Absolute quantification - determining the absolute amount of concentrations of proteins within a sample. Further these two categories can be divided on the basis of using stable isotope labeling or label free approach for quantitation, as summarized in Table 1.2.

Table 1.2 An overview of different label-based and label free approaches used for absolute and relative quantification highlighting important examples in each category. Adapted from [82]

Relative quantification				**Absolute quantification**	
Label-Based			Label-free	Label-based	Label-free
Metabolic	Chemical	Enzymatic	Ion intensities (XIC)	AQUA peptides	iBAQ
^{15}N	ITRAQ/TMT	O^{18}	Spectral counting		
SILAC	DML				

SILAC - Stable isotope labeling by amino acids in cell culture, ITRAQ - Isobaric tags for relative and absolute quantification, TMT - Tandem mass tags, DML - Dimethyl labeling, XIC - Extracted ion chromatogram, AQUA - Absolute quantification and iBAQ - Intensity based absolute quantification.

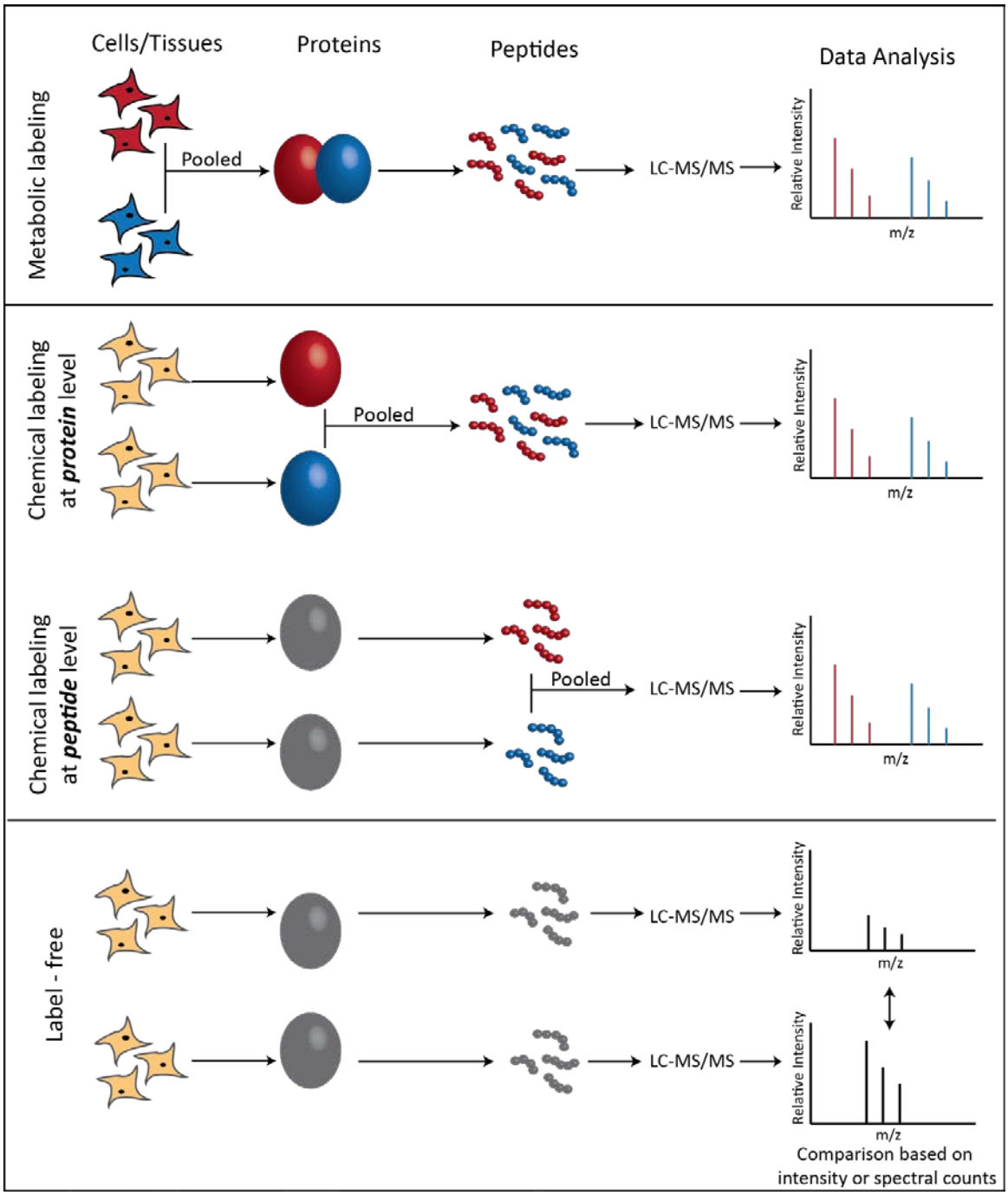


Figure 1.9 Common workflows used for quantitative proteomics.

Most commonly used approaches for protein quantification include: i) Metabolic labeling, ii) Chemical labeling and iii) Label-free approach. The labeling is achieved with stable isotope incorporation which can be at the metabolic level as the cells grow or using chemical reagents. The later can be performed at the protein or the peptide level using isotopically encoded reagents. After labeling the proteins or the peptides, the respective samples are pooled, followed by MS analysis. The relative level of expression is obtained on comparing the signal of labeled and unlabeled peptides. In the label-free approach the MS data from two samples is compared with respect to the intensity of peptide precursor ions or spectral counts of a particular peptide, for a given protein. Colors red and blue indicate the light and heavy labels respectively, at the level of cells, proteins or peptides [83].

For the investigation of prokaryotic immune defense system two quantitative proteomics approaches were used as described below.

1.2.5.1 Differential isotope labeling using dimethyl labeling of peptides.

Chemical labeling of peptides using differential isotope labeling has been widely used in proteomics research. Dimethyl labeling is a very fast and straightforward approach using inexpensive chemical reagents that provide almost 100% labeling efficiency and multiplex quantification [84].

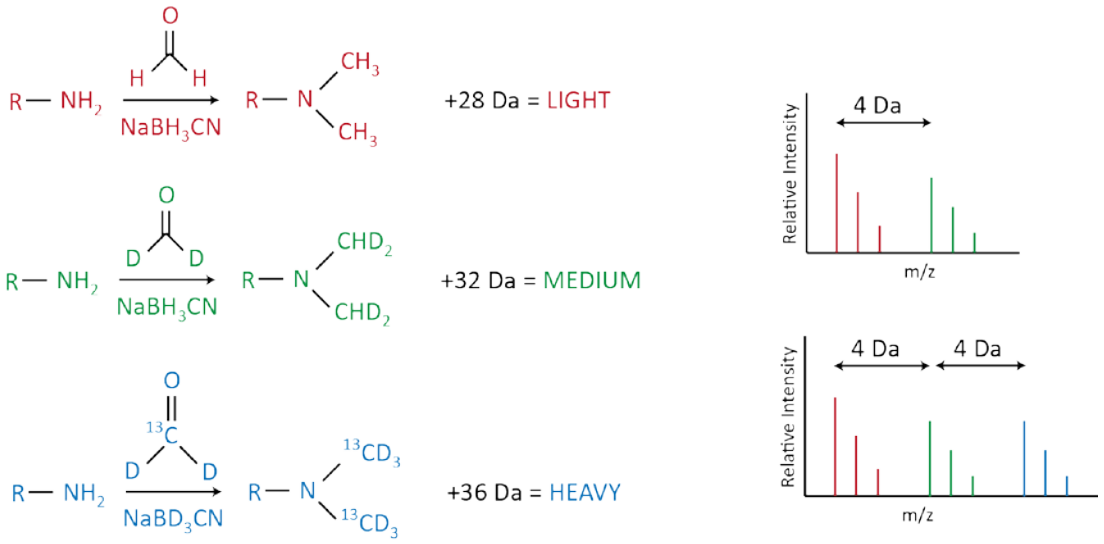


Figure 1.10 Labeling scheme for triplex stable isotope dimethyl labeling.

Three different samples can be analyzed in parallel by labeling the peptides using three different isotopomers of formaldehyde. The labels result in light, intermediate or heavy labeled peptides with an additional mass of 28, 32 or 36 Da. These labels provide a minimum of 4 Da mass differences between the peaks in the generated peptide triplets. Adapted from [84] with permission from the publisher.

In dimethyl labeling, all the primary amines in a peptide mixture are converted into dimethylamines, which includes the N-terminus and the side-chain of lysine residues. A combination of isotopomers of formaldehyde and cyanoborohydride can be used to generate peptide triplets that differ in mass by a minimum of 4 Da between different samples (Figure 1.10). The peptide mixture from a particular sample is labeled with light, medium or heavy labels and the labeled peptides from different samples are pooled and simultaneously analyzed by MS. The peptide abundance from different samples is then compared depending upon the mass difference of the dimethyl labels [85]. Moreover, dimethyl labeling can also be applied at the level of intact proteins, but this would limit the choice of proteases as trypsin and Lys-C would not be able to cleave modified lysine residues [86].

1.2.5.2 Label-free approach (iBAQ)

Quantitative proteomics also aim at determining the absolute amount of proteins in a sample. Intensity based absolute quantification (iBAQ) is a label-free approach in quantitative MS [87]. It estimates the absolute amount of a particular protein by summing the peak intensities of all detected peptides of the protein dividing it by the number of theoretically observable peptides [88]. iBAQ intensities are the most accurate measurement of the absolute abundance of all the proteins identified in a sample. In addition to determining the protein abundance, iBAQ has also been used to determine the protein stoichiometry in a multi-protein complex [89]. A reference protein mixture such as universal protein standard (UPS) is spiked into the sample and the iBAQ intensities of reference proteins are plotted against their known amounts to prepare a regression curve. The amount of different proteins in the sample is then determined from this regression curve using their experimental iBAQ intensities. Once the absolute amount of different proteins in a complex is determined, the stoichiometry of proteins in the complex can be calculated.

The quantitative proteomics approaches such as dimethyl labeling and iBAQ have gained immense popularity due to the availability of computational platforms such as MaxQuant software [90] that have a provision for processing the raw data, performing database searches, quantification of peptides and proteins and statistical evaluation of the data.

1.2.6 Structural Proteomics

Different MS-based approaches have been used to study the structure and dynamics of macromolecular assemblies that comprise physically interacting proteins with/without nucleic acids. Determination of structural organization of these complexes has always contributed to the understanding of various biological functions. A variety of techniques such as NMR, X-ray crystallography and cryoEM have been widely used to determine the structure of protein complexes [91]. A major challenge for the structural biologists is to study the three dimensional structural organization of these complexes due to conformational dynamics, heterogeneous composition, asymmetric structure and the large complex size. Furthermore, for a complete understanding of the biological role and the mode of action of such macromolecular assemblies it has become important to have high resolution structural information about the identity, shape and structure of individual components, stoichiometry of different components and interactions between different components present in the complex [92].

A majority of structural MS investigations are based on the principle that the non-covalent interactions can be maintained in the gas phase [93, 94]. Native MS approach has made it possible to analyze entire protein complexes in intact form in the mass spectrometer, for e.g., large MDa complexes such as ribosomes [95]. In addition, protein-protein cross-linking approach based on chemical cross-linkers that covalently connect the functional groups on proteins or protein complexes to create structurally defined interactions between proteins is also becoming a method of choice to study protein-protein interactions (Section 1.2.6.2). The advantage of using MS for structural studies, compared to other methods, lies in the requirement of very low sample amounts and the fast analysis speed enabling real time monitoring of molecular interactions [92]. To study the protein-RNA and protein-protein interactions in multi-subunit ribonucleoprotein complexes such as the crRNPs (CRISPR ribonucleoprotein complexes), we used two structural proteomics approaches as described below.

1.2.6.1 UV induced protein-RNA cross-linking

Ribonucleoprotein (RNP) complexes play a key role in mediating biological processes such as gene expression and regulation. A vast array of RNA binding proteins (RBPs) have been reported in eukaryotes that stabilize the RNA structure and also mediate its interactions with other biomolecules when they are part of a macromolecular assembly [96]. The RBPs can bind single or double stranded RNA through their conserved structural motifs known as the RNA binding domains (RBDs). The computational analysis has led to identification of such structural motifs in these RBPs, such as RNA-recognition motifs (RRMs) [97], K homology (KH) domains [98], zinc-finger domains [99], G-patch domains [100], Sm motifs [101], etc. The three-dimensional structures of recently crystallized prokaryotic RBPs show presence of RBDs similar to eukaryotes, e.g., the Sm-fold in bacterial HfQ proteins [102] and the RRMs in Cas7 protein family [55, 103]. However, there is very little information available on interaction between RBDs and their cognate RNA. In order to understand the molecular details of these processes it becomes important to characterize the interactions between proteins and RNA.

Structural studies using co-crystallization, NMR and high resolution EM are the gold standards for characterization of molecular interactions between RBDs and the cognate RNA molecules, as shown in the past [46, 104-106] and from the co-structures of various protein-RNA complexes available in the PDB. In recent years UV induced protein-RNA cross-linking coupled

with MS has emerged as a more specialized and direct approach for obtaining information about the protein-RNA interactions in RNPs [107].

UV induced protein-RNA cross-linking allows the identification of cross-linked peptides and RNA moieties and the exact contact sites within the RNA and protein at single nucleotide and single amino acid resolution [108]. This approach can be applied to single proteins such as the recombinant proteins that interact with RNA and to the complex assemblies of RNPs that have been reconstituted or purified from the cells (endogenous).

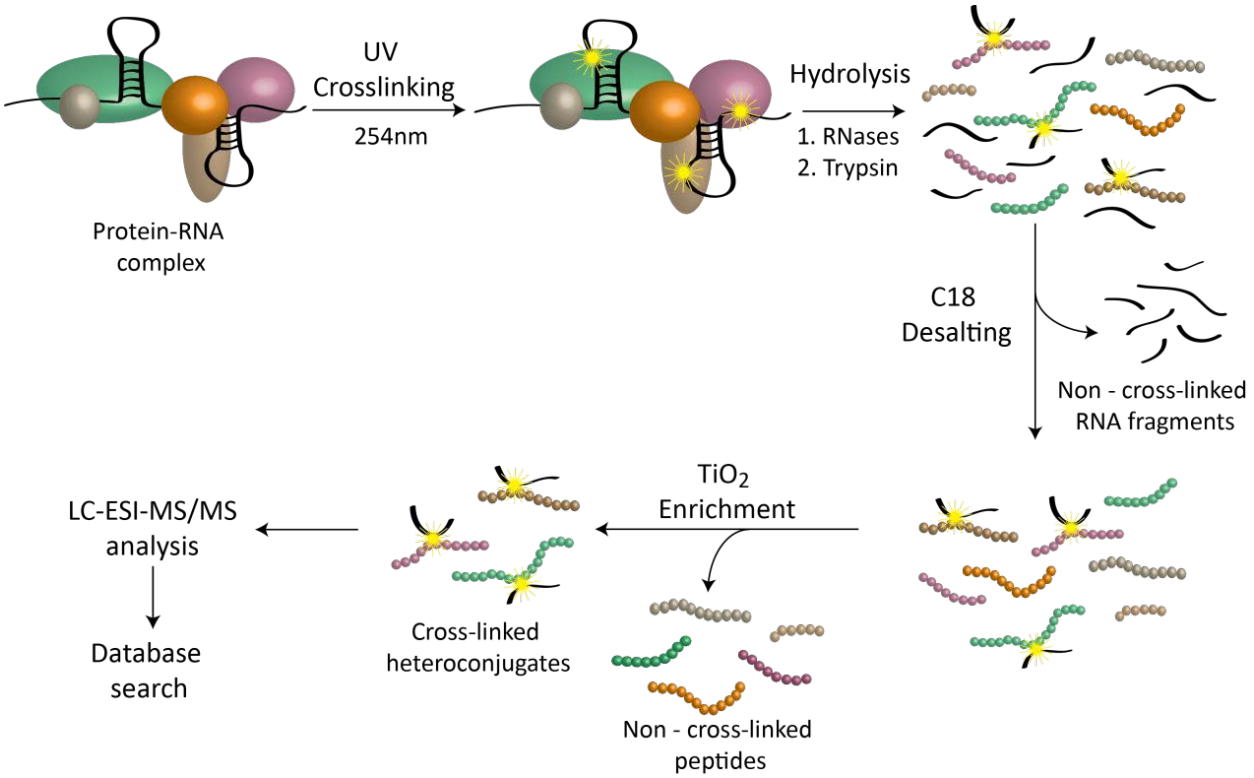


Figure 1.11 Overview of the protein-RNA cross-linking workflow.

Protein-RNA complex is UV irradiated at 254 nm and hydrolyzed by RNases and trypsin resulting in a mixture of non-cross-linked peptides and RNA fragments and cross-linked heteroconjugates. The non-cross-linked RNA fragments are removed with C18 desalting and the cross-linked heteroconjugates are enriched via TiO_2 columns that remove non-cross-linked peptides. These enriched cross-linked samples are then analyzed by LC-ESI-MS/MS followed by database search. Adapted from [109] with permission.

The principle underlying this approach is that upon UV induction the amino acid side chains of proteins cross-link to the nucleobases of RNA in close spatial proximity [109]. The cross-linked region and cross-linked amino acid and nucleotide can then be identified with high resolution MS. MS fragmentation allows sequence determination of the cross-linked peptide and the

composition of cross-linked RNA. An overview of the workflow for UV induced protein-RNA cross-linking is shown in Figure 1.11.

After UV cross-linking, the cross-linked peptide and RNA are required to be isolated for the subsequent LC-MS/MS analysis. Both the RNA and protein moieties are completely digested with endonucleases and endoproteinases under denaturing conditions. Most commonly used endonucleases include combination of RNase A and T1 for single stranded RNA or benzonase which digests both single and double stranded RNA/DNA in an unspecific manner generating short fragments of mostly single nucleotides. When the RNA moiety is short, the MS analysis to determine the sequence of cross-linked peptide becomes more sensitive [110].

For the proteolysis, trypsin is the most commonly and widely used endoproteinase in MS-based proteomics. The UV cross-linking yield and the efficiency of RNA and protein hydrolysis determines the yield of cross-linked peptide-RNA heteroconjugates.

The yield of UV induced cross-linking between proteins and RNA is relatively low [109], therefore it is essential to enrich the cross-linked species for subsequent MS analysis. The mixture obtained after digestion of protein-RNA comprises of mainly non-cross-linked peptides, non-cross-linked RNA oligonucleotides and cross-linked peptide-RNA heteroconjugates. For the removal of non-cross-linked oligonucleotides and the non-cross-linked peptides two successive purification steps are performed (Figure 1.11). The non-cross-linked RNA oligonucleotides are removed by C18 reversed-phase chromatography because they do not bind to the C18 material whereas both the cross-linked and non-cross-linked peptides have a strong affinity towards the C18 material. This step is also referred to as C18-desalting as it allows salts and other contaminants to be washed off from the sample [111]. After removal of the non-cross-linked RNA oligonucleotides, the sample mainly consists of cross-linked peptide-RNA heteroconjugates, non-cross-linked peptides, and residual non-cross-linked RNA oligonucleotides. To remove non-cross-linked peptides and enrich peptide-RNA heteroconjugates, titanium dioxide (TiO_2) enrichment is used. It has been established as a method for enrichment of phosphopeptides in MS-based proteomics experiments [112, 113].

If the proteins and RNA differ considerably in their size then the proteins are hydrolyzed prior to RNA hydrolysis. After proteolysis the intact RNA with or without cross-linked peptides is enriched using size exclusion chromatography. The approach has been used earlier in studying the protein-RNA interface of different RNP complexes [108, 114].

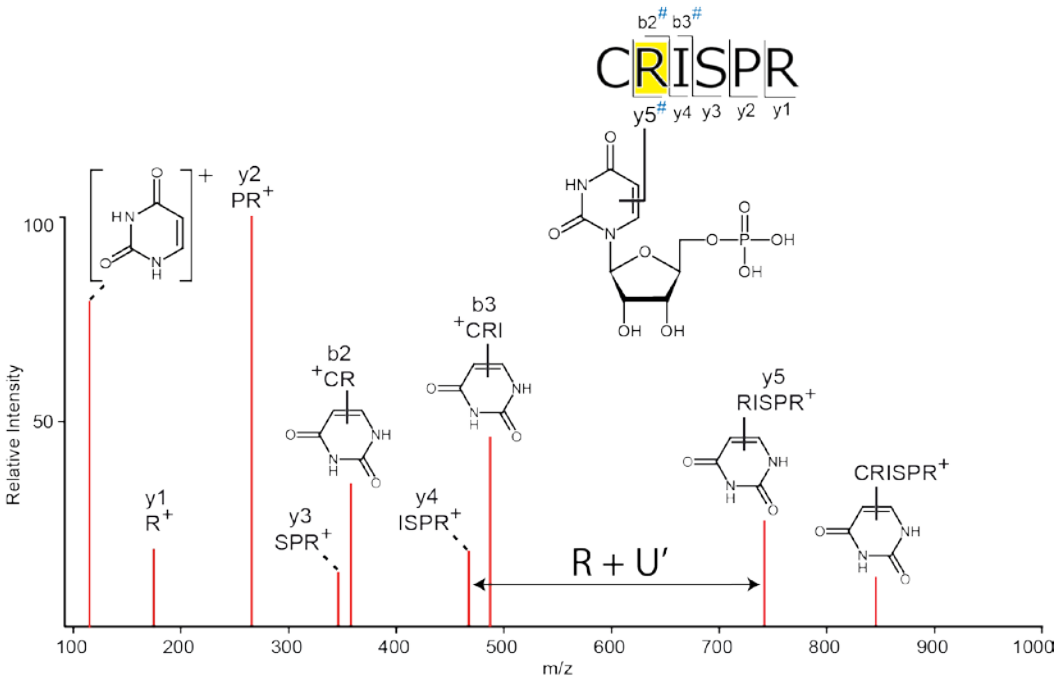


Figure 1.12 An artificial MS/MS spectrum of a peptide-RNA cross-link.

Sample spectrum of a peptide 'CRISPR' cross-linked to a uracil nucleotide to indicate the characteristic peaks and shifts observed during fragmentation of a peptide-RNA cross-link. The peptide sequence and fragment ions are indicated on the top and cross-linked residue is highlighted in yellow. Distinct fragment ions containing nucleic acid base and peptide fragments are indicated in red. Some of the b- and y- ions are shown with a mass shift of # corresponding to U': U marker ion adduct of 112.0273 Da.

The MS analysis of the purified peptide–RNA cross-links allows sequencing of the cross-linked peptide and RNA moieties in the gas phase. ESI coupled to a nano-LC system is a method of choice for analyzing such cross-links. The aim is to determine not only the amino-acid sequence of the cross-linked oligopeptide but also to identify the cross-linked amino acid. The mass of entire cross-linked species is a simple additive of the mass of cross-linked nucleotide and the mass of cross-linked peptide [109]. Thus to determine the cross-linked nucleotide, mass difference between the cross-linked species (experimental precursor) and the cross-linked peptide is calculated. An artificial MS/MS spectrum from a typical peptide-RNA cross-link fragmented under HCD conditions is shown in Figure 1.12. The signals from the fragment ions of the cross-linked peptide pre-dominate the MS/MS spectrum. The fragment ions from the peptide that contain the cross-linked residue are shifted by the mass of this nucleotide residue, when compared to the regular peptides fragments. Additionally, marker ions of the cross-linked nucleotides are also observed in the lower m/z regime of the fragment spectrum corresponding to the bases (e.g., U' = 113.0351, Uracil base).

For protein-RNA cross-linking studies high resolution MS instruments are required that provide a high mass accuracy in determining the precise mass of precursors (the cross-linked species) and the product ions (the fragment ions from sequencing of cross-linked peptide and oligonucleotide). Orbitrap instruments that carry out fragmentation in HCD mode such as the LTQ-Orbitrap Velos, Q Exactive and Q Exactive High Field instruments, were used for MS analysis of the protein-RNA cross-linking experiments performed during the course of this thesis. The data analysis was carried out using the RNP^{xl} tool [108] implemented in OpenMS [115, 116] using OMSSA [117] as the search engine. The detailed description for the data analysis workflow has been described in [108].

Nonetheless, the interpretation of MS data from protein-RNA cross-linking experiments is challenging as every MS/MS fragment spectrum for an identified cross-link is manually validated. Up to now there has been no suitable software that can handle different features observed in fragment spectra of cross-linked heteroconjugates (as depicted in Figure 1.12). However the efforts are being made for automated identification of the cross-links in collaboration with the group of Prof. Oliver Kohlbacher, University of Tübingen, Tübingen.

1.2.6.2 Protein-protein cross-linking using chemical cross-linker

Interactions between different proteins can occur when these proteins co-exist in organized structural complexes or during transient encounters for mediating various biological processes. In order to derive more information on the spatial proximity and the arrangement of proteins in an assembled complex, chemical cross-linkers such as BS3 (Bis(sulfosuccidimidyl)suberate) are used which introduce specific and stable chemical linkages (covalent interactions) in the otherwise transiently associated proteins.

BS3 is a widely used homobifunctional cross-linker. It contains two identical functional groups i.e., the N-hydroxysulfosuccinimide esters (NHS esters) at both reactive sites connected with a carbon chain spacer that bridges the defined distance of 11.4 Å, allowing the identical groups of the proteins to be cross-linked. It is water soluble due to the terminal sulfonyl substituents, thus excluding the need to use organic solvents that might interfere with protein structure. It is amine reactive i.e., the NHS esters at the two ends react specifically with primary amines. It targets the primary amines in the side chains of lysine residues on protein's surfaces for targeted protein-protein cross-linking [118] and protein N-termini. It is highly reactive with a half-life of 20-30 min at pH 7.5. Therefore the BS3 solution for experimental purposes is always

prepared fresh and used immediately. The cross-linking reactions are carried out at room temperature and quenched with buffers containing primary amine reagents such as ammonium bicarbonate or Tris-HCl to consume the non-reacted cross-linker.

After cross-linking the complex is hydrolyzed with endoproteinases such as trypsin, to generate different types of cross-linked and non-cross-linked peptides as shown in Figure 1.13. The cross-linked peptides are a very small proportion of the overall mixture of different peptide species generated after tryptic digestion [119, 120]. For enrichment, size exclusion chromatography is used to separate the cross-linked peptides from linear and non-cross-linked peptides, considering the cross-linked peptide will have a higher mass than linear peptides [121]. This approach is usually applied for larger complexes. The cross-linking workflow applied for investigating the protein-protein interactions in Type I-B Cascade complex from *C. thermocellum* is described in Figure 1.13.

Alternatively, for smaller complexes (<400 kDa) the cross-linked complex can be separated from non-cross-linked proteins using SDS-PAGE. Following which the bands of higher molecular weights corresponding to cross-linked complex can be digested [119].

The cross-linked peptides have higher charge states and higher masses as compared to the linear peptides. The analysis of peptide-RNA cross-links by MS requires high resolution at MS1 level for correct charge determination. The fragmentation can be performed using either CID or HCD. Hybrid Orbitrap instruments are often used for the MS analysis of cross-linked peptides providing a linear ion trap for CID fragmentation, identifying more number of cross-links and determination of fragment masses in the Orbitrap for higher accuracy. The Orbitrap Fusion Tribid mass spectrometer (Thermo Fischer Scientific, Schwerte, DE) was used for the analysis of protein-protein cross-linking experiments in the course of this thesis.

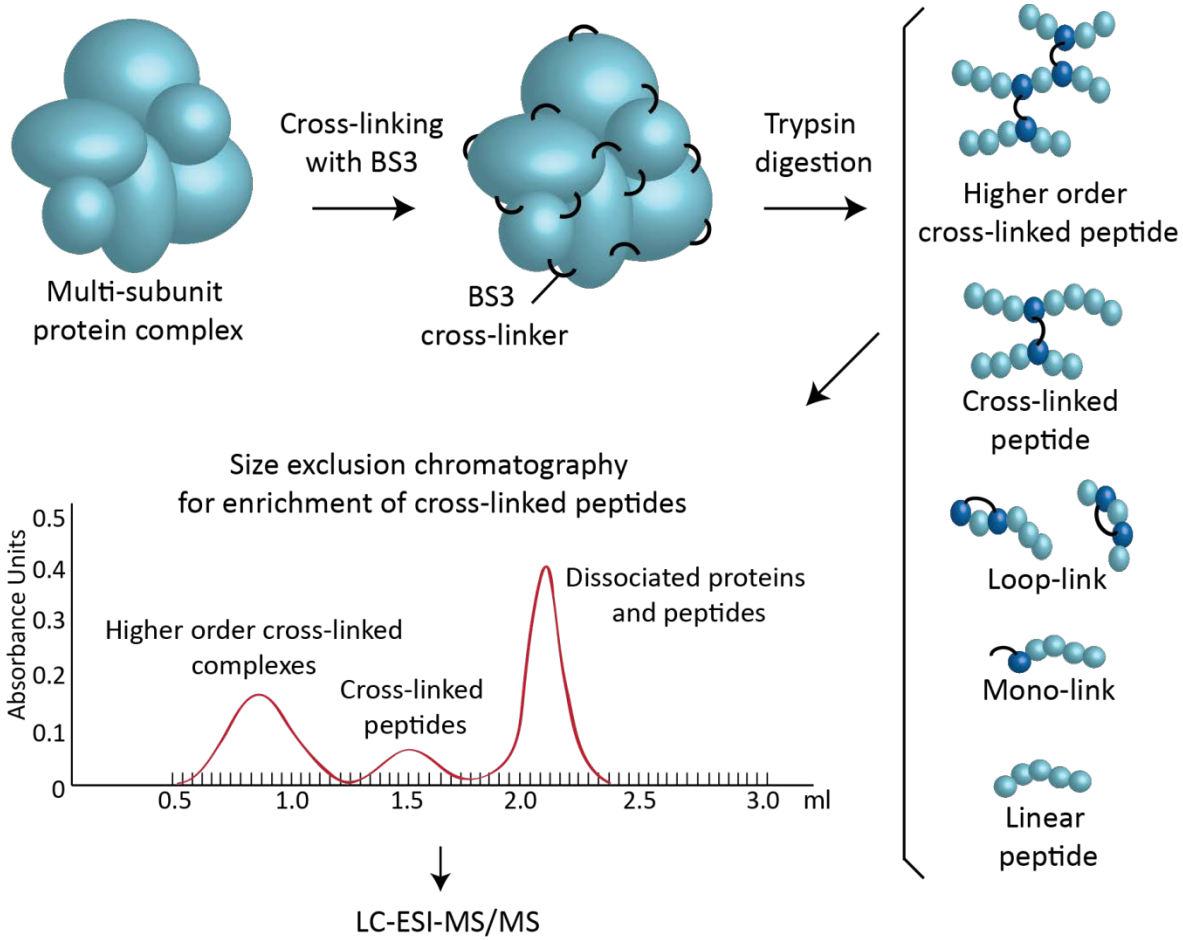


Figure 1.13 Overview of the protein-protein cross-linking workflow.

Multi-subunit protein complex are cross-linked with chemical cross-linking reagents such as BS3. After cross-linking the cross-linked complex is digested with endoproteinases such as trypsin, generating different kinds of cross-linked and non-cross-linked peptide species. The cross-linked peptides are enriched using size exclusion chromatography followed by MS analysis. Adapted from [122] with permission.

Additionally, the increased sample complexity due to protein-protein cross-linking results in a very large search space that requires dedicated algorithms for the analysis of data derived from the protein-protein cross-linking experiments. Standard database search engines might not be able to handle such advanced algorithms. Data analysis for the identification of cross-links in the protein-protein cross-linking experiments reported in this thesis was performed using pLink which is accessible through pfind.ict.ac.cn/software/pLink [123].

1.3 CRISPR-Cas systems studied with mass spectrometry

In this section, the CRISPR-Cas proteins, crRNAs and the crRNP complexes that were investigated during the progress of this thesis are briefly described. Quantitative proteomics studies were carried out to determine the stoichiometry of Cas protein subunits in different Cascade complexes and to perform comparative proteome analysis between a wild type and a mutant strain of *Haloferax volcanii*. Structural proteomics studies were carried out to gain insights into the protein – RNA interactions in (i) recombinant Cas proteins and their cognate crRNAs and (ii) multi-subunit crRNP complexes, and also protein – protein interactions in multi-subunit crRNP complexes. An overview of the objectives behind using these particular complexes is also discussed here; however the MS approaches are described under section 1.2

1.3.1 Type I-B CRISPR-Cas system

The Type I-B system is most commonly found in archaea such as methanogens and halophiles and some bacteria such as *Clostridia*. The characteristic features of this subtype are very similar to other Type I subtypes with respect to three stages of the mechanism of CRISPR-Cas action, however the composition of Type I-B complex has not been described yet. Here we describe some characteristics of Type I-B system in two different organisms, *Haloferax volcanii* and *Clostridium thermocellum*. Both organisms contain Cas proteins Cas5, Cas6, Cas7 and Cas8b.

1.3.1.1 Haloferax volcanii Type I-B system

H. volcanii is a halophilic archaeon that requires high salt concentration for growth. The CRISPR-Cas system comprises of three different CRISPR loci as shown in Figure 1.14. Two loci are located on the chromosomal plasmid pHV4 (locus P1 and P2) and one on the main chromosome. The *cas* genes encoding for Cas protein 1 - 8b are located between the P1 and P2 loci on the chromosomal plasmid [124, 125].

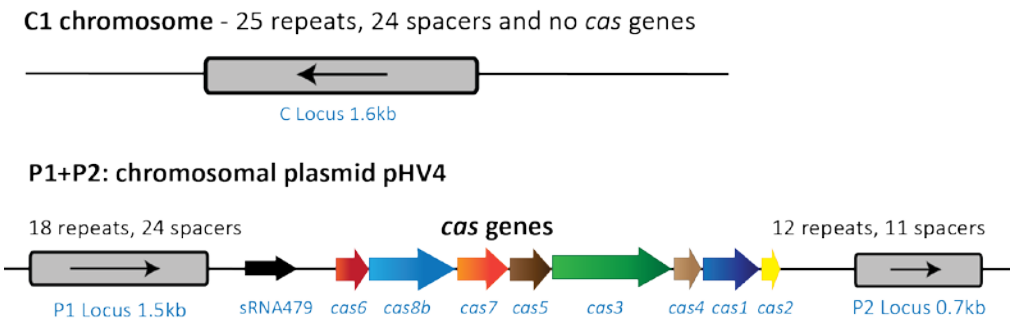


Figure 1.14 Illustration of the CRISPR loci in *H. volcanii.*

The CRISPR locus located on the main chromosome comprises of 25 repeats and 24 spacers. No *cas* genes are encoded next to this locus. In addition there are two CRISPR loci are located on the chromosomal plasmid pHV4.The *cas* gene cluster encoding the genes *cas1* - *cas8b* is located between these two loci. Adapted from [126] with permission.

Recently it was reported that the *H. volcanii* Type I-B system consists of Cascade like complex comprising at least Cas5, Cas6 and Cas7 proteins along with a crRNA [89]. Also to confirm the association of these proteins, a FLAG-tag version of Cas7 was expressed and purified along with different interaction partners. The proteins were identified by MS and the stoichiometry of these proteins in the complex was determined by using the intensity based absolute quantification (iBAQ) approach (Section 1.2.5.2). Also one of the Cas protein genes, *cas7,* was deleted to monitor the changes in the overall proteome upon deletion or mutation of a Cas protein. To achieve this, the wild type (WT) and Cas7 knock-out (Cas7KO) strains were compared at the proteome level using dimethyl labeling followed by MS analysis (Section 1.2.5.1).

1.3.1.2 Clostridium thermocellum type I-B Cascade complex

The Type I-B complex found in *C. thermocellum* consists of the same set of Cas proteins like the archaeon *H. volcanii*. The Cascade complex comprising of proteins Cas6, Cas8b, Cas5 and Cas7 assembled around the crRNA was used to determine the stoichiometry of these Cas proteins in the complex using iBAQ approach. Also to go one step further in terms of identifying the interaction sites between different Cas proteins in the Cascade complex, we used the protein-protein cross-linking approach using a chemical cross-linker BS3 as described before under section 1.2.6.2.

1.3.1.3 Cas6 endonucleases from Clostridium thermocellum and Methanococcus maripaludis

Cas6 enzymes are one of the most highly diverged families of Cas proteins. Two Cas6 endonucleases were identified in subtype I-B, in the archaeal and bacterial model organisms, *M. maripaludis* and *C. thermocellum*. These two Cas6 variants were also investigated for the crRNA processing *in vitro* [127].

It has been reported that the repeat sequences and Cas6b enzymes of *M. maripaludis* (archaea) and *C. thermocellum* (bacteria) are very similar [127], indicating a horizontal gene transfer event for these CRISPR-Cas systems. Furthermore to compare the two Cas6 enzymes in terms of RNA interaction sites, the Cas6b proteins were incubated with deoxy variants of their respective cognate crRNAs (deoxy variants, where first unprocessed nucleotide in the crRNA is replaced with a deoxynucleotide) to form a crRNP complex. These complexes were then analyzed for protein-RNA interactions using UV induced cross-linking and MS (Section 1.2.6.1), highlighting the similarities between an archeal and bacterial CRISPR/Cas subtype.

1.3.2 Type I-E Cascade complex in *Escherichia coli*

Recently published crystal structures of 3.03 Å [106] and 3.24 Å resolution [105] and the Cryo-EM structure of 8-9 Å resolution [43] of the fully assembled, Type I-E Cascade complex of *E. coli* has been a major breakthrough in understanding the crRNA guide-surveillance complex. In one of the most extensively studied Cascade complexes, the 61 nucleotide long crRNA assembles into a 405 kDa multi-subunit surveillance complex along with eleven Cas proteins with a stoichiometry of $Cse1_1Cse2_2Cas5e_1Cas7_6Cas6e_1$. The 3’ and 5’ end of the crRNA are at the two ends of the complex and are anchored by Cas6e and Cas5e proteins respectively. The 32 nucleotide long spacer along with six Cas7 proteins forms the helical backbone, giving the overall complex a seahorse-shaped architecture (Figure 1.15).

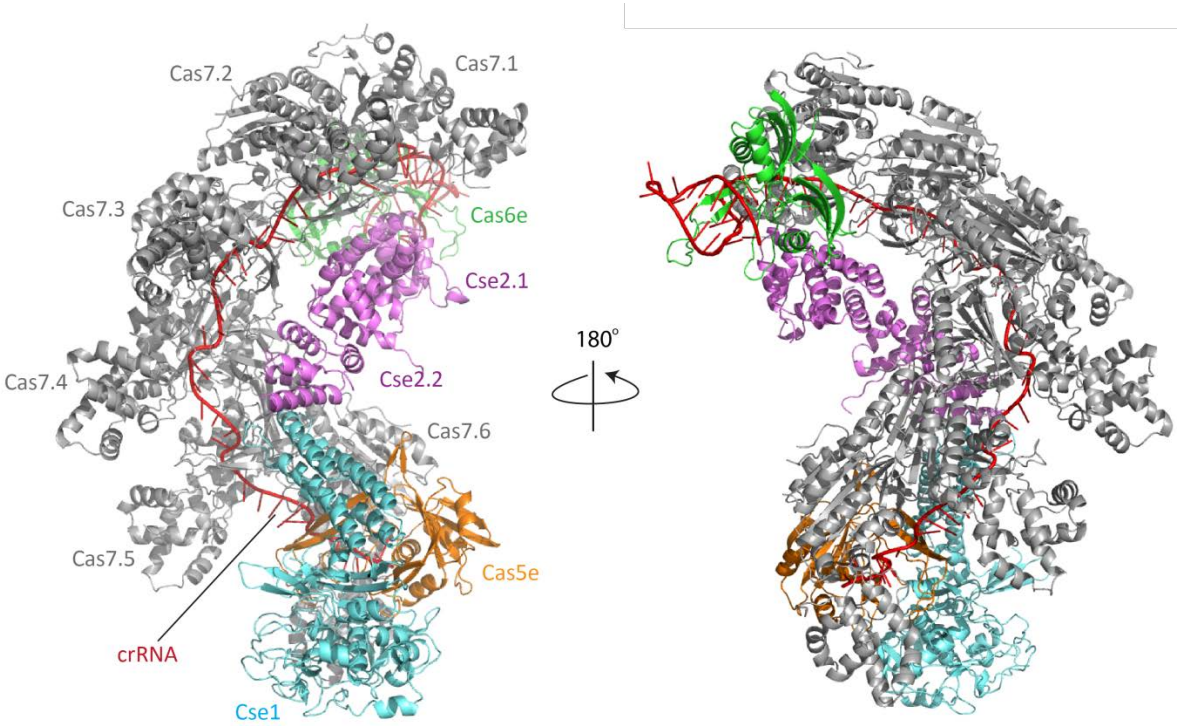


Figure 1.15 Crystal structure of *E.coli* Type I-E Cascade complex.

Two views of the fully assembled, crRNA bound complex (180° rotation, PDB 1VY8) [105]. As per the stoichiometry of the complex, the respective Cas protein subunits are indicated in the crystal structure (Cse1: Blue, two Cse2 subunits: Violet, six Cas7 subunits: Grey, Cas6e: Green, Cas5e: Orange) with the crRNA (red) in center.

The pre-assembled complex from *E.coli* was pulled down with the help of a StrepTag on Cse2 [29]. Further the protein-RNA cross-linking experiments were carried out on the assembled complex to gain deeper insight into the structural organization of the complex, by investigating the interactions between different Cas proteins and the crRNA.

1.3.3 Type III-A Csm complex in *Thermus thermophilus*

The Type III-A Csm complex in bacterium *T. thermophilus* consists of five different protein subunits Csm1 – Csm5 and a crRNA of size 35-53 nucleotides. Two models have been proposed for the stoichiometry of Csm proteins in the assembled complex: $Csm1_1Csm2_3Csm3_6Csm4_2Csm5_1crRNA_1$ (model 1, 427.04 kDa complex) or $Csm1_1Csm2_3Csm3_2Csm4_4Csm5_2crRNA_1$ (model 2, 427.462 kDa complex). The cryoEM structure of model 1 is depicted in Figure 1.16). The cryoEM structure of the complex shows a sea worm-shaped architecture which is strikingly similar to the architecture of Type I-E Cascade complex, with the Csm3 protein forming a helical Cas7-like backbone [59].

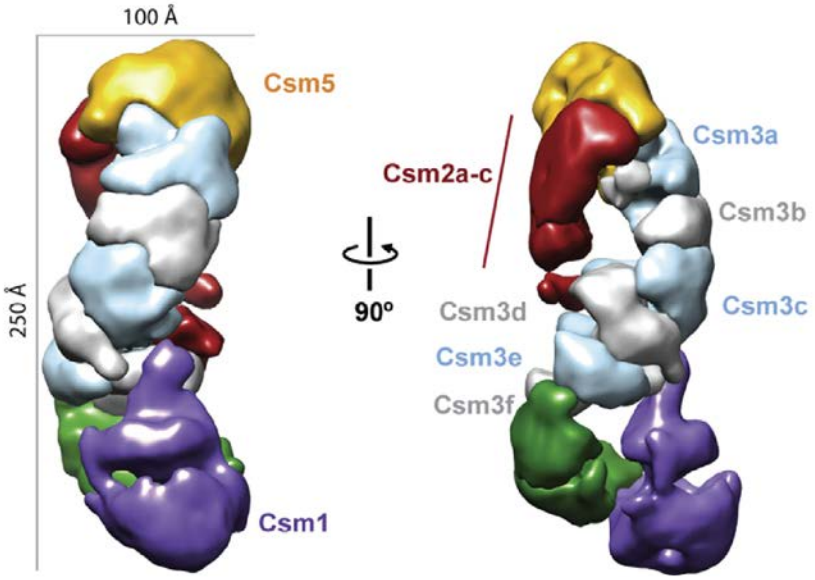


Figure 1.16 Molecular architecture of the *T. thermophilus* Type III-A Csm complex.

The CryoEM structure of the *Tt Csm* complex, reconstructed at 17 Å resolution, showing the "sea worm" architecture. Different segmented regions of the complex are indicated with different colors and respective labels: Csm1 (purple), Csm2 (red), Csm3 (alternating light blue and gray), Csm4 (green), and Csm5 (orange). Adapted from [59] with permission from the publisher.

The major difference between the *T. thermophilus*, Csm complex (*Tt Csm* complex) and the *E.coli* Cascade complex is during the interference step. Unlike Cascade complex, the Csm complex has been reported to harbor RNase activity rather than DNase activity under *in vitro* conditions with multiple sites present along the backbone. Interestingly, the Csm complex in *T. thermophilus* catalyzes the cleavage of RNA targets in a similar mechanism like the one reported for Type III-B Cmr complex [52, 59, 128]. Here also the protein-RNA interactions were investigated using the UV induced protein-RNA cross-linking method. The results were also compared with the protein-RNA cross-linking studies done in the *E. coli* Cascade complex.

1.3.4 Type III-B Cmr complex in *Thermus thermophilus*

The Type III-B Cmr complex in *T. thermophilus* (*Tt Cmr* complex) is composed of six different subunits Cmr1-Cmr6 and one crRNA in $Cmr1_1Cmr2_1Cmr3_1Cmr4_4Csm5_3Cmr6_1$:$crRNA_1$ stoichiometry. Two breakthroughs in understanding the architecture of Type III-B Cmr complex have been: i) the sea-worm shaped cryoEM structure of the *Tt Cmr* complex [52] and ii) a pseudo-atomic model of the complex where the crystal structure of individual subunits were combined with the previous structural information to generate a cryoEM map of the

Pyrococcus furiosus Cmr complex [56] (Figure 1.17). The overall sea-worm shaped architecture of TtCmr complex resembles Type I-E Cascade, however the composition is different. It comprises of a heterodimer tail of Cmr2 and Cmr3 subunits, a helical backbone of Cmr4 subunits capped with a Cmr5 subunit and a head comprising of Cmr1 and Cmr6 subunits [52].

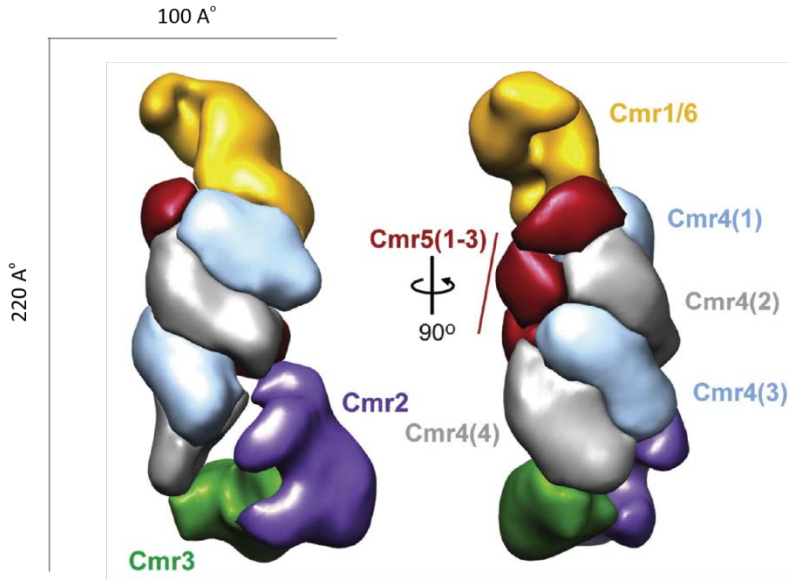


Figure 1.17 Molecular architecture of the *T. thermophilus* Type III-B Cmr complex.

The crystal structure of the *Tt Cmr* complex, reconstructed at 22 Å resolution, showing the "sea worm" architecture. Different segmented regions of the complex are indicated with different colors and respective labels: Cmr1/Cmr6 (orange), Cmr2 (purple), Cmr3 (green), Cmr4 (alternating light blue and gray) and Cmr5 (red). Adapted from [52] with permission from the publisher.

The distinct characteristic of Type III-B Cmr system, making it different from all the other CRISPR-Cas system is that it targets ssRNA and not DNA. Both endogenous and reconstituted Cmr complexes have been used to study the degradation of ssRNA target. The cleavage of target RNA starts at the 3' end with sequential endonucleolytic activity after every six nucleotides, proceeding towards the 5' end, also known as the 5' ruler mechanism [52].

Considering a similar approach we used both endogenous and reconstituted Cmr complexes to look for protein-RNA interaction sites using the UV induced protein-RNA cross-linking procedure. This way we could also check the reproducibility of our approach in the differentially assembled multi-subunit crRNP complexes.

1.3.5 The Cas7 protein family

The Cas proteins have been classified into different families on the basis of their sequences and functions. As mentioned above, the different Cas7 proteins are a characteristic feature of Type I and Type III systems, forming the backbone of the large multi-subunit interference complexes. The Cas7-family proteins belong to the RAMP superfamily comprising a highly conserved RRM domain and having a distinct nomenclature in every subtype (Table 1.1). Recently published crystal structures of Cas7 proteins from different subtypes indicate it is one of the most frequently investigated Cas protein for structural studies, at the same time highlighting a structural homology between different Cas7 proteins.

The overall architecture of these Cas7 proteins comprises of four domains, the core domain which has a β-α-β-β-α-β topology typical of RRM-like and ferredoxin-like folds and three insertion domains (a lid domain, a metal binding domain and a helical domain) flanking the core domain. The core domain is highly conserved within the three proteins and the topology of the insertion domains also show striking similarities [103] (Figure 1.18).

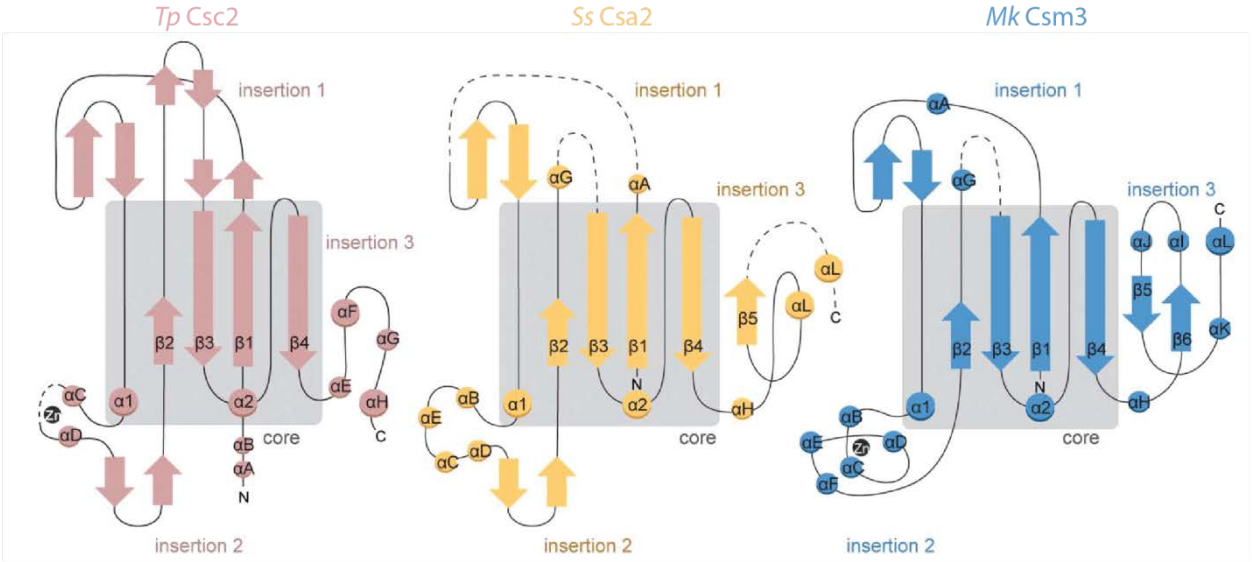


Figure 1.18 Comparison between the topology of three Cas7-family proteins *Tp Csc2*, *Ss Csa2* and *Mk Csm3*.

The highly conserved core domain is highlighted in the center (grey) also showing the connectivity with three insertion domains. The topological arrangement of the three insertion domains is also similar in the three proteins. Adapted from [103] with permission from the publisher.

Some of these examples include: the 2.4 Å resolution crystal structure of Type III-A *Methanopyrus kandleri* Csm3 (*Mk Csm3*) [55], 2.0 Å resolution structure of Type I-A *Sulfolobus solfataricus* Csa2 (*Ss Csa2*) [19] and 1.8 Å resolution structure of Type I-D *Thermofilum pendens* Csc2 (*Tp Csc2*) [103]. The topology diagrams of these three proteins show a high structural conservation within the core RRM-like fold (Figure 1.19).

Figure 1.19 Crystal structures of the three Cas7-family proteins: *Tp Csc2*, *Ss Csa2* and *Mk Csm3*.

Cartoon representation of the recently published crystal structures of Cas7 orthologs depicted after superimposition of their RRM-like domains. *Tp Csc2* (salmon) (PDB 4TXD) [103], *Ss Csa2* (orange) (PDB 3PS0) [19] and *Mk* Csm3 (blue) (PDB 4N0L) [55].

In order to identify the RNA-binding interface of Cas7 proteins, we used UV induced protein-RNA cross-linking approach. Two Cas7 proteins, *T. pendens* Csc2 and *Thermoproteus tenax* Cas7 were used for this study. Moreover the results were also compared with the cross-links identified in other Cas7 homologs: Cas7 protein in Type I-E Cascade complex and Csm3 protein in Type III-A Csm complex to look for putative RNA-protein interfaces in structurally conserved RNA interacting regions of this protein family.

2. Materials and Methods

2.1 Materials

2.1.1 Chemicals and Solvents

Acetone	Merck, Darmstadt, DE
Acetonitrile, LiChrosolv (ACN)	Merck, Darmstadt, DE
Ammonium Hydroxide 28-30% (w/v)	Acros Organics, New Jersey, USA
Ammonium persulfate	Serva Electrophoresis, Heidelberg, DE
BS3 (bis(sulfosuccinimidyl)suberate)	Pierce, Thermo Fischer, Schwerte, DE
Chloroform	Merck, Darmstadt, DE
Coomasie Brilliant Blue G-250	Sigma-Aldrich, Steinheim, DE
Dihydroxybenzoic acid (DHB)	Sigma- Aldrich, Steinheim, DE
Dimethylsulfoxide (DMSO)	Roth, Karlsruhe, DE
Dithiothrietol (DTT)	Roth, Karlsruhe, DE
Dodecyl sulfate sodium salt (SDS)	Merck, Darmstadt, DE
Ethanol	Merck, Darmstadt, DE
Ethylenediaminetetraacetic acid, (EDTA, disodium salt)	Roth, Karlsruhe, DE
Formic acid (FA)	Fluka, Sigma- Aldrich, Steinheim, DE
Formaldehyde (CH_2O) 37% (v/v)	Sigma-Aldrich, Steinheim, DE
Formaldehyde (CD_2O) 28% wt., 98%D	Isotec, Miamisburg, OH, USA
Glycogen	Roche, Mannheim, DE
Glycerol	Merck, Darmstadt, DE
Hepes	Merck, Darmstadt, DE
Iodoacetamide (IAA)	Sigma-Aldrich, Steinheim, DE
Magnesium chloride	Merck, Darmstadt, DE
Methanol, Lichrosolv	Merck, Darmstadt, DE
n-Hexane	Merck, Darmstadt, DE
Ortho-phosphoric acid	Merck, Darmstadt, DE
RapiGest	Waters, Manchester, UK
Sodium acetate	Merck, Darmstadt, DE
Sodium chloride	Merck, Darmstadt, DE
Sodium cyanoborohydride ($NaBH_3CN$)	Fluka, Sigma- Aldrich, Steinheim, DE
N,N,N',N'-tetramethylenethyldiamide (TEMED)	Sigma-Aldrich, Steinheim, DE
TFA (trifluoroacetic acid)	Roth, Karlsruhe, DE
TRIS tris(hydroxymethyl)aminomethane	Roth, Karlsruhe, DE
Urea	Sigma-Aldrich, Steinheim, DE
Universal Protein Standard 2 (UPS2)	Sigma-Aldrich, Steinheim, DE
Water, LiChrosolv	Merck, Darmstadt, DE

2.1.2 Commercial buffers and solutions

Acrylamide/Bis-acrylamide, Rotiphorese Gel 30 (37.5:1)	Roth, Karlsruhe, DE
Bradford solution	Bio-Rad Protein Assay, Bio-Rad, Munich, DE
Imperial protein stain	Pierce, Thermo Fischer, Schwerte, DE
IPG ampholytes buffer	GE Healthcare, Munich, DE
NuPAGE Antioxidant	Invitrogen, Karlsruhe, DE
NuPAGE LDS sample buffer (4X)	Invitrogen, Karlsruhe, DE
NuPAGE MOPS SDS running buffer (10X)	Invitrogen, Karlsruhe, DE
NuPAGE Sample reducing agent (10X)	Invitrogen, Karlsruhe, DE
PCI solution (phenol/chloroform/isoamyl alcohol 25:24:1)	Roth, Karlsruhe, DE
PNK buffer (10X)	New England Biolabs, Frankfurt, DE
Protein marker, Precision Plus (Unstained)	Bio-Rad, Munich, DE
Triethylammonium bicarbonate buffer (TEAB) 1.0 M, pH 8.5 ± 0.1	Sigma-Aldrich, Steinheim, DE

2.1.3 Enzymes and Enzyme inhibitors

Benzonase	Novagen, EMD Chemicals, CA, USA
Protease inhibitors, EDTA free	Roche, Mannheim, DE
RNase A	Ambion, Applied Biosystems, Darmstadt, DE
RNase T1	Ambion, Applied Biosystems, Darmstadt, DE
T4 polynucleotide kinase (T4 PNK)	New England Biolabs, Frankfurt, DE
Trypsin, modified (sequencing grade)	Promega, Madison, WI, USA
Trypsin	Serva Electrophoresis, Heidelberg, DE

2.1.4 Proteins, peptides and oligonucleotides

ATP, [Υ^{-32}P]-labeled	PerkinElmer, Waltham, MA, USA
BSA standard, Bradford	Thermo Fischer, Schwerte, DE
d(-1) repeat *M. maripaludis*: 5'-CUAAAAGAAUAACUUGCAAAAUAACA AG(dC)AUUGAAAC-3'	MWG, Eurofins Genomics, Ebersberg, DE
d(-1) repeat *C. thermocellum*: 5'-GUUUUUAUCGUACCUAUGAGG(dA)AU UGAAAC-3'	MWG, Eurofins Genomics, Ebersberg, DE
poly(U)$_{15}$	Purimex, Grebestein, DE

2.1.5 CRISPR proteins and protein – RNA complexes for quantitative and structural proteomics studies

Proteins and protein – RNA complexes	Provided By
For quantitative proteomics studies	
H. volcanii (H119 WT and H119 Δ*Cas7*) protein extract	Britta Stoll, Prof. Anita Marchfelder, Ulm University, Ulm, DE
H. volcanii FLAG-Cas7 purified fraction with Cas5 and Cas6 proteins co purified.	Jutta Brendel, Prof. Anita Marchfelder, Ulm University, Ulm, DE
C. thermocellum Type I-B Cascade complex	Judith Zöphel, Prof. Lennart Randau, MPI Terrestrial Microbiology, Marburg, DE
For structural proteomics studies	
M. maripuldis Cas6b and *C. thermocellum* Cas6	Hagen Richter, Prof. Lennart Randau, MPI Terrestrial Microbiology, Marburg, DE
T. tenax Cas7 and *T. pendes* Csc2	Ajla Hrle, Prof. Elena Conti, MPI Biochemistry, Martinsried, DE
C. thermocellum Type I-B Cascade complex	Judith Zöphel, Prof. Lennart Randau, MPI Terrestrial Microbiology, Marburg, DE
E.coli Type I-E Cascade complex	Tim Kunne, Dr. Stan Brouns, Wageningen Univesity, NL
T. thermophilus Type III-A Csm Complex (endogenous)	Raymond Staals, Prof. Jon van der Oost, Wageningen Univesity, NL
T. thermophilus Type III-B Cmr Complex (both endogenous and reconstituted)	Raymond Staals, Yifan Zhou, Prof. Jon van der Oost, Wageningen Univesity, NL

2.1.6 Commonly used buffers and solutions

10 mM DTT	1.54 mg/ml DTT in water (prepared fresh)
60 mM IAA	11.3 mg/ml IAA in water (prepared fresh, kept in dark)
1 M Tris-HCl buffer, pH 7.9	Tris base, desired pH adjusted with 37% (v/v) HCl
8M Urea	9.6g Urea in 12.8 ml water, Filtered (prepared fresh)
BS3 Stock solution	2 mg BS3, 100 µl DMSO (prepared fresh)
Colloidal Coomasie staining solution	0.08% (w/v) Coomasie Brilliant Blue G-250 20% (v/v) Methanol 1.6% (v/v) Orthophosphoric acid 8% (w/v) Ammonium sulfate
CE Buffer	10 mM Cacodylic acid pH 7.0, 0.2 mM EDTA pH 8.0
LC-MS Sample Loading buffer	5% (v/v) acetonitrile, 0.1% (v/v) formic acid

SDS Resolving gel buffer 4x	1.5 M Tris-HCl, pH 8.8
SDS Stacking gel buffer 4x	0.5 M Tris-HCl, pH 6.8
SDS Running buffer 1X	25mM Tris, 192 mM Glycine, 0.1% (w/v) SDS
SDS Sample buffer	60 mM Tris, 1 mM EDTA, 16% (v/v) Glycerine, 2% SDS, 0.1% (w/v) Bromophenol blue, 50 mM DTT

2.1.7 Other consumables

C18 Column Material • Reprosil-Pur basic C18-HD, 120 Å, 3 or 5µm	Dr. Maisch, Ammerbuch, DE
C18 Empore Octadecyl 47 mm Extraction Discs	3M, Bellefonte, PA, USA
IPG Strip 7 cm and 18 cm	GE Healthcare, Munich, DE
IPG Dry Strip cover fluid, Plus One	Pharmacia Biotech, Uppsala, Sweden
IPG Strip Holders for 7 cm and 18 cm strips	GE Healthcare, Munich, DE
IPG Strip Holder cleaning solution	GE Healthcare, Munich, DE
Microtiterplates for UV cross-linking • Black polypropylene 96-well microtiter plates (#655209)	Greiner Bio-One, Frichenhausen, DE
MicroSpin Columns G-25 (for RNA isolation)	GE Healthcare, Munich, DE
NuPAGE Novex 4-12% Bis-Tris Mini Gels, 1mm	Invitrogen, Karlsruhe, DE
Phosphoimager Screens • Storage Phosphor Screens	Molecular Dynamics, GE Healthcare, Munich, DE
Sep-Pak Vac C18 Columns	Waters, Eschborn, DE
Superdex 200 PC 3.2/30	GE Healthcare, Munich, DE
Superdex 75 PC 3.2/30	GE Healthcare, Munich, DE
Superdex Peptide PC 3.2/30	GE Healthcare, Munich, DE
Titanium dioxide TiO_2 Column Material • Titansphere 5 µm	GL Sciences, Tokyo, JP
Whatman 3mm, CHR	GE Healthcare, Munich, DE

2.1.8 Instruments and Laboratory equipments

Autoclave, Varioklav steam sterilizer H+P	Thermo Fisher Scientific, Schwerte, DE
Centrifuges	
• Eppendorf Centrifuge 5415R	Eppendorf, Hamburg, DE
• Heraeus Fresco 17	Thermo Fisher Scientific, Schwerte, DE
• Heraeus Pico 17	Thermo Fisher Scientific, Schwerte, DE
• Heraeus Biofuge pico	Thermo Fisher Scientific, Schwerte, DE
• Heraeus Megafuge 1.0 R	Thermo Fisher Scientific, Schwerte, DE
Clean Bench, HeraSafe, Heraeus	Thermo Fisher Scientific, Schwerte, DE
Cross-linking Apparatus (Build In-house) operated with four 8W lamps, 254 nm, G8T5	Sankyo Denki, Japan
Gel Dryer Model 583	Bio-Rad, Munich. DE
IPGphor for pIEF	Pharmacia Biotech, San Francisco, CA, USA
LC-MS	
• HPLC, 1100 series	Agilent Technologies, Böblingen, DE
• Dionex, Ultimate 3000 UHPLC	Thermo Fisher Scientific, Schwerte, DE
• LTQ Orbitrap XL	Thermo Fisher Scientific, Schwerte, DE
• LTQ Orbitrap Velos	Thermo Fisher Scientific, Schwerte, DE
• Q Exactive HF	Thermo Fisher Scientific, Schwerte, DE
• Orbitrap Fusion	Thermo Fisher Scientific, Schwerte, DE
Thermomixer Comfort	Eppendorf, Hamburg, DE
ThermoStat plus	Eppendorf, Hamburg, DE
Pharmacia SMART system	Pharmacia, GE Healthcare, Munich, DE
Phosphoimager, Typhoon 8600	GE Healthcare, Munich, DE
Scintillation Counter, Tri-Carb 2100TR	Packard, PerkinElmer, Waltham, MA, USA
Spectrophotometer, Ultrospec, 3000 pro	Pharmacia, GE Healthcare, Munich, DE
SpeedVac Concentrator	
• Savant SPD121P	Thermo Fisher Scientific, Schwerte, DE
• Eppendorf Concentrator 5301	Eppendorf, Hamburg, DE
Vortex Genie 2	Scientific Industries, Roth, Karlsruhe, DE

2.2 Methods

2.2.1 Cell culture, expression and purification of proteins and protein – RNA complexes

2.2.1.1 Protein extracts for quantitative proteome analysis of H. volcanii wild type and cas7 knock out

The following steps for generating the *cas7* KO strain and preparation of protein extracts for proteome analysis were performed by Britta Stoll, in the lab of Prof. Anita Marchfelder, Ulm University, Ulm, DE.

***H. volcanii* strain**	***Special features for auxtrophic selection* [129]**
H119 wild type (H119 WT)	DS70 Wildtype (ΔpHV2) Δ*pyrE2* Δ*trpA* Δ*leuB*
Cas7 knock out (Δ*cas7* KO)	DS70 Wildtype (ΔpHV2) Δ*pyrE2* Δ*trpA* Δ*leuB* Δ*cas7*

The *cas7* gene was knocked out by transformation of H119 WT using the Pop In/ Pop Out method [130] and the knockouts were verified with PCR and Southern blot. The method for generation of knock out strains and preparation of protein extracts has been described previously in [126]. Briefly, the cells were grown in YPC medium and harvested and washed in ice-cold salt-water. The cells were lysed in Sodium Taurodeoxycholate (0.006% final concentration) and the insoluble cell components were separated with ultracentrifugation. The supernatant (extract) was dialyzed (in 10 mM Tris-HCl, 6 mM MgCl2, pH 7.5) overnight and was treated with DNase1, Exonuclease III and RNaseA. The extract was further dialyzed (in 2 mM Tris-HCl, pH 7.5) overnight. The proteins were precipitated from the extract with 100% acetone (at -20 °C overnight), washed with ice-cold 80% (v/v) ethanol and frozen in liquid nitrogen and stored at -80 °C.

2.2.1.2 Expression and purification of H. volcanii Cas5, Cas6 and Cas7 proteins for iBAQ analysis

The following steps were performed by Jutta Brendel, in the lab of Prof. Anita Marchfelder, Ulm University, Ulm, DE. Further details for every step have been described in [89]. Briefly, *H. volcanii* strain H26Δ*cas*Cluster28 (with no cas genes at all) [129] carrying a *Haloferax*-overexpression-vector comprising cas-genes *cas6*, *cas8, cas7* and *cas5* was used in this study. The *cas7* was fused to a combined His-and-FLAG-tag for purification [89]. To express the tagged Cas7 protein, the cells were grown in Hv-Ca medium (enhanced casamino broth) containing tryptophan. The cells were harvested and lysed (in 1 M NaCl, 100 mM Tris-HCl, pH 7.5, 1 mM EDTA, 10 mM $MgCl_2$, 1 mM $CaCl_2$, 8 units/µl DNase RQ1, 13 µl/ml protease inhibitor mixture) by sonification. Insoluble cell debris was removed by ultracentrifugation. The 3×FLAG tagged

protein was purified using the anti-FLAG M2 affinity gel (Sigma-Aldrich, Steinheim, DE). The final elution was performed using the 3×FLAG peptide (Sigma-Aldrich, Steinheim, DE). This FLAG-Cas7 purified fraction comprising co-purified Cas5 and Cas6 proteins was used for the iBAQ analysis.

2.2.1.3 Expression and purification of M. maripaludis Cas6b and C. thermocellum Cas6b

The following steps were performed by Hagen Richter, in the lab of Prof. Lennart Randau, MPI Terrestrial Microbiology, Marburg, DE. The details for expression and purification of *M. maripaludis* Cas6b have been described in [127]. Briefly, the respective *cas6* genes were cloned into pET-20b vector for the protein expression with a C-terminal His-tag. To generate Cas6 variants, mutations were induced using site-directed mutagenesis. These Cas6 variant proteins were expressed in *E. coli* (Rosetta2 DE3) cells with IPTG induction. After induction the cells were harvested and lysed (in 10 mM Tris–HCl pH 8.0, 300 mM NaCl, 10% (v/v) glycerol and 0.5 mM DTT) with lysozyme (1 mg/g cell pellet) using sonication. The lysate was cleared with centrifugation and the supernatant was applied to a Ni–NTA–Sepharose column for purification. Elution of the proteins was performed by a linear imidazole gradient (0–500 mM) and the purified proteins were dialyzed into lysis buffer (without glycerol) and used for further analysis.

2.2.1.4 Expression and purification of T. tenax Cas7 and T. pendens Csc2

The following steps were performed by Ajla Hrle, in the lab of Prof. Elena Conti, MPI Biochemistry, Martinsried, DE. The details for expression and purification of *T. pendens* Csc2 and *T. tenax* Cas7 have been described in [103, 131]. Briefly, the gene construct for *T. tenax* Cas7 was cloned in pET24a (+) and the gene for *T. pendens* Csc2 was ordered as a synthetic construct (GeneArt, Life technologies). The full-length proteins were expressed as a recombinant His-SUMO-tagged using *E. coli* BL21-Gold (DE3) Star pRARE cells (Stratagene). The cells were harvested and lysed in 50 mM Tris-HCl, pH 7.5, 1 M NaCl, 10 mM imidazole and 10% glycerol, supplemented with protease inhibitors. The lysate was cleared with centrifugation and the supernatant was applied to Ni^{2+} affinity chromatography and further a HiTrap Heparin column (GE Healthcare) for purification. The proteins were treated with SUMO protease for the removal of His-tag. Final purification was performed on a Superdex 75 column using 20 mM Hepes pH 7.5, 150 mM NaCl and 5 mM DTT and 10% glycerol.

2.2.1.5 Preparation of E.coli Type I-E Cascade complex

The following steps were performed by Tim Künne, in the lab of Dr. Stan J. J. Brouns, Wageningen University, Wageningen, NL. The details for the assembly and purification of Type I-E Cascade complex from *E. coli* have been described in [29, 30, 44]. Briefly, the cas genes and

CRISPRs were PCR amplified from *E. coli* K12 genomic DNA and directionally cloned into compatible expression vectors as described in [29]. The Plasmids were transformed into *E. coli* BL21 (DE3) lacking endogenous *cas* genes. Cells were harvested, resuspended in lysis buffer (20 mM Hepes, 75 mM NaCl, 1 mM DTT, pH 7.5) and disrupted using a French Pressure Cell. The pre-assembled complex was pulled down using the Strep-Tactin column (IBA, Germany) following manufacturer's instructions using different elution buffer (20mM Hepes, 75mM NaCl, 1mM DTT, 2.5 mM (Desthiobiotin).

2.2.1.6 Preparation of T. thermophilus Type III-A Csm complex

The following steps were performed by Raymond Staals, in the lab of Prof. John van der Oost, Wageningen University, Wageningen, NL. The details for the assembly and purification of Type III-A Csm complex from *T. thermophilus* are described in [59]. Briefly, the C-terminal $(His)_6$-tagged Csm5 was produced by inserting the tag-coding sequence within the genome of *T. thermophilus* HB8 by homologous recombination using pUC-csm5h plasmid. The cells were resuspended in lysis buffer (20 mM Tris-HCl, pH 8.0, 50 mM NaCl, 0.1 mM phenylmethylsulfonyl fluoride) and disrupted by sonication. The lysate was separated by ultracentifugation, and the supernatant was applied to a series of columns in a sequential manner for purification and desalting. The columns used included: HisTrap HP column (GE Healthcare), HiPrep 26/10 desalting column (GE Healthcare), RESOURCE Q column (GE Healthcare), HiLoad 16/60 Superdex 200 pg column (GE Healthcare), HiPrep 26/10 desalting column, HiTrap Heparin column (GE Healthcare), HiPrep 26/10 desalting column and finally CHT2-1 column (Bio-Rad Laboratories, Inc.). The final purified complex was resuspended in 20 mM Tris-HCl, pH 8.0, 150 mM NaCl.

2.2.1.7 Preparation of T. thermophilus Type III-B Cmr complex

The following steps were performed by Yifan Zhu and Raymond Staals, in the lab of Prof. John van der Oost, Wageningen University, Wageningen, NL. The details for assembly and purification of Type III-B Cmr complex from *T. thermophilus* are described in [52]. Briefly, The C-terminal $(His)_6$-tagged Cmr6 was produced by inserting the tag-coding sequence within the genome of *T. thermophilus* HB8 by homologous recombination using pUC-cmr6h plasmid. The cells were resuspended in lysis buffer (20 mM Tris-HCl, pH 8.0, 50 mM NaCl, 0.1 mM phenylmethylsulfonyl fluoride) and disrupted by sonication. The lysate was separated by ultracentifugation, and the supernatant was applied to a series of columns in a sequential manner for purification. The columns used included: HisTrap HP column (GE Healthcare), HiPrep 26/10 desalting column (GE Healthcare), RESOURCE Q column (GE Healthcare) and finally, HiLoad 16/60 Superdex 200 pg column (GE Healthcare). The final purified protein-complex was resuspended in 20 mM Tris-HCl, pH 8.0, 150 mM NaCl.

For the reconstituted Cmr complex, *cmr1*, *cmr2*, *cmr3*, *cmr4*, *cmr5*, and *cmr6* genes were each amplified by genomic PCR. Each recombinant protein was expressed in *E. coli* by means of pET-expression system (Merck). The purified proteins were then mixed *in vitro* with the crRNA in the required stoichiometry.

2.2.1.8 Preparation of C. thermocellum Type I-B Cascade complex

The following steps were performed by Judith Zöphel, in the lab of Prof. Lennart Randau, MPI Terrestrial Microbiology, Marburg, DE. The Cas5-SUMO, Cas6-HIS, Cas7-SUMO and Cas8b-HIS proteins were expressed as described in [127, 131] and purified with a Nickel-NTA column. The elution was carried out with a linear imidazole gradient of 0-500 mM imidazole. Cas5 and Cas7 were dialyzed together with SUMO-protease (since both proteins contain a N-terminal SUMO-tag that needs to be cleaved off) overnight. Cas8b was further purified using a Heparin column (as it was highly contaminated with nucleic acids) and the elution was carried out with a linear salt gradient of 0-1 M NaCl [127]. In order to generate the crRNA, a substrate consisting of Spacer-Repeat-Spacer-Repeat-Spacer was cloned in puc19. The precursor RNA was generated via in-vitro transcription using the linearized vector [127] and was cleaved with Cas6 protein to generate the crRNA.
For the assembly, Cas6 protein along with crRNA, was mixed with Cas 8b and the dialyzed Cas7 and Cas5 proteins and incubated at 50°C for 30 min.

2.2.2 Standard molecular biology methods

2.2.2.1 PCI extraction

In order to separate the nucleic acids from proteins, phenol-chloroform-isoamylalcohol (PCI) extraction was used. The sample was mixed with 1 volume of PCI solution and 1 µl of 1 µg/µl glycogen followed by vigorous shaking for 15 min. The sample was then centrifuged for 5 min, 13000 rpm at room temperature. The upper aqueous phase containing the nucleic acids, was transferred into a fresh tube. This aqueous phase was further purified with addition of 1 volume of chloroform, followed by vigorous shaking and phase separation as mentioned above. Again the upper phase was collected and the nucleic acids were recovered from this aqueous mixture by ethanol precipitation.

2.2.2.2 Ethanol precipitation

Proteins, nucleic acids or RNP complexes were precipitated by the addition of 3 volumes of ice cold (-20 °C) ethanol and 1/10 volume of 3 M NaOAc pH 5.3 and incubation at -20 °C for at least 2 h. The precipitated sample was then centrifuged for 30 min at 13000 rpm and 4 °C and the supernatant was removed with thin tips. The pellet was washed with 2 volumes of 80% ice cold

(-20 °C) ethanol and centrifuged as above. The supernatant was removed and the pellet was dried in a SpeedVac concentrator only for a short duration until all the solvent was gone.

2.2.2.3 5' labeling of RNA with γ-[^{32}P]-ATP

RNA oligonucleotides were labeled at the 5' end with γ-[^{32}P]-ATP and T4 polynucleotide kinase (PNK). The reaction mixture comprised of 5 pmol of RNA oligo, 1 µl of 10X PNK buffer, 5 µl γ-[^{32}P]-ATP and 1 µl of T4 polynucleotide kinase and the final volume was made up to 10 µl with water. This mixture was incubated at 37 °C for 1 h. After the reaction, 40 µl CE buffer were added to the mixture and free γ-[^{32}P]-ATP was removed by loading the mix on a G-25 MicroSpin column, followed by centrifugation at 3000 rpm for 2 min. To the elute, 150 µl CE buffer and 1 µl glycogen were added and the RNA was purified with PCI extraction (see 2.2.2.1). Final RNA pellet was dissolved in CE buffer.

2.2.3 Standard protein biochemical methods

2.2.3.1 Determination of Protein concentration

Protein concentrations were determined by using the Bradford protein assay [132]. The protein sample was diluted with water to make a final volume of 800 µl, followed by the addition of 200 µL Bradford solution. The mix was kept in dark, for 10 min at room temperature and the absorbance was measured at 595 nm. The protein concentration was determined by comparison with a standard curve. For this purpose a standard curve was determined for a standard dilution series of BSA 0-20 µg using BSA standard stock solution (0.2 mg/mL), to which the Bradford solution was added. Three independent measurements of three different protein concentrations were performed for the same protein sample and averaged to determine the final protein concentration.

2.2.3.2 Denaturing polyacrylamide gel electrophoresis using NuPAGE system

Proteins were separated using the NuPAGE system according to the manufacturer's protocols. Briefly, the protein samples were mixed with 10X NuPAGE sample reducing agent, 4X NuPAGE sample buffer heated for 10 min at 70 °C. Running buffer was prepared by diluting 20X NuPAGE MOPS SDS running buffer. Samples were loaded onto pre-cast 4-12% Bis-Tris 1.0 mm gels and run for 50 min at 200V with the addition of NuPAGE antioxidant in the center buffer chamber.

2.2.3.3 Denaturing polyacrylamide gel electrophoresis using self-cast gels

The protein separation was performed using self-cast gels, samples were mixed with the SDS sample buffer in 1:1 ratio (v/v) and heated at 95 °C for 2 min prior to loading. Both stacking and resolving gels were prepared using the recipe below. Gels were run with 1X SDS buffer at 28 mA for stacking and 45 mA during protein separation in resolving gel.

	4% Stacking Gel (10 ml)	12% Resolving Gel (10ml)
H_2O	6.1 ml	3.1 ml
Acrylamide/Bis-acrylamide	1.3 ml	4 ml
4x buffer	2.5 ml	2.5 ml
10% SDS	100 µl	100 µl
10% APS	100 µl	100 µl
TEMED	10 µl	10 µl

2.2.3.4 Colloidal Coomassie staining

The proteins separated by SDS-PAGE were stained with colloidal coomassie [133]. The gel was covered in colloidal coomassie staining solution and incubated overnight with gentle shaking. The destaining was performed with several wash and rinse cycles with water. For fast staining, the gel was incubated with Imperial protein stain for 30 min and destained the same way as mentioned above.

2.2.4 Quantitative proteomics by differential isotope labeling

Quantitative proteome analysis of *H. volcanii* wild type and *cas7* knock out (H119 WT vs H119 Δ*cas7*) was carried out using the dimethyl labeling approach. Protein extracts were digested in the presence of 8M Urea. The sample was free of any primary-amine containing molecules, other than the peptides, to achieve maximum labeling efficiency. The peptide mixture was then separated into fractions using peptide isoelectric focusing to overcome the high complexity of sample prior to MS analysis.

2.2.4.1 In-solution digestion in presence of 8M Urea

The dried protein pellets or protein complexes were dissolved in 20 µl 8 M Urea, with vigorous shaking for 30 min at 25 °C. The disulfide bridges in proteins were reduced with addition of 1 µl 200 mM DTT (in 100 mM TEAB), followed by incubation for 1 h at 600 rpm in a thermomixer at 25 °C. Further, the –SH groups were alkylated with addition of 1 µl 1.2 mM IAA (in 100 mM TEAB), followed by incubation as in the previous step. The reaction mixture was taken to a final volume of 200 µl with 100 mM TEAB and modified trypsin was added in a 1:40 w/w ratio for overnight hydrolysis at 600 rpm in a thermomixer at 25 °C.

2.2.4.2 Dimethyl labeling of peptides from In-solution digestion

The peptides were labeled with dimethyl labeling as described in [84] with slight modifications. To the in-solution digested mix from above, the dimethyl labeling reagents were added directly

and all the steps for labeling were carried out in a fume hood. To generate light dimethylation 8 µl of 4% (v/v) formaldehyde and to generate medium dimethylation 8 µl of 4 % (v/v) CD_2O (heavy formaldehyde) was added to the sample. The samples were mixed with gentle vortexing, followed by addition of 8 µl of 0.6 M $NaBH_3CN$ and incubation in a fume hood for 1 h at 600 rpm (in a thermomixer) at 17 °C. The labeling reaction was quenched by addition of 32 µl of 1 % (v/v) Ammonia solution. After mixing and spinning the solutions down, 10 µl formic acid was added to further quench the reaction and to acidify the samples for further steps. Both light and medium labeled samples were then pooled in a 1:1 ratio.

2.2.4.3 Desalting with Sep-Pak Vac C18 column

Desalting was carried out in Sep-Pak Vac C18 column placed in a 15 ml falcon. All washing, loading and elution steps were performed by centrifugation at 500 g for 1 min. The column was equilibrated successively by passing 500 µl of methanol, 500 µl of 80% (v/v) ACN, 1.0% (v/v) FA and 500 µl of 1.0% (v/v) FA. The pooled sample after dimethyl labeling was loaded onto the column, followed by washing the column twice with 500 µl of 1.0% FA. The peptides were eluted with 500 µl of 80% ACN, 1.0% FA in a fresh 15 ml falcon. The elute was dried in a SpeedVac concentrator.

2.2.4.4 Peptide iso-electric focusing (pIEF) and LC-MS/MS analysis

The pIEF was performed using 18cm IPG strips (pH 3-10). The dried peptides were dissolved in 350 µl 8 M Urea, 0.2% (v/v) IPG ampholytes buffer and the peptide solution was pipetted into the middle of an 18 cm strip holder, between the electrodes in a drop wise manner. The IPG strip was placed on top (gel facing the electrodes) and it was covered with 1.5 ml IPG Dry Strip cover fluid. The gel was left to rehydrate over night at 20 °C without applied voltage. Following the rehydration, the peptides were separated on an IPGPhor for a total of 30,000 Vh at max 50 µA per strip. The following parameters were used: 500 Vh at 500 V, then 1750 Vh at a gradient from 500 V to 3000 V and 27750 Vh at 8000V, all the steps were performed at 20 °C. Immediately after pIEF, the IPG strip was removed and immersed in n-hexane for 10 sec to remove excessive cover fluid. Afterwards, the gel on the strip was manually cut into 18 slices (1 cm each) and each slice was further cut into three small pieces before putting in reaction tubes. The peptides were extracted from these slices in a sequential manner at 26 °C, 1050 rpm with 30 min for each step. Gel slices were subjected to a series of extraction steps with 50% (v/v) ACN, 1% (v/v) FA and 100% (v/v) ACN, 1% (v/v) FA and the supernatant from each step was pooled in a reaction tube and dried in a SpeedVac.

2.2.4.5 C18 Desalting and LC-MS/MS analysis

The desalting and removal of interfering substances was performed using STAGE-Tips as described in [134] with two C18 discs. The eluted peptides were dried in a SpeedVac until all the solvent was removed. The dried peptides were frozen at -20 °C.

Before submission for MS analysis the samples were dissolved in 20 µl LC-MS sample loading buffer. For every slice 5 out of 20 µl was injected into the LTQ Orbitrap XL (details under Section 2.2.8). The data analysis was performed using MaxQuant software (details under Section 2.2.9.1).

2.2.5 Absolute quantification using iBAQ

The iBAQ analysis was performed to determine the stoichiometry of Cas proteins in Type I-B Cascade complexes from *H. volcanii* and *C. thermocellum* respectively.

The in-solution digestion was performed using 1 µg of protein sample, mixed with UPS2 protein mix in a 1:1 (w/w) ratio, in 10 µl 1% (w/v) Rapigest (in 100 mM TEAB). In order to reduce disulfide bridges, 10 µl of 10 mM DTT (in 100 mM TEAB) was added, and the sample was incubated for 1 h at 600 rpm. In order to alkylate –SH groups, 10 µl of 60 mM IAA (in 100 mM TEAB) was added and incubated as in the previous step. The reaction mixture was taken to a final volume of 100 µl with 100 mM TEAB and modified trypsin was added in a 1:20 w/w ratio for overnight hydrolysis at 600 rpm in a thermomixer at 37 °C. After overnight proteolysis, 20 µl 5% (v/v) TFA was added to decompose the Rapigest. Samples were centrifuged at 13000 rpm for 30 min and the supernatant was transferred to a new tube, dried in a SpeedVac and either stored at -20 °C or directly dissolved in 20 µl LC-MS sample loading buffer.

For the LC-MS analysis, the dissolved sample was measured in three technical replicates of 5 µl, in a LTQ Orbitrap Velos (Section 2.2.8). The data analysis was performed using iBAQ function in MaxQuant software (Section 2.2.9.2).

2.2.6 UV induced protein-RNA cross-linking

The UV induced cross-linking was performed with recombinant RNA-binding proteins and their (cognate) RNA oligonucleotides and also with the endogenous protein-RNA complexes isolated from prokaryotic cells

2.2.6.1 UV induced cross-linking of γ-[^{32}P]-ATP labeled crRNA and Cas6b proteins

The single stranded crRNA oligonucleotide (5 pmol) was radiolabeled with γ-[^{32}P]-ATP (6000 Ci/mmol) following standard procedures (Section 2.2.2.3). The resulting labeled RNA was incubated for 30 min on ice with approximately 250 pmol (50 fold molar excess) of purified Cas6b proteins. The resulting protein-RNA complex was purified with ethanol precipitation. The

complex was then UV irradiated at 254 nm for 10 min, using a cross-linking apparatus build in-house. For cross-linking, the samples were placed on ice at approximately 1 cm distance from the lamps. The cross-linked samples were analyzed by SDS-PAGE and the gel was coomassie stained. The bands containing radiolabeled RNA were visualized with a STORM phosphoimager (Figure 3.4).

2.2.6.2 Protein-RNA cross-linking for Cas6b proteins with their cognate crRNAs

UV induced protein-RNA cross-linking and enrichment of cross-linked heterconjugates was performed using the standard workflow as described in [108, 109, 135]. Briefly, 1 nmol of the crRNA was incubated in a 1:1 molar ratio with the Cas6b protein and incubated on ice for 30 min. The volume of the resulting mixture was raised to 100 µl in buffer containing 10 mM Tris–HCl pH 8.0, 300 mM NaCl and 0.5 mM DTT and transferred to black polypropylene micro-plates (Greiner Bio-One) positioned under the UV lamp at a distance of about 1 cm. The sample was then irradiated at 254 nm for 10 min (maximum) and then transferred back into a reaction tube. Cross-linked samples were ethanol precipitated and redissolved in 50 µl 4 M Urea, 50 mM Tris-HCl pH 7.9. Sample volume was raised to 200 µl with Tris-HCl pH 7.9 buffer and RNA hydrolysis was performed with Benzonase (25 U/µl) (Novagen), for which $MgCl_2$ was added to the digestion buffer to a final concentration of 1 mM and the hydrolysis was carried out at 37 °C for 1 h. RNA digestion was followed by overnight trypsin proteolysis at 37 °C. Modified trypsin was added at a 1:20 w/w ratio. Samples were then subjected to enrichment via C18 and TiO_2 chromatography. The dried sample pellets after enrichment, were dissolved in 2 µl 50% ACN, 0.1% FA and diluted to final concentration of 10% ACN, 0.1% FA by addition of 10 µl 0.1% FA. Of the 12 µl sample volume, 8 µl were injected for a single LC-MS run. The LC-MS analysis was performed on an LTQ Orbitrap Velos mass spectrometer.

2.2.6.3 Protein-RNA cross-linking for T. tenax Cas7 and Tp Csc2 proteins with poly(U)15

The *T. tenax* Cas7 and *Tp* Csc2 proteins were purified as described in 2.2.1.4. For cross-linking, 1 nmol of the purified protein was incubated in a 1:1 molar ratio with poly$(U)_{15}$ and incubated at 50 °C for 15 min. Final volume of the mixture was raised to 100 µl in a buffer containing 20 mM HEPES pH 7.5, 150 mM NaCl and 5 mM DTT. UV cross-linking was performed at 50 °C for 10 min as described in Section 2.2.6.2. Cross-linked samples were purified by ethanol precipitation and redissolved in 50 µl 4M Urea, 50 mM Tris-HCl pH 7.9. The sample volume was raised to 200 µl with Tris-HCl pH 7.9 buffer and RNA hydrolysis was performed with 1 µl RNase A (1 µg/µl) (Ambion, Applied Biosystems) at 52 °C for 2 h. RNA digestion was followed by overnight trypsin proteolysis at 37° C. Modified trypsin was added at a 1:20 w/w ratio. Enrichment of cross-links was performed as described in 2.2.6.2 and the LC-MS analysis was performed on an LTQ Orbitrap Velos mass spectrometer.

2.2.6.4 Protein-RNA cross-linking for the E. coli Cascade complex

The endogenous complex from *E. coli* was purified as described in 2.2.1.5. Prior to cross-linking, 1 nmol of the pre-assembled complex was incubated at 37 °C for 10 min. The sample volume was taken to 100 µl in buffer containing 20 mM HEPES, 75 mM NaCl, 1 mM DTT, 2.5 mM (Desthiobiotin). UV cross-linking was performed as described in 2.2.6.2. Cross-linked samples were purified by ethanol precipitation and redissolved in 50 µl 4 M Urea, 50 mM Tris-HCl pH 7.9. The sample volume was raised to 200 µl with Tris-HCl pH 7.9 buffer and RNA hydrolysis was performed for 2 h at 52 °C using 1 µl each of (1 µg/µl) and T1 (1 U/µl) (Ambion, Applied Biosystems). In addition, further hydrolysis was performed with 1 µl Benzonase, for which $MgCl_2$ was added to the digestion buffer to a final concentration of 1 mM and the hydrolysis was carried out at 37 °C for 1 h. After RNA hydrolysis, trypsin (Promega) was added in a protein-to-enzyme ratio of 20:1 (w/w) followed by an overnight incubation at 37 °C. Enrichment of cross-links was performed as described in 2.2.6.2 and LC-MS analysis was performed using Q Exactive HF mass spectrometer.

2.2.6.5 Protein-RNA cross-linking for Tt Csm and Tt Cmr complex

The *Tt* Csm and *Tt* Cmr complexes from *T. thermophilus* were purified as described in 2.2.1.6 and 2.2.1.7. Prior to cross-linking, 1 nmol of the pre-assembled complex was incubated at 65 °C for 10 min. The sample volume was raised to 100 µl in buffer containing 20 mM Tris-HCl, pH 8.0, 150 mM NaCl. UV cross-linking was performed at room temperature for 10 min as described above. Cross-linked samples were purified by ethanol precipitation and redissolved in 50 µl 4 M Urea, 50 mM Tris-HCl pH 7.9. Samples were then taken to 200 µl with Tris-HCl pH 7.9 buffer and RNA hydrolysis was performed for 2 h at 52 °C using 1 µl each of RNase A (1 µg/µl) and T1 (1 U/µl). RNA digestion was followed by overnight trypsin proteolysis at 37 °C. Modified trypsin was added at a 1:20 w/w ratio. Enrichment of cross-links was performed as described in 2.2.6.2 and LC-MS analysis was performed using LTQ Orbitrap Velos mass spectrometer.

2.2.7 Protein-protein cross-linking

The protein-protein cross-linking experiments were performed for the investigation of protein-protein interactions in Type I-B Cascade complex from *C. thermocellum*. The complex was incubated at 50 °C for 10 min prior to any *in vitro* experiments.

2.2.7.1 Determination of optimal cross-linker to protein ratio

The optimal cross-linker to protein ratio was determined by incubating 30 pmol aliquots of the purified Cascade complex with freshly prepared BS3 cross-linker in a series of molar excesses of 5, 10, 25, 50, 100 and 200 as well as a control. Samples were allowed to react with the cross-linker for 30 min at room temperature. The reaction was quenched by adding 2 µl of 2 M Tris-Cl

pH 7.9 to the reaction mix. The cross-linked samples were analyzed by SDS page on a 4-12% Bis-Tris gel (Invitrogen) using the manufacturer's protocol and stained with coomassie (Figure 3.15). From the SDS-PAGE analysis the cross-linker to protein ratio of 75:1 was considered optimal for carrying out further experiments.

2.2.7.2 Protein-protein cross-linking of Cascade complex

For MS analysis, 150 pmol Cascade complex was cross-linked with freshly prepared BS3 in a cross-linker to protein ratio 75:1, for 30 min at room temperature. The reaction was quenched by adding 2 µl of 2 M Tris-Cl pH 7.9 to the reaction mix. The reaction volume was reduced to 2-5 µl by drying in a SpeedVac.

Further, the cross-linked sample was digested in-solution. The dried sample was dissolved in 20 µl 8 M Urea (in water), with vigorous shaking for 30 min at 25 °C. The disulfide bridges in proteins were reduced with the addition of 20 µl 10 mM DTT (in water), followed by incubation for 30 min at 600 rpm in a thermomixer at 25 °C. Further the–SH groups were alkylated with addition of 20 µl 60 mM IAA (in water), followed by incubation as in the previous step. The final volume of reaction mixture was raised to 200 µL with water and modified trypsin was added in a 1:20 w/w ratio for overnight hydrolysis.

2.2.7.3 Enrichment of cross-linked peptides and LC-MS

The cross-linked peptides were enriched as previously reported in [136]. The peptide mixture was desalted with C18 chromatography as described under the Section 2.2.4.3. Peptides were reconstituted in 30% ACN, 0.1% TFA and injected onto a Superdex Peptide column and eluted at 50 µl/min collecting fractions of 50 µl. These fractions were dried in a SpeedVac and reconstituted in 20 µl LC-MS sample loading buffer of which 8 µl were injected for a single LC-MS run. MS analysis was performed using Orbitrap Fusion mass spectrometer.

2.2.8 LC-ESI-MS/MS

The LC-ESI-MS/MS analysis was carried out using nano-liquid chromatography (nano-LC) system directly coupled to the electrospray (ESI) source of a mass spectrometer. Four different mass spectrometers were used in this thesis.

The LTQ Orbitrap XL and LTQ Orbitrap Velos instruments (Thermo Fisher Scientific) coupled to an Agilent LC-system (Agilent 1100 series) and the Orbitrap Fusion and Q Exactive HF instruments (Thermo Fisher Scientific) coupled to Thermo Fisher Scientific LC-system (Dionex Ultimate 3000, UHPLC). Details for the LC separation and MS analysis are described below.

2.2.8.1 Nano-LC separation

Nano-LC separation (Agilent 1100 series, Agilent Technologies)

The samples were injected onto a nano-LC system including a C18 trapping column (length ~2 cm, inner diameter 150 µm) in-line with a C18 analytical column (length ~15 cm, inner diameter 75 µm). Both packed in-house by Uwe Pleßmann using C18 AQ, 120 Å, 5 µm (Dr. Maisch GmbH). Analytes were loaded on the trapping column at a flow rate of 10 µL/min in buffer A (0.1% v/v FA) and subsequently eluted and separated on the analytical column with a gradient of 7–38% buffer B (95% v/v acetonitrile, 0.1% v/v FA) for 33 min in a 50 min gradient, followed by a column wash with 90% buffer B at a flow rate of 300 nl/min.

UHPLC separation (Dionex, Ultimate 3000, Thermo Fisher Scientific)

The samples were injected onto a nano-liquid chromatography system including a C18 trapping column (length ~2 cm, inner diameter 150 µm) in-line with a C18 analytical column (length ~30 cm, inner diameter 75 µm), both packed in-house by Uwe Pleßmann. The trapping column was packed as above, however the analytical column was packed using C18 AQ 120 Å 1.9 µm (Dr. Maisch GmbH). Analytes were loaded on the trapping column at a flow rate of 10 µl/min in buffer A (0.1% v/v FA) and subsequently eluted and separated on the analytical column with a gradient of 8–46% buffer B (80% v/v acetonitrile, 0.08% v/v FA) with an elution time of 45 min in a 50 min gradient or 75 min in a 90 min gradient, followed by a column wash with 90% buffer B at a flow rate of 300 nl/min.

2.2.8.2 ESI-MS/MS

LTQ Orbitrap XL (Thermo Fisher Scientific)

The instrument was operated in data-dependent mode using a TOP8 method. MS scans were recorded in the Orbitrap (m/z range 350-1600) with a resolution of 30,000 at 400 m/z and automatic gain control (AGC) target at 10^6. For subsequent MS/MS, top 8 most intense ions were selected. Fragment ions were generated in the ion trap by CID activation (collision induced dissociation, normalized collision energy=35). In order to avoid re-fragmentation, the dynamic exclusion was set to 60 s.

LTQ Orbitrap Velos (Thermo Fisher Scientific)

The instrument was operated in data-dependent mode using a TOP10 method. MS scans were recorded in the Orbitrap (m/z range 350-1600) with a resolution of 30,000 at 400 m/z and AGC target 10^6. For subsequent MS/MS, top 10 most intense ions were selected. Both precursor ions as well as fragment ions were scanned in the Orbitrap. Fragment ions were generated by HCD activation (higher energy collision dissociation, normalized collision energy=40). In order to

avoid re-fragmentation, the dynamic exclusion was set to 60 s. The MS/MS fragment spectra were recorded with a first fixed mass of m/z=100 and a resolution of 7500.

Q Exactive HF (Thermo Fisher Scientific)

The instrument was operated in data-dependent mode using a TOP20 method. MS scans were recorded in the Orbitrap (m/z range 350-1600) with a resolution of 60,000 and AGC target 10^6. For subsequent MS/MS, top 20 most intense ions were selected. Both precursor ions as well as fragment ions were scanned in the Orbitrap. Fragment ions were generated by HCD activation (higher energy collision dissociation, normalized collision energy=35). In order to avoid re-fragmentation, the dynamic exclusion was set to 30 s. The MS/MS fragment spectra were recorded with a first fixed mass of m/z=110 and a resolution of 15000, AGC target 10^5.

Orbitrap Fusion (Thermo Fisher Scientific)

The instrument was operated in data-dependent mode using a TOP20 method. MS scans were recorded in the Orbitrap (m/z range 350-1600) with a resolution of 120,000 and AGC target $5x10^5$. For subsequent MS/MS, top 20 most intense ions were selected. Both precursor ions as well as fragment ions were scanned in the Orbitrap. Fragment ions were generated by HCD activation (higher energy collision dissociation, normalized collision energy=30). In order to avoid re-fragmentation, the dynamic exclusion was set to 10 s. The MS/MS fragment spectra were recorded with a first fixed mass of m/z=110 and a resolution of 30000, AGC target $5x10^4$.

2.2.9 Data analysis

2.2.9.1 Quantitative proteome analysis after dimethyl labeling using MaxQuant

Raw MS data were analyzed using MaxQuant software v1.2.2.5 incorporated with Andromeda [90, 137]. The following settings were used as default settings: MS/MS tolerance 0.5 Da, FDR at both peptide and protein level 1 %, maximum peptide posterior error probability (PEP) 1.0, minimum peptide length 6 amino acids, minimum ratio count 2, maximum number of modifications per peptide 4, maximum precursor charge 5, “re-quantify” True, “keep low-scoring versions of identified peptides” False, “use razor and unique peptides” True. Carbamidomethylation of cysteine was used as fixed modification and oxidation of methionine and acetylation of N-terminal of protein were used as variable modifications. Trypsin specificity with no proline restriction and up to 2 missed cleavages were allowed. Specifically for the dimethyl labeling analysis: Multiplicity was set to 2, with maximum 3 labeled amino acids per peptide, DimethLys0 and DimethNter0 were used as light labels and DimethLys4 and DimethNter4 were used as heavy labels. The search was performed using *H. volcanii* protein database from UniPROT.

The output from MaxQuant (proteinGroups.txt) was imported in Perseus. All "Reverse" and "Contaminant" entries were deleted. P-value (Significance B) was calculated and set as a main criterion for the data interpretation. Total summed peptide intensities were plotted in $\log_{10}$ scale, normalized enrichment ratios in $\log_2$ scale as scatterplots using R [138] (R scripts were previously published in [139].

2.2.9.2 Quantitative proteome analysis after iBAQ using MaxQuant

Results were analyzed using the MaxQuant software v1.2.7.4, using the default settings as above. The multiplicity for labels was set to 1 as this was a label free analysis and the iBAQ function was set to true. The MS data was matched against the *H. volcanii* protein database from UniPROT supplemented with the sequences of the 48 proteins contained in the UPS2 standard. The iBAQ values from three replicates were averaged and the standard deviation was calculated to judge the precision of analysis. The UPS2 standard proteins observed in all the three replicate analyses were used for calibration. A calibration curve was obtained by linear regression from a double logarithmic plot (log(iBAQ) vs. log(amount). The calibration function was then used to calculate the amount of different Cas proteins in the respective samples.

2.2.9.3 Identification of protein-RNA cross-links with RNP^{xl}

The MS .raw files were converted into the .mzML format with msconvert, part of the ProteoWizard software bundle [140] or with Proteome discoverer 1.10 software provided by Thermo Fischer Scientific (http://www.thermoscientific.com/en/product/proteome-discoverer-software.html). Protein-RNA cross-links were analyzed using RNP^{xl} tool [108] implemented in OpenMS [115, 116] and using OMSSA [117] as search engine. Data analysis workflows were assembled especially for our laboratory by Timo Sachsenberg (Prof. Oliver Kohlbacher, University of Tübingen, Tübingen). The high scoring cross-linked peptides were manually annotated for confirmation.

The cross-linked regions/residues identified were mapped on the crystal structures (where available) using PyMOL (an open source software maintained and distributed by Schrödinger (http://www.pymol.org), structure modelling was performed using Phyre2 [141] and superposition was performed using Secondary structure matching (SSM) in COOT [142].

Online tools used for the calculation of monoisotopic masses of peptides, RNA oligonucleotides and their fragments:

ProteinProspector v5.14.0 – University of California, San Francisco:

http://prospector.ucsf.edu/prospector/mshome.htm

Peptide Mass Calculator v3.2 – Immunology Division, University of Utrecht, NL:

http://immweb.vet.uu.nl/P&P_fac/pepcalc.htm

Mongo Oligo Mass Calculator v2.06 – University at Albany, State University of New York: http://mods.rna.albany.edu/masspec/Mongo-Oligo

2.2.9.4 Identification of protein-protein cross-links with pLink

The protein-protein cross-links were identified with pLink using the data analysis workflow described in their publications and manuals [123]. Briefly, the .raw files from the MS instruments were converted to the .mgf format using MSConvertGUI and submitted to database search with standard parameters. Oxidation of methionine was selected as a variable modification whereas carbamidomethylation of cysteine was selected as a fixed modification. Spectra were searched against a database containing the UniPROT sequences of the protein complex components. False discovery rate was set to a maximum of 1%.

3. Results

The MS investigation of prokaryotic immune defense system in the course of this work was based on two key aspects of proteomics.

- Quantitative proteomics - For comparison between the proteomes of wild-type strains with the strains that carry deletion mutants of Cas proteins and for the absolute quantification of Cas proteins in the multi-subunit CRISPR-Cas complexes to determine their stoichiometry.
- Structural proteomics - For investigating the protein-RNA and protein-protein interactions within various CRISPR complexes.

3.1 Quantitative MS investigations in the CRISPR-Cas system

Quantitative proteomics have numerous applications as described under Section 1.2.5. Here I present the results from protein quantification using differential stable-isotope labeling (dimethyl labeling) and label-free absolute quantification (iBAQ) carried out in *Haloferax volcanii* comprising Type I-B CRISPR-Cas system. Both quantitative analyses were performed in collaboration with the group of Prof. Anita Marchfelder, Ulm Univeristy, Ulm.

3.1.1 Quantitative proteome analysis of *H. volcanii WT* and *Δcas7 KO* using dimethyl labeling

The Cas7 proteins are one of the key proteins of the CRISPR ribonucleoprotein complexes as they constitute the core of these complexes in the form of a helical backbone [22]. The goal of this project was to investigate the effects of deletion of *cas7* gene in *H. volcanii* on the expression of other Cas proteins and at the proteome level, the proteomes of wild-type strain (H119 WT) and *cas7* deletion strain *(Δcas7* KO, where *cas7* gene was knocked out) were compared using a dimethyl labeling strategy. The protein extracts for both strains were prepared by Britta Stoll in Prof. Anita Marchfelder's Lab in Ulm (Section 2.2.1.1). For relative quantification, the peptides from the two samples were chemically labeled using 'Light' and 'Medium' dimethyl labeling reagent [84]. Although dimethyl labeling is a triplex labeling approach with a provision of 'Heavy' label in addition to the 'Light' and 'Medium' [84], in this experiment only two labels were used for the comparison between two samples (Section 2.2.4). After this point the 'Medium' labeled sample is referred to as 'Heavy' and all the comparisons

thereafter are attributed as between a 'Light' labeled and a 'Heavy' labeled sample. Two experimental workflows were designed for quantitative analysis, the "Forward" and "Reverse" experiments as shown in Figure 3.1.

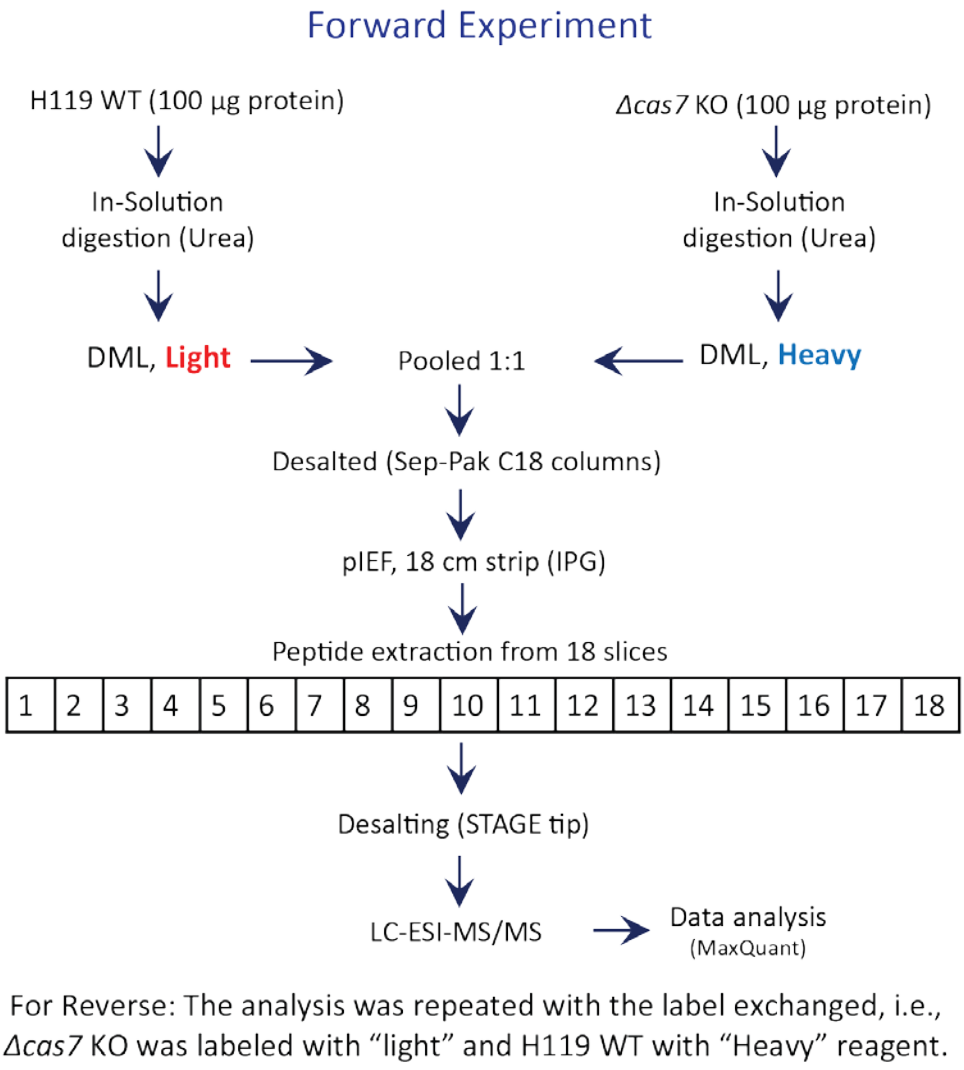


Figure 3.1 Workflow for the H119 WT vs *Δcas7* KO, quantitative analysis: Forward Experiment.

For the wild-type and cas7 deletion mutant, the proteins were digested in solution using trypsin. The peptides were isotopically labeled with dimethyl labeling (DML) reagents and pooled in a 1:1 ratio. Excess of salts and the unused labeling reagents were removed with desalting using C18 columns. Peptides were separated using peptide iso-electric focusing and the peptide fractions were further desalted using STAGE-tips, followed by MS analysis. For the Reverse experiment only the labeling reagents were exchanged at the starting point and the same procedure was followed thereafter.

For quantitative proteome analysis ~100 µg dried acetone precipitated protein extracts from both H119 WT and *Δcas7* KO were dissolved in 8M Urea followed by in-solution digestion using

trypsin (in presence of 1 M urea). The labeled peptides derived from the two samples, were pooled in 1:1 ratio and the complex peptide mixture was then separated using pIEF (Section 2.2.4). This results in focusing of peptides into very sharp regions of pH gradient on an IPG gel strip, based on their iso-electric points [143].

Pooling the peptide mix from two differentially labeled samples before enrichment ensures that all the processing steps such as fractionation and MS analysis are performed simultaneously for all the peptides from both the samples thereby avoiding the introduction of undesired variability. The intensities of both 'light' and 'heavy' labeled versions of a particular peptide derived from a protein will be measured at the same time. Therefore the ratio of the signal from the 'heavy peptide' and the signal from the 'light peptide' can be used to derive relative amount of differences in the proteins from the wild-type and the *cas7* deletion strains. High (>1) H/L (heavy/light) peptide ratios in the forward experiment would indicate that the corresponding proteins are up-regulated or more abundant upon *cas7* deletion with respect to the wild-type. Conversely, peptides with low (<1) H/L ratio would indicate down-regulation or less abundance upon *cas7* deletion relative to the wild-type. This can be further confirmed with a reverse experiment. The results of relative quantification are shown in Figure 3.2.

In both forward and reverse experiments approximately 1800 proteins (after removing contaminants such as keratin) were identified and quantified. For relative quantification the $\log_2$ normalized ratio (H/L) was determined for different proteins, using MaxQuant software (Section 2.2.9.2). For the forward experiment this ratio would be calculated from the intensities of 'heavy'-labeled *Δcas7* KO proteins divided by intensities of 'light'-dimethyl-labeled H119 WT proteins and for the reverse experiment this ratio would be intensities of 'heavy'-dimethyl-labeled H119 WT proteins divided by intensities of 'light'-labeled *Δcas7* KO proteins.

In both the experiments, majority of the proteins (approximately 90% of the total number of identified proteins) presented a $\log_2$ normalized ratio (H/L) close to zero, i.e. a ratio of heavy/light close to 1 (as indicated in the Figure 3.2 A and B based on the MaxQuant significance B, p-value >0.05). This result is expected, as the major part of the proteome for both wild-type and deletion strain should remain unchanged.

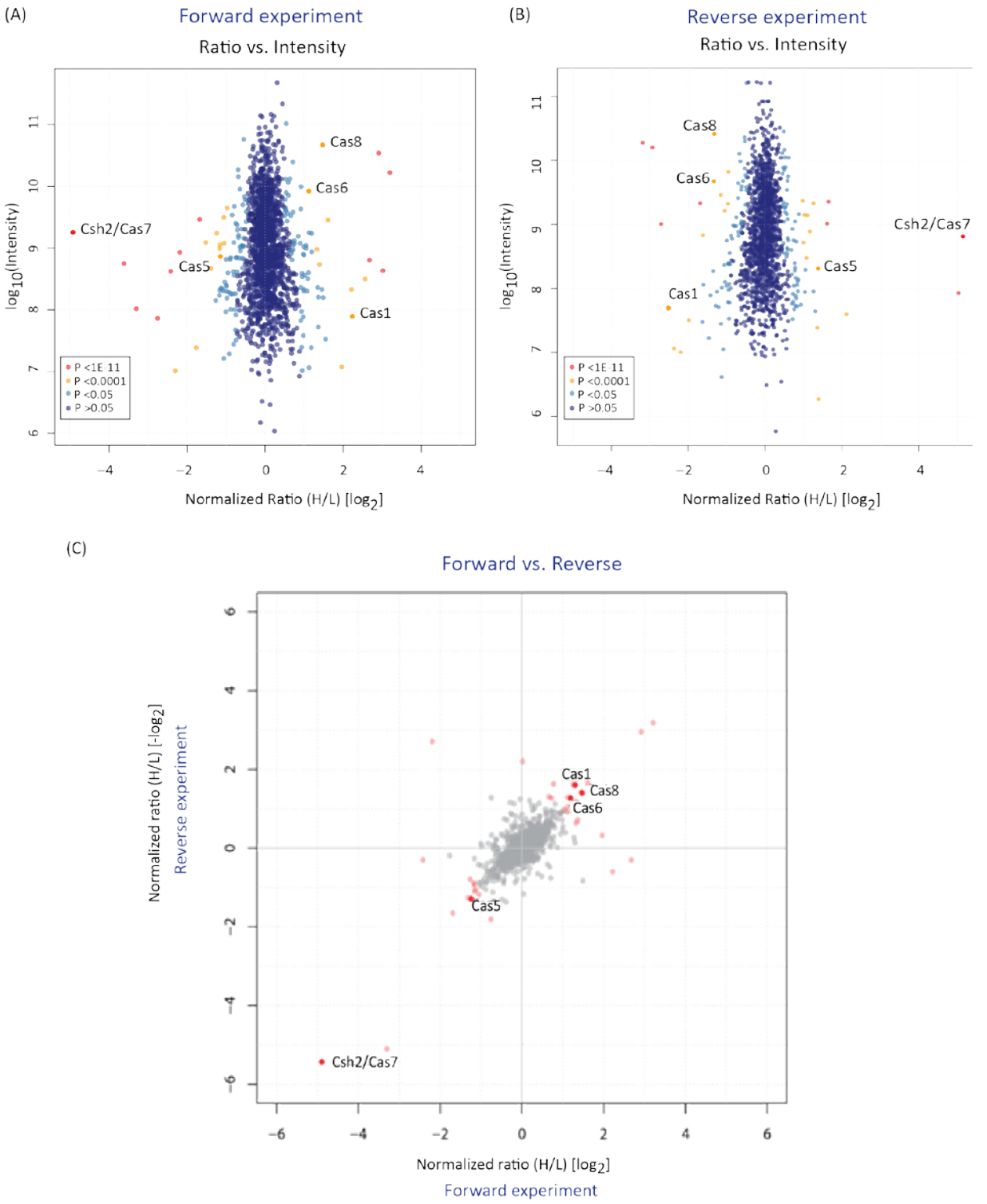


Figure 3.2 Scatter-plot analysis of protein quantification in H119 WT and *Δcas7* KO mutants.

(A) Forward experiment and (B) Reverse experiment; the $\log_2$ normalized ratios (H/L) of the proteins identified are plotted against the $\log_{10}$ of the Intensity. Significantly up or downregulated proteins are colored according to their corresponding p-value as shown in the legend. (C) Forward vs. Reverse experiment; the $\log_2$ ratios of the "Forward experiment" are plotted against the $-\log_2$ ratios of the "Reverse experiment". The up-regulated proteins appear in upper-right quadrant and the down-regulated proteins appear in the lower-left quadrant. All proteins showing an average ratio higher than 1 ($\log_2$ scale) are indicated in light-red. The Cas proteins identified in this experiment are shown in red.

For the forward experiment, from all the quantified proteins (~1800), approximately 155 proteins were significantly regulated i.e., either up- or down- based on the MaxQuant significance B ($p < 0.05$) (Figure 3.2 A) and for the reverse experiment approximately 150 proteins were significantly regulated (Figure 3.2 B). In order to increase the confidence of the proteins with significant difference in the wild-type and deletion strain and to restrict the list of potential proteins presenting biological significance, the $\log_2$ normalized ratio (H/L) for the forward and reverse experiment were plotted against each other (Figure 3.2 C). The proteins which were significantly regulated in both the experiments are summarized in Table 3.1 for the down-regulated proteins and Table-3.2 for the up-regulated proteins. These include mainly certain cytosolic proteins and proteins belonging to ABC-transport system in addition to the Cas proteins. Two parameters considered for summarizing this list of significant proteins included the Ratio H/L normalized and the PEP value. The latter being the posterior error probability (PEP) of the identification. The PEP value essentially operates as a p-value, where smaller is more significant.

Table 3.1 Proteins "Down-regulated" upon *cas7* deletion, significant in both forward and reverse experiments. The proteins of interest are shaded in orange.

		Forward Experiment		Reverse Experiment	
UniPROT ID	Protein Name	Ratio H/L Normalized	PEP value	Ratio H/L Normalized	PEP value
D4GQN6	CRISPR-associated protein, Csh2 family (Cas7)	0.03	2.92E-126	42.92	6.25E-18
D4GYK7	IMP cyclohydrolase	0.10	0.027485	34.37	0.007626
D4GPI4	Aspartate racemase	0.15	0.028474	2.06	2.77E-09
D4GWI7	Homoserine kinase	0.29	9.76E-17	1.74	2.21E-28
D4GPP7	Short-chain family oxidoreductase	0.31	3.57E-173	3.14	3.60E-107
D4GQN7	CRISPR-associated protein Cas5	0.38	6.12E-128	2.60	2.59E-18
D4GPE5	ABC-type transport system periplasmic substrate- binding protein (Probable substrate iron-III)	0.41	6.95E-16	1.74	8.83E-07
D4GW66	Coenzyme PQQ synthesis protein E homolog	0.42	4.51E-141	2.40	9.79E-88
D4GVN2	Trk potassium uptake system protein	0.44	8.77E-46	1.66	1.14E-33
D4GSJ2	OsmC-like protein superfamily	0.48	5.45E-61	2.08	0.000249
D4GW62	Putative uncharacterized protein	0.48	1.75E-102	2.03	1.18E-78
D4GW08	Translation initiation factor aIF-2B delta subunit	0.48	9.88E-92	1.73	9.02E-107
D4GS83	Thioredoxin reductase	0.49	4.95E-67	2.25	1.24E-36
D4GYN6	Ornithine carbamoyltransferase	0.50	4.23E-32	2.12	0.000182
D4GYV1	Flavoprotein reductase homolog	0.51	2.51E-26	1.81	1.03E-39

Table 3.2 Proteins "Up-regulated" upon *cas7* deletion, significant in both forward and reverse experiments. The proteins of interest are shaded in orange.

		Forward Experiment		Reverse Experiment	
UniPROT ID	Protein Name	Ratio H/L Normalized	PEP value	Ratio H/L Normalized	PEP value
D4GWP5	ABC-type transport system ATP-binding protein (Probable substrate zinc)	9.20	0	0.11	0
D4GWP4	ABC-type transport system periplasmic substrate-binding protein (Probable substrate zinc)	7.55	0	0.13	0
D4GQP0	CRISPR-associated protein Cas1	4.70	5.58E-10	0.33	1.34E-20
D4GPV2	Putative uncharacterized protein	3.06	3.07E-71	0.31	9.07E-95
D4GQN5	CRISPR-associated protein Cas8	2.79	0	0.37	0
D4GZR8	ATP-dependent DNA helicase	2.75	8.26E-23	0.37	8.92E-13
D4GZR4	Cupin superfamily	2.51	9.20E-42	0.33	7.39E-81
D4GZR3	Archaea-specific helicase AshA	2.29	5.39E-46	0.39	2.20E-38
D4GWP3	Glutaredoxin-like protein	2.29	6.68E-40	0.45	1.05E-37
D4GZR5	SpoIVFB-type metallopeptidase, transmembrane (TBD)	2.27	2.07E-12	0.22	1.48E-08
D4GTW9	Predicted protein, putative	2.18	1.10E-94	0.48	8.05E-119
D4GQN4	CRISPR-associated protein Cas6	2.17	1.04E-259	0.40	1.48E-225
D4GPA4	Glycosyl Hydrolase Family 88 superfamily	2.17	4.61E-72	0.54	5.61E-72
D4GRI4	Acetyl-CoA C-acyltransferase	2.07	2.00E-182	0.50	8.65E-253
D4GPX0	Oxidoreductase	1.93	1.70E-24	0.49	6.38E-08
D4GQB5	Rieske-type [2Fe-2S] iron-sulfur protein	1.73	4.24E-103	0.42	2.36E-11

The proteins of interest i.e., the Cas proteins could be identified and quantified. As expected, Cas7 was observed as the most significantly down-regulated protein. Although the *cas7* gene was deleted, the ratio of Cas7 can be determined against the background noise and hence it was observed as the most distant point in all the three scatterplots in Figure 3.2. Other Cas proteins whose expression was affected upon *cas7* deletion include Cas5 protein, which was down-regulated or weakly expressed in the *cas7* KO strain and Cas1, Cas6 and Cas8 which were up-regulated i.e., more abundantly expressed in the *cas7* KO strain. This observation suggests that the deletion of a single gene (*cas7*) within the CRISPR locus, in a single operon, strongly affects the expression of other genes within the same operon.

Another significant protein belonging to CRISPR-Cas system which was identified in this experiment was the archaeal-specific helicase (AshA), which was observed to be up-regulated

upon Cas7 deletion (Table 3.2). In addition, this protein has been shown to play a significant role in CRISPR interference step (Marchfelder Lab, unpublished data).

3.1.2 Determination of stoichiometry of Cas5:Cas6:Cas7 in *H. volcanii* with iBAQ

The *H. volcanii* Type I-B system contains a Cascade-like complex comprising Cas7, Cas5 and Cas6 that is required for the biogenesis and stability of crRNA. It has been shown that when the whole *cas* gene cluster (*cas1-8*) is removed, the organism is not able to produce and stably maintain mature crRNAs. The crRNA production and stability can be rescued only if *cas5, cas6* and *cas7* are present. The Cas6 protein is necessary for crRNA generation but is not sufficient for its maintenance and Cas5 and Cas7 are additionally required for crRNA generation or stabilization [89].

To confirm the association of these three Cas proteins, a FLAG-Cas7 fusion protein was expressed and purified together with all potential interaction partners (Section 2.2.1.2). The FLAG-purified fraction when analyzed with SDS-PAGE showed co-purification of two additional proteins when analyzed with SDS-PAGE. With western blot analysis using anti-FLAG antibody, the largest band was confirmed to be Cas7 protein (Cas7-FLAG fusion protein) (Figure 3.3 A and B). These experiments were performed by Jutta Brendel in Prof. Anita Marchfelder's Lab in Ulm. The two additional proteins in SDS-PAGE were identified as Cas5 and Cas6 with MS analysis.

Furthermore, the stoichiometry of Cas5, Cas6 and Cas7 proteins in the complex was determined using a label-free absolute quantification approach (iBAQ). A mixture of quantified standard proteins (UPS2) was spiked into the complex of three Cas proteins isolated in a co-purification (Section 2.2.5). UPS2 is a universal protein standard comprising 48 different human proteins of various molecular weight and abundances that span a concentration range of five orders of magnitude and is used to generate a calibration curve. Together this mixture of proteins is digested in-solution with trypsin followed by MS analysis, as described in the workflow in Figure 3.3 C. The iBAQ intensity for a protein is measured by summing the peak intensities of all detected peptides for the protein dividing it by the number of theoretically observable peptides. iBAQ intensities were determined using MaxQuant software.

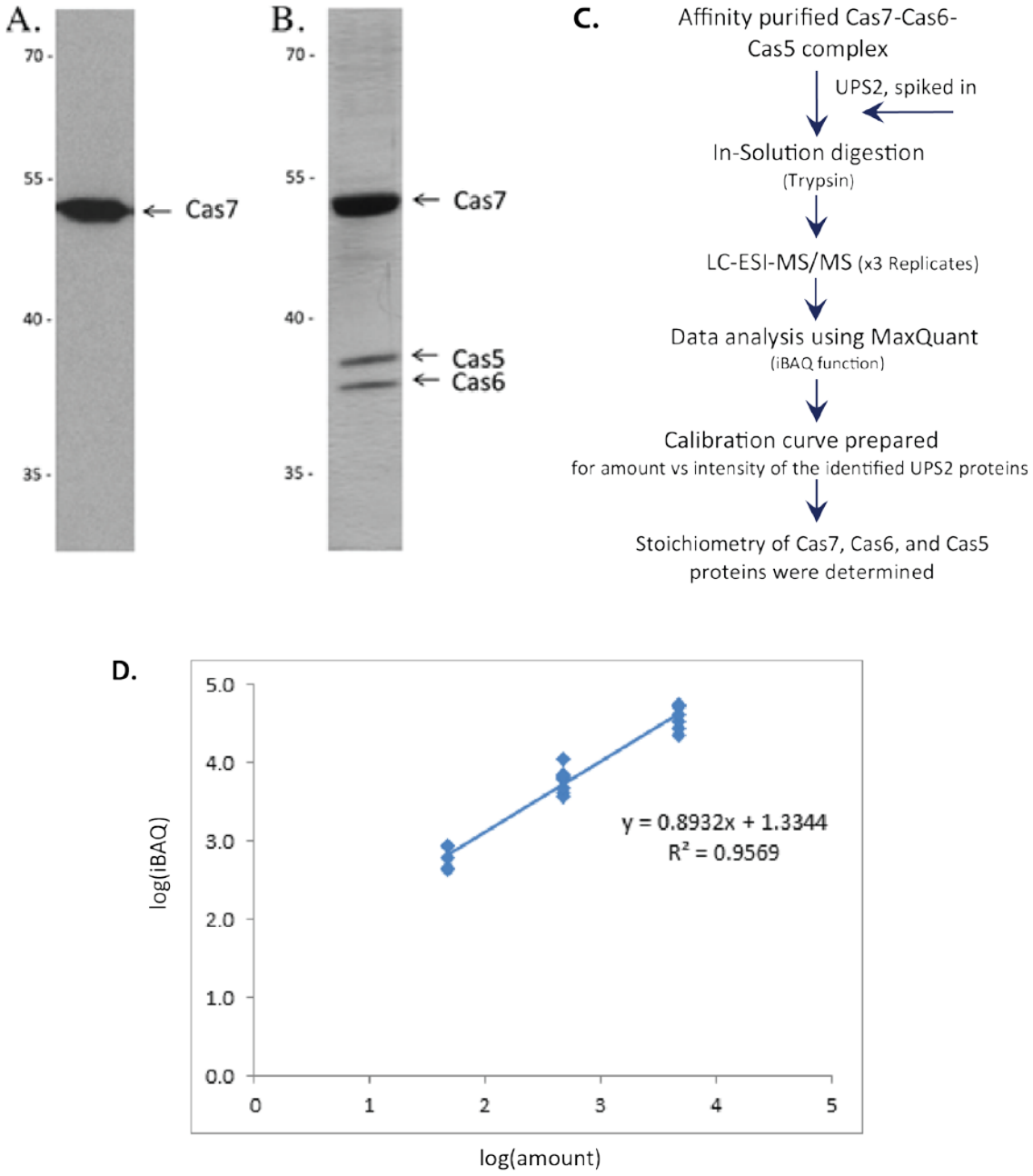


Figure 3.3 Determination of stoichiometry of Cas5:Cas6:Cas7 in *H. volcanii* with iBAQ

A FLAG-tagged Cas7 protein was expressed in *Haloferax* cells and purified together with all potential interaction partners using the FLAG-tag. (A) Western blot analysis of the FLAG-purified fraction, probed with an anti-FLAG antibody. (B) SDS-PAGE analysis of the FLAG-purified fraction, showing three proteins visible on the Silver-stained gel. According to the western blot the largest band corresponds to Cas7 protein (Cas7-FLAG fusion protein probed with FLAG antibody). According to the MS analysis the two smaller proteins are Cas5 and Cas6 respectively. The protein size marker depicted on the left in kDa. Figure originally reported in [89] and reproduced with permission. (C). Overview of the iBAQ analysis workflow. (D). iBAQ calibration curve of UPS2 proteins used in determining the stoichiometry of Cas5, Cas6 and Cas7 from *H. volcanii.* Double logarithmic plot of log(iBAQ) *vs* log(amount) using 20 out of 48 UPS2 proteins observed in all three technical replicates for the absolute quantification (Table 3.3)

Table 3.3 iBAQ quantitative mass spectrometry analysis of Cas7 co-purification to determine the absolute amounts of Cas5, Cas6 and Cas7 proteins.

Protein	Source	Protein amount (fmol)	iBAQ	IBAQ StdDev	log (amount)	log (iBAQ)	Ratio
>P12081ups| SYHC_HUMAN	UPS2 Standard	47	426	13	1.674	2.630	
>P16083ups| NQO2_HUMAN	UPS2 Standard	47	457	83	1.674	2.660	
>P61626ups| LYSC_HUMAN	UPS2 Standard	47	858	85	1.674	2.934	
>P06732ups| KCRM_HUMAN	UPS2 Standard	47	862	133	1.674	2.936	
>P02753ups| RETBP_HUMAN	UPS2 Standard	47	617	275	1.674	2.790	
>P63165ups| SUMO1_HUMAN	UPS2 Standard	472	4074	254	2.674	3.610	
>P02144ups| MYG_HUMAN	UPS2 Standard	472	3745	345	2.674	3.573	
>P15559ups| NQO1_HUMAN	UPS2 Standard	472	4791	558	2.674	3.680	
>P01133ups| EGF_HUMAN	UPS2 Standard	472	6039	395	2.674	3.781	
>P62937ups| PPIA_HUMAN	UPS2 Standard	472	6963	854	2.674	3.843	
>Q06830ups| PRDX1_HUMAN	UPS2 Standard	472	6475	170	2.674	3.811	
>P04040ups| CATA_HUMAN	UPS2 Standard	472	6633	463	2.674	3.822	
>P00167ups| CYB5_HUMAN	UPS2 Standard	472	11428	1643	2.674	4.058	
>P69905ups| HBA_HUMAN	UPS2 Standard	4717	22246	1469	3.674	4.347	
>P68871ups| HBB_HUMAN	UPS2 Standard	4717	27773	4132	3.674	4.444	
>P62988ups| UBIQ_HUMAN	UPS2 Standard	4717	33625	3406	3.674	4.527	
>P00918ups| CAH2_HUMAN	UPS2 Standard	4717	40906	6958	3.674	4.612	
>P41159ups| LEP_HUMAN	UPS2 Standard	4717	52631	1567	3.674	4.721	
>P00915ups| CAH1_HUMAN	UPS2 Standard	4717	54385	6349	3.674	4.735	
>P01031ups|CO5_HUMAN	UPS2 Standard	4717	53181	5761	3.674	4.726	
H. volcanii **Cas6**	FLAG-purification	249	1259	353	2.396	3.100	**1.0**
H. volcanii **Cas5**	FLAG-purification	419	2004	228	2.622	3.302	**1.7**
H. volcanii **Cas7**	FLAG-purification	2119	8522	362	3.326	3.931	**8.5**

Originally published in [89], reproduced with permission from the publisher.

The proteins that were confidently identified and quantified from three technical replicates are listed under Table 3.3. iBAQ values for the standard proteins, from the three replicates, were averaged. The known amounts of the standard proteins in UPS2 and the average iBAQ values were used to prepare a calibration curve. A double logarithmic plot where the log (amount) of the standard proteins identified was plotted against the log (iBAQ value). The concentration of Cas5, Cas6 and Cas7 proteins were calculated using the log/log linear regression (Appendix, Figure 3.3 D).

The derived protein concentrations indicate a Cas5:Cas6:Cas7 stoichiometry of 1.7: 1: 8.5 (Table 3.3). This low:low:high type stoichiometry is in agreement with previously observed stoichiometry for Cascade-type protein complexes in *E. coli* [30] and *P. aeruginosa* [40].

3.2 UV induced protein-RNA cross-linking for investigation of protein-RNA interactions in the CRISPR-Cas systems

As a part of the structural proteomics based investigations into the CRISPR-Cas system, the UV induced protein-RNA cross-linking strategy was used for identifying the protein-RNA contact sites in the CRISPR ribonucleoprotein (crRNP) complexes [108]. It is a straightforward approach for the identification of not only the proteins that cross-link to RNA but also to unambiguous identification of the cross-linked peptide or amino-acid and the cross-linked nucleotide(s) [108]. The method is applicable to single (e.g., recombinant) Cas proteins that interact with crRNAs but could not be co-crystallized in complex with the crRNA. Furthermore, it can be also applied to assembled crRNPs of varying complexity, obtained either by reconstitution or by purification from extracts. In this section, I present the results obtained during the investigation of protein-RNA interactions in both fully assembled multi-subunit crRNP complexes as well as single Cas protein-crRNA complexes. These projects were a part of extensive collaboration with the members of DFG Forschergruppe 1680 and the groups of Prof. John van der Oost and Stan Brouns (Wageningen University, Wageningen, NL).

3.2.1 Protein-RNA cross-linking in Cas6b proteins from *M. maripaludis* and *C. thermocellum* with their cognate crRNA

The CRISPR-Cas system presents a broad diversity with different types and subtypes, the general aspects and the mechanism of three major CRISPR types are however very similar. The Type I CRISPR systems have a similar mechanism for the processing of the transcribed CRISPR locus to yield mature crRNA. All the Type I subtypes use a Cas6 homolog to process pre-crRNA yielding a mature form of interfering RNA. The Cas6 endoribonucleases share some basic features, e.g., a ferredoxin-like fold in the structure, metal independent processing of pre-crRNA and always yielding mature crRNA with a 5’ terminal repeat tag of 8 nucleotides [18, 144-146]. Recently, novel Cas6 enzymes have been identified from the archaeal and bacterial model organisms *M. maripaludis* and *C. thermocellum* [127] referred to as “Cas6b” (corresponding to the subtype I-B). The crRNA processing for both these Cas6 enzymes was also characterized with the identification of individual spacer sequences [127].

As a follow up of this investigation, UV induced protein-RNA cross-linking was used to identify the RNA interaction sites in the bacterial and archaeal Cas6 proteins. The experiments were

performed in collaboration with Hagen Richter and Judith Zöphel from the group of Prof. Lennart Randau (MPI Terrestrial Microbiology, Marburg), who expressed and purified the Cas6b proteins from *M. maripaludis C5* (*Mm* Cas6b) and *C. thermocellum* 3205 (*Ct* Cas6b). The cognate crRNA for these proteins with a substitution of the first unprocessed nucleotide against a deoxynucleotide were synthesized by Eurofins MWG Operon.

M. maripaludis Cas6b cognate crRNA: 5'-CUAAAAGAAUAACUUGCAAAAUAACAAG**(dC)**AUUGAAAC-3'

C. thermocellum Cas6b cognate crRNA: 5'-GUUGAAGUGGUACUUCCAGUAAAACAAG**(dG)**AUUGAAAC-3'

One deoxy-nucleotide is introduced in the sequence to allow the Cas6b protein and the crRNA to form a complex when incubated together under optimal conditions as well as to prevent the Cas6b (an endoribonuclease) from hydrolyzing the crRNA.

3.2.1.1 Analysis of Cas6b-crRNA cross-linking with SDS-PAGE

The UV induced cross-linking between the Cas6b proteins and their cognate crRNAs was first analyzed using SDS-PAGE. Since the MS based approach requires a comparatively higher sample amount owing to the low cross-linking yield upon UV induction, it becomes essential to first check the cross-linking efficiency on a gel before proceeding with the MS analysis. The crRNA was labeled at the 5' end with ϒ-^{32}P-ATP and purified using PCI extraction (Section 2.2.2). The labeled crRNA was incubated with Cas6b protein (50 times molar excess of the protein) to form a binary complex of Cas6b-crRNA which was UV irradiated at 254 nm (Section 2.2.6.1). The cross-linked complex and a non-cross-linked control which also comprised Cas6b protein and crRNA (^{32}P labeled) complex but with no UV treatment, were analyzed using SDS-PAGE. The results of SDS-PAGE analysis were visualized with both coomassie staining and autoradiography as shown in the Figure 3.4.

The SDS-PAGE analysis clearly showed cross-linking between *Mm* Cas6b and the cognate crRNA, presenting a single band at 30 kDa on the coomassie stained gel. In the autoradiograph the ^{32}P signal was observed at approximately the same molecular weight in the cross-linked sample and no signal was observed in the non-cross-linked sample. In the analysis of the *Ct* Cas6b there was a strong ^{32}P signal present in the cross-linked sample which was absent in the control. Nonetheless, the *Ct* Cas6b protein purification was not sufficient pure as evident from multiple bands on the coomassie stained gel. From the presence of the ^{32}P signal in the autoradiograph it can be speculated that UV induced cross-linking might have resulted in higher-order complex

formation with Cas6b and some contaminants in the sample. These results are sufficient to confirm that cross-linking between Cas6 proteins and crRNA took place and the interaction can be further analyzed with MS.

3.2.1.1 Analysis of Cas6b-crRNA cross-linking with mass spectrometry

UV cross-linking in combination with MS was performed to assess which regions of the Cas6b proteins are involved in direct interactions with the crRNA. The purified Cas6b proteins were reconstituted *in vitro* with their cognate crRNAs (in 1:1 ratio, one nmol each) to form the binary Cas6b-crRNA complex. The complex was UV irradiated as described using the standard protocol (Section 2.2.6.2). LC-MS analysis was carried out with an LTQ-Orbitrap Velos instrument and data analysis was performed as described in Section 2.2.9.3. In both *Mm* Ca6b and *Ct* Cas6b a single region of the protein was identified as cross-linked. Close inspection of the corresponding spectra enabled identification of the cross-linked residues in both cases (Table 3.4). Representative spectra of the cross-linked region for both the proteins are shown in Figure 3.5.

Table 3.4 Cross-links identified for the *Cas6b-crRNA cross-linking.*

Protein (Uniprot ID)	Peptide	Amino acid	RNA	Figure
Mm Cas6b (A4FXZ3)	^{182}NQNM(ox)VGFR^{189}	M^{185}	UG-PO_3, UG, UGC, UUCA-HPO_3, **UUGC-HPO_3**, UUGC	3.5 (A)
Ct Cas6b (A3DKC1)	^{184}MIGFK^{188}	M^{184}	**UGA**, UUG	3.5 (B)

Protein: Cross-linked protein (Uniprot ID); Peptide: Sequence of the cross-linked peptide, specified with the position of the peptide in the protein sequence; Amino acid: One-letter code specified with the position of cross-linked amino acid, RNA: composition of the RNAs observed cross-linked.
For peptides cross-linked to the RNA depicted in bold, the corresponding MS/MS spectra are shown in Figure 3.5.

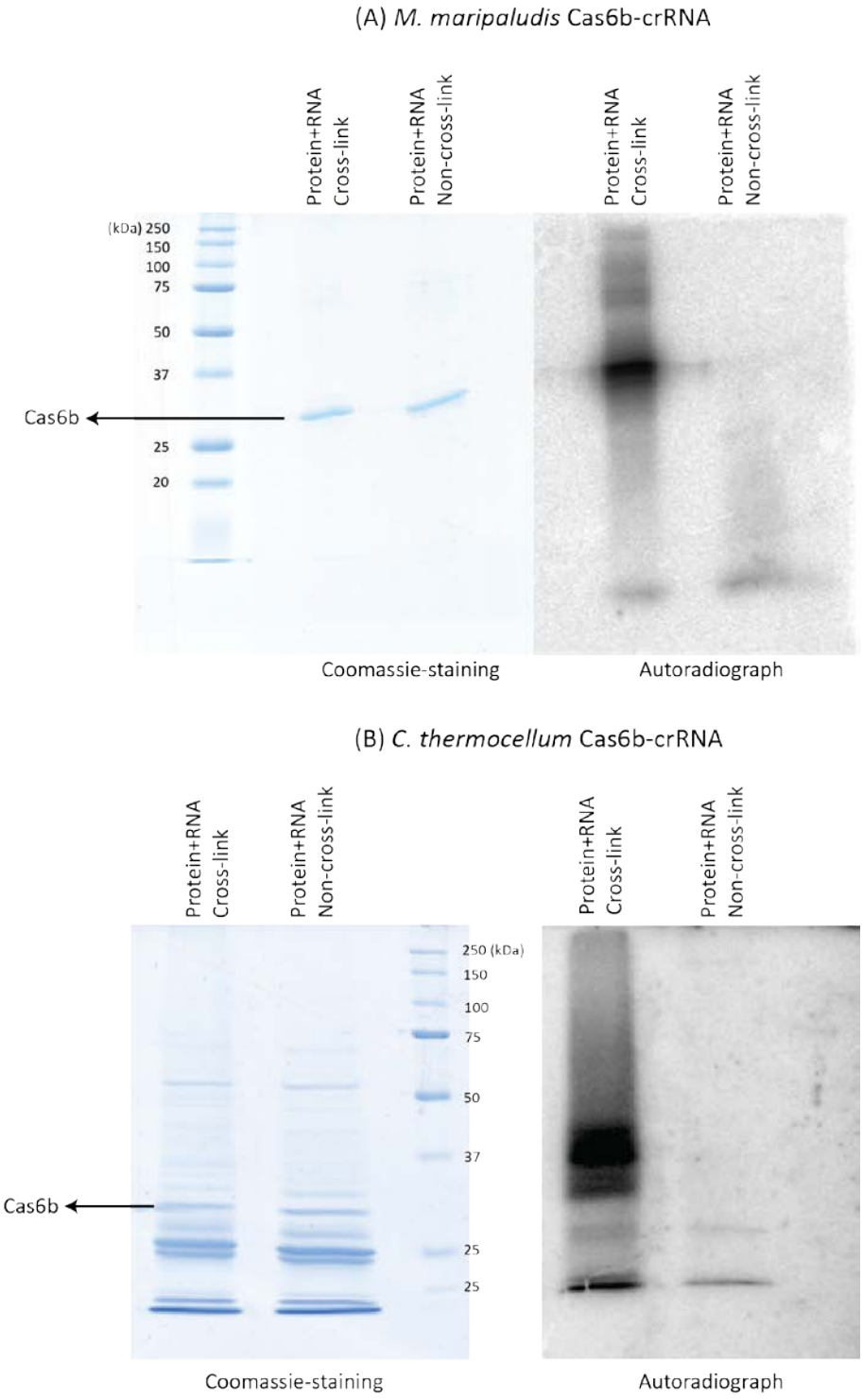


Figure 3.4 SDS-PAGE analysis of UV cross-linked Cas6b protein and ϒ-^{32}P-ATP labeled crRNA.

(A) *M. maripaludis* Cas6b cross-linked ϒ-^{32}P labeled crRNA, (B) *C. thermocellum.* Cas6b cross-linked with ϒ-^{32}P-ATP labeled crRNA. The protein-RNA cross-linking was analyzed using regular SDS-PAGE. Left panel: Coomassie stained gel and Right Panel: Autoradiography of the gel after 30 minutes exposure to the Phosphoimager screen. The Coomassie staining shows the purified Cas6b protein and the autoradiography shows cross-linking products of Cas6b-crRNA cross-linking.

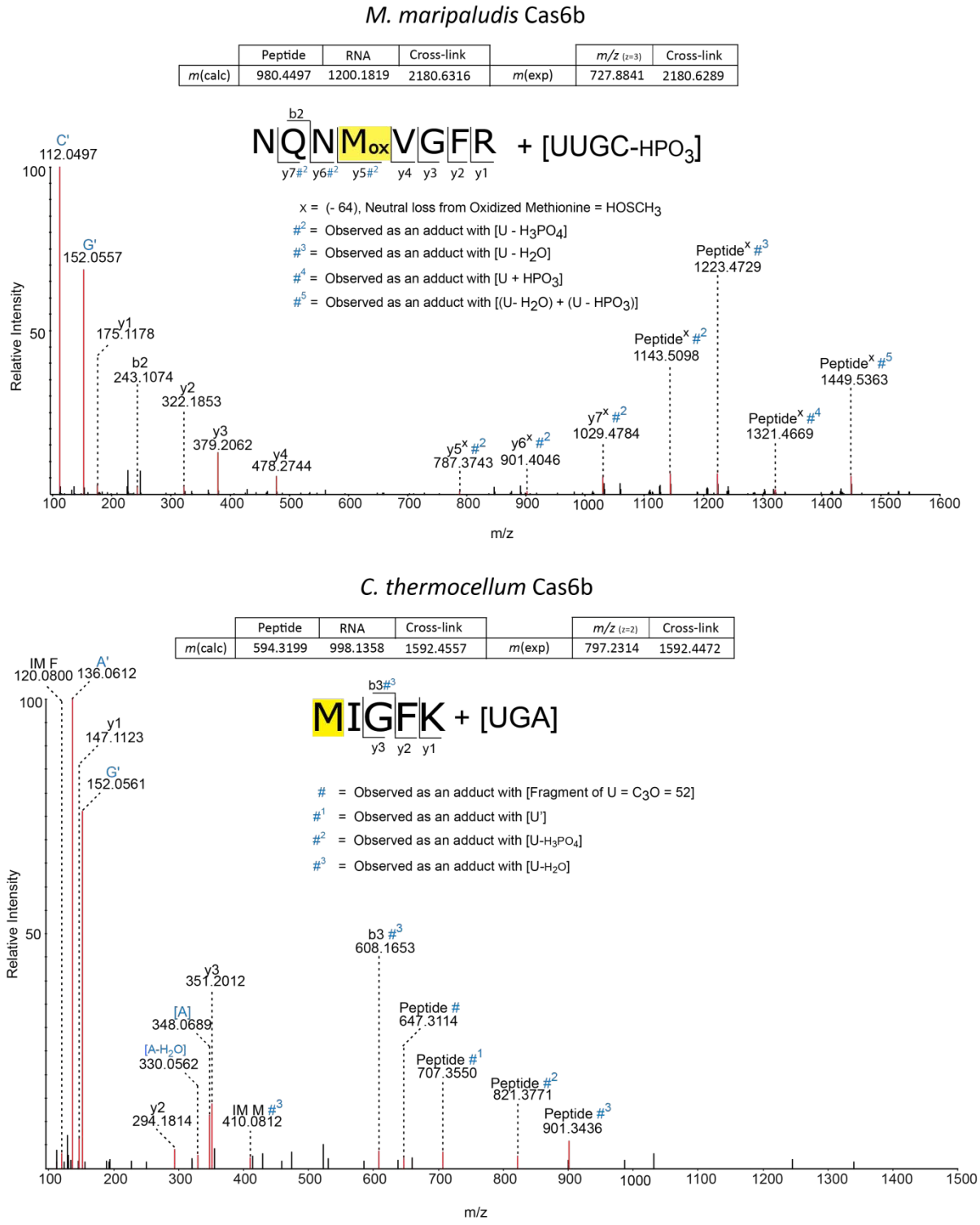


Figure 3.5 MS/MS spectra of the *M. maripaludis* Cas6b peptide ^{182}NQNM(ox)VGFR^{189} cross-linked to UUGC-PO$_3$ and *C. thermocellum* Cas6b peptide ^{184}MIGFK^{188} cross-linked UGA.

The cross-linked peptide sequence and its corresponding y- and b- type fragment ions are indicated at the top, with the cross-linked amino acid highlighted in yellow. The fragment ion peaks with their corresponding m/z values are marked in the spectrum. Ions with a mass shift corresponding to the cross-linked nucleotides are indicated with #, #^{1}, #^{2}, #^{3}, #^{4} and #^{5} with the respective adduct composition mentioned below the peptide sequence. The cross-linked nucleotide in both the cases is one of the uracil residues as the bases of nucleotides A, G and C are present as marker ions in the lower m/z regime of the spectrum. A': Base of A with marker ion of 136.06 m/z, G': Base of G with marker ion of 152.05 m/z, C': Base of C with marker ion of 112.05 m/z. U': Base of U observed in this case as adduct of 112.02 Da.

In *Mm* Cas6b, the peptide ^{182}NQNM(ox)VGFR^{189} was cross-linked to different RNA moieties as listed in Table 3.4. The methionine residue (M^{185}) in its oxidized state was identified as the cross-linked amino-acid residue. The cross-link spectra were manually validated and one example spectrum is shown in Figure 3.5 A. The b2 ion and the y- ion series until the y4 ion was observed without any mass shifts. All y- ions following y4 were observed shifted by a mass corresponding to the neutral loss from the oxidized methionine plus the RNA adduct [U-H_3PO_4; 226.0490 Da]. The neutral loss from oxidized methionine is a commonly observed phenomenon that occurs upon CID or HCD fragmentation, and it corresponds to a loss of methanesulfenic acid (CH_3SOH, 64 Da) as reported earlier [147-149]. Additionally, the cross-linked nucleotide could also be identified from the different RNA moieties that were observed cross-linked to ^{182}NQNM(ox)VGFR^{189}. From the manually annotated spectrum of peptide ^{182}NQNM(ox)VGFR^{189} cross-linked to UUGC-HPO_3, the cross-linked nucleotide was identified as one of the uracil residues, due to the shifted y- ions presenting mass-shifts corresponding to different uracil adducts (Figure 3.5 A) and the bases of nucleotides G and C were being present as marker ions (G': Base of G with marker ion of 152.05 m/z, C': Base of C with marker ion of 112.05 m/z) in the lower m/z regime of the spectrum. In all the different RNA moieties identified to be cross-linked to this particular peptide, the uracil was identified as the cross-linked nucleotide. When mapping the different cross-linked RNA moieties identified (UG, UGC, UUCA and UUGC) in the crRNA sequence, the 5' end U^{15} was identified as the only possible cross-linked nucleotide. (Figure 3.6A, crRNA sequence).

In *Ct* Cas6b, the peptide ^{184}MIGFK^{188} was observed to be cross-linked to UGA and UUG (Table 3.4) and M^{184} was identified as the cross-linked amino acid residue. The cross-links were validated manually and one example spectrum is shown in Figure 3.5 B. The y- ion series until the y3 was observed without any mass shifts. The b3 ion however, was observed with U-H_2O adduct (306.012 Da) indicating M^{184} or I^{185} as the probable cross-linked amino acid residue. In addition, the immonium ion of methionine was observed shifted by the mass of U-H_2O adduct at 410.0812 m/z, which confirmed M^{184} as the cross-linked amino acid residue. The cross-linked nucleotide was mapped on the crRNA sequence using the same approach as discussed above and U^3 and U^{32} from the 5'end were identified as possible candidates (Figure 3.6 B, crRNA sequence).

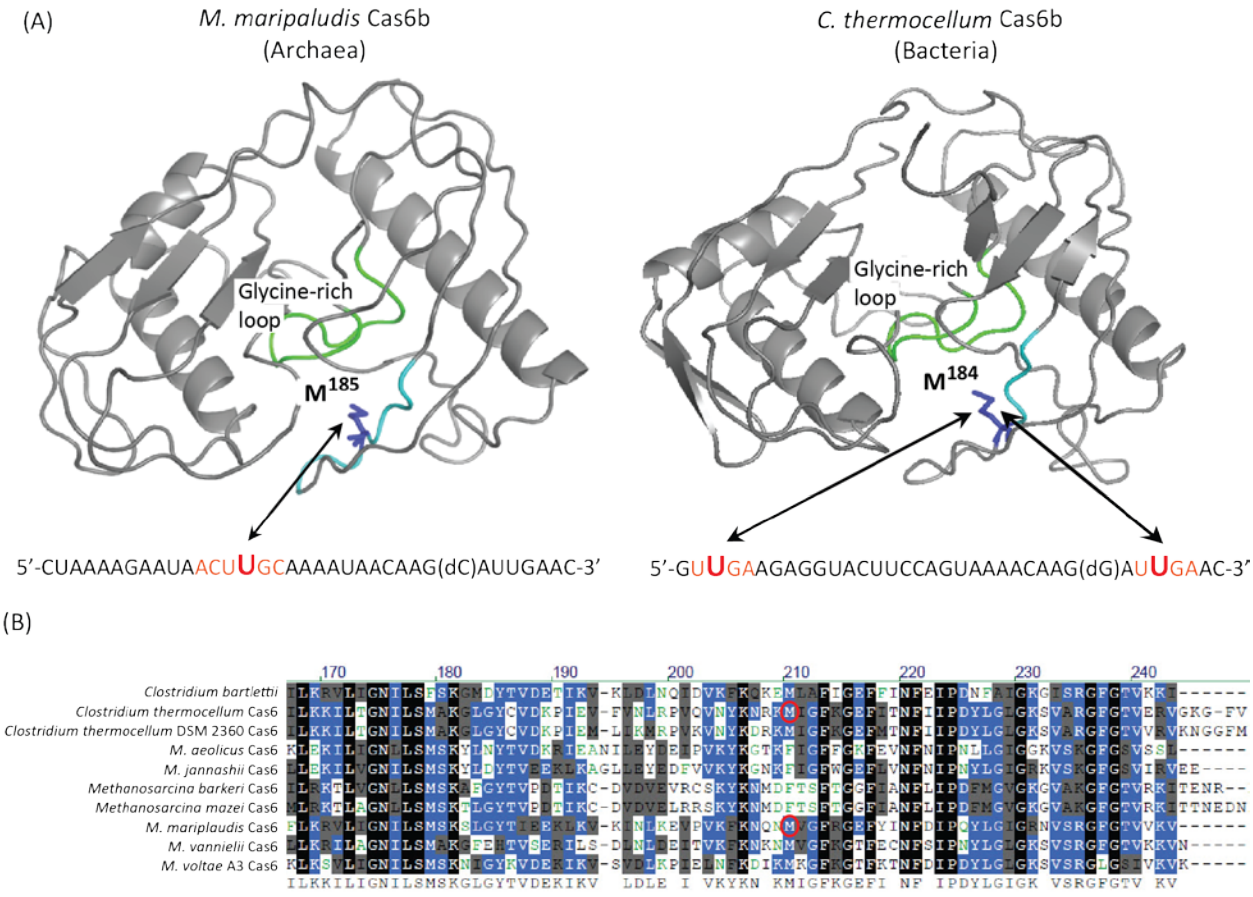


Figure 3.6 Cas6b-crRNA cross-linking in the archaeal and bacterial Cas6 proteins.

(A) Homology models of *Mm* Cas6b and *Ct* Cas6b (Phyre2 modeling on the basis of the *Pyrococcus furiosus* Cas6 [PDB: 3I4H]) as cartoon representations. The cross-linked peptides mapped on these models are indicated in cyan and the side-chains of the cross-linked amino acid residues (methionine) are indicated in blue and the conserved glycine-rich region is indicated in green. The RNA region observed as cross-linked moiety (in Table 3.3) is indicated in orange and the possible cross-linked nucleotide residue is highlighted in red on the respective crRNA sequence. (B) Alignment of Cas6b homologs (Clustal W) of subtype I-B showing the conserved amino acid residues (black). The cross-linked residues are indicated in the red circle. Figure 3.6 B was kindly provided by Prof. Lennart Randau (MPI, Terrestrial Microbiology, Marburg).

The Cas6 proteins comprise a conserved glycine-rich loop at the C-terminal of the protein [24] which is known for its RNA binding properties. The conserved glycine-rich loop for the two Cas6b proteins is shown in green in Figure 3.6A. The sequence alignment of the Cas6b homologues indicates the presence of a conserved glycine-rich region near the C-terminal of the protein. In addition, the cross-linked methionine residue identified for both *Mm* Cas6b and *Ct* Cas6b is also conserved across different archaeal and bacterial Cas6b proteins. These methionine residues are not an integral part of the C-terminal glycine-rich loop however they are in a close spatial proximity (Figure 3.6). Nonetheless, the structural and sequence alignment of Cas6b homologues indicate that with the protein-RNA cross-linking a new conserved RNA-

binding region was identified in addition to the evolutionary conserved gycine-rich region in the Cas6b proteins.

Further, our collaborators confirmed the results for the cross-linked amino acid residue identified in *Mm* Cas6b, with mutagenesis and co-crystallization studies as discussed later in Section 4.2.1.

3.2.2 Protein-RNA cross-linking in the Cas7 family proteins, *Thermofilum pendens* Csc2 and *Thermoproteus tenax* Cas7

In continuation with the approach for identification of protein-RNA interactions in single (recombinant) proteins and RNA, the investigation was extended to the two Cas7 proteins: *T. pendens* Csc2 (*Tp* Csc2) and *T. tenax* Cas7 from Type I-D and Type I-A CRISPR-Cas systems respectively. Both the proteins are representatives of the RAMP superfamily [21, 24]. They contain at least one RRM-like domain and although there is limited sequence conservation, they have a conserved glycine rich region between the α2 and β2 of the RRM domain [150]. The Cas7 proteins are characteristic of Type I and Type III systems and constitute the core subunit of interference complexes. Multiple copies of these proteins assemble in a helical fashion around the processed crRNA to form the backbone of large multi-subunit crRNP complexes [19, 40, 41, 43]. Recently they have become an important subject for structural studies. The first published structure of a Cas7 representative was of Type I-A *Sulfolobus solfataricus* Csa2 (*Ss* Csa2) [19], followed by *Methanopyrus kandleri* Csm3 (*Mk* Csm3) [55] and *Tp* Csc2 [103]. In addition, given the interesting results involving glycine rich regions in the Cas6b proteins, I wanted to verify whether this was a general mode of RNA binding in the extended protein family including the Cas7 proteins.

In this section I report the results for the investigation of RNA binding regions in the *Tp* Csc2 and *T. tenax* Cas7 using UV induced cross-linking and MS. For this study, the main focus was on the determination of specific regions of the two Cas7 proteins that interact with RNA along with the identification of the exact amino acid residues in these regions that cross-link to RNA. The experiments were performed in collaboration with Ajla Hrle from the group of Prof. Elena Conti (MPI Biochemistry, Martinsried), who expressed and purified the two Cas7 proteins. The protein-RNA cross-linking study was carried out after our collaborators successfully crystallized the *Tp* Csc2 [103]. In addition, these results were compared with the cross-links identified for the Cas7 family proteins in the Type I-E *E. coli* Cascade complex and Type III-A *T. thermophilus* Csm complex, where the protein-RNA cross-linking was performed at the level of fully assembled crRNP complexes (Details under Section 4.2.2).

In the EMSA studies performed by our collaborators, the *Tp* Csc2 was able to bind a polyU-RNA of 15 nucleotide length (polyU$_{(15)}$) [103] in a comparable fashion as previously reported for *Mk* Csm3 [55] where the latter was observed to bind polyU$_{(15)}$ RNA about ten times stronger than

the repeat sequence in the crRNA. *Tp* Csc2 was observed to bind to both polyU$_{(15)}$ and the *Tp* crRNA in a similar way [103]. The RNA used in this experiment was a synthetic polyU$_{(15)}$, more easily available compared to the *Tp* crRNA.

Both the purified protein and the polyU$_{(15)}$ RNA were incubated together in 1:1 ratio (1 nmol each) at 50 °C for 15 minutes for the protein-RNA cross-linking experiment. The UV cross-linking was performed using the standard protocol (Section 2.2.6.3). LC-MS analysis was carried out with the Orbitrap Velos and data analysis was performed as described in Section 2.2.9.3.

3.2.2.1 Cross-linked regions observed in the two Cas7 proteins

Three protein regions in *Tp* Csc2 and five protein regions in *T. tenax* Cas7 were identified cross-linked to RNA (Table 3.5). The cross-links were manually validated and the corresponding spectrum for each cross-link identified is shown in the Appendix, Figure 6.1.

In the *Tp* Csc2, the three cross-linked peptides identified were ^{82}LMAVTR^{87}, ^{124}KVSEEWNCTIQPPLAEFGEK^{143} and ^{346}WVEELKGGGQK^{356} and for each of them the cross-linked residue could also be identified. These included M^{83}, C^{131} and W^{346} respectively.

In the *T. tenax* Cas7 five cross-linked regions were identified. For the peptides ^{3}VAPPYVR^{9} and ^{164}SKEEQEGTEMMVFK^{177} the cross-linked amino acid residues were identified as Y^{7} and M^{163} respectively. However, for the peptide ^{14}FEAQLSVLTGAGNMGNYNMHAVAK^{37} and ^{127}VSFAVPVLEEK^{137} the cross-linked residue could not be identified. In the former peptide ^{28}G-N^{29} was observed as a probable cross-linked region. Since asparagine has never been observed to cross-link [108] we speculate the G^{28} to be the cross-linked amino acid residue. The peptide ^{145}FAVVHNRVDPFKR^{158} was observed as with a missed cleavage at R^{151} and also as two separate tryptic peptides ^{145}FAVVHNR^{151} and ^{152}VDPFKR^{158}. In the former V^{148} was identified as the cross-linked residue and in the latter two V^{148} and F^{155} were identified as the cross-linked amino acid residues after manual validation of the fragment spectra (Figure 6.1 G and H).

The solved crystal structure of *Tp* Csc2 and the modelled structure of *T. tenax* Cas7 were used for mapping the RNA binding regions identified in these two proteins (Figure 3.7).

Table 3.5 List of cross-links identified for the *T. pendens* Csc2 *and T. tenax* Cas7.

Protein (Uniprot ID)	Peptide	Amino Acid	RNA	Figure
Tp Csc2 (A1RZU2)	^{82}LMAVTR^{87}	M^{83}	U, **UU**	6.1 A
	^{124}KVSEEWNCTIQPPLAEFGEK^{143}	C^{131}	**U**	6.1 B
	^{346}WVEELKGGGQK^{356}	W^{346}	**U**	6.1 C
T. tenax Cas7 (G4RJZ1)	^{3}VAPPYVR^{9}	Y^{7}	**U**, UU	6.1 D
	^{14}FEAQLSVLTGAGNMGNYNMHAVAK^{37}	^{28}G-N^{29}	**U**	6.1 E
	^{127}VSFAVPVLEEK^{137}	-	U-H_2O, **U**	6.1 F
	^{145}FAVVHNRVDPFKR^{158}	V^{148}	U, UU	-
	^{145}FAVVHNR^{151}	V^{148}	U, **UU**, UUU	6.1 G
	^{152}VDPFKR^{158}	F^{155}	**U**, UU	6.1 H
	^{164}SKEEQEGTEMMVFK^{177}	M^{173}	**U**, UU	6.1 I

Protein: Cross-linked protein (Uniprot ID); Peptide: Sequence of the cross-linked peptide, specified with the position of the peptide in the protein sequence; Amino acid: One-letter code specified with the position of cross-linked amino acid, RNA: composition of the RNAs observed cross-linked.
For peptides cross-linked to mono- di- or tri-nucleotides depicted in bold, the corresponding MS/MS spectra are given in Appendix, Figure 6.1.

3.2.2.2 Mapping the cross-linked residues on Tp Cscs2 crystal structure and T. tenax Cas7 model

The cross-linked residues identified for the Cas7 proteins were mapped on their respective structures as dpecited in a cartoon representation in Figure 3.7. For *Tp* Csc2 a high resolution crystal structure was available [PDB 4TXD] [103] whereas for *T. tenax* Cas7 a sturctural model was generated using Phyre2 server [141].

The three cross-linked residues identified in *Tp* Csc2 are present in a central positively-charged groove on the surface of the protein (Figure 3.8), suggesting that this region is important in mediating the Csc2 binding to RNA. Meanwhile, our collaborators performed mutagensis studies on the surface exposed regions of *Tp* Csc2, targeting K^{179} and R^{183} and observed a reduction in the RNA binding activity compared to the wild-type protein [103].

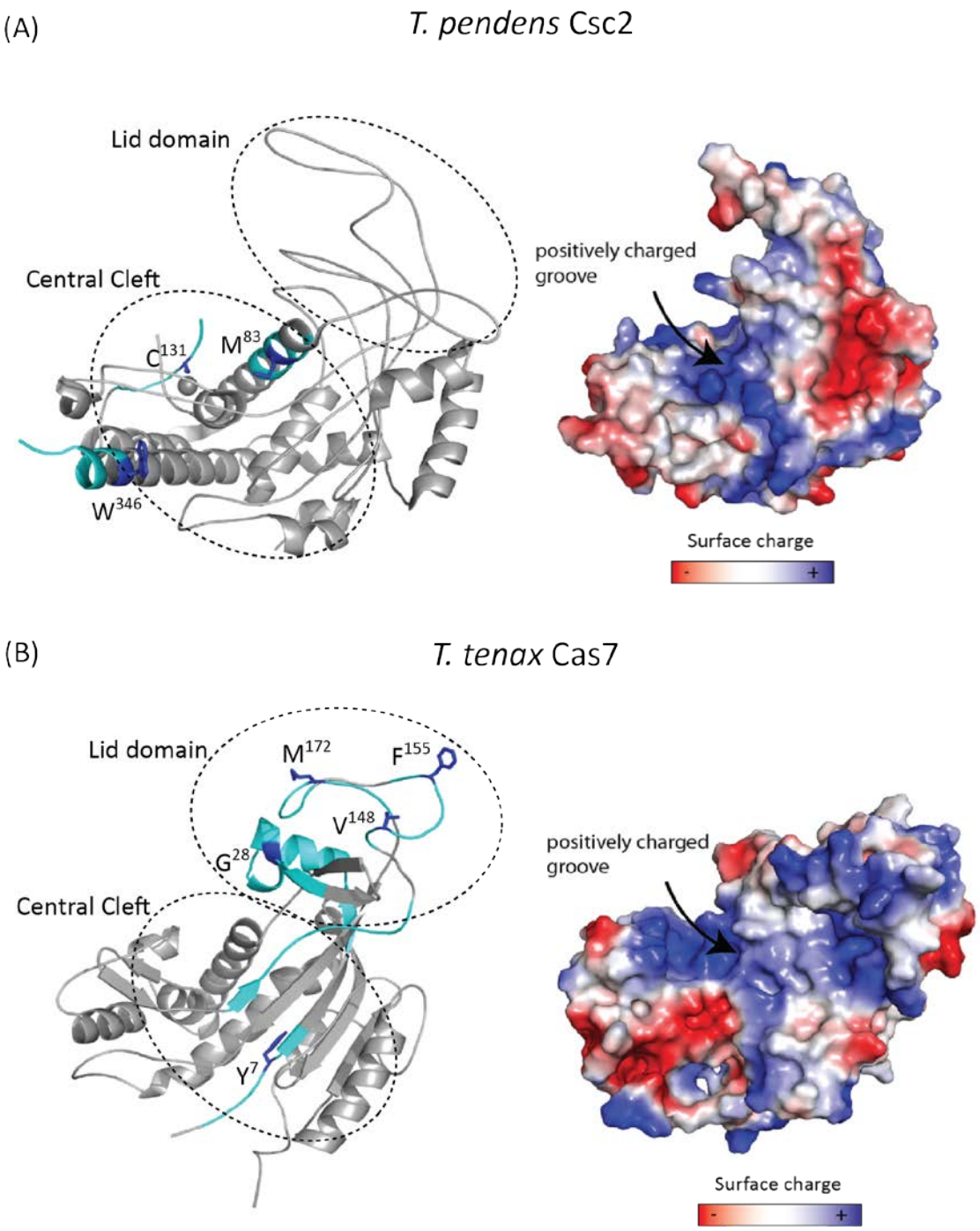


Figure 3.7 Cross-linked regions mapped on the *Tp* Csc2 crystal structure and *T. tenax* Cas7 model.

Left Panel: A cartoon representation (gray) of (A) *Tp* Csc2 solved crystal structure [PDB 4TXD] and (B) *T. tenax* Cas7 modelled structure generated using Phyre2 server [141] with the lid domain and the central cleft domain indicated in circles, with the cross-linked peptides colored in cyan and the cross-linked residues colored in blue (stick representation). Right panel: Surface representation of *Tp* Csc2 (A) and *T. tenax* Cas7 (B) (in the same orientation as in the left panel) depicting the electrostatic potential (red for electronegative and blue for electropositive). An arrow points to the prominent positively charged groove. Upper panel: Figure (A) originally published in [103] and reproduced with permission.

A model of four copies of *Tp* Csc2 arranged in the Type I-D Cascade assembly was derived from the interpretation of the cryo-EM structure of the Type I-E Cascade complex [43]. Both the cross-linked residues and the mutated residues were mapped on this model (Figure 3.8) and compared with the surface representation of the model showing the electrostatic potential. The surface representation showed that both the cross-linked and the mutated residues are placed within a central positively charged channel along the four copies of the protein.

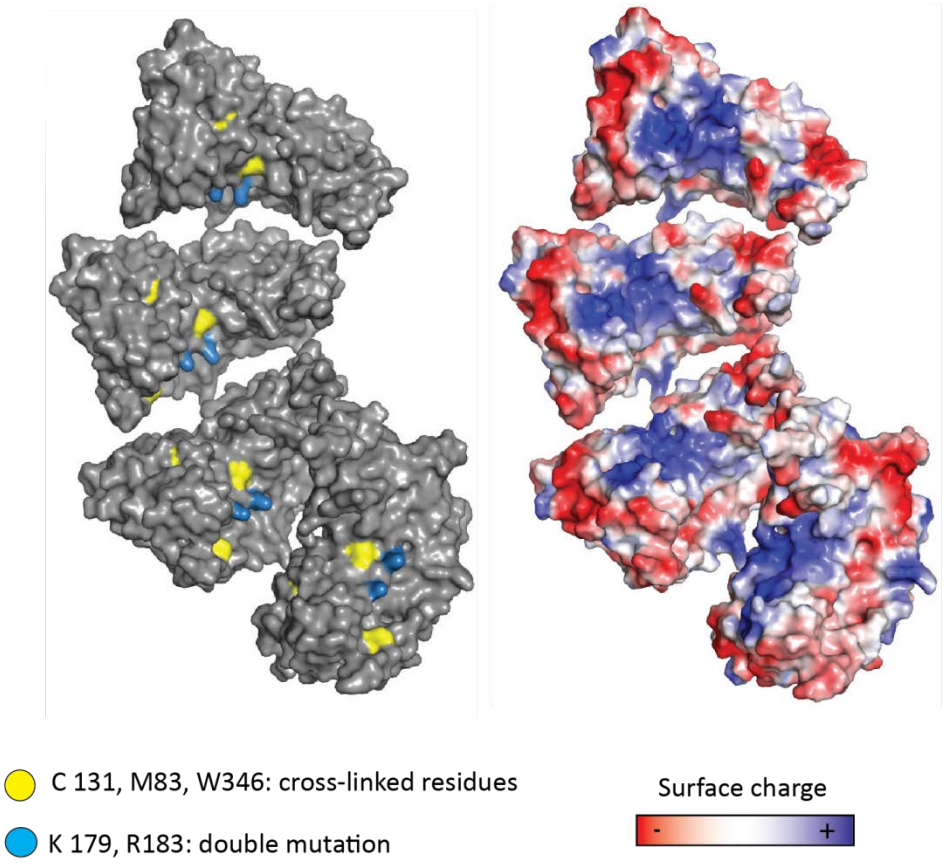


Figure 3.8 Cross-linked residues mapped on the model arrangement of four copies of *Tp* Csc2.

A model arrangement of multiple copies of *Tp*Csc2 (gray) upon Type I-D Cascade assembly; based on fitting a *Tp Csc*2 monomer into the EM density of the Type I-E Cascade [43]. Left Panel: Surface representation of the model showing the position of cross-linked (yellow) and mutated residues (blue). Right panel: Surface representation of the model (in the same orientation as the left panel) depicting the electrostatic potential (red for electronegative and blue for electropositive). Figure originally published in [103] and reproduced with permission.

Next, the five cross-linked amino acid residues identified in the *T. tenax* Cas7 were also mapped on a Phyre2 generated homology model structure of the protein (Figure 3.7). The mapped residues were localized in two regions of the protein i.e., the lid domain and the central cleft. The surface representation shows that cross-linked regions are placed in the positively charged central groove and the lid domain. Four out of five cross-linked residues identified are in the lid domain, which is in agreement with the previous reports from *Ss* Csa2 and *Mk* Csm3 in which the lid domain is involved in the nucleic acid binding [19, 55]. However, no such residues cross-linking to RNA were identified in the lid domain of *Tp* Csc2. The positively charged surface groove appears to be a conserved functional site for crRNA recognition whereas the influence of the lid domain in crRNA interaction varies with respect to different Cas family proteins.

3.2.3 Protein-RNA cross-linking in Type I-E Cascade complex from *E. coli*

The UV induced protein-RNA cross-linking approach is not limited to the investigation of single protein and RNA complexes and can also be applied to multi-subunit RNP assemblies of varied complexity. Here, I report the results from protein-RNA cross-linking investigations in the multi-subunit CRISPR ribonucleoprotein (crRNP) complex in *E. coli*.

The Type I-E surveillance complex in *E. coli* also referred to as Cascade complex was the first crRNP complex to be identified. It is a 405 kDa complex comprising 11 Cas protein subunits (from five Cas proteins) in a $Cse1_1Cse2_2Cas5e_1Cas7_6Cas6e_1$ stoichiometry and a crRNA. The 61 nucleotide crRNA spans the entire length of the complex. Amongst these 11 protein subunits, all except the two Cse2 subunits make a direct contact with the crRNA and all RNA-binding proteins except Cse1 contain a modified RRM [105]. The recently published high resolution crystal structures of the sea-horse shaped Cascade complex constituted a significant development in understanding the mechanism of crRNP complex assembly and target recognition [105, 106].

To this end, my part of the work focused on identification of RNA binding regions in the Cas protein and to identify the exact amino acid residues that interact with RNA using UV cross-linking and MS and to validate the interactions determined by the co-crystallization. A pre-assembled Cascade complex comprising 11 protein subunits and a crRNA was used for this study. The proteins and RNA components were cloned and co-expressed in *E. coli*. The complex was allowed to self-assemble *in vivo* and then affinity purified using the StrepTag on Cse2 [29]. The complex was purified by Tim Künne, from the group of Stan Brouns (Wageningen University, Wageningen, NL).

Approximately, one nmol of pre-assembled crRNP complex was used for the UV cross-linking and MS analysis. The standard protocol for cross-linking and enrichment of cross-linked peptides was followed as described in Section 2.2.6.4. LC-MS analysis was carried out in a Q Exactive HF mass spectrometer and data analysis was performed as described in Section 2.2.9.3. The cross-linked regions identified in different Cas proteins are listed in Table 3.6. In 90% of the cross-linked regions (10 out of 12) the exact cross-linked residue could also be identified and was validated with manual annotation of the corresponding spectrum (Appendix, Figure 6.2).

The 3.24 Å crystal structure of *E. coli* Cascade [105] (Figure 3.9A) was used as the core for our protein-RNA cross-linking investigations in the Type I-E *E. coli* Cascade complex. All the cross-links identified in our study were mapped on this crystal structure. However, there was a potentially important difference in the Cascade complex used for crystallization in [105] and the one used for protein-RNA cross-linking study. The two complexes were assembled around a different crRNA, i.e., the crRNA in both the complexes had a different spacer (as shown in the sequence below) which would influence the RNA binding in solution.

The sequence for the two crRNAs is indicated below. Both the crRNA had same 5' and 3' repeat sequence (upper case) derived from the conserved repeat regions but different spacer sequence (lower case). The crRNA in crystallized Cascade complex [105] (upper sequence) and the crRNA in UV cross-linked Cascade complex (crRNA*: lower sequence). The potential sites for cross-linking (uracil) are indicated in red.

crRNA : 5' AUAAACCGacgguauuguucagauccuggcuugccaacagGAGUUCCCCGCGCCAGCGGGG 3'

crRNA*: 5' AUAAACCGcugacgaccgggucuccgcaaguggcacuuuuGAGUUCCCCGCGCCAGCGGGG 3'

This difference in the crRNA led to some ambiguity while mapping the cross-linked residues on the crystal structure as some of the cross-linked residues could not be confidently assigned. However in such cases the most feasible interaction between the identified cross-linked amino acid residue and the closest uracil residue was indicated (Figure 3.9 – 3.12), as the cross-linked nucleotide in all the cross-links identified was uracil (Table -3.6). The representative spectra for each cross-linked region are shown in the Appendix, Figure 6.2. Below, a detailed comparison for the different regions of the complex will be provided.

Table 3.6 List of cross-links identified for the *E. coli* Type I-E Cascade complex.

Protein (Uniprot ID)	Peptide	Amino Acid	RNA	Figure
Cas6e (Q46897)	^{103}LDSKGNIK^{110}	K^{106}	**U-H_2O**	6.2 A
	^{136}VEDVHPISERPQYFSGDGK^{154}	^{145}R-Y^{148} *	**U**, UU	6.2 B
Cse2 (P76632)	^{13}AWQQLDNGSCAQIR^{26}	C^{22}	**U**, UG	6.2 C
	^{43}LVQPFGWENPR^{53}	W^{49}	**U**	6.2 D
	^{61}M(Ox)VFCLSAGK^{69}	F^{63}	**U**	6.2 E
	^{78}KSEQTTGISLGR^{89}	K^{78}	**U-H_2O**	6.2 F
	^{136}MLTWWGK^{142}	W^{139}	**U**	6.2 G
Cas7 (I2ZSV0)	^{21}DDMNMQKDAIFGGK^{34}	K^{27}	**U-H_2O**	6.2 H
	^{51}SGYYAQNIGESSLRTIHLAQLR^{72}	-	U, **UG-H_2O**	6.2 I
	^{83}FDQKIIDK^{90}	K^{86}	**U-H_2O**	6.2 J
	^{91}TLALLSGKSVDEAEK^{105}	^{97}G-K^{98}*	**U-H_2O**	6.2 K
	^{129}AEADNLDDKK^{138}	K^{137}	**U-H_2O**	6.2 L
	^{139}LLKVLK^{144}	K^{141}	**U-H_2O**	6.2 M
	^{142}VLKEDIAAIR^{151}	K^{144}	**U-H_2O**	6.2 N
	^{166}MATSGMMTELGK^{177}	M^{166}	U-H_2O, **U,** UU, AU, UC, UG, UUC	6.2 O
	^{166}M(Ox)ATSGMMTELGK^{177}	-	U-H_2O, U, UA, UC, UG, UU	-
	^{166}MATSGM(Ox)MTELGK^{177}	M^{166}	U-H_2O, U, UA-H_2O, UA, UC, UG, UU, UUC	-
	^{166}MATSGMM(Ox)TELGK^{177}	M^{166}	U-H_2O, U, UG-H_2O, UA	-
	^{166}M(Ox)ATSGM(Ox)MTELGK^{177}	-	U-H_2O, U, UA	-
	^{166}M(Ox)ATSGMM(Ox)TELGK^{177}	-	U, UC, UA	-
	^{166}MATSGM(Ox)M(Ox)TELGK^{177}	M^{166}	U, UA ,UC, UU, UUC	-
	^{166}M(Ox)ATSGM(Ox)M(Ox)TELGK^{177}	-	U, UU, UA, UC	-
Cas5e (H0Q9G2)	^{9}LAGPMQAWGQPTFEGTRPTGR^{29}	W^{16}	**U**	6.2 P
	^{9}LAGPM(Ox)QAWGQPTFEGTRPTGR^{29}	W^{16}	U	-
	^{84}DYHTVLGAR^{92}	Y^{85}	**U**, UA	6.2 Q
	^{142}YTPYLGR^{148}	Y^{145}	U, **UA**, UU,UUA	6.2 R
	^{198}DEPMITLPR^{206}	P^{200}	**U**,UU	6.2 S
	^{198}DEPM(Ox)ITLPR^{206}	P^{200}	U,UU	-
Cse1 (Q46901)	^{395}ALYTFAEGFK^{404}	F^{403}	**U**	6.2 T

Protein: Cross-linked protein (Uniprot ID); Peptide: Sequence of the cross-linked peptide, specified with the position of the peptide in the protein sequence; Amino acid: One-letter code specified with the position of cross-linked amino acid, RNA: composition of the RNAs observed cross-linked.
For peptides cross-linked to mono- di- or tri-nucleotides depicted in bold, the corresponding MS/MS spectra are given in the Appendix, Figure 6.2.
* Exact cross-linked residue could not be unambiguously identified; the residues which are most likely cross-linked are indicated instead.

3.2.3.1 Cross-links identified in Cas6e (head of the crRNP complex)

The Cas6 proteins are metal dependent endoribonculeases that process the long crRNA transcripts [39]. In *E. coli* Cascade complex the Cas6e protein binds 3' end of the crRNA at the head of the complex (Figure 3.9B). Upon UV cross-linking of the entire crRNP complex, two cross-linked regions were identified for the Cas6e protein. In the first region ^{103}LDSKGNIK^{110}, K^{106} was identified cross-linked to U-H_2O. The K^{106} is located in the positive charged groove-loop (A^{98}-E^{119}) on the C-terminal RRM domain of the Cas6e, that has been reported to make extensive electrostatic contacts with the 3' crRNA stem-loop [105].

The second cross-linked region identified for Cas6e was ^{136}VEDVHPISERPQYFSGDGK^{154}, but the exact cross-linked residue could not be unambiguously identified. Mass shifts present in the fragment spectrum identify ^{145}R-Y^{148} as the cross-linked region. The exact amino acid residue could not be identified due to a lack of fragment ions in that region of the ion-series (Figure 6.2 B).

Both cross-linked regions are located in the flexible loops that are in a close spatial proximity to the two adjacent uracil residues (U^{44} and U^{45}) present in the conserved 3' repeat sequence of the crRNA. The cross-linked regions mapped on the crystal structure of Cas6e are depicted in Figure 3.9B.

3.2.3.2 Cross-links identified in Cse1 and Cas5e (tail of the crRNP complex)

Together the three proteins Cse1, Cas5e and Cas7.6 form the tail of the crRNP complex and the 5' end of crRNA is sandwiched between these three proteins (Figure 3.9). One RNA binding region was identified for Cse1 and four for Cas5e in the protein-RNA cross-linking studies (Figure 3.9 C and D).

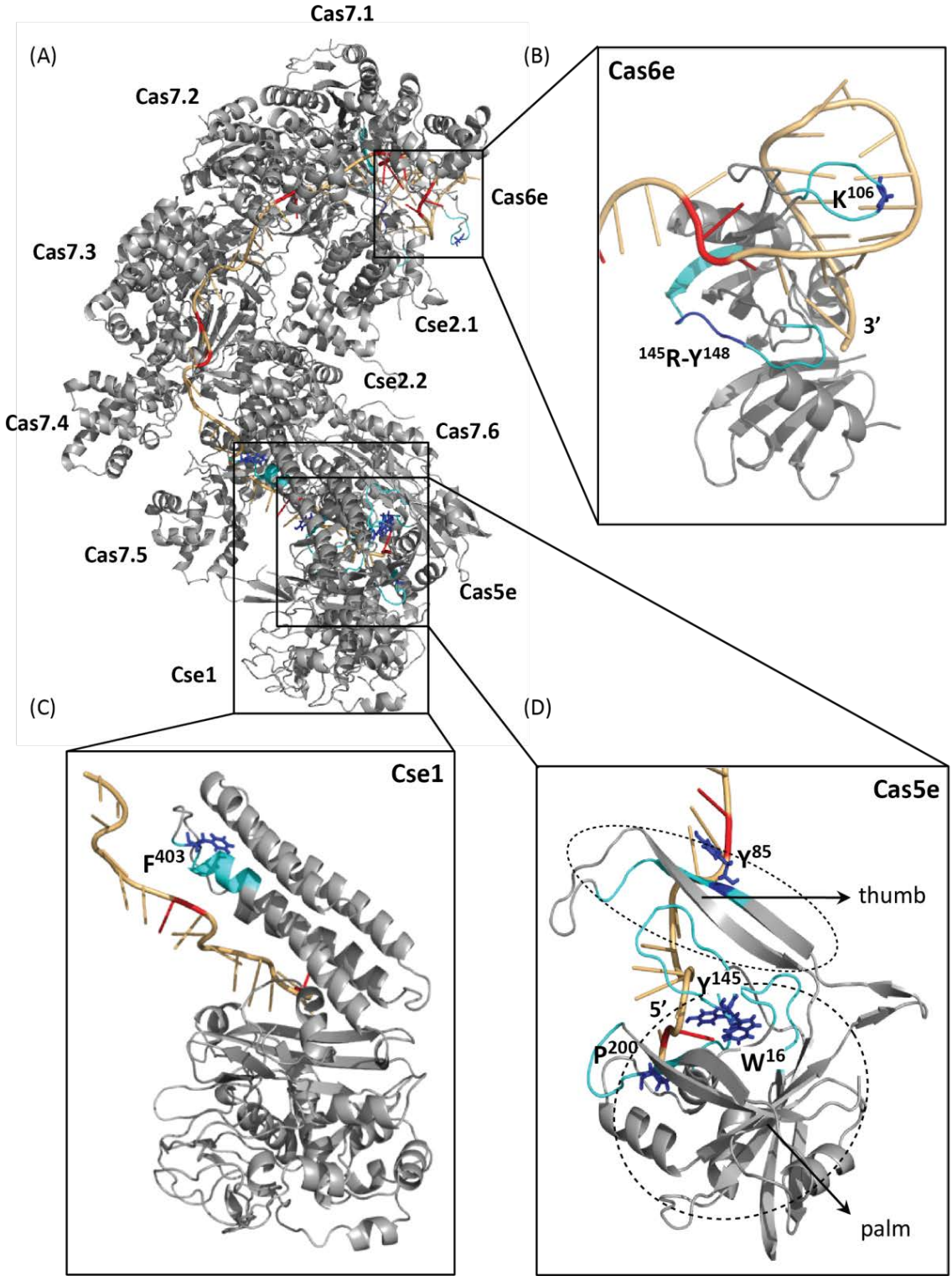


Figure 3.9 Cross-linked regions identified for proteins Cas6e, Cas5e and Cse1 mapped on the crystal structure.

(A) Cartoon representation of the Type I-E Cascade complex from *E. coli* [PDB 1VY8] [105] comprising 11 Cas proteins (gray) and one crRNA (light orange). The nucleotide positions on the crRNA where a corresponding uracil residue is present in the crRNA* are indicated in red. The cross-linked regions for Cas6e, Cas5e and Cse1 are indicated in cyan and the side-chains of cross-linked residues as sticks (blue). The zoomed in structures of Cas6e (B), Cse1 (C) and Cas5e (D) with the cross-linked regions and the cross-linked residues as depicted in (A) (other subunits were removed for clarity).

Cse1

Cse1 is the largest protein in the Cascade complex comprising two-domains. A globular fold that interacts with RRM domain of Cas5e and C-terminal four-helix bundle that contacts the C-terminal of Cse2.2. Only one cross-linked region was identified in Cse1. The peptide ^{395}ALYTFAEGFK^{404} was found cross-linked to a uracil nucleotide and F^{403} was identified as the cross-linked residue (Figure 6.2 T). In addition, the cross-linked region was mapped on the crystal structure of Cse1 and it was observed that F^{403} was in close spatial proximity to the U^{10} (nearest uracil residue) on the crRNA (Figure 3.9 C). The cross-linked region is located in a significant domain of the Cse1 protein i.e., the four-helix bundle at C-terminal of Cse1 that connects with C-terminal of Cse2.2, forming a structural bridge between the Cse1 tail and the Cas6e head [105].

Cas5e

The Cas5e protein has a “right-handed fist-shape” structure with the thumb arching across the top of the fist/palm (Figure 3.9D) [105, 106]. Out of the four cross-linked regions identified for Cas5e; the peptide ^{84}DYHTVLGAR^{92} is located in the arch of the Cas5e thumb. The arch is considered a modified RRM comprising 50 amino acid insertion between two β strands, the residues from V^{79}-A^{114} form together the thumb [106]. The cross-linked residue identified in this peptide was Y^{85} (Figure 6.2 Q), which was observed to be in close spatial proximity to U^{10} residue when mapped onto the crystal structure.

The protein regions M^{1}-S^{78} and S^{115}-Q^{224} form the fist/palm of Cas5e. Three cross-linked regions were identified for the palm domain of Cas5e. The peptides ^{9}LAGPMQAWGQPTFEGTRPTGR^{29}, ^{142}YTPYLGR^{148} and ^{198}DEPMITLPR^{206} were observed cross-linked to uracil residue and W^{16}, Y^{145} and P^{200} were identified as the cross-linked residues in the three peptides, respectively (Figure 6.2 P, R and S). In the crystal structure, all three cross-linked residues are present in close spatial proximity to the U^{2} residue, in the conserved 3’ repeat sequence of the crRNA. As observed in the crystal structure, the side-chain of U^{2} is adjacent to the aromatic side-chains of Y^{145} and W^{16} (Figure 3.9 D).

The peptide ^{42}YTPYLGR^{148} was observed cross-linked to UA RNA, and in the crRNA the nucleotides adjacent to U^{2} are adenine. This further confirms U^{2} as the RNA nucleotide cross-linked to Y^{145}, in agreement with previous reports [106].

3.2.3.3 Cross-links identified in Cas7 (helical backbone of the crRNP complex)

Cas7 proteins are one of the major constituents of the crRNP complex as they form the helical backbone of the complex and connect the head and the tail proteins. In the *E. coli* Cascade complex six Cas7 subunits (Cas7.1, Cas7.2, Cas7.3, Cas7.4, Cas7.5 and Cas7.6) assemble along the crRNA as depicted in Figure 3.9 A and Figure 3.10 A. The Cas7 proteins fold into a right hand shaped structure with a modified RRM domain forming the palm, residues G^{59}-A^{181} forming a helical domain that resembles the fingers, residues V^{193}-V^{223} forming a loop in the shape of a thumb [105] (Figure 3.10 B and C). The Cas7 filament is organized in such a way that the thumb of one Cas7 subunit extends towards the fingers of the adjacent subunit (Figure 3.10 A).

With protein-RNA cross-linking five cross-linked regions were identified for Cas7, ^{21}DDMNMQKDAIFGGK^{34}, ^{51}SGYYAQNIGESSLRTIHLAQLR^{72}, ^{83}FDQKIIDKTLALLSGKSVDEAEK^{105}, ^{129}AEADNLDDKKLLKVLKVLKEDIAAIR^{151} and ^{166}MATSGMMTELGK^{177}. The third and fourth regions were observed as different tryptic peptides resulting from missed cleavages at lysine residues, because most of these lysine residues were the sites for cross-linking (Table 3.6). Only for the second protein region no cross-linked amino acid residue could be identified. Details of the cross-linked peptides identified are listed in Table 3.6 and the corresponding fragment spectra and corresponding manual annotation are depicted in the Appendix (Figure 6.2 H-O).

The six cross-linked residues in the Cas7 protein included K^{27}, K^{86}, K^{98}, K^{137}, K^{141}, K^{144} and M^{166}. On the crystal structure these cross-linked residues could not be mapped unambiguously owing to the multiple stoichiometry of the Cas7 proteins. Given that six copies of the Cas7 protein are present in the complex, the cross-linked residues could not be assigned to a specific residue/protein in the crystal structure. Possible representations are depicted in Figure 3.10 and 3.11.

Cas7.1

All the cross-linked residues were mapped on Cas7.1 due to this protein being adjacent to the stretch of four consecutive uracil residues U^{33}-U^{36} and U^{31} in the crRNA used in the cross-linked Cascade complex. Two residues, K^{27} located in the palm domain and M^{166} in the finger domain present their side-chains in close proximity to this uracil-rich region (Figure 3.10 C). It can therefore be expected that these two particular residues in Cas7.1 are cross-linked.

The K^{86}, K^{98}, K^{137}, K^{141} and K^{144} are located in the lysine-rich helix i.e., the finger domain of Cas7.1. Although the finger domain has never been reported to have a direct contact with the

crRNA, our investigation revealed that the lysine rich-loop is a favorable region for cross-linking. In addition, the presence of the four uracil residue stretch in close proximity gives an indication that the crRNA (in the Cascade complex used in our study) weakly interacting with the lysine-rich region could potentially flip over to the finger domain of the Cas7.1 when in solution. In previous protein-RNA cross-linking studies both the lysine and uracil have been reported as likely candidates for cross-linking (Urlaub lab, unpublished data). In addition, in all other Cas7 subunits, the lysine-rich helix is not in a close spatial proximity to the crRNA, which limits the number of feasible option for assigning cross-linked amino acid residues to these regions.

Other Cas7 subunits

The crRNA sequence close to the Cas7.2 subunit contains a single uracil residue (Figure 3.10 B). Two cross-linked amino acid residues (K^{27} and M^{166}) were identified in Cas7.2, both in close spatial proximity to U^{31} in the crRNA.

None of the cross-linked amino acid residues was mapped onto Cas7.3, due to its proximal crRNA region not having any uracil residue present.

In Cas7.4, Cas7.5 and Cas7.6 only one cross-linked residue could be unequivocally assigned to the structure. In Cas7.4, K^{27} amino acid residue is proximal to U^{21} nucleotide residue on the crRNA (Figure 3.11 B). In addition the proposed cross-linked regions from Cas7.5 and Cas7.6 are depicted in Figure 3.11C. The U^{10} residue is tightly located between the side chains of M^{166} in Cas7.5 and K^{27} in Cas7.6, which indicates this contact constitutes one of the identified protein-RNA cross-linking regions.

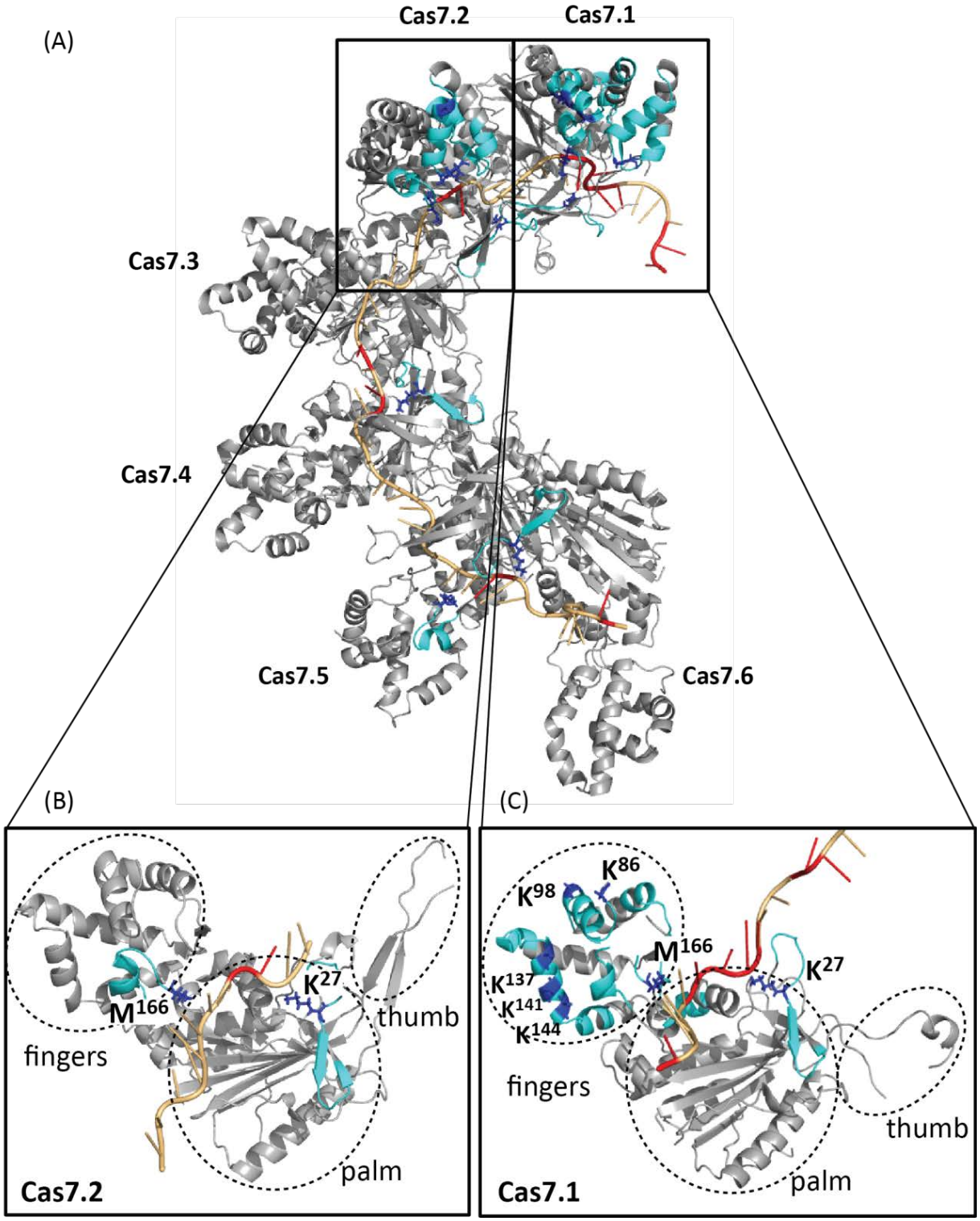


Figure 3.10 Cross-linked regions identified for the Cas7 proteins mapped on the crystal structure at the possible sites for cross-linked residues in Cas7.1 and Cas7.2.

(A) Cartoon representation of the six Cas7 subunits (gray) wrapped around the crRNA (light orange) in the Type I-E Cascade complex from *E. coli* [PDB 1VY8] [105] (other subunits were removed for clarity). The nucleotide positions on the crRNA where a corresponding uracil residue is present in the crRNA* are indicated in red. The proposed cross-linked regions are indicated in cyan and the side-chains of cross-linked residues as sticks (blue). The zoomed in structures of Cas7.2 (B) and Cas7.1 (C) with the cross-linked regions and the cross-linked residues as depicted in (A).

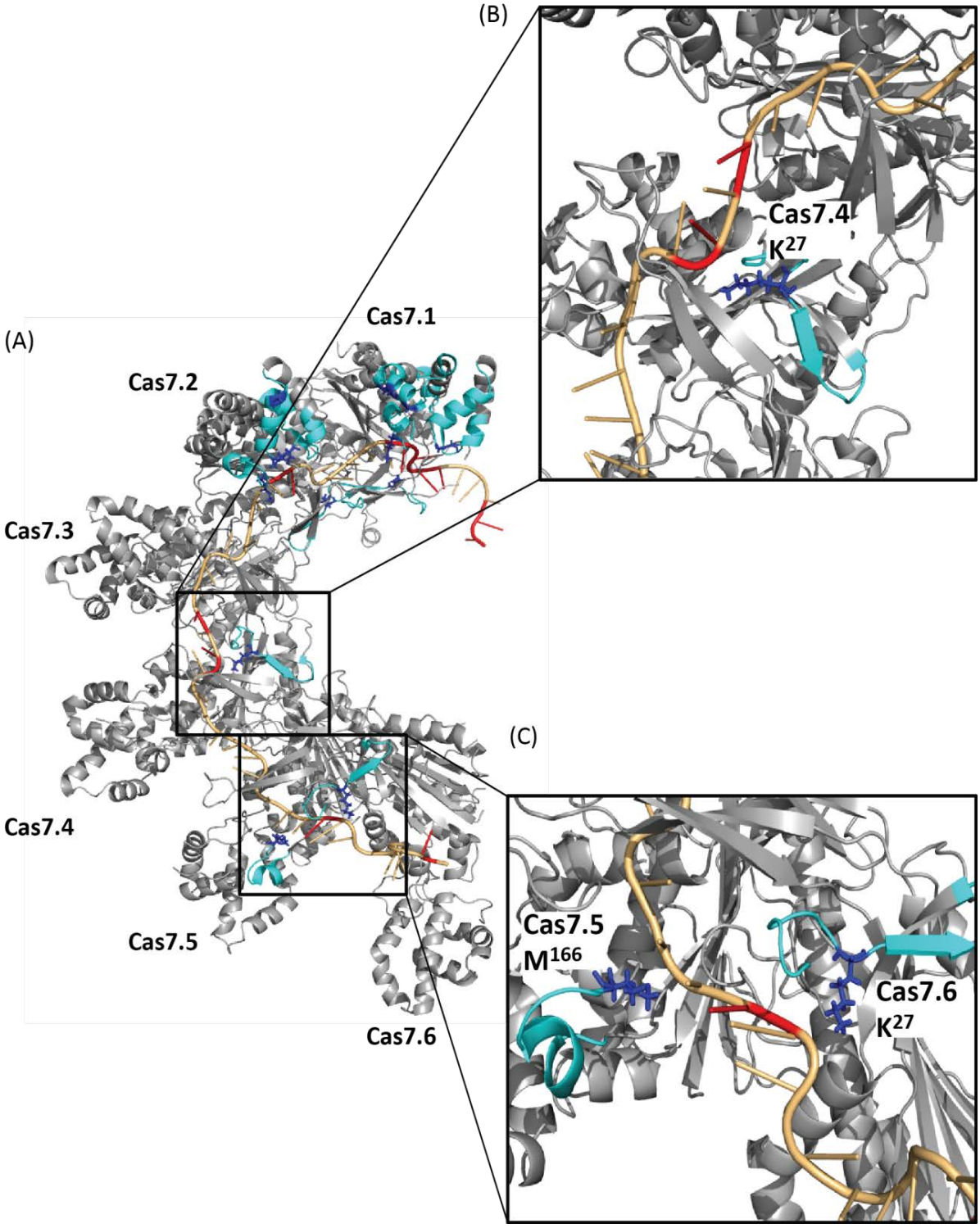


Figure 3.11 Cross-linked regions identified for the Cas7 proteins mapped on the crystal structure at the possible sites for cross-linking in Cas7.4, Cas7.5 and Cas7.6.

(A) Cartoon representation of the six Cas7 subunits (gray) wrapped around the crRNA (light orange) in the Type I-E Cascade complex from *E. coli* [PDB 1VY8] [105] (other subunits were removed for clarity). The nucleotide positions on the crRNA where a corresponding uracil residue is present in the crRNA* are indicated in red. The proposed cross-linked regions are indicated in cyan and the side-chains of cross-linked residues as sticks (blue). The zoomed in structures of Cas 7.4 (B) and Cas7.5/Cas7.6 (C) with the cross-linked regions and the cross-linked residues as depicted in (A).

Additionally, the peptide ^{166}MATSGMMTELGK^{177} and several modified versions of the peptide with oxidized methionine/s were observed cross-linked to a wide range of RNA moieties, with M^{166} identified as the cross-linked residue in 40 out of 70 cross-links observed for the *E. coli* Cascade complex (Table 3.6). Therefore, the M^{166} residue indicated in Cas7.1, Cas7.2, and Cas7.5 (Figure 3.10 B, C and Figure 3.11 C) is proposed to cross-link to the indicated uracil residues. This is in agreement with the previous X-ray studies showing the side-chain of highly conserved M^{166} intercalating between the 3rd and 4th bases of each crRNA segment close to the respective Cas7 subunit [105, 106].

In general, the six Cas7 subunits together forming the Cas7 filament have been reported to have extensive interactions with the crRNA in the Cascade complex. A major portion of the sugar-phosphate backbone of crRNA is buried by the Cas7 proteins, leaving only the bases exposed to the solvent [105, 106].

3.2.3.4 Cross-links identified in Cse2 (belly of the crRNP complex)

In the fully assembled crRNP complex the Cse2 protein is present as two copies Cse2.1 and Cse2.2 that assemble along the "belly" of the Cascade complex. The two subunits form a head-to-tail dimer that contacts the backbone of Cas7 proteins as shown in Figure 3.9A. Additionally, both available crystal structures [105, 106] show that the two Cse2 subunits do not form any direct contacts with the crRNA (Figure 3.12).

However, in the protein-RNA cross-linking studies five different cross-linked regions were identified in Cse2 protein and in all five peptides the exact cross-linked residue could be identified (Table 3.6, Figure 6.2 C-G). It cannot be unambiguosly indicated which of the cross-linked region/residue corresponds to which of the two Cse subunits, but the possibilities are considered in Figure 3.12.

When mapped on the crystal structure, all the cross-linked amino acid residues are located on the surface opposite to the surface facing the crRNA. Despite the fact the two Cse2 subunits do not contact the RNA directly, the electrostatic calculations indicate that both the faces of the Cse2 dimer are positively charged and constitute potential RNA binding regions [43, 151]. In addition, the Cse2 proteins have been reported to play an important role in stabilizing the interaction with the target DNA [106]. Presence of residual crRNA or residual nucleotides from purification and the conformational variability of the Cascade complex in solution are potential explanations for the identification of protein-RNA cross-links in Cse2.

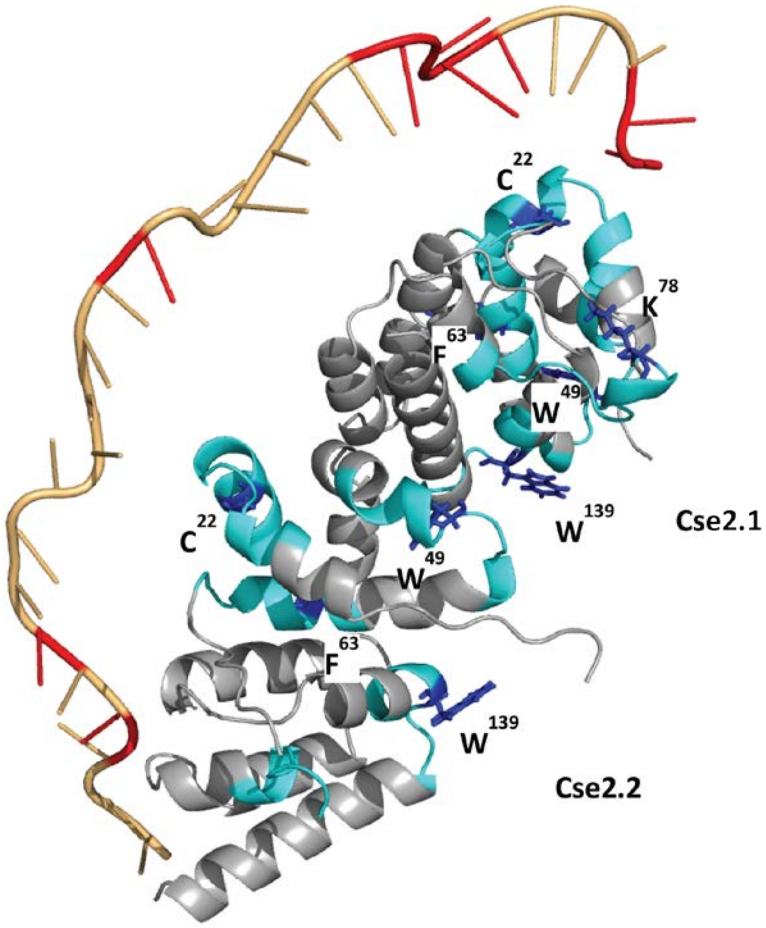


Figure 3.12 Cross-linked regions identified for the Cse2 proteins mapped on the crystal structure of both Cse2.1 and Cse2.2.

Cartoon representation of the six Cse2 subunits (gray) and the neighboring crRNA (light orange) in the Type I-E Cascade complex from *E. coli* [PDB 1VY8] [105] (other subunits were removed for clarity). The nucleotide positions on the crRNA where a corresponding uracil residue is present in the crRNA* are indicated in red. The cross-linked regions are indicated in cyan and the side-chains of cross-linked residues are indicated as sticks (blue). The mass spectrometric analysis does not allow for an unambiguous identification of the individual cross-links in the respective subunits. Cross-links are represented in both.

Overall this cross-linking study shows the power of UV induced protein-RNA cross-linking to identify direct protein-RNA interaction sites on the peptide or amino-acid level in a large, heterogeneous assemblies such as the multi-subunit *E. coli* Cascade complex. In addition, it provides invaluable in solution information, which can reflect the conformational landscape that occurs in the cell.

3.2.4 Protein-RNA cross-linking in Type III-A Csm complex from *T. thermophilus*

In continuation with our investigation of protein-RNA interactions in complex systems such as the multi-subunit crRNP complexes, the UV induced protein-RNA cross-linking approach was also tried on the Type III-A Csm complex from *T. thermophilus* (*Tt* Csm complex). This crRNP complex is composed of five protein subunits (Csm1-Csm5) in a proposed stoichiometry of $Csm1_1Csm_23Csm3_6Csm4_2Csm5_1$ and one crRNA of variable size (35-53 nucleotides) [59]. The complex has not been crystallized till date, however an EM structure of the complex (Figure 1.16) [59] exhibits the characteristic architecture of Type I-E Cascade and Type III-B Cmr complexes, as reported in [43, 52]. I set out to investigate whether the mode of RNA interaction was similar to the previous observations in Type I-E Cascade complex (Section 3.2.3). The endogenous complex was purified from its native host *Thermus thermophilus* HB8 using a genomic tag (His-tag) as described in [59]. The purified complex was provided by Raymond Staals from the group of Prof. John van der Oost (Wageningen University, Wageningen, NL).

UV cross-linking was performed using one nmol of pre-assembled *Tt* Csm complex. The cross-linking and data analysis was performed as described in the standard protocol (Section 2.2.6.5 and 2.2.9.3). Using this approach, nine cross-linked regions (corresponding to 12 cross-linked peptides) were identified in the entire crRNP complex. The composition of the cross-linked peptides and the RNA moieties is listed under Table 3.7. However, the *Tt* Csm complex used for cross-linking contained a mixture of different crRNAs (natural guides) isolated from *T. thermophilus*, therefore the exact cross-linked nucleotide on the crRNA could not be identified unambiguously.

Table 3.7 List of cross-links identified for the *T. thermophilus* Type III-A Csm complex.

Protein (Uniprot ID)	Peptide	Amino Acid	RNA	Figure
Csm1 (Q53W19)	^{371}RLHEALAR^{378}	-	**UUA**	6.3 A
Csm2 (Q53WF6)	^{35}LKSSQFR^{41}	K^{36}	**U-H_2O**, U	6.3 B
Csm3 (Q53WF5)	^{21}IGMSRDQMAIGDLDNPVVR^{39}	-	U, **UU**, UG	6.3 C
	^{40}NPLTDEPYIPGSSLK^{54}	^{49}P-K^{54*}	U-H_2O, U, **UG**, UA	6.3 D
	^{91}IFGLAPENDER^{101}	P^{96}	**U,** UU, UC, UG	6.3 E
	^{136}GGLYTEIKQEVFIPR^{150}	Q^{144}	U, UU, **UC**, UG, UCG, UUC, UUG	6.3 F

Csm3 (Q53WF5)	^{151}LGGNANPR^{158}	G^{153}	**UA**, UC, UG UGG, UCA, UUA	6.3 G
	^{159}TTERVPAGAR^{168}	R^{162}	**U**, UG, UGG, UUG	6.3 H
Csm4 (Q53WF4)	^{69}LPPVQVEETTLRK^{81}	P^{71}	U, **UG**, UUA	6.3 I
	^{126} TRVGVDR^{132}	V^{128}	**UU**, UC	6.3 J
Csm5 (Q53W18)	^{132}SPLGAYLPGSSVK^{144}	P^{139}	**U**, UA, UG, UUA	6.3 K
	^{255}MVLLAETFR^{263}	M^{255}	U-H_2O, **U**, UG	6.3 L

Table originally published in [59] and reproduced with permission.
Protein: Cross-linked protein (Uniprot ID); Peptide: Sequence of the cross-linked peptide, specified with the position of the peptide in the protein sequence; Amino acid: One-letter code specified with the position of cross-linked amino acid, RNA: composition of the RNAs observed cross-linked.
For peptides cross-linked to mono- di- or tri-nucleotides depicted in bold, the corresponding MS/MS spectra are provided in the Appendix, Figure 6.3.
* Exact cross-linked residue could not be unambiguously identified; the two potential cross-linked amino acid residues are indicated instead.

3.2.4.1 Cross-links identified in Csm3 (a Cas7 family protein)

Three different cross-linked regions were identified in Csm3. In the first protein region ^{21}IGMSRDQMAIGDLDNPVVRNPLTDEPYIPGSSLK^{54} two different tryptic peptides ^{21}I-R^{39} and ^{40}N-K^{54} were observed cross-linked to different RNA moieties (Table 3.7). In both the cross-linked peptides the exact cross-linked residue could not be identified. However for the latter ^{49}P-K^{54}, the six amino acid sequence was identified as the cross-linked peptide. The fragment spectrum did not provide sufficient ion signals to identify which of the amino acids carried the RNA mass shift (Figure 6.3 D).

In the second region ^{91}IFGLAPENDER^{101}, P^{96} was identified as the cross-linked residue. When the two regions were mapped on the *Tt* Csm3 model, it was observed that both regions lie in the central cleft of the protein, which is located in the conserved positively charged groove and is therefore a potential region for RNA interaction.

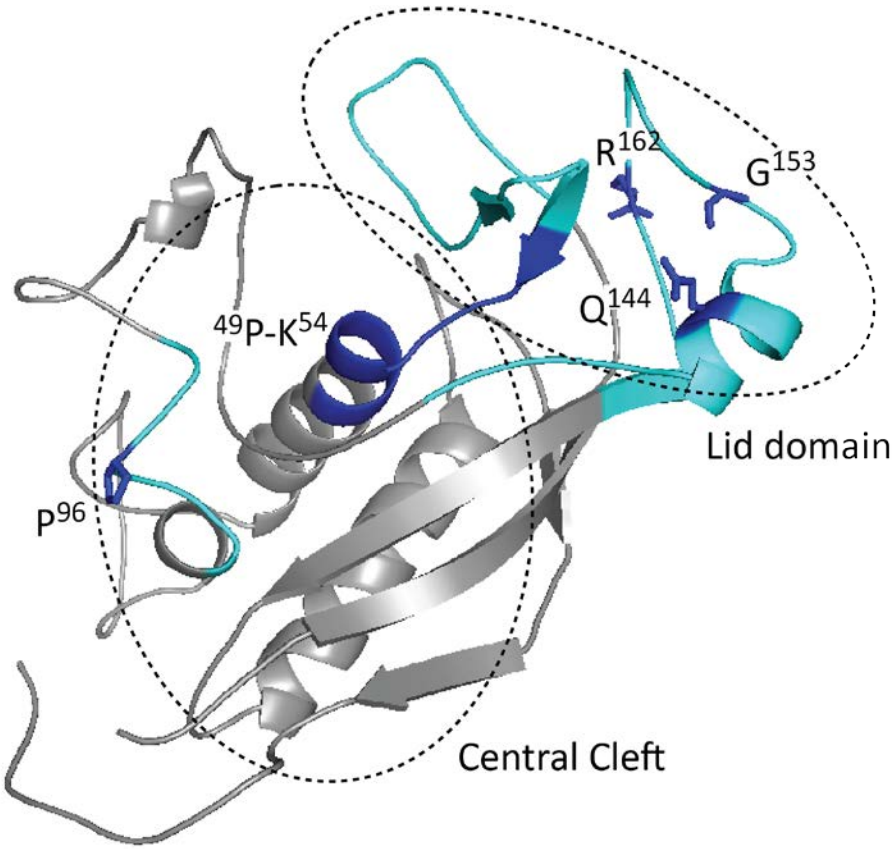


Figure 3.13 Cross-linked regions mapped on a *Tt* Csm3 homology model.

Cartoon representation of the *Tt* Csm3 homology model (gray) derived by Phyre2 homology modeling based on *E. coli* Cas7 [PDB 1VY8] [105]. The cross-linked regions are indicated in cyan and the side-chains of cross-linked are indicated as sticks (blue). The two universal domains of the Cas7 family proteins are indicated in the two circles.

The third cross-linked protein region identified consisted of ^{136}GGLYTEIKQEVFIPRLGGNAN-PRTTERVPAGAR^{168} peptide, which forms the lid-domain of the Csm3 complex. It was observed as three separate tryptic peptides ^{136}G-R^{150}, ^{151}L-R^{158} and ^{159}T-R^{168} cross-linked to different RNA moieties (Table 3.7). The corresponding cross-linked amino acid residue identified for each peptide was Q^{144}, G^{153} and R^{162} respectively. The lid-domain of the Cas7 family proteins constitutes yet another promising region for RNA-interaction, as this has been observed for the *T. tenax* Cas7 (Figure 3.7 B) and also in other Cas7 family proteins [103]. In general all the cross-linked residues mentioned above cannot be assigned to one single Csm3 subunit, as shown in Figure 3.13 because there are in total six Csm3 subunits in the *Tt* Csm complex.

Since the proteins Csm1, Csm2, Csm4 and Csm5 could not be crystallized till date, there was no crystal structure available to map the cross-linked amino acid residues identified for these proteins.

Overall, the results of cross-linking studies are very similar for both Type III-A Csm and Type I-E Cascade complex with at least one cross-linked region identified in all the Cas proteins in both the complexes. In addition, 28 out of 44 cross-links identified for the entire Csm complex, corresponded to Csm3, a Cas7 family protein. This result is expected as these proteins are present in the highest copy number in a crRNP complex and are in a close proximity to the crRNA, which provides a higher likelihood for crRNA-protein interaction.

3.2.5 Protein-RNA cross-linking in Type III-B Cmr complex from *T. thermophilus*

Thus far, the protein-RNA cross-linking investigation for multi-subunit crRNP complexes was performed only with the endogenous complexes (purified from endogenous sources). In this section I report the results from the investigation of protein-RNA interactions in both the endogenous and reconstituted Type III-B Cmr complex from *T. thermophilus* (*Tt* Cmr).

The *Tt* Cmr complex is composed of six different protein subunits (Cmr1-6) with a stoichiometry $Cmr1_1Cmr2_1Cmr3_1Cmr4_1Cmr5_3Cmr6_1$ and one crRNA. Electron microscopy studies revealed that the structure of *Tt* Cmr complex resembles a sea-worm [52], unlike the sea-horse model of Type I-E *E. coli* Cascade complex [105, 106]. In *Tt* Cmr, Cmr2 and Cmr3 form the tail, Cmr4 (Cas7 family protein), capped by Cmr5, form the helical backbone and Cmr1 and Cmr6 form a curled head (Figure 3.14). The overall architecture resembles notwithstanding the Type I-E Cascade complex.

The endogenous complex was purified from the host *T. thermophilus* HB8 using a genomic tag (His-tag) as described in [52]. For the obtention of reconstituted complex, the individual components were expressed in *E. coli* and purified separately; the complex was reconstituted in a reaction tube by mixing the components one at a time in the required stoichiometry. The complex was reconstituted using two different crRNAs:

46 nt crRNA
5’ AUUGCGACCCGUAGAUAAGGCGCCCGGGGACGACCACGUCAAGGCG 3’
40 nt crRNA (lacking the six nucleotides at the 3’ end)
5’ AUUGCGACCCGUAGAUAAGGCGCCCGGGGACGACCACGUC 3’

Deep sequencing analysis identified several different crRNAs that bind the endogenous complex. In most cases, however, the 40 nt crRNA was the most abundant followed by the 46 nt crRNA [52]. Due to this reason these two crRNAs were used in the reconstitution of the complex *in vitro*. Nevertheless for the endogenous complex, it cannot be distinguished which of the crRNAs is a part of the final complex that was cross-linked, because the complex was purified from the endogenous sources i.e., from the *T. thermophilus* HB8 where several different crRNAs are present.

Both endogenous and reconstituted complexes for this study were provided by Yifan Zhu from the group of Prof. John van der Oost (Wageningen University, Wageningen, NL). For both

complexes UV cross-linking was performed using one nmol of the respective complex. The cross-linking and data analysis was performed as described in (Section 2.2.6.5 and 2.2.9.3). In each of the two complexes, four cross-linked regions were identified. Three of these regions were observed in both complexes and the cross-linked residue identified in these three regions were also the same in both the endogenous and reconstituted complexes as summarized in Table 3.8. The corresponding fragment spectra for the cross-links identified are shown in the Appendix (Figure 6.4).

Table 3.8 List of cross-links identified for the endogenous and reconstituted *T. thermophilus* Type III-B Cmr complex.

Endogenous complex				
Protein (Uniprot ID)	**Peptide**	**Amino Acid**	**RNA**	**Figure**
Cmr2 (Q53W09)	^{160}DFAPVSWGSPAYK^{172}	Y^{171}	**U**	6.4 A
Cmr3 (Q53W08)	^{156}GHEGPVPETRTHVALDPAAQTAR^{178}	-	**U**, UU, UUA	6.4 B
	^{166}THVALDPAAQTAR^{178}	L^{170}	U, **UU**, UA, UUA, UGC	6.4 C
Cmr4 (Q53W06)	^{204}IRLDDETK^{211}	L^{206}	**U**, UU, UA	6.4 D
Cmr6 (Q53W04)	^{169}LHPDILNPHHPDYYGSVK^{186}	-	**UU**	6.4 E
Reconstituted complex				
Protein (Uniprot ID)	**Peptide**	**Amino Acid**	**RNA**	**Figure**
Cmr1 (Q53W07)	^{34}TYTLLTPLFGGGVEPREADPVSVVR^{58}	-	U, **UC**	6.4 F
Cmr2 (Q53W09)	^{160}DFAPVSWGSPAYK^{172}	Y^{171}	U	-
Cmr3 (Q53W08)	^{156}GHEGPVPETRTHVALDPAAQTAR^{178}	-	UC	-
	^{166}THVALDPAAQTAR^{178}	L^{170}	U, UU, UA, UUA, UGC	-
Cmr4 (Q53W06)	^{204}IRLDDETK^{211}	L^{206}	UU, UUA	-

Protein: Cross-linked protein (Uniprot ID); Peptide: Sequence of the cross-linked peptide, specified with the position of the peptide in the protein sequence; Amino acid: One-letter code specified with the position of cross-linked amino acid, RNA: composition of the RNAs observed cross-linked.
For peptides cross-linked to mono- di- or tri-nucleotides depicted in bold, the corresponding MS/MS spectra are given in the Appendix, Figure 6.4.

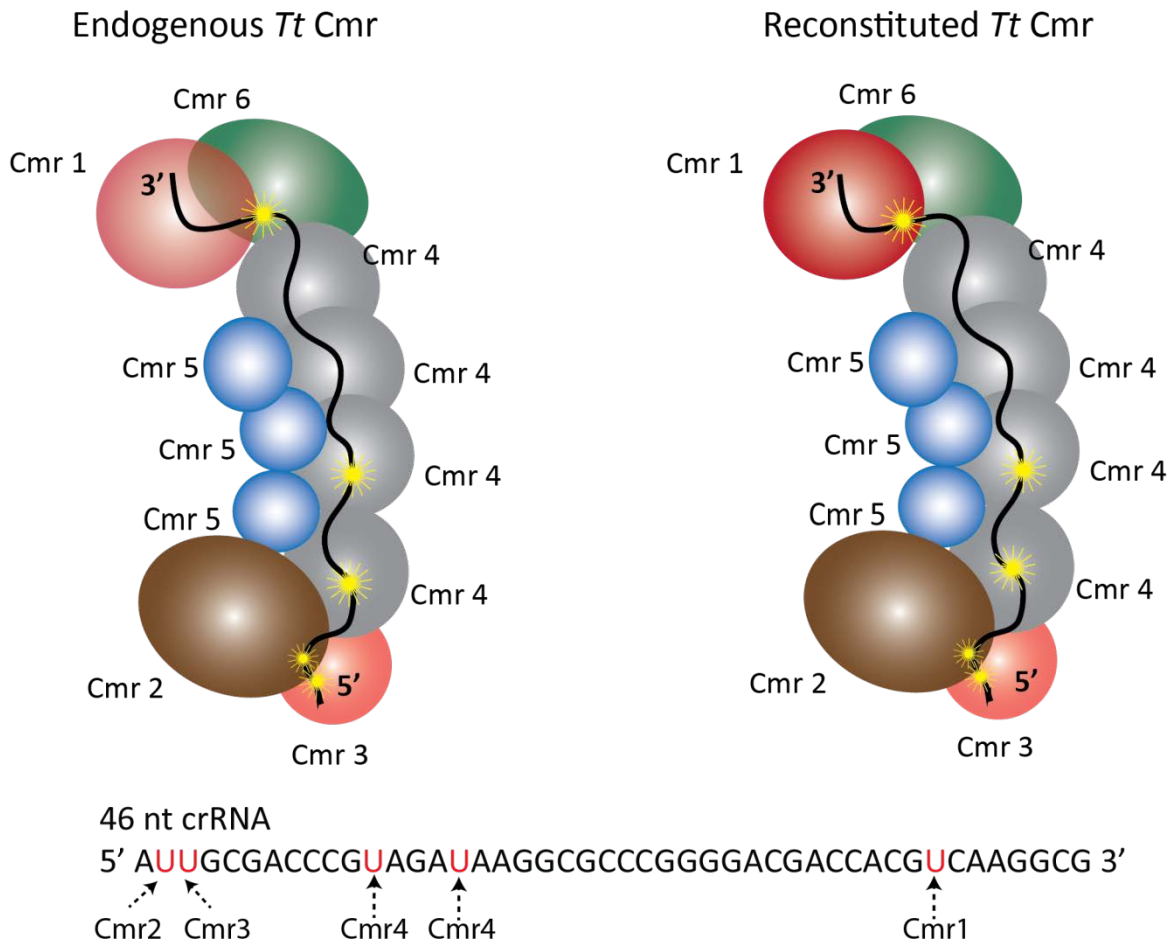


Figure 3.14 Schematic representation of the cross-linked regions mapped on a model of the *Tt* Cmr complex.

Model for the *Tt* Cmr complex based on the stoichiometry and structural organization of the sea-worm model reported in [52]. The cross-linked residues identified are mapped on the possible cross-link sites (yellow star) in the complex based on the position of the uracil residues in the crRNA sequence (in red). The Cmr subunits predicted to cross-link to different uracil residues are indicated below. The Cmr1 in the endogenous complex is depicted as transparent red shape because it is speculated that it is loosely attached to the complex.

In each of the Crm2, Cmr3 and Cmr4 proteins only one cross-linked region/residue was identified. Depending on the location of the Cmr subunits in the EM structure and the sequence of the crRNA these results were mapped on a model of the Cmr complex (Figure 3.14). In Cmr2, Y^{171} was observed cross-linked to U, which could be any of the uracil residues close to the 5' end of the crRNA as Cmr2 is the tail protein present close to the 5' end of the crRNA. In Cmr3, L^{170} was observed cross-linked to different RNA moieties, with uracil as the cross-linked nucleotides (Table 3.8). The composition of these RNA moieties indicates that the cross-linked uracil could be either of the U^2 or U^3 residues at the 5' end. In Cmr4, L^{206} was observed cross-linked to uracil. As Cmr4 is present in four copies in the complex, the cross-linked residue could

not be unambiguously assigned to one of the four copies. Considering that there are only five uracil nucleotides in the entire crRNA which are not uniformly distributed, it can be speculated that only the Cmr subunits in close proximity to these uracil residues are likely to cross-link as shown in a model of the *Tt* Cmr complex in Figure 3.14.

In addition to the above mentioned cross-links, one cross-linked region was observed in the Cmr1 subunit of the reconstituted complex. The peptide ^{34}TYTLLTPLFGGGVEPREADPVSVVR^{58} was cross-linked to a UC RNA dinucleotide, where uracil was the cross-linked nucleotide. According to the crRNA sequence U^{39} is the only uracil with a cytosine residue adjacent to it and it is in proximity to the 3' end of the crRNA which is the position of Cmr1 in the *Tt* Cmr complex as reported in the EM structure [52].

One significant difference observed between the two complexes was the cross-linked protein region ^{169}LHPDILNPHHPDYYGSVK^{186} in Cmr6 protein of the endogenous complex and the cross-linked protein region ^{34}TYTLLTPLFGGGVEPREADPVSVVR^{58} in Cmr1 protein in the reconstituted complex. With native MS analysis on the *Tt* Cmr complex it has been shown that the Cmr1 subunit is located at the periphery of the complex and has a very lose attachment with the complex under *in vitro* conditions [52]. This gives rise to the ambiguity in identifying the subunit cross-linked to the 3' end of the crRNA.

The results from this study show high reproducibility between the endogenous and reconstituted Cmr complexes, and prove this approach as suitable for studying other complexes of similar complexity, size and affinity.

3.3 Quantitative and structural investigation of the Type I-B Cascade complex from *C. thermocellum*

After establishing the protein-RNA cross-linking workflow for the identification of protein-RNA interactions in the prokaryotic crRNP complexes, the structural investigations were further extended to determine the interactions within different protein subunits constituting the crRNP complex. The structural investigation was carried out in combination with a quantitative analysis to estimate the absolute amount of the protein subunits present in the complex. In this way the interactions identified by MS analysis could be explained with comparison to the stoichiometry of the different protein subunits in the complex.

The project was undertaken in collaboration with Judith Zöphel and Prof. Lennart Randau (MPI Terrestrial Microbiology, Marburg). It focused on performing quantitative analysis using iBAQ and investigating the interactions between different protein subunits using protein-protein cross-linking in Type I-B Cascade-like complex in *Clostridium thermocellum*_3205.

Type I-B Cascade complex is one of the least investigated assemblies of the CRISPR-Cas systems. The complex used in this study consists of four Cas proteins, Cas5, Cas6, Cas7 and Cas8b in an unknown stoichiometry. The presence of the signature protein Cas8b, classifies this complex as a Type I-B crRNP complex [21, 22]. For the complex assembly, different components of the complex were cloned and expressed separately. The entire complex was reconstituted *in vitro* by addition of the crRNA and the four Cas proteins one at a time, further purification details of the assembly and purification of Cascade complex are described in Section 2.2.1.8. The purified assembled complex was provided by Judith Zöphel.

3.3.1 Stoichiometry determination in the *C. thermocellum* Cascade complex

To investigate the stoichiometry of Cas5, Cas6, Cas7 and Cas8b proteins in the complex, the label-free absolute quantification approach (iBAQ) was performed using the universal protein standard UPS2. The same workflow was used for iBAQ as described under Section 3.1.2 for the *H. volcanii* Type I-B Cascade complex. All the proteins that were identified and quantified from three technical replicates are listed under Table 6.1 in the Appendix. iBAQ values for three replicates of the standard proteins were averaged. A calibration curve was prepared using the known amounts of standard proteins in UPS2 and the average iBAQ values. A double logarithmic plot where the log (amount) of the standard proteins identified was plotted against

the log iBAQ value. The concentration of Cas5, Cas6, Cas7 and Cas8b proteins were calculated using the log/log linear regression (Appendix, Figure 6.5).

The derived protein concentrations indicate a Cas5:Cas6:Cas8b:Cas7 stoichiometry of 1:1:2.5:6 (Appendix, Table 6.1). This low:low:high ratio for the Cas5, cas6 and Cas7 stoichiometry is in agreement with previously observed stoichiometry for Cascade-type protein complexes in *H. volcanii* [89], *E. coli* [30] and *P. aeruginosa* [40].

In addition, from the stoichiometry of 2.5 observed for Cas8b protein, it can be assumed that there are two Cas8b subunits in the Cascade complex and one truncated fragment corresponding to half a Cas8b subunit. However, when the purified preparation of Cascade complex was analyzed by SDS-PAGE (Figure 3.15, Lane1), Cas8b was observed as two bands after coomassie staining, one corresponding to the full length version at 72 kDa and a short fragment of ~15kDa (Cas8b*) which was also confirmed as Cas8b by MS analysis. This conspicuous nature of Cas8b is currently under investigation (Randau Lab, unpublished data).

In other Type I CRISPR systems, it has been reported that the large subunit protein (Cas8) is present in one single copy as depicted in a schematic representation of the Type I crRNP complexes in Figure 1.4.

3.3.2 Protein-protein cross-linking in the *C. thermocellum* Cascade complex

The goal of this investigation was to characterize protein-protein interactions within the Cas proteins in the Type I-B Cascade complex from *C. thermocellum*. The project sought to evaluate the feasibility of chemical protein-protein cross-linking for the investigation of purified crRNP complexes. Although there is not enough protein-protein interaction data available to validate the results of this study, the information available from the common architectural features of the Type I crRNP complexes was used to explain the outcome from protein-protein cross-linking study.

Chemical cross-linking was performed using BS3, which targets primary amines in the side chains of lysine residues on protein surface as well as protein N-termini. In order to optimize the cross-linker:protein ratio to be used for cross-linking, the purified Cascade complex was incubated with increasing molar excesses of BS3 cross-linker in the cross-linker to protein ratio 5:1, 10:1, 25:1, 50:1 , 100:1 and 200:1. The results of this titration were analyzed with SDS-

PAGE as shown in Figure 3.15. The details of the cross-linking reaction are described under Section 2.2.7.

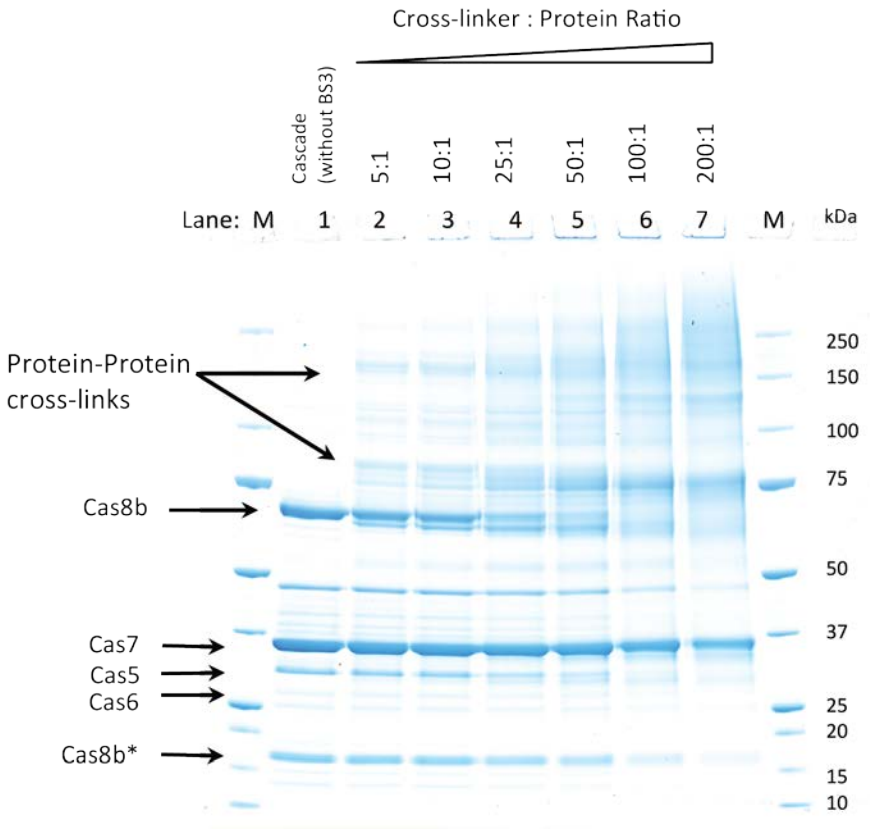


Figure 3.15 Analysis of protein-protein cross-linking by SDS-PAGE.

Coomassie stained SDS-PAGE gel showing the results of protein-protein cross-linking performed in the Cascade complex using different concentration of BS3 cross-linker. Lane 1: The purified Cascade complex alone (without any cross-linker) consisted of Cas7 (35 kDa), Cas5 (28 kDa), Cas6 (26 kDa) and two fragments for Cas8b the larger fragment 72 kDa and a smaller ~15 kDa fragment Cas8b*. Lane 2-7: Optimization of cross-linker to protein complex ratio. As the molar excess of BS3 over Cascade increases, the intensity of bands corresponding to Cas proteins diminishes. At the same time bands corresponding to protein-protein cross-links appear with even 5:1 cross-linker to protein ratio. The ratio 100:1 and later corresponds to high molecular aggregates which can be neglected. lM: Protein marker (BioRad).

Upon the SDS-PAGE analysis it was observed that the purified Cascade complex from *C. thermocellum* comprised four Cas proteins Cas7, Cas5, Cas6 and Cas8b (Figure 3.15 Lane 1). Throughout the titration the protein-protein cross-links could be observed in all the cases where even a minimal amount (5 molar excess) of BS3 cross-linker was present. With increasing amounts of cross-linker the band corresponding to protein-protein cross-links also enhanced (Figure 3.15 Lane 2-7). At very high cross-linker amounts there were higher order aggregates and artefacts appearing as a smear at high molecular weight (Figure 3.15 Lane 6 and 7).

For the MS analysis the cross-linker to protein ratio of 75:1 was selected as the optimal ratio for the final cross-linking experiment. This ratio was selected to avoid the complex high-order aggregates and at the same time to get enough protein-protein cross-links, for the MS analysis. The cross-linked complexes were hydrolyzed with trypsin and the resulting peptide mixture was further enriched for cross-linked peptides using size exclusion chromatography (Figure 1.13), details in section 2.2.7.3.

The enriched cross-links were analyzed by LC-MS/MS in DDA mode on an Orbitrap Fusion instrument and the cross-linked peptides were identified using pLink, using 1% FDR and the results are presented in a cross-link map (Figure 3.16) after applying a score cut-off of 2.0 (based on the p-value). The spectra for the cross-links identified were also manually checked to confirm the assigned cross-link. Further the cross-link species such as mono-links or loop links (Figure 1.13) were filtered out from the final list of cross-links, as they carry limited information not significant in the investigation of protein-protein interactions. In the end only unique inter-protein and intra- protein cross-links are reported in this work.

3.3.2.1 Inter-protein cross-links identified for Cas5, Cas6, Cas7 and Cas8b proteins

From the MS analysis 126 unique inter-protein cross-links shown in Figure 3.16 were identified within the four Cas proteins (Appendix, Table 6.2). However due to a lack of structures available for these proteins, it was not possible to map the cross-link sites on three dimensional structures. Nonetheless, the cross-links identified in the Cas proteins were in agreement to the Cas protein stoichiometry and the position of a particular Cas protein in the Type I Cascade complexes [22], as determined in recent studies on the architecture and organization of *E. coli* Cascade complex [105, 106].

Two inter-protein cross-links were identified for the Cas6 protein, K^{22} was observed cross-linked to K^{155} in the N-terminal region of Cas7 protein and K^{106} was observed cross-linked to K^{73} in the N-terminal region of Cas8b. The Cas6 protein forms the head of an assembled Cascade complex interacting with the 3’ end of crRNA, as reported earlier in the Type I-E *E. coli* Cascade complex. Further, the head protein interacts with the first Cas7 protein that forms the backbone of a Cascade complex and also the large subunit Cas8b which spans the entire length of the complex. In addition no cross-link between Cas6 and Cas5 (the tail protein) was observed, which is in agreement with the positioning of Cas5 and Cas6 proteins in the Cascade complex as

shown in (Figure 1.4 and 1.15) i.e., the head and tail proteins are distant enough for cross-linking via BS3 to occur.

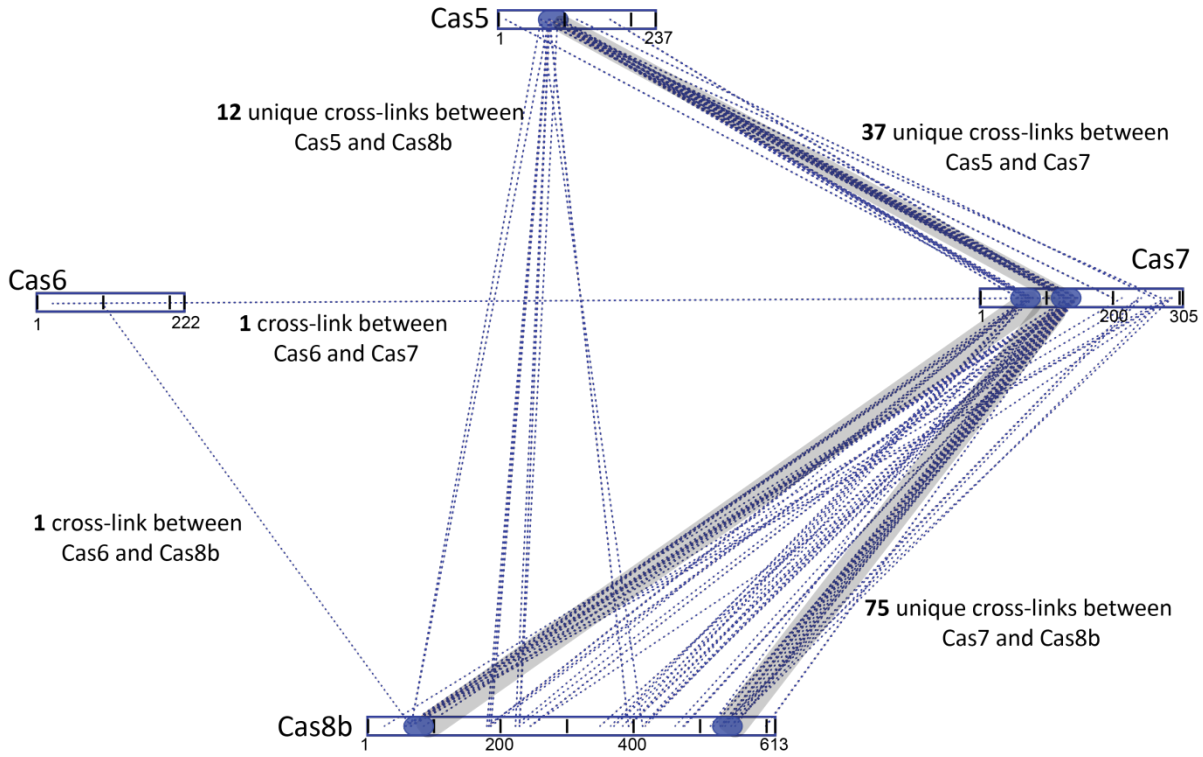


Figure 3.16 Protein-protein cross-linking map for the *C. thermocellum* Cascade complex.

A protein-protein cross-link map made using xiNET. Only the inter-protein cross-links identified for Cas6, Cas5, Cas7 and Cas8b proteins are shown. The cross-links were identified using pLink with 1% FDR. The cross-links indicated are the one remaining after applying a score cut off 2.0. The gray shaded regions indicate high density of cross-links between those regions. The blue shaded regions indicate lysine rich regions of the proteins with a high density of cross-links observed for the lysine residues. Further details of all the identified cross-links are provided in the Appendix (Table 6.2)

In Cas5 a higher number of cross-links were identified. There were 12 unique cross-links identified between Cas5 and Cas8b and 37 between Cas5 and Cas7. Most of these cross-links were confined to a lysine rich region in Cas5 with four residues K^{72}, K^{84}, K^{85} and K^{97} as shown in Figure 3.16. Cas5 is a part of the tail assembly in Cascade complex [105] which comprises Cas5, Cas7 and in this case Cas8b subunits.

Between Cas7 and Cas8b proteins 75 unique cross-links were identified corresponding to lysine residues distributed throughout the protein sequence. In both Cas7 and Cas8b a very high number of cross-links were identified in the two lysine rich regions of both the proteins as

shown in Figure 3.16. The high number of cross-links identified is in agreement with the stoichiometry determined. Both the proteins are present in more than one copy; therefore the cross-links identified cannot be assigned to any one subunit of the two proteins. Rather it can only be speculated that the regions with high density of cross-links are in close spatial proximity, likely to form cross-links in the presence of a chemical cross-linker such as BS3. In addition, the Cas7 protein has been reported to form the helical backbone of the Cascade-like complexes and is supported by the large subunit Cas8b throughout the length of the complex (Figure 1.4), therefore so many protein-protein interactions between the two proteins are likely to occur.

3.3.2.2 Intra-protein cross-links identified for Cas5, Cas6, Cas7 and Cas8b proteins

In addition to the inter-protein cross-links described above, a large number of intra-protein cross-links were also identified for all the four Cas proteins. A total of 16 unique intra-protein cross-links were observed in Cas5 and six in Cas6 (Appendix, Table 6.3 and 6.4). Only for these two proteins Cas5 and Cas6 protein which are present in a single copy in the Cascade complex, the identified cross-links could be mapped unambiguously on the protein sequence (Appendix, Figure 6.6).

In Cas7 there were 112 unique intra-protein cross-links and for Cas8b there were 230 unique intra-protein cross-links. However these numbers cannot be assigned to any one subunit of the two proteins in particular. Considering the stoichiometry of the two proteins, there is a high ambiguity in mapping these cross-link sites to the protein sequence and assigning them to different protein subunits. Hence further details of the intra-protein cross-links identified for Cas7 and Cas8b are not included in this work.

These protein-protein cross-linking studies provide the first experimental evidence for the protein-protein interactions in Type I-B Cascade complex from *C. thermocellum* and can be used in the development of a structural model.

4. Discussion

The work presented in this thesis focused on two main mass-spectrometric approaches in the investigation of prokaryotic immune defense system. (i) Quantitative approach to compare the proteomes of wild-type strains with those of strains that carry deletion mutants of Cas proteins and to determine the stoichiometry of Cas proteins in a multi-subunit crRNP complex. (ii) Structural mass-spectrometric approach to elucidate protein–RNA contact sites among the various single and multi-subunit crRNP complexes from prokaryotes using UV induced cross-linking and a lysine directed cross-linking approach to look into the protein-protein interactions in Type I-B Cascade complex from *C. thermocellum*. In this chapter the applicability, advantages and limitations of both mass spectrometric approaches are discussed in conjunction with the implications of results obtained in the different CRISPR-Cas systems.

4.1 Quantitative approach for the investigation of CRISPR-Cas system

The two methods used for quantitative analysis in this work allow for a comprehensive evaluation of quantitative proteomics using relative quantification at the proteome level and absolute quantification at the protein level.

4.1.1 Relative quantification using dimethyl labeling to investigate the effect of *cas7* deletion on other Cas proteins in *H. volcanii*

The quantitative MS based proteome study of *H. volcanii* revealed how the deletion of *cas7* gene influences the expression of other Cas proteins and the entire cellular proteome. Using differential stable isotope dimethyl labeling, approximately 1800 proteins were identified and quantified in both forward and reverse experiments. The level of Cas7 protein in the *cas7* deletion strain was observed to be the most downregulated (Table 3.1). The normalized ratio H/L of Cas7 in both the experiments indicated a residual level of Cas7 protein against the background noise. The quantitation approach showed that the gene deletion was effective as only negligible amounts of protein were detected. Therefore the expression of the gene was efficiently suppressed.

One of the proteins remarkably affected by the *cas7* deletion was the Cas5 protein. It was the only Cas protein other than Cas7 that was observed to be downregulated. It has been recently

reported that the proteins Cas5, Cas6 and Cas7 form together a Cascade-like complex in *H. volcanii* for the processing and stabilization of crRNA [89]. Due to the association between Cas7 and Cas5 protein it can be proposed that in the wild-type strain Cas5 is a part of multi-subunit protein complex, hence it might be protected from degradation by cellular proteases. However in the deletion strain where an integral part of the Cascade complex, the Cas7 protein is missing, it can be speculated that the Cas5 protein is degraded by proteases and therefore it appears downregulated with the relative quantification analysis. Additionally, the *cas5* gene is located just 11 bp downstream of the *cas7* gene in the CRISPR loci (Figure 1.14). The removal of *cas7* sequence could mean that the possible additional promoters [152] in the *cas7* gene were also removed which adversely affects the production of Cas5 protein. The possible additional promoters in *cas7* gene are currently under investigation in the Marchfelder lab, University of Ulm.

The Cas proteins which were upregulated upon *cas7* deletion, included Cas1, Cas8 and Cas6 (Table 3.2). In a nutshell all the Cas proteins forming an integral part of the Cascade complex in *H. volcanii* were upregulated upon deletion of *cas7* gene, except the Cas5. The CRISPR-Cas system plays a very significant role in the prokaryotic immune defense. The removal of *cas7* is compensated by increased expression level of the *cas* gene cluster to make up for the effectiveness of the immune system. The Cas1 protein was the most upregulated protein, which is expected because it is an endonuclease which in conjunction with the Cas2 is responsible for the selection of a new protospacer and its incorporation as a spacer in the genome in the starting (adaptation) phase of the CRISPR based immune defense [23]. In addition, the archaeal specific helicase protein AshA was also observed to be upregulated upon *cas7* deletion. The protein has not yet been shown to have a direct association with the CRISPR-Cas system, interference tests performed with a deletion strain of this helicase resulted an inadequate immune response against artificially generated invader plasmids (Marchfelder lab, unpublished observation).

Other Cas proteins in *H. volcanii*, Cas2, Cas3 and Cas4 were not identified in this experiment, which could correspond to the levels of these proteins being below the limit of detection of the method used. A specialized workflow has been developed earlier in our group, for in-depth proteome analysis using 1D-PAGE coupled to pIEF prior to LC-MS/MS, resulting in increased protein identification with higher sequence coverage [153]. A similar approach was also

developed for the *H. volcanii* in-depth proteome analysis, to look especially for low abundant Cas proteins using the in-depth proteomics approach recently established in the Urlaub lab [153].

Differential stable isotope labeling using dimethyl labels was used for the relative quantification study in this thesis. Stable isotope dimethyl labeling performed in-solution is best suited for experiments where more than one samples have to be labeled because it is straightforward, easy and cheap for the comparison of two or three different samples [84]. 99% percent of the proteins quantified in this experiment, had a ratio close to 1:1 for the H119 wild-type and the Δ*cas7KO* deletion strain in both forward and reverse experiments. This demonstrates a good reproducibility between both the forward and reverse experiments and that the majority of the proteome is unaffected by the *cas7* deletion.

As the protein samples were derived from an archaeon and there was no auxotrophic strain available for the incorporation of isotopic labels at the cellular level, popular quantification methods such as SILAC could not be used in this study. In addition, the labeling was performed at the peptide level, and not at the protein level to allow proteolysis with trypsin which in latter situation would not be able to cleave the modified lysine residues. Lastly, the dimethyl labeling is a faster and cheaper alternative to the strategies routinely used for relative quantification.

4.1.2 Absolute quantification using iBAQ to determine the stoichiometry of Cas proteins in *H. volcanii and C. thermocellum* Cascade complex

In this thesis the absolute quantification experiments were performed using iBAQ to determine the absolute amount (copy number) of Cas proteins in different multi-subunit protein complexes assembled around a crRNA (Cascade-like complexes) from *H. volcanii* and *C. thermocellum* [89, 127]. These CRISPR associated complexes for antiviral defense are a characteristic feature of Type I CRISPR-Cas systems. A well-organized assembly of multiple Cas proteins with individual components in different copy numbers. For any investigation related to the identification of protein-RNA or protein-protein interactions between different components of these complexes and to gain insights into the architectural organization of these complexes it is a prerequisite to know the copy number of different proteins present in the complex. With the recent discoveries of a Cascade-like complex in *H. volcanii* and *C. thermocellum* and limited information available on the copy number of individual proteins, the iBAQ approach was used to determine the absolute amount of Cas proteins in these complexes.

The iBAQ approach was used as the method for absolute quantification for two reasons: (i) it is inexpensive in comparison to the use of labeled standard peptides in the AQUA (Absolute quantification) method and (ii) it does not require a dedicated MS instrumental set-up or method development like in AQUA and/or native MS experiments. However iBAQ exhibits limited accuracy in the determination of high stoichiometries in protein complexes due to a lower accuracy in the determination of high protein copy numbers [87, 89].

In *H. volcanii* type I-B system a Cascade-like complex was observed, comprised of at least Cas5, Cas6 and Cas7 proteins and a crRNA. The purification of FLAG-Cas7 revealed two potential interacting partners (Figure 3.3 B) which were identified as Cas5 and Cas6 proteins with MS analysis. The co-purification of these proteins indicates the presence of these Cas proteins in the *Haloferax* Cascade complex. The absolute amounts of proteins derived from the iBAQ analysis show a Ca6:Cas5:Cas7 stoichiometry 1:1.7:8.5 (Table 3.3). The low:low:high stoichiometry fits very well to the previous observations in the Cascade complexes from Type I-A, Type I-E and Type I-F systems [19, 30, 41]. In all these complexes the Cas7 protein is present in multiple copies and forms the backbone of the complex.

In *C. thermocellum* the Cascade-like complex was observed to comprise of four proteins, Cas5, Cas6, Cas7 and the large subunit Cas8b which is the signature protein of Type I-B systems. The protein components were also identified and confirmed with MS analysis. The absolute amounts of proteins derived from the iBAQ analysis show a Cas6:Cas5:Cas8b:Cas7 stoichiometry 1:1:2.5:6 (Appendix, Table 6.1). The ratio of Cas5, Cas6 and Cas7 was similar to the previous observations in different Cascade complexes as described above. However, the stoichiometry of 2.5 for Cas8b is in contrast to the proposed models for Type I-A, I-C, I-E and I-F Cascade complexes where the large subunit Cas8b appears as a single copy [22]. This conspicuous behaviour of Cas8b was also observed during the SDS-PAGE analysis (Figure 3.15) where the protein was present as a full 72 kDa protein and as a short ~15 kDa fragment. The corresponding bands for both protein fragments from the SDS-PAGE were identified as Cas8b with the MS analysis. This observation is still under investigation in the Randau lab.

Nonetheless, the 2.5 stoichiometry of Cas8b can be confirmed with a widely used absolute quantification approach, AQUA. It is a targeted approach that makes use of chemically synthesized peptide containing stable isotopes which are added to the protein sample as internal standards and the absolute amount of proteins (comprising these peptides) can be

estimated in a complex mixture [154, 155]. Likewise, two specific labeled peptides can be used for determining the absolute amount of Cas8b, with one peptide against the C-terminal part of the protein which is speculated to form the short fragment and the second peptide against the N-terminal part of the protein. However, a drawback of this method would be the very high cost of the labeled peptide standards, requirement of a dedicated MS instrument set-up and in the end only two peptides per protein would be quantified.

4.2 CRISPR-Cas: a mass spectrometry based structural perspective

The three major types of CRSIPR-Cas systems are divided into eleven subtypes on the basis of organization of CRISPR locus and the *cas* gene content [21]. Each of these comprises a different crRNP complex composed of Cas proteins and a crRNA. The building-blocks of these complexes are the Cas proteins which vary in their copy numbers in the different complexes. The fully assembled complexes are mechanistically very diverse in the three types of CRISPR-Cas systems. The Type I and Type III systems have multi-subunit assemblies resembling a sea-horse structure as in the Type I-E *E. coli* Cascade [43] or a sea-worm structure as in the Type III-B *T. thermophilus* Cmr complex [52], in contrast to a single multi-functional protein Cas9 in the Type II system. However for all the crRNP complexes the architectural assembly has one thing in common, that is the assembly of Cas proteins around a crRNA which acts as a guide for the target recognition by the crRNP complex. Moreover, they provide an elaborate platform for the investigation of protein-RNA interactions as reported in the recently co-crystallized *E. coli* Cascade complex comprising 11 Cas protein subunits assembled around a crRNA [105, 106]. The proteins Cas5, Cas6 and Cas7 contain RNA binding domains such as modified RRMs, belong to the RAMP (repeat associated mysterious protein) superfamily and have been shown to interact with the crRNA [18, 19, 29, 30, 105]. The structural proteomics studies performed in this work using UV induced protein-RNA cross-linking and MS allowed the identification of RNA-binding sites on a peptide or amino acid level and helped validate some of these interactions. In addition, new RNA binding regions were identified in these proteins which were not a part of the characteristic RRMs. The details of these investigations are discussed below with respect to protein-RNA cross-linking in the individual (recombinant) Cas protein and their cognate crRNA and the entire multi-subunit crRNP complexes. In general, the considerations and potentials of the protein-RNA cross-linking approach are discussed later in a separate Section 4.3.

4.2.1 Cas6b-crRNA cross-linking

Cas6 is one of the most widely distributed Cas proteins present in both bacteria and archaea. The Cas6 proteins are endoribonucleases that catalyze the reaction of pre-crRNA cleavage into smaller processed crRNAs [39]. The cleavage reaction yields mature crRNA with a 5' terminal tag of 8 nucleotides derived from the repeat [39, 127, 146]. The crRNA processing activity of all the Cas6 proteins is similar but the sequence similarity in these proteins is limited. They typically contain two RAMP domains (ferredoxin-like folds, similar to RRM domains) with a glycine-rich loop located in between [39, 156, 157]. The sequence of this glycine-rich loop often fits the consensus sequence GΦGXXXXXGΦG, where Φ is a hydrophobic residue and X is any residue with the variable region containing at least one positively charged residue [15, 24]. An alignment for the Cas6b homologues using Clustal W also indicated the presence of a conserved glycine rich loop (Figure 3.6) in the two Cas6b proteins used in this work. In both prokaryotes and eukaryotes such glycine rich domains like the G-patch domains have been reported to be involved in RNA-interactions [100, 158, 159]. However, the cross-linked residues identified for the two Cas6b proteins were located in close proximity to the G-rich loop but were not an integral part of the loop.

In the Cas6b protein from archaeon *M. maripaludis*, M^{185} residue in its oxidized state was identified cross-linked to U^{15} residue on the cognate crRNA (Figure 3.5 and 3.6). To further confirm that M^{185} played an important role in interaction with the crRNA, mutation studies were performed on the *Mm* Cas6b by Hagen Richter from the group of Prof. Lennart Randau, MPI Terrestrial Microbiology, Marburg. The Cas6 protein with a single point mutation of M185A was tested for RNA-binding by electrophoretic mobility shift assays (EMSAs) which exhibited a significant decrease in the RNA affinity of the mutated protein and the RNA binding was considerably affected (Appendix, Figure 6.7). The experiment was performed as a qualitative analysis only to confirm the efficacy of UV induced cross-linking in identifying protein-RNA interaction sites. Meanwhile, our collaborators were successful in the co-crystallization of *Mm* Cas6b with the cognate crRNA. The co-crystal structure showed that the motif II of the crRNA repeat recognized by the *Mm* Cas6b, comprises a two base pair stem (G^{16}-C^{29}/C^{17}-G^{28}) and an adenine-rich loop (A^{18}-A^{27}). The major groove edge of U^{15}, forms a polar interface with the protein at which the amino acid residue R^{206} of G-rich loop interacts with the C^{29} nucleotide residue on the crRNA and the amino acid residue L^{124}, with U^{15} nucleotide of the crRNA. The

base of U^{15} nucleotide stacks with M^{185} amino acid residue (Randau lab, unpublished data) in agreement to the protein-RNA cross-linking results.

In the Cas6b protein from bacterium *C. thermocellum*, M^{184} was identified as the cross-linked amino acid residue. However on the crRNA the cross-linked nucleotide could not be identified unambiguously and both U^3 and U^{32} from the 5'end were identified as the possible candidates (Figure 3.6). The peptide ^{184}MIGFK^{188} was observed cross-linked to UGA and UUG and in both RNA moieties uracil was the cross-linked nucleotide. When mapped on to the crRNA sequence these correspond to the U^3 and U^{32} residues. In this case a cross-link with a longer RNA moiety would have been more informative to map the correct uracil residue on the crRNA sequence. However the method is developed for identification of the protein region in contact with RNA and the presence of longer RNA moieties suppresses the intensity of peptide fragment. The RNA length comes at the expense of the identification of the peptide and therefore extensive RNAse digestion is performed so that the RNA fragments are not longer than 2-3 nucleotides. Further details regarding potentials of UV induced cross-linking are discussed under Section 4.3. Despite low primary sequence conservation between Cas6 proteins, there is a structural homology between *Mm* Cas6b and *Pyrococcus furiosus* (*Pf*) Cas6 [127]. Both the archaeal and bacterial Cas6b proteins were modeled based on the *Pf* Cas6 crystal structure (Figure 3.6) and when the two models were aligned structurally the identified cross-linked residues were mapped onto the same location in both proteins (Figure 3.6). The results from sequence alignment for Cas6b homologues indicated the presence of a conserved glycine-rich region near the C-terminal of the protein, which constitutes the glycine-rich loop that is known for its interaction with the RNA. In addition, the methionine residue identified as the cross-linked amino acid in both *Mm* Cas6b and *Ct* Cas6b proteins is also conserved across the different archaeal and bacterial Cas6 (Figure 3.6). Lastly, the co-crystallization studies in *Mm* Cas6b indicate that in addition to the glycine-rich loop, an evolutionary conserved RNA binding region comprising the M^{185} amino acid residue also interacts with the crRNA.

4.2.2 Protein-RNA interactions in the Cas7 protein family

In addition to the Cas6 protein family, the Cas7 family also belongs to one of the best understood Cas proteins both structurally and functionally. Four different Cas7 proteins were used for this investigation. Two recombinant Cas7 proteins, Type I-A Cas7 from *T. tenax* and Type I-D Cas7 (Csc2) from *T. pendens* were cross-linked to polyU RNA. Both these proteins have

been observed before to bind both polyU RNA as well as the cognate crRNA [103, 131]. However due to the unavailability of crRNA, the cross-linking was performed with $polyU_{(15)}$. The results were compared to the cross-linked regions identified in Cas7 proteins belonging to the fully assembled crRNP complexes. These included Type I-E Cas7 in the Cascade complex from *E. coli* and Type III-A Cas7 (Csm3) in the Csm complex from *T. thermophilus*, the cross-linking was performed on fully assembled crRNP complexes where these Cas proteins were bound in multiple copies to a single crRNA (Figure 3.7 and 3.8).

The identified cross-links were mapped onto the crystal structure of *E. coli* Cas7 (PDB: 1VY8) [105] and *T. pendens* Csc2 (PDB ID: 4TXD) [103] and to the predicted 3D-structure models of *T. tenax* Cas7 and *T. thermophilus* Csm3, generated using Phyre2 [141]. The results were compared with the crRNA-binding surface of Type I-E *E. coli* Cas7, which was crystallized in context of the fully assembled crRNP complex from *E. coli* [105]. The crystal structure of *T. pendens* Csc2 and the homology models (*T. tenax* Cas7 and *T. thermophilus* Csm3) were superposed onto two copies of *E. coli* Cas7 (PDB ID: 1VY8) using secondary-structure matching (SSM) superposition in COOT [142]. The structure of two copies of *E. coli* Cas7 bound to crRNA was used for the superposition (Figure 4.1).

***E. coli* Cas7**

In the *E. coli* Cas7 protein, due to a right hand shaped structure, different structural domains are referred to as thumb, finger and palm domains (Figure 3.10), however to follow a uniform nomenclature for all the Cas7 homologues these domains are named here as insertion domain1, insertion domain2 and the central cleft respectively. Two out of seven cross-links amino acid residues K^{27} and M^{166} were present in the central cleft region of the protein. The universal central cleft is a positively charged groove on the protein surface and is important in mediating the binding to RNA as also observed in *T. pendens* Csc2 [103] and *T. tenax* Cas7. However, five out of seven cross-linked amino acid residues identified K^{86}, K^{98}, K^{137}, K^{141} and K^{144} were present in the finger domain *i.e.*, the insertion domain 2 of the Cas7 protein (Figure 4.1). All the cross-links were observed in the lysine-rich helix. Although there have been no earlier reports for RNA interaction in this domain, the results described here provide a strong indication that this region also has a role in RNA interaction. Moreover, it should be noted that the Cascade complex was cross-linked in solution and therefore there is a strong possibility that in solution the RNA can flip over the Cas7 protein, interacting with these lysine residues in the insertion domain 2, allowing in this manner for the identification of novel RNA binding regions.

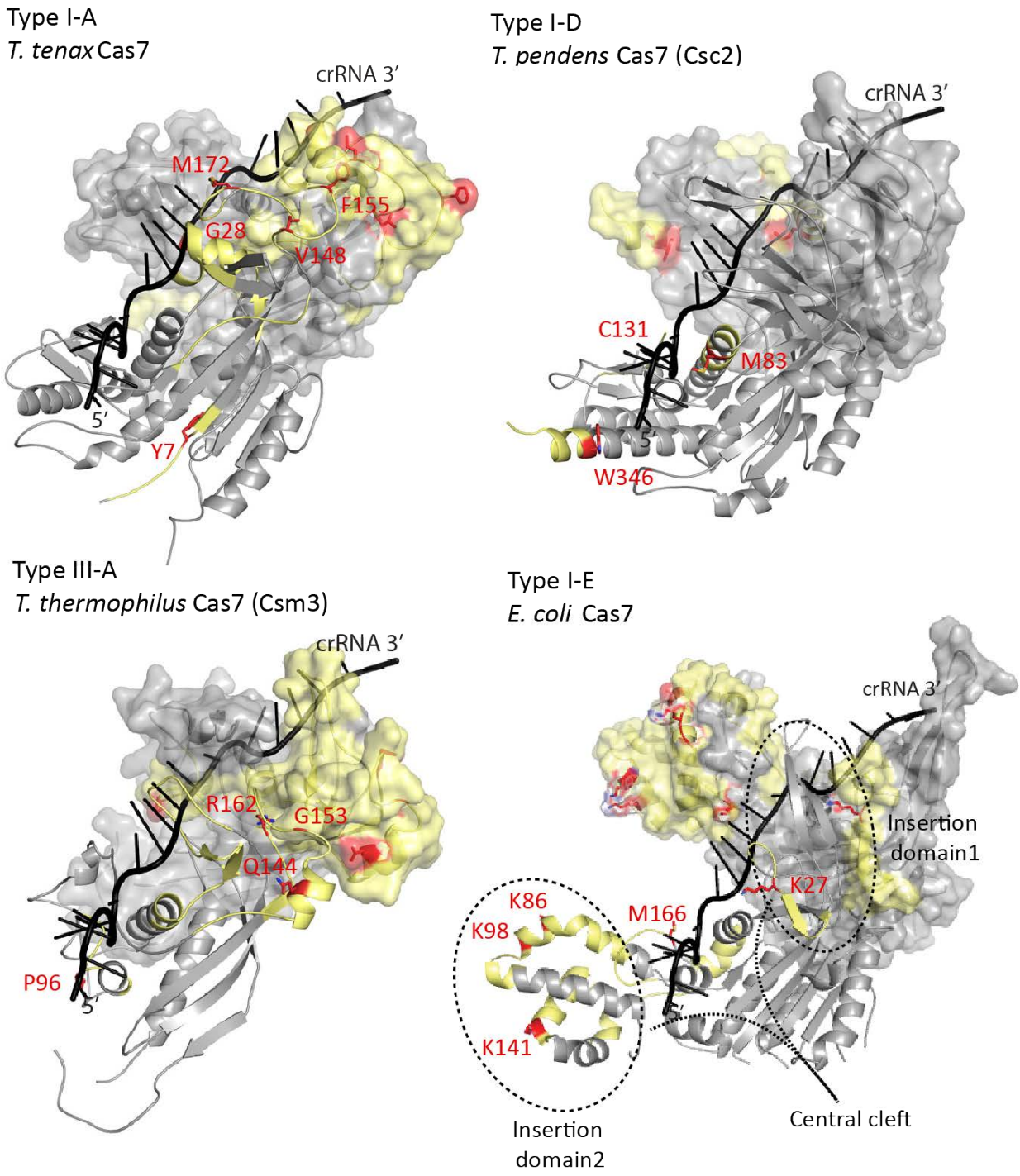


Figure 4.1 Mapping the protein-RNA cross-links identified in different Cas7 proteins to the crystal and modeled structures.

Predicted Cas7 3D-models (Type I-A *T. tenax* Cas7 and Type III-A *T. thermophilus* Cas7) and Type I-D *T. pendens* Cas7 (Csc2) crystal structure (PDB ID: 4TXD) [103] superposed to two copies of crRNA-bound Type I-E Cas7 (PDB ID: 1VY8) [105]. The front copy is shown in cartoon, the back copy additionally in surface representation. The crRNA is colored black, the protein structures are colored grey, the cross-linked peptide yellow and cross-linked sites are highlighted in red. The universally present central cleft defined by the RRM and the insertion domain 1 and 2 (circles) are labeled in Type I-E Cas7. Modified from Sharma, K. *et al.,* manuscript under revision.

***T. pendens* Csc2**

In Type I-D *T. pendens* Csc2 all the three cross-linked amino acid residues M^{83}, C^{131} and W^{346} were observed in the universal central cleft region. When compared to the surface structure of the protein depicting the electrostatic potential, it was observed that the central cleft region is a positive charged patch on the protein surface and therefore likely to interact with the negatively charged RNA moiety. These results were also mapped on a model of four *Tp* Csc2 protein subunits arranged in a chain. When the three residues were mapped on each of the four subunits, the location of the cross-links coincided with a positively charged channel along the four proteins (Figure 3.8) indicating that this constitutes an RNA binding patch on the surface of these proteins. In addition, the biochemical analysis shows that conserved residues in these positively charged grooves contribute significantly to RNA binding [103].

***T. tenax* Cas7**

In the superimposed model of the Type I-A *T. tenax* Cas7, the crRNA uniformly contacts secondary structure elements of the lid domain as well as the central cleft defined by the core RRM. Four out of five cross-linked amino acid residues, G^{28}, V^{148}, F^{155} and M^{172} were located in the lid-domain (insertion domain 1) and one cross-linked amino acid residue Y^{7} was located in the universally present central cleft.

***T. thermophilus* Csm3**

In Type III-A *T. thermophilus* Csm3 the cross-linked amino acid residues Q^{144}, G^{153} and R^{162} were present in the lid domain and P^{96} was present in the central cleft. Previous studies on the respective Type I-A and Type III-A homologues *Sulfolobus solfataricus* Cas7 [19] and *Methanopyrus kandleri* Csm3 [55] also report similar RNA binding properties in the insertion domain 1 (lid domain) of Cas7 proteins in full agreement with the cross-links observed in this study.

In conclusion, the positively charged surface groove appears to be a conserved functional site for crRNA recognition and the influence of the lid domain in crRNA interaction varies with respect to specific Cas7 family proteins. Also it can be hypothesized that the Cas7 family proteins constituting the backbone of crRNP complexes harbor the active sites for RNA binding.

4.2.3 Structural insights into the protein-RNA interactions in multi-subunit crRNP complexes

The protein-RNA cross-linking approach was extended to investigate protein-RNA interactions in fully assembled multi-subunit crRNP complexes. The three crRNP complexes investigated in this work included Type I-E *E. coli* Cascade complex, Type III-A *T. thermophilus* Csm complex and Type III-B *T. thermophilus* Cmr complex.

Type I-E Cascade complex from *E. coli*

The *E. coli* Cascade complex comprising 11 protein subunits (corresponding to five different Cas proteins) and a crRNA was recently crystallized (Figure 1.15) [105], providing a major breakthrough in the understanding of molecular interactions within a crRNP complex. With the protein-RNA cross-linking investigation of *E. coli* Cascade complex, 70 different cross-linked peptide-RNA heteroconjugates were identified (Table 3.6), corresponding to 17 unique cross-linked regions across the five different Cas proteins in the complex. The high yield of cross-links is indicative of the significant contribution of the protein-RNA cross-linking approach. Two important points were taken into consideration during the interpretation of these results when mapped on the crystal structure of *E. coli* Cascade complex: (i) The cross-linking experiments were performed in solution, which would result in a conformational variability of the complex in solution and (ii) The crystal structure used to map the cross-links was obtained from the Cascade complex comprising a different crRNA (with a different spacer) than the one used in the cross-linking study, therefore the mode of RNA binding would be different in the two complexes. Together these two considerations were helpful in interpreting the differences between the co-crystallization and the in solution protein-RNA interactions studies.

In Cas6e, the head of the Cascade complex, two cross-linked regions were observed (Figure 3.9B). Firstly, K^{106} present in the positive charged groove-loop in the C-terminal RRM domain was identified as a cross-linked residue. This region has been reported to present extensive contacts with the 3' crRNA stem-loop and the results of protein-RNA cross-linking validate this interaction. In the second cross-linked region $^{145}R\text{-}Y^{148}$ the exact cross-linked residue could not be identified. Nonetheless, this region is a part of a flexible loop in close spatial proximity to uracil residue and therefore it is likely to interact with the RNA in solution.

In Cse1, the largest protein in Cascade complex, one cross-linked amino acid residue F^{403} was identified (Figure 3.9C). The cross-linked region is located at C-terminal of Cse1 that connects with C-terminal of Cse2.2. To date there have been no reports on the RNA binding properties of this domain. However, in the cross-linking study F^{403} was found to interact with a uracil residue in close spatial proximity indicating that in addition to providing a structural bridge between Cse1 and Cse2.2 the C-terminal part of Cse1 is also involved in RNA interaction.

In Cas5e, the tail protein with a right-hand fist-shape, four different cross-linked amino acid residues were identified. The W^{16}, Y^{145} and P^{200} were observed in the palm domain of the protein. From the crystal structure it can be observed that the side-chain of U^{2} nucleotide residue is adjacent to the side-chain of amino acid residues W^{16} and Y^{145} which makes it highly possible that these residues form a cross-link (Figure 3.9 D). The Y^{85} amino acid residue present in the arch of the thumb domain was also observed cross-linked to a uracil residue. It has been postulated that the Y^{85} amino acid residue stabilized the RNA-DNA duplex formation between the crRNA and the target DNA [106]. Considering the structural variability of the Cascade complex in solution it can be speculated that the cross-linked nucleotide residue U^{10} is in close spatial proximity to this amino acid residue. In addition, the arch is a modified RRM which is a characteristic RNA binding domain.

In Cas7 proteins, which form the backbone of Cascade complex, six cross-linked residues were identified corresponding to five different cross-linked regions (Table 3.6). However the cross-linked residues could not be assigned to a specific residue or protein in the crystal structure because there are six copies of Cas7 present in the Cascade complex. Nevertheless possible protein-RNA interactions were mapped on the crystal structure depending on the location of uracil residues on the crRNA which were in close proximity to the possible cross-linked residues in Cas7 proteins. The location of these cross-linked residues/regions has been discussed earlier (Section 4.2.2) with comparison to other Cas7 family proteins. The K^{27} amino acid residue in the palm domain of the Cas7 protein lies in the universal central cleft that interacts with the RNA as observed in other Cas7 homologues. According to the crystal structure, in the proteins Cas7.1, Cas7.2, Cas7.4 and Cas7.6 the K^{27} residue is in close spatial proximity to a uracil residue making it highly likely that these result in protein-RNA cross-links upon UV irradiation. In addition, an interesting observation was made with regard to the M^{166} amino acid residue. Approximately, 40 out of 70 cross-links observed for the entire Cascade complex corresponded to the cross-

linked region ^{166}MATSGMMTELGK^{177}. The peptide was observed in eight different modified versions with one or more of the three methionine residues oxidized and each of the eight peptides was observed cross-linked to a wide range of RNA moieties (Table 3.6). In 25 out of 40 cross-links it was the M^{166} amino acid residue identified as the cross-linked residue, however in the remaining 15 exact residue could not be identified due to a lack of fragment ions in that region of the ion series. When mapped to different Cas proteins on the crystal structure, the M^{166} residue in Cas7.1, Cas7.2 and Cas7.5 was observed to be in close spatial proximity to a uracil residue (Figure 3.10 and 3.11). Although with the protein-RNA cross-linking approach it is not possible to identify exactly in which of these Cas proteins M^{166} residue was cross-linked, from the considerably high number of cross-links identified it can nonetheless be speculated that in all the three Cas proteins the M^{166} amino acid residue interacts with the RNA. This result validates the previous observation that the M^{166} amino acid interacts with the crRNA, and the side chain of methionine intercalates between the 3rd and 4th base of the crRNA segment in close proximity to the respective Cas7 protein [105].

In Cse2 proteins, the belly of the Cascade complex, five different cross-linked regions were identified. As there are two Cse2 subunits present in the Cascade complex, the cross-linked amino acid residues cannot be unambiguously assigned to either of these subunits. Novel RNA binding regions were identified in the Cse2 proteins in contrast to the previous X-ray studies where direct contacts between the crRNA and Cse2 proteins were not reported (Figure 3.12). All the identified cross-linked residues were located on the protein surface opposite to the one facing the crRNA. However, both the surfaces of the Cse2 dimer are positively charged as determined by electrostatic calculations [43, 151] suggesting that both surfaces constitute potential RNA binding regions.

Overall, the Cascade complex is an excellent platform to study protein-RNA interactions especially with the recently published crystal structure of the Cascade complex now available. The crystal structures do not represent an active structure as in solution, because a crystal structure is obtained only after the flexibility is reduced. However the experiments performed in this study were in solution and the results reflect the conformational variability the complex can have in solution. The protein-RNA cross-linking helps validate the interactions determined by co-crystallization e.g., the Y^{145} amino acid residue in Cas5e cross-linked to the U^{2} nucleotide residue on the crRNA at the 5’ end. In addition new RNA interaction sites were identified in

Cse1, Cse2 and Cas7 proteins, by in solution protein-RNA cross-linking approach that might correspond to a different conformation than the one showed in X-ray studies.

Type III-A Csm complex from *T. thermophilus*

In the Type III-A Csm complex composed of twelve different protein subunits corresponding to five proteins (Csm1-Csm5), 43 different peptide-RNA cross-links were observed. Overall these corresponded to ten different cross-linked protein regions across the five Csm proteins (Table 3.7), with at least one cross-linked region identified in every protein.

In Csm3 (a Cas7 family protein) alone 28 cross-links were identified, corresponding to three cross-linked protein regions and within these regions four cross-linked amino acid residues could be identified. The results of Csm3 protein-RNA cross-linking have been discussed earlier (Section 4.2.2) in comparison with other Cas7 homologues. The cross-linked amino acid residues Q^{144}, G^{153} and R^{162} were present in the lid-domain and P^{96} and ^{49}P-K^{54} in the central cleft and both these regions have been observed to interact with RNA e.g., in the *T. tenax* Cas7 and *Tp* Csc2. However, there are six copies of the Csm3 protein in the Csm complex and it was therefore not possible to assign these cross-linked regions/residues to a single Csm3 subunit.

In the *E. coli* Cascade complex the sequence of the crRNA present in the crRNP complex was known, therefore most of the cross-linked nucleotide residues could also be mapped on the crRNA sequence. However, the *Tt* Csm complex is known to bind crRNAs of variable lengths as verified by the deep sequencing analysis [59] and it could not be unambiguously determined which of the crRNAs was present in the final assembled Csm complex used for cross-linking experiments. As the sequence of crRNA present in the Csm complex was not known the cross-linked nucleotide residues could not be mapped on the RNA level.

In addition, the results from protein-RNA cross-linking studies are more explanatory when analyzed in conjunction with a high resolution crystal structure like in the *E. coli* Cascade complex, so that all the identified residues can be mapped on the protein structure and the predicted protein-RNA interactions can be visualized in three-dimension. Nonetheless, the extensive number of cross-links observed in *Tt* Csm complex reveals a strong potential of protein-RNA cross-linking in determining protein-RNA interactions *in vitro* and as a source for constrains when modelling RNA strands on crystal structures of proteins.

Type III-B Cmr complex from *T. thermophilus*

The protein-RNA cross-linking investigations in the *Tt* Cmr complex were carried out in both endogenous and reconstituted complexes. Five cross-linked protein regions were identified in both the complexes. In the endogenous complex the cross-linked regions were observed in proteins Cmr2, Cmr3, Cmr4 and Cmr6 and in the reconstituted complex in proteins Cmr1, Cmr2, Cmr3 and Cmr4 (Table 3.8). In all the observed cross-links a uracil residue was identified as the cross-linked nucleotide. Based on the location of the uracil residues in the 46 nt crRNA (the crRNA used in the assembly of reconstituted Cmr complex), the cross-links were mapped on the Cmr proteins in close proximity to these uracil residues as shown in the model of *Tt* Cmr complex (Figure 3.14).

The results of cross-linking analysis were reproducible for the cross-links identified in endogenous complex and the same cross-links identified in reconstituted complex, with only one exception in each case. A cross-linked protein region ^{34}T-R^{58} was identified in the Cmr1 protein in the reconstituted complex however no cross-link was identified for the Cmr1 protein in the endogenous complex, the cross-linked nucleotide residue was identified as the U^{39} residue close to the 3' end of the crRNA. Similarly, a cross-linked protein region ^{169}L-K^{186} was identified in the Cmr6 protein in the endogenous complex and not in the reconstituted complex, however here the cross-linked residue could not be mapped on the crRNA.

The electron microscopy studies have shown that both Cmr1 and Cmr6 together form a 'curled-head' of the 'sea-worm' shaped Cmr complex (Figure 1.17) [52]. From these observations it can be speculated that Cmr1 protein cross-links to the U^{39} residue at the 3' end of crRNA. However, due to Cmr1 being weakly associated with the complex it can dissociate in solution [52]. When Cmr1 dissociates from the complex, Cmr6 is able to form a cross-link to the U^{39} residue as it is also in close proximity to this uracil residue at the 3' end of crRNA (Figure 3.14). Furthermore, this hypothesis is in agreement with the previous reports from the native MS analysis, showing that Cmr1 protein has a loose association with the Cmr complex under *in vitro* conditions [52].

The protein-RNA cross-linking studies performed with the endogenous and reconstituted *Tt* Cmr complex demonstrate both complexes interact with RNA in an equivalent manner, to the level of cross-linked nucleotide and amino acid. In addition, these results revealed the first five direct protein-RNA contacts in the multi-subunit *Tt* Cmr complex, providing valuable structural information for further structural and functional studies.

4.2.4 Protein-protein interactions in Type I-B *C. thermocellum* Cascade complex

In addition to the protein-RNA interactions discussed so far, the protein-protein interactions in a multi-subunit crRNP complex were also investigated as a part of the structural proteomics studies in this work. Type I-B *C. thermocellum* Cascade complex, a ~330 kDa comprising four Cas proteins Cas5, Cas6, Cas8b and Cas7 was used for this study. The stoichiometry of Cas6:Cas5:Cas8b:Cas7 in this complex was determined as low:low:medium:high with iBAQ analysis (further details of the quantitative analysis have been discussed earlier in Section 4.1.2).

The protein-protein interactions were investigated using a lysine directed chemical cross-linker BS3, because of the high reactivity of primary amines and high abundance and extensive distribution of lysine residues over the protein surface. However, it is not possible to identify all theoretically possible lysine-lysine cross-links because all the lysine residues might not be equally reactive or accessible to the cross-linker. They could be involved in cross-links that result in very long or very short peptides that do not fly/fragment well in the mass spectrometer and also not all cross-linked peptides are detectable by MS. The molar excess of cross-linker (75:1) was selected to avoid formation of higher-order aggregates and refrain from interactions deriving from experimental artifacts.

From the MS analysis, 126 unique inter-protein cross-links were identified within the four Cas proteins (Appendix, Table 6.2). The number of inter-protein cross-links was in agreement with the stoichiometry determined for the complex. The Cas7 protein which is speculated to be present in six copies in the Cascade complex was observed to form inter-protein contacts with all the other proteins. No inter-protein cross-link was identified between the two proteins Cas5 and Cas6 which are present in single copies and this is also in agreement with their position in the Cascade complex, as Cas6 forms the head and Cas5 the tail which are unlikely to interact.

In addition, from the crystal structure of a fully assembled Cascade complex such as the Type I-E *E. coli* Cascade, it can be observed that Cas7 forms the helical backbone of the complex which interacts with the head protein Cas6, the tail protein Cas5 and across the length of the complex the Cas7 backbone is supported by a belly of two Cse2 proteins (Figure 1.15) [43, 105]. The Cse2 dimer in case of Type I-B complexes is replaced by a single large subunit Cas8b (Figure 1.4) [22]. The protein-protein cross-links identified are also in agreement with this structural

arrangement, with 75 inter-protein cross-links identified between the Cas7 backbone and Cas8b, 37 between Cas7 and Cas5 (tail) and one between the Cas7 and Cas6 (head).

The intra-protein cross-links identified did not provide much information as the stoichiometry of Cas7 and Cas8b indicate multiple copies of these proteins are present and there it is not possible to distinguish the intra-protein cross-links from inter-protein cross-links. Nonetheless for the Cas5 and Cas6 proteins which are present in a single copy in the complex, the intra-protein cross-links identified were mapped on the respective protein sequence (Appendix, Figure 3.7).

An overall higher number of cross-links were identified for Cas7 and Cas8b which might indicate a high structural flexibility of these proteins in solution. In addition the presence of lysine rich regions distributed across the protein length also indicates a high possibility of protein-protein interactions.

A high number of protein-protein cross-links were identified in this investigation (126) which has become possible at present due to the availability of high sequencing speed such as that provided by the Orbitrap Fusion mass spectrometer and advanced search engines such as pLink used in this study. The cross-links identified show a heterogeneous distribution indicating different protein interaction networks within the Cascade complex. This information in conjunction with additional structural data can be used to propose a structural model for the three dimensional description of the protein complex.

4.3 Considerations in the identification of protein-RNA interactions by UV induced cross-linking and MS

A major part of the presented work focused on the investigation of protein-RNA interactions in crRNP complexes using UV induced protein-RNA cross-linking and MS. Some important considerations and potentials of the protein-RNA cross-linking approach will be discussed in this section in reference to the results obtained during this work.

1. Fragmentation mode: The fragmentation of peptide-RNA heteroconjugates was performed using high-energy collision dissociation (HCD) in Orbitrap instruments where the MS/MS fragmentation occurs in the Orbitrap, therefore benefiting from the high-mass accuracy. In addition the limitations of ion traps could be overcome because here the spectra obtained do not exhibit a low mass cut-off and the important ions like RNA marker ions, a2-b2 pairs and

immonium ions (usually located in the lower m/z region) were detected. Furthermore, as the yield of a cross-linking reaction is low and cross-links have a poor ionization efficiency compared to linear peptides, the MS analysis were performed using instruments with high resolution, high mass accuracy and better sequencing speeds such as LTQ-Orbitrap Velos and Q Exactive HF. This would provide a stronger likelihood that a peptide-RNA precursor would be picked for fragmentation out of a complex sample.

2. Identification of cross-link: The spectra for cross-links are poor quality compared to standard peptides because of low signal intensities. The automated proteomics search engines are directed toward peptide identification considering a cross-linked RNA moiety as a modification on the peptide. Also, a large number of potential RNA oligonucleotides can form a cross-link and together these two aspects make the data analysis difficult. The basic principle applied in cross-link identification was that the mass of a peptide-RNA cross-link precursor is purely additive of the masses of peptide and RNA moieties [160]. A new data analysis tool, RNP^{xl} (Section 2.2.9.3) was used for the analysis which followed a precursor variant approach for automated identification of cross-linked heteroconjugates by subtraction of calculated RNA masses from the experimental precursor mass. The results were further checked manually to validate the cross-link identification.

3. Cross-linked nucleotides: The cross-linked nucleotide is determined by calculating the mass difference between the mass of cross-linked heteroconjugate and the mass of cross-linked peptide. There is a substantial difference in the reactivity of uridine and other nucleotides as observed in the cross-links reported in this work, in previous studies related to protein-RNA cross-linking in the Urlaub lab and also in the studies where the cross-linking yields of different amino-acids and nucleotides were investigated by Shetlar *et al*. [59, 108, 109, 160-163]. Conversely, if a cross-linked heteroconjugate comprise adenosine, guanosine or cytidine cross-linked to the amino acid residue it is less likely to be identified by the MS analysis. In the cross-links observed for the *E. coli* Cascade complex, the peptide ^{51}SGYYAQNIGESSLRTIHLAQLR^{72} was observed cross-linked to UG-H_2O, where G nucleotide was identified as the cross-linked nucleotide residue (Appendix, Figure 6.2 I). This was the only example of a cross-link where a nucleotide other than uracil was observed to be cross-linked. Considering that uracil is the most reactive nucleotide, some of the cross-linking experiments in this work were also performed with polyU oligonucleotide, but the aim was only to identify regions on the protein level that

interact with RNA. In addition, the information derived from protein-RNA cross-linking studies is limited when attempting to unambiguously map the cross-linked nucleotide at the level of RNA because in most cases the cross-linked RNA was a single nucleotide residue. Nonetheless, there were cross-links identified with longer RNA moieties, but the current cross-linking workflow only provides information about the composition of the cross-linked RNA moiety and not their position in the sequence.

4. RNA marker ions: A typical observation in the cross-link spectra reported in the work was presence of distinct marker ions of the intact nucleotides with a neutral loss of water (C = 306.05, U = 307.03, A = 330.06, G = 346.05) or the nucleic acid bases (C' = 112.05, U' = 113.03, A' = 136.06, G' = 152.06) in the lower m/z region of the spectrum. In previous studies fragmentation of pure RNA also has been observed to produce nucleobases due to the cleavage of N-glycosidic bond [164]. When only one nucleotide was cross-linked to a peptide, no marker ion signal is visible and the MS/MS spectra are dominated by peptide fragments. Some of the examples where marker ions were observed include A' marker ion as depicted in *C. thermocellum* Cas6b peptide ^{184}MIGFK^{188} cross-linked UGA and G' and C' marker ions in *M. maripaludis* Cas6b peptide ^{182}NQNM(ox)VGFR^{189} cross-linked to UUGC-PO_3 and (Figure 3.5).

5. Cross-linked amino acids: In the previous studies all amino acids except E, N and D have been demonstrated to cross-link with RNA moieties [59, 108, 161]. In this work 12 different amino acids have been reported to cross-link with RNA and these include M, Y, L, P, F, W, K, C, V, G, Q and R as shown in different examples in the Appendix, Figure 6.1-6.5. The aromatic residues are the most commonly observed cross-linked amino acid residues. Usually the cross-linked amino acid residues were identified when the fragment ions comprising this amino acid within the b- or y- series show a mass-shift corresponding to the mass of an RNA adduct and the other fragment ions are observed without any mass shifts e.g., amino acid residue M^{83} in the peptide ^{82}LMAVTR^{87} cross-linked to uracil (Figure 6.1 A). In addition, when the fragment ion information was not sufficient, the immonium ion of the cross-linked amino acid was observed with a mass-shift corresponding to mass of RNA adducts e.g., amino acid residue W^{346} in the peptide ^{346}WVEELKGGGQK^{356} cross-linked to uracil (Figure 6.1 C). However, for nine peptide-RNA cross-links out of the 48 reported in this work, the cross-linked amino acid residue could not be identified e.g., the peptide ^{136}VEDVHPISERPQYFSGDGK^{154} cross-linked to uracil (Figure 6.2 B).

When all considerations above are taken into account, the MS approaches have been very helpful in identifying interactions which are otherwise very difficult to do. In this work, approximately 50 protein-RNA cross-links and 126 inter-protein cross-links were identified and the stoichiometries of Type I-B Cascade complexes from *H. volcanii* and *C. thermocellum* were determined. These results when integrated with other structural techniques in collaboration resulted in three dimensional views of the complexes, as demonstrated by the more than five joint publications.

4.4 Conclusions and future perspectives

In this work, the prokaryotic immune defense system was investigated using both quantitative and structural aspects of mass spectrometry. With relative quantitation using dimethyl labeling the effect of deletion of a significant *cas7* gene on the regulation of other Cas proteins could be determined. It provides a strong indication that removal of an essential *cas* gene is compensated by increased expression of other *cas* genes in the system. In addition, the absolute amounts of different Cas protein components in *H. volcanii* and *C. thermocellum* Cascade complex were determined using iBAQ. These results provide a straightforward and first step in determining the stoichiometry of Cas proteins in multi-subunit crRNP complexes for further investigation of molecular interactions in this system. The presented protein-protein cross-linking investigation provides molecular insights into the interactions between different Cas proteins subunits in the *C. thermocellum* Cascade complex. The limitation in the assignment of exact cross-link sites in some of the proteins provides a scope for future improvements that could be implemented in the investigation of such multi-subunit complexes by MS. Furthermore, the protein-RNA interactions in both single (recombinant) protein-RNA complexes and multi-subunit crRNP complexes were investigated using UV induced cross-linking. The extensive amounts of protein-RNA cross-links reported in this work are the first reported evidence of such interactions to be determined in the different CRISPR-Cas systems by MS. Although this study mainly focused on the structural investigations in Type I and Type III crRNP complexes, a well-established workflow is available and ready to explore the widely popular Type II CRISPR-Cas system comprising a single player Cas9. Moreover, the cross-linking approach can be further extended to investigate protein-DNA interactions between the fully assembled crRNP complexes and the target DNA for an improved understanding of the CRISPR interference and the overall mechanism of the CRISPR-Cas based immune defense system.

5. References

1. Breitbart, M. and F. Rohwer, *Here a virus, there a virus, everywhere the same virus?* Trends Microbiol, 2005. **13**(6): p. 278-84.
2. Samson, J.E., et al., *Revenge of the phages: defeating bacterial defences.* Nat Rev Microbiol, 2013. **11**(10): p. 675-87.
3. Nakamura, Y., et al., *Biased biological functions of horizontally transferred genes in prokaryotic genomes.* Nat Genet, 2004. **36**(7): p. 760-6.
4. Thomas, C.M. and K.M. Nielsen, *Mechanisms of, and barriers to, horizontal gene transfer between bacteria.* Nat Rev Microbiol, 2005. **3**(9): p. 711-21.
5. Barrangou, R., et al., *CRISPR provides acquired resistance against viruses in prokaryotes.* Science, 2007. **315**(5819): p. 1709-12.
6. Sorek, R., V. Kunin, and P. Hugenholtz, *CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea.* Nat Rev Microbiol, 2008. **6**(3): p. 181-6.
7. Kunin, V., R. Sorek, and P. Hugenholtz, *Evolutionary conservation of sequence and secondary structures in CRISPR repeats.* Genome Biol, 2007. **8**(4): p. R61.
8. Marraffini, L.A. and E.J. Sontheimer, *CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea.* Nat Rev Genet, 2010. **11**(3): p. 181-90.
9. Ishino, Y., et al., *Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product.* J Bacteriol, 1987. **169**(12): p. 5429-33.
10. Nakata, A., M. Amemura, and K. Makino, *Unusual nucleotide arrangement with repeated sequences in the Escherichia coli K-12 chromosome.* J Bacteriol, 1989. **171**(6): p. 3553-6.
11. Jansen, R., et al., *Identification of genes that are associated with DNA repeats in prokaryotes.* Mol Microbiol, 2002. **43**: p. 1565-1575.
12. Westra, E.R., A. Buckling, and P.C. Fineran, *CRISPR-Cas systems: beyond adaptive immunity.* Nat Rev Microbiol, 2014. **12**(5): p. 317-26.
13. Bult, C.J., et al., *Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii.* Science, 1996. **273**(5278): p. 1058-73.
14. Clery, A., M. Blatter, and F.H. Allain, *RNA recognition motifs: boring? Not quite.* Curr Opin Struct Biol, 2008. **18**(3): p. 290-8.
15. Makarova, K.S., et al., *A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis.* Nucleic Acids Res, 2002. **30**(2): p. 482-96.
16. Makarova, K.S., et al., *A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action.* Biol Direct, 2006. **1**: p. 7.

17. Sakamoto, K., et al., *X-ray crystal structure of a CRISPR-associated RAMP module [corrected] Cmr5 protein [corrected] from Thermus thermophilus HB8.* Proteins, 2009. **75**(2): p. 528-32.
18. Haurwitz, R.E., et al., *Sequence- and structure-specific RNA processing by a CRISPR endonuclease.* Science, 2010. **329**(5997): p. 1355-8.
19. Lintner, N.G., et al., *Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE).* J Biol Chem, 2011. **286**(24): p. 21643-56.
20. Haft, D.H., et al., *A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes.* PLoS Comput Biol, 2005. **1**(6): p. e60.
21. Makarova, K.S., et al., *Evolution and classification of the CRISPR-Cas systems.* Nat Rev Microbiol, 2011. **9**(6): p. 467-77.
22. van der Oost, J., et al., *Unravelling the structural and mechanistic basis of CRISPR-Cas systems.* Nat Rev Microbiol, 2014. **12**(7): p. 479-92.
23. Wiedenheft, B., et al., *Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense.* Structure, 2009. **17**(6): p. 904-12.
24. Reeks, J., J.H. Naismith, and M.F. White, *CRISPR interference: a structural perspective.* Biochem J, 2013. **453**(2): p. 155-66.
25. Pourcel, C., G. Salvignol, and G. Vergnaud, *CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies.* Microbiology, 2005. **151**(Pt 3): p. 653-63.
26. Deveau, H., et al., *Phage response to CRISPR-encoded resistance in Streptococcus thermophilus.* J Bacteriol, 2008. **190**(4): p. 1390-400.
27. Mojica, F.J., et al., *Short motif sequences determine the targets of the prokaryotic CRISPR defence system.* Microbiology, 2009. **155**(Pt 3): p. 733-40.
28. Shah, S.A., et al., *Protospacer recognition motifs: mixed identities and functional diversity.* RNA Biol, 2013. **10**(5): p. 891-9.
29. Brouns, S.J., et al., *Small CRISPR RNAs guide antiviral defense in prokaryotes.* Science, 2008. **321**(5891): p. 960-4.
30. Jore, M.M., et al., *Structural basis for CRISPR RNA-guided DNA recognition by Cascade.* Nat Struct Mol Biol, 2011. **18**(5): p. 529-36.
31. Bhaya, D., M. Davison, and R. Barrangou, *CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation.* Annu Rev Genet, 2011. **45**: p. 273-97.
32. Marraffini, L.A. and E.J. Sontheimer, *Invasive DNA, chopped and in the CRISPR.* Structure, 2009. **17**(6): p. 786-8.
33. Beloglazova, N., et al., *A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats.* J Biol Chem, 2008. **283**(29): p. 20361-71.
34. Yosef, I., M.G. Goren, and U. Qimron, *Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli.* Nucleic Acids Res, 2012. **40**(12): p. 5569-76.

35. Sinkunas, T., et al., *Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system.* EMBO J, 2011. **30**(7): p. 1335-42.
36. Sinkunas, T., et al., *In vitro reconstitution of Cascade-mediated CRISPR immunity in Streptococcus thermophilus.* EMBO J, 2013. **32**(3): p. 385-94.
37. Sashital, D.G., B. Wiedenheft, and J.A. Doudna, *Mechanism of foreign DNA selection in a bacterial adaptive immune system.* Mol Cell, 2012. **46**(5): p. 606-15.
38. Hochstrasser, M.L., et al., *CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference.* Proc Natl Acad Sci U S A, 2014. **111**(18): p. 6618-23.
39. Carte, J., et al., *Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes.* Genes Dev, 2008. **22**(24): p. 3489-96.
40. Wiedenheft, B., et al., *RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions.* Proc Natl Acad Sci U S A, 2011. **108**(25): p. 10092-7.
41. van Duijn, E., et al., *Native tandem and ion mobility mass spectrometry highlight structural and modular similarities in clustered-regularly-interspaced shot-palindromic-repeats (CRISPR)-associated protein complexes from Escherichia coli and Pseudomonas aeruginosa.* Mol Cell Proteomics, 2012. **11**(11): p. 1430-41.
42. Westra, E.R., et al., *The CRISPRs, they are a-changin': how prokaryotes generate adaptive immunity.* Annu Rev Genet, 2012. **46**: p. 311-39.
43. Wiedenheft, B., et al., *Structures of the RNA-guided surveillance complex from a bacterial immune system.* Nature, 2011. **477**(7365): p. 486-9.
44. Westra, E.R., et al., *CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3.* Mol Cell, 2012. **46**(5): p. 595-605.
45. Jinek, M., et al., *Structures of Cas9 endonucleases reveal RNA-mediated conformational activation.* Science, 2014. **343**(6176): p. 1247997.
46. Nishimasu, H., et al., *Crystal structure of Cas9 in complex with guide RNA and target DNA.* Cell, 2014. **156**(5): p. 935-49.
47. Deltcheva, E., et al., *CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III.* Nature, 2011. **471**(7340): p. 602-7.
48. Jinek, M., et al., *A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity.* Science, 2012. **337**(6096): p. 816-21.
49. Karvelis, T., et al., *crRNA and tracrRNA guide Cas9-mediated DNA interference in Streptococcus thermophilus.* RNA Biol, 2013. **10**(5): p. 841-51.
50. Gasiunas, G., et al., *Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria.* Proc Natl Acad Sci U S A, 2012. **109**(39): p. E2579-86.
51. Mali, P., K.M. Esvelt, and G.M. Church, *Cas9 as a versatile tool for engineering biology.* Nat Methods, 2013. **10**(10): p. 957-63.
52. Staals, R.H., et al., *Structure and activity of the RNA-targeting Type III-B CRISPR-Cas complex of Thermus thermophilus.* Mol Cell, 2013. **52**(1): p. 135-45.

53. Rouillon, C., et al., *Structure of the CRISPR interference complex CSM reveals key similarities with cascade.* Mol Cell, 2013. **52**(1): p. 124-34.
54. Spilman, M., et al., *Structure of an RNA silencing complex of the CRISPR-Cas immune system.* Mol Cell, 2013. **52**(1): p. 146-52.
55. Hrle, A., et al., *Structure and RNA-binding properties of the type III-A CRISPR-associated protein Csm3.* RNA Biol, 2013. **10**(11): p. 1670-8.
56. Benda, C., et al., *Structural model of a CRISPR RNA-silencing complex reveals the RNA-target cleavage activity in Cmr4.* Mol Cell, 2014. **56**(1): p. 43-54.
57. Hatoum-Aslan, A., et al., *Genetic characterization of antiplasmid immunity through a type III-A CRISPR-Cas system.* J Bacteriol, 2014. **196**(2): p. 310-7.
58. Deng, L., et al., *A novel interference mechanism by a type IIIB CRISPR-Cmr module in Sulfolobus.* Mol Microbiol, 2013. **87**(5): p. 1088-99.
59. Staals, R.H., et al., *RNA targeting by the type III-A CRISPR-Cas Csm complex of Thermus thermophilus.* Mol Cell, 2014. **56**(4): p. 518-30.
60. Cocozaki, A.I., et al., *Structure of the Cmr2 subunit of the CRISPR-Cas RNA silencing complex.* Structure, 2012. **20**(3): p. 545-53.
61. Hoffmann, E., et al., *Wiley - Mass Spectrometry: Principles and Applications, 3rd Edition*. 2007. John Wiley & Sons.
62. Fenn, J.B., et al., *Electrospray ionization for mass spectrometry of large biomolecules.* Science, 1989. **246**(4926): p. 64-71.
63. Tanaka, K., *Protein and polymer analysis up to m/z 100,000 by laser ionization time-of-flight mass spectrometry.* Rapid Commun Mass Spectrom, 1988. **2**(8): p. 151-153.
64. Steen, H. and M. Mann, *The ABC's (and XYZ's) of peptide sequencing.* Nat Rev Mol Cell Biol, 2004. **5**(9): p. 699-711.
65. Cole, R.B., *Wiley - Electrospray and MALDI Mass Spectrometry: Fundamentals, Instrumentation, Practicalities, and Biological Applications, 2nd Edition.* 2010. John Wiley & Sons.
66. Li, K.Y., H. Tu, and A.K. Ray, *Charge limits on droplets during evaporation.* Langmuir, 2005. **21**(9): p. 3786-94.
67. Dass, C., *Fundamentals of contemporary mass spectrometry.* 2007. John Wiley & Sons.
68. Holcapek, M., *Mass Analayzers.* Chromedia Analytical Sciences: www.chromedia.org.
69. Zubarev, R.A. and A. Makarov, *Orbitrap mass spectrometry.* Anal Chem, 2013. **85**(11): p. 5288-96.
70. Makarov, A., et al., *Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer.* Anal Chem, 2006. **78**(7): p. 2113-20.
71. Scientific, T.F., *LTQ Orbitrap XLBiotech Operations,* (Training Course Manual).
72. Roepstorff, P. and J. Fohlman, *Proposal for a common nomenclature for sequence ions in mass spectra of peptides.* Biomed Mass Spectrom, 1984. **11**(11): p. 601.
73. Biemann, K., *Contributions of mass spectrometry to peptide and protein structure.* Biomed Environ Mass Spectrom, 1988. **16**(1-12): p. 99-111.
74. McLuckey, S.A., G.J. Van Berkel, and G.L. Glish, *Tandem mass spectrometry of small, multiply charged oligonucleotides.* J Am Soc Mass Spectrom, 1992. **3**(1): p. 60-70.

75. Wu, J. and S.A. McLuckey, *Gas-phase fragmentation of oligonucleotide ions.* International Journal of Mass Spectrometry, 2004. **237**(2-3): p. 197-241.
76. Aebersold, R. and M. Mann, *Mass spectrometry-based proteomics.* Nature, 2003. **422**(6928): p. 198-207.
77. Kim, M.S., et al., *A draft map of the human proteome.* Nature, 2014. **509**(7502): p. 575-81.
78. Wilhelm, M., et al., *Mass-spectrometry-based draft of the human proteome.* Nature, 2014. **509**(7502): p. 582-7.
79. Laemmli, U.K., *Cleavage of structural proteins during the assembly of the head of bacteriophage T4.* Nature, 1970. **227**(5259): p. 680-5.
80. Cargile, B.J., D.L. Talley, and J.L. Stephenson, Jr., *Immobilized pH gradients as a first dimension in shotgun proteomics and analysis of the accuracy of pI predictability of peptides.* Electrophoresis, 2004. **25**(6): p. 936-45.
81. Mann, M., R.C. Hendrickson, and A. Pandey, *Analysis of proteins and proteomes by mass spectrometry.* Annu Rev Biochem, 2001. **70**: p. 437-73.
82. Nikolov, M., C. Schmidt, and H. Urlaub, *Quantitative mass spectrometry-based proteomics: an overview.* Methods Mol Biol, 2012. **893**: p. 85-100.
83. Bantscheff, M., et al., *Quantitative mass spectrometry in proteomics: a critical review.* Anal Bioanal Chem, 2007. **389**(4): p. 1017-31.
84. Boersema, P.J., et al., *Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics.* Nat Protoc, 2009. **4**(4): p. 484-94.
85. Boersema, P.J., et al., *Triplex protein quantification based on stable isotope labeling by peptide dimethylation applied to cell and tissue lysates.* Proteomics, 2008. **8**(22): p. 4624-32.
86. Hsu, J.L., et al., *Enhanced a1 fragmentation for dimethylated proteins and its applications for N-terminal identification and comparative protein quantitation.* J Proteome Res, 2007. **6**(6): p. 2376-83.
87. Schwanhausser, B., et al., *Global quantification of mammalian gene expression control.* Nature, 2011. **473**(7347): p. 337-42.
88. Wilhelm, B.G., et al., *Composition of isolated synaptic boutons reveals the amounts of vesicle trafficking proteins.* Science, 2014. **344**(6187): p. 1023-8.
89. Brendel, J., et al., *A complex of Cas proteins 5, 6, and 7 is required for the biogenesis and stability of clustered regularly interspaced short palindromic repeats (crispr)-derived rnas (crrnas) in Haloferax volcanii.* J Biol Chem, 2014. **289**(10): p. 7164-77.
90. Cox, J. and M. Mann, *MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.* Nat Biotechnol, 2008. **26**(12): p. 1367-72.
91. Berman, H., K. Henrick, and H. Nakamura, *Announcing the worldwide Protein Data Bank.* Nat Struct Biol, 2003. **10**(12): p. 980.
92. Sharon, M., *How far can we go with structural mass spectrometry of protein complexes?* J Am Soc Mass Spectrom, 2010. **21**(4): p. 487-500.

93. Ganem, B., Y.T. Li, and J.D. Henion, *Detection of Noncovalent Receptor Ligand Complexes by Mass-Spectrometry.* Journal of the American Chemical Society, 1991. **113**(16): p. 6294-6296.
94. Ganem, B., Y.T. Li, and J.D. Henion, *Observation of Noncovalent Enzyme Substrate and Enzyme Product Complexes by Ion-Spray Mass-Spectrometry.* Journal of the American Chemical Society, 1991. **113**(20): p. 7818-7819.
95. Videler, H., et al., *Mass spectrometry of intact ribosomes.* FEBS Lett, 2005. **579**(4): p. 943-7.
96. Gerstberger, S., M. Hafner, and T. Tuschl, *A census of human RNA-binding proteins.* Nat Rev Genet, 2014. **15**(12): p. 829-45.
97. Maris, C., C. Dominguez, and F.H. Allain, *The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression.* FEBS J, 2005. **272**(9): p. 2118-31.
98. Valverde, R., L. Edwards, and L. Regan, *Structure and function of KH domains.* FEBS J, 2008. **275**(11): p. 2712-26.
99. Quintal, S.M., Q.A. dePaula, and N.P. Farrell, *Zinc finger proteins as templates for metal ion exchange and ligand reactivity. Chemical and biological consequences.* Metallomics, 2011. **3**(2): p. 121-39.
100. Aravind, L. and E.V. Koonin, *G-patch: a new conserved domain in eukaryotic RNA-processing proteins and type D retroviral polyproteins.* Trends Biochem Sci, 1999. **24**(9): p. 342-4.
101. Hermann, H., et al., *snRNP Sm proteins share two evolutionarily conserved sequence motifs which are involved in Sm protein-protein interactions.* EMBO J, 1995. **14**(9): p. 2076-88.
102. Schumacher, M.A., et al., *Structures of the pleiotropic translational regulator Hfq and an Hfq-RNA complex: a bacterial Sm-like protein.* EMBO J, 2002. **21**(13): p. 3546-56.
103. Hrle, A., et al., *Structural analyses of the CRISPR protein Csc2 reveal the RNA-binding interface of the type I-D Cas7 family.* RNA Biol, 2014. **11**(8).
104. Duss, O., et al., *Structural basis of the non-coding RNA RsmZ acting as a protein sponge.* Nature, 2014. **509**(7502): p. 588-92.
105. Jackson, R.N., et al., *Structural biology. Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli.* Science, 2014. **345**(6203): p. 1473-9.
106. Mulepati, S., A. Heroux, and S. Bailey, *Structural biology. Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target.* Science, 2014. **345**(6203): p. 1479-84.
107. Urlaub, H., E. Kuhn-Holsken, and R. Luhrmann, *Analyzing RNA-protein crosslinking sites in unlabeled ribonucleoprotein complexes by mass spectrometry.* Methods Mol Biol, 2008. **488**: p. 221-45.
108. Kramer, K., et al., *Photo-cross-linking and high resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins.* Nat Methods, 2014. **11**(10): p. 1064-70.

109. Kramer, K., et al., *Mass-spectrometric analysis of proteins cross-linked to 4-thio-uracil- and 5-bromo-uracil-substituted RNA.* International Journal of Mass Spectrometry, 2011. **304**(2-3): p. 184-194.
110. Kuhn-Holsken, E., et al., *Improved identification of enriched peptide RNA cross-links from ribonucleoprotein particles (RNPs) by mass spectrometry.* Nucleic Acids Res, 2007. **35**(15): p. e95.
111. Naldrett, M.J., et al., *Concentration and desalting of peptide and protein samples with a newly developed C18 membrane in a microspin column format.* J Biomol Tech, 2005. **16**(4): p. 423-8.
112. Larsen, M.R., et al., *Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns.* Mol Cell Proteomics, 2005. **4**(7): p. 873-86.
113. Pinkse, M.W., S. Lemeer, and A.J. Heck, *A protocol on the use of titanium dioxide chromatography for phosphoproteomics.* Methods Mol Biol, 2011. **753**: p. 215-28.
114. Urlaub, H., et al., *Protein-rRNA binding features and their structural and functional implications in ribosomes as determined by cross-linking studies.* EMBO J, 1995. **14**(18): p. 4578-88.
115. Bertsch, A., et al., *OpenMS and TOPP: open source software for LC-MS data analysis.* Methods Mol Biol, 2011. **696**: p. 353-67.
116. Sturm, M., et al., *OpenMS - an open-source software framework for mass spectrometry.* BMC Bioinformatics, 2008. **9**: p. 163.
117. Geer, L.Y., et al., *Open mass spectrometry search algorithm.* J Proteome Res, 2004. **3**(5): p. 958-64.
118. Hermanson, G.T., *Bioconjugate Techniques, 2nd Edition.* Academic Press, 2008.
119. Sinz, A., *Chemical cross-linking and mass spectrometry for mapping three-dimensional structures of proteins and protein complexes.* J Mass Spectrom, 2003. **38**(12): p. 1225-37.
120. Sinz, A., *Investigation of protein-protein interactions in living cells by chemical crosslinking and mass spectrometry.* Anal Bioanal Chem, 2010. **397**(8): p. 3433-40.
121. Fritzsche, R., et al., *Optimizing the enrichment of cross-linked products for mass spectrometric protein analysis.* Rapid Commun Mass Spectrom, 2012. **26**(6): p. 653-8.
122. Hofele, R.V., *Structural investigation of protein (-RNA) assemblies by mass spectrometry.* PhD Thesis, George-August-Universität Göttingen, 2013: p. 21.
123. Yang, B., et al., *Identification of cross-linked peptides from complex samples.* Nat Methods, 2012. **9**(9): p. 904-6.
124. Maier, L.K., et al., *The immune system of halophilic archaea.* Mob Genet Elements, 2012. **2**(5): p. 228-232.
125. Maier, L.K., et al., *The ring of confidence: a haloarchaeal CRISPR/Cas system.* Biochem Soc Trans, 2013. **41**(1): p. 374-8.
126. Stoll, B., *Analyse des prokaryotischen Immunsystems CRISPR/Cas Typ I-B imarchaealen Modellorganismus Haloferax volcanii.* PhD Thesis, Universität Ulm., 2013.

127. Richter, H., et al., *Characterization of CRISPR RNA processing in Clostridium thermocellum and Methanococcus maripaludis.* Nucleic Acids Res, 2012. **40**(19): p. 9887-96.
128. Hale, C.R., et al., *RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex.* Cell, 2009. **139**(5): p. 945-56.
129. Fischer, S., et al., *An archaeal immune system can detect multiple protospacer adjacent motifs (PAMs) to target invader DNA.* J Biol Chem, 2012. **287**(40): p. 33351-63.
130. Allers, T. and M. Mevarech, *Archaeal genetics - the third way.* Nat Rev Genet, 2005. **6**(1): p. 58-73.
131. Plagens, A., et al., *In vitro assembly and activity of an archaeal CRISPR-Cas type I-A Cascade interference complex.* Nucleic Acids Res, 2014. **42**(8): p. 5125-38.
132. Bradford, M.M., *A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding.* Anal Biochem, 1976. **72**: p. 248-54.
133. Neuhoff, V., et al., *Improved staining of proteins in polyacrylamide gels including isoelectric focusing gels with clear background at nanogram sensitivity using Coomassie Brilliant Blue G-250 and R-250.* Electrophoresis, 1988. **9**(6): p. 255-62.
134. Rappsilber, J., Y. Ishihama, and M. Mann, *Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics.* Anal Chem, 2003. **75**(3): p. 663-70.
135. Luo, X., et al., *Structural and functional analysis of the E. coli NusB-S10 transcription antitermination complex.* Mol Cell, 2008. **32**(6): p. 791-802.
136. Leitner, A., et al., *Expanding the chemical cross-linking toolbox by the use of multiple proteases and enrichment by size exclusion chromatography.* Mol Cell Proteomics, 2012. **11**(3): p. M111 014126.
137. Cox, J., et al., *Andromeda: a peptide search engine integrated into the MaxQuant environment.* J Proteome Res, 2011. **10**(4): p. 1794-805.
138. R-Development-Core-Team, *R: a language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2008.
139. Nikolov, M., et al., *Chromatin affinity purification and quantitative mass spectrometry defining the interactome of histone modification patterns.* Mol Cell Proteomics, 2011. **10**(11): p. M110 005371.
140. Kessner, D., et al., *ProteoWizard: open source software for rapid proteomics tools development.* Bioinformatics, 2008. **24**(21): p. 2534-6.
141. Kelley, L.A. and M.J. Sternberg, *Protein structure prediction on the Web: a case study using the Phyre server.* Nat Protoc, 2009. **4**(3): p. 363-71.
142. Krissinel, E. and K. Henrick, *Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions.* Acta Crystallogr D Biol Crystallogr, 2004. **60**(Pt 12 Pt 1): p. 2256-68.
143. Branca, R.M., et al., *HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics.* Nat Methods, 2014. **11**(1): p. 59-62.

144. Reeks, J., et al., *Structure of a dimeric crenarchaeal Cas6 enzyme with an atypical active site for CRISPR RNA processing.* Biochem J, 2013. **452**(2): p. 223-30.
145. Sashital, D.G., M. Jinek, and J.A. Doudna, *An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3.* Nat Struct Mol Biol, 2011. **18**(6): p. 680-7.
146. Gesner, E.M., et al., *Recognition and maturation of effector RNAs in a CRISPR interference pathway.* Nat Struct Mol Biol, 2011. **18**(6): p. 688-92.
147. Jiang, X.Y., J.B. Smith, and E.C. Abraham, *Identification of a MS-MS fragment diagnostic for methionine sulfoxide.* Journal of Mass Spectrometry, 1996. **31**(11): p. 1309-1310.
148. Lagerwerf, F.M., et al., *Identification of oxidized methionine in peptides.* Rapid Communications in Mass Spectrometry, 1996. **10**(15): p. 1905-1910.
149. Srikanth, R., et al., *Improved sequencing of oxidized cysteine and methionine containing peptides using electron transfer dissociation.* Journal of the American Society for Mass Spectrometry, 2007. **18**(8): p. 1499-1506.
150. Makarova, K.S., et al., *Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems.* Biol Direct, 2011. **6**: p. 38.
151. Nam, K.H., Q. Huang, and A. Ke, *Nucleic acid binding surface and dimer interface revealed by CRISPR-associated CasB protein structures.* FEBS Lett, 2012. **586**(22): p. 3956-61.
152. Pul, U., et al., *Identification and characterization of E. coli CRISPR-cas promoters and their silencing by H-NS.* Mol Microbiol, 2010. **75**(6): p. 1495-512.
153. Atanassov, I. and H. Urlaub, *Increased proteome coverage by combining PAGE and peptide isoelectric focusing: comparative study of gel-based separation approaches.* Proteomics, 2013. **13**(20): p. 2947-55.
154. Desiderio, D.M. and M. Kai, *Preparation of stable isotope-incorporated peptide internal standards for field desorption mass spectrometry quantification of peptides in biologic tissue.* Biomed Mass Spectrom, 1983. **10**(8): p. 471-9.
155. Gerber, S.A., et al., *Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS.* Proc Natl Acad Sci U S A, 2003. **100**(12): p. 6940-5.
156. Park, H.M., et al., *Crystal structure of a Cas6 paralogous protein from Pyrococcus furiosus.* Proteins, 2012. **80**(7): p. 1895-900.
157. Wang, R., et al., *The impact of CRISPR repeat sequence on structures of a Cas6 protein-RNA complex.* Protein Sci, 2012. **21**(3): p. 405-17.
158. Christian, H., et al., *Insights into the activation of the helicase Prp43 by biochemical studies and structural mass spectrometry.* Nucleic Acids Res, 2014. **42**(2): p. 1162-79.
159. Ghisolfi, L., et al., *Concerted activities of the RNA recognition and the glycine-rich C-terminal domains of nucleolin are required for efficient complex formation with pre-ribosomal RNA.* Eur J Biochem, 1992. **209**(2): p. 541-8.
160. Kuhn-Holsken, E., et al., *Complete MALDI-ToF MS analysis of cross-linked peptide-RNA oligonucleotides derived from nonlabeled UV-irradiated ribonucleoprotein particles.* RNA, 2005. **11**(12): p. 1915-30.

161. Schmitzova, J., et al., *Crystal structure of Cwc2 reveals a novel architecture of a multipartite RNA-binding protein.* EMBO J, 2012. **31**(9): p. 2222-34.
162. Shetlar, M.D., et al., *Photochemical addition of amino acids and peptides to polyuridylic acid.* Photochem Photobiol, 1984. **39**(2): p. 141-4.
163. Shetlar, M.D., et al., *Photochemical addition of amino acids and peptides to homopolyribonucleotides of the major DNA bases.* Photochem Photobiol, 1984. **39**(2): p. 135-40.
164. Andersen, T.E., F. Kirpekar, and K.F. Haselmann, *RNA fragmentation in MALDI mass spectrometry studied by H/D-exchange: mechanisms of general applicability to nucleic acids.* J Am Soc Mass Spectrom, 2006. **17**(10): p. 1353-68.

6. Appendix

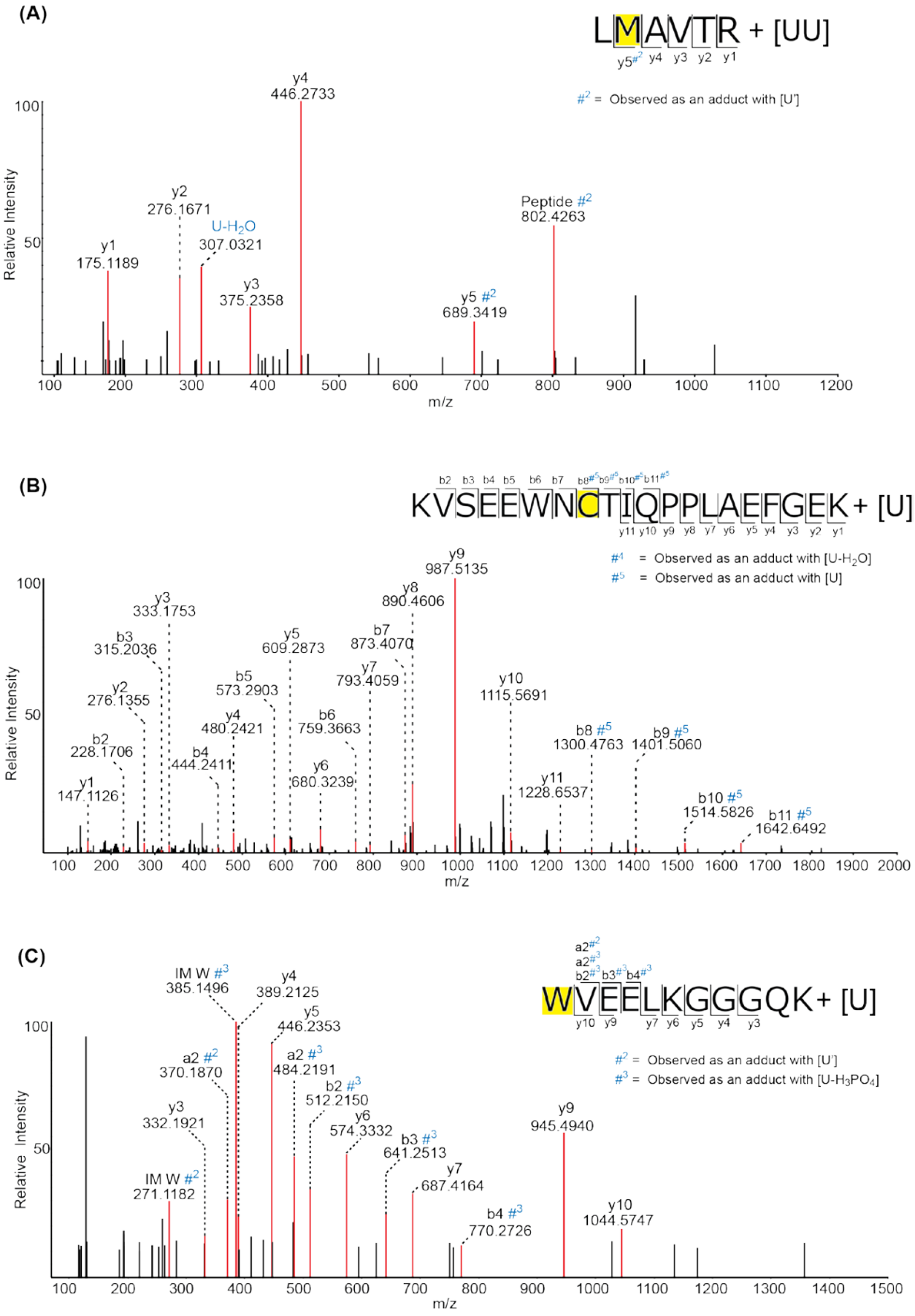
6.1 Additional Information

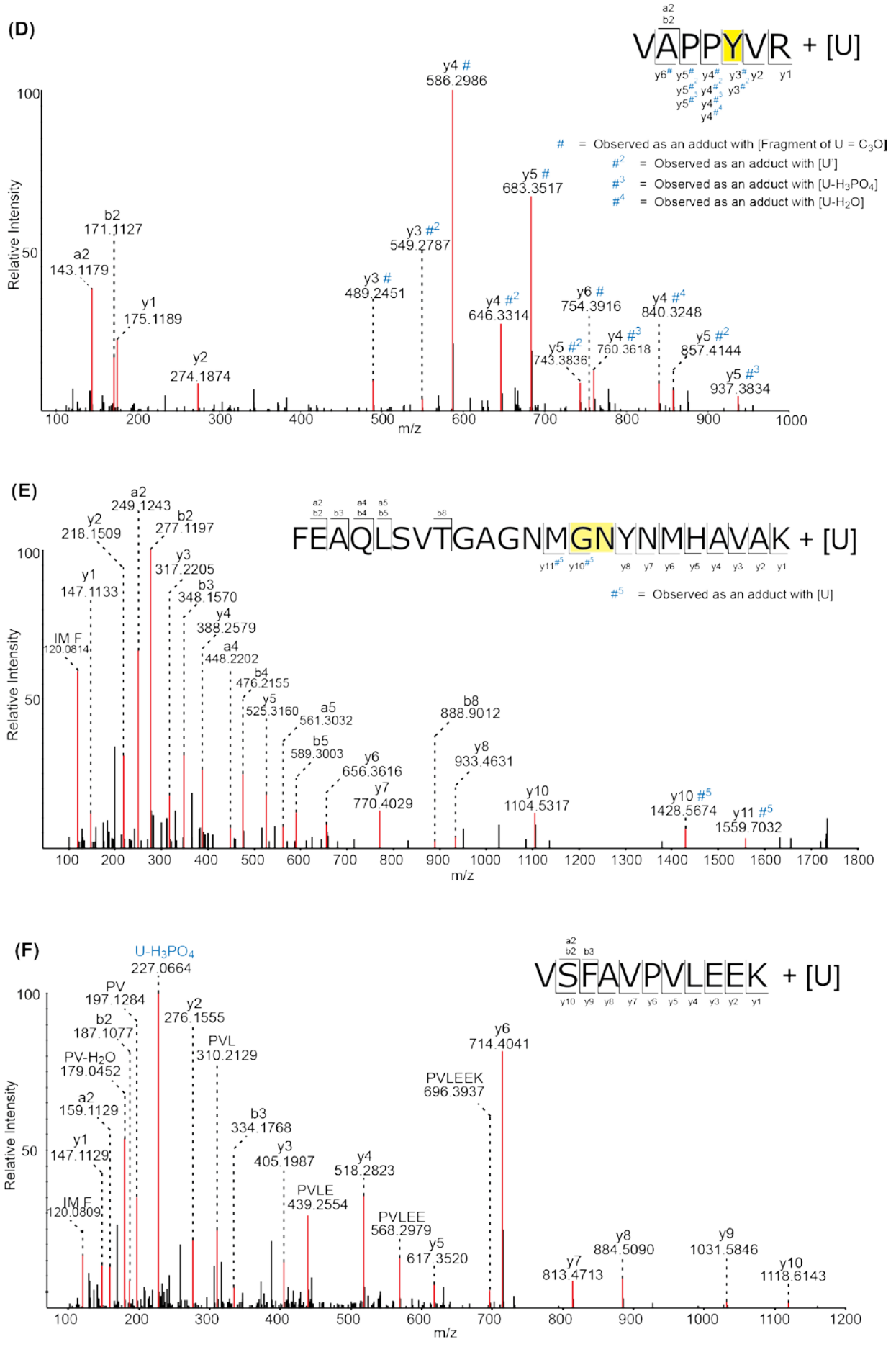
Figure 6.1 Protein-RNA cross-link spectra identified in *T. tenax* Cas7 cross-linking with poly(U)$_{15}$ and *T. pendens* Csc2 cross-linking with poly(U)$_{15}$.

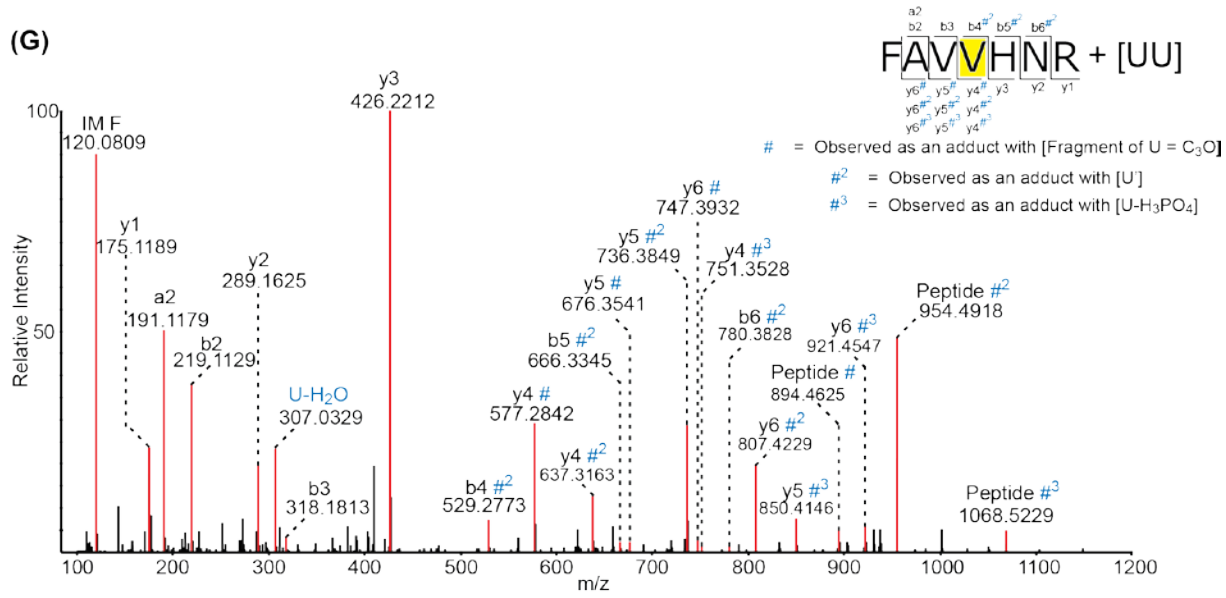
An overview of the cross-links identified and the corresponding mass values is provided in the table below. In each spectrum, the cross-linked peptide sequence and its corresponding y- and b- type fragment ions are indicated at the top. These refer to ions which retain the charge on the N-terminus or C-terminus, respectively. All the fragment ion peaks are marked with their corresponding m/z values. Ions with a mass shift corresponding to the cross-linked nucleotides are indicated with #: C_3O, #^{2}: U', #^{3}: U-H_3PO_4, #^{4}: U-H_2O and #^{5}: U. Mass shifts in the sequence tags help identify the site of cross-linking and are indicated for the corresponding fragments. The cross-linked amino acid highlighted in yellow. Adducts or maker ions corresponding to RNA component of the cross-link are indicated in blue. U': Base of U, 112.02 Da, IM: Immonium ion.

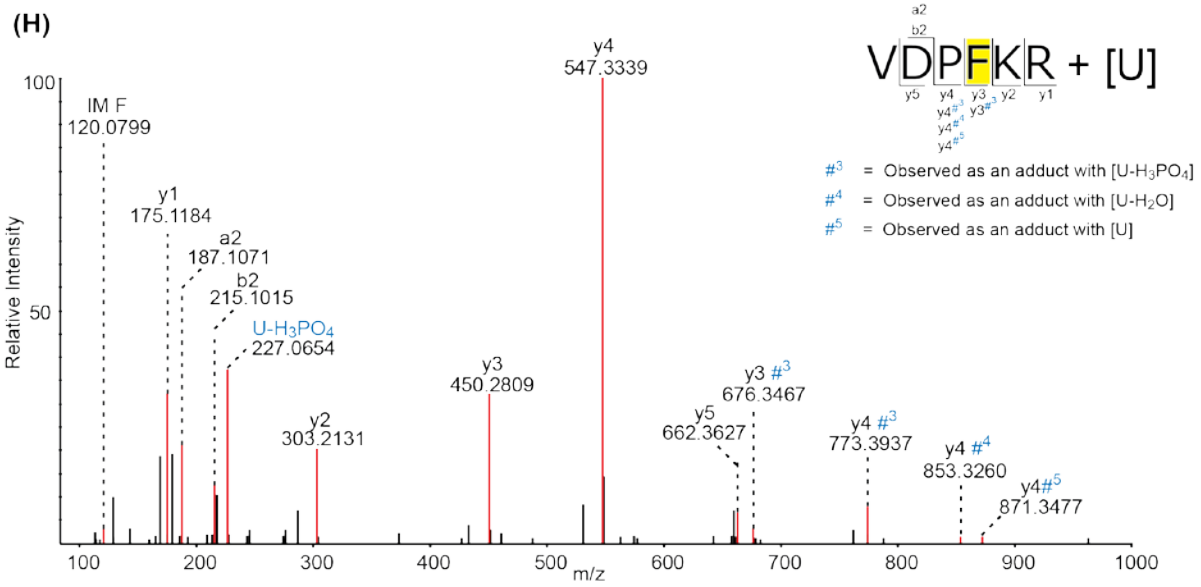
Protein (Uniprot ID)	Peptide	aa	RNA	m(Peptide)	m(RNA)	m(XL) Calc	*z*	*m/z*	m(XL) Exp	Fig
Tp Csc2 (A1RZU2)	^{82}LMAVTR^{87}	M^{83}	UU	689.3894	630.0612	1319.4596	2	660.7323	1319.4490	(A)
	^{124}KVSEEWNCTIQPPLAEFGEK^{143}	C^{131}	U	2304.1905	324.0359	2628.2246	3	877.0552	2628.1442	(B)
	^{346}WVEELKGGGQK^{356}	W^{346}	U	1229.6404	324.0359	1553.6763	2	777.8446	1553.6736	(C)
T. tenax Cas7 (G4RJZ1)	^{3}VAPPYVR^{9}	Y^{7}	U	800.4544	324.0359	1124.4903	2	536.2520	1124.4884	(D)
	^{14}FEAQLSVLTGAGNMGNYNMHAVAK^{37}	^{28}G-N^{29}	U	2522.2045	324.0359	2846.2404	3	949.7532	2846.2362	(E)
	^{127}VSFAVPVLEEK^{137}	-	U	1216.6702	324.0359	1540.7061	2	771.3601	1540.7046	(F)
	^{145}FAVVHNR^{151}	V^{148}	UU	841.4558	630.0612	1471.5170	2	736.7654	1471.5152	(G)
	^{152}VDPFKR^{158}	F^{155}	U	760.4231	324.0359	1084.4590	2	543.2367	1084.4578	(H)
	^{164}SKEEQEGTEMMVFK^{177}	M^{173}	U	1671.7483	324.0359	1995.7842	3	666.2685	1995.7821	(I)

Protein: Name of the protein (Uniprot ID), Peptide: amino acid sequence of the cross-linked peptide, aa: position of cross-linked amino acid residue, RNA: composition of cross-linked RNA, m: mass, m(Peptide): calculated mass of cross-linked peptide, m(RNA): calculated mass of cross-linked RNA, m(XL) Calc: calculated mass of cross-link [m(Peptide)+m(RNA)], z: charge state in which cross-link was observed, *m/z*: experimentally observed mass to charge ratio, m(XL) Exp: experimentally observed mass of cross-link [(*m/z*) $*$ *z*) - ((mass of proton) $*$ *z*] and Fig: reference to figure of the annotated MS/MS fragment spectrum.


(A)
LMAVTR + [UU]
y5#2 y4 y3 y2 y1
#2 = Observed as an adduct with [U']
Relative Intensity
m/z
y4 446.2733
y2 276.1671
U-H2O 307.0321
y1 175.1189
y3 375.2358
y5 #2 689.3419
Peptide #2 802.4263
(B)
b2 b3 b4 b5 b6 b7 b8#5 b9#5 b10#5 b11#5
KVSEEWNCTIQPPLAEFGEK + [U]
y11 y10 y9 y8 y7 y6 y5 y4 y3 y2 y1
#4 = Observed as an adduct with [U-H2O]
#5 = Observed as an adduct with [U]
Relative Intensity
m/z
y9 987.5135
y8 890.4606
y3 333.1753
b3 315.2036
y5 609.2873
b7 873.4070
y7 793.4059
y2 276.1355
b5 573.2903
y10 1115.5691
y4 480.2421
b6 759.3663
b8 #5 1300.4763
b9 #5 1401.5060
b2 228.1706
b4 444.2411
y6 680.3239
y1 147.1126
y11 1228.6537
b10 #5 1514.5826
b11 #5 1642.6492
(C)
a2#2
a2#3
b2#3 b3#3 b4#3
WVEELKGGGQK + [U]
y10 y9 y7 y6 y5 y4 y3
#2 = Observed as an adduct with [U']
#3 = Observed as an adduct with [U-H3PO4]
Relative Intensity
m/z
IM W #3 385.1496
y4 389.2125
y5 446.2353
a2 #2 370.1870
a2 #3 484.2191
b2 #3 512.2150
y3 332.1921
y6 574.3332
y9 945.4940
b3 #3 641.2513
IM W #2 271.1182
y7 687.4164
b4 #3 770.2726
y10 1044.5747


(D)
VAPPYVR + [U]
= Observed as an adduct with [Fragment of U = C3O]
#2 = Observed as an adduct with [U]
#3 = Observed as an adduct with [U-H3PO4]
#4 = Observed as an adduct with [U-H2O]
Relative Intensity
m/z
(E)
FEAQLSVTGAGNMGNYNMHAVAK + [U]
#5 = Observed as an adduct with [U]
Relative Intensity
m/z
(F)
VSFAVPVLEEK + [U]
U-H3PO4
227.0664
Relative Intensity
m/z


(G)
FAVVHNR + [UU]
= Observed as an adduct with [Fragment of U = C3O]
#2 = Observed as an adduct with [U]
#3 = Observed as an adduct with [U-H3PO4]
IM F 120.0809
y3 426.2212
y1 175.1189
y2 289.1625
a2 191.1179
b2 219.1129
U-H2O 307.0329
b3 318.1813
b4 #2 529.2773
y4 # 577.2842
y4 #2 637.3163
b5 #2 666.3345
y5 # 676.3541
y5 #2 736.3849
y6 # 747.3932
y4 #3 751.3528
b6 #2 780.3828
Peptide #2 954.4918
y6 #3 921.4547
Peptide # 894.4625
y6 #2 807.4229
y5 #3 850.4146
Peptide #3 1068.5229
Relative Intensity
m/z


(H)
VDPFKR + [U]
#3 = Observed as an adduct with [U-H3PO4]
#4 = Observed as an adduct with [U-H2O]
#5 = Observed as an adduct with [U]
IM F 120.0799
y4 547.3339
y1 175.1184
a2 187.1071
b2 215.1015
U-H3PO4 227.0654
y3 450.2809
y2 303.2131
y3 #3 676.3467
y5 662.3627
y4 #3 773.3937
y4 #4 853.3260
y4#5 871.3477
Relative Intensity
m/z

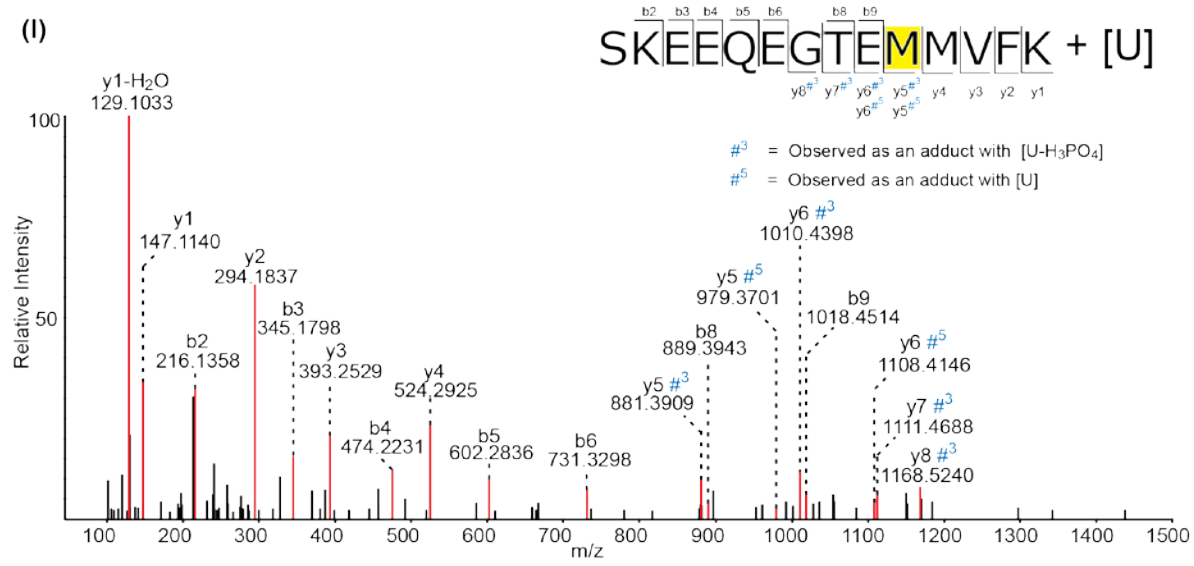
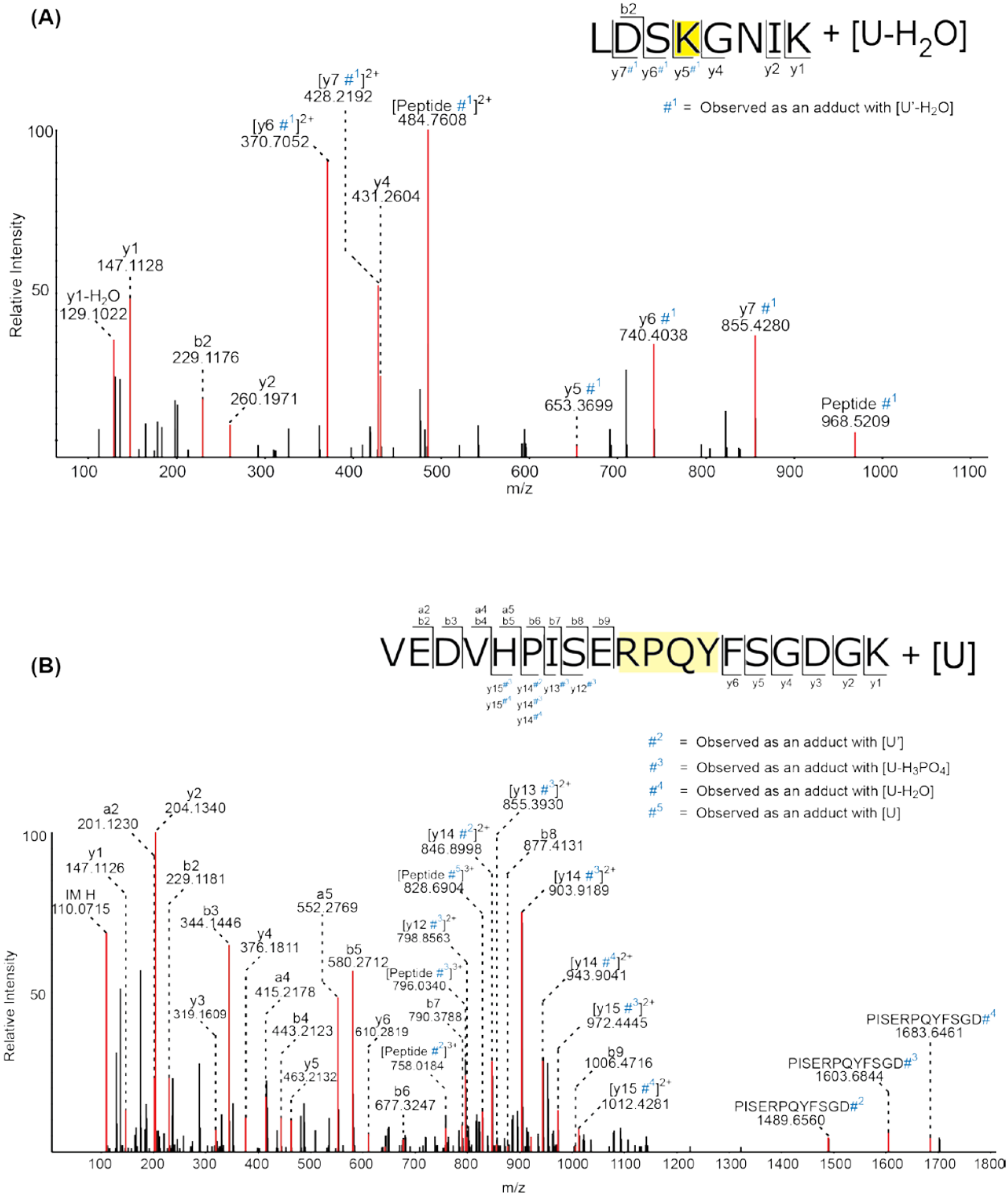

(I)
SKEEQEGTEMMVFK + [U]
#3 = Observed as an adduct with [U-H3PO4]
#5 = Observed as an adduct with [U]
y1-H2O 129.1033
y1 147.1140
y2 294.1837
b3 345.1798
b2 216.1358
y3 393.2529
y4 524.2925
b4 474.2231
b5 602.2836
b6 731.3298
y5 #3 881.3909
b8 889.3943
y5 #5 979.3701
y6 #3 1010.4398
b9 1018.4514
y6 #5 1108.4146
y7 #3 1111.4688
y8 #3 1168.5240
Relative Intensity
m/z

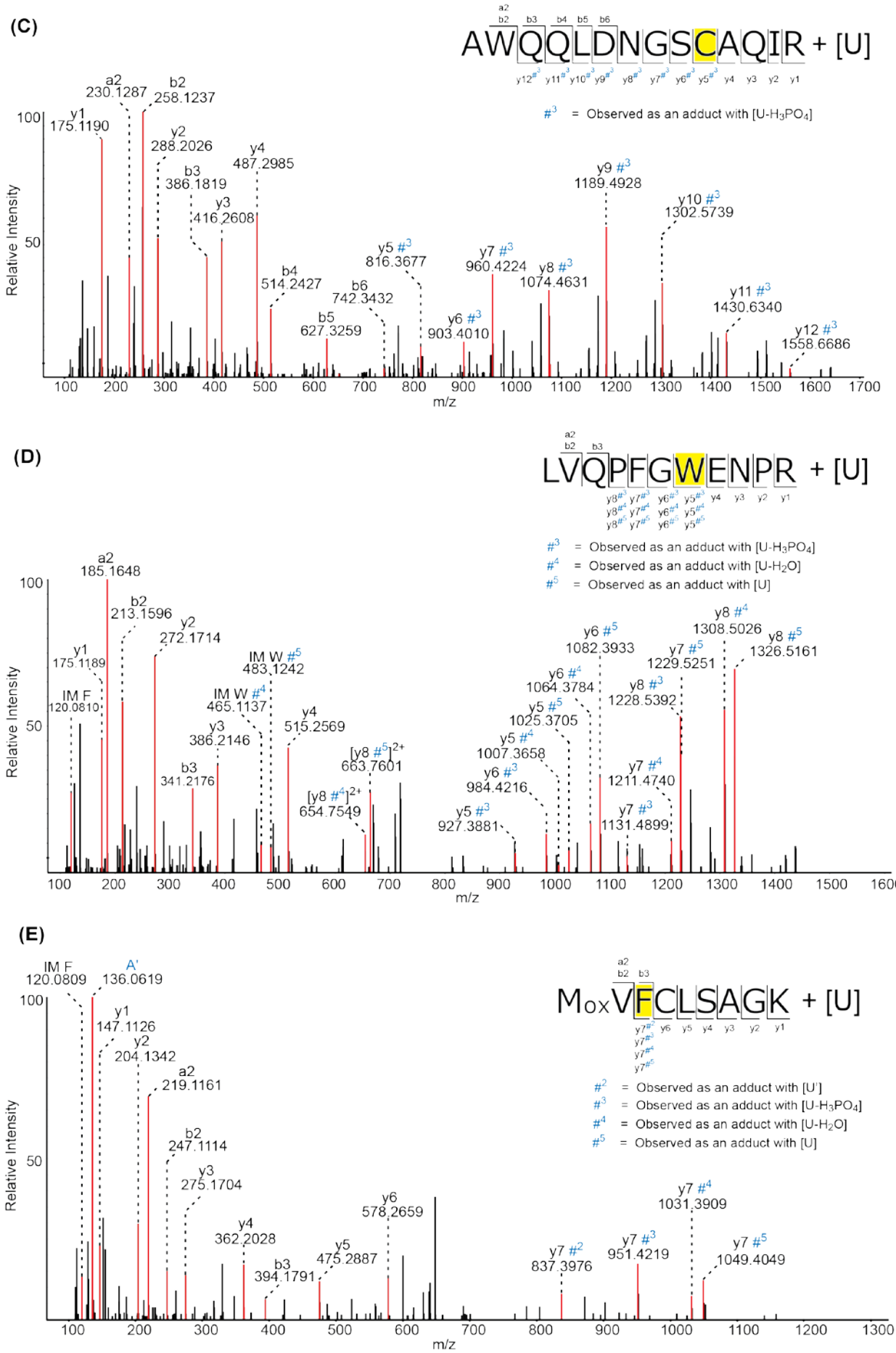
Figure 6.2 Protein-RNA cross-link spectra identified in Type I-E *E. coli* Cascade complex.

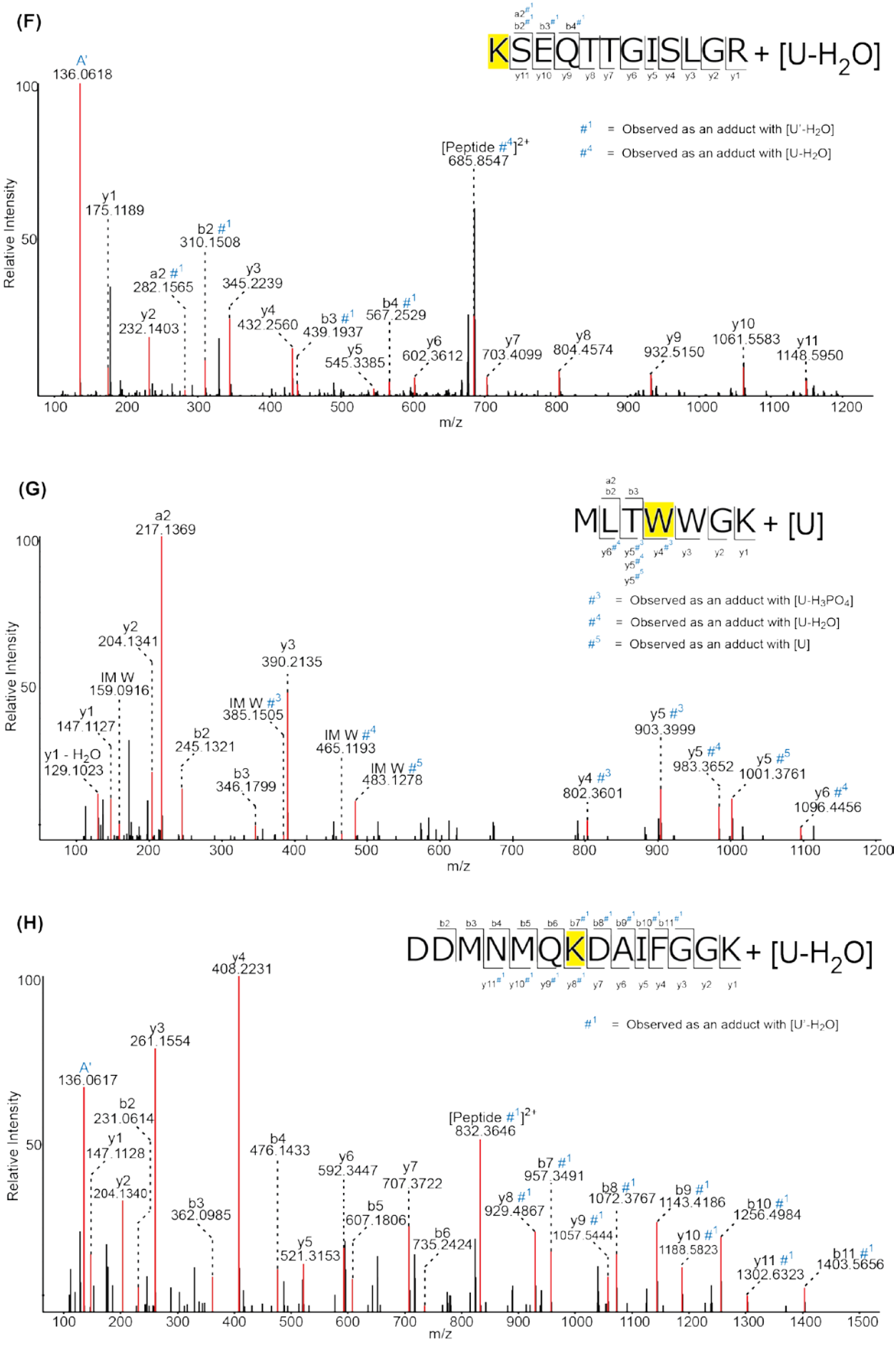
An overview of the cross-links identified and the corresponding mass values is provided in the table below. In each spectrum, the cross-linked peptide sequence and its corresponding y- and b- type fragment ions are indicated at the top. These refer to ions which retain the charge on the N-terminus or C-terminus, respectively. All the fragment ion peaks are marked with their corresponding m/z values. Ions with a mass shift corresponding to the cross-linked nucleotides are indicated with #: C_3O, #[1]: U'-H_2O, #[2]: U', #[3]: U-H_3PO_4, #[4]: U-H_2O and #[5]: U. Mass shifts in the sequence tags help identify the site of cross-linking and are indicated for the corresponding fragments. The cross-linked amino acid highlighted in yellow. Adducts or maker ions corresponding to RNA component of the cross-link are indicated in blue. U': Base of U, 112.02 Da; A': Base of A, 136.06 Da; C': Base of C, 112.05 Da, G': Base of G, 152.05 Da, IM: Immonium ion.

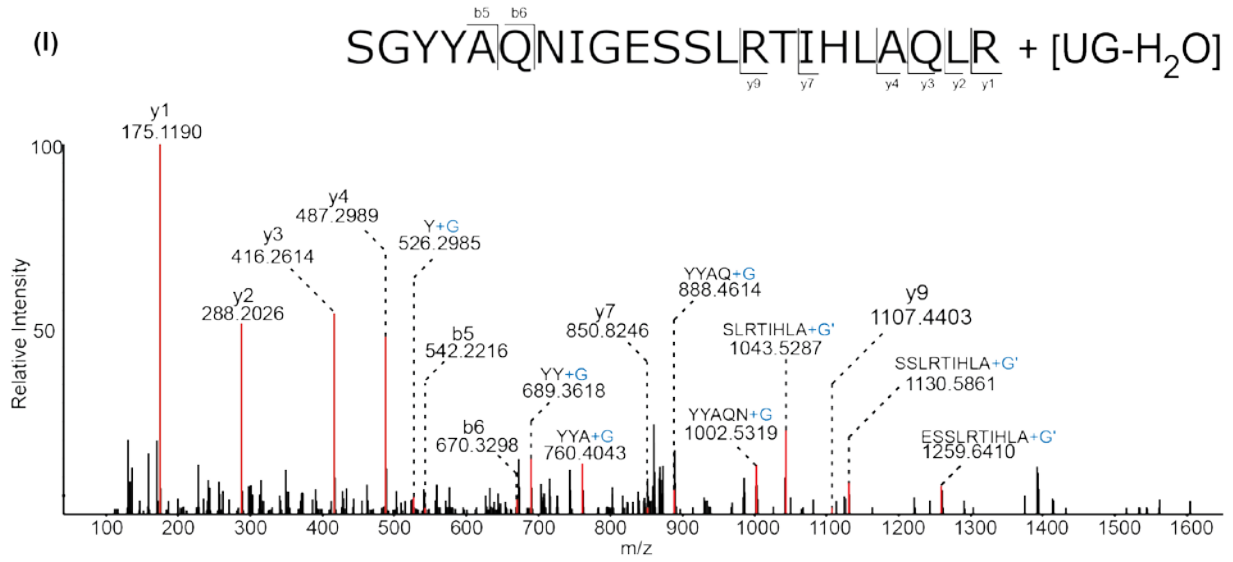
Protein (Uniprot ID)	Peptide	aa	RNA	m(Peptide)	m(RNA)	m(XL) Calc	*z*	*m/z*	m(XL) Exp	Fig
Cas6e (Q46897)	^{103}LDSKGNIK^{110}	K^{106}	U-H_2O	873.4919	306.0253	1179.5172	2	590.7661	1179.5166	(A)
	^{136}VEDVHPISERPQYFSGDGK^{154}	^{145}R-Y^{148}	U	2159.0823	324.0359	2483.1182	3	828.6961	2483.0649	(B)
Cse2 (P76632)	^{13}AWQQLDNGSCAQIR^{26}	C^{22}	U	1588.7416	324.0359	1912.7775	2	957.3967	1912.7778	(C)
	^{43}LVQPFGWENPR^{53}	W^{49}	U	1341.6829	324.0359	1665.7188	2	833.8672	1665.7188	(D)
	^{61}M(Ox)VFCLSAGK^{69}	F^{63}	U	970.4616	324.0359	1294.4975	2	648.2559	1294.4962	(E)
	^{78}KSEQTTGISLGR^{89}	K^{78}	U-H_2O	1275.6782	306.0253	1581.7035	2	791.8590	1581.7024	(F)
	^{136}MLTWWGK^{142}	W^{139}	U	920.4578	324.0359	1244.4937	2	623.2540	1244.4924	(G)
Cas7 (I2ZSV0)	^{21}DDMNMQKDAIFGGK^{34}	K^{27}	U-H_2O	1568.6962	306.0253	1874.7215	2	938.3693	1874.7230	(H)
	^{51}SGYYAQNIGESSLRTIHLAQLR^{72}	-	UG-H_2O	2476.2822	651.0727	3127.3549	3	1043.4525	3127.3341	(I)
	^{83}FDQKIIDK^{90}	K^{86}	U-H_2O	1005.5494	306.0253	1311.5747	2	656.7949	1311.5742	(J)
	^{91}TLALLSGKSVDEAEK^{105}	^{97}G-K^{98}	U-H_2O	1559.8406	306.0253	1865.8659	2	933.9406	1865.8656	(K)
	^{129}AEADNLDDKK^{138}	K^{137}	U-H_2O	1117.5250	306.0253	1423.5503	2	712.7828	1423.5500	(L)
	^{139}LLKVLK^{144}	K^{141}	U-H_2O	712.5210	306.0253	1018.5463	2	510.2805	1018.5454	(M)
	^{142}VLKEDIAAIR^{151}	K^{144}	U-H_2O	1126.6709	306.0253	1432.6962	2	717.3553	1432.6950	(N)
	^{166}MATSGMMTELGK^{177}	M^{166}	U	1255.5610	324.0359	1579.5969	2	790.8062	1579.5968	(O)
Cas5e (H0Q9G2)	^{9}LAGPMQAWGQPTFEGTRPTGR^{29}	W^{16}	U	2257.1061	324.0359	2581.1420	3	861.3889	2581.1433	(P)
	^{84}DYHTVLGAR^{92}	Y^{85}	U	1030.5195	324.0359	1354.5554	2	678.2844	1354.5532	(Q)
	^{142}YTPYLGR^{148}	Y^{145}	UA	868.4442	653.0884	1521.5326	3	508.1855	1521.5331	(R)
	^{198}DEPMITLPR^{206}	P^{200}	U	1070.5430	324.0359	1394.5789	2	698.2974	1394.5792	(S)
Cse1(Q46901)	^{395}ALYTFAEGFK^{404}	F^{403}	U	1145.5756	324.0359	1469.6115	2	735.8139	1469.6122	(T)

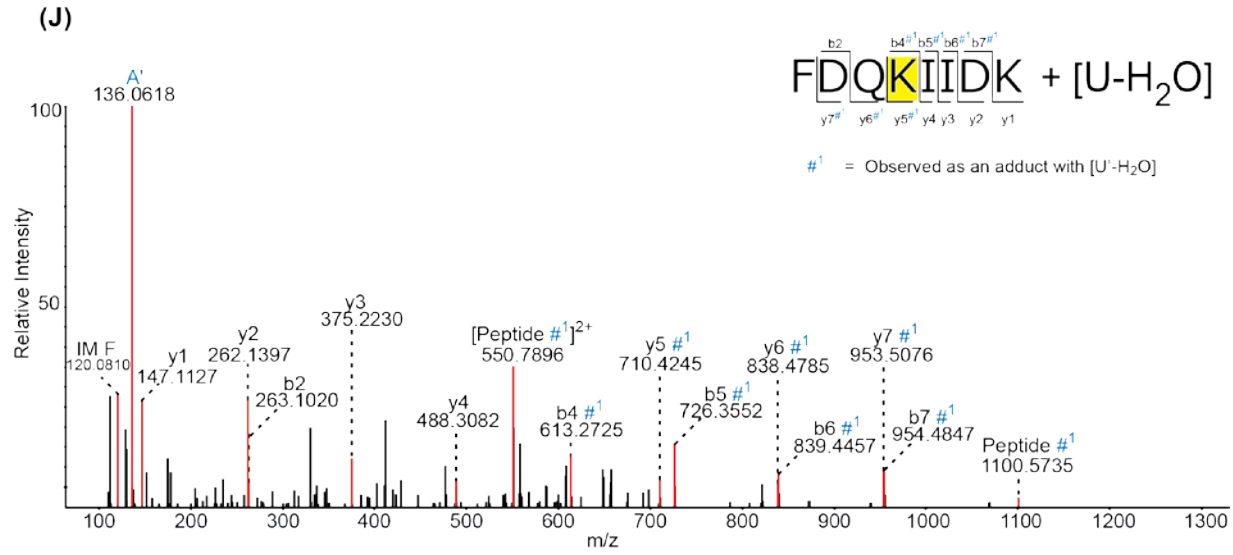
Protein: Name of the protein (Uniprot ID), Peptide: amino acid sequence of the cross-linked peptide, aa: position of cross-linked amino acid residue, RNA: composition of cross-linked RNA, m: mass, m(Peptide): calculated mass of cross-linked peptide, m(RNA): calculated mass of cross-linked RNA, m(XL) Calc: calculated mass of cross-link [m(Peptide)+m(RNA)], z: charge state in which cross-link was observed, *m/z*: experimentally observed mass to charge ratio, m(XL) Exp: experimentally observed mass of cross-link [(*m/z*) ∗ *z*) - ((mass of proton) ∗ *z*] and Fig: reference to figure of the annotated MS/MS fragment spectrum.

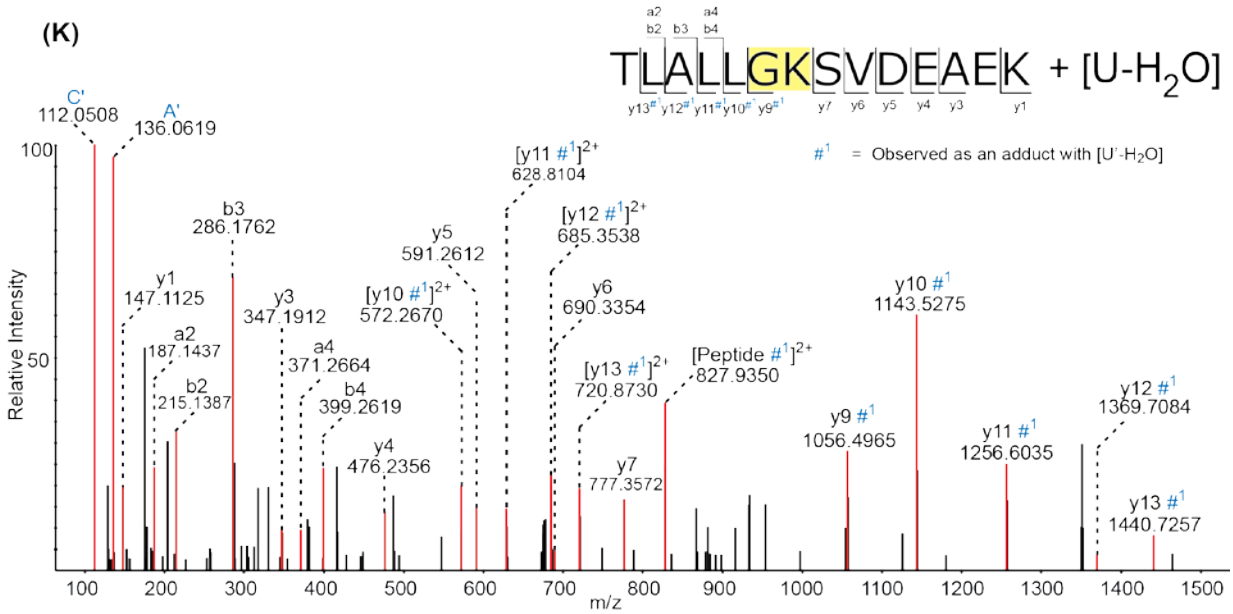

(A)
LDSKGNIK + [U-H2O]
#1 = Observed as an adduct with [U'-H2O]
[y7 #1]2+ 428.2192
[Peptide #1]2+ 484.7608
[y6 #1]2+ 370.7052
y4 431.2604
y1 147.1128
y1-H2O 129.1022
b2 229.1176
y2 260.1971
y5 #1 653.3699
y6 #1 740.4038
y7 #1 855.4280
Peptide #1 968.5209
Relative Intensity
m/z
(B)
VEDVHPISERPQYFSGDGK + [U]
#2 = Observed as an adduct with [U']
#3 = Observed as an adduct with [U-H3PO4]
#4 = Observed as an adduct with [U-H2O]
#5 = Observed as an adduct with [U]
[y13 #3]2+ 855.3930
y2 204.1340
a2 201.1230
[y14 #2]2+ 846.8998
b8 877.4131
y1 147.1126
b2 229.1181
[Peptide #5]3+ 828.6904
[y14 #3]2+ 903.9189
IM H 110.0715
b3 344.1446
a5 552.2769
[y12 #3]2+ 798.8563
y4 376.1811
b5 580.2712
[y14 #4]2+ 943.9041
[Peptide #3]3+ 796.0340
a4 415.2178
b7 790.3788
[y15 #3]2+ 972.4445
y3 319.1609
b4 443.2123
y6 610.2819
PISERPQYFSGD#4 1683.6461
[Peptide #2]3+ 758.0184
b9 1006.4716
y5 463.2132
PISERPQYFSGD#3 1603.6844
[y15 #4]2+ 1012.4281
b6 677.3247
PISERPQYFSGD#2 1489.6560
Relative Intensity
m/z

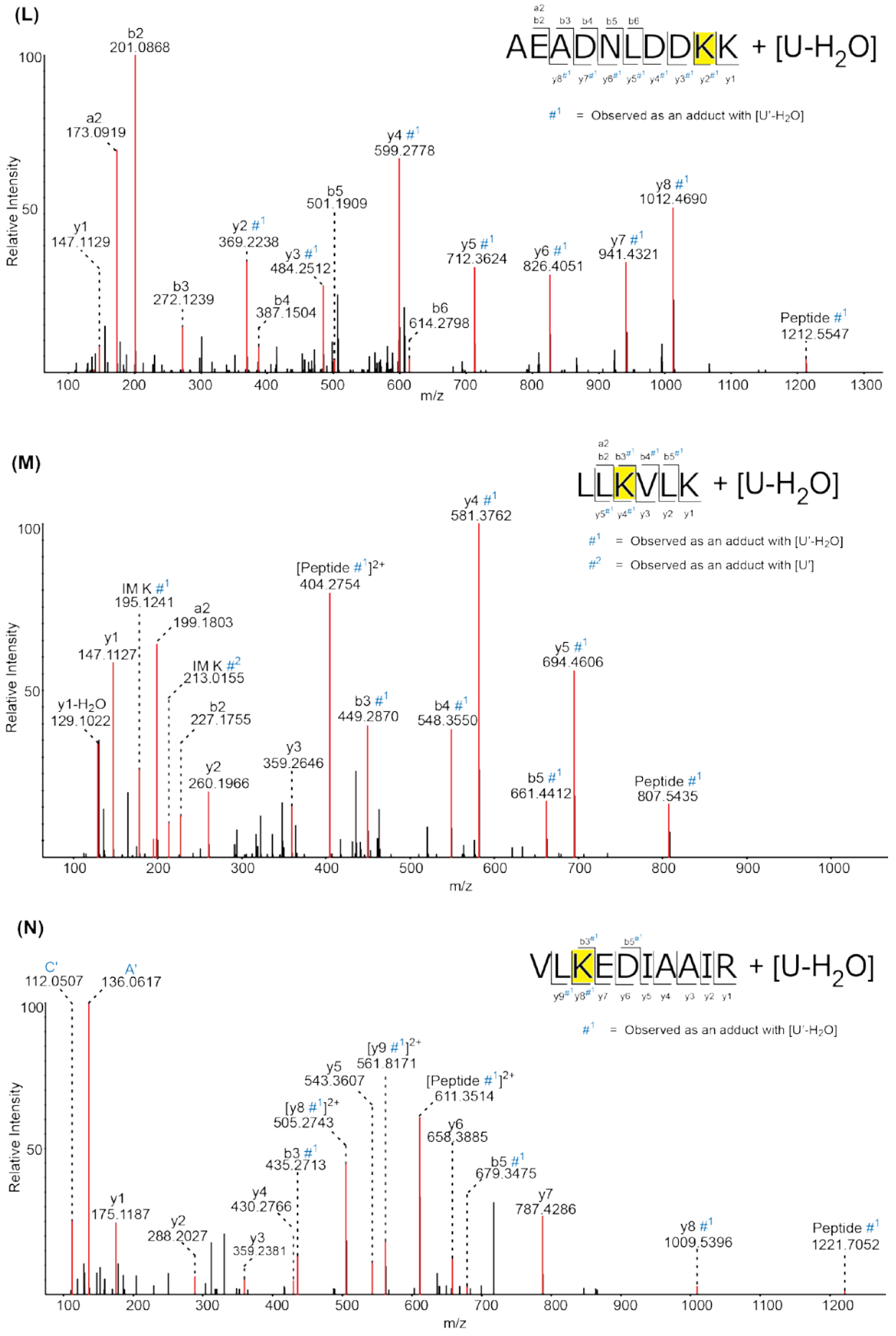

(C)
AWQQLDNGSCAQIR + [U]
#3 = Observed as an adduct with [U-H3PO4]
a2 230.1287
b2 258.1237
y1 175.1190
y2 288.2026
y4 487.2985
b3 386.1819
y3 416.2608
b4 514.2427
y5 #3 816.3677
b6 742.3432
b5 627.3259
y6 #3 903.4010
y7 #3 960.4224
y8 #3 1074.4631
y9 #3 1189.4928
y10 #3 1302.5739
y11 #3 1430.6340
y12 #3 1558.6686
Relative Intensity
m/z
(D)
LVQPFGWENPR + [U]
#3 = Observed as an adduct with [U-H3PO4]
#4 = Observed as an adduct with [U-H2O]
#5 = Observed as an adduct with [U]
a2 185.1648
b2 213.1596
y2 272.1714
y1 175.1189
IM W #5 483.1242
IM F 120.0810
IM W #4 465.1137
y4 515.2569
y3 386.2146
b3 341.2176
[y8 #5]2+ 663.7601
[y8 #4]2+ 654.7549
y5 #3 927.3881
y6 #3 984.4216
y5 #4 1007.3658
y5 #5 1025.3705
y6 #4 1064.3784
y6 #5 1082.3933
y8 #3 1228.5392
y7 #5 1229.5251
y7 #4 1211.4740
y7 #3 1131.4899
y8 #4 1308.5026
y8 #5 1326.5161
Relative Intensity
m/z
(E)
MoxVFCLSAGK + [U]
#2 = Observed as an adduct with [U']
#3 = Observed as an adduct with [U-H3PO4]
#4 = Observed as an adduct with [U-H2O]
#5 = Observed as an adduct with [U]
IM F 120.0809
A' 136.0619
y1 147.1126
y2 204.1342
a2 219.1161
b2 247.1114
y3 275.1704
y4 362.2028
b3 394.1791
y5 475.2887
y6 578.2659
y7 #2 837.3976
y7 #3 951.4219
y7 #4 1031.3909
y7 #5 1049.4049
Relative Intensity
m/z


(F)
KSEQTTGISLGR + [U-H2O]
#1 = Observed as an adduct with [U'-H2O]
#4 = Observed as an adduct with [U-H2O]
A' 136.0618
y1 175.1189
b2 #1 310.1508
a2 #1 282.1565
y3 345.2239
y2 232.1403
y4 432.2560
b3 #1 439.1937
b4 #1 567.2529
y5 545.3385
y6 602.3612
[Peptide #4]2+ 685.8547
y7 703.4099
y8 804.4574
y9 932.5150
y10 1061.5583
y11 1148.5950
Relative Intensity
m/z
(G)
MLTWWGK + [U]
#3 = Observed as an adduct with [U-H3PO4]
#4 = Observed as an adduct with [U-H2O]
#5 = Observed as an adduct with [U]
a2 217.1369
y2 204.1341
y3 390.2135
IM W 159.0916
IM W #3 385.1505
y1 147.1127
y1 - H2O 129.1023
b2 245.1321
IM W #4 465.1193
b3 346.1799
IM W #5 483.1278
y4 #3 802.3601
y5 #3 903.3999
y5 #4 983.3652
y5 #5 1001.3761
y6 #4 1096.4456
Relative Intensity
m/z
(H)
DDMNMQKDAIFGGK + [U-H2O]
#1 = Observed as an adduct with [U'-H2O]
y4 408.2231
y3 261.1554
A' 136.0617
b2 231.0614
y1 147.1128
y2 204.1340
b3 362.0985
b4 476.1433
y6 592.3447
y7 707.3722
b5 607.1806
y5 521.3153
b6 735.2424
[Peptide #1]2+ 832.3646
b7 #1 957.3491
y8 #1 929.4867
b8 #1 1072.3767
y9 #1 1057.5444
b9 #1 1143.4186
b10 #1 1256.4984
y10 #1 1188.5823
y11 #1 1302.6323
b11 #1 1403.5656
Relative Intensity
m/z

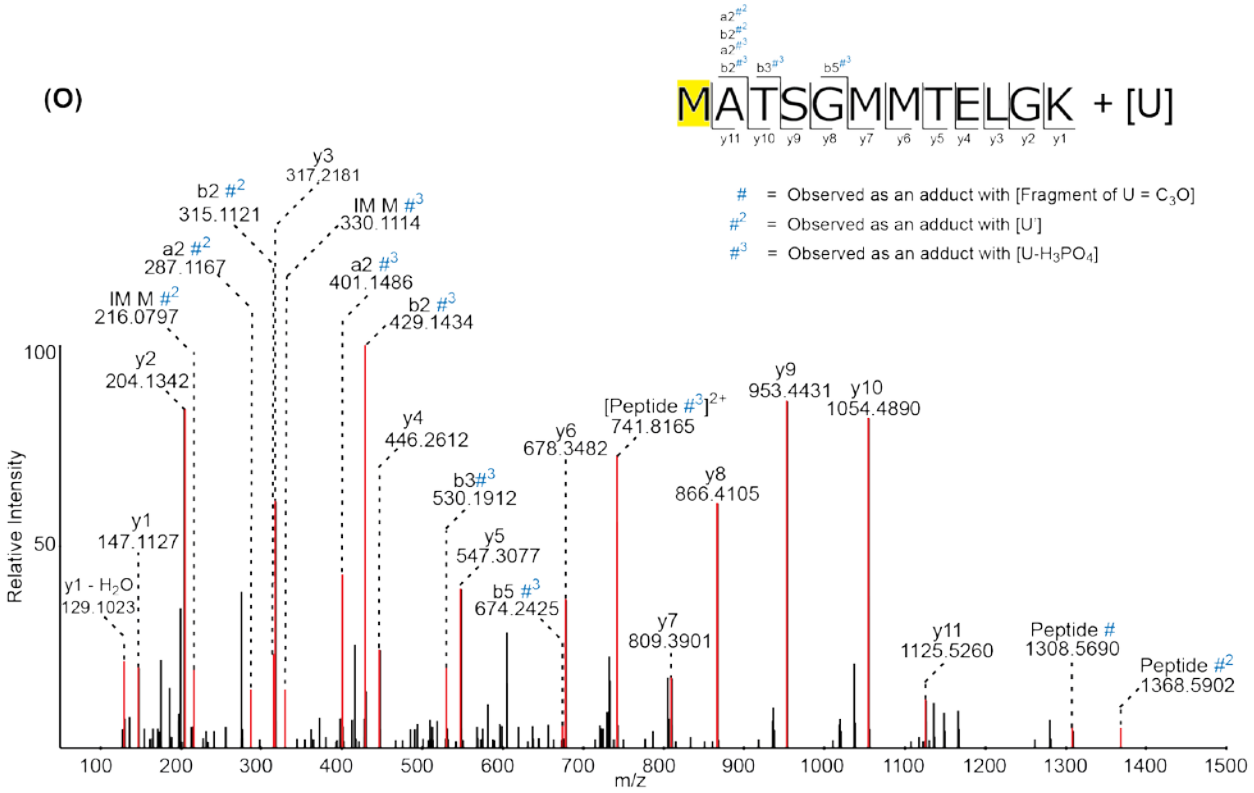

(I)
SGYYAQNIGESSLRTIHLAQLR + [UG-H2O]
Relative Intensity
m/z
y1 175.1190
y4 487.2989
y3 416.2614
Y+G 526.2985
y2 288.2026
b5 542.2216
y7 850.8246
YYAQ+G 888.4614
y9 1107.4403
SLRTIHLA+G' 1043.5287
SSLRTIHLA+G' 1130.5861
YY+G 689.3618
b6 670.3298
YYA+G 760.4043
YYAQN+G 1002.5319
ESSLRTIHLA+G' 1259.6410


(J)
FDQKIIDK + [U-H2O]
#1 = Observed as an adduct with [U'-H2O]
Relative Intensity
m/z
A' 136.0618
IM F 120.0810
y1 147.1127
y2 262.1397
y3 375.2230
b2 263.1020
y4 488.3082
[Peptide #1]2+ 550.7896
b4 #1 613.2725
y5 #1 710.4245
b5 #1 726.3552
y6 #1 838.4785
b6 #1 839.4457
y7 #1 953.5076
b7 #1 954.4847
Peptide #1 1100.5735

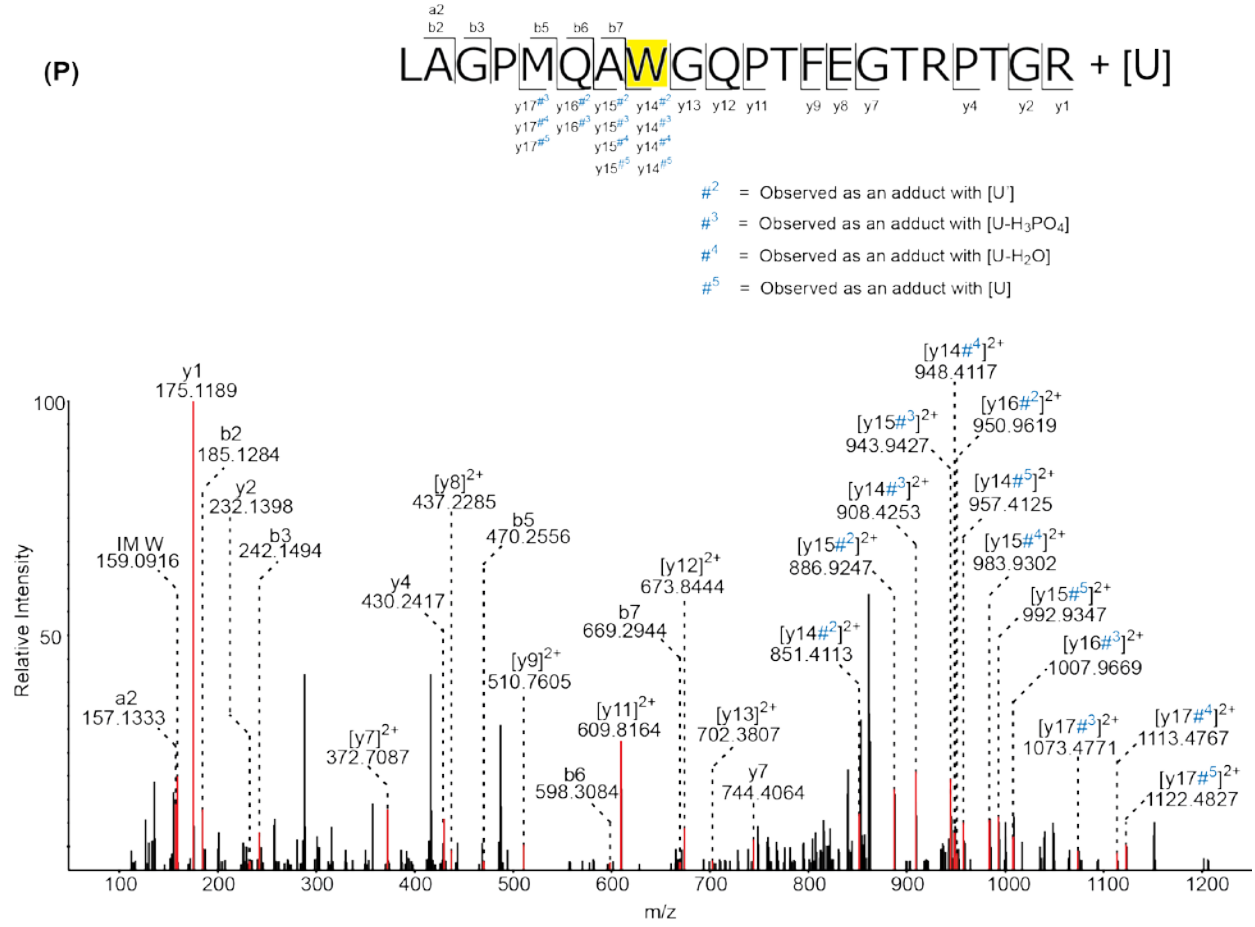

(K)
TLALLGKSVDEAEK + [U-H2O]
#1 = Observed as an adduct with [U'-H2O]
Relative Intensity
m/z
C' 112.0508
A' 136.0619
y1 147.1125
b3 286.1762
y3 347.1912
a2 187.1437
b2 215.1387
a4 371.2664
b4 399.2619
y4 476.2356
y5 591.2612
[y10 #1]2+ 572.2670
[y11 #1]2+ 628.8104
[y12 #1]2+ 685.3538
y6 690.3354
[y13 #1]2+ 720.8730
y7 777.3572
[Peptide #1]2+ 827.9350
y9 #1 1056.4965
y10 #1 1143.5275
y11 #1 1256.6035
y12 #1 1369.7084
y13 #1 1440.7257

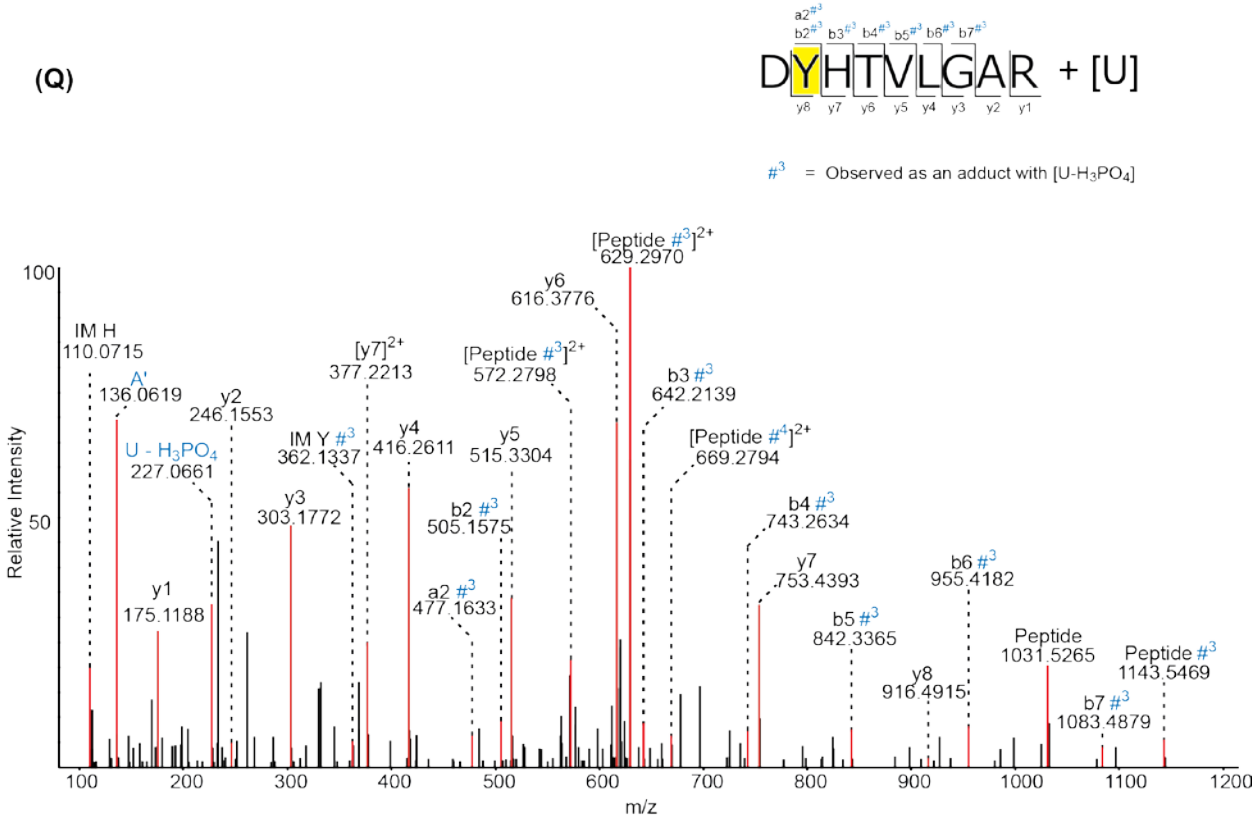

(L)
AEADNLDDKK + [U-H2O]
#1 = Observed as an adduct with [U'-H2O]
b2 201.0868
a2 173.0919
y4 #1 599.2778
b5 501.1909
y8 #1 1012.4690
y1 147.1129
y2 #1 369.2238
y3 #1 484.2512
y5 #1 712.3624
y6 #1 826.4051
y7 #1 941.4321
b3 272.1239
b4 387.1504
b6 614.2798
Peptide #1 1212.5547
Relative Intensity
m/z
(M)
LLKVLK + [U-H2O]
#1 = Observed as an adduct with [U'-H2O]
#2 = Observed as an adduct with [U']
y4 #1 581.3762
[Peptide #1]2+ 404.2754
IM K #1 195.1241
a2 199.1803
y1 147.1127
IM K #2 213.0155
y5 #1 694.4606
y1-H2O 129.1022
b2 227.1755
b3 #1 449.2870
b4 #1 548.3550
y3 359.2646
y2 260.1966
b5 #1 661.4412
Peptide #1 807.5435
Relative Intensity
m/z
(N)
VLKEDIAAIR + [U-H2O]
#1 = Observed as an adduct with [U'-H2O]
C' 112.0507
A' 136.0617
[y9 #1]2+ 561.8171
y5 543.3607
[Peptide #1]2+ 611.3514
[y8 #1]2+ 505.2743
y6 658.3885
b3 #1 435.2713
b5 #1 679.3475
y4 430.2766
y7 787.4286
y1 175.1187
y2 288.2027
y3 359.2381
y8 #1 1009.5396
Peptide #1 1221.7052
Relative Intensity
m/z

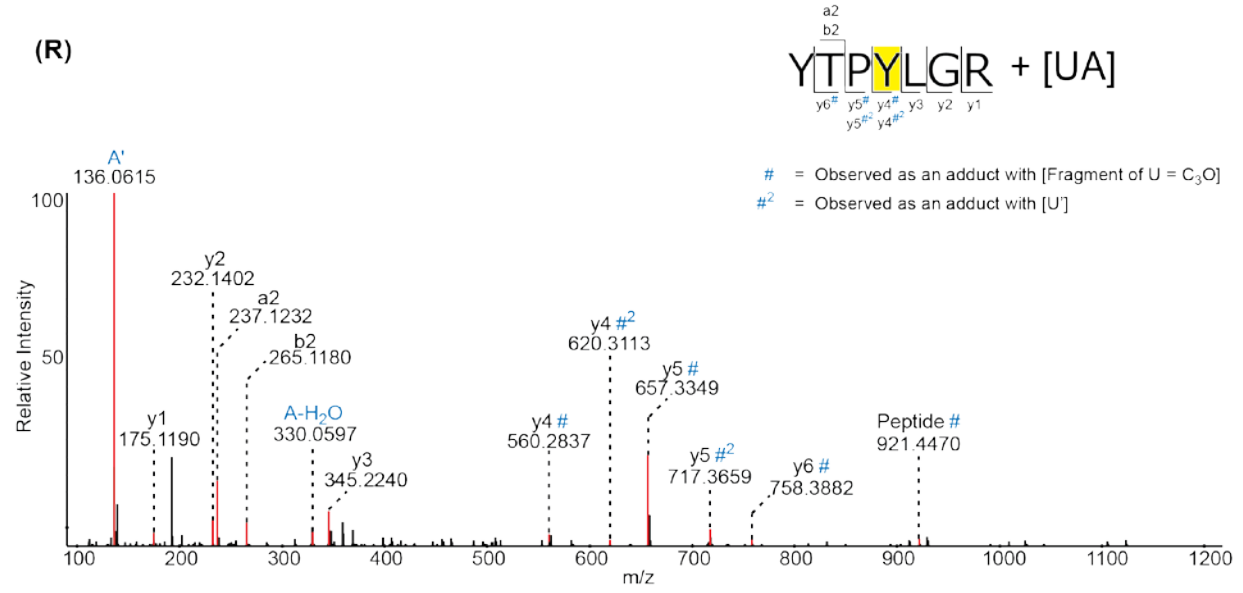
(O)

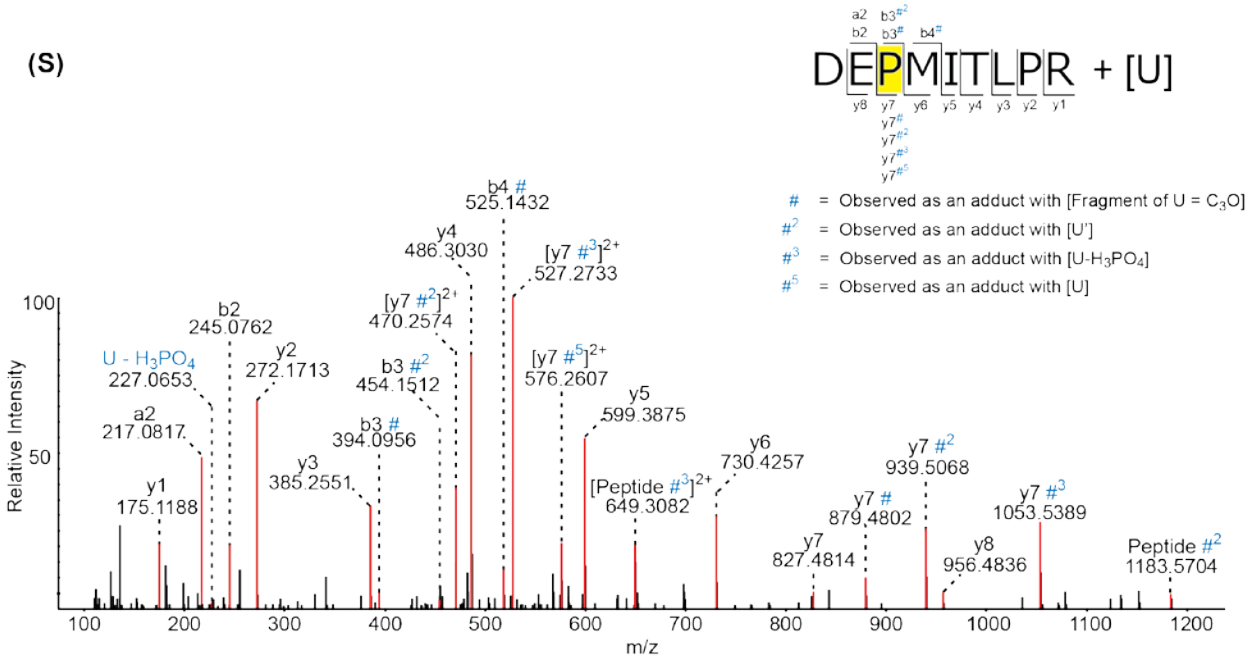

MATSGMMTELGK + [U]
= Observed as an adduct with [Fragment of U = C3O]
#2 = Observed as an adduct with [U˙]
#3 = Observed as an adduct with [U-H3PO4]
y3 317.2181
b2 #2 315.1121
IM M #3 330.1114
a2 #2 287.1167
a2 #3 401.1486
IM M #2 216.0797
b2 #3 429.1434
y2 204.1342
y4 446.2612
[Peptide #3]2+ 741.8165
y6 678.3482
y9 953.4431
y10 1054.4890
b3#3 530.1912
y8 866.4105
y1 147.1127
y5 547.3077
y1 - H2O 129.1023
b5 #3 674.2425
y7 809.3901
y11 1125.5260
Peptide # 1308.5690
Peptide #2 1368.5902
Relative Intensity
m/z

(P)


LAGPMQAWGQPTFEGTRPTGR + [U]
#2 = Observed as an adduct with [U˙]
#3 = Observed as an adduct with [U-H3PO4]
#4 = Observed as an adduct with [U-H2O]
#5 = Observed as an adduct with [U]
y1 175.1189
b2 185.1284
y2 232.1398
IM W 159.0916
b3 242.1494
[y8]2+ 437.2285
b5 470.2556
y4 430.2417
[y12]2+ 673.8444
b7 669.2944
[y9]2+ 510.7605
a2 157.1333
[y7]2+ 372.7087
[y11]2+ 609.8164
b6 598.3084
[y13]2+ 702.3807
y7 744.4064
[y14#4]2+ 948.4117
[y15#3]2+ 943.9427
[y16#2]2+ 950.9619
[y14#3]2+ 908.4253
[y14#5]2+ 957.4125
[y15#2]2+ 886.9247
[y15#4]2+ 983.9302
[y15#5]2+ 992.9347
[y14#2]2+ 851.4113
[y16#3]2+ 1007.9669
[y17#3]2+ 1073.4771
[y17#4]2+ 1113.4767
[y17#5]2+ 1122.4827
Relative Intensity
m/z


(Q)
DYHTVLGAR + [U]
#3 = Observed as an adduct with [U-H3PO4]
Relative Intensity
m/z
IM H 110.0715
A' 136.0619
y2 246.1553
U - H3PO4 227.0661
y1 175.1188
y3 303.1772
IM Y #3 362.1337
[y7]2+ 377.2213
y4 416.2611
y5 515.3304
b2 #3 505.1575
a2 #3 477.1633
[Peptide #3]2+ 572.2798
[Peptide #3]2+ 629.2970
y6 616.3776
b3 #3 642.2139
[Peptide #4]2+ 669.2794
b4 #3 743.2634
y7 753.4393
b5 #3 842.3365
y8 916.4915
b6 #3 955.4182
Peptide 1031.5265
b7 #3 1083.4879
Peptide #3 1143.5469


(R)
YTPYLGR + [UA]
= Observed as an adduct with [Fragment of U = C3O]
#2 = Observed as an adduct with [U']
Relative Intensity
m/z
A' 136.0615
y2 232.1402
a2 237.1232
b2 265.1180
y1 175.1190
A-H2O 330.0597
y3 345.2240
y4 # 560.2837
y4 #2 620.3113
y5 # 657.3349
y5 #2 717.3659
y6 # 758.3882
Peptide # 921.4470


(S)
DEPMITLPR + [U]
= Observed as an adduct with [Fragment of U = C3O]
#2 = Observed as an adduct with [U']
#3 = Observed as an adduct with [U-H3PO4]
#5 = Observed as an adduct with [U]
b4 # 525.1432
y4 486.3030
[y7 #3]2+ 527.2733
[y7 #2]2+ 470.2574
b2 245.0762
U - H3PO4 227.0653
y2 272.1713
b3 #2 454.1512
[y7 #5]2+ 576.2607
y5 599.3875
a2 217.0817
b3 # 394.0956
y3 385.2551
y6 730.4257
y7 #2 939.5068
y1 175.1188
[Peptide #3]2+ 649.3082
y7 # 879.4802
y7 #3 1053.5389
y7 827.4814
y8 956.4836
Peptide #2 1183.5704
Relative Intensity
m/z

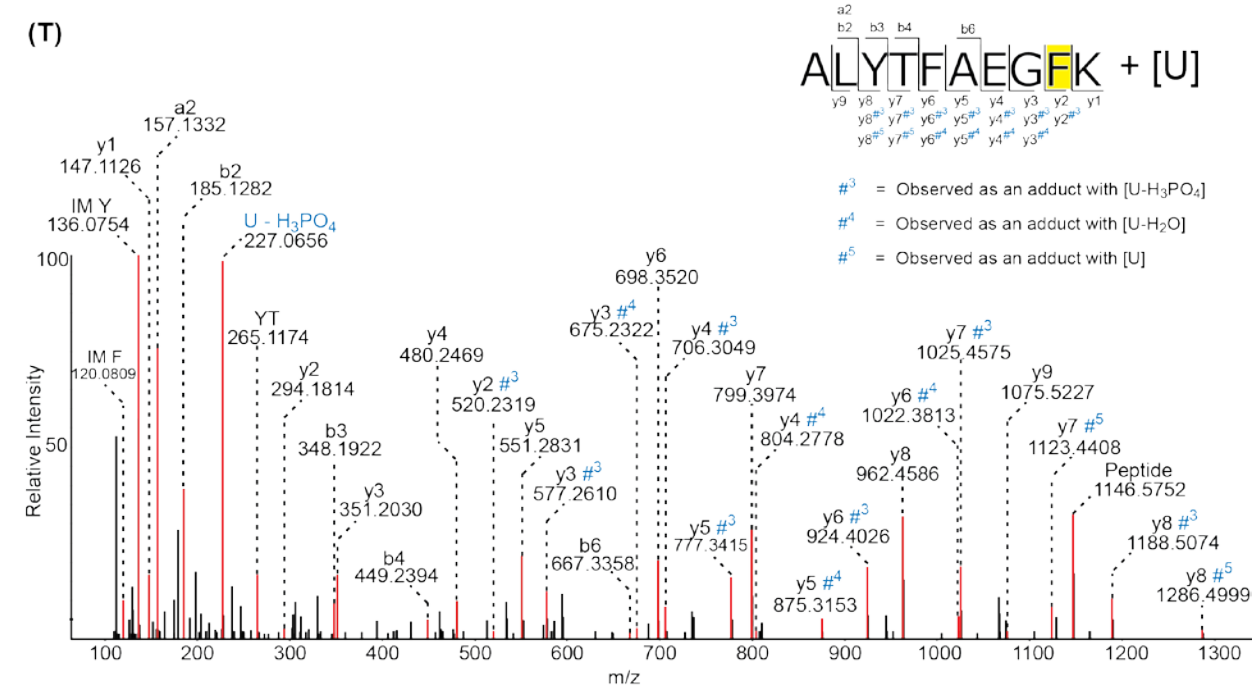

(T)
ALYTFAEGFK + [U]
#3 = Observed as an adduct with [U-H3PO4]
#4 = Observed as an adduct with [U-H2O]
#5 = Observed as an adduct with [U]
a2 157.1332
y1 147.1126
b2 185.1282
IM Y 136.0754
U - H3PO4 227.0656
y6 698.3520
y3 #4 675.2322
y4 #3 706.3049
YT 265.1174
y4 480.2469
y7 799.3974
y7 #3 1025.4575
IM F 120.0809
y2 294.1814
y2 #3 520.2319
y6 #4 1022.3813
y9 1075.5227
y4 #4 804.2778
b3 348.1922
y5 551.2831
y8 962.4586
y7 #5 1123.4408
y3 #3 577.2610
y3 351.2030
y5 #3 777.3415
Peptide 1146.5752
y6 #3 924.4026
b6 667.3358
y8 #3 1188.5074
b4 449.2394
y5 #4 875.3153
y8 #5 1286.4999
Relative Intensity
m/z

Figure 6.3 Protein-RNA cross-link spectra identified in Type III-A *T. thermophilus* Csm complex.

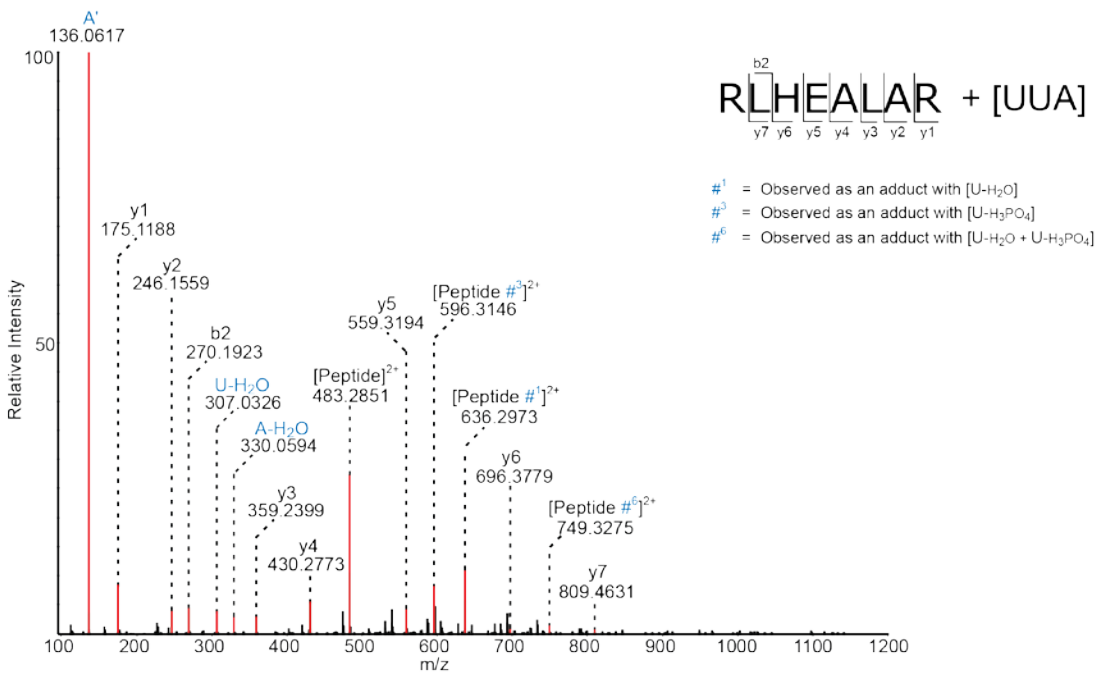
An overview of the cross-links identified and the corresponding mass values is provided in the table below. In each spectrum, the cross-linked peptide sequence and its corresponding y- and b- type fragment ions are indicated at the top. These refer to ions which retain the charge on the N-terminus or C-terminus, respectively. All the fragment ion peaks are marked with their corresponding m/z values. Ions with a mass shift corresponding to the cross-linked nucleotides are indicated with #: C_3O; #[1]: U'-H_2O; #[2]: U'; #[3]: U-H_3PO_4; #[4]: U-H_2O; #[5]: U and #[6]: U-H_2O + U-H_3PO_4. Mass shifts in the sequence tags help identify the site of cross-linking and are indicated for the corresponding fragments. The cross-linked amino acid highlighted in yellow. Adducts or maker ions corresponding to RNA component of the cross-link are indicated in blue. U': Base of U, 112.02 Da; A': Base of A, 136.06 Da; C': Base of C, 112.05 Da, G': Base of G, 152.05 Da, IM: Immonium ion.

Protein (Uniprot ID)	Peptide	aa	RNA	m(Peptide)	m(RNA)	m(XL) Calc	*z*	*m/z*	m(XL) Exp	Fig
Csm1 (Q53W19)	^{371}RLHEALAR^{378}	-	UUA	964.5566	959.1137	1923.6703	3	642.2306	1923.6684	(A)
Csm2 (Q53WF6)	^{35}LKSSQFR^{41}	K^{36}	U-H_2O	864.4817	306.0253	1170.5070	2	586.2637	1170.5118	(B)
Csm3 (Q53WF5)	^{21}IGMSRDQMAIGDLDNPVVR^{39}	-	UU	2086.0299	630.0612	2716.0911	3	906.3689	2716.09112	(C)
	^{40}NPLTDEPYIPGSSLK^{54}	^{49}P-K^{54}	UG	1629.8249	669.0833	2298.9082	3	767.3115	2298.9189	(D)
	^{91}IFGLAPENDER^{101}	P^{96}	U	1259.6145	324.0359	1583.6504	2	792.8327	1583.6498	(E)
	^{136}GGLYTEIKQEVFIPR^{150}	Q^{144}	UC	1748.9460	629.0772	2378.0232	3	793.6842	2378.0292	(F)
	^{151}LGGNANPR^{158}	G^{153}	UA	797.4143	653.0884	1450.5027	2	726.2637	1450.5118	(G)
	^{159}TTERVPAGAR^{168}	R^{162}	U	1056.5675	324.0359	1380.6034	3	461.2064	1380.5958	(H)
Csm4 (Q53WF4)	^{69}LPPVQVEETTLRK^{81}	P^{71}	UG	1508.8562	669.0833	2177.9395	3	726.9895	2177.9253	(I)
	^{126} TRVGVDR^{132}	V^{128}	UU	801.4456	630.0612	1431.5068	2	716.7577	1431.4998	(J)
Csm5 (Q53W18)	^{132}SPLGAYLPGSSVK^{144}	P^{139}	U	1274.6870	324.0359	1598.7229	2	800.3674	1598.7192	(K)
	^{255}MVLLAETFR^{263}	M^{255}	U	1078.5844	324.0359	1402.6203	2	702.3160	1402.6164	(L)

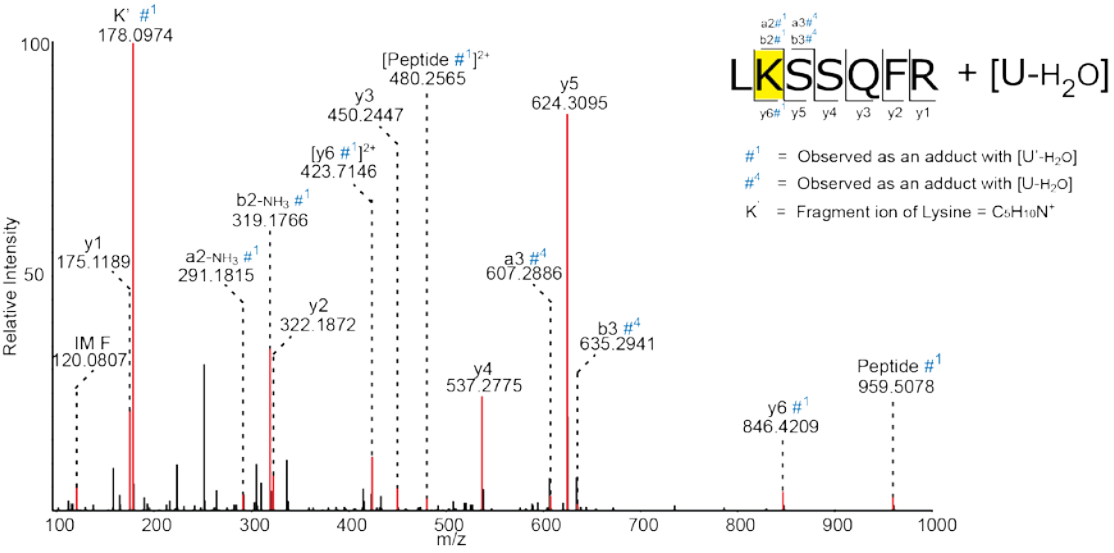
Protein: Name of the protein (Uniprot ID), Peptide: amino acid sequence of the cross-linked peptide, aa: position of cross-linked amino acid residue, RNA: composition of cross-linked RNA, m: mass, m(Peptide): calculated mass of cross-linked peptide, m(RNA): calculated mass of cross-linked RNA, m(XL) Calc: calculated mass of cross-link [m(Peptide)+m(RNA)], z: charge state in which cross-link was observed, *m/z*: experimentally observed mass to charge ratio, m(XL) Exp: experimentally observed mass of cross-link [(*m/z*) $*$ *z*) - ((mass of proton) $*$ *z*] and Fig: reference to figure of the annotated MS/MS fragment spectrum.

Figure originally published in [59] and reproduced with permission.

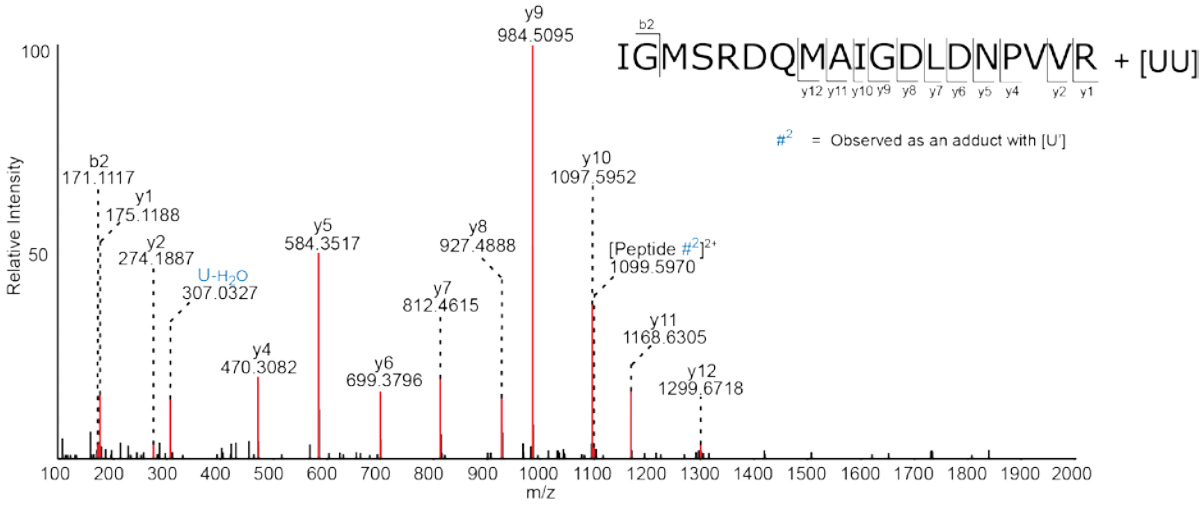
(A)

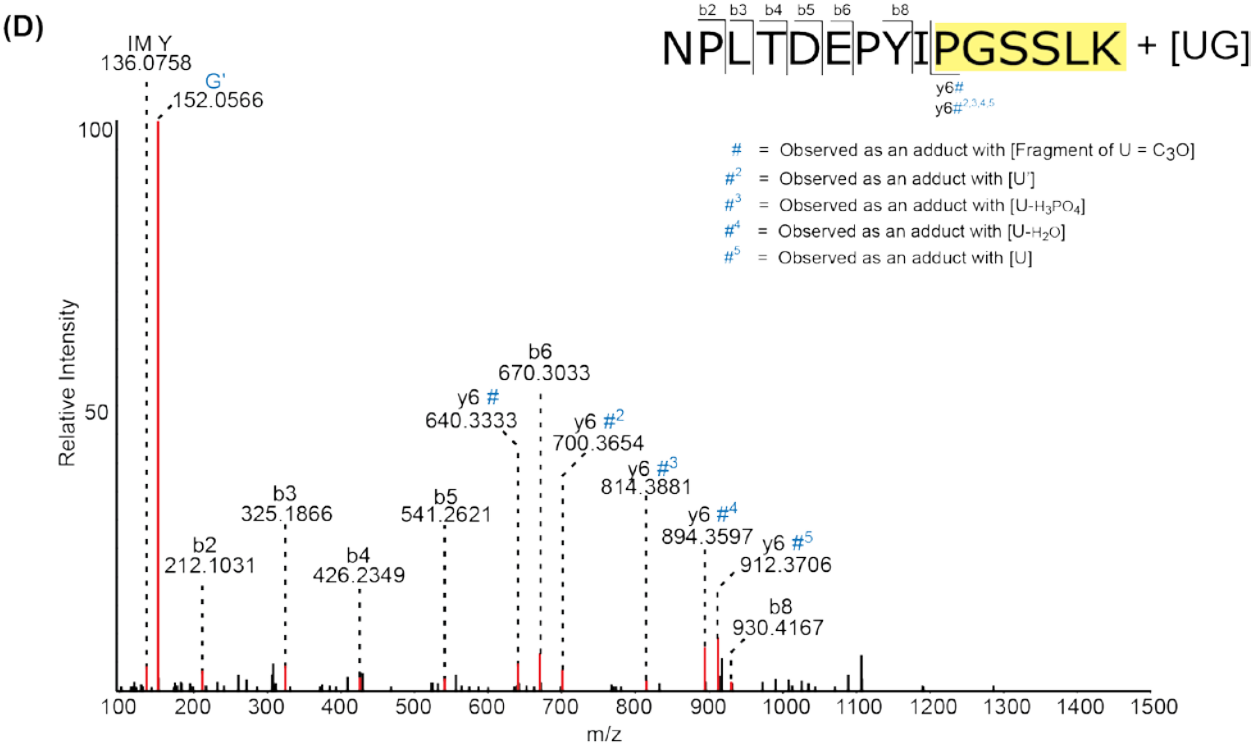

A'
136.0617
y1
175.1188
y2
246.1559
b2
270.1923
U-H2O
307.0326
A-H2O
330.0594
y3
359.2399
y4
430.2773
[Peptide]2+
483.2851
y5
559.3194
[Peptide #3]2+
596.3146
[Peptide #1]2+
636.2973
y6
696.3779
[Peptide #6]2+
749.3275
y7
809.4631
Relative Intensity
m/z
b2
RLHEALAR + [UUA]
y7 y6 y5 y4 y3 y2 y1
#1 = Observed as an adduct with [U-H2O]
#3 = Observed as an adduct with [U-H3PO4]
#6 = Observed as an adduct with [U-H2O + U-H3PO4]

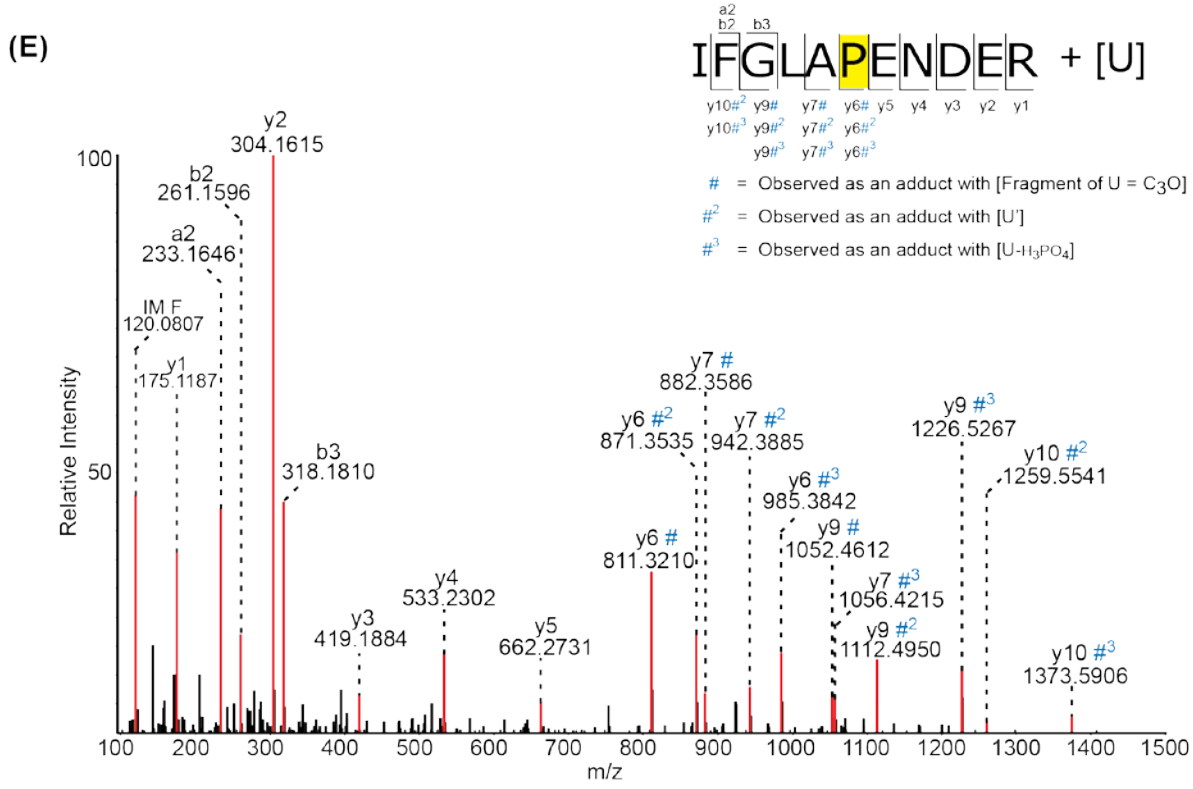
(B)


K' #1
178.0974
[Peptide #1]2+
480.2565
y3
450.2447
y5
624.3095
[y6 #1]2+
423.7146
b2-NH3 #1
319.1766
y1
175.1189
a2-NH3 #1
291.1815
a3 #4
607.2886
y2
322.1872
b3 #4
635.2941
IM F
120.0807
y4
537.2775
Peptide #1
959.5078
y6 #1
846.4209
Relative Intensity
m/z
a2#1 a3#4
b2#1 b3#4
LKSSQFR + [U-H2O]
y6#1 y5 y4 y3 y2 y1
#1 = Observed as an adduct with [U'-H2O]
#4 = Observed as an adduct with [U-H2O]
K = Fragment ion of Lysine = C6H10N+

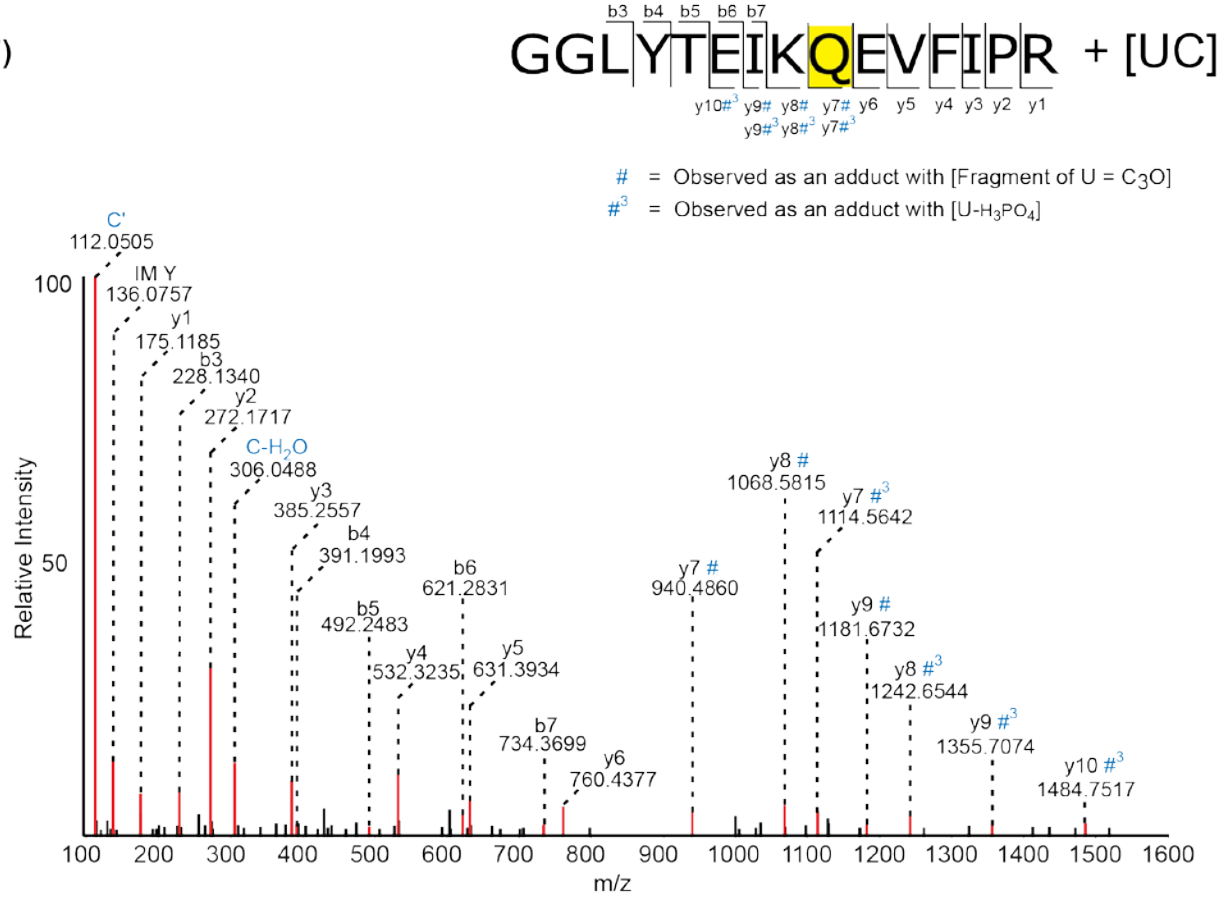
(C)


y9
984.5095
b2
171.1117
y1
175.1188
y2
274.1887
U-H2O
307.0327
y5
584.3517
y4
470.3082
y6
699.3796
y8
927.4888
y7
812.4615
y10
1097.5952
[Peptide #2]2+
1099.5970
y11
1168.6305
y12
1299.6718
Relative Intensity
m/z
b2
IGMSRDQMAIGDLDNPVVR + [UU]
y12 y11 y10 y9 y8 y7 y6 y5 y4 y2 y1
#2 = Observed as an adduct with [U']

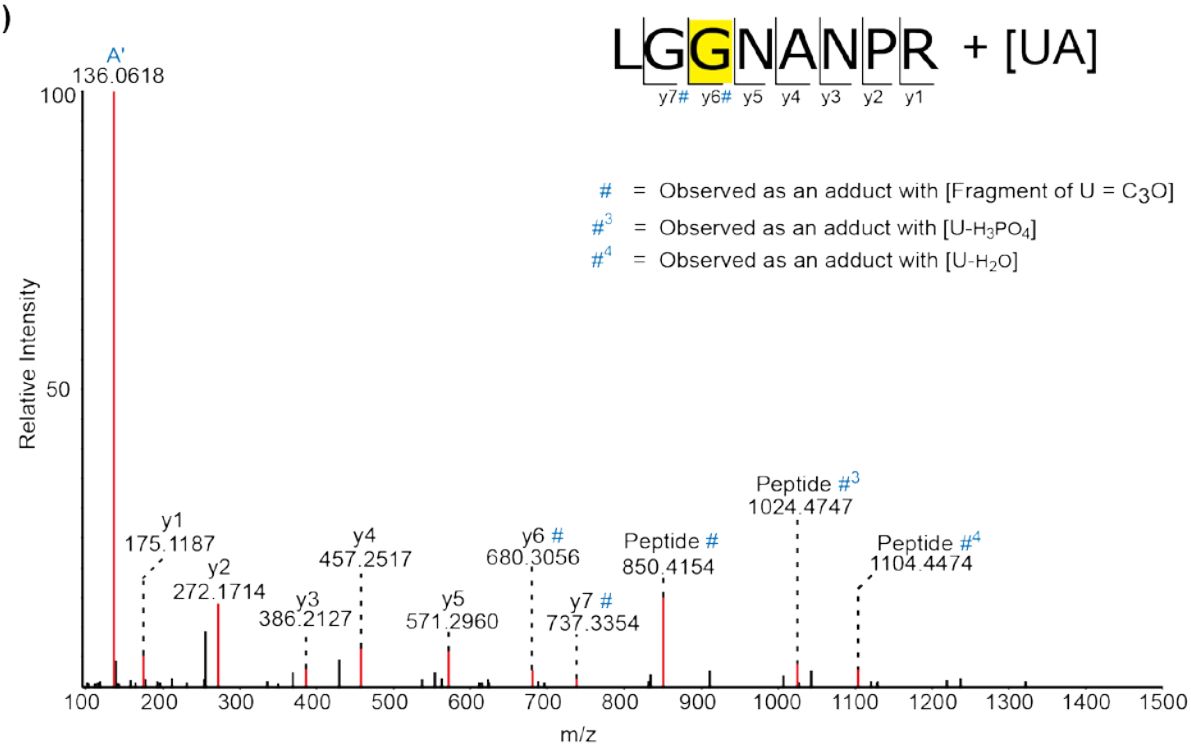

(D)
b2 b3 b4 b5 b6 b8
NPLTDEPYIPGSSLK + [UG]
y6#
y6#2,3,4,5
= Observed as an adduct with [Fragment of U = C3O]
#2 = Observed as an adduct with [U']
#3 = Observed as an adduct with [U-H3PO4]
#4 = Observed as an adduct with [U-H2O]
#5 = Observed as an adduct with [U]
IM Y
136.0758
G'
152.0566
b6
670.3033
y6 #
640.3333
y6 #2
700.3654
y6 #3
814.3881
b3
325.1866
b5
541.2621
y6 #4
894.3597
y6 #5
912.3706
b2
212.1031
b4
426.2349
b8
930.4167
Relative Intensity
m/z


(E)
a2
b2 b3
IFGLAPENDER + [U]
y10#2 y9# y7# y6# y5 y4 y3 y2 y1
y10#3 y9#2 y7#2 y6#2
y9#3 y7#3 y6#3
= Observed as an adduct with [Fragment of U = C3O]
#2 = Observed as an adduct with [U']
#3 = Observed as an adduct with [U-H3PO4]
y2
304.1615
b2
261.1596
a2
233.1646
IM F
120.0807
y1
175.1187
b3
318.1810
y7 #
882.3586
y6 #2
871.3535
y7 #2
942.3885
y9 #3
1226.5267
y10 #2
1259.5541
y6 #3
985.3842
y9 #
1052.4612
y6 #
811.3210
y7 #3
1056.4215
y4
533.2302
y3
419.1884
y5
662.2731
y9 #2
1112.4950
y10 #3
1373.5906
Relative Intensity
m/z

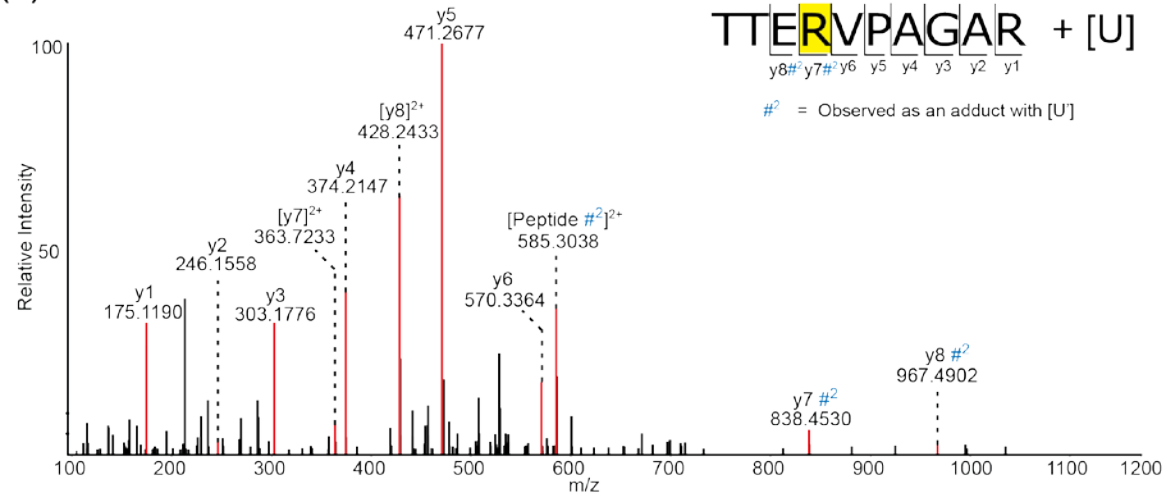
(F)


b3 b4 b5 b6 b7
GGLYTEIKQEVFIPR + [UC]
y10#3 y9# y8# y7# y6 y5 y4 y3 y2 y1
y9#3 y8#3 y7#3
= Observed as an adduct with [Fragment of U = C3O]
#3 = Observed as an adduct with [U-H3PO4]
C'
112.0505
IM Y
136.0757
y1
175.1185
b3
228.1340
y2
272.1717
C-H2O
306.0488
y3
385.2557
b4
391.1993
b6
621.2831
b5
492.2483
y4
532.3235
y5
631.3934
b7
734.3699
y6
760.4377
y7 #
940.4860
y8 #
1068.5815
y7 #3
1114.5642
y9 #
1181.6732
y8 #3
1242.6544
y9 #3
1355.7074
y10 #3
1484.7517
Relative Intensity
m/z

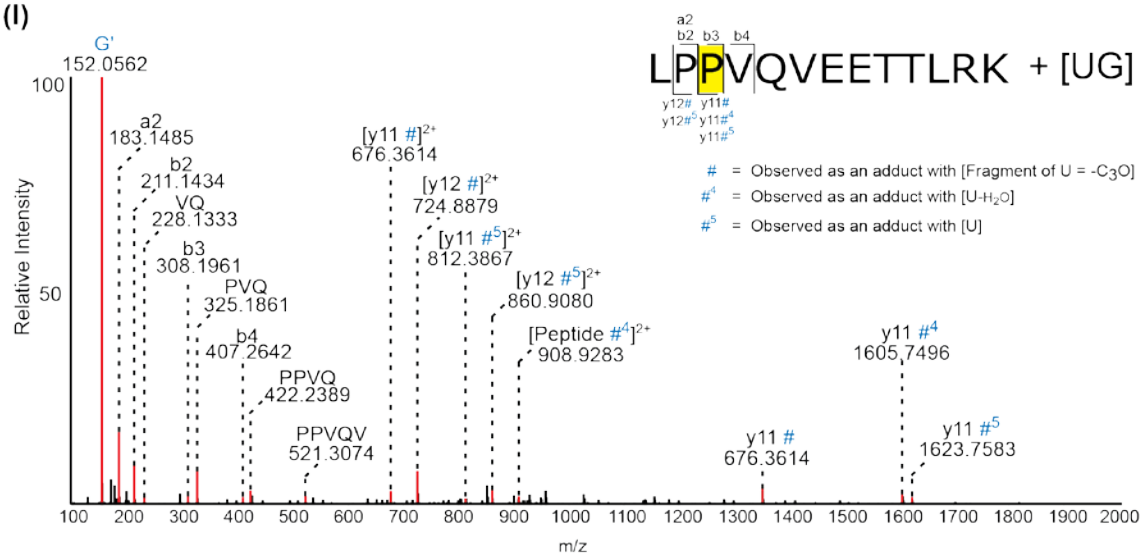
(G)


LGGNANPR + [UA]
y7# y6# y5 y4 y3 y2 y1
= Observed as an adduct with [Fragment of U = C3O]
#3 = Observed as an adduct with [U-H3PO4]
#4 = Observed as an adduct with [U-H2O]
A'
136.0618
y1
175.1187
y2
272.1714
y3
386.2127
y4
457.2517
y5
571.2960
y6 #
680.3056
y7 #
737.3354
Peptide #
850.4154
Peptide #3
1024.4747
Peptide #4
1104.4474
Relative Intensity
m/z

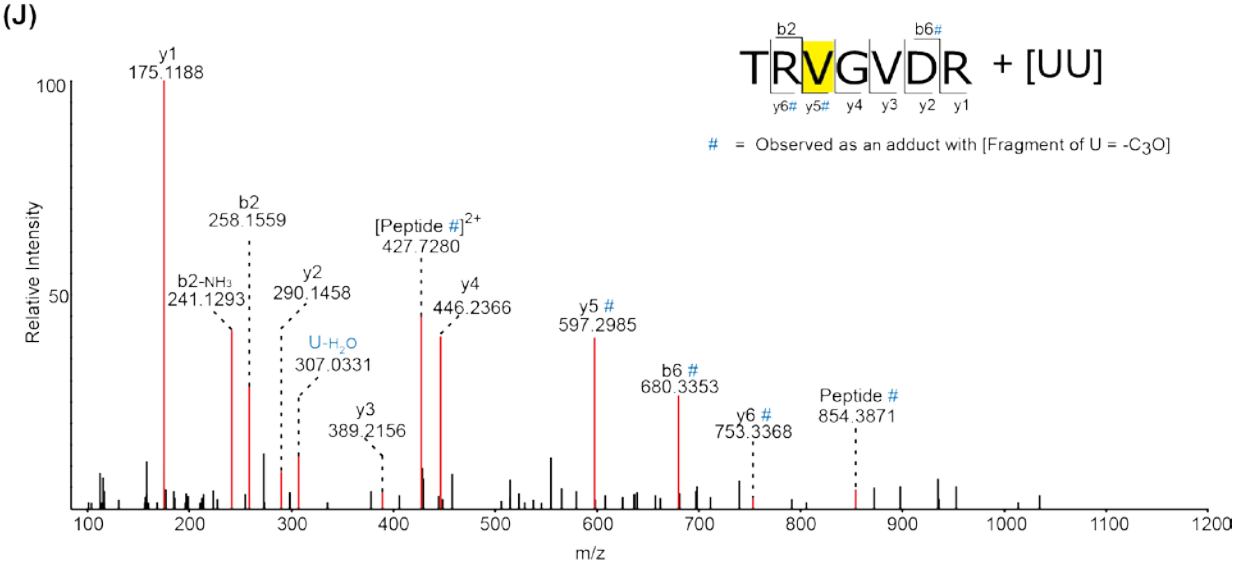
(H)


TTERVPAGAR + [U]
y8#² y7#² y6 y5 y4 y3 y2 y1
#² = Observed as an adduct with [U]
Relative Intensity
m/z
y5 471.2677
[y8]²⁺ 428.2433
y4 374.2147
[y7]²⁺ 363.7233
y2 246.1558
[Peptide #²]²⁺ 585.3038
y6 570.3364
y1 175.1190
y3 303.1776
y8 #² 967.4902
y7 #² 838.4530

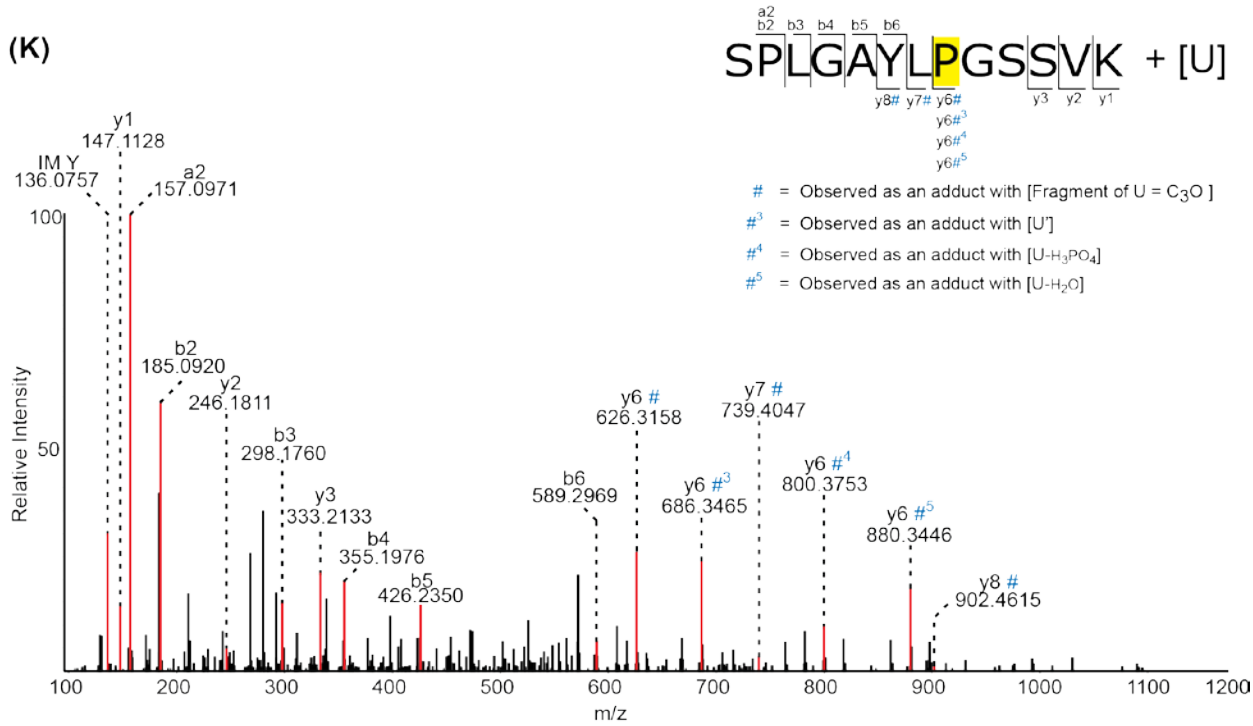
(I)


LPPVQVEETTLRK + [UG]
a2 b2 b3 b4
y12# y11#
y12#⁵ y11#⁴
y11#⁵
= Observed as an adduct with [Fragment of U = -C3O]
#⁴ = Observed as an adduct with [U-H2O]
#⁵ = Observed as an adduct with [U]
Relative Intensity
m/z
G' 152.0562
a2 183.1485
b2 211.1434
VQ 228.1333
b3 308.1961
PVQ 325.1861
b4 407.2642
PPVQ 422.2389
PPVQV 521.3074
[y11 #]²⁺ 676.3614
[y12 #]²⁺ 724.8879
[y11 #⁵]²⁺ 812.3867
[y12 #⁵]²⁺ 860.9080
[Peptide #⁴]²⁺ 908.9283
y11 # 676.3614
y11 #⁴ 1605.7496
y11 #⁵ 1623.7583

(J)


TRVGVDR + [UU]
b2 b6#
y6# y5# y4 y3 y2 y1
= Observed as an adduct with [Fragment of U = -C3O]
Relative Intensity
m/z
y1 175.1188
b2 258.1559
b2-NH3 241.1293
y2 290.1458
U-H2O 307.0331
y3 389.2156
[Peptide #]²⁺ 427.7280
y4 446.2366
y5 # 597.2985
b6 # 680.3353
y6 # 753.3368
Peptide # 854.3871

(K)


a2
b2 b3 b4 b5 b6
SPLGAYLPGSSVK + [U]
y8# y7# y6#
y6#3
y6#4
y6#5
y3 y2 y1
= Observed as an adduct with [Fragment of U = C3O]
#3 = Observed as an adduct with [U']
#4 = Observed as an adduct with [U-H3PO4]
#5 = Observed as an adduct with [U-H2O]
Relative Intensity
m/z
y1 147.1128
IM Y 136.0757
a2 157.0971
b2 185.0920
y2 246.1811
b3 298.1760
y3 333.2133
b4 355.1976
b5 426.2350
b6 589.2969
y6 # 626.3158
y6 #3 686.3465
y7 # 739.4047
y6 #4 800.3753
y6 #5 880.3446
y8 # 902.4615

(L)

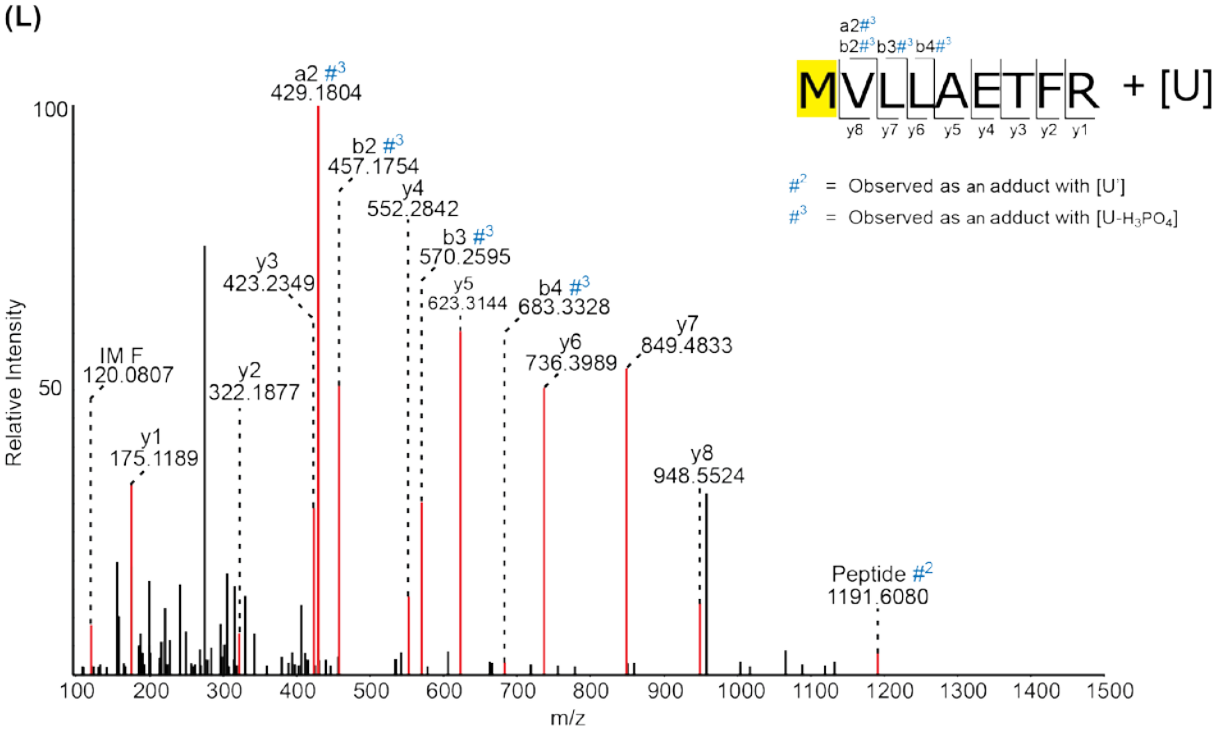
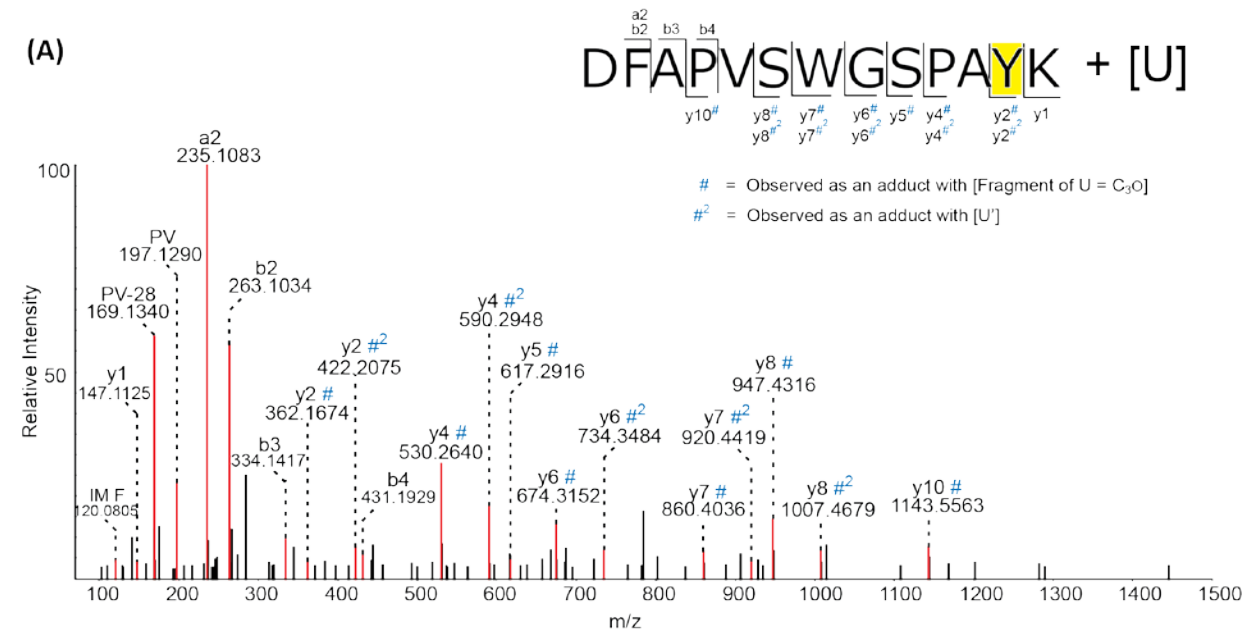

a2#3
b2#3 b3#3 b4#3
MVLLAETFR + [U]
y8 y7 y6 y5 y4 y3 y2 y1
#2 = Observed as an adduct with [U']
#3 = Observed as an adduct with [U-H3PO4]
Relative Intensity
m/z
a2 #3 429.1804
b2 #3 457.1754
y4 552.2842
b3 #3 570.2595
y3 423.2349
y5 623.3144
b4 #3 683.3328
y6 736.3989
y7 849.4833
IM F 120.0807
y2 322.1877
y1 175.1189
y8 948.5524
Peptide #2 1191.6080

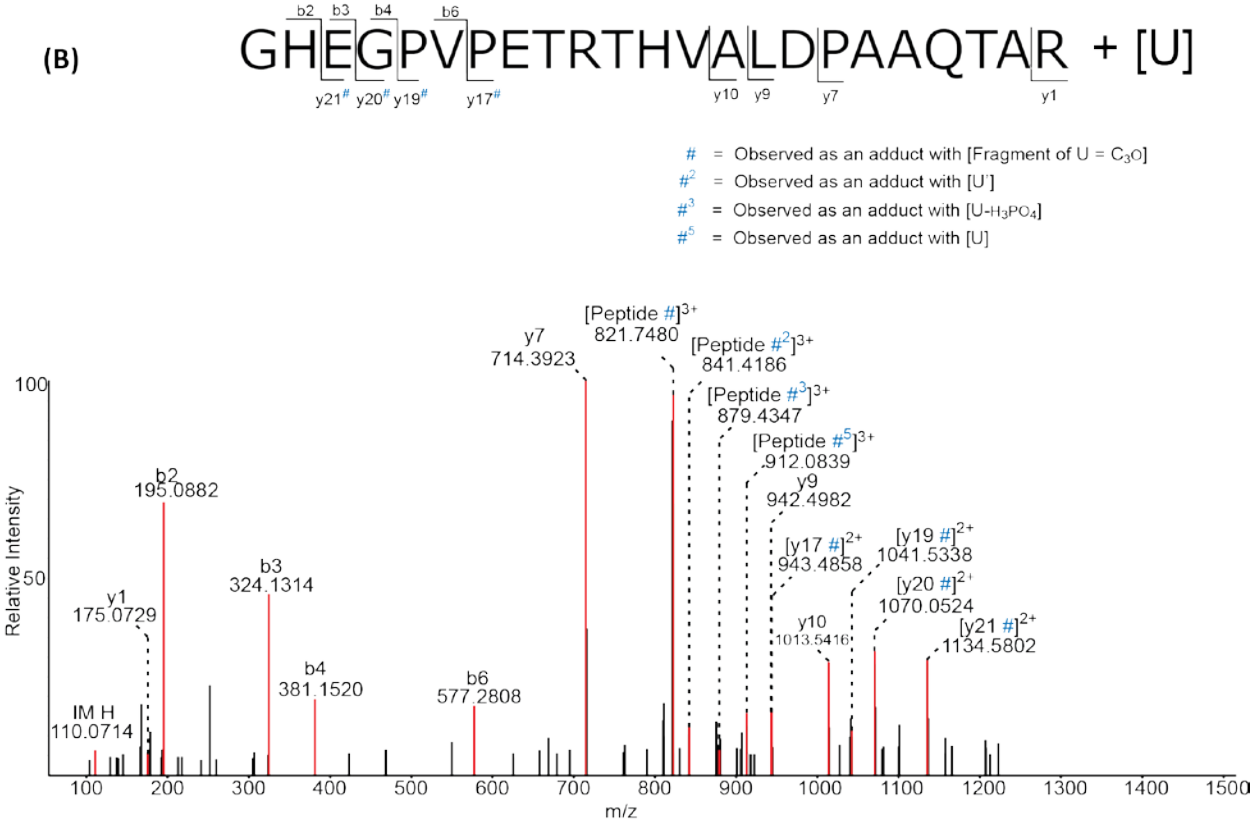
Figure 6.4 Protein-RNA cross-link spectra identified in Type III-B *T. thermophilus* Cmr complex.

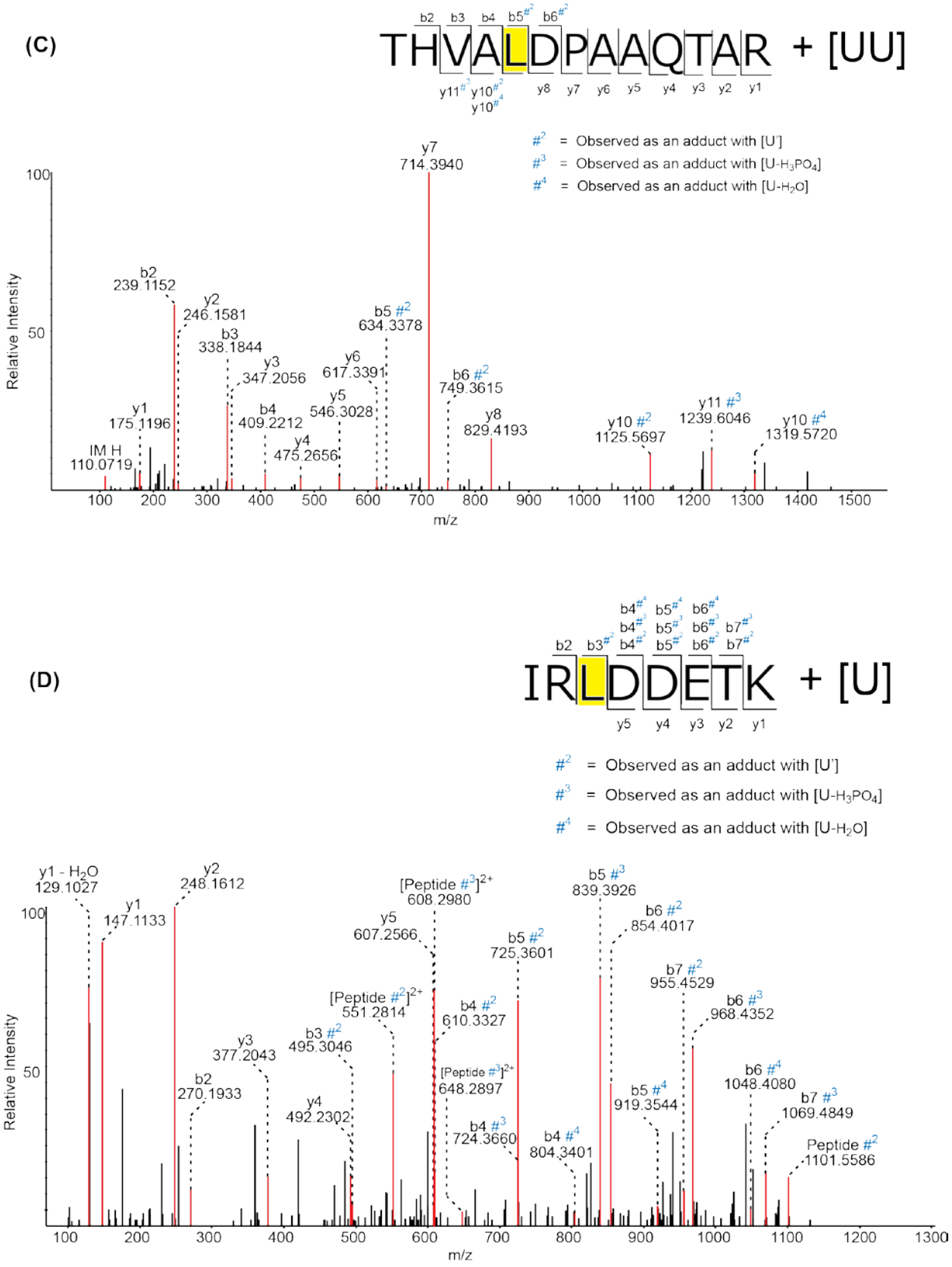
An overview of the cross-links identified and the corresponding mass values is provided in the table below. In each spectrum, the cross-linked peptide sequence and its corresponding y- and b- type fragment ions are indicated at the top. These refer to ions which retain the charge on the N-terminus or C-terminus, respectively. All the fragment ion peaks are marked with their corresponding m/z values. Ions with a mass shift corresponding to the cross-linked nucleotides are indicated with #: C_3O; #[1]: U'-H_2O; #[2]: U'; #[3]: U-H_3PO_4; #[4]: U-H_2O and #[5]: U. Mass shifts in the sequence tags help identify the site of cross-linking and are indicated for the corresponding fragments. The cross-linked amino acid highlighted in yellow. Adducts or maker ions corresponding to RNA component of the cross-link are indicated in blue. U': Base of U, 112.02 Da; C': Base of C, 112.05 Da, IM: Immonium ion.

Endogenous complex										
Protein (Uniprot ID)	**Peptide**	**aa**	**RNA**	**m(Peptide)**	**m(RNA)**	**m(XL) Calc**	***z***	***m/z***	**m(XL) Exp**	**Fig**
Cmr2 (Q53W09)	^{160}DFAPVSWGSPAYK^{172}	Y^{171}	U	1423.6771	324.0359	1747.7130	2	874.5654	1747.1152	(A)
Cmr3 (Q53W08)	^{156}GHEGPVPETRTHVALDPAAQTAR^{178}	-	U	2409.2148	324.0359	2733.2507	3	912.0919	2733.2523	(B)
	^{166}THVALDPAAQTAR^{178}	L^{170}	UU	1349.7051	630.0612	1979.7663	2	990.8915	1979.7674	(C)
Cmr4 (Q53W06)	^{204}IRLDDETK^{211}	L^{206}	U	988.5189	324.0359	1312.5548	2	657.2855	1312.5554	(D)
Cmr6 (Q53W04)	^{169}LHPDILNPHHPDYYGSVK^{186}	-	UU	2101.0380	630.0612	2731.0992	3	911.3740	2731.0986	(E)
Reconstituted complex										
Protein (Uniprot ID)	**Peptide**	**aa**	**RNA**	**m(Peptide)**	**m(RNA)**	**m(XL) Calc**	***z***	***m/z***	**m(XL) Exp**	**Fig**
Cmr1 (Q53W07)	^{34}TYTLLTPLFGGGVEPREADPVSVVR^{58}	-	UC	2672.4172	629.0772	3301.4944	3	1101.5060	3301.4946	(F)

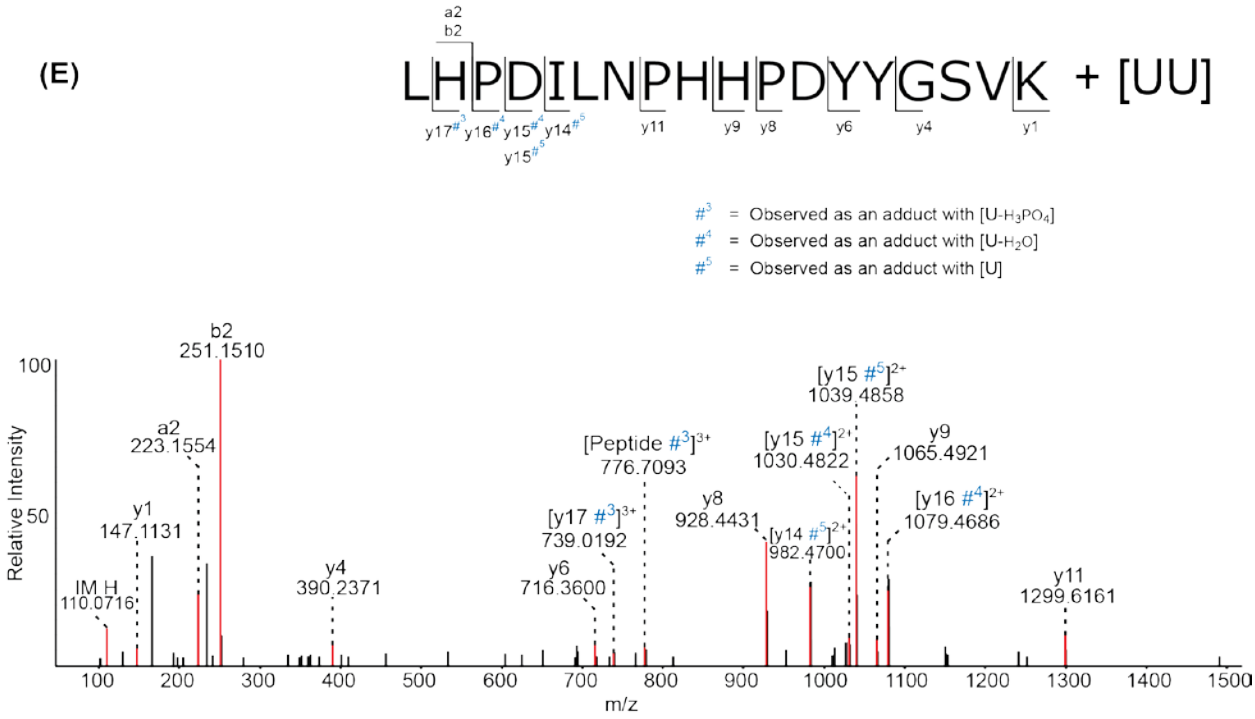
Protein: Name of the protein (Uniprot ID), Peptide: amino acid sequence of the cross-linked peptide, aa: position of cross-linked amino acid residue, RNA: composition of cross-linked RNA, m: mass, m(Peptide): calculated mass of cross-linked peptide, m(RNA): calculated mass of cross-linked RNA, m(XL) Calc: calculated mass of cross-link [m(Peptide)+m(RNA)], z: charge state in which cross-link was observed, *m/z*: experimentally observed mass to charge ratio, m(XL) Exp: experimentally observed mass of cross-link [(*m/z*) $*$ *z*) - ((mass of proton) $*$ *z*] and Fig: reference to figure of the annotated MS/MS fragment spectrum.


(A)
DFAPVSWGSPAYK + [U]
= Observed as an adduct with [Fragment of U = C3O]
#2 = Observed as an adduct with [U']
a2 235.1083
PV 197.1290
PV-28 169.1340
b2 263.1034
y1 147.1125
y2 #2 422.2075
y2 # 362.1674
b3 334.1417
y4 #2 590.2948
y5 # 617.2916
y4 # 530.2640
b4 431.1929
y6 #2 734.3484
y6 # 674.3152
y7 #2 920.4419
y8 # 947.4316
y7 # 860.4036
y8 #2 1007.4679
y10 # 1143.5563
IM F 120.0805
Relative Intensity
m/z

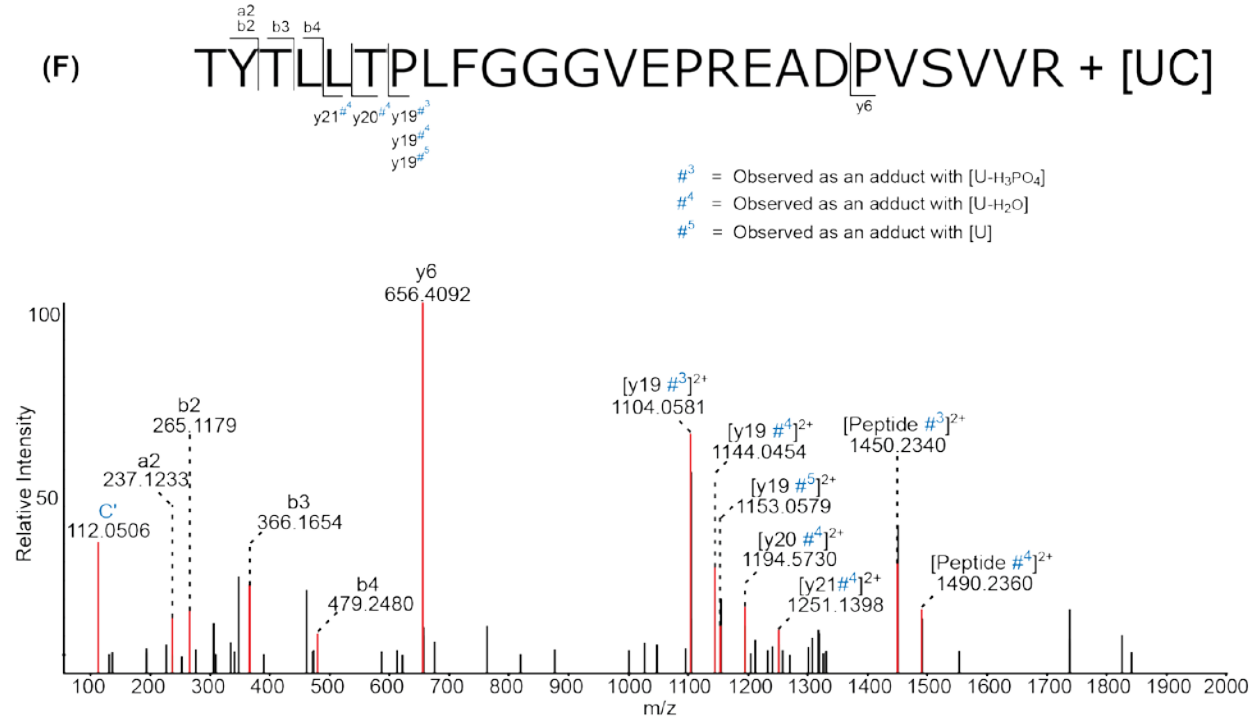

(B)
GHEGPVPETRTHVALDPAAQTAR + [U]
= Observed as an adduct with [Fragment of U = C3O]
#2 = Observed as an adduct with [U']
#3 = Observed as an adduct with [U-H3PO4]
#5 = Observed as an adduct with [U]
[Peptide #]3+ 821.7480
y7 714.3923
[Peptide #2]3+ 841.4186
[Peptide #3]3+ 879.4347
[Peptide #5]3+ 912.0839
y9 942.4982
b2 195.0882
[y17 #]2+ 943.4858
[y19 #]2+ 1041.5338
[y20 #]2+ 1070.0524
[y21 #]2+ 1134.5802
b3 324.1314
y1 175.0729
b4 381.1520
b6 577.2808
y10 1013.5416
IM H 110.0714
Relative Intensity
m/z


(C)
THVALDPAAQTAR + [UU]
#2 = Observed as an adduct with [U']
#3 = Observed as an adduct with [U-H3PO4]
#4 = Observed as an adduct with [U-H2O]
Relative Intensity
m/z
y7 714.3940
b2 239.1152
y2 246.1581
b3 338.1844
y3 347.2056
b5 #2 634.3378
y6 617.3391
y5 546.3028
b6 #2 749.3615
y1 175.1196
b4 409.2212
y4 475.2656
y8 829.4193
IM H 110.0719
y10 #2 1125.5697
y11 #3 1239.6046
y10 #4 1319.5720
(D)
IRLDDETK + [U]
#2 = Observed as an adduct with [U']
#3 = Observed as an adduct with [U-H3PO4]
#4 = Observed as an adduct with [U-H2O]
Relative Intensity
m/z
y1 - H2O 129.1027
y2 248.1612
y1 147.1133
[Peptide #3]2+ 608.2980
y5 607.2566
b5 #3 839.3926
b6 #2 854.4017
b5 #2 725.3601
b7 #2 955.4529
[Peptide #2]2+ 551.2814
b4 #2 610.3327
b6 #3 968.4352
y3 377.2043
b3 #2 495.3046
[Peptide #3]2+ 648.2897
b6 #4 1048.4080
b2 270.1933
y4 492.2302
b5 #4 919.3544
b7 #3 1069.4849
b4 #3 724.3660
b4 #4 804.3401
Peptide #2 1101.5586

(E)


a2
b2
LHPDILNPHHPDYYGSVK + [UU]
y17#3 y16#4 y15#4 y14#5
y15#5
y11
y9
y8
y6
y4
y1
#3 = Observed as an adduct with [U-H3PO4]
#4 = Observed as an adduct with [U-H2O]
#5 = Observed as an adduct with [U]
b2 251.1510
a2 223.1554
y1 147.1131
IM H 110.0716
y4 390.2371
y6 716.3600
[Peptide #3]3+ 776.7093
[y17 #3]3+ 739.0192
y8 928.4431
[y14 #5]2+ 982.4700
[y15 #5]2+ 1039.4858
[y15 #4]2+ 1030.4822
y9 1065.4921
[y16 #4]2+ 1079.4686
y11 1299.6161
Relative Intensity
m/z

(F)


a2
b2
b3
b4
TYTLLTPLFGGGVEPREADPVSVVR + [UC]
y21#4 y20#4 y19#3
y19#4
y19#5
y6
#3 = Observed as an adduct with [U-H3PO4]
#4 = Observed as an adduct with [U-H2O]
#5 = Observed as an adduct with [U]
y6 656.4092
[y19 #3]2+ 1104.0581
b2 265.1179
[y19 #4]2+ 1144.0454
[Peptide #3]2+ 1450.2340
a2 237.1233
[y19 #5]2+ 1153.0579
C' 112.0506
b3 366.1654
[y20 #4]2+ 1194.5730
b4 479.2480
[y21#4]2+ 1251.1398
[Peptide #4]2+ 1490.2360
Relative Intensity
m/z

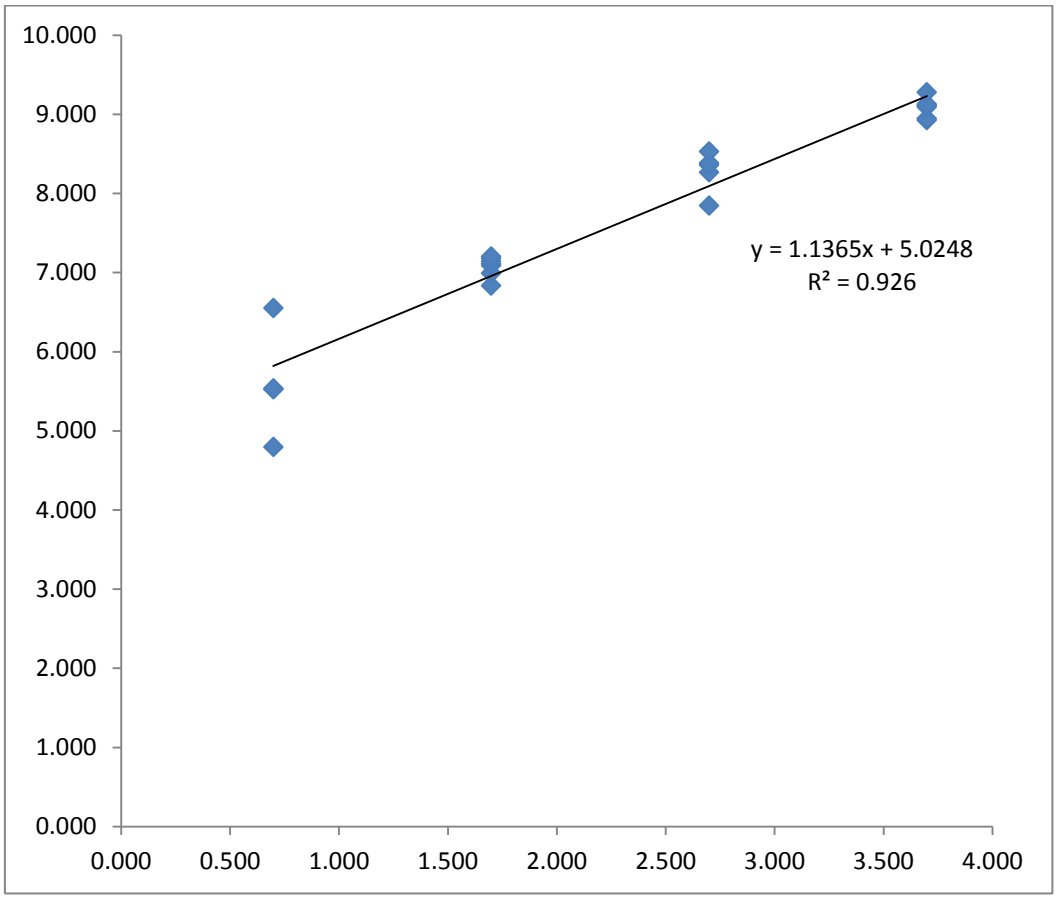


Figure 6.5 iBAQ calibration curve of UPS2 proteins used in determining the stoichiometry of Cas5, Cas6, Cas7 and Cas8b in *C. thermocellum* Cascade complex.

Double logarithmic plot of log(iBAQ) *vs* log(amount) using 22 out of 48 UPS2 proteins observed in all three technical replicates for the absolute quantification. The log (iBAQ) and log (amount) for all the 22 proteins is presented in Table 6.1.

Table 6.1 iBAQ quantitative mass spectrometry analysis of *C. thermocellum* Cascade complex.

Protein	Source	Amount (fmol)	iBAQ	log (amount)	log (iBAQ)	Ratio
>P01127ups| PDGFB_HUMAN	UPS2 Standard	5	62227	0.699	4.794	
>P01008ups| ANT3_HUMAN	UPS2 Standard	5	333550	0.699	5.523	
>P08263ups| GSTA1_HUMAN	UPS2 Standard	5	341790	0.699	5.534	
>P01344ups| IGF2_HUMAN	UPS2 Standard	5	3559200	0.699	6.551	
>P02753ups| RETBP_HUMAN	UPS2 Standard	50	6824400	1.699	6.834	
>Q15843ups| NEDD8_HUMAN	UPS2 Standard	50	9824600	1.699	6.992	
>P06732ups| KCRM_HUMAN	UPS2 Standard	50	1.2E+07	1.699	7.088	
>P16083ups| NQO2_HUMAN	UPS2 Standard	50	1.3E+07	1.699	7.106	
>P63279ups| UBC9_HUMAN	UPS2 Standard	50	1.4E+07	1.699	7.142	
>P61626ups| LYSC_HUMAN	UPS2 Standard	50	1.5E+07	1.699	7.177	
>P00709ups| LALBA_HUMAN	UPS2 Standard	50	1.6E+07	1.699	7.201	
>P01133ups| EGF_HUMAN	UPS2 Standard	500	7E+07	2.699	7.845	
>P02144ups| MYG_HUMAN	UPS2 Standard	500	1.8E+08	2.699	8.267	
>P15559ups| NQO1_HUMAN	UPS2 Standard	500	2.3E+08	2.699	8.360	
>P04040ups| CATA_HUMAN	UPS2 Standard	500	2.4E+08	2.699	8.383	
>Q06830ups| PRDX1_HUMAN	UPS2 Standard	500	3.4E+08	2.699	8.529	
>P00918ups| CAH2_HUMAN_UPS	UPS2 Standard	5000	8.4E+08	3.699	8.925	
>P68871ups| HBB_HUMAN	UPS2 Standard	5000	8.8E+08	3.699	8.945	
>P41159ups| LEP_HUMAN	UPS2 Standard	5000	1.2E+09	3.699	9.089	
>P01031ups| CO5_HUMAN	UPS2 Standard	5000	1.3E+09	3.699	9.112	
>P69905ups| HBA_HUMAN	UPS2 Standard	5000	1.3E+09	3.699	9.126	
>P00915ups| CAH1_HUMAN	UPS2 Standard	5000	1.9E+09	3.699	9.277	
C. thermocellum **Cas6**	Cascade	788.860	2.1E+08	2.897	8.317	**1.0**
C. thermocellum **Cas5**	Cascade	1009.253	2.7E+08	3.004	8.439	**1.28**
C. thermocellum **Cas8b**	Cascade	1896.706	5.6E+08	3.278	8.750	**2.40**
C. thermocellum **Cas7**	Cascade	4731.513	1.6E+09	3.675	9.201	**5.99**

Table 6.2 Inter-protein cross-links identified in *C. thermocellum* Cascade complex.

Protein1 and Protein2 are the two proteins in which the inter-protein cross-link was identified, Residue1: Cross-linked residue in protein1, Residue2: Cross-linked residue in protein2, Peptide1: Cross-linked peptide corresponding to protein1 and Peptide2: Cross-linked peptide corresponding to protein2 Cross-linked lysine residues in the peptides, are highlighted in blue.

Protein1	Protein2	Residue 1	Residue 2	Peptide1	Peptide2
Cas7	Cas8b	K76	K73	EIESE**K**GGIK	**K**GSAR
Cas7	Cas8b	K76	K190	EIESE**K**GGIKDGK	YG**K**TSK
Cas7	Cas8b	K76	K230	EIESE**K**GGIK	IGFVTGGF**K**QENAWK
Cas7	Cas8b	K76	K378	EIESE**K**GGIKDGK	NLPG**K**DNATYDLK
Cas7	Cas8b	K76	K391	EIESE**K**GGIKDGK	FGFD**K**IR
Cas7	Cas8b	K76	K400	EIESE**K**GGIK	TFFPNN**K**TEGNFDK
Cas7	Cas8b	K76	K421	EIESE**K**GGIK	**K**ISYK
Cas7	Cas8b	K76	K523	EIESE**K**GGIKDGK	LLNIQY**K**ER
Cas7	Cas8b	K80	K190	GGI**K**DGK	YG**K**TSK
Cas7	Cas8b	K83	K65	DG**K**AR	**K**INNYAYK
Cas7	Cas8b	K83	K72	DG**K**AR	INNYAY**K**K
Cas7	Cas8b	K83	K190	DG**K**AR	YG**K**TSK
Cas7	Cas8b	K83	K400	DG**K**AR	TFFPNN**K**TEGNFDK
Cas7	Cas8b	K87	K65	A**K**DFNENVDEILQK	**K**INNYAYK
Cas7	Cas8b	K87	K73	A**K**DFNENVDEILQK	**K**GSAR
Cas7	Cas8b	K87	K373	A**K**DFNENVDEILQK	VDETVLF**K**NLPGK
Cas7	Cas8b	K133	K421	SLN**K**VNLK	**K**ISYK
Cas7	Cas8b	K133	K425	SLN**K**VNLK	ISY**K**FLLGR
Cas7	Cas8b	K133	K470	SLN**K**VNLK	G**K**EVQK
Cas7	Cas8b	K133	K538	VNL**K**HIK	LNGL**K**LNK
Cas7	Cas8b	K140	K65	HI**K**GTGAFASGEGK	**K**INNYAYK
Cas7	Cas8b	K140	K73	HI**K**GTGAFASGEGK	**K**GSAR
Cas7	Cas8b	K140	K92	HI**K**GTGAFASGEGK	YTDSM**K**TLNK
Cas7	Cas8b	K140	K400	HI**K**GTGAFASGEGK	TFFPNN**K**TEGNFDK
Cas7	Cas8b	K140	K528	HI**K**GTGAFASGEGK	GS**K**PFYSR
Cas7	Cas8b	K140	K538	HI**K**GTGAFASGEGK	LNGL**K**LNK
Cas7	Cas8b	K140	K541	HI**K**GTGAFASGEGK	LN**K**NIVK
Cas7	Cas8b	K140	K554	HI**K**GTGAFASGEGK	IYTEAIN**K**LNEYNK
Cas7	Cas8b	K140	K560	HI**K**GTGAFASGEGK	LNEYN**K**NYYK
Cas7	Cas8b	K151	K33	GTGAFASGEG**K**AQK	STDTEEYLQVIENPND**K**GNYNHVLK
Cas7	Cas8b	K151	K64	GTGAFASGEG**K**AQK	GVEYEEFSS**K**K
Cas7	Cas8b	K151	K65	GTGAFASGEG**K**AQK	**K**INNYAYK
Cas7	Cas8b	K151	K73	GTGAFASGEG**K**AQK	**K**GSAR
Cas7	Cas8b	K151	K92	GTGAFASGEG**K**AQK	YTDSM**K**TLNK
Cas7	Cas8b	K151	K187	GTGAFASGEG**K**AQK	ILNEQYYN**K**YGK
Cas7	Cas8b	K151	K193	GTGAFASGEG**K**AQK	TS**K**GK
Cas7	Cas8b	K151	K236	GTGAFASGEG**K**AQK	QENAW**K**NYPVCSCCAQK
Cas7	Cas8b	K151	K253	GTGAFASGEG**K**AQK	**K**YIR
Cas7	Cas8b	K151	K357	GTGAFASGEG**K**AQK	V**K**NILR
Cas7	Cas8b	K151	K378	GTGAFASGEG**K**AQK	NLPG**K**DNATYDLK
Cas7	Cas8b	K151	K400	GTGAFASGEG**K**AQK	TFFPNN**K**TEGNFDK
Cas7	Cas8b	K151	K421	GTGAFASGEG**K**AQK	**K**ISYK
Cas7	Cas8b	K151	K482	GTGAFASGEG**K**AQK	TE**K**NKK
Cas7	Cas8b	K151	K484	GTGAFASGEG**K**AQK	TEKN**K**
Cas7	Cas8b	K151	K523	GTGAFASGEG**K**AQK	LLNIQY**K**ER
Cas7	Cas8b	K151	K528	GTGAFASGEG**K**AQK	GS**K**PFYSR
Cas7	Cas8b	K151	K538	GTGAFASGEG**K**AQK	LNGL**K**LNK
Cas7	Cas8b	K151	K545	GTGAFASGEG**K**AQK	NIV**K**R

Cas7	Cas8b	K151	K554	GTGAFASGEGKAQK	IYTEAINKLNEYNK
Cas7	Cas8b	K151	K606	GTGAFASGEGKAQK	FFNEEKKDGEDEELEHHHHHH
Cas7	Cas8b	K151	K607	GTGAFASGEGKAQK	FFNEEKKDGEDEELEHHHHHH
Cas7	Cas8b	K154	K64	AQKTFR	GVEYEEFSSKK
Cas7	Cas8b	K154	K65	AQKTFR	KINNYAYK
Cas7	Cas8b	K154	K73	AQKTFR	KGSAR
Cas7	Cas8b	K154	K187	AQKTFR	ILNEQYYNKYGK
Cas7	Cas8b	K154	K252	AQKTFR	LEQGKK
Cas7	Cas8b	K154	K357	AQKTFR	VKNILR
Cas7	Cas8b	K154	K400	AQKTFR	TFFPNNKTEGNFDK
Cas7	Cas8b	K154	K482	AQKTFR	TEKNKK
Cas7	Cas8b	K154	K523	AQKTFR	LLNIQYKER
Cas7	Cas8b	K154	K528	AQKTFR	GSKPFYSR
Cas7	Cas8b	K154	K538	AQKTFR	LNGLKLNK
Cas7	Cas8b	K154	K554	AQKTFR	IYTEAINKLNEYNK
Cas7	Cas8b	K154	K560	AQKTFR	LNEYNKNYYK
Cas7	Cas8b	K221	K65	VVYKPGENFFIGDLQNR	GVEYEEFSSKKINNYAYK
Cas7	Cas8b	K245	K541	ISLNFDVEEEKIR	LNKNIVK
Cas7	Cas8b	K245	K545	ISLNFDVEEEKIR	NIVKR
Cas7	Cas8b	K250	K73	SIKDFSIK	KGSAR
Cas7	Cas8b	K250	K554	SIKDFSIK	IYTEAINKLNEYNK
Cas7	Cas8b	K286	K65	LSYKGR	KINNYAYK
Cas7	Cas8b	K286	K554	LSYKGR	IYTEAINKLNEYNK
Cas7	Cas8b	K293	K73	EINLKDIK	KGSAR
Cas7	Cas8b	K296	K252	DIKDIR	LEQGKK
Cas7	Cas8b	K296	K538	DIKDIR	LNGLKLNK
Cas7	Cas5	K65	K84	GFDGSNGKDIFVR	ISENLINTKK
Cas7	Cas5	K65	K85	GFDGSNGKDIFVR	KSMNIIHER
Cas7	Cas5	K76	K84	EIESEKGGIKDGK	ISENLINTKK
Cas7	Cas5	K76	K85	EIESEKGGIKDGK	KSMNIIHER
Cas7	Cas5	K76	K97	EIESEKGGIKDGK	TQIKIEFLK
Cas7	Cas5	K80	K19	EIESEKGGIKDGK	KPYTTTSPLTYSIPTR
Cas7	Cas5	K80	K84	EIESEKGGIKDGK	ISENLINTKK
Cas7	Cas5	K80	K85	GGIKDGK	KSMNIIHER
Cas7	Cas5	K80	K97	GGIKDGK	TQIKIEFLK
Cas7	Cas5	K83	K84	DGKAR	ISENLINTKK
Cas7	Cas5	K83	K85	DGKAR	KSMNIIHER
Cas7	Cas5	K87	K84	AKDFNENVDEILQK	ISENLINTKK
Cas7	Cas5	K87	K85	AKDFNENVDEILQK	KSMNIIHER
Cas7	Cas5	K133	K72	SLNKVNLK	NPVKK
Cas7	Cas5	K133	K84	SLNKVNLK	ISENLINTKK
Cas7	Cas5	K133	K102	SLNKVNLK	IEFLKDACYR
Cas7	Cas5	K137	K72	VNLKHIK	NPVKK
Cas7	Cas5	K137	K84	VNLKHIK	ISENLINTKK
Cas7	Cas5	K137	K97	VNLKHIK	TQIKIEFLK
Cas7	Cas5	K137	K176	VNLKHIK	KGDIEFEDDREYFTETIPVEMDAER

Cas7	Cas5	K140	K72	HIKGTGAFASGEGK	NPVKK
Cas7	Cas5	K140	K84	HIKGTGAFASGEGK	ISENLINTKK
Cas7	Cas5	K140	K85	HIKGTGAFASGEGK	KSMNIIHER
Cas7	Cas5	K140	K97	HIKGTGAFASGEGK	TQIKIEFLK
Cas7	Cas5	K140	K102	HIKGTGAFASGEGK	IEFLKDACYR
Cas7	Cas5	K140	K176	HIKGTGAFASGEGK	KGDIEFEDDREYFTETIPVEMDAER
Cas7	Cas5	K151	K72	GTGAFASGEGKAQK	NPVKK
Cas7	Cas5	K151	K84	GTGAFASGEGKAQK	ISENLINTKK
Cas7	Cas5	K151	K85	GTGAFASGEGKAQK	KSMNIIHER
Cas7	Cas5	K151	K97	GTGAFASGEGKAQK	TQIKIEFLK
Cas7	Cas5	K154	K84	AQKTFR	ISENLINTKK
Cas7	Cas5	K154	K85	GTGAFASGEGKAQKTFR	KSMNIIHERTQIK
Cas7	Cas5	K154	K97	AQKTFR	TQIKIEFLK
Cas7	Cas5	K207	K84	SKMGHMPR	ISENLINTKK
Cas7	Cas5	K207	K85	SKMGHMPR	KSMNIIHER
Cas7	Cas5	K269	K126	LDELIDELANYGDKIEK	ESLKEHR
Cas7	Cas5	K279	K85	VVFVADKNLR	KSMNIIHER
Cas7	Cas6	K151	K22	GTGAFASGEGKAQK	DIPKIR
Cas8b	Cas5	K73	K72	KGSAR	NPVKK
Cas8b	Cas5	K187	K84	ILNEQYYNKYGK	ISENLINTKK
Cas8b	Cas5	K190	K84	YGKTSK	ISENLINTKK
Cas8b	Cas5	K193	K84	TSKGK	ISENLINTKK
Cas8b	Cas5	K193	K85	TSKGK	KSMNIIHER
Cas8b	Cas5	K230	K85	IGFVTGGFKQENAWK	KSMNIIHER
Cas8b	Cas5	K236	K84	QENAWKNYPVCSCCAQK	ISENLINTKK
Cas8b	Cas5	K236	K85	QENAWKNYPVCSCCAQK	KSMNIIHER
Cas8b	Cas5	K236	K97	QENAWKNYPVCSCCAQK	TQIKIEFLK
Cas8b	Cas5	K400	K84	TFFPNNKTEGNFDK	ISENLINTKK
Cas8b	Cas5	K400	K85	TFFPNNKTEGNFDK	KSMNIIHER
Cas8b	Cas5	K421	K72	KISYK	NPVKK
Cas8b	Cas6	K73	K106	KGSAR	ELKTTER

Table 6.3 Intra-protein cross-links identified in the Cas5 *protein in C.thermocellum* Cascade complex.

Residue1 and Residue2: The two residues in the protein which were observed to be cross-linked. Peptide1 and Peptide2: The two peptides which were observed to be cross-linked. Cross-linked lysine residues in the peptides, are highlighted in blue.

Residue 1	Residue 2	Peptide1	Peptide2
K18	K84	YLVFDISASYGHF**K**K	ISENLINT**K**K
K18	K85	YLVFDISASYGHFKK	**K**SMNIIHER
K19	K84	**K**PYTTTSPLTYSIPTR	ISENLINT**K**K
K19	K97	**K**PYTTTSPLTYSIPTR	TQI**K**IEFLK
K58	K176	EDYQEHFT**K**PQAK	**K**GDIEFEDDREYFTETIPVEMDAER
K62	K157	EDYQEHFTKPQA**K**IAIGIR	EV**K**GNK
K72	K84	NPV**K**K	ISENLINT**K**K
K72	K102	NPV**K**K	IEFL**K**DACYR
K84	K97	ISENLINT**K**K	TQI**K**IEFLK
K84	K102	ISENLINT**K**K	IEFL**K**DACYR
K85	K85	**K**SMNIIHER	**K**SMNIIHER
K85	K97	**K**SMNIIHER	TQI**K**IEFLK
K85	K102	**K**SMNIIHER	IEFL**K**DACYR
K115	K122	IYFTHTD**K**QIYER	L**K**ESLK
K115	K126	IYFTHTD**K**QIYER	ESL**K**EHR
K157	K176	EV**K**GNK	**K**GDIEFEDDR

Table 6.4 Intra-protein cross-links identified in the Cas6 *protein in C.thermocellum* Cascade complex.

Residue1 and Residue2: The two residues in the protein which were observed to be cross-linked. Peptide1 and Peptide2: The two peptides which were observed to be cross-linked. Cross-linked residues in the peptides, are highlighted in blue.

Residue 1	Residue 2	Peptide1	Peptide2
M1	K106	**M**DIK	EL**K**TTER
K22	K87	DIP**K**IR	EIDM**K**DK
K22	K89	DIP**K**IR	D**K**VMSILEK
K101	K106	GYVL**K**TR	EL**K**TTER
K101	K219	GYVL**K**TR	VG**K**GFVLEHHHHHH
K106	K219	EL**K**TTER	VG**K**GFVLEHHHHHH

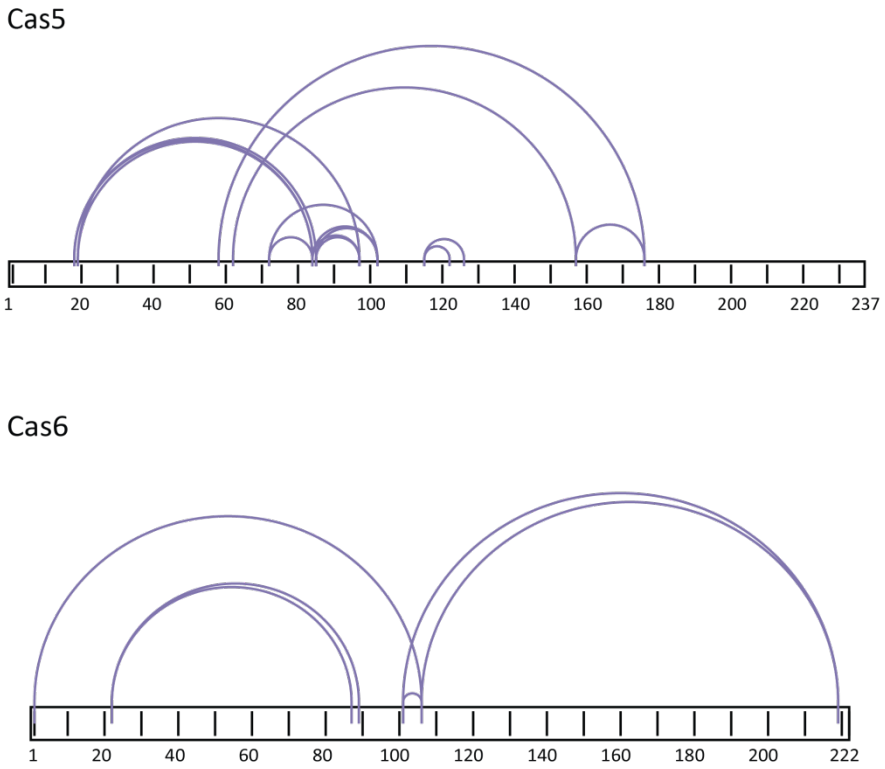


Figure 6.6 Intra-protein cross-links identified in Cas5 and Cas6 protein in the Type I-B Cascade complex from *C. thermocellum*.

Intra-protein cross-links for the Cas5 and Cas6 protein (as summarized in Table 5.3 and 5.4), mapped on the respective protein sequence using XiNET.

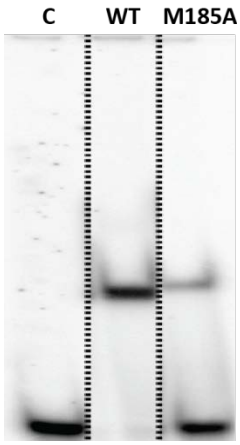


Figure 6.7 EMSA to confirm M185 residue in *Mm* Cas6b binds the cognate crRNA.

The Mm Cas6b proteins (Wild-type and mutated) were analyzed for binding with their cognate crRNA using EMSA and visualized by autoradiography. Lane - C: Control only γ-[^{32}P]-ATP labeled crRNA, Lane - WT: Cas6b Wild-type binding to the labeled crRNA, Lane - M185A: Cas6b protein with M185A mutation binding to the labeled crRNA. Figure provided by Hagen Richter (MPI Terrestrial Microbiology, Marburg).

6.2 Abbreviations

ACN	acetonitrile
AGC	automatic gain control
APS	ammonium per sulfate
ATP	adenosine-5'triphosphate
AQUA	absolute quantification
bp	base pairs
BS3	bis(sulfosuccinimidyl)suberate
BSA	bovine serum albumin
Cas	CRISPR associated
Cascade	CRISPR associated antiviral defense
CID	collision induced dissociation
COOT	Crystallographic Object-Oriented Toolkit
CRISPR	clustered regulatory interspaced short palindromic repeats
crRNA	CRISPR RNA
crRNP	CRISPR ribonucleoprotein
DDA	data dependent acquisition
DHB	dihydroxy benzoic acid
DML	dimethyl labeling
DNA	deoxyribonucleic acid
DTT	dithiothreitol
e.g.	for example, *exempli gratia*
EDTA	ethylene diamine tetraacetic acid
EM	electron microscopy
EMSA	electrophoretic mobility shift assay
ESI	electrospray ionization
et al	and others, *et alli*
FA	formic acid
FDR	false discovery rate
FLAG-tag	FLAG-octapeptide of sequence DYKDDDDK for labeling and purification of fusion proteins
FTMS	Fourier-Transform mass spectrometer
HCD	high energy collision dissociation
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HPLC	high performance liquid chromatography

i.e.	that is, *id est*
IAA	iodoacetamide
IPG	Immobiline DryStrip gel
IPTG	isopropyl β-D-1-thiogalactopyranoside
KO	knock-out (deletion)
KH domain	K homology domain
iBAQ	intensity based absolute quantification
IM	immonium ion
LC	liquid chromatography
LTQ	linear trap quadrupole
M	Molar, mol/l
m/z	mass-to-charge (ratio)
MALDI	matrix assisted laser desorption/ionization
MS	mass spectrometry
MS/MS	tandem mass spectrometry
NMR	nuclear magnetic resonance
OMSSA	Open Mass Spectrometry Search Algorithm
ORF	open reading frame
PAGE	polyacrylamide gel electrophoresis
PAM	protospacer adjacent motif
PDB	protein data bank
PCI	phenol-chloroform-isoamyl alcohol
PCR	polymerase chain reaction
Phyre	Protein Homology/analogY Recognition Engine
pIEF	peptide isoelectric focusing
ppm	parts per million
Q Exactive HF	Q Exactive High Field
RAMP	repeat associated mysterious protein
RBD	RNA binding domain
RF	radio frequency
RNA	ribonucleic acid
RNase	ribonuclease
RNP	ribonucleoprotein
rpm	rounds per minute
RRM	RNA recognition motif
SDS	sodium dodecyl sulfate

SILAC	stable isotope labeling with amino acid in cell culture
STAGE-Tips	Stop and Go Extraction - Tips
TFA	trifluoroacetic acid
Tof	time-of-flight
UHPLC	ultra high performance liquid chromatography
UniPROT	universal protein resource
UPS2	universal protein standard 2
UV	ultra violet
v/v	volume/volume
w/v	weight/volume
w/w	weight/weight
WT	wild type
XIC	extracted ion chromatogram
YPC	yeast-peptone casaminoacids
Δ	delta (deletion)
µ	micro

Ct	*Clostridium thermocellum*
Mk	*Methanopyrus kandleri*
Mm	*Methanococus maripaludis*
Pf	*Pyrococcus furiosus*
Ss	*Sulfolobus solfataricus*
Tt	*Thermus thermophilus*

Acknowledgements

I would like to thank many people who have helped and supported me during my journey for a PhD in the last few years.

First, I would like to express my deepest gratitude to my supervisor Prof. Henning Urlaub for giving me this opportunity to be a part of the amazing CRISPR field, one of the hottest topics in research today. Thank you, Henning for your constant support, guidance and trust in my abilities and the opportunities to visit so many international conferences. You are indeed the coolest boss one could ever ask for.

I would like to thank my committee members Prof. Jörg Stülke and Prof. Peter Rehling for their time and useful suggestions in the committee meetings and Prof. Rolf Daniel, Prof. Uwe Groß and Prof. Patrick Cramer for being a part of the extended thesis committee.

Sincere thanks to all my collaborators and the members of DFG forschergruppe 1680. To all the fellow graduate students Ajla, Judith, Britta, Jutta, Hagen who have always been generous in sharing their samples and answering my innumerable queries related to the CRISPR field. I would like to thank Anita, the lifeline of the FOR 1680 who has been very generous with her time, always guided me in the scientific discussions related to different projects and publications, helped me organize the fantastic CRISPR meeting in Göttingen and also for critically reading the parts of this thesis. I also thank the people from Netherlands, especially Raymond, for sharing the precious CRISPR complexes and critically reading the manuscripts and helping me with the posters. I am deeply thankful for all these wonderful collaborations that introduced me to the fascinating world of CRISPR-Cas systems.

It is my pleasure to acknowledge all my current and previous colleagues of the Bioanalytical Mass Spectrometry group for providing a stimulating and fun environment to learn and grow. First, I would like to thank our excellent multi-tasking technicians Uwe, Moni, Annika and Lisa, for keeping the day-to-day jobs on track, the true powerhouse of the lab. I would like to thank Carla, for supervising me during the lab rotation days, my first steps in the mass spec world. All the people in the office and the cave, Christof, Kuan-Ting, Chung-Tien, Samir for their useful advices and for helping me innumerable times with the data analysis and statistics. Olex for his help with the protein-protein cross-linking experiments, Jasmin and Christin for providing a nice working atmosphere in the office and Juliane, for taking care of all the non-scientific work and managing the international trips and reimbursements time and again. I would like to thank the cross-linkers Katha, Uzma and Saadia for answering the endless number of questions that I asked and continue to ask till date and teaching me all what I know today about protein-RNA cross-linking.

Life in the office would have been undeniably boring without the coolest Bulgarians around. Ilian, Miro and Aleks I will always remember all our endless laughing sessions and the craziest things we did in the office. Also for the tremendous help you gave me scientifically with all sorts of data analysis and troubleshooting. Sunit, my desi connection in the lab, I thank you for all

your guidance in various ways within and outside the lab and our intense discussions about Indian politics, cricket, IPL, movies and especially Bigg Boss have always been so entertaining. A special thanks to the sweetest person of our group, Romina, for her help and encouragement throughout, for the fun at ASMS and for critically reading this thesis and giving me numerous suggestions for improvement. Romi, I won't be wrong if I say you were the best part of this group. The time spent with you guys is what I will remember the most.

The journey through PhD was like a rollercoaster ride and I was very lucky to have started it with the IMPRS Molecular Biology program. The excellent co-ordination from Steffen Burkhardt and Kerstin Grüniger and the company of the highly enthusiastic and fun-loving fellow MolBios is irreplaceable. All the PhD retreats, trip to the Weizmann Institute and organizing the fantastic Horizons symposium was a great learning experience when one is among the best young scientists from all over the world. Also, I want to thank my methods course group MolBio4, Bernard, Maria and Ewa for the fun moments in all the methods courses, movie nights, gaming nights and get-togethers. You guys are the best and I will always stay in touch with you.

The city of Göttingen has given me several good friends and in this regard I would like to thank all the desi PhD students with whom I never felt that I am staying away from home. I want to thank Hema and Vinay for making my days in the MPI so much fun with all the entertaining lunch sessions. Right from the very first days in Göttingen, the best part of my journey was spent with very special friends Ankit, Ananya T, Avani, Nicee, Panchhi, Pawan, Priyanaka, Soham, Sumir and Upasana. First of all, thanks to all of you for the rocking Indian Culture Night, the most memorable event we all organized together. All the pot-lucks, movie-nights, playing Cards, UNO, dumb-charades, Siedler and trips all over Europe, the time spent with you guys will always be one of the best memories of my life. You guys are amazing, stay the way you are, always.

My best friends Veena and Vinita, the two crazy minions I met here that have now become an essential part of my life. You guys are a true gems and I can never thank you enough for giving me a family away from home. A very big thanks to Heena, thank you for reading every word of this thesis. You are the best person in the whole world and I am really happy and lucky to have you by my side when I achieve this milestone. This friendship with you guys is priceless and I am sure it will last for a lifetime.

The biggest thanks of all goes to my family, all my cousins especially Naveen, Hitesh, Deepak, Neha and Khushboo, your unconditional love is my biggest support. I cannot imagine my current position without the love and support from my parents. Mumma you are the best teacher one can ever have, you are the strength and the reason for my positive thinking because of which I am able to accomplish this goal and Papa I know how difficult it was for you to let me come to Göttingen, one of the biggest decisions of your life. But you guys have made me what I am today and I know you must be proud.

Lastly, I thank the almighty for giving me the wisdom and blessing, protecting and guiding me throughout this period.

Curriculum-Vitae

Personal details

Name	**Kundan Sharma**
Date of Birth	30th May 1987
Place of Birth	New Delhi
Nationality	India

Education

10/2011 - present	IMPRS Molecular Biology, Göttingen, Germany
	PhD Thesis "Investigation of prokaryotic immune defense system with quantitative and structural mass spectrometry"
	Bioanalytical Mass Spectrometry Group,
	Max Planck Institute for Biophysical Chemistry, Göttingen
09/2010 - 09/2011	IMPRS Molecular Biology, Göttingen, Germany
07/2008 - 05/2010	MSc. Microbiology, Department of Microbiology, University of Delhi, New Delhi, India
07/2005 - 05/2008	BSc. Microbiology, University of Delhi, New Delhi, India
04/1991 - 05/2005	New Green Fields Public School, Saket, New Delhi, India

Scholarships/ Awards

09/2010 - 12/2012	Stipend of the Excellence Foundation for the Promotion of the Max Planck Society
2008 - 2010	Gold Medal for 1st rank in MSc. Microbiology
2008 - 2010	Monstato Scholarship during MSc. Microbiology
2005 - 2008	1st Rank in Bsc. Microbiology

Conference Contributions (Posters)

Protein-RNA contacts in CRISPR-Cas protein complexes of the prokaryotic adaptive immune systems, 62nd **ASMS** Conference, Baltimore, USA, 06/2014

MS monitoring of protein-RNA interactions within the CRISPR-Cas system, 3rd European **CRISPR Meeting**, Berlin, Germany, 05/2014

MS monitoring of protein-RNA interactions within the CRISPR-Cas system, Biochemical Society **CRISPR Meeting**, St. Andrews, UK, 06/2013

Quantitative proteomics of the prokaryotic immune defense system including the analysis of protein-RNA interactions within the CRISPR-Cas system, 61st **ASMS** Annual Conference on Mass Spectrometry and Allied Topics, Minneapolis, MN, USA 06/2013

Quantitative proteome analysis of the prokaryotic immune system, 6th **EU Summer School** in Proteomics Basics, Brixen, South Tyrol, Italy, 08/2012

Publications

Sharma K, Hrle A, Kramer K, Sachsenberg T, Staals R, Randau L, Marchfelder A, Oost J, Kohlbacher O, Conti E, Urlaub H. *Analysis of protein-RNA interactions by UV-induced cross-linking and mass spectrometry.* Methods, May 2015.

Cass S, Haas K, Stoll B, Alkhnbashi O, **Sharma K**, Urlaub H, Backofen R, Marchfelder A, Bolt E. *The role of Cas8 in type I CRISPR interference*. Biosci. Rep., March 2015.

Staals R, Zhu Y, Taylor D, Kornfeld JE, **Sharma K**, Barendregt A, Koehorst J, Vlot M, Varossieau K, Neupane N, Sakamoto K, Suzuki T, Dohmae N, Yokoyama S, Schaap P, Urlaub H, Heck AJR, Nogales E, Doudna JA, Shinkai A, Oost J. *RNA-targeting Type III-A CRISPR-Cas complex of Thermus thermophilus.* Mol. Cell, Oct 2014.

Hrle A, Maier LK, **Sharma K**, Ebert J, Basquin C, Urlaub H, Marchfelder A, Conti E. *Structural analyses of the CRISPR protein Csc2 reveal the RNA-binding interface of the type I-D Cas7 family.* RNA Biology. July 2014.

Plagens A, Tripp V, Daume M, **Sharma K**, Klingl A, Hrle A, Conti E, Urlaub H, Randau L. *In vitro assembly and activity of an archaeal CRISPR-Cas type I-A Cascade interference complex.*Nucleic Acids Res. 2014 Feb 5.

Brendel J, Stoll B, Lange SJ, **Sharma K**, Lenz C, Stachler AE, Maier LK, Richter H, Nickel L, Schmitz RA, Randau L, Allers T, Urlaub H, Backofen R, Marchfelder A. *A complex of Cas proteins 5, 6, and 7 is required for the biogenesis and stability of crRNAs in Haloferax volcanii.* J Biol Chem. 2014 Jan 23.

Bhaskar V, Roudko V, Basquin J, **Sharma K**, Urlaub H, Séraphin B, Conti E. *Structure and RNA-binding properties of the Not1-Not2-Not5 module of the yeast Ccr4-Not complex.* Nat Struct Mol Biol. 2013 Nov;20(11):1281-8.

Sharif H, Ozgur S, **Sharma K**, Basquin C, Urlaub H, Conti E. *Structural analysis of the yeast Dhh1-Pat1 complex reveals how Dhh1 engages Pat1, Edc3 and RNA in mutually exclusive interactions.* Nucleic Acids Res. 2013 Sep;41(17):8377-90.