EXPRESSION AND PURIFICATION OF TWO CRISPR-CAS PROTEINS, Csm3 AND Csm5 FROM MYCOBACTERIUM *TUBERCULOSIS*

by

Marziah Hashimi

A thesis submitted in partial fulfillment
of the requirements for the degree

of

Master of Science

in

Biochemistry

MONTANA STATE UNIVERSITY
Bozeman, Montana

November 2015

ACKNOWLEDGEMENTS

I would not have achieved what I have with this thesis without the help and support of numerous mentors, co-worker, and my amazing friends and family. I would like thank my advisor Dr. Martin Lawrence for his guidance and Dr. Valérie Copié for her encouragement. My sister, Razia Dawlaty, who is both my best friend and the mother I never had, has always supported me. She and her beautiful daughters, Ranna, Hosna, and Sana have brought joy to my everyday life and pushed me to complete this thesis and I could never thank them enough.

TABLE OF CONTENTS

TABLE OF CONTENTS-CONTINUED

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

LIST OF FIGURES –CONTINUED

Figure Page

ABSTRACT

One third of the World's population is infected with tuberculosis (TB). TB disease is caused by bacterium called *Mycobacterim tuberculosis*, which is a facultative intracellular parasite that is transferred through the air from one person to another in close contact. A six month course of four antimicrobial drugs is the only current treatment for drug-sensitive TB. However, multi-drug resistance TB is difficult to treat. Phage therapy might be one answer as a treatment for multi-drug resistance TB. In order for phage therapy to have a chance against TB, the immune system of bacteria, known as CRISPR/Cas needs to be inhibited.

Our lab has taken a structural and biochemical approach to try to understand the CRISPR/Cas system in *M. tuberculosis*. We have cloned, expressed, and purified individual Csm proteins from the H37Rv M. tuberculosis strain. Two Csm protein, Csm3 and Csm5 were successfully purified to homogeneity in yields suitable for structure and biochemical studies. While to date, each has failed to produce crystals, the ability to the express and purify each of these proteins will allow further biochemical characterization of Csm3 and Csm5.

INTRODUCTION

According to the Center for Disease Control and Prevention, one third of the world's population is infected with tuberculosis (TB), which is the leading cause of death just behind HIV/AIDS worldwide. In 2014 there were 9,421 TB cases in United States resulting in a rate of 2.96 cases per 100,000 people[1]. TB is a disease that is caused by a bacterium called *Mycobacterium tuberculosis*, which can infect any part of the body, but usually attacks the lungs. It can reproduce within the phagocytic cells, including dendritic cells and alveolar macrophages [2]. *M. tuberculosis* is a facultative intracellular parasite that is transferred through the air from one person to another within close contact [1, 2].

A person can be infected with TB without showing symptoms of the disease; this is known as latent TB. Latent TB occurs because the immune system of the infected patient is strong enough to prevent the growth of bacteria. However, if the immune system is weakened, the bacteria becomes active and the person becomes ill [2, 3]. People with compromised immune systems, such as people with the HIV infection, are at a higher risk of developing TB [2].

Treatment of Tuberculosis

Drug-sensitive TB can be treated with a standard six-month course of four antimicrobial drugs. These drugs have two important roles; one is antibacterial activity and the other is the capacity to inhibit the onset of resistance. There is also one vaccine

against TB, Bacille Calmetta-Guerin (GCG), which is controversial because it can potentially have no effect at all.

Multi-drug resistance TB is a strain of TB that is resistant to standard drug therapy. This makes treatment difficult since there are no alternative treatments. Phage therapy might be the answer to multi-drug resistance TB as the next treatment option. Bacteriophages (Phages) are bacterial viruses that invade bacterial cells and, in the case of lytic phages, disrupt bacterial metabolism which causes the bacterium to lyse. Phage therapy uses phages as therapeutic agents to treat bacterial infections. Compared to antibiotics, phage therapy has many benefits: phage therapy is very specific, since usually one virus has a specific bacterial target, and it has fewer side effects. No serious side effects have been described with phage therapy yet, but there can be multiple side effects with antibiotics. While it can take up to several years to develop new antibiotics, selecting a new virus as the phage agent is a rapid process, only taking days or weeks [4]. Phage therapy has great potential to be a successful treatment against TB. However, this bacterium has an adaptive immune system against viruses called Clustered Regulary Interspaced Short Palindromic Repeats (CRISPR)and its Associated Proteins (CRISPR-Cas) system [5]. How the CRISPR/Cas system functions and more importantly how is it regulated in *M. tuberculosis* is currently unknown. Understanding the function and regulation of the CRISPR/Cas system in *M. tuberculosis* could lead to developing better phage therapy against TB. If we knew how to how to inhibit the activity of the CRISPR/Cas system in *M. tuberculosis*, we might make phage therapy more plausible.

This thesis is focused on understanding the structure function relationships in the CIRSPR/Cas (type IIIA) system of *M. tuberculosis.*

Overview of CRISPR

As bacteria encounter a phage or plasmid, they gain a memory of the foreign DNA by integrating a short fragment of foreign nucleic acid into the host chromosome at one end of the CRISPR locus (See 1.6.1 Acquisition). Upon the second encounter with the plasmid or phage, the CRISPR loci are transcribed and processed into a library of small crRNAs (See 1.6.2 CRISPR RNA Biogenesis). Next, the crRNAs are loaded into large surveillance complexes that detect and destroy the foreign DNA in a sequence based manner (See 1.6.3 CRISPR RNA-guided Interference) (Figure I) [6-11].

CRISPR loci consist of an array of highly conserved short DNA repeat sequences that are typically 21-48 base pairs (bp) long. Each repeat is separated by a unique stretch of variable sequence called a spacer which is typically between 26 and 72bp long. Each CRIPSR locus consist of this architecture of repeat-spacer-repeat [12, 13].

Although the repeat sequence within CRIPSR loci are highly conserved, the repeats in different CRISPR loci can vary in both sequence and length. Some of the repeats are palindromic which generate predicated RNAs with stable hairpin structures, while other repeats are predicted to be unstructured. However, despite this diversity, most repeats have a conserved sequence, (GAAA(C/G), at the 3' end. Aside from the sequence of repeats being diverse, the number of CRIPSR loci varies as well. It is common for a prokaryotic chromosome to have multiple repeat-spacer units in the

CRISPR loci which range from 2 to 249. The number of CRISPR loci does not depend on the genomic size and as of 2014, 45% of bacterial and 84% of archaeal genome have CRISPR-Cas loci that have been identified [14]. A leader that is adenine and thymine rich often flanks CRISPR loci, which serves as a promoter and binding site for regulatory proteins [13, 15].

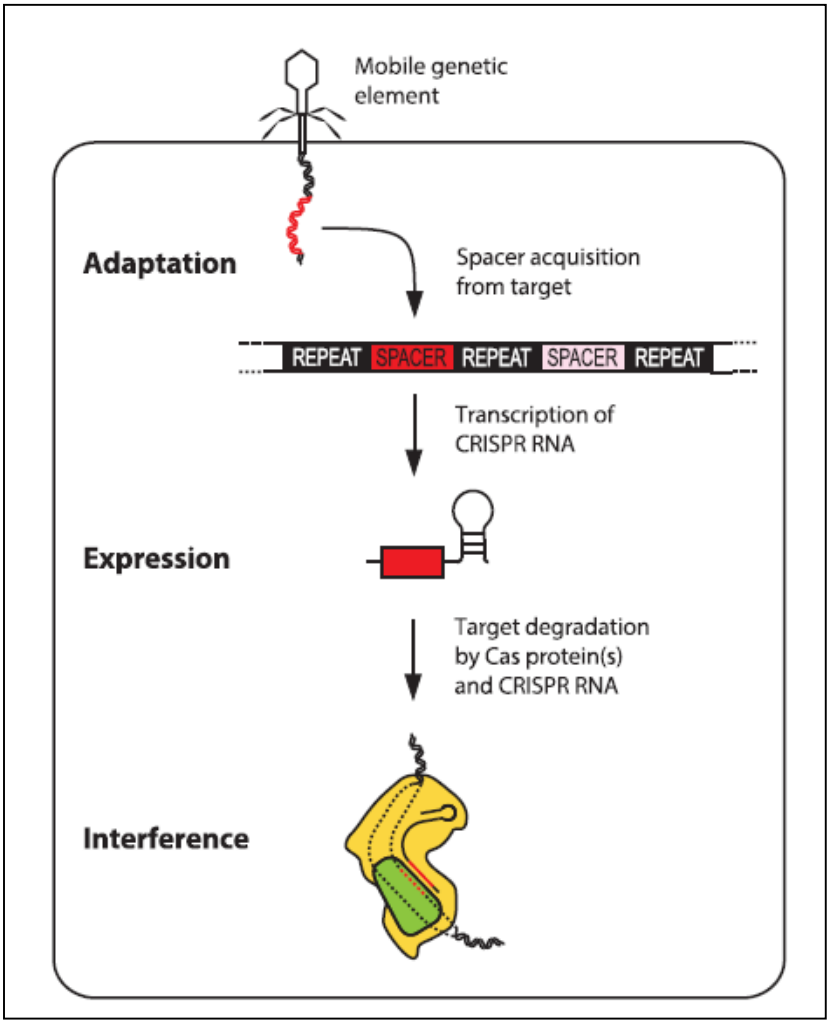


Figure 1.1: Overview of CRISPR-Cas system with permission from reference [16]

Types of CRISPR/Cas Systems

Previously the CRISPR/Cas system was divided into three types based on gene conservation and locus organization. However, currently the CRISPR/Cas system is divided into five types that are categorized under two different classes [17, 18]. The five types are subdivided to 16 subtypes. If the *cas* genes encode proteins that form into a multisubunit complex such as the CRISPR associated complex for antiviral defence (Cascade), the Csm complex, or the Cmr complex, they are labeled class 1. Member of this class include type I, III and a new putative type IV. Type I contains the signature gene *cas3* and it is divided into seven subtypes. The Cas3 protein contains an N-terminal histidine and aspartates (HD) endocuclease domain and a C-terminal helicase domain. In some cases the helicase and nuclease domains are encoded by a separate gene [18, 19]. Type III is subdivided into four subtypes, IIIA (csm complex), IIIB (cmr complex), IIIC and IIID, and all of them have *cas10* as a signature gene. Cas10 is a multidomain protein that contains an RNA recognition Motif (RRM) and is the largest subunit of type III crRNA-effector complexes. Type IIIC and D are the upgraded forms of type IIIA and IIIB respectfully [18, 20]. Type IIIC has an inactivated cyclase-like domain of Cas10, where the type IIID encoded Cas10 protein lacks the HD domain.

Class 2 CRISPR/Cas systems are defined by the presence of a single subunit crRNA-effector module which includes the type II CRISPR/Cas system and the putative new type V. Type II is the simplest CRISPR/cas system in terms of the number of genes, as it has only four, one being the signature gene *cas9* [21]. Type V of this class has a gene called *cpf1*, which is adjacent to *cas1* and *cas2* on the CRISPR array.

Despite the diversity of CRISPR-Cas systems, they all share a common molecular principle for genome silencing. They all use the mature CRISPR RNA containing a unique spacer sequence that guides one or more Cas proteins to recognize and destroy the invading nucleic acid in a sequence-specific manner [22]. Figure 2 gives an overview of the three main types of CRISPR-cas systems along with three stages of the CRISPR/Cas that will be described below in detail [23].

Mechanism of CRISPR/Cas Systems

The mechanism of CRISPR/Cas adaptive immune system occurs in three stages: (1) adaptation or spacer acquisition, (2) expression and crRNA biogenesis, and (3) interference or silencing.

Acquisition

The first step of the CRIPSR/cas immunity system is the acquisition of a new spacer (Figure 2). A short segment of foreign DNA, called a protospacer, is incorporated in the host's CRISPR array which requires the universal proteins Cas1 and Cas2 [24]. The selection of the protospacer is not random. A short segment of DNA called a protospacer adjacent motif (PAM), a 2-4 bp long sequence, is found near the protospacer. The PAM is important for correct target DNA binding and cleavage and it varied based on the CRISPR/Cas types [25]. The new spacer is added at the leader-end of the CRISPR locus. With the addition of a new spacer, the leader-end of the repeat is duplicated, which keeps the architecture of the repeat-spacer-repeat [24, 26].

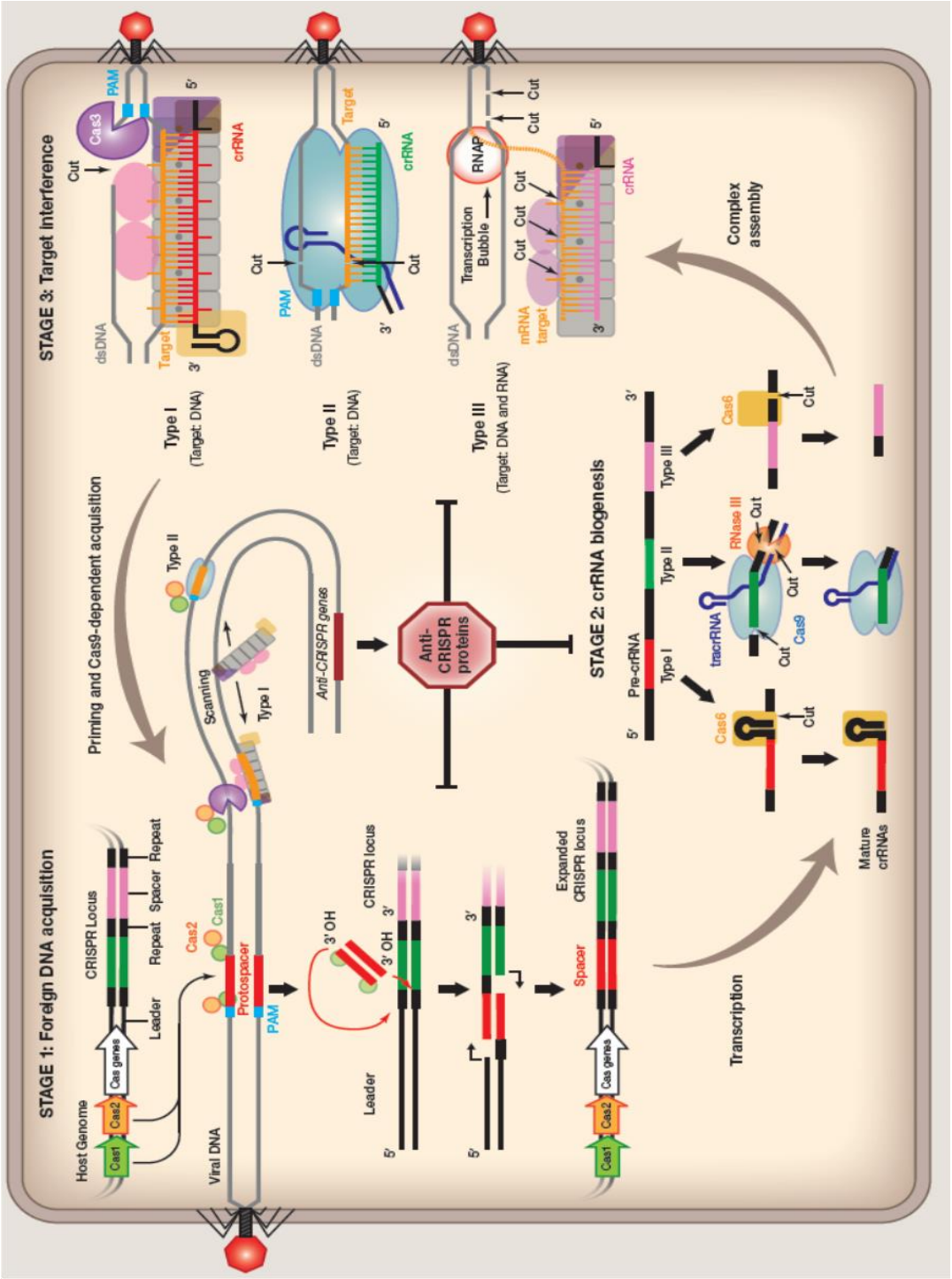


Figure 1.2: Shows the three types with their signature protein, and the three stages of CRISPR/Cas system in the three types with permission from reference [22].

Cas1 and Cas2, which are required for this step, are conserved in every genome that has a CRISPR/Cas system [17, 26, 27]. Cas1 is a metal-dependent endonuclease and capable of catalyzing the cleavage of double-stranded DNA, single-stranded DNA, and branched DNA in a sequence-independent manner [8]. Cas2 is also a metal-dependent nuclease that contains a Repeat Associated Mysterious Protein (RAMP)-like fold [28]. Cas2 proteins mainly catalyze the cleavage of dsDNA which indicates that they have deoxyribonucleases activity [29].

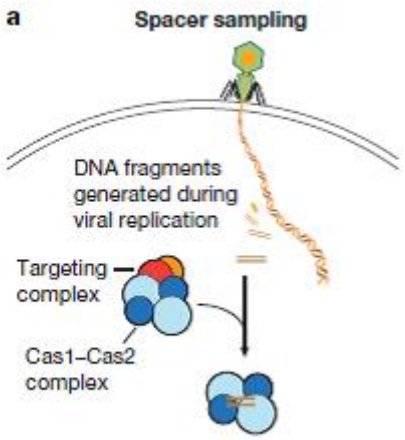


Figure 1.3: Spacer Integration: Integration requires Cas1 and Cas2 used with permission from reference [30].

James *et al.* [27] has shown that Cas1-Cas2 from *Escherichi coli* form a stable complex that is required for acquisition. The mechanism of integration of the spacer into the CRISPR array resembles retroviral integration and it is catalyzed by the Cas1-Cas2 complex. The complex provides the orientation specificity of integration of the new spacer by a nucleophilic attack at both 3' ends of the repeat. The first nucleophilic attack

is on the minus strand of the repeat distal to the leader end by carbon of the 3' end of the protospacer (half-site intermediate). The second attack is on the pulse strand at the leader-repeat border, which results in the single stranded DNA gaps as well as integration of the new spacer (Figure 3). The DNA polymerase and ligase are required for the repair the repeat gaps by an unknown mechanism [31]. Although the catalytic activity of Cas2 is not required for the acquisition of a new spacer or for the formation of Cas1 and Cas2 complex, it recruits Cas1 to the leader sequence through an indirect mechanism as James *et al.* [27] have shown. They also have shown that the Cas1-Cas2 complex has a preference for binding to the CRISPR locus which serves as the target site for spacer acquisition [27]. The final product of adaption is a new spacer which is added to the leader-end and the repeat is duplicated.

CRISPR RNA Biogenesis

The CRISPR/Cas immunity system activation starts with the second step, crRNA biogenesis of mature crRNAs that get bound to CRISPR-associated proteins forming a crRNP complex. The process of producing mature crRNA starts with transcription of the CRISPR locus from a promoter located within the leader sequence producing precursor crRNA (pre-crRNA). These pre-crRNAs undergo primary cleavage at a specific site within the repeats to yield crRNAs that consist of the entire spacer sequence flanked by partial repeat sequence on each site. In type III and type II, the primary cleavage yields intermediates that undergo further cleavage to yield two mature crRNA, while type I does not.

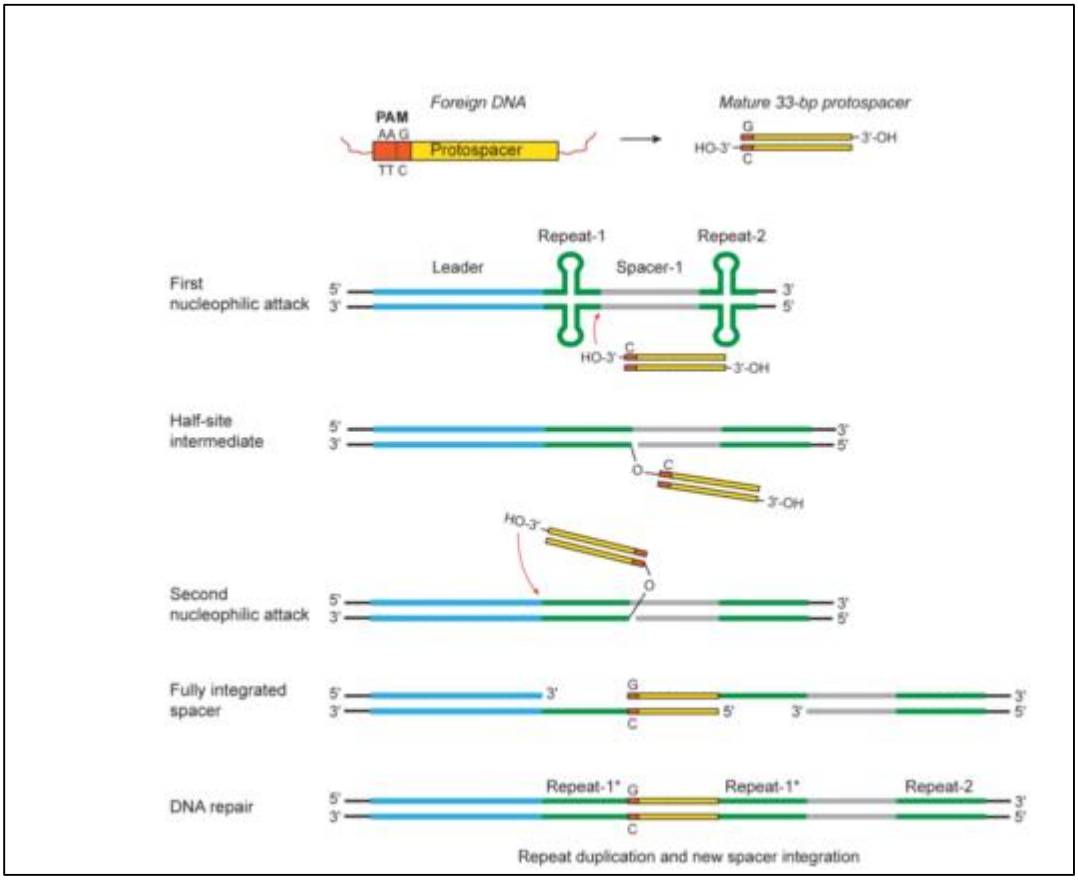


Figure1.4: The Mechanism of Protospacer Integration. The four reactions that the protospacer and spacer undergo are showing [31]

The primary processing is achieved by endoribonucleolytic cleavage within repeats. Cas6 is metal-independent endoribonuclease that cleave the repeat sequence at a conserved position typically 8 nts upstream of the repeat-spacer boundary [32]. The final product of crRNA biogenesis is a mature crRNA that contains the spacer sequence flanked by partial repeats which is wrapped by Cas proteins to form crRNP complex. The Cas proteins that are found in the crRNA effector complex share a similar structural feature and that most of them have a RNA-recognition motif (RRM) a nucleic acid-binding domain [20, 33].

CRISPR RNA-Guided Interference

In the last step of CRISPR/Cas system interference, the spacer in the mature crRNA guides the Cas proteins and destroys the invading foreign nucleic acid [6, 24, 34]. Cas proteins forming CRISPR-associated complexes are directly involved in target binding. These complexes facilitate target recognition by enhancing sequence-specific hybridization between the CRISPR RNA and a complementary target sequence [35]. At the 5' end of crRNA, a short high-affinity binding site is located that governs the efficiency of target binding. This is functionally analogous to the "seed" sequence that has been identified in eukaryotic mircroRNA(miRNA)[36, 37] . The Argonaute protein facilitated target recognition by pre-ordering the nucleotides at the 5' end of miRNA in a helical configuration. This reduces the entropic penalty by this pre-ordering which is associated with helix formation. The pre-ordering then provides a thermodynamic advantage for target binding [38]. A similar mechanism may occur during crRNA target binding which would provide an example of converging evolution between CRISPR-based immunity in prokaryotes and RNAi in eukaryotes [11]. Viruses that acquire a single mismatch in this region were able to escape detection by the CRISPR-Cas immune system [36].

CRISPR/Cas immunity in type I is mediated by Cascade and its signature protein Cas3 [6]. *E. coli* Cascade is composed of 11 protein subunits that wrap around a 61 nucleotide mature crRNA which causes kinks in the crRNA at 6 nts intervals (Figure 2). The stoichiometry of the complex is in the following order: $(Cse1)_1$, $(Cse2)_2$, $(Cas5)_1$, $(Cas7)_6$, and $(Cas6e)_1$. [6, 8].

The first 8 nts of the 5' end of crRNA, the "seed" sequence, is critical for immunity. A single mismatch in this region will lead to bacteriophages escaping Cascade-mediated immunity [36]. The Protospace Adjacent Motif (PAM) is recognized by Cse1 which promotes Cascade binding to its target [19, 39]. The seed sequence in crRNA is most likely a nucleation point for progressive hybridization of crRNA and target DNA. This leads to a local unwinding of the double stranded proteospaer DNA in an ATPase dependent manner [35, 36]. This then leads to the formation of an R-loop in which the crRNA spacer base pairs with one DNA strand of the duplex, leaving the displaced strand single stranded [10, 36]. The formation of the R-loop corresponds to a conformational change in Cascade that triggers the recruitment of the Cas3 nuclease-helicase[6]. Cas3 contains a N-terminal histidine and aspartate an HD phosphohydrolase domain, and a C-terminal helicase domains\. The domains can be encoded as one subunit or as individual subunits [17, 33]. The R-loop serves as the binding site for Cas3. Binding of Cas3 to ssDNA triggers ATPase/helicase activity which cleaves the DNA strand in the protospacer region. After the cleavage of the DNA, Cas3 is translocated along the non-target strand in the 3' to 5' direction in an ATP-independent manner. Cas3 cleaves the non-target strand with its HD-nuclease domain [40, 41].

Type II CRISPR-Cas immunity is executed by only one *cas* gene, *cas9.* The two small RNAs, crRNA, and trans-encoded crRNA (tracrRNA) that are produced from crRNA biogenesis are required for immunity [42]. A PAM is also required for target recognition by Cas9, and promotes the initial recognition and melting of the two DNA strands immediately upstream of the PAM. Following complete hybridization of the

crRNA with the target DNA, the nuclease domain of Cas9, HNH and RuvC, each cleave one DNA strand of the protospacer sequence resulting in a double stranded cut [43].

Type III CRISPR-Cas systems require target transcription, in contrast to Type I and Type II. [44]. Type III systems target both DNA and RNA, which results in co-transcriptional crRNA-guided cleavage of the target DNA and its transcripts [45-47]. The Cas10.Csm ribonuleoprotein complex requires the transcription of the target DNA; only guide crRNA complementary to the non-template strand can be directly cleaved. The Cas10 palm polymerase domain is required for the cleavage of non-template DNA[48]. The backbone subunits of type IIIA and type IIIB are Csm3 and Cmr4 respectively. [45, 46]. To date, no PAM requirements have been observed for type III CRISPR targeting [49].

Csm Proteins in *M. Tuberculosis*

M. Tuberculosis H37Rv, has two long CRISPR loci: one is adjacent to *cas* genes, CRISPR1, and the other is more distant, CRISPR2. They have 24 and 18 repeats respectively and their spacers are not duplicated between them. This indicates that they are distinct CRISPR(s)[5]. The CRISPR locus is followed by nine consecutive *cas* gens, cas2, cas1, csm6, csm5, csm4, csm3, csm2, cas10(csm1) and cas6, from 5' to 3' (Figure 1.5). Since H37Rva has cas10, it is categorized as Class 1 Type IIIA [5, 18].

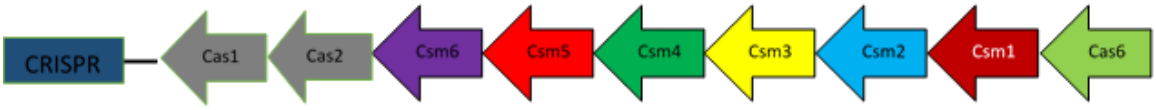


Figure 1.5: Schematic diagram of CRISPR-Cas locus in *M. Tuberculosis* H37Rv.

Electron microscopy single particle reconstruction of the type IIIA Csm complex shows that there are similarities to Cascade architecture (Figure 1.6). The type IIIA Csm complex from *Sulfolobus solfataricus* is composed of 13 subunits with the stoichiometry of: $(Cas10)_1$, $(Csm2)_{3,}$ $(Csm3)_8$, and $(Csm4)_1$. The complex is arranged into a basal body with a major and minor intertwining the crRNA. Cas10, the large subunit, serves as the anchor for the two filaments and is located at the base of the Csm complex are . While the backbone of Cascade consists of Cas7, Csm3 is the backbone of Csm complex. Csm4's position and structure correlate with Cas5 of Cascade as they both make the base of the complexes [50].

Cas6 is responsible for primary cleavage of pre-crRNA in both types I and type III. In type I-E, after the primary cleavage of crRNA transcript, Cas6 stays bound to the 3' hairpin of the mature crRNA at the head of the complex at the 3' end of crRNA [51]. In the type IIIA, Cas6 just performs the primary processing and it is not part of the complex.

Cas 6 cleaves the pre-crRNA and results in an intermediate crRNA containing 8 nts on the 5' end of crRNA (crRNA 5' handle tag) which is followed by the complete sequence of the spacer and a repeat-derived 3'handle of variable size [32]. The intermediate crRNA undergoes a final maturation at the 3' end by an unknown catalytic mechanism.

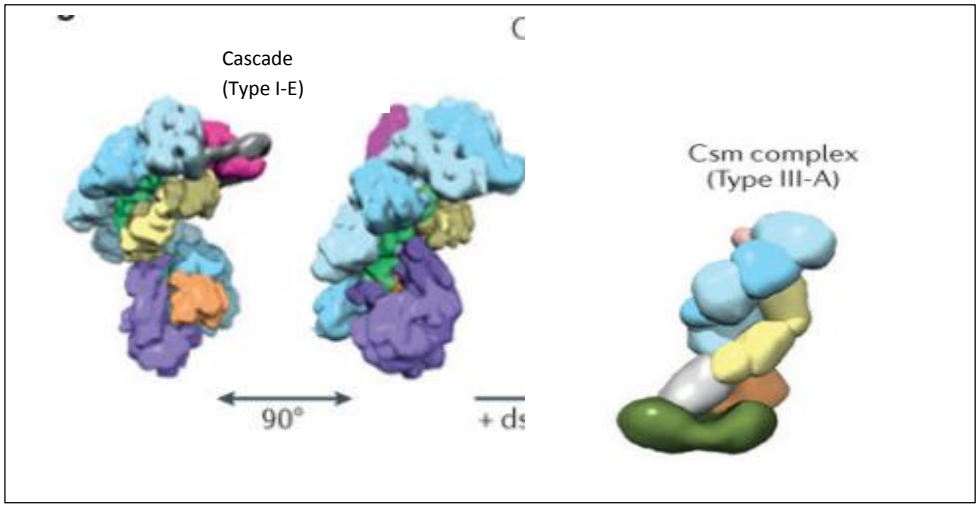


Figure 1.6: The overall shape comparison of Cascade (Type 1-E) and Csm (TypeIIA)with permission from reference [52]

In *S. epidermidis* the 3' end trimming of intermediate crRNA produces two mature crRNAs with a difference of 6 nts in size. The trimming of the 3' end of intermediate crRNA does not depend on the sequence, structure, or length of intermediate crRNA [53]. Maturation of the intermediate crRNA is achieved by a mechanism that measures the overall crRNA length using the 5' end primary processing site as the reference point [53]. Csm3 protein interacts with crRNA once every 6 nt at multiple sites. Higher levels of Csm3 results in longer crRNA as each additional copy of Csm3 extends the crRNA length by 6 nts. This indicates that Csm3 is the ruler that sets the length of crRNA [54]. In *S. epidermidis,* maturation of pre-crRNA generates two crRNA species 43 and 37 (nt) long after nuclease cleavage of the 3' end of a 71-nt long intermediate. The Csm3 subunits bind to intermediate crRNA every 6 nts segment leaving the 3' ends exposed for cleavage by an unknown nuclease. [7, 54].

Csm3 also acts as the RNase in *S. thermophilus* which targets RNA in a PAM independent manner. Csm3 produces a regular 6 nucleotide cleavage pattern of the target RNA. Cas10 is also proposed to be involved in the targeting process as it has an HD domain and shows 3'-5' exonuclease activity for ssDNA and RNA in vitro [55]. Csm2 has structural similarities with the Cse2 of Cascade and is proposed to be involved in the targeting as well [28, 53, 56].

Distinguishing Self from Non-Self

The bacteria should be able to differentiate between self and non-self when targeting foreign nucleic acids. Type I and II depend on the PAM sequence to differentiate between self and non-self, whereas Type III systems lack PAM [25, 57]. Instead, type III uses a transcription-dependent mechanism, thus there is no need for PAM for DNA melting. Type III uses a DNA targeting mechanism in a PAM-independent manner which most likely increases the flexibility and specificity as it depends on antisense protospacer transcription [44]. The RNA polymerase exposes 6-14 nucleotides of the non-template DNA strand which then provides a potential initial annealing window for the 5'-spacer sequence of crRNA to base pair with the protospacer [58]. It has been shown that *S.epidermidis* depends on the 5' handle of mature crRNA in the Csm complex to differentiate self from non-self. The base pairing between the 5' handle of mature crRNA and the repeat sequence on the CRISPR locus signals it as self which results in a "switching off" of the interference process. This may be due to the prevention of nuclease recruitment [49]. The type III systems that target RNAs may not

need the presence of PAM to distinguished self from non-self as there is no need for complementing to the RNA target [11].

Our Approach

Our approach for investigation of the structure, function and regulation of the Csm complex from *M. tuberculosis* was to express and purify the individual proteins for both biochemical and structural analysis. To determine the individual Csm protein's structure, individual *csm* genes were cloned using the Gateway system and were expressed in *E. coil*. Currently we are attempting to express the entire *csm* operon in *M. smegmatsis*. The *csm* operon from *M. tuberculous* strain H37Rva has nine csm proteins which include cas1, cas2, csm1 (cas10), csm2, csm3, csm4, csm6, and cas6. H37Rv has 18 repeats with 36 bp length and with spacers 35-40 bp long [5]. One repeat-spacer-repeat will be duplicated seven times and cloned along with the Csm genes. By having one unit of repeat-spacer-repeat in multiple copies, we will know the sequence of the crRNA that will co-purify with the csm complex. *M. smegmatis* may be the ideal expression system for expressing Mycobacterium proteins. It is an easy model to use because it shares more than 2000 homologs with *M. tuberculosis*, it is fast growing, and it is non-pathogenic. The *M. smegmatis* mc^2155 strain is hyper-transformable, only taking 3-5 days to form visible colonies [59].

CLONING, EXPRESSION, AND PURIFICATION OF *Mycobacterium tuberculosis* Csm PROTEINS

Introduction

This chapter outlines the materials and methods for cloning, expressing, and purifying Csm proteins in *E. coli*. The chapter will start by describing the cloning of *cas1*, *cas2*, and *csm* genes, *csm1-csm6.*

Cloning

The Polymerase Chain Reaction (PCR) product of *cas* and *csm* genes were cloned using gateway technology. Gateway technology is based on site-specific recombination that integrates bacteriophage lambda into *E. coli*. The Gateway technology also allows for a highly efficient and rapid way to express protein in various expression system. The integration of bacteriophage lambda into *E. coli* is facilitated by a mixture of lambda and *E. coli*-encoded recombination proteins that bring together the target sites, cleave them, and covalently attach the DNA in two separate reactions: BP and LR reaction. The site-specific attachment site (att) serves as the binding site for recombination proteins. In the Gateway technology, a donor plasmid is constructed with the gene of interest (BP reaction) first, then the expression clone is constructed from the donor clone in the LR reaction.

PCR

In our lab *M. Tuberculosis cas* and *csm* genes were amplified by PCR. The genomic DNA of *M. tuberculosis* H37Rv was used as template for nested PCR amplification reaction. Two sets of primers were used for the PCR: internal primers and external primers. The forward external primer was designed to add the following: the attB1 site, a His-tag N-terminus and the Shine-Dalgarno (SD) and Kozak on the 5' (primer 1-3) end of the gene. AttB2 sites and a stop codon were added on the 3' end (primer 1-4). The attBI and attB2 are required for the correct integration of the PCR product into the desired vector as they are the specific recombination sites. The Shine-Dalgarno and Kozak are sequences which direct the initial binding of the mRNA to the ribosome and initiate translation in prokaryotes and eukaryotes respectively. The His-tag is used for the efficient purification of soluble expressed proteins using immobilized meta-affinity chromatography (IMC).

External Primer for N-Terminus His-tag:

Up 1-3= ATTB1+SK+Kozak+MET+6X HIS

Up 1-4= ATTB2+STOP

External Primer for C-Terminus His-tag

pk 1-3= ATTB1+SK+Kozak+MET

pk 1-4= ATTB2+6XHis+STOP

The internal primers consist of two parts: a 5' portion that overlaps with the sequence from the external primers and a 3' portion that is gene specific. The properties of the internal primers are:

Forward Primers: overlap with 1-3 (external primer) + Gene specific sequence

Reverse Primer: overlap with 1-4 (external primer) +Gene specific sequence

PCR was done as described in the Gateway Manual, with the exception of variation in annealing temperatures for the different primers that were used in the reaction. The annealing temperatures were adjusted depending on the melting temperature (Tm) of each set of primers. The PCR conditions and primer sequence could not be recorded because documentation of previous research was transferred improperly. A high fidelity and high processivity polymerase, KOD HiFi from Novagen, was used in the PCR. Two step PCR was performed. The first step was done with the internal primers in 10-cycles of PCR with H37Rva genomic DNA as the template. The product of this step was used as template for the second round of PCR with external primers. The finial product of the PCR was gel-purified using QIAquick gel extraction kit (Qiagen, Catalog# 28704) and contained attB1, SD, Kozak, and 6X His-tag on N-terminus (Figure 2-1) or C-terminus (Figure 2-2) of the gene of interest and attB2.

Figure 2-1: Schematic representation of the desired PCR product His-tag being on N-His tag

Figure 2-2: Schematic representation of the desired PCR product His-tag being on C-His tag

BP Reaction and Construction of the Entry Clone

The BP reaction was performed using the purified PCR product and pDONR201 (Invitrogen) to create entry clones. The BP reaction requirements were made by having attB1 and attB2 on the PCR product and attP1 and attP2 on the donor vector. The site-specific reaction was facilitated by BP clonase enzyme. The PCR products were mixed with the donor vector and one-fourth of the recommended volume of BP clonase enzyme. The mixtures were then incubated overnight at room temperature (25℃). To stop the reaction, 1 μl of Proteinase K was added to the mixture and the reaction was incubated for 1-2 hours at 37 ℃. The product was then immediately transformed into DH5α competent cells and was plated on Luria Broth (LB) plate with 50 μg/ml kanamycin. The pDONR201 contains a ccdB killer gene which selectively targets the *E. coli* DNA gyrase and kills the cell. The ccdB gene is replaced with the gene of the interest in the BP reaction. As a result, theoretically only cells that took plasmids with the gene of interest will survive on the LB plate. Three colonies were selected for colony PCR to confirm the presence of the gene of interest using internal primers for each clone. The colonies that were positive for the correct size of PCR product were grown over night in LB media with kanymycin as an antibiotic selection. The clones were then purified using the QIAprep spin mini prep kit (Qiagen, Catalog # 27106). The positive, purified clones with the correct size by PCR were sequenced by Nevada Genomic Center to ensure there were no mutations. The results were analyzed using DNAStar (Lasergene Inc, Madison, WI) software.

LR Reaction and Construction of Expression Clone

The clones from the BP reaction that were positive for the correct size and were mutation free were used for the LR reaction because they contained the attL sites. The different destination vector contain different promoter and fusion tags which allows expression of proteins in different systems such as *E. coli*, yeast, or mammalian systems with a verified of different tag . With destination vectors, the His-tag can be added by primers which could not be cleaved off after the protein is expressed. Another possible way to incorporate a His-tag is to use a destination vector that already contains the tag that can be cleaved off. Since the tag was incorporated into the PCR constructs, a non-fusion destination vector, pDest14 typically used for native protein expression, was used as the expression vector. The LR reactions were carried out similarly to the BP reaction with the exception of using LR clonase enzyme. Since the destination vector is resistant to ampicillin, the product of the LR reaction was plated on LB plate with 100 µl/ml ampicillin. For the confirmation of the gene of interest, three clones were selected and colony PCR was performed using internal primers. Positive clones were grown in LB media with ampicillin overnight and were extracted and purified similarly to the BP clones. The expression clones were stored at -20 ℃. In summary, we attempted to clone each of the *cas* and *csm* genes twice, once with the His-tag on the N-terminus and the other with His-tag on the - terminus, because we wanted to test the effect of the His-tag on both termini on the expression of proteins. However while, 16 construct were attempted, but only 14 different constructs were actually cloned. The Csm4 and Csm6

proteins were only cloned with their His-tag at the N-terminus; the C-terminus construct was never successfully cloned.

Small Scale Expression

All 14 constructs were then tested in small scale expression trials. The pDEST14 plasmids that contained the desired gene with the His-tag were transformed into chloramphenicol resistant cells BL21-CodonPlus-(DE3)-RIL (Stratagene) and were plated on LB plates with 100 μg/ml ampicillin (AMP) and 34 μg/ml chloramphenicol (CAM). A single colony was used to inoculate 5ml of ZYP-0.8G media with AMP and CAM and was grown for 4-6 hours at 37℃ with constant shakering. These cultures were used to inoculate 500 ml of ZYP-2025 auto-inducing media with 1 to 1000 fold dilution with both antibiotics, AMP and CAM. After 18 hours of growth, the cultures had about 3.2 Optical Density at 600 nm (OD_{600}). The cells were pelleted by centrifugation at 6,000 x g and were stored at -20 ℃ until needed.

Purification of Cas and Csm Proteins

As described under the Cloning section (Page 17), the His-tag facilitates easy protein purification using immobilized metal-affinity chromatography (IMAC). A gravity column with 1ml of Ni-NTA resin was used to purify the proteins. Cell pellets were thawed and resuspended in the lysis buffer, 20 mM Tris-Cl pH 8.0, 400 mM NaCl, and 5 mM imidazole with a ratio of 5 ml per gram of cell pellet. To inhibit protease activity, freshly prepared phenyl methyl sulfonyl fluoride (0.1 mM PMSF) was added to resuspended cell pellet throughout the purification steps. The cells were then passed

through a microfluidizer (Microfludrics Corp. Model 110L) three times to ensure complete lysis of the cells. The cell lysates were clarified by centrifugation at 22,000 × g for 30 minutes. The supernatants were applied to the gravity column containing 0.5 ml bed volume of Ni-NTA resin (Qiagen) and the flow-through was collected for further analysis. The column was washed five times the column volume with lysis buffer having 10 mM imidazole, these washes were also collected for further analysis. The proteins were eluted from the column with the elution buffer, 10 mM Tris-Cl, pH 8.0, 50 mM NaCl, and 200 mM imidazole, in five fractions of bed column volume. To determine the protein purity, SDS-Poly Acrylamide Gel Electrophoresis (SDS-PAGE) was used on the flow-through, washes, and elution fractions. To determine the protein concentration of the purified protein, the Bradfrod assay was performed and bovine serum albumin (BSA) was used as the standard. Out of the 14 proteins that were screened, Cas2 and Csm5 having the C-terminal His-tag were expressed at low levels. Cas1, Csm2, and N-Terminal Csm4 were insoluble. Expression of Csm1 and Cas6 were not detected.

PURIFICATION AND CRYSTALLIZATION TRIALS FOR Csm5

Introduction

This chapter outlines the steps that were taken to scale up the expression of Csm5 and to purify it to homegensity for crystallization trials. From the small scale purification, it was concluded that Csm5 expresses at low levels and it also had many contaminants that co-eluted with Csm5 from the Ni-NTA column. Different steps were taken to increase the expression and purity of Csm5 are described later in this chapter. This chapter also outlines the initial crystallization trials that have been tried to crystalize Csm5.

Large Scale Expression and Purification of Csm5

The SDS-PAGE gel from the small scale purification showed many contaminant that co-purified with Csm5. To further purify Csm5 and to exchange the buffer to a lower salt, the eluted Csm5 was then applied to a calibrated Superdex S-75 (GE Healthcare Life Science) gel filtration column equilibrated with 10 mM Tris-Cl, pH 8.0 and 300 mM NaCl. Csm5 having a molecular weight (MW) of 42.33 kDa, its expected elution volume as monomer is between 9.75 and 10 ml. The elution of Csm5 from the column started at 12 ml and it completely eluted by 15 ml (Figure 3-1A and B). with the 2.25 ml behind its expected volume which indicated the Csm5 might be interacting with the column. The eluted fractions from the SEC columns were pooled together and Amicon spin concentrator (Amicon Ultra ™ concentrator Millipor) with 30,000 Da (MW) cutoff was used to concentrate.

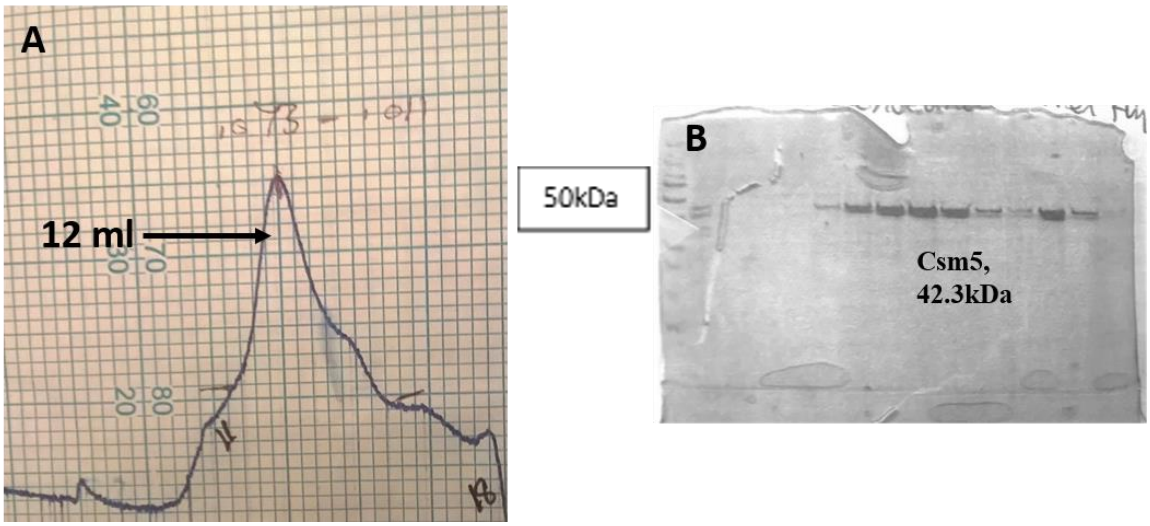


Figure 3-1: Chromotogram and SDS-Page of Csm5:
A: The peak in the chromatogram correspond with Csm5' elution.
B: 15% SDS-PAGE on the fraction that Csm5 was eluted (11 ml to 16 ml).

However, Csm5 did not concentrate well as most of the protein was lost in the process. In order to take Csm5 to a large scale purification, some changes were made in the purification step to increase the expression and purity level. The Ni-NTA column was washed five times which were then collected and analyzed by SDS-PAGE. The elution of Csm5 was seen in all five washes. The lysis and wash buffers had 10 mM imidazole which was strong enough to elute the weakly bound Csm5 from the column. To prevent the elution of Csm5 in washes, imidazole was removed from both the lysis and the wash buffers. The yeild level of purified Csm5 was low, as only 0.3 mg of protein was obtain from 1.5 L of auto-induciton media. However, Csm5 was also seen in the cell pellet in the form of inclusion bodies. To help solubilize Csm5, a high salt (1 M NaCl) lysis buffer, a lysis buffer with detergent (12% octyl glucoside), and a lower pH (pH 7.0)

buffer were tried to lyse the cell. These three different conditions did not help to solubilize Csm5.

To increase purity of Csm5, as even after size exclusion chromatography, some contaminates were eluted with Csm5, a linear imidazole gradient was used to elute the protein by Ni^{2+}-affinity chromatography. Chromatography was performed on the FPLC system using a Ni-charged HiTRap™ Chelating HP column. A 20 ml linear gradient running from 0.0 M imidazole to 1M at 1ml/min flow rate was used to elute Csm5. Both pumps of the FPLC were used for the imidazole gradient; Pump A was in elution buffer with 0.0 M imidazole and Pump B in 1 M imidazole elution buffer. The Csm5 eluted with no contaminants when pump B was at 43% of 1M imidazole (430 mM imidazole). Figure 3.2 shows a comparison of two gels; one is the elution from the Ni-NTA gravity column and the other is the elution fractions from the linear imidazole gradient Ni^{2+}-NTA affinity chromatography. As it can be seen from the gel, Csm5 eluted with no contaminates using the imidazole gradient. Csm5 only had a 10% recovery from size exclusion column so a higher salt concentration, 400 mM NaCl, was tried to improve the recovery from SEC. The eluted Csm5 was identified by MALDI-TOF MS after being digested with trypsin in-solution.

Lysis buffer without imidazole, 20 mM Tris-Cl, pH 8.0 and 400 mM NaCl, a linear imidazole gradient method to elute the protein from the Ni^{2+}-affinity column, and a higher salt concentration for the SEC were established as the standard purification protocol for Csm5. To speed the process of growing cells on a large scale, unused 2L Coca-Cola bottles were used instead of using glass flasks to grow cells.

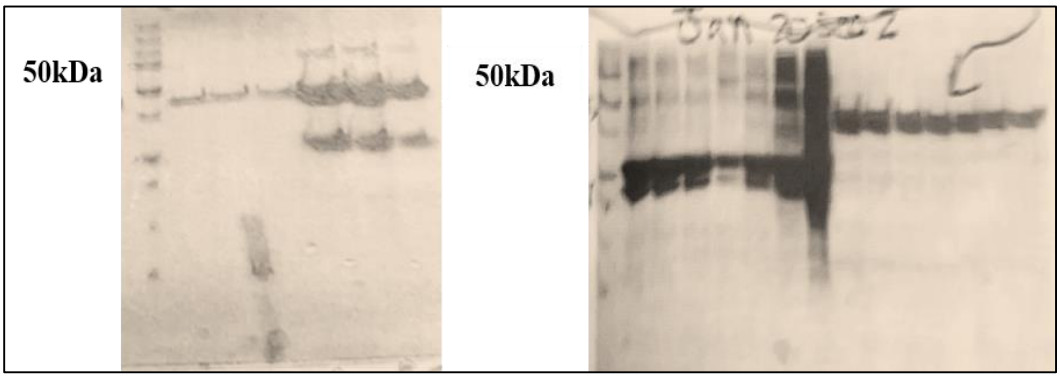


Figure 3-2: Comparison of Two Gels: The left hand gel shows the elution (last three lanes) from Ni-NTA gravity column. The right hand gel shows the elution (last seven lanes) from imidazole gradient

As compared to growing cells using flasks, bottles required less time to prepare media and were easier to dispose of after. The media did not need to be autoclaved as the bottles were already sterilized. Later in the experiment, fermenters were used to grow 8 L of cells in auto-induction media. Freshly transformed cells were found to be absolutely criticall for decent yields of expressed Csm5.

Crystallization Trials of Csm5

For the initial screening, it is recommend that the protein concentration should be 10mg/ml. However, I was unable to concentrate Csm5 above 3 mg/ml, so the initial screening was tried at this concentration. The initial screening was carried out using the sitting drop vapor diffusion method with drop composition of 400 nl well solution + 400 nl Csm5 with the commercial screens in 96-well format. Trays were setup using the Honeybee Crystallization Robot (Genomic Solutions). Only Screens 1 and 2 and the Salt Screen from the Hampton Research were set up because there was not enough protein to

try all screens. Negative controls, buffers without the protein, were also tested against the two screens. Two different conditions having phosphate yielded crystals.

The two conditions that yield crystals were:

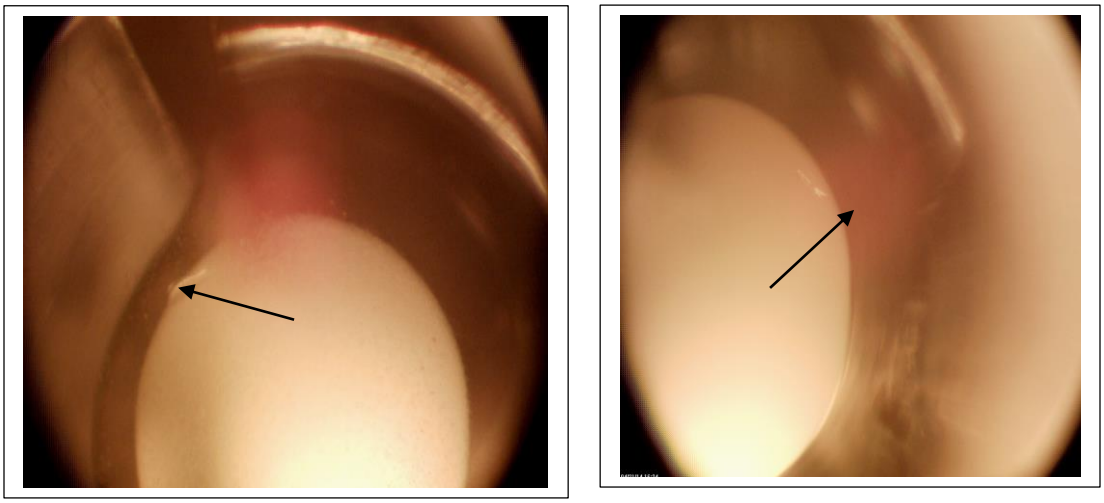
1) Hampton Crystal Screen 1&2 B6: 1.4 M, pH 6.9, 0.49 M sodium phosphate, 0.91 M postassium phosphate dibasic (Figure 2-1A) and
2) Hampton Crystal Screen 1&2 E2: 1.8 M, 0.1 M Ammonium phosphate, 4.6 M sodium acetate (Figure 2-1B).

There were no crystals in the corresponding negative control. Since both conditions that yielded crystals had phosphate, this could be an indication that phosphate helped to stabilize Csm5. To increase the recovery of Csm5 from SEC, 200mM potassium phosphate was added to SEC buffer making the SEC buffer 200 mM K_2PO4, pH 8.0 and 300 mM NaCl. Having the phosphate in the SEC buffer not only helped with recovery from the column, but it also helped with the concentration step as Csm5 could be then concentrated up to 10 mg/ml.

A second screening trial was tried with a concentration of 10 mg/ml Csm5 in a buffer with phosphate. The screens used were: Crystal Screens 1&2, Salt Screen, PEG Screen, PEG/Ion Screen, and Crystal Screen Lite Screen from the Hampton Research.

Three different conditions yielded crystals and had the following properties:

1) Hampton Crystal Screen 1&2 D2: 0.1 HEPES, pH 7.0, 30% v/v Jeffamine M-600 pH 7.0 (picture not shown),



A B

Figure 3-3: Crystals of Csm5 From the First Initial Screening: (A) is B6 and (B) is E2. Arrows are point to the crystals.

2) Hampton Crystal Screen 1&2 E1: 0.2 Calcium chloride dehydrate 0.1 M BIS-TRIS, pH 6.5, 45% (+/-)-2-methyl-2,4-pentanediol (Figure 3-3A), and
3) Emerlad Cryo Screen 2 B9: 40% (v/v) 1,2-propandediol, 0.1 M acetate, pH 4.5, 0.05 M $Ca(OAc)_2$ (Figure 3-3B).

The initial crystallization conditions were optimized by hand in 24 wells, 2 ul + 2 ul hanging-drop vapor diffusion format. Negative controls were also set up in SEC buffer without Csm5. Crystals were seen in all negative controls, and solutions with protein (Figure 3-1C). The fact that crystals were seen in every drop indicates that the crystals were probably not the Csm5 crystals, but instead might be the calcium phosphate crystals. Phosphate is negatively charged and the calcium is positively charged, thus they will yield crystals. In addition calcium phosphate is notoriously insoluble

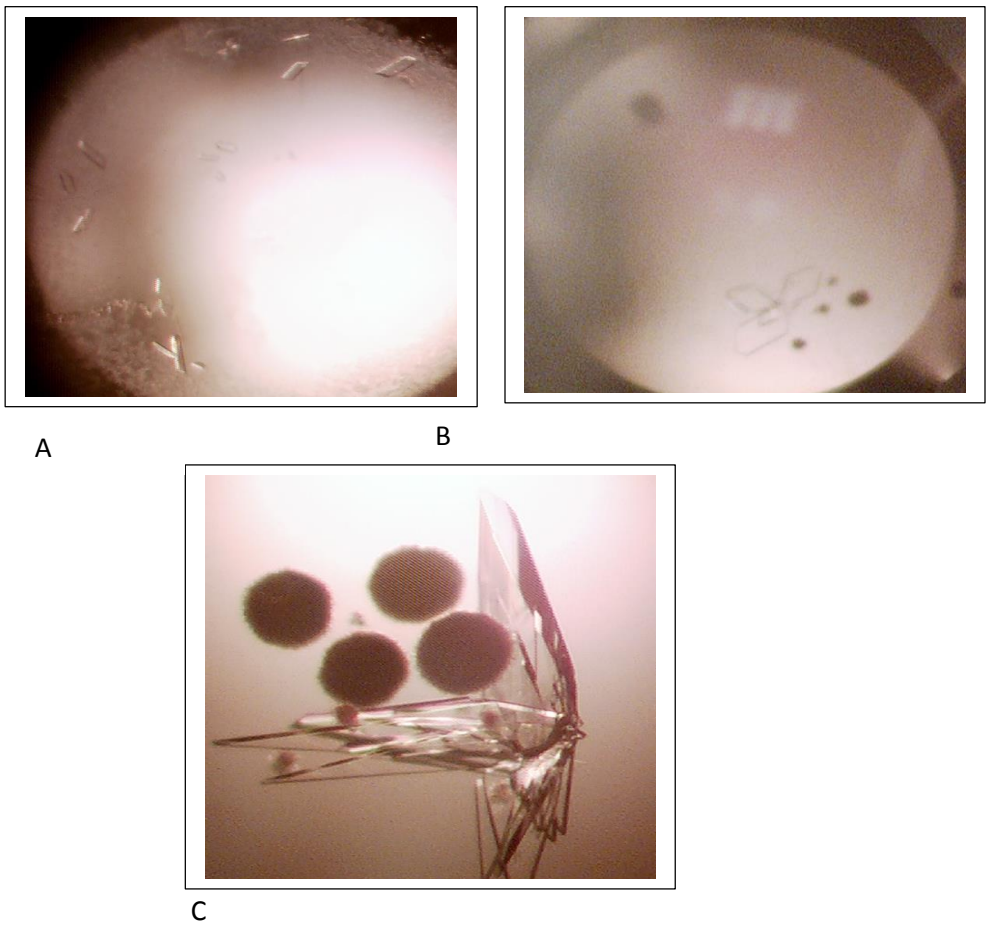


Figure 3.4: Crystals of Csm5 with 200mM Phosphate: 3-A: 0.2 Calcium chloride dehydrate 0.1 M BIS-TRIS, pH 6.5, 45% (+/-)-2-methyl-2,4-pentanediol and B: 40% (v/v) 1,2-propandediol, 0.1 M acetate, pH 4.5, 0.05 M Ca(OAc)2 are the crystals from initial screening. 3-1C one of the crystals that were seen in negative control from optimizing the crystals by hand.

Conclusion

One purpose of purifying and crystalizing Csm5 was to determine its structure which could give us some insight into its function and role in the Csm complex. Even though the conditions for expression and purification of Csm5 were optimized, it failed to produce crystals. Other biochemistry experiments could be done to determine its

function. One possible experiment could be to perform an electrophoretic mobility shift assay (EMSA) to explore Csm5 binding to crRNA.

PURIFICATION AND CRYSTALLIZATION TRIALS FOR Csm3

Introduction

This chapter outlines the steps that were taken to extract Csm3 from inclusion bodies. From the Small Scale Purification section in Chapter 2 (Page 19). Csm3 was in the cell pellet as possible inclusion bodies. To prevent protein from entering inclusion bodies, different expression temperatures were tried. When that did not work, we solubilized the protein from the inclusion bodies using a denaturing agent. The solubilized denatured protein was refolded on the column which yielded soluble Csm3. This chapter will also described the crystallization trials for Csm3.

Temperature

To prevent Csm3 from entering inclusion bodies, a small scale expression and purification experiments was carried out. The expression was carried out as described under Chapter 2, Small Scale Expression (Page 19), with the exception of growing cells at 28℃. The cells were lysed and protein was purified as described in Chapter 2, Purification of Cas and Csm Proteins. It was confirmed by SDS-PAGE that lowering the temperature did not help to prevent Csm3 from entering the cell pellet.

Denaturing and Refolding Csm3 on the Column

I decided to solubilize the inclusion bodies with denaturing agent and to extract Csm3, and to then refold protein on the column as described [60]. The cells were expressed as described in Chapter 2, Small Scale Expression (Page 19). The cells were

resuspended in the lysis buffer that had 6M guanidinium hydrochloride to solubilize the inclusion bodies, 20 mM Tris-Cl, pH 8.0 and 100 mM NaCl with 1 ml per 5 gram of cell pellet. The cells were passed through the microfluidizer three times to break them open. The cell lysate was clarified by centrifugation at 22,00 x g for 30 minutes. The solubilized protein was then bound to Ni-NTA resin in a gravity flow column. The column was washed with the lysis buffer having 10 mM imidazole, pH 8.0, 10 times the column volume (10 x CV) to remove nonspecifically bound contaminants. Next, the column was washed with 10 CV of buffer containing 0.1% Triton® X-100. To remove the detergent from the protein-detergent complex and allow the protein to refold, the column was washed with 5 mM β-cyclodextrin (Sigma, Lousi, Missouri, USA). To remove any remaining impurities and β-cyclodextrin, the column was washed with 10 x CV with 20 mM Tris-Cl, pH 8.0 and 0.5 M NaCl. Finally, the refolded protein was eluted from the column with 300 mM imidazole, pH 8.0, 20 mM Tris-Cl, and 100 mM NaCl. SDS-PAGE was used to check the purity of eluted Csm3 and the concentration of the eluted Csm3 was determined by the Bradford assay. To confirm that the eluted Csm3 was refolded, size exclusion chromatography was used. Csm3, as monomer, has a molecular weight (MW) of 26.7 kDa, thus the expected elution volume from Superdex S-75 column should be 12.5 ml. The Superdex S-75 column was equilibrated with 20 mM Tris-Cl, pH 8.0 and 300 mM NaCl. The Csm3 was eluted from the column at 13 ml, which is a 0.5 ml delay from its expected eluted volume (Figure 4.1). This may indicates that Csm3 was refolded successfully after being denatured. If the Csm3 was not refolded, it would probably remain aggregated and this have eluted in the void volume in the first 8

ml. The fractions containing Csm3 were pooled together and concentrated by using the Amicon spin concentrator with 3, 000 Da MW cutoff was used. The denaturing and refolding protocol was carried out for both constructs of Csm3 with the His-tag on the N and C termini. Both constructs yielded the same results, however we chose to carry out the crystallization with His-tag on the N-Terminus with no obvious preference.

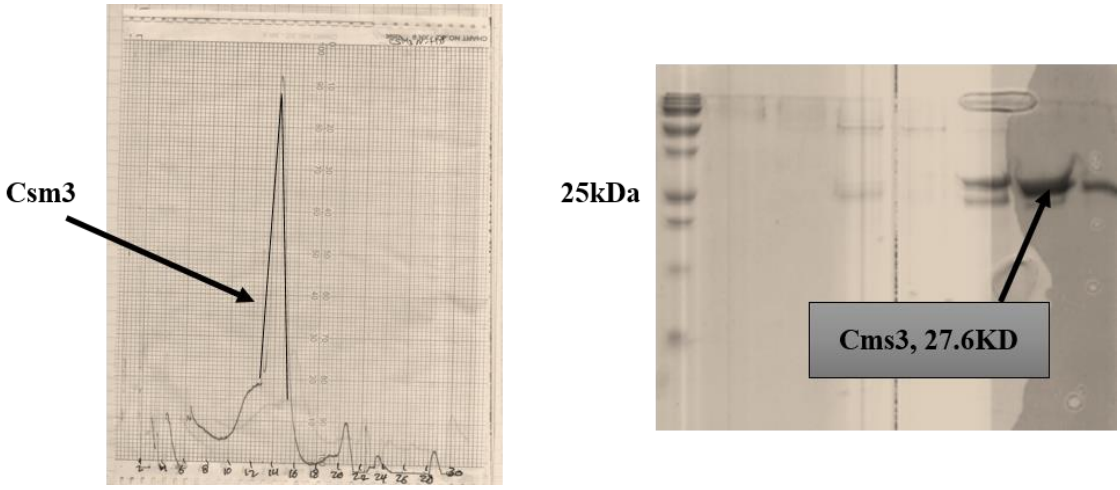


Figure4.1: Chromatogram and SDS-Page of Csm3: A shows the peak corresponding with the elution of Csm3 B: 15% SDS-Page gel of fraction 13 and 14 ml.

Crystallization Trials for Csm3

Csm3 was concentrated to 10 mg/ml for initial crystallization screens. The seven screens from the Hampton Research were set up and the trays were stored at 22℃. The trays were checked after 24 hours, two days, and a week later for the presence of crystals. There were no crystals present in any of the conditions. A second screen trial was tried by increasing the concentration of Csm3 to 15 mg/ml, but no crystals were found. Csm3 is the backbone of the Csm complex and interacts with crRNA, thus having a single

nucleotide might help it to crystalize. A third initial screen trail was carried out by adding 10 mM AMP to 7 mg/ml of Csm3. The lower concentration of Csm3 with AMP was used as it was convenient. The trays were checked for the presence of crystals, but none were found.

Conclusion

Csm3 was denatured and was refolded successfully with β-cyclodextrin. High levels of Csm3 were purified and concentrated upto 15 mg/ml. Although no crystals were produced with the three crystallization trials of initial screening, other biochemical experiments could be done to learn more about Csm3.

Denaturing and Refolding Csm2 on the Column

The inclusion bodies of Csm2 were solubilized as described for Csm3. However, Csm2 did not bind to the Ni-NTA column, which indicates that it's His-tag may not be present. A western blot with Anti-His tag needs to be done, to confirm the presence of His-tag.

EXPRESSING *CSM* GENES IN MYCOBACTERIM

Introduction

This chapter outlines the process of cloning the individual c*sm* genes, and the entire *csm* operon, csm1-csm6 and cas6, in vectors suitable for expression in *Mycobacterium smegmatis*. The PCR design and amplification conditions will be described as well as some preliminary steps that have been taken for cloning.

Cloning

There are two plasmids that could be used to express proteins in Mycobacterium. pST-2k was chosen to clone the entire *csm* operon, which is a dual-expression vector. pST-2K has a P_{myc1}^{tetO} and P_{smyc} promoters and each has multiple cloning sites (MCS) that carry five unique restriction sites (Figure 5-1). The entire *csm* operon without *cas1* and *cas 2* will be cloned under P_{myc1}^{tetO} into the EcoRV and NotI (NEB) restriction sites. These two restriction sites are ideal since their sites are absent in the Csm operon. As a result, the first protein in the operon will have an N-terminus His-tag. The of crRNA will be cloned under control of the Ps_{myc} promoter using PacI and HpaI sites. The pST-2k (purchased from Addgen) has a Kanamycin antibiotic selection and has origin of replication in *E .coli* and Mycobacterium.

The second vector, pYUB1062, will be used to clone individual *csm* genes. It was developed and kindly provided by Dr. William Jacobs, Albert Einstein College of Medicine. This vector has *oriE* for replication in *E. coli*, *oriM* for replication in

mycobacteria, a hygromysin gene as antibiotic selection, a T7 promoter/terminator, and a histidine tag region for purification. The individual Csm proteins will be cloned using NdeI and EcoRV. A Strep-tag was added to the individual Csm proteins at the N-terminus and the His-tag from the vector was eliminated by using the NdeI restriction sites. Table 5.1 gives the list of constructs, primers, and the annealing temperatures used for PCR.

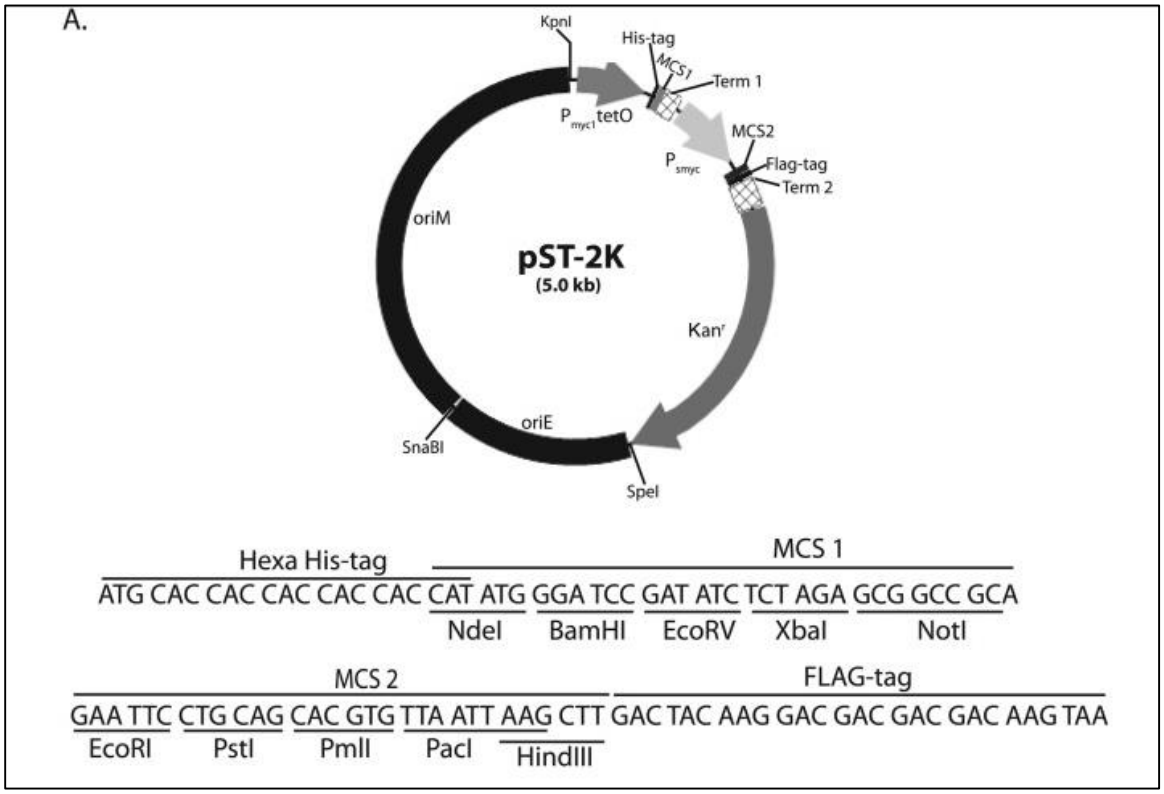


Figure 5-1: Map of pST-2K. A Map of pST-K and sequence of the multiple cloning sites 1 and 2 [61].

PCR Amplification of the Csm Operon

The segment of the Csm operon without c*as1* and c*as2* was amplified by PCR using the H37Rva genomic DNA as a template. The primers were designed to add the EcoRV and NotI (NEB) restriction sites. A three (bp) GC clamp and three AT bases were also added at the 5' end of both forward and reverse primes to promote specific binding or better annealing to the template. The PCR was carried out as recommend by Phusion High-Fidelity DNA Polymerase from the Thermo Fisher Scientific company. The annealing temperatures for all the constructs were calculated using the online calculator provided by Thermo Fisher Scientific and are listed in Table 5-1.
Following the initial denaturing at 95℃ for 30 sec the PCR cycles were run for 35 cycles and had the following steps:

1) Denaturation at 95℃ for 10 sec
2) Primer annealing at the indicated temperature, see Table 5.1
3) Extension by Phusion polymerase at 72℃ for indicated time, see Table 5.1

Following this, there was a final extension step at 72℃ for 10 min
Construct 1A, which consists of the entire operon without *Cas* 1 and *Cas* 2 was tried first.

The default HF buffer was used in PCR and the expected size of construct 1A is 7.706 Kbp. Following the standard protocol failed to amplify segment 1A. Two temperature gradients over range of 40 ℃ to 70 ℃ were then tried to amplify this segment, but both trials failed. Different buffers, using additives such as DMSO and MgCl were tried to amplify construct 1A. Construct 1A was finally amplified by having 2% DMSO as additive, lowering the annealing temperature from 69.2℃ to 65.0℃, and by

using GC buffer. The PCR product was gel-purified. To confirm that the product was the right segment, *csm4*, which is in the middle segment of 1A was amplified. The purified PCR product construct 1A needs to be digested with EcoRV and NotI, and ligated into pST-2K. This is a future step that is needed for this project. Construct 2A is the truncated version of 1A, so the PCR product of 1A could be used as template for 2A PCR reaction after it is sequenced.

PCR Design of Individual Csm Protein Constructs

cas10 (csm1), csm2, csm3, and csm4 were PCR amplified from the genomic H37Rva DNA. Table 5.1 provides the list of the primers that were used with their corresponded annealing temperatures. The restriction sites EcoRV and NdeI were added by primers that were used in the PCR reaction (Figure 5.2). A three base pair GC clamp and three AT bases were added to the primers to promote specific binding to primers as well. The PCR reaction was set up as described above (Page # 38), using the extension time from Table 5.1. The PCR products were then gel-purified (Figure 5.3).

Then 1 μg of purified PCR product (Figure 5.3) and 1 μg of plasmid (pYUB1062) (Figure 5.2) were singly digested with five units of NdeI and five unites of EcoRV. Since NdeI requires a longer sticky end for a better activity, the plasmid and insert were cut first with NdeI. The digestion was carried for an hour at 37 ℃ in Buffer 2 and the reaction was stop by incubating the reaction at 65 C for 20 minutes. The products were cleaned up using QIAquick Purification kit. The purified products were digested for the second time with five units EcoRV in recomened buffer by NEB, Buffer 3. The digestion was carried for an hour at 37 ℃ and was stopped by heat inactivation of the enzyme by

incubating the reaction at 80 ℃ for 20 minutes. The doubly digested plasmid and inserts were gel-purified.

The purified digested PCR products were ligated into the purified digested vector using T4 ligase (Promega). Both of these enzymes were tested with a positive control, to ensure their activity.

The ligation reaction was carried out at room temperature for three hours. The ligated vector was then transformed in DH5α cells and pelleted on LB with Hygromycin antibiotic selection and incubated at 37℃ overnight. The digested vector, without the PCR product, was ligated and transformed in the DH5α cells as a negative control. The next day, plates were checked for the growth of clones and were compared with the negative control. The number of colonies were compared between the two. For example for one of transformation, there were 17 colonies in negative control and 76 in the plate of gene of interest.

Three colonies were picked from the insert plate and colony PCR was done with gene specific primers. The positive clones were grown in 5 ml of LB with hygromycine. The clones were then extracted and purified using the Qiagen mini prep kits. These clones were sent to GenScript for sequencing using the T7 promoter primers.

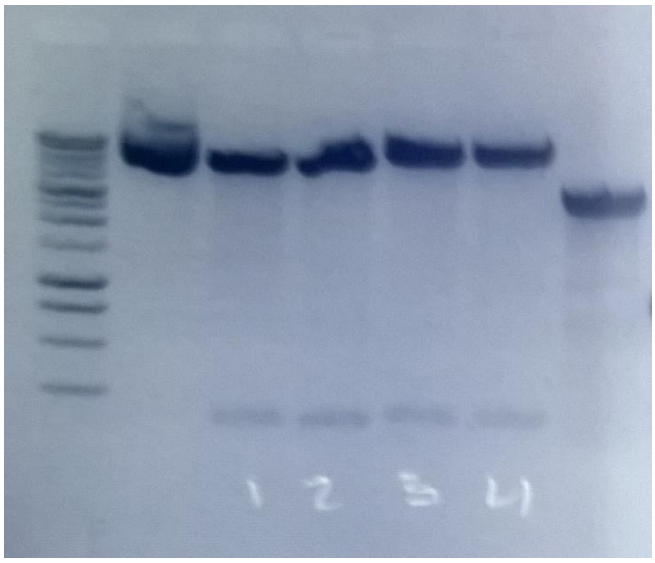


Figure 5.2: The Digest pYUB1062 Vector: 1% agarose gel, Lane 1: DNA Ladder (1Kb). Lane 2: uncut vector. Lanes 3-6: are the double digest vector with NdeI and EcoRV. The faint band in the bottom of Lanes 3-6 could be the cut piece (138bp) from the vector.

The PCR results indicated that the genes of interest were successfully cloned while sequencing results indicated otherwise. Sequencing results showed that the vector was cut and ligated back together by itself and not with the gene of interest, and it was not intact and possible some bases were missing from the MSC. One possibility why the gene of interest was not ligated with the vector might be because the gene of interest did not have a sticky end to ligate with the sticky end of the vector. To ensure that the gene was digested, the digestion of inserts were carried out overnight at 4℃ instead of one hour at 37℃. The digested inserts were gel-purified and ligated with the digested vector and same the results were obtained. Thus further optimization in the digestion and/or the ligation step(s) is needed in order to successfully clone the four c*sm* genes of interest individually

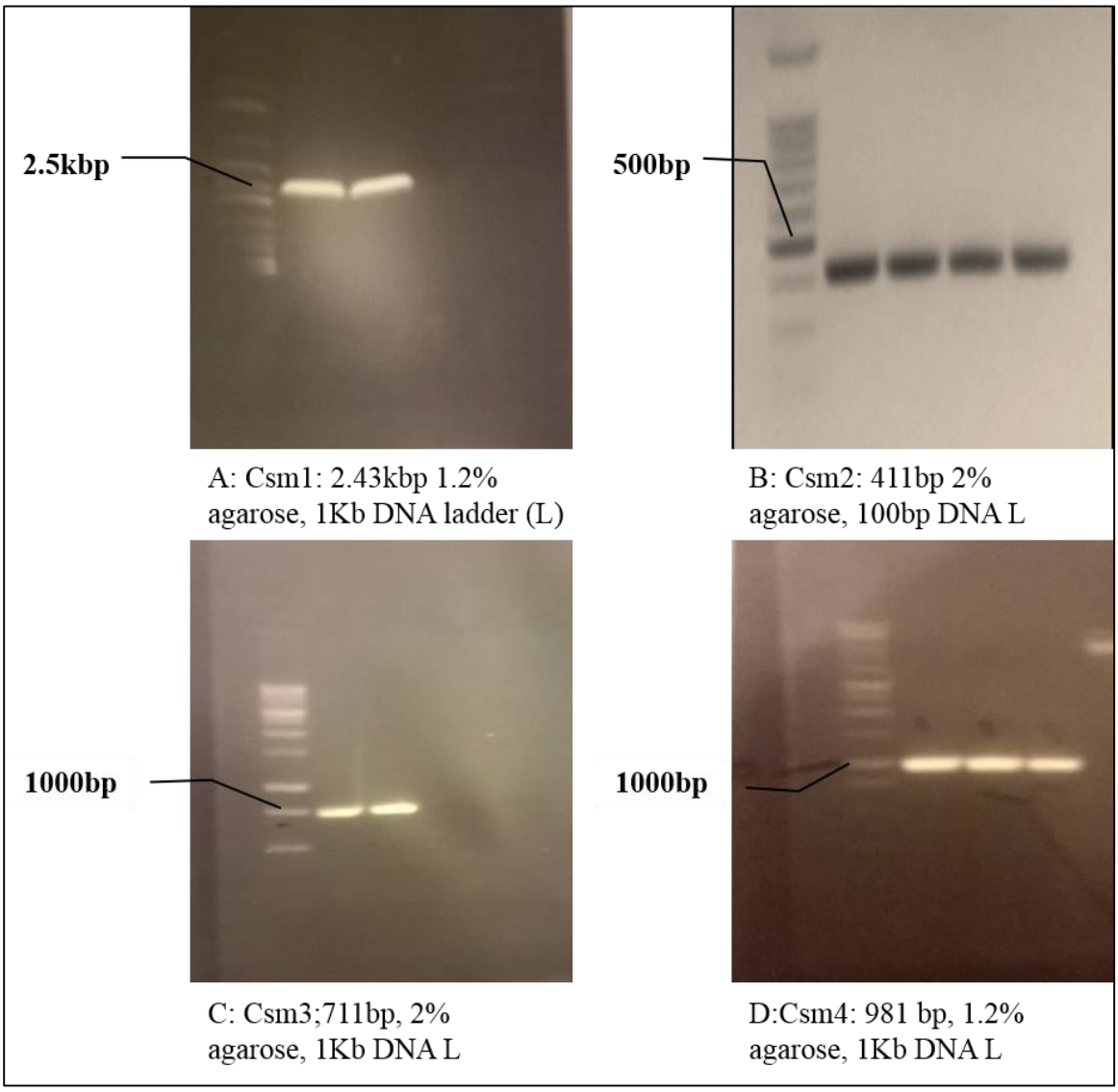


Figure 5.3: PCR product of *csm1, csm2, csm3,* and c*sm4*: Each gene is shown with their corresponding sizes, the percentage of agarose gel and the size of DNA Ladder.

:

Table 5-1: The Csm Constructs, Primers and Annealing Temperatures

Clone Name	Name of Protein	Positon of Strep-tag	Positon of His-tag	Name of Primer	Primer sequence	Annealing temp for PCR ℃	Extension time (sec)
1A	Csm1-Csm6 +Cas6	C-Terminus Of Cas6	N-Terminus Of Csm1	1AF*	GCGTATGATATCATGGCTGC TCGCCGAGG	69.2	120
				1AR*	GCGTAT GCGGCCGC CTA TTT TTC GAA CTG CGG GTG GCT CCA GCC CAG CGG TGC CAT ATC		
2A	Csm1-Csm6	C-Terminus of Csm6	N-Terminus of Csm1	2AF	GCGTATGATATCATGGCTGC TCGCCGAGGC	85.6	120
				2AR	CGC ATA GCGGCCGC CTA TTT TTC GAA CTG CGG GTG GCT CCA TTC GGC TCT CCT GAT CGA		
3B	Csm4	N-Terminus	None	3BF	GCG TAT CATATG TGG AGC CAC CCG CAG TTC GAA AAA ATG AAC TCG CGG CTG TTT	63.9	20
				3BR	GCG TAT GATATC CTA TGC GGC GGA CTC CGG		

Table 5-1 Continued

Clone Name	Name of Protein	Positon of Strep-tag	Positon of His-tag	Name of Primer	Primer sequence	Annealing temp for PCR ℃	Extension time (sec)
5B	Csm2	N-Terminus	None	5BF	GCG TAT CATATG TGG AGC CAC CCG CAG TTC GAA AAA ATG AGC GTC ATC CAA GAC	53.9	15
				5BR	GCG TAT GATATC CTA CTT GTC CTT CGG ATC GAG		
7B	Csm1	N-Terminus	None	7BF	GCG TAT CATATG TGG AGC CAC CCG CAG TTC GAA AAA ATG AAC CCG CAA CTC ATC	61.7	36
				7BR	GCG TAT GATATC CTA TTC GGA CTC CTC CTT GCG		
9B	Csm3	N-Terminus	None	9BF	GCG TAT CATATG TGG AGC CAC CCG CAG TTC GAA AAA ATG ACT ACG AGC TAC GCC	58.5	16
				9BR	GCG TAT GATATC CTA AAC AGC CGC GAG TTC ATG		

Table5.1: Give the list of all the constructs with their corresponding Primers, Annealing Temperature, and Extension time. F* stands for forward and R* stands for Reverse. The colors coordinates with the composition of the primers with Figure 5.3

FUTURE WORK

Csm3 and Csm5

Csm3 makes up the backbone of the Csm complex and it is possible that Csm3 has a direct interaction with Csm5. One biochemical experiment that can be done is to explore how Csm3 and Csm5 bind to each other by using size exclusion chromatography. A 1:1 mole ration of both protein can be loaded on S-75 column, based on the elution from the SEC column in order to explore whether they form a complex or not. Different ratios of Csm3 to Csm5 can be explored to learn more about their stoichiometry with each other.

Cloning Csm Proteins in *M. Smegmatis*

It can be clearly seen from the sequence results that the plasmid was cut twice and it ligated to itself. Figure 6.1 shows that where the expected site to be cut with NdeI and EcoRV. The segments that is predicted to be cleaved off from the vector is cleaved and cannot be seen in the sequencing results (Figure 6.2). As it can be seen from figure 6.2 there is no NdeI and EcoRV sites, but the ribose binding site is there as well as the His-tag. This indicates that the vector is ligated to itself rather than to the gene of interest. The digested vector can be treated with phosphatase, which will remove the 5' phosphate preventing the ligase to ligate the two ends of vector. After treating the digested vector with phosphatase, the ligation with the gene of interest can be carried out according the T4 ligase.

Forward Primer's Composition with Strep tag on N-terminus:
GC clamp, restriction cite, start codon, strep tag, gene specific

Reverse Primer's Composition:
GC clamp, restriction cite, stop codon, gene specific

Figure 6.1: pYUB1062's MSC: The expected restriction that needs to be cut is highlighted in red

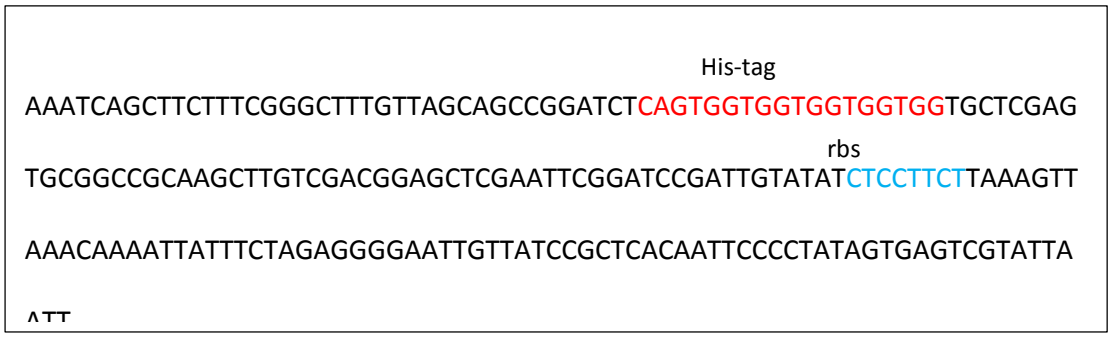


Figure 6.2: Sequencing results: His-tag and ribosome binding sites are indicated

Cloning the Entire Operon

The ligation of the entire operon with the pST-2K plasmid carrying a sequence coding for crRNA needs to be completed. However, once this project gets to the expression phase, a couple of the experiments must be done first before taking it to crystallography. The expression of construct 1A with the expression of crRNA will indicate two important thing about the composition of the complex. Csm6 and Cas6 have not been co-purified for any of the Csm complexes that have been studied so far. Construct 1A has a His-tag on the N-Terminus of Csm6 and a Strep-tag on C-terminus of Cas6. If Csm6 is a part of the complex, then the rest of the proteins will co-purify with

Csm6. If Csm6 is not part of the complex, only the Csm6 will be purified using the Ni-NTA column. Cas6 has a Strep-tag, so if it is part of the complex, the protein in construct 1A will co-purify with Cas6 using the Strep-Tactin column. From purifying the construct 1A, two important conclusions can hopefully be made and that is whether or not Csm6 and/or Cas6 are part of the complex. If Csm6 and Cas6 is not part of the complex then from construct 1A the Csm complex cannot be purified, as result construct 2A will be cloned.

The expression of construct 2A, which is composed of Csm1-Csm6, has the His-tag on the N-terminus of Csm6 and a Strep-tag on Csm1. When purifying construct 2A with Strep-Tactin, Csm5-Csm4 could co-purify as they make up the Csm complex. A Blue Native SDS-PAGE can be used to evaluate the complex. The stoichiometry for the complex denaturing and tandem the Mass Spectrometry can be used. Knowing the MW of the Csm complex and the stoichiometry will be an excellent start toward understanding this complex in *M. tuberculosis*.

REFERENCES CITED

1. Prevention, C.f.D.C.a., *Tuberculosisi, Basic TB Facts.* 2015.

2. Russell, D.G., C.E. Barry, 3rd, and J.L. Flynn, *Tuberculosis: what we don't know can, and does, hurt us.* Science, 2010. **328**(5980): p. 852-6.

3. Smith, I., *Mycobacterium tuberculosis Pathogenesis and Molecular Determinants of Virulence.* Clinical Microbiology Reviews, 2003. **16**(3): p. 463-496.

4. Sulakvelidze, A., Z. Alavidze, and J.G. Morris, Jr., *Bacteriophage therapy.* Antimicrob Agents Chemother, 2001. **45**(3): p. 649-59.

5. He, L., X. Fan, and J. Xie, *Comparative genomic structures of Mycobacterium CRISPR-Cas.* J Cell Biochem, 2012. **113**(7): p. 2464-73.

6. Brouns, S.J.J., et al., *Small CRISPR RNAs guide antiviral defense in prokaryotes.* Science, 2008. **321**(5891): p. 960-964.

7. Hale, C., et al., *Prokaryotic silencing (psi) RNAs in Pyrococcus furiosus.* Rna-a Publication of the Rna Society, 2008. **14**(12): p. 2572-2579.

8. Jore, M.M., et al., *Structural basis for CRISPR RNA-guided DNA recognition by Cascade.* Nature Structural & Molecular Biology, 2011. **18**(5): p. 529-U141.

9. Lawrence, C.M., et al., *Structure and mechanism of the CRISPR associated complex for antiviral defense (CASCADE).* Faseb Journal, 2012. **26**: p. 1.

10. Wiedenheft, B., et al., *Structures of the RNA-guided surveillance complex from a bacterial immune system.* Nature, 2011. **477**(7365): p. 486-U141.

11. Wiedenheft, B., S.H. Sternberg, and J.A. Doudna, *RNA-guided genetic silencing systems in bacteria and archaea.* Nature, 2012. **482**(7385): p. 331-8.

12. Mojica, F.J.M., et al., *Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements.* Journal of Molecular Evolution, 2005. **60**(2): p. 174-182.

13. Kunin, V., R. Sorek, and P. Hugenholtz, *Evolutionary conservation of sequence and secondary structures in CRISPR repeats.* Genome Biology, 2007. **8**(4): p. 7.

14. Grissa, I., G. Vergnaud, and C. Pourcel, *The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats*. 2007, Bmc Bioinformatics.

15. Sorek, R., V. Kunin, and P. Hugenholtz, *CRISPR - a widespread system that provides acquired resistance against phages in bacteria and archaea.* Nature Reviews Microbiology, 2008. **6**(3): p. 181-186.

16. Rath, D., et al., *The CRISPR-Cas immune system: Biology, mechanisms and applications.* Biochimie, 2015. **117**: p. 119-128.

17. Makarova, K.S., et al., *Evolution and classification of the CRISPR-Cas systems.* Nature Reviews Microbiology, 2011. **9**(6): p. 467-477.

18. Makarova, K.S., et al., *An updated evolutionary classification of CRISPR-Cas systems.* Nature reviews. Microbiology, 2015. **13**(11): p. 722-36.

19. Sinkunas, T., G. Gasiunas, and V. Siksnys, *Cas3 nuclease-helicase activity assays.* Methods in molecular biology (Clifton, N.J.), 2015. **1311**: p. 277-91.

20. Makarova, K.S., et al., *Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems.* Biology Direct, 2011. **6**: p. 27.

21. Jinek, M., et al., *A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity.* Science, 2012. **337**(6096): p. 816-821.

22. Charpentier, E., et al., *Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity.* Fems Microbiology Reviews, 2015. **39**(3): p. 428-441.

23. Carter, J. and B. Wiedenheft, *SnapShot: CRISPR-RNA-Guided Adaptive Immune Systems.* Cell, 2015. **163**(1): p. 260-U284.

24. Barrangou, R., et al., *CRISPR provides acquired resistance against viruses in prokaryotes.* Science, 2007. **315**(5819): p. 1709-1712.

25. Mojica, F.J.M., et al., *Short motif sequences determine the targets of the prokaryotic CRISPR defence system.* Microbiology-Sgm, 2009. **155**: p. 733-740.

26. Yosef, I., M.G. Goren, and U. Qimron, *Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli.* Nucleic Acids Research, 2012. **40**(12): p. 5569-5576.

27. Nunez, J.K., et al., *Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity.* Nature Structural & Molecular Biology, 2014. **21**(6): p. 528-534.

28. Reeks, J., J.H. Naismith, and M.F. White, *CRISPR interference: a structural perspective.* Biochemical Journal, 2013. **453**: p. 155-166.

29. Nam, K.H., et al., *Double-stranded Endonuclease Activity in Bacillus halodurans Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-associated Cas2 Protein.* Journal of Biological Chemistry, 2012. **287**(43): p. 35943-35952.

30. Marraffini, L.A., *CRISPR-Cas immunity in prokaryotes.* Nature, 2015. **526**(7571): p. 55-61.

31. Nunez, J.K., et al., *Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity.* Nature, 2015. **519**(7542): p. 193-+.

32. Carte, J., et al., *Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes.* Genes & Development, 2008. **22**(24): p. 3489-3496.

33. Haft, D.H., et al., *A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes.* Plos Computational Biology, 2005. **1**(6): p. 474-483.

34. Marraffini, L.A. and E.J. Sontheimer, *CRISPR Interference Limits Horizontal Gene Transfer in Staphylococci by Targeting DNA.* Science, 2008. **322**(5909): p. 1843-1845.

35. Wiedenheft, B., et al., *RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(25): p. 10092-10097.

36. Semenova, E., et al., *Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(25): p. 10098-10103.

37. Bartel, D.P., *MicroRNAs: Genomic, Biogenesis, Mechanism, and Function.* Cell Press, 2004. **116**: p. 281-291.

38. Parker, J.S., et al., *Enhancement of the Seed-Target Recognition Step in RNA Silencing by a PIWI/MID Domain Protein.* Molecular Cell, 2009. **33**(2): p. 204-214.

39. Sternberg, S.H., R.E. Haurwitz, and J.A. Doudna, *Mechanism of substrate selection by a highly specific CRISPR endoribonuclease.* Rna-a Publication of the Rna Society, 2012. **18**(4): p. 661-672.

40. Gong, B., et al., *Molecular insights into DNA interference by CRISPR-associated nuclease-helicase Cas3.* Proceedings of the National Academy of Sciences of the United States of America, 2014. **111**(46): p. 16359-16364.

41. Huo, Y.W., et al., *Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation.* Nature Structural & Molecular Biology, 2014. **21**(9): p. 771-777.

42. Deltcheva, E., et al., *CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III.* Nature, 2011. **471**(7340): p. 602-+.

43. Rimantas Sapranauskas, G.G., Christophe Fremaux, Roodolphe, Philippe Horvath, and Virginijus Siksys, *The Streptococcus thermophilus CRISPR/Cas system providesimmunity inEscherichia coli.* Nucleic Acids Res., 2011. **39**(21): p. p275-9282.

44. Deng, L., et al., *A novel interference mechanism by a type IIIB CRISPR-Cmr module in Sulfolobus.* Molecular Microbiology, 2013. **87**(5): p. 1088-1099.

45. Staals, R.H.J., et al., *RNA Targeting by the Type III-A CRISPR-Cas Csm Complex of Thermus thermophilus.* Molecular Cell, 2014. **56**(4): p. 518-530.

46. Tamulaitis, G., et al., *Programmable RNA Shredding by the Type III-A CRISPR-Cas System of Streptococcus thermophilus.* Molecular Cell, 2014. **56**(4): p. 506-517.

47. Goldberg, G.W., et al., *Conditional tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting.* Nature, 2014. **514**(7524): p. 633-+.

48. Samai, P., et al., *Co-transcriptional DNA and RNA Cleavage during Type III CRISPR-Cas Immunity.* Cell, 2015. **161**(5): p. 1164-1174.

49. Marraffini, L.A. and E.J. Sontheimer, *Self versus non-self discrimination during CRISPR RNA-directed immunity.* Nature, 2010. **463**(7280): p. 568-U194.

50. Rouillon, C., et al., *Structure of the CRISPR Interference Complex CSM Reveals Key Similarities with Cascade.* Molecular Cell, 2013. **52**(1): p. 124-134.

51. Niewoehner, O., M. Jinek, and J.A. Doudna, *Evolution of CRISPR RNA recognition and processing by Cas6 endonucleases.* Nucleic Acids Research, 2014. **42**(2): p. 1341-1353.

52. van der Oost, J., et al., *Unravelling the structural and mechanistic basis of CRISPR-Cas systems.* Nature Reviews Microbiology, 2014. **12**(7): p. 479-492.

53. Hatoum-Aslan, A., I. Maniv, and L.A. Marraffini, *Mature clustered, regularly interspaced, short palindromic repeats RNA (crRNA) length is measured by a ruler mechanism anchored at the precursor processing site.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(52): p. 21218-21222.

54. Hatoum-Aslan, A., et al., *A Ruler Protein in a Complex for Antiviral Defense Determines the Length of Small Interfering CRISPR RNAs.* Journal of Biological Chemistry, 2013. **288**(39): p. 27888-27897.

55. Cocozaki, A.I., et al., *Structure of the Cmr2 Subunit of the CRISPR-Cas RNA Silencing Complex.* Structure, 2012. **20**(3): p. 545-553.

56. Hatoum-Aslan, A., et al., *Genetic Characterization of Antiplasmid Immunity through a Type III-A CRISPR-Cas System.* Journal of Bacteriology, 2014. **196**(2): p. 310-317.

57. Shah, S.A., et al., *Protospacer recognition motifs: Mixed identities and functional diversity.* Rna Biology, 2013. **10**(5): p. 891-899.

58. Andrecka, J., et al., *Nano positioning system reveals the course of upstream and nontemplate DNA within the RNA polymerase II elongation complex.* Nucleic Acids Research, 2009. **37**(17): p. 5803-5809.

59. Kahn, J.-M.R.a.D., *Mycobacterim smegmatis: an absurd model for terculosis?* TRENDS in Microbiology, 2001. **9**(10): p. 472-73.

60. Natalia Oagesyan, S.-H.K.a.R.K., *<On column Chemical Refolding of Proteisn.pdf>*. PharmaGenomics, 2004.

61. Parikh, A., et al., *Development of a New Generation of Vectors for Gene Expression, Gene Replacement, and Protein-Protein Interaction Studies in Mycobacteria.* Applied and Environmental Microbiology, 2013. **79**(5): p. 1718-1729.