CRISPR-Cas immunity: Analysis of adaptation and interference reactions in prokaryotes

Simon David Ben Cass, Bsc. (Hons)

The University of Nottingham

UNITED KINGDOM · CHINA · MALAYSIA

Thesis submitted to the University of Nottingham for the Degree of Doctor of Philosophy

September 2015

Abstract

Mobile genetic elements (MGEs, e.g. transposons, plasmids and phage) are an important driver of genetic diversity in microorganisms, and have diverse effects on microbe populations. Adaptation of Bacteria and Archaea to overcome negative effects of phage infection is sometimes referred to as an "arms race" that provokes the development of systems to protect against phage attack. One such defence is CRISPR-Cas, the topic of this research thesis. CRISPR (Clustered Regular Interspersed Short Palindromic Repeat) loci and Cas (CRISPR-associated) proteins are the molecular basis of this resistance mechanism. CRISPR-Cas can protect against phage and other foreign MGEs by incorporating a fragment of novel DNA into CRISPR (spacer acquisition) and using this as a template to generate a small RNA molecule, CRISPR RNA (crRNA), which targets the degradation of complementary sequences (interference). Effective interference requires formation of R-loop nucleic acid structure of crRNA base-pairing to homologous DNA, at positions flanked by PAM (Protospacer Adjacent Motif) sequence within the invader.

This thesis investigates actions of CRISPR-Cas interference proteins, with focus on archaeal species *Methanothermobacter thermautotrophicus* (*Mth*) and *Haloferax volcanii* (*Hvo*). *Mth* and *Hvo* catalyse interference by utilizing a Cascade (CRISPR-associated Complex for Antiviral DEfence) protein-crRNA complex. Cas8, the large subunit protein in Cascade, was investigated to explain it's essential role in interference. It is a PAM sensing protein that stabilizes R-loop formation to bring about interference. In addition, this analysis identified a surprising RNase activity of Cas8 that remains of unknown function. The thesis also details recent work on adaptation by Cas1 and Cas2 in *Escherichia coli*. Cas1 nuclease and transesterification activities upon replication fork intermediates are presented alongside a new model for spacer acquisition.

Acknowledgements

I would like to take this opportunity to thank Dr. Edward Bolt for giving me the chance to study for a PhD and for all the help, support and assistance that he has provided over the last four years. I would also like to thank Jamie Howard, Jamie Webster, James Cavey, Steven Werner, Neil Walker and Emily Walker for both the technical assistance and advice and providing a really fun and enjoyable lab experience.

Of particular note is the work carried out by our collaborators making much of this work possible. The research group of Dr. Anita Marchfelder of the University of Ulm, Germany deserve thorogh recognition for their work, specifically Karina Haas and Britta Stoll who carried out some of the experimental work displayed in Chapter 4 where indicated that forms the key genetic data for the first research publication listed in the appendix. The hard work of Dr. Ivana Ivancic-Bace of the University of Zagreb, Croatia also contributed significantly to the second research publication here, again much genetic analysis demonstrated the key targets for analysis in Chapter 5. I am very greatful for the opportunity to work alongside such passionate and exciting researchers.

Friends, family and fellow PhD students, specifically Lucy Dunajova and Matt HardCastle have helped provide support and encouragement throughout my studies. My parents have thankfully been there to support me and allow me to squat while finishing up my thesis writing. The staff and friends of the Johnson Arms have provided a great place to relax and unwind and celebrate as and when appropriate.

Finally, I would like to thank the University of Nottingham and the BBSRC for providing the funding for this research, and making it all possible.

Table of Contents

Figures

Tables

1 Introduction

1.1 General overview of genome dynamics and stability

The dynamic nature of prokaryotic genomes is readily detectable as genetic variation from numerous (6814) genome-sequencing projects, as of September 2015 available from a "Genome Database Search" on the National Centre for Biotechnology Information*,* (search for details on specific genomes by organism name and strain). Examples include acquisition of mobile genetic elements (e.g. horizontal transfer of transposons, plasmids) and variation of conserved gene neighbourhoods. Yet prokaryotes also have numerous systems to maintain genome integrity by eliminating invasive genetic elements or repressing expression (e.g. histone-like nucleoid-structuring protein [H-NS]) (*1, 2*). In addition, there are various prokaryotic DNA repair systems that control mutagenesis arising from endogenous and exogenous genotoxic agents (*1, 2*). Many genome rearrangements result in the appearance of new sequences that have the potential to alter protein expression or function, either by introducing promoters or interrupting regulatory regions of the host chromosome (*3*). The innate restriction-modification (RM) systems and adaptive CRISPR-Cas systems (comprising clustered regularly interspaced short palindromic repeats [CRISPR] and CRISPR-associated [Cas] proteins) constitute prokaryotic defences against invading genetic elements. Genome instability can be random or programmed causing phase (general protein expression) and antigenic variation (cell surface protein alterations) within a population (*4-7*).

1.1.1 Mechanisms of Horizontal Gene Transfer

New sequences often derive from horizontal gene transfer (HGT), the movement of non-parental genetic information between cells (*8*). HGT can become fixed into the new recipient cell if it proves advantageous during selection pressure, but is likely to be lost if its effects are metabolically negative or neutral (*9*). Genetic advantages of HGT can include attenuation of gene expression, alteration of chromosomes by rearrangements, insertions or deletions (*10*) or generation of novel responses to cellular stress to drive niche adaptation (*11*). Below is a brief overview of the types of HGT observed in prokaryotes.

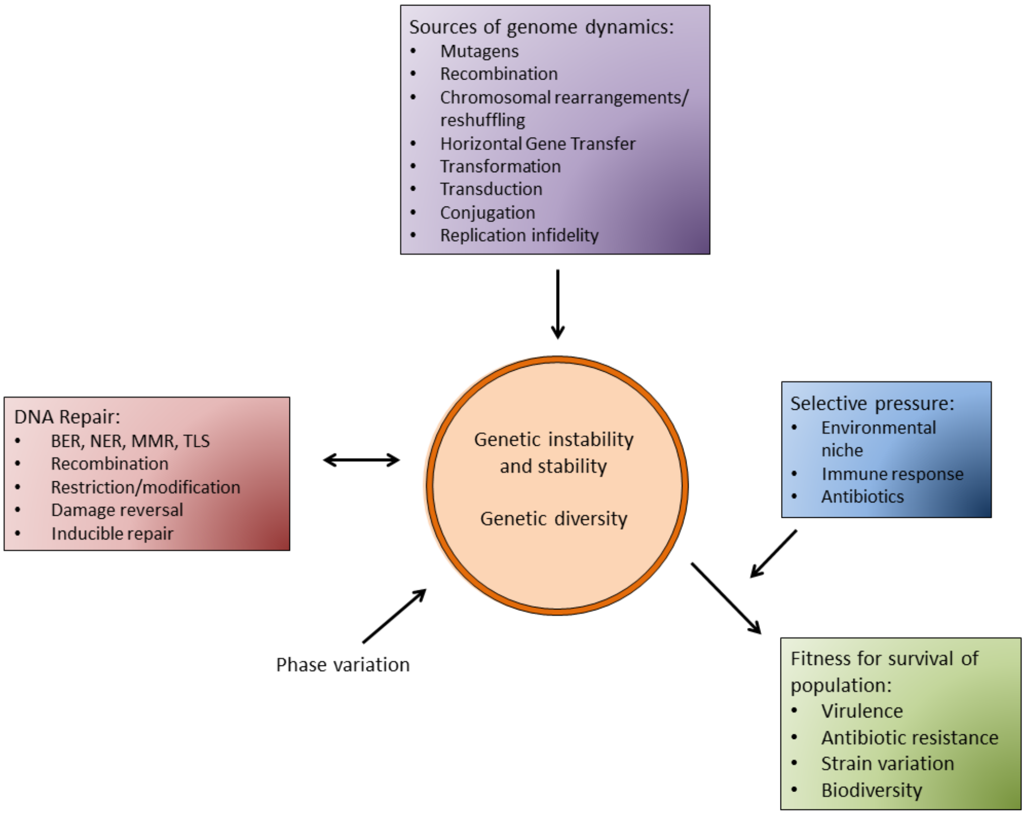


Figure 1-1. Flow of genetic information in prokaryotes. The fitness of prokaryotic populations is determined by the balance of genetic instability, DNA repair and selection pressure. Endogenous and exogenous stresses cause mutations and DNA damage, generating genetic variation that is often harmful. Diversity is directed by recombination and horizontal gene transfer (HGT). Phase variation and antigenic variation are important for cell surface alterations affecting virulence factors and host immunity avoidance. A delicate balance of genetic variation is maintained by DNA-repair pathways. Selective pressures such as environmental factors and antibiotics, influence fitness and survival. BER, base excision repair; MMR, mismatch repair; NER, nucleotide excision repair; TLS, translesion synthesis. Adapted from (*12*).

1.1.1.1 Transformation

Transformation is the direct uptake of free DNA from the environment (*13*). Transformation can be triggered by cellular stress that allows the uptake of extracellular DNA which has adhered to specific cell surface regions (Bayer junctions)

(*14*), and its recombination into the host genome. Analysis of the *Escherichia coli* (*E.coli*) genome determined 17% of genes were acquired by HGT, as calculated by comparing the GC content and codon bias from deep sequencing data (*15*). Transformation was first reported by Frederick Griffith in *Streptococcus pneumoniae* where addition of heat-killed virulent strains of *S. pneumoniae* made harmless strains virulent (1928). Avery, MacLeod and McCarty (1944) identified that this acquired virulence stemmed from transfer of DNA between the two strains. Transformation is now a routine tool for molecular biology employed to introduce a desired genetic element into a cell for further study (*16, 17*).

1.1.1.2 Transduction

Prokaryotic transduction is the process of DNA (viral or prokaryotic) packaging into a viral envelope, release of viral particles from a lysed cell and transfer (by infection) to another cell (*18, 19*). DNA has three fates upon insertion into another cell: absorption and degradation, plasmid recirculation or recombination with a homologous region of the recipient's genome. Imprecise excision of viral DNA from a host genome (prophage) might result in packaging of host genes adjacent to the prophage into new virus particles, thereby potentiating HGT into a new host via transduction. This specialised transduction occurs in lambdoid (λ) phage infection of *E.coli*. Viral particles act indirectly as gene transfer agents (GTA) (*20, 21*). Gene flow from virus to host is overwhelmingly unidirectional; viral genomes rarely incorporate prokaryotic genes. Up to 15-20% of prokaryotic genomes consist of viral or plasmid DNA (*22-24*) and at least 50% of eukaryotic genomes are derived from mobile elements and endogenous viruses (*25, 26*). Mobile elements are usually subject to strong repression although novel genes can be advantageous by exaptation, the utilisation of a gene that once served one function but was subsequently adapted to serve another (*27-29*).

1.1.1.3 Conjugation

Conjugation is the transfer of DNA through direct cell-to-cell ’bridging’ by pili (*30-32*). Conjugative replication transfers a copy of the pili encoding F-plasmid to the recipient that can integrate onto the genome by homologous recombination. Fragments of donor genetic information can be transferred by inaccurate disintegration of the F-plasmid.

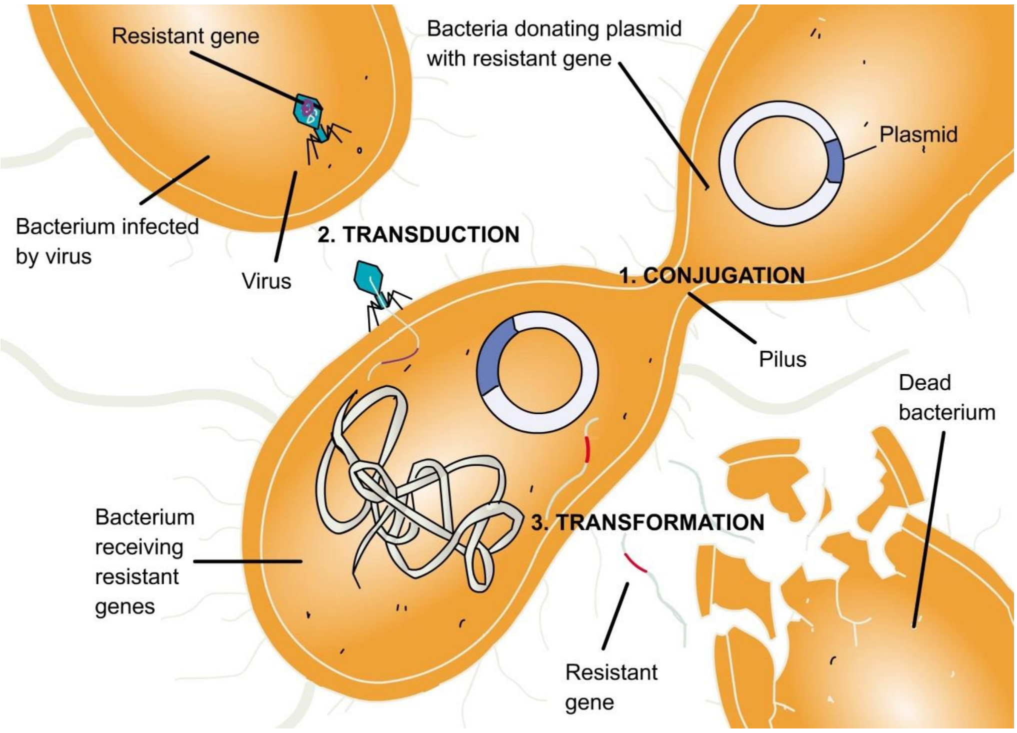

Resistant gene
Bacteria donating plasmid with resistant gene
Plasmid
Bacterium infected by virus
2. TRANSDUCTION
Virus
1. CONJUGATION
Pilus
Dead bacterium
Bacterium receiving resistant genes
3. TRANSFORMATION
Resistant gene

Figure 1-2. The three general mechanisms of gene transfer in prokaryotes. Transformation, transduction and conjugation are the mechanisms of genetic transfer. Antibiotic resistance genes and other selective advantages can be shared by horizontal gene transfer. (Adapted from “The limits of horizontal gene transfer”, Dan Rhodes 2007)

1.1.1.4 Phase and antigenic variation effects of genome instability on prokaryotes

Variation is essential for evolution. Several programmed genetic and epigenetic mechanisms induce specific, adaptive and reversible phenotypic alterations (*4, 6*). Phase variants are important for gene expression variation (typically on/off) (*33*) whereas antigenic variation alters protein structure and function. Phase and antigenic variation create differing subpopulations that can create adaptive advantages (*34*). The mechanisms of phase and antigenic variation are either through DNA replication, repair and recombination dependent events or independently through excision and integration events (*5*).

1.2 Overview of prokaryotic genome defence against HGT

Prokaryotes have evolved several strategies to control or prevent effects of HGT on recipient cell metabolism. The major defence mechanisms are described in the next sections, with emphasis on how nucleic acids are manipulated or enzymatically processed to establish defence against the mobile genetic element. However, phages have also developed methods to avoid prokaryotic immune systems (Table 1).

Table 1. The adaptive mechanisms employed by phage to escape detection upon bacterial infection. Adapted from (*35*).

Phage Escape strategy	Phage	Escape mechanism
Inhibition of phage adsorption		
Adapting to new receptors	Coliphage ϕX174 and T7	Mutations in the RBP-encoding gene lead to adsorption to a modified LPS
Digging for receptors	Coliphage K1F and K1-5	Encoded endosialidase or glycosidase degrades *E.coli* capsule
Stochastic recognition of variable host receptors	Coliphage T4	Duplication of His Box element in tail proteins, shuffling tail specificity to host receptors
Abortive-infection (Abi) system		
Mutation in phage genes	Coliphage T4	Mutation in *gol* prevents activation of Abi Lit
Encoding antitoxin molecule	Coliphage T4	Dmd neutralises RnlA and LsoA toxins during phage replication
Restriction modification (RM) systems		
Fewer restriction sites or unrecognisable orientation	Coliphage T7 and T3	*Eco*RII sites distant on genome preventing REase *Eco*RII cleavage
Modified restriction sites	Coliphage P1	DarA and DarB co-injected and bind restriction sites of phage genome, protecting against type I RM systems.
Mimicry of phage DNA	Coliphage T7	Ocr mimics DNA backbone and sequesters REase
Stimulation of modification enzymes	Coliphage λ	Ral enhances *Eco*KI methyltransferase, rapidly methylating phage DNA, protecting phage DNA from *Eco*KI recognition
Degradation of an R-M cofactor	Coliphage T3	S-adenosyl-L-methionine hydrolase removes S-adensoyl-L-methionine, inhibiting REase
CRISPR-Cas systems		
Mutation in PAM or protospacer	*Streptococcus thermophilus phage*	Mutations or deletions in PAM or protospacer result in CRISPR avoidance.
Antibacterial CRISPR-Cas system	*Vibrio cholera* serogroup 01ICP1 phage	Phage encoded CRISPR system targets phage inducible bacterial CRISPR system

1.2.1 Abortive infections systems: Toxin-antitoxin

Toxin-antitoxin (TA) systems are often auto regulated mobile elements that encode a stable toxin gene and an unstable antitoxin gene (*36-38*). Toxins are always small proteins, but antitoxins can be either protein or RNA. An antitoxin can act in two ways, as a transcriptional repressor or to sequester its toxin counterpart. Loss of the antitoxin part of TA systems by mutation, recombination or segregation is toxic to the host as the stable toxin protein can persistently interfere with host DNA replication and cell surface molecules (*39-41*).

The precise role of TA systems varies. They may contribute to the maintenance of other mobile genetic elements (MGEs), stabilise chromosomal regions susceptible to deletion, become integrated into host cell regulation network, or contribute to pathogenicity and stress induced altruistic cell suicide (*42*). TA systems have been reported as 'junk' DNA or selfish elements only promoting self-survival, but some play critical roles in prokaryotic cell biology. Some antitoxins can be degraded by stress-induced proteases allowing toxin activation to induce abortive infection (cell suicide) (*43*), biofilm formation (*44*), differentiation into persistors (dormant multi-drug resistance strains) or stop infection (*45*).

1.2.2 Restriction-modification systems

Restriction-Modification systems are present in over 90% of all bacterial and archaeal sequenced genomes. They encode a target-specific DNA endonuclease and its cognate modification activity, typically a DNA methyltransferase (*46-48*). These two genes are usually located in the same neighbourhood of the host genome, virus or MGE. Both the nuclease and the modification activity are sequence-specific, unique to each R-M system. Mutations that affect the regulation of expression or targeting of these restriction endonucleases can cause lethal unregulated degradation of host genomes.

Five different function have been associated with the role of R-M systems: (a) defence systems (*49*), (b) selfish genetic elements , (c) stabilisation of genomic islands (*50, 51*), (d) roles in recombination and genome rearrangements controlling speciation and evolution (*47*) and (e) regulation of host chromosome DNA methylation (*52*). Endonucleases identify their cognate DNA modification and degrade elements that do not display it. Therefore inactivated R-M systems allow MGE integration into host as there is no specific degradation. R-M systems are often connected to MGE and stabilize these elements in a population by targeted degradation of competing DNA, acting as a selfish genetic element. Other systems are hypothesised to target host chromosomes to facilitate transposition. Targeting of host chromosomes allows genome rearrangements or methyl-DNA modifications that are linked to epigenetic regulation and diversity.

1.2.3 CRISPR-Cas

Recently identified CRISPR-Cas systems are the only adaptive immune system of prokaryotes. CRISPR-Cas systems are predicted to be mobile selfish genetic elements (defence islands) that confer resistance to invading phage and mobile genetic elements (*53, 54*). CRISPR-Cas systems are discussed in greater detail later.

1.2.4 Transcriptional repression of horizontally acquired genes

Small and abundant histone-like nucleoid structuring (H-NS) proteins are involved in chromosome organisation and gene regulation (*55*). H-NS binds to AT-rich DNA sequences with high affinity and induces sequence specific DNA curvature (*56-58*). Mutation or deletion of H-NS increases cellular transcription levels, indicating H-NS has a gene silencing effect. H-NS are known to silence *virF (59),* a horizontally acquired virulence factor, and the *Cas* genes of the *E.coli* CRISPR-Cas locus. H-NS negatively regulates genes by binding over promoters, occluding RNA polymerase or trapping RNA polymerase in H-NS bridged DNA loops (*60-62*).

1.3 Roles of Homologous recombination in horizontal gene transfer

Homologous recombination (HR) pathways play important roles in meiosis (*63*), repair of DNA strand breaks and in overcoming blocked or broken DNA replication forks (*64-66*). A corollary to the roles of HR in underpinning DNA replication is that it promotes replication of mobile genetic elements such as transposons and phage, which rely on host cell replisome for their propagation. There is also emerging evidence that HR enzymes in bacteria are important for establishment of CRISPR immunity (*67*), and therefore I present a review of relevant aspects of HR in the following sections, and summarised in Figure 1-3.

1.3.1 Homologous recombination

Homologous recombination (HR) descirbes the exchange of nucleotide sequences between either identical (homologous) or near-identical (homeologous) sequences of DNA. HR is initiated by invasion of single stranded DNA into a homologous DNA duplex, termed ‘strand exchange’, generating a ‘D-loop’ intermediate. This reaction is catalysed by RecA-family enzymes that are found in bacteria (RecA) (*68*), Archaea (RadA) (*69*), eukaryotes (Rad51) (*70*) and viruses (Cre) (*71*). D-loop formation can set in motion a number of subsequent DNA processing reactions depending on the context of strand exchange. For example, a D-loop can be extended into a Holliday junction by the actions of heliCases (RuvAB and RecG) (*72-77*) and so activating the classical double-strand break or ‘long tract’ HR pathways. However, these require Holliday junction resolution (RuvC, RusA, Mus81 etc.) (*78-80*) or dissolution (RecQ-TopoI) (*81-84*) to restore duplex DNA after synthetic repair (*85*).

Holliday junctions ("chicken-feet") are proposed to form in HR based repair of stalled replication forks, resulting from encountering a blocking lesion (*86*). RecG, RecA and RuvAB enzymes can catalyse Holliday junction formation. RecG recognises and unwinds a variety of substrates including strand invasion products generated by RecA. RuvAB drives Holliday junction formation at D- and R-loops to allow resolution by RuvC (*85*). Removal of D- and R-loops is important for blocked replication fork progression. Typically, DnaA initiates bacterial replication at a replication origin (*87, 88*). However, in strains deleted for DnaA and RecG, a distinct pathway independent of DnaA, known as stable DNA replication (SDR), follows, whereby PriA triggers chromosomal replication at 3'flaps (*89*). RecG acts as an antagonist to PriA most likely to prevent pathological over-replication by controlling replication restart by SDR (*90-92*). PriA-PriB facilitates DnaB (the replicative heliCase) loading and replisome assembly at branched substrates; however, heliCase defective PriA300 reduces SDR (*93*).

At unstable integrated transposable elements which create gap regions and long flanking DNA stall replication forks. Replication stalling initiates co-ordinated repair at fork-dependent double strand ends by gap filling, concomitant with degradation of the extraneous flanking DNA leaving invading elements intact and stable (*94, 95*). This is thought to occur with other mobile genetic elements including HIV-1 DNA and lambda phage. Lambda phage can integrate into the host genome through lambda P protein recruiting the replicative heliCase DnaB to lambda O initiation site (*96, 97*), acting analogously to *E.coli* DnaC and as a competitive inhibitor of DnaB-DnaC complex formation (*98, 99*). DnaB-DnaC associate to DnaA bound origins, DnaC acts as a loading factor for DnaB.

Current understanding of homologous replication and replication fork restart stems from research carried out in yeast and bacteria. Homologs of RuvABC and RecG are absent from most archaeal species as archaeal DNA and RNA metabolism is more closely related to eukaryotic systems. However, archaea do encode some analogous proteins such as Holliday junction resolvase Hjc (*100, 101*) and branch migration protein Rad54 (*102, 103*).

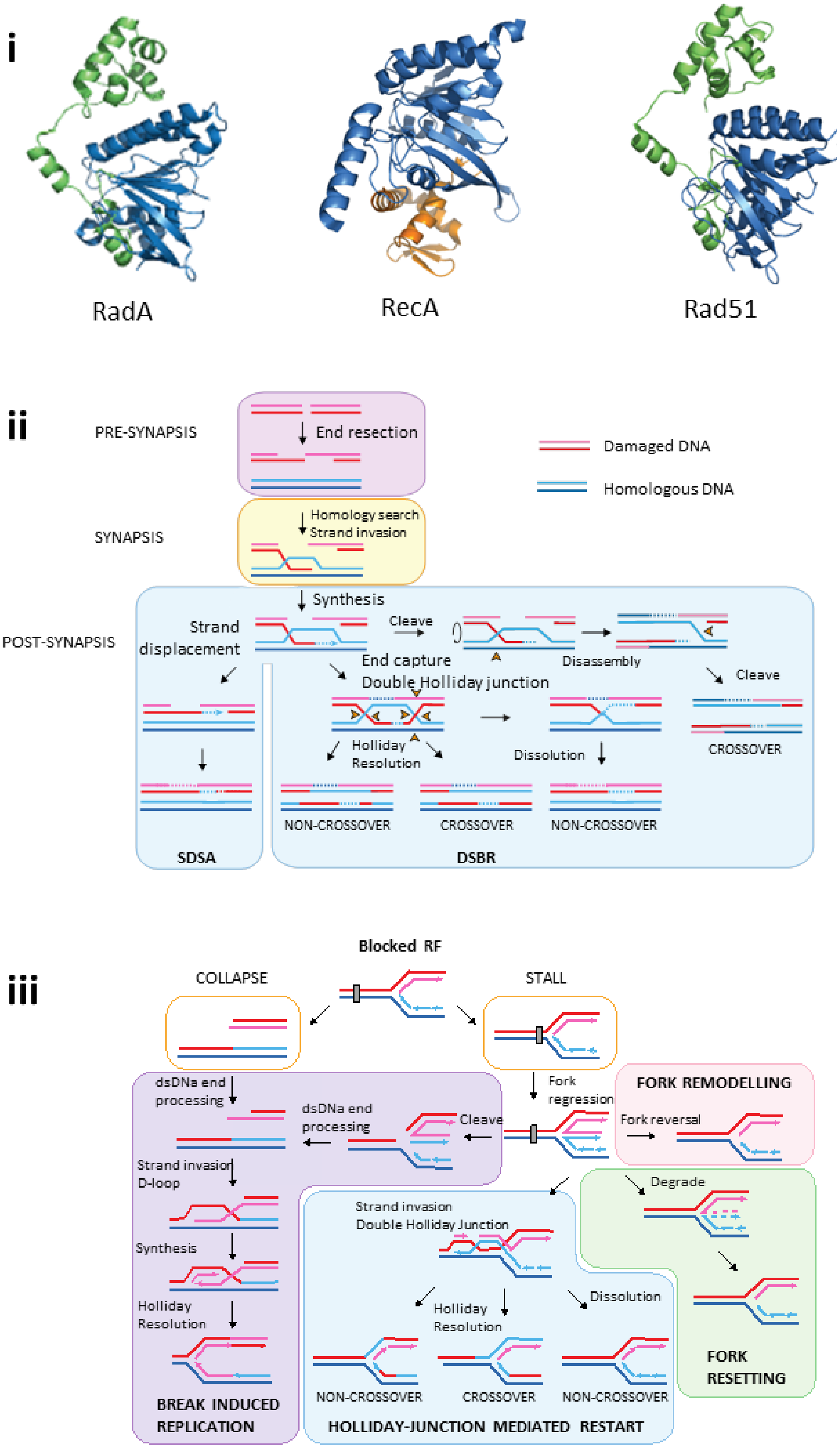


Figure 1-3. An overview of homologous recombination (HR) and its role in DNA repair and blocked replication fork restart. (i) Comparative crystal structures of recombinases from archaea (RadA), bacteria (RecA) and eukaryotes (Rad51) demonstrates archaeal and eukaryotic recombinases are very similar in tertiary structure. (ii) HR-mediated DNA repair is either synthesis dependent stand annealing (SDSA) or double strand break repair (DSBR). DNA damage (here a double strand break [DSB]), is successively repaired through pre-synapsis (end resection), synapsis (strand invasion) and post-synapsis (D-loop or Holliday junction resolution). (iii) During replication, replication forks (RF) encounter blockages that either collapse or stall the fork. RFs can be restarted by the mechanisms summarised.

1.3.2 Recombination at repeated sequences drives genome instability

Homologous or illegitimate recombination at repeat regions in a genome can cause DNA deletions or amplifications. Amplifications usually revert to their original state unless they pose an advantage to the organism, for example, an antibiotic resistance gene amplification improves viability. During recombination subtle junction alterations may alter gene expression or function. Stability of repeats is dependent upon repeat sequence (*104*), distance (*105*), and genomic context (*106*). Palindromic sequences form hairpins in genomes and stimulate strand slippage, promoting illegitimate recombination and repair (micro-homology mediated end joining, MMEJ). Hairpins that form in the lagging strand during replication can be deleted if situated between Okazaki fragments (*107, 108*). Nucleases such as SbcCD cleave hairpins and initiate homologous recombination mediated repair (*109, 110*). If hairpins are flanked by direct repeats incorrect recombination can stimulate genome instability, creating deletions as in single strand annealing. Homologous recombination mediated instability can also occur through rolling-circle replication and non-equal recombination. Non-equal recombination or unequal crossover deletes a sequence in one strand and replaces it with a duplication from the sister chromatid during mitosis or its homologous chromosome in meiosis (*111, 112*).

1.4 CRISPR

1.4.1 Background to the discovery of CRISPR-Cas defence system

A 1987 analysis of the genomic context of the *E. coli* alkaline phosphatase gene (*iap*) noted a region of repetitive DNA that is now referred to as CRISPR (Clustered regularly interspaced short palindromic repeats) (*113*). There is great diversity in CRISPR-Cas loci across bacteria and archaea but some common principles are described in this section and shown in Figure 1-5, which summarises CRISPR loci relevant to this study (*114*). The wide range of sequence variations found in CRISPR loci allowed their application in serotyping methodology in clinical settings. The sequence variability arises from the presence of distinct ‘spacer’ regions of 24-48 base pars (bp) that alternate with the repeat sequences (*113*). Two key observations from bioinformatics have helped to advance our understanding of the biological role of CRISPR loci: First, the identification of spacer sequences as matched sequences from extant mobile genetic elements (*115*)and second, the recognition of a genetic linkage between CRISPR loci and several different open reading frames (ORFs) predicted to encode DNA processing enzymes (*116, 117*). The CRISPR associated (Cas) ORFs suggested a conserved biological function of CRISPR through a diverse range of bacteria and archaea. Further bioinformatics analyses of predicted functions of Cas proteins led to the suggestion that CRISPR-Cas was a DNA repair system or, later, an RNA processing system similar to RNA interference (RNAi) found in many eukaryotes (*118*). In 2007, Barrangou *et al.* demonstrated CRISPR-acquired resistance to phage (*119*). Strains of the yoghurt-producing bacteria *Streptococcus thermophilus* were inoculated with phage isolates from industrial fermenters. Some *S. thermophilus* were phage resistant and sequencing identified acquisition of new DNA matching phage DNA, i.e. novel spacers. Only *S. thermophilus* genomes containing perfectly matched spacers correlated to phage resistance.

At the time of writing the CRISPR database (CRISPR finder and CRISPRI (*120*)) has analysed 2762 prokaryotic genomes identifying 4065 CRISPRs. CRISPR loci are widely distributed in prokaryotes and have now been discovered in 126 (out of 150) archaeal and 1167 (out of 2612) bacterial genomes. Given that similar CRISPR loci have been identified between different organisms, HGT and selective adaptation have most likely taken place (*121*). The origin of CRISPR is not known, but various hypotheses indicate a MGE that is by nature selfish and acts through non-self targeting. Some *Cas* genes (*Cas1* and *Cas2*) have been identified as transposons (*122, 123*) and so could be the origin of CRISPR-Cas immune systems (*124*).

1.4.1.1 CRISPR-Cas loci

CRISPR loci share common architecture: direct repeats (24 to 48 nt) separated by spacers (22 to 33 nt) (*124, 125*). Two organisms that were the subject of this study, the bacterial *E.coli K-12* strain *MG1655* has 29bp long direct repeats and 32bp long spacers and the archaeal *Methanothermobacter thermautotrophicus* (*Mth*) ΔH strain possesses repeats of 2-60bp and spacers of 30bp. Since the spacers provide a historical account of an organism's exposure to invading genetic elements, it is unsurprising the number of spacers varies between organisms. Indeed, *E.coli* has 13 spacers, whereas *Mth* contains 123. Upstream of the CRISPR array is the AT-rich 'leader' (20-543 bp in length) which acts as a promoter for RNA polymerase-catalysed transcription, generating a transcript called pre-CRISPR RNA (pre-crRNA). Leaders form cruciform structures that are proposed to be essential for novel spacer integration. *Cas* genes are located in close proximity to the CRISPR locus, either up or downstream. *Cas* genes are diverse in organisation and the different constituents at each locus give rise to different 'Type' definitions of each CRISPR system.

Since 2007 discovery that CRISPR-Cas systems provide resistance to phage, a great deal of further research lead to uncovering the notable diversity in CRISPR systems, and the various subtypes encoded by different genes (*126*). A summary of the main CRISPR-Cas types (Figure 1-4) demonstrates the universal conservation of *Cas1* and *Cas2* genes and the variation of the other genes which gives rise to different protein complexes and activities. The core 'chassis' of CRISPR interference is similar between types, a protein and RNA complex targets invading nucleic acids for degradation (*127*). Type I systems CRISPR gene clusters encode several Cas proteins that form a ribonucleoprotein complex, Cascade (*128, 129*). Since Type I systems were the focus of this study it is expanded upon in later sections. Type II CRISPR systems vary by encoding only a single protein Cas9, with together with guide RNA (gRNA) forms an analogous complex for DNA targeting (*130, 131*). Given the greater simplicity of the Cas9 Type II system it is currently being exploited as a genome-editing tool in mainstream research (*132-136*). The greatest diversity can be seen in type III systems. Type III-A CRISPR systems are analogous to the type I CRISPR systems, encoding similar proteins which assemble into a so-called Csm complex that targets DNA for degradation (*137*). Type III-B systems are unique, as the Cmr complexes formed can target only RNA for degradation (*138*).

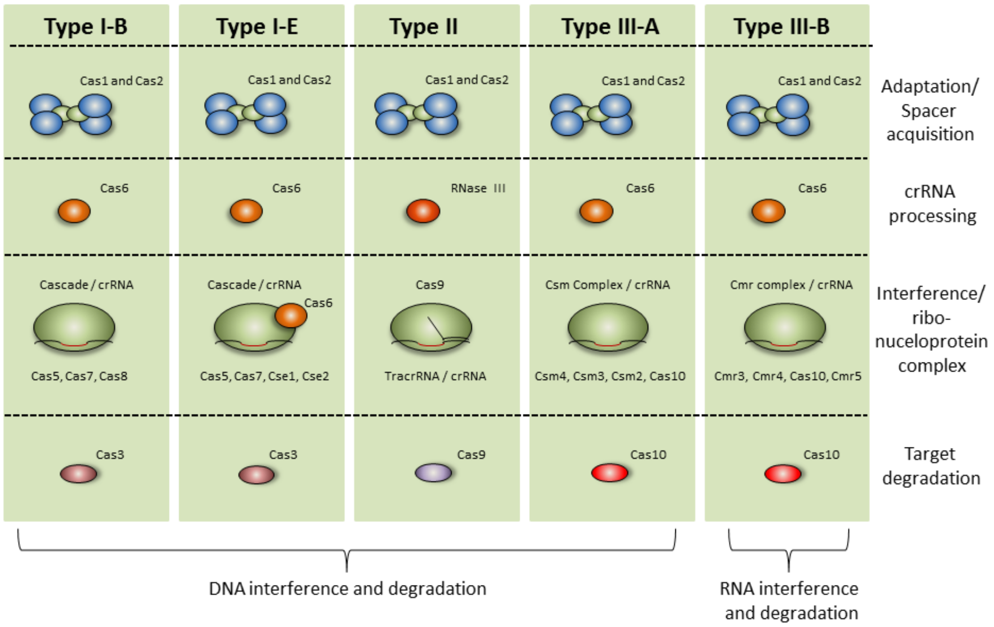


Figure 1-4. A summary of CRISPR adaptation, expression and interference proteins comparing the difference between the major CRISPR types. The universally conserved Cas1 and Cas2 are for spacer acquisition – adaptation. Expression and processing of CRISPR RNA (crRNA) is typically carried out by Cas6 proteins except in Type II system where host RNase III and Cas9 mature the crRNA molecules. Interference complexes have varied protein constituents depending on the subtype, Type I systems encode a Cascade complex with varied subunits, Type III-A encode a Csm complex. Whereas Type I, Type II and Type III-A systems target DNA for interference, the Type III-B type targets RNA. Type II systems differ from all the other CRISPR types in that the formed ribonucleoprotein complex involves only a single protein: Cas9. CRISPR systems are all mechanistically homologous in the overall three step process for resistance to invading genetic elements and only have nuances in the specific catalytic pathways and mechanisms. (Adapted from Van der Oost *et al* 2014) (*139*).

Prokaryotic genomes can contain between 1 and 20 different CRISPR loci with rare duplications such as those seen in *Sulfolobus solfataricus (140)*. Within these loci more than 20 different *Cas* genes have been identified displaying diversity between organisms and CRISPR subtypes. While the different subtypes have been defined by bioinformatics analysis there is however a conserved overall mechanism of CRISPR-mediated immunity. The Type I-B CRISPR-Cas system was recently sub-divided to include types I-G and I-B V_1-2, henceforth referred to as I-H. *Mth* contains two subtypes: a Type I-H (*141, 142*) and a Type III-A subtype. The Type I-H system contains *Cas3, Cas5, Cas6, Cas7, Cas8'* and *Cas8''* encoding genes. The *E.coli* system is a Type I-E subtype, encodes *CasA, CasB, CasC, CasD, CasE* and *Cas3*. The present study explores the two Type I subsystems of *E.coli* and *Mth* (shown in Figure 1-5), with further analysis via comparison with *Haloferax volcanii* Type I-B CRISPR system.

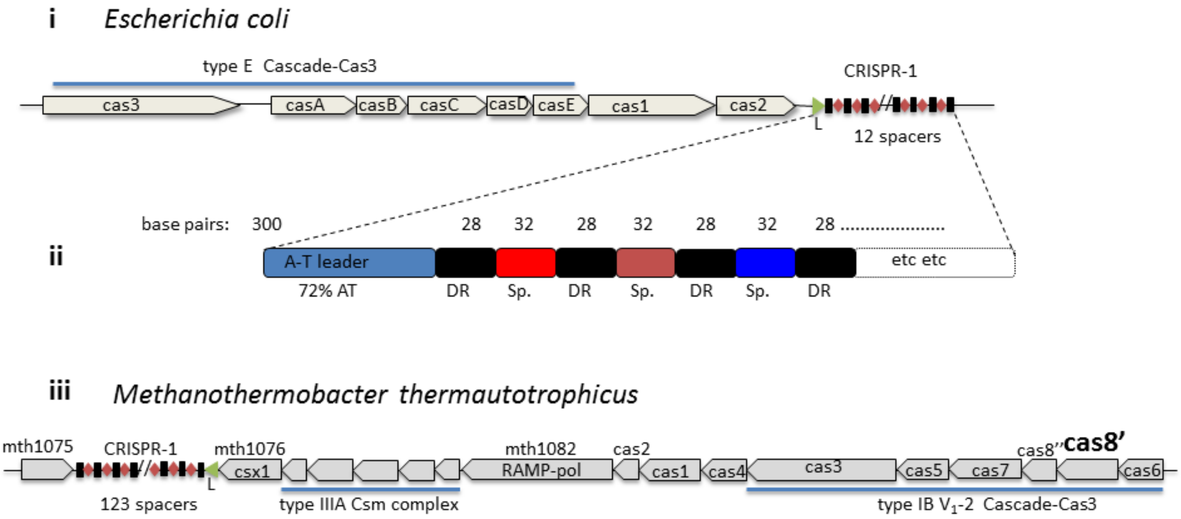


Figure 1-5. Genomic arrangement of CRISPR array and *Cas* genes in *E.coli* and *Methanothermobacter thermautotrophicus*. (i) *E.coli* Type I-E CRISPR system encodes adaptation proteins (Cas1 and Cas2) and interference proteins (Cas3 and CasABCDE) with a CRISPR array downstream (ii). *E.coli* CRISPR array initiates with the AT leader region followed by 28 nt direct repeats (DR) and then 12 different spacers (Sp). (iii) CRISPR locus arrangement in *Mth*, a fusion of Type I-H (Cas5, Cas7, Cas8'', Cas8' and Cas6) and Type III-A (Csm genes) and including the conserved adaptation proteins (Cas1 and Cas2). The *Mth* CRISPR array is laid out as in *E.coli*, only with 123 spacers instead. (Adapted from (*143, 144*)).

1.4.2 CRISPR-Cas immunity in three stages

CRISPR-mediated immunity consists of three stages: Adaptation, processing and interference as summarised in Figure 1-6. This response is triggered by phage adhesion or other stress responses (*145, 146*). Adaptation requires Cas1 and Cas2 and host DNA metabolism factors (*67, 147, 148*). Cas1 and Cas2 proteins capture a small region of foreign genetic information: the process known as target capture. Following capture, DNA is then integrated into the host genome at the CRISPR locus by an as of yet unknown mechanism. Novel spacers are usually acquired proximal to

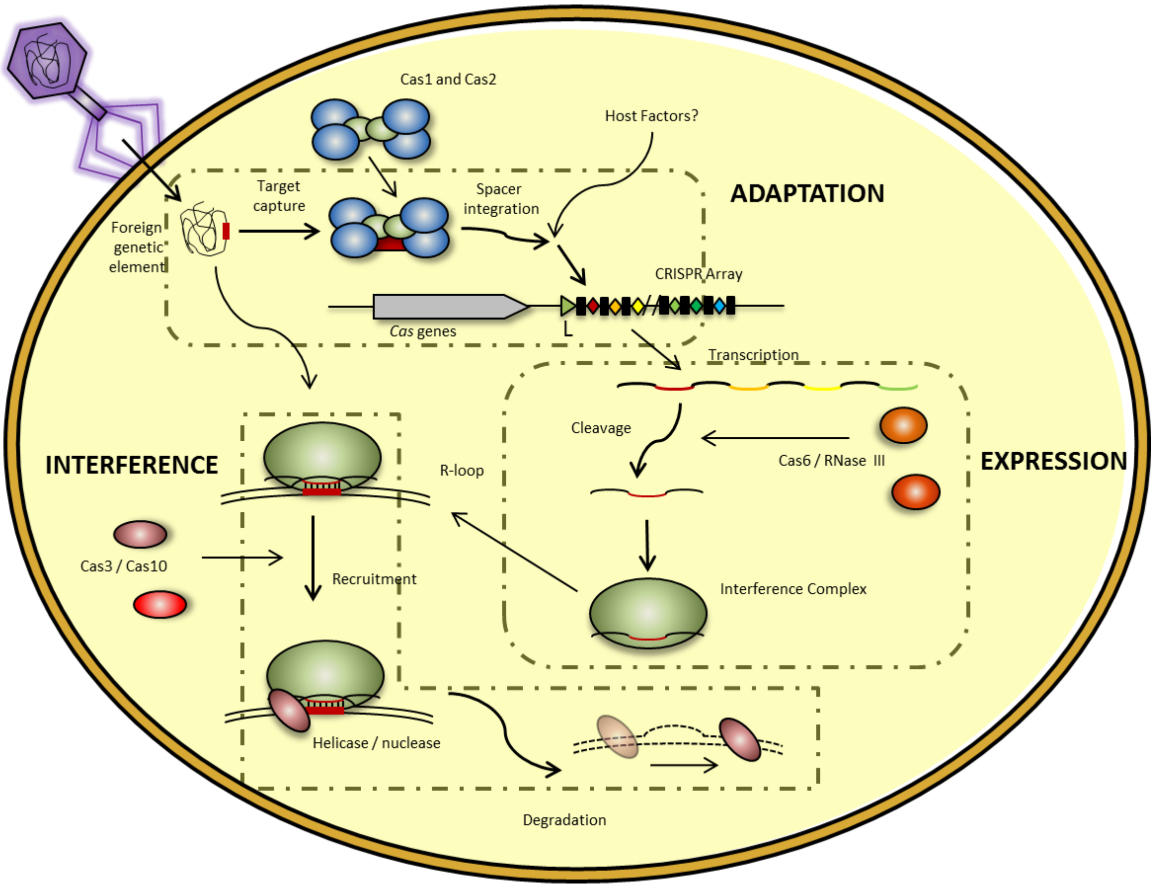


Figure 1-6. Overview of the CRISPR-Cas immune system of prokaryotes. CRISPR-Cas mediated adaptive immunity is split into three stages: (1) Adaptation, the acquisition of a novel short sequence of DNA originating from the invading DNA element – known as a spacer, here represented by viral infection. (2) Expression, the transcription of the CRISPR locus into a long pre-CRISPR RNA (pre-crRNA) and subsequent processing into small mature crRNA molecules that are packaged into the interference complex (consisting of Cas proteins). (3) Interference; target identification by sequence homology between the spacer and the invading element followed by strand invasion forming R-loops and the recruitment of the nuclease component to degrade the invading element.

the leader region (*149-152*). Once spacer acquisition is complete, the CRISPR locus is transcribed to long pre-crRNA (pre-CRISPR RNA). Cas6 or Cas9 and RNases cleave pre-crRNA into mature crRNA (*153-156*). Mature crRNA is assembled into a ribonucleoprotein complex with Cas proteins, generally termed Cascades (CRISPR associated complex for antiviral defence) or interference/effector complexes. The ribonucleoprotein interference complex then targets invader genetic elements in a sequence specific manner and upon correct identification, an R-loop is formed (*157-159*). The interference complex-stabilised R-loop triggers degradation of the invading elements, either directly through Cas9 or through recruitment and activation of Cas3 or Cas10 (*127, 160-168*).

CRISPR and *Cas* gene regulation varies between organisms. In *Salmonella enterica* Serova Typhi and other bacteria the 'master' repressor H-NS represses CRISPR-Cas, in combination with the leucine-responsive regulatory protein (LRP) (*1, 169*). H-NS and LRP respond to environmental factors and affect *Cas* gene repression. LeuO activates both *E.coli* and *Senterica typhi* CRISPR-Cas systems. The two-component regulatory system BaeSR responds to envelope stress in *E.coli* to stimulate *Cas* gene expression (*170, 171*). BaeS senses envelope stress and modulates the phosphorylation state of BaeR. BaeR is a transcription factor which activates the *CasA* promotor, when overexpressed *(146)*. CRISPR gene expression can be sensitive to metabolic change; in energy metabolism cyclic AMP (cAMP) and CRP (adenylate cyclase) form a complex that increase expression of 100 genes in *E.coli*, including *Cas1* and *Cas2-3* genes (*172*). These mechanisms of CRISPR regulation illustrate the general principle of gene expression regulation.

1.4.2.1 Adaptation / Spacer acquisition

The assembly of invader DNA fragments as spacers into CRISPR loci underpins CRISPR immunity, by providing interfering crRNA that can target the invader in a sequence specific manner. Cas1 and Cas2 proteins fulfil an essential role in this process as they catalyse spacer acquisition into CRISPR via an 'adaptation reaction' (*148, 149, 173*). The current model for this integration event is summarised in Figure 1-7. However, evidence is emerging that additional non-Cas host proteins are also required for this reaction. Indeed, in *E.coli* RecBCD processing of double strand breaks at replication forks has been suggested to generate DNA fragments used by Cas1 and Cas2 for spacer integration (*67*). *E.coli* Cas1 and Cas2 overexpression in BL21AI – an *E.coli* strain lacking *Cas* genes but including a CRISPR loci, induced spacer acquisition of sequences derived from the inducible plasmid or host genome (*147*). Part of the adaptation reactions involves Cas1 and Cas2 searching for short target sequences know as PAMs (protospacer adjacent motifs) in the invading nucleic acids (*150, 174*). Subsequently, PAMs direct Cas1 and Cas2 to initiate spacer integration. However, PAMs are required in both adaptation and interference and are thus classified into two sub-sets: SAMs (spacer acquisition motifs) and TIMs (target interference motifs) (*175, 176*). Whereas SAMs constitute the 5’ sequences flanking the novel spacers, TIMs are the 5’ or 3’ sequences proximal to the R-loop formed in the interference step (*157, 163, 177-179*). SAMs are important in the selection of spacer sequences and constitutes the final nucleotides of the repeat elements found at CRISPR loci. On the other hand, TIMs prevent interference complexes targeting self CRISPR loci (*179*).

Adaptation can be stimulated by inefficient interference (‘primed’) (*180, 181*), or can act independently (‘naïve’) (*182, 183*). ‘Primed’ adaptation is more efficient and requires invading elements evading interference through mutation in either PAM or protospacer sequences. This allows interference complexes to bind to but not degrade the invader. Therefore novel spacer acquisition is required to re-establish immunity against the invader. Primed adaptation requires Cas*1*, Cas*2*, *Cas3*, *Cascade* and a spacer that was originally complementary to the protospacer sequence flanked by PAM, as shown genetically. Cas1 and Cas2 mediate naïve adaptation independently of Cascade both *in vivo* and *in vitro*.

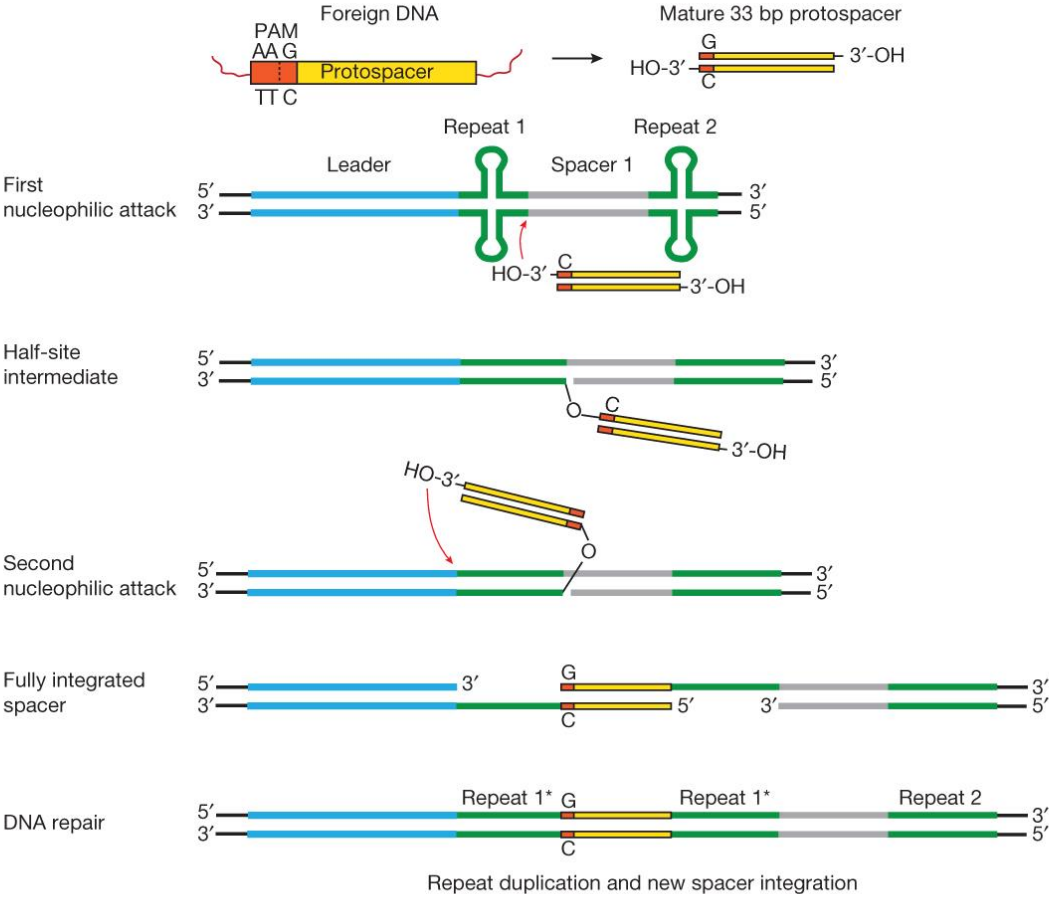


Figure 1-7. The current model of Cas1 mediated spacer acquisition in CRISPR adaptation. Cas1 and Cas2 are involved in the capture of a short DNA molecule from the invading element by a mechanism unknown creating a mature protospacer. The protospacer is then targeted to the CRISPR array by cruciform structures formed by the repeat sections. A series of transesterification reactions at either end of repeat 1 integrates a novel spacer at position 0. DNA repair enzymes then repair the resulting gaps. (Taken from (*184*).

Several hypotheses exist for the action of Cas1 and Cas2 proteins and the mechanism of adaptation. Models require isolation of a short piece of duplex DNA and then an integration event into the host genome at the 0 position of the CRISPR loci (Figure 1-7). The leader and the repeats of the CRISPR loci form cruciform structures, which function either as target recognition sites for the adaptation machinery or as road blocks that stall replication forks (*148, 184*). Through a series of transesterification reactions repeat 1 is split and the new spacer inserted into this region, a process known as spacer integration (SpIN). During preparation of this report the catalytic activity of Cas1 has been reported as transesterification (*150*) hence facilitating disintegration of a short (ds)DNA fragment at replication fork structures. This reflects the difficulties in separating the two process of adaptation, protospacer capture and spacer integration. Finally, DNA duplex gaps are filled, most likely by gap-filling polymerases.

1.4.2.1.1 Cas1

Further analysis identified some *Cas1* genes independent of CRISPR arrays. Specifically, two distinct groups of 'Cas1-solo' were described (*123*). The first group was found exclusively in the *Methanomicrobiales* and the second group was shown in Euryarchaeota and Thaumarchaeota clades. The latter has a patchy distribution, suggestive of horizontal gene transfer (as with transposases). The gene neighbourhood of Cas1-solos in Thaumarcheota contains PolBs, specific DNA polymerases that are protein-primed and are encoded by viruses and self-synthesising transposons in eukaryotes. Cas1-solo elements act in a manner akin to DNA transposons, genomic islands containing TIRs (tandem inverted repeats). No known transposases or recombinases are encoded within this island suggesting Cas1-solo genes are novel transposases, that have been designated Casposons (*122*).

Despite substantial amino acid sequence diversity in bacterial and archaeal Cas1 proteins, the available crystal structures all conform to the same core dimeric overall shape, adorned with additional N-terminal domain (NTD) or C-terminal domain (CTD) regions. In each Case, the NTD of the two monomers constitute the dimer interface. There is a conserved hinge like region between these domains, as shown in *Pseudomonas aeruginosa* PA14 Cas1, a DNA specific endonuclease (*185*). Analysis of Cas1 catalytic activity shows nuclease and transesterase activities upon forked structures. *E.coli* Cas1 (*YgbT* gene) has been characterised with cleavage activity on Holliday junction, flapped substrate, ssRNA and dsDNA in a divalent metal cation dependent manner (*118, 185*). This DNA and RNA nuclease activity is also observed in

Archaeoglobus fulgidus Cas1 (*186*). Despite disputed catalytic activities, *in vivo s*pacer acquisition is dependent on catalytically active Cas1 protein. There is no published information of the *Mth* Cas1 protein function.

1.4.2.1.2 Cas1 homology to Integrase enzymes

Strikingly, *Cas1* genes (Y*gbT* in *E.coli*) have sequence homology to HIV-1 integrase (*184*). Integrases (IN) have two biochemical activities: 3' DNA end processing and strand transfer (*187*). The two reactions have been separated by *in vitro* analysis (*188*). Synthetic 3' end pre-processed DNA molecules were provided as target DNA and integration was observed. Recently, analysis of the IN enzyme of Prototype Foamy Virus (PFV) has revealed mechanistic detail of 3' end processing and strand transfer reactions (*189*). 3' ends are joined with host DNA by two proposed mechanisms. A host phosphodiester bond is broken and the resulting energy is stored in an intermediate between DNA and the integrase with the strand transfer reaction actually requiring only little energy. Alternatively, an isoenergetic transesterification reaction concomitantly directs the new bond formation in a single step. The remaining 5' flaps from this integration reaction are removed by host DNA repair enzymes and the ensuing gap ligated. From the hypothetical models presented, it appears the activities of IN are likely to be very similar to the CRISPR spacer acquisition reaction (*190*).

1.4.2.1.3 Cas2

The crystallisation of Cas2 proteins gave important insights into their role during spacer integration. Identical structures of *Streptococcus pyogenes* Cas2 were deduced at different pHs (5.6-7.5) demonstrating pH dependent DNA duplex nuclease activity (*191*). DNA duplex nuclease activity is also observed in *Bacillus halodurans* Cas2 (*192*), in contradiction to previous findings that *Sulfolobus* *solfataricus* (*Sso*) Cas2 is an endoribonuclease with preference for single-stranded (ss) RNA (*116*). Nonetheless, catalytically inactive mutants of *E.coli* Cas2 in *E.coli* show no loss of function in *in vivo* spacer acquisition which suggests Cas2 has a structural rather than a catalytic role in the process (*148, 184*).

1.4.2.1.4 Cas1 and Cas2

E.coli Cas1 and Cas2 have been shown to interact *in vivo* and can be seen to co-purify in pull-down assays and carry out spacer acquisition. ITC and AUC results from *E.coli* proposed it was dimers of Cas1 and Cas2 which interact to form a heterotetramer. However, *in vitro* reconstitution and crystallisation established Cas1 and Cas2 in fact organise into a hexameric arrangement, where a Cas2 dimer is sandwiched between two Cas1 dimers (Figure 1-8) (*148*). Interference with these interaction surfaces exhibits a loss of *in vivo* spacer acquisition. Therefore, interaction between Cas1 and Cas2 is essential for adaptation in *E.coli*. Sequence comparison revealed conserved catalytic residues of Cas1 proteins in bacteria: E141, H208, D218, D221 and K224. Mutation of any of these residues abolishes spacer acquisition *in vivo* (*148*). Whereas, in Cas2 only mutations that disrupt the Cas1:Cas2 interaction surfaces have a significant effect on acquisition.

Complex formation of Cas1 and Cas2 dictates DNA binding specificity. Co-expression of Cas2 with Cas1 indicates preferential binding to DNA containing the leader and the first repeat of the *E.coli* CRISPR locus rather than random DNA sequences. This corresponds with Cas1 and Cas2 targeting the cruciform structure of the leader and repeat sequences.

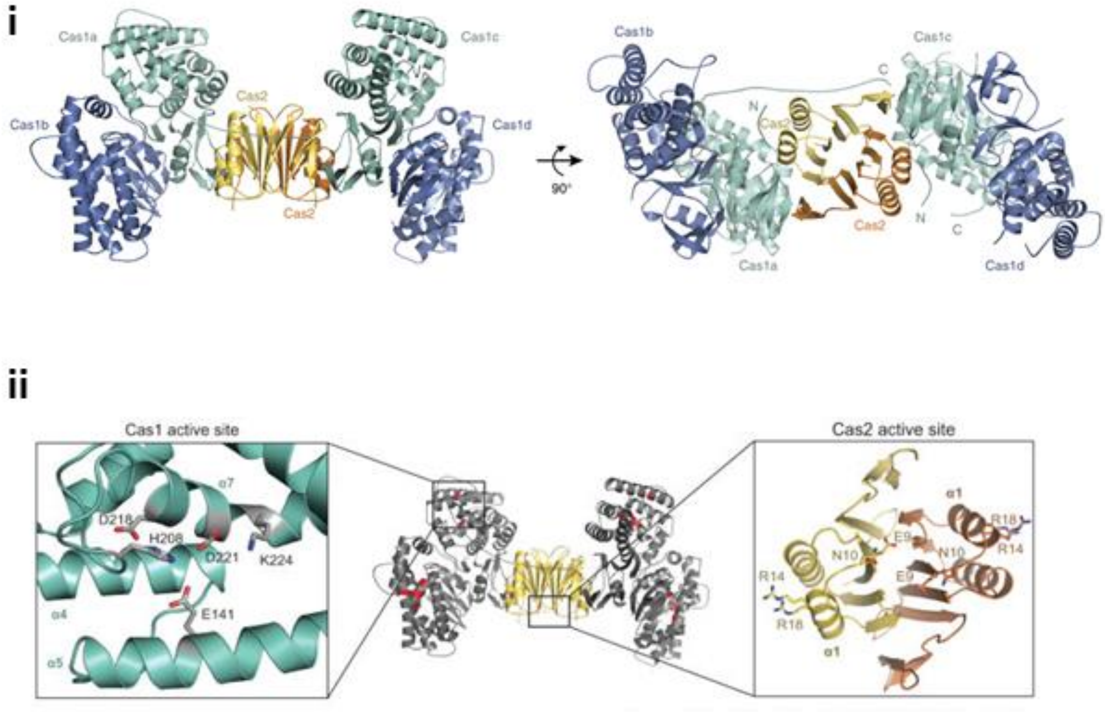


Figure 1-8. Co-crystal structure of Cas1 and Cas2 heterohexamer. (i) Overall structure of the adaptation machinery of the CRISPR immune system, consisting of a Cas1-Cas2 complex where a Cas2 dimer (yellow and orange) forms a linking bridge between two Cas1 dimers (blue and teal). (ii) Detailed views of both the Cas1 and Cas2 active sites with the conserved residues highlighted. (Adapted (*148*)).

1.4.2.1.5 Other factors involved in spacer acquisition

Analysis of the *Cas4* gene from *Sso* revealed 5' to 3' exonuclease activity (*193, 194*). Cas4 end resects DNA creating 3' flaps suggestive of a recombinogenic precursor important for spacer integration, similar to IN proteins. This activity was observed on single strands of duplex DNA as well as single stranded DNA. However, the *Cas4* gene is not well conserved between CRISPR systems and can be found only in bacterial strains lacking *recB.* For instance, it is reported Cas4 forms functional complexes with Cas1 and Cas2 in the archaeon *Thermoproteus tenax (195)*. In *Enterococcus faecalis* and *Streptococcus agalactiae* Csn2 proteins have been linked to CRISPR adaptation, carrying out a similar role to Cas4 (*196*).

1.4.2.2 Expression and processing of crRNA

1.4.2.2.1 Cas6 and Cas9/RNase III

Transcription of CRISPR generates a single mRNA of varied lengths called pre-crRNA. Repeat-spacer units within pre-crRNA generate RNA secondary structure that may be identified by nucleolytic processing enzymes. In type I and III CRISPR systems Cas6 carries out the maturation of pre-crRNA to crRNA (*153, 154, 197-201*). Cas6 proteins of *Sso* have been characterised. In *Sso* there are 6 CRISPR loci and 4 Cas6 genes, two of which have been studied and characterized at the protein level: Sso1437 and Sso2004 (*197*). Both of these enzymes form dimeric structures that cleave pre-crRNA hairpin structures into the mature crRNA product 8 nucleotides from the end of a repeat section (Figure 1-9). However, other Cas6 proteins display an alternate mechanism. PfuCas6 (*Pyrococcus furiosus*) has two RAMP domains that wrap the RNA molecule around the Cas6 protein itself, Which in this manner acts as a molecular ruler to generate the mature crRNA (*198*), schematically shown in Figure 1-9. Cas6 proteins have significantly higher affinities for their cleaved RNA products, indicative of a single turnover event ensuring correct crRNA processing. The three possible fates of Cas6 after completing crRNA maturation are: (i) the release of crRNA, or shifting either (ii) up or (iii) downstream of the cleavage point creating an anchor for effector complex assembly. In the Type I-E systems of *E.coli* and *Thermus thermophilus* (Tt) Cas6 (Cas6e/CasE) remains bound to the mature crRNA and forms part of the Cascade complex. On the other hand, Cas6 family proteins of archaeal Csm and Cmr complexes release the mature crRNA for interference complex assembly without these Cas6 proteins (*202-204*).

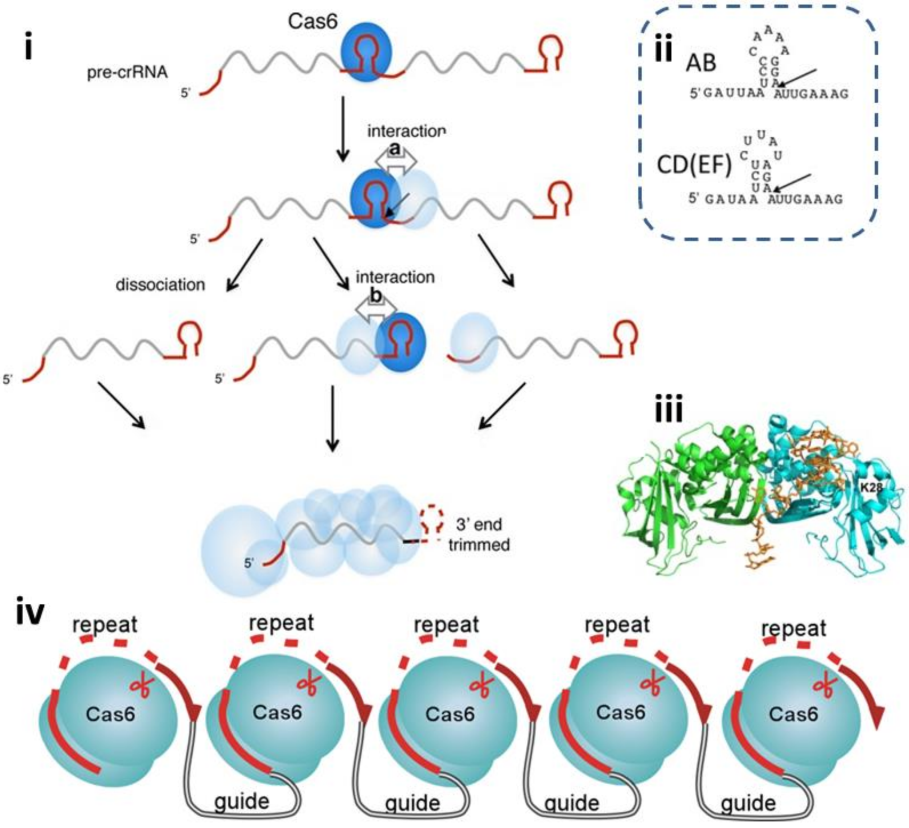


Figure 1-9. Cas6 proteins mature CRISPR-RNA (crRNA) in two different ways. Some Cas6 proteins identify the RNA stem loop structure in pre-crRNA long transcripts (i). This allows targeted cleavage, which has been mapped in *Sulfolobus solfataricus* (ii) and also shown in crystallised Cas6 to map the active site (iii). Other Cas6 proteins perform a molecular ruler type mechanism to cleave pre-crRNA in mature crRNA. (iv). Each Cas6 can then follow several fates depending on the subtype and on the interference complex that follows (i) either by disassociating from the crRNA, or sliding up or down the crRNA molecule and forming part of the CRIPSR interference complex. (Adapted from (*199*) and (*154*)).

The catalytic site of Cas6 proteins shows variability. There is no conserved active site and the mechanism of cleavage site identification also varies. Interestingly, Cas6 generates crRNA molecules with a 3' –OH and 5' cyclic phosphate (*205*). This specific processing may be important for interference complex assembly or identification of crRNAs for complex integration. RNA that was extracted and sequenced from the active *Methanopyrus kandleri* CRISPR system demonstrated crRNA sequences proximal to the leader where the highest transcribed (*206*). A second leader-like AT rich region was found distal from the CRISPR locus that showed a second spike in transcription levels. This suggests there is little regulation in completing transcription of the entire CRISPR array.

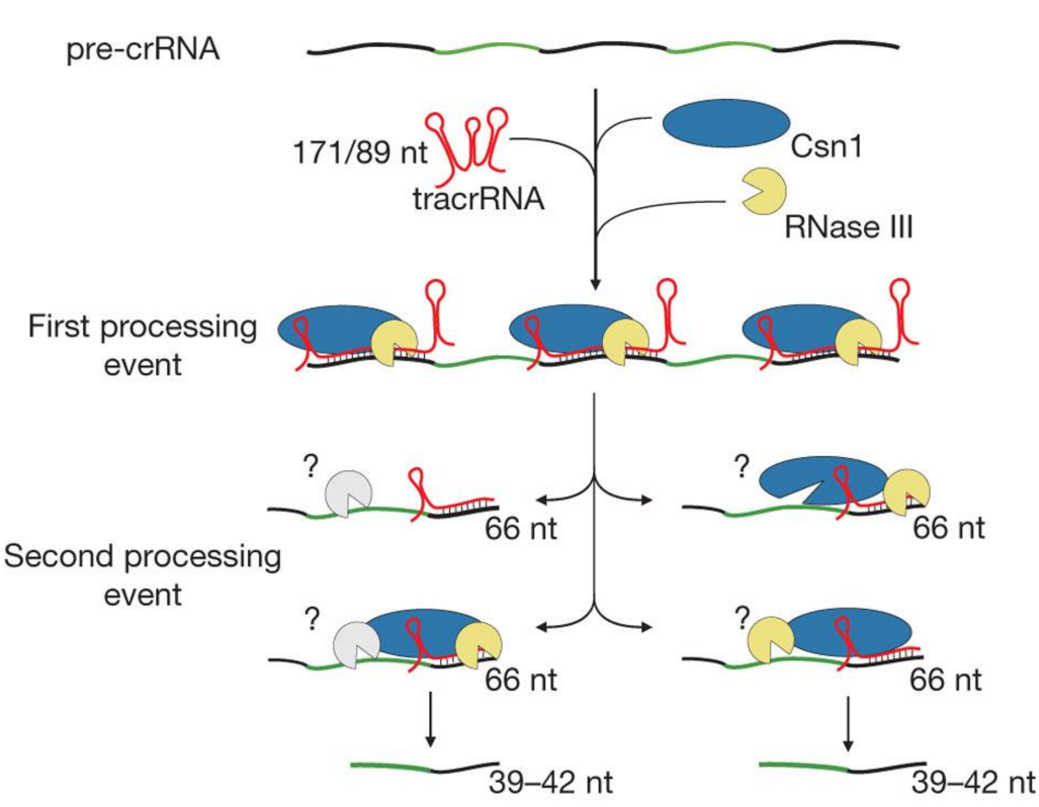


Figure 1-10. Type II CRISPR type maturation of CRISPR-RNA (crRNA) by Csn1 and host RNase factors. Type II systems differ from other CRISPR types as the recruit host RNase III for maturation of crRNA. Type II systems involve two RNA molecules for structural assembly: trans activating crRNA (tracrRNA) and guide RNA (gRNA), similar to crRNA. TracrRNA allows the first processing event to target the pre-crRNA transcript at repeat regions. (Taken from (*207*)).

Type II systems without Cas6 proteins utilise host RNases to prepare crRNA (*155*). In *Streptococcus pyogenes* an extra RNA molecule is required for the processing of crRNA (Figure 1-10). Differential RNA sequencing (dRNA-seq) detected an RNA species that is transcribed 210 nt upstream on the anti-sense strand of the coding sequence for the *Cas* genes. Transcripts were termed tracrRNA (*trans*-activating CRISPR RNA). Within tracrRNA a 25nt stretch has almost complete (one mismatch) complementation to the CRISPR repeats of *S. pyogenes*. Deletion of tracrRNA inhibits the generation of mature crRNA; tracrRNA is therefore essential in Type II crRNA maturation.

The importance of host RNases was identified through RNA analysis of Cas9 CRISPR (Type II) systems revealing the presence of 3' overhangs reminiscent of products of RNase III cleavage (*155, 168, 208*). In some systems, *RNaseIII* or *Csn1* deletions impeded mature crRNA formation (*207*). Csn1 is proposed to act as a molecular anchor protecting pre-crRNA and tracrRNA and carrying out a second cleavage event through a RuvC-like fold (RNase H)(*209*). Host RNase III and anti-CRISPR transcripts are conserved between CRISPR Type II systems. Evolutionarily, CRISPR Cas9 systems may be a precursor to eukaryotic Dicer and Drosha nucleases and the production of small interfering RNAs (siRNA) and micro RNAs (miRNAs) important for gene regulation (*210-212*).

1.4.2.3 Interference - degradation of invading genetic elements

Mature crRNA is incorporated into interference complexes that target invader genetic elements. Atomic resolution structures have shed light on the existence of four interference pathways.

1.4.2.3.1 Cascades and multi-protein interference complexes

The structure of the *E.coli* Cascade complex was determined both with crRNA alone and with crRNA/protospacer RNA duplex. The core of the interference complex is a backbone constructed of six Cas7 proteins (*213, 214*). These Cas7 monomers form a hexameric helical structure with a binding groove that crRNA resides in. In the archaeal type I system from *Sso* homologs of Cas5 and Cas7 show a similar helical arrangement (*213*). Interestingly, *Sso* helix length was dependent on the RNA molecule length and other Cascade proteins present. Other variants of Cascades include *Pseudomonas aeruginosa* in which six Csy3 organise with a similar helical pitch and RNA topology (*215*). Structural similarities suggest that Csy3 and other proteins in the type I systems such as Csc2 in fact all belonging to a Cas7 superfamily (*216*). Therefore, it is Cas7 family proteins which invariably form the core of Cascade complexes across different organisms. Moreover, Cas7 proteins are proposed to perform the same structural role in both the *Mth* I-H and *Hvo* I-B systems as well.

As previously discussed in *Expression and processing of crRNA*, Cas6 and Cas9 coupled to tracrRNA and RNase III execute crRNA maturation in Types I and III and Type II, respectively. Variants of Cas6 (and some Cas5s) contain a typical RNase ferredoxin-like fold. Whereas, Cas5c of *Bacillus halodurans* is thought to process pre-crRNA, thus presenting with catalytic activity, it is not the Case for other Cas5 proteins (*217, 218*). Generally Cas5 has a structural crRNA capping role. In *E.coli* Cas5 (CasD) interacts stably with Cas7 and Cse1 (CasA) but additionally interacts with the Cas7 superfamily protein Cse2 (CasB) (*214*). This series of interactions combined with PAM/TIM interactions are important for R-loop stabilisation. CRISPR systems that lack a Cas6 contain a catalytically active Cas5 (Cas5c) instead, combined with 'regular' Cas5 for its structural role. There are only a few examples where no Cas5 gene can be found, specifically CRISPR subtypes I-D and I-F (*125*). However it is predicted that Csc1 and Csy2 from the respective subtypes belong to the Cas5 protien family. Nonetheless, as shown by SAXS analysis, these proteins don't interact with Cas7 (*214, 219*).

The 'small' subunits of some interference complexes and are predicted to belong to the Cas11 superfamily. Csa5 (I-A), Cse2 (I-E), Csm2 (III-A) and Cmr5 (III-B) are all homologous in their N-terminal domains but diverge in their C-terminal domains (*216*). The *E.coli* Cse2 (CasB) protein dimer is an integral part of the *E.coli* Cascade (*220, 221*), stabilising the formed R-loop and increasing overall affinity of Cascade for dsDNA by an order of magnitude. The 'small' subunit of the Cmr complex Cmr5, is not essential for interference in the Type III-B system (*222*), unlike its counterparts CasB and Cas5 from other CRISPR systems. These results indicate that the 'small' subunits of interference complexes maintain structural similarities but lack close functional links.

1.4.2.3.2 ‘Large’ subunits of interference complexes (Cascades)

Interference complexes of Type I and III systems contain a ‘large’ subunit: Cas8 (I-A, I-B, I-C), Cse1 (I-E), Csy1 (I-F) and Cas10 (I-D, III-A and III-B). Originally, Cas10 proteins were believed to be novel polymerases based on sequence homology to other polymerases and cyclase palm domains. No homology is observed between Cas10 proteins and those of other Type I ‘large’ subunits (e.g. Cas8) (*223*). Cas10 proteins contain a phosphohydrolase domain, similar to that found in Cas3 proteins described later. Some ‘large’ subunit proteins interact with crRNA. Cas10 of Cmr complexes interact with Cmr3 (the Cas7 family protein) (*224, 225*), indicating an interaction with crRNA at the protein: protein interface. Cas8 family protein CasA (Cse1) interacts with the 5’ handle of crRNA and is discussed in greater detail a later section (*226, 227*): *Biochemical analysis of archaeal Cas8 in CRISPR interference*. The importance of CasA in crRNA binding and PAM recognition is explained, along with the role of the ‘large’ subunit in CRISPR mediated immunity. The ‘large’ subunits of these interference complexes have important roles in identifying PAM/TIM sequences (*228*), exemplified by the *E.coli* CasA (*162*). By analogy, Cas8 and Csy1 proteins are suggested to localise to similar regions of the ribonucleoprotein complex (*214, 229*). The association of the ‘large’ subunit with the rest of the complex seems to be transient and may well indicate that proteins like CasA scan invading genetic elements and when a target sequence is identified acts as a trigger to recruit the Cascade complex to test for crRNA homology.

Through a combination of crystal structure analysis and Cryo-EM, *E.coli* CasA has been shown to interact with the PAM region of target DNA identifying it as non-self DNA, and helping to prevent host DNA targeting (*162, 227*). CasA interaction with the CasBCDE-crRNA complex doesn’t alter DNA binding affinity, as K_Ds determined for CasBCDE and CasABCDE binding to dsDNA are indistinguishable (5 ± 1nM and 8 ±

4nM, respectively)(*128*). CasA protein did not interact with DNA. When modelled into the electron density map the L1 region (residues 130-143) interacted with the 5' handle of the crRNA. The suggestion is that CasA scans for PAM, triggering duplex destabilisation and crRNA/Cascade binding. Target binding by Cascade triggers a conformational change of CasA, CasB1-2 and CasE around the axis of CasC helical bundle. CasA and Cas3 have been shown to co-localise *in vivo*, supported by some CRISPR systems encoding Cas3-CasA fusions (*167*). Cas3 and the interaction with the 'large' subunit of Cascade complexes are essential for transfer of information and interference. The interaction of CasA with PAM and CasD (Cas5 family) facilitates seeding of the crRNA. It is also believed that L1 loop of CasA is responsible for the transfer of information from CasA and the Cascade complex to Cas3 (*227*). Mutations made in the L1 region, specifically N131A, perturbed Cas3 nuclease activity.

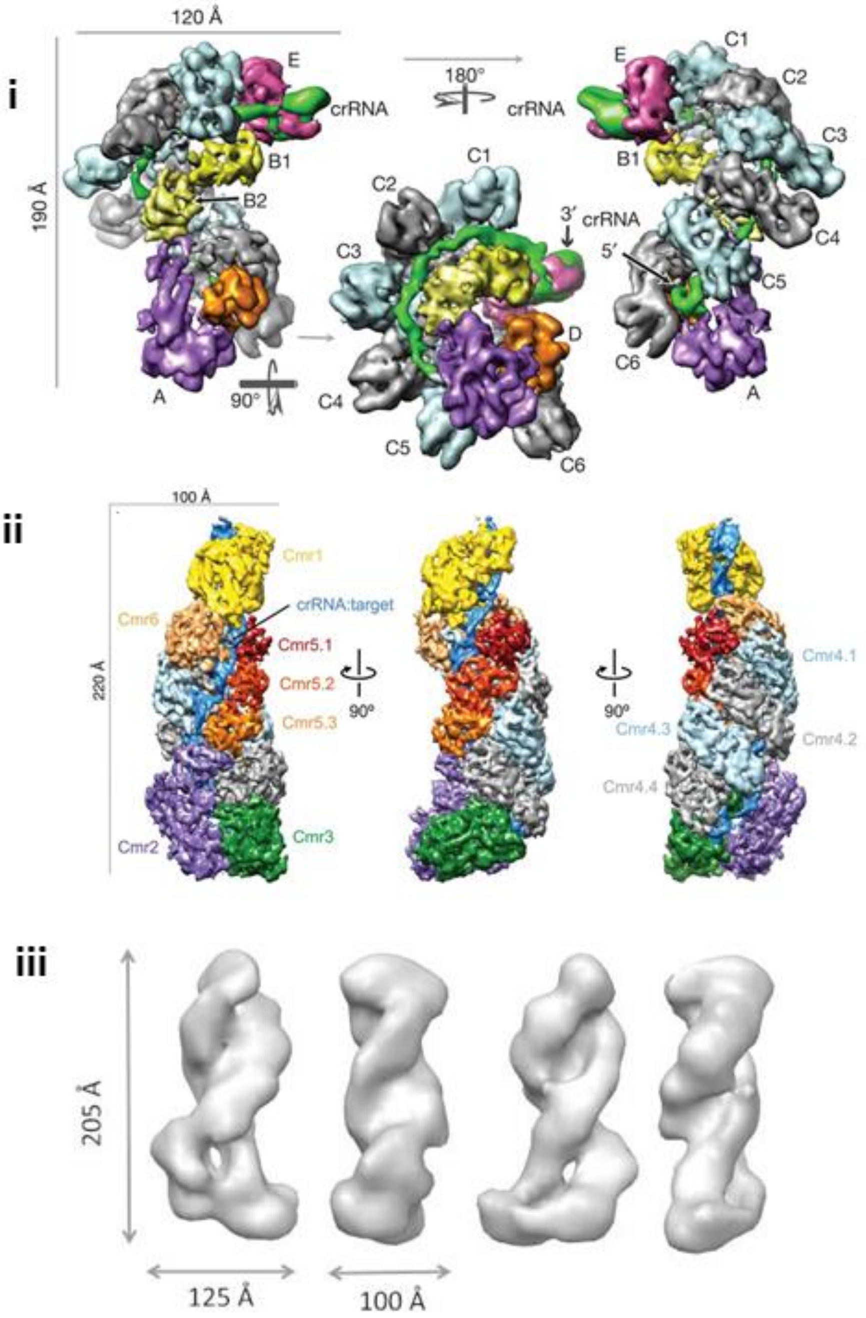


Figure 1-11. Comparison of the structures of different interference complexes from Type I and Type III CRISPR systems. The third stage of CRISPR-Cas immunity is orchestrated by interference complexes. In Type I systems these are called Cascades, the best studied example is that of *E.coli* (i). This Cascade has been co-crystallised with CRISPR-RNA (crRNA) and is said to structurally represent a 'sea-horse'. Type III systems have two general interference complexes: Csm and Cmr. The Cmr complex of *P. furiosus* (ii) has been well studied and also co-crystallised with crRNA. The Csm complex is less well understood and is typical of archaeal CRISPR immune systems. Atomic force microscopy (AFM) structures of the Csm complex have been determined to a much lower resolution (iii). Overall conservation of key architecture of various interference complexes presented here can be seen: Helical stacking of the crRNA and capping at the 3' and 5' ends of the RNA molecule. (Adapted from (*214*), *(230)* and (*137*)).

The diversity of ribonucleoprotein complexes explained above and structurally represented in Figure 1-11 does not alter the targeted identification of foreign genetic elements. Type III-B system Cmr interference complexes are the only CRISPR defence mechanism against foreign RNA elements (*224, 231*). In Type I CRISPR systems Cas3 is recruited to the R-loop formed by Cascade complexes (*167*). Transfer of correct R-loop and PAM information triggers heliCase/nuclease activity of Cas3 to degrade the invader DNA. In Type II systems, the interference complex consists of the Cas9 protein and one or two RNA molecules, tracrRNA and crRNA (guide RNA) or a chimeric RNA molecule. Cas9 protein has two nuclease catalytic domains: RuvC-like and HNH, and transloCase domains to degrade the invading DNA duplex (*209, 232*). Type III systems utilise Cas10 family proteins to carry out the degradation function (*233*).

1.4.2.3.3 Cas9

Recently, Cas9 has been exploited for its genome editing potential (*132-136, 234-236*). As a commercial successor to TALEN's (transcription activator-like effector nuclease) and transposases co-expression of Cas9 with synthetic crRNA's (or guide RNA [gRNA]) can attenuate gene expression (*134, 237-241*). TALEs were first discovered in the plant pathogen *Xanthomonas sp*. bound to DNA RVDs (repeat variable di-residues) altering gene expression in a sequence specific manner (*242*). Fusion of a TALE with *Fok I* nuclease (TALEN) bought about the first genome editing tool (*243*), which, through generation of targeted double-strand DNA breaks (DSBs) subsequently repaired by non-homologous end joining (NHEJ), could induce insertion or deletion mutations (*244*). These mutations cause frame shifts and subsequent gene expression attenuation. However repair by homologous recombination reduces mutation frequency. While TALENs were cost effective, mutation rates were low leaving the stage open to the development of more efficient new tools for genetic engineering such as Cas9.

Atomic resolution of Cas9 proteins has identified two nuclease domains: a RuvC-like RNase H fold and an HNH fold (similar to that found in T4 Endo VII) (*130, 168, 245*), shown in Figure 1-12. These folds are essential for nuclease degradation, shown by mutational analysis. Type II Cas9 has been harnessed for site-specific DSB generation. Through the use of programmable gRNAs, the efficiency of genome editing in human and mouse cell lines and zebrafish embryos has improved from 2-4% to 20% (*246*). Nuclease-deficient mutants (dCas9) have been developed for regulating transcription without DSB induction (*247*). Cas9 mediated genome engineering requires PAM sequences for stable target binding (*157, 177*). Cloning and transfection of multiple gRNAs can generate simultaneous targeting and expression regulation. Cas9 is an attractive mammalian editing tool with approximately 40.5% of human exons being suitable unique targets.

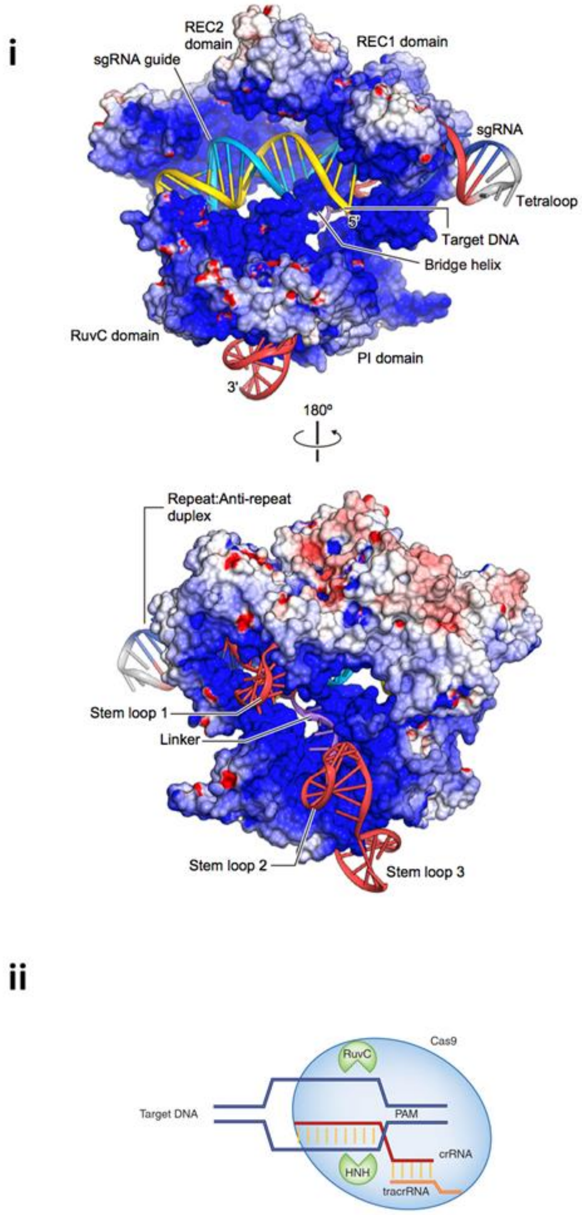


Figure 1-12. Crystal Structure of Cas9 in complex with Guide RNA and Target DNA. (i) Electrostatic surface potential of Cas9 protein in complex with guide RNA (gRNA – cyan), target DNA (yellow) and trans-activating CRISPR RNA (tracrRNA - red). HNH domain omitted for clarity. (ii) Cas9 forms R-loops through CRISPR-RNA (crRNA) or guide-RNA (gRNA) that is structurally arranged within the Cas9 by trans-activating crRNA (TracrRNA). Cas9 cleaves at R-loops through its RuvC-like domain and HNH fold. Pam sequence is essential for HNH activated cleavage. (Adapted from (*131*) and (*248*)).

PAM and gRNA are essential for complete double strand cleavage (two nicking reactions). Without PAM in a double strand context, only the HNH domain nicks the strand bound by the gRNA, the displaced strand remains intact. A single strand can be repaired by ligases without following DSBR, decreasing the likelihood of a phenotypically deleterious mutation. Mutant Cas9s (dCas9) are now used in gene regulation across a host of organisms by transfection and gene knockdown experiments.

1.4.2.3.4 CRISPR interference via R-loops

DNA/RNA hybrids and R-loops are important for regulating genome stability (*153, 249*), with their roles in various processes summarised in Table 2. An RNA strand invades a DNA duplex (typically supercoiled) and base pairs with one DNA strand, forming an R-loop. R-loops were first identified in bacteria, priming ColE1 plasmid replication (*250*). The importance of R-loops has also been recognised in mitochondrial and viral DNA replication priming. Persistent R-loops affect genome stability by blocking replication forks and initiating additional DNA replication, provoking illegitimate HR (*251*). Regulation and reversal of R-loops is carried out by host proteins, summarised in

Table 3.

Table 2. RNA molecules involved in genome dynamics. (Adapted from (*159*))

Process	RNA–DNA hybrid/R-loop
Transcription	Nascent RNA synthesis from DNA templates "Thread-back" and "extended hybrid" models of R-loop persistence during transcription (*251*)
Epigenetics: methylation	Protection of CpG promoter islands from methylation (*252*)
DNA replication	Priming of lagging strand synthesis. Priming of plasmid (ColE1), viral and mitochondrial replication (*253*)
Genome instability	R-loops provoking illegitimate recombination
CSR	R-loops at G-rich sequence provoke immunoglobulin diversity (*254*)
Telomere processing	"t-loop" of RNA and G-rich DNA (*255*)

CRISPR targeting	R-loop of crRNA targeting invasive DNA

Table 3. Proteins involved in the regulation and reversal of R-loops. (Adapted from (*159*))

Protein	**Activity**
RNaseHI	Ribonuclease on RNA strands base paired to DNA (*256*)
Topoisomerase I	Relaxation of negatively supercoiled DNA (*257*)
RecG	HeliCase unwinding R-loops (*258*)
Pif1	HeliCase unwinding RNA–DNA hybrids (*259*)
Senataxin/Sen1	HeliCases unwinding RNA–DNA hybrids (*260*)
Various transcription termination and mRNA processing factors	Bind to nascent RNA to prevent thread-back R-loops (*251*)
Cas3	HeliCase unwinding R-loops (*143*)

R-loops are essential to the successful targeted degradation of invading genetic elements. The action of the type I CRISPR interference system is summarised in a schematic overview from *E.coli*, shown in Figure 1-13. This summary suggests the process of interference: (a) assembly of the interference complex, (b) scanning of the invader by CasA, (c) R-loop formation and (d) Cas3 recruitment and degradation. Cas proteins from the same families have been highlighted in the figure legend to compare the proposed similarities with the *Mth* Type I-H system.

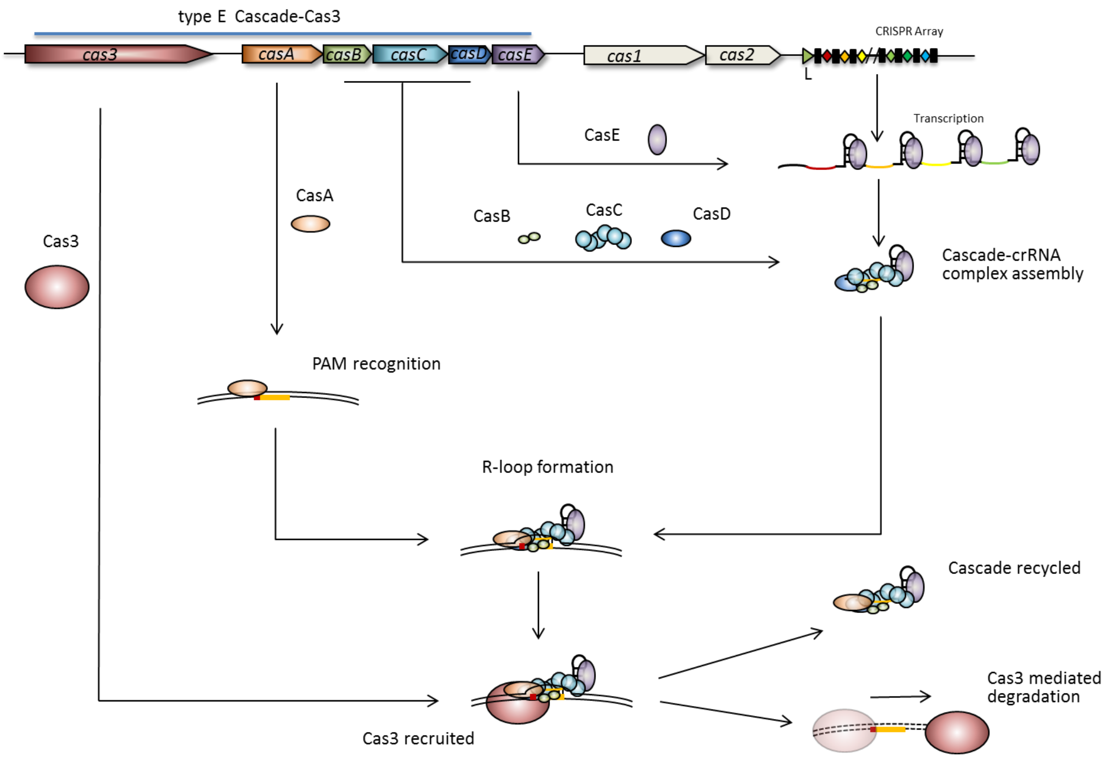


Figure 1-13. Comparing the *E.coli* Type I-E Cascade complex assembly and interference to the possible *Mth* Type I-H. This is an overview of the proposed role of Cascade proteins from *E.coli*: CasE (Cas6) processes pre-CRISPR-RNA (pre-crRNA) into mature crRNA, CasB (Cas8''), CasC (Cas7) and CasD (Cas5) assemble into the Cascade complex with CasE and crRNA. Each bracketed protein is the equivalent in *Mth*. CasA (Cas8') surveys DNA for Protospacer adjacent motif (PAM) and recruits the Cascade complex to initiate strand invasion and R-loop stabilisation. The stable R-loop recruits Cas3 to trigger degradation on the invading element and recycling of Cascade.

1.4.2.3.5 Cas3

Cas3 proteins are dual function enzymes consisting of an N-terminal HD phosphohydrolase domain and a C-terminal SFII (super family II) DExD/H-box heliCase domain (*261*). In some systems these two domains are expressed as separate ORFs but interact and act similarly. Other Cas3 variants exist for example, Cas3–Cas2 in I-F subtype (*181*) and Cas3–Cse1 (CasA) in some I-E systems (*167*). Cas3, unlike Cas9, is recruited to the R-loop generated by interference complexes and then degrades the invading genetic element. The nuclease activity of Cas3 requires conserved divalent metal cation coordination, typically magnesium (*165*). Superimposed HD domains of Cas3s demonstrate characteristic HD super family motifs in their amino acid sequence: H-HD-H-H-HD, and core 5 alpha-helical bundles as shown in TtCas3^{HD} (*Thermus thermophilus)* (*164*) and MjaCas3” (*Methanocaldococcus jannaschii*) (*161*).

Cas3 carries out interference via its two distinct nuclease activities: endonuclease cleavage of the R-loop and the 3’-5’ exonuclease activity (*164, 165, 167*). Cas3 crystal structures (Figure 1-14) aid understanding of how the exonuclease activity is coupled to the heliCase/transloCase activity of Cas3. Whereby two RecA-like domains coordinate ATP and so provide the chemical energy required for translocation through a ssDNA binding channel within the protein. The C-terminal heliCase domain of *T. terrenum* contains structural homology to the archaeal DNA heliCase Hel308 (*262*) and flavivirus RNA heliCase N53 (*263*). Cas3 recruited to the stabilised R-loop interacts with the ‘large’ subunit of the Cascade complexes (*162*). In *E.coli*, CasA (Cse1) directly interacts with Cas3 and transfers correct R-loop and PAM information triggering endonuclease cleavage of the R-loop, followed by transloCase/nuclease degradation of the invading genetic element (*162, 227, 264*).

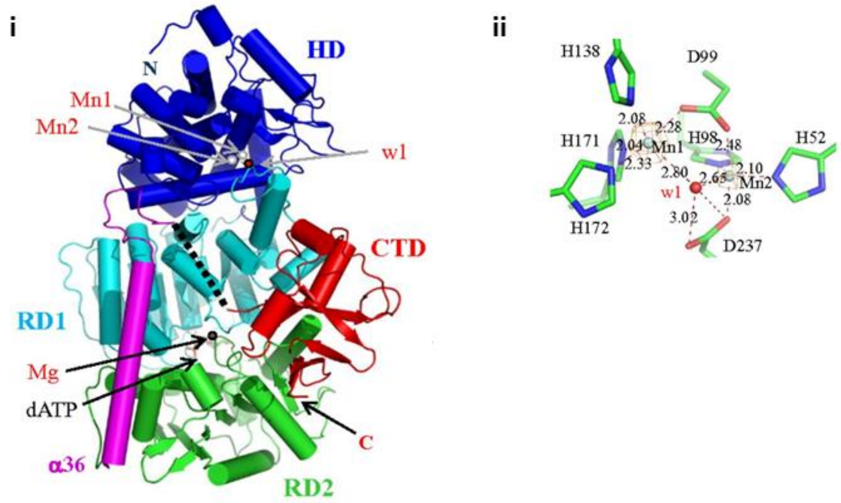


Figure 1-14. Crystal structure of Cas3 from *Thermobaculum terrenum.* (i) N-terminal HD nuclease domain (blue), unknown function CTD (red), and two RecA like motor domains (Cyan and green) are distinguished. (ii) HD nuclease domain active site, interaction lengths and distance are presented as dotted lines. Each metal ion label is associated with a water molecule. (Adapted from (*265*))

1.4.3 Outstanding questions about CRISPR-Cas

CRISPR-Cas research is still a young field and as such several important questions remain unanswered about the function and activity of these ‘immune’ systems:

- CRISPRs are similar to selfish MGE as some organisms repress CRISPR gene expression with H-NS. Whereas some archaea encode constitutively active CRISPR arrays, which are then further induced by cell stress responses (SOS). Through silencing of this immune system implies no selective advantage during cellular evolution (*169*). It is rare for such strains to lose CRISPR arrays. Therefore, are CRISPR elements selfish or evolutionary beneficial?
- The majority of spacers originate from MGE and phage, but others correspond to host chromosome sequences. Cas9 and synthetic CRISPR systems have been shown to effectively regulate gene expression in a host but how profound is this host regulation effect?
- Spacer acquisition can either be primed or naïve. Naïve spacer acquisition has a very low efficiency; the efficacy of CRISPR-Cas immunity to lytic phage may therefore require an additional stage of CRISPR immunity – Facilitation (*266*), a mechanism that prevents viral replication while a novel spacer can be integrated to target the invader. Primed acquisition, while improved from naïve is still less effective than would be expected from an efficient immune response. Why are these processes so inefficient? What is the mechanism controlling CRISPR spacer library expansion (*267*)?
- The mechanism of spacer integration is unknown; delineation of this process may lead to understanding the key details of this pathway. Cas1 proteins which act as transposons may improve our understanding on the function of CRISPR-Cas systems. Transposable DNA elements are important for driving diversity and evolution in a cell population, and often provide a selective

advantage. Do Cas proteins have additional roles distinct from simply producing an immune response (*118*)?

1.5 Aims

This body of work had three main objectives; as little is known about archaeal Cascade complexes, (i) analysis of archaeal Type I CASCADEs from *Methanothermobacter thermautotrophicus* and *Haloferax volcanii*, specifically the role of the essential Cas8 proteins in both of these similar systems, and understanding the mysterious heliCase/nuclease activity of Cas3. Cas proteins are speculated to interact with host DNA metabolism proteins with some existing evidence. Here, the aim was to (ii) identify interactions of *E.coli* Cas1, Cas2 and Cas3 with other *E.coli* proteins. At the outset of this study nothing was known about the mechanism of spacer acquisition, therefore (iii) the biochemical activities of *E.coli* Cas1 and Cas2 were tested *in vitro* to create a testable model of target capture and spacer integration.

2 Materials and Methods

2.1 Materials

2.1.1 *E.coli* Strains

Table 4. A list of bacterial strains used throughout this study. Each strain genotype is detailed and descriptions for use listed.

Strain	Genotype	Details and source
DH5α	*F- 80dlacZ M15 (lacZYA-argF) U169 recA1 endA1hsdR17(rk-, mk+) phoAsupE44 -thi-1 gyrA96 relA1*	Most suitable as a cloning stain, used for some protein overexpression. High quality plasmid DNA. Restriction, endonuclease and recombination deficient. *Dam+*. From NEB.
BL21Ai	*F- ompT hsdSB(rB*$^{-}$ *mB*$^{-}$*) gal dcm araB::T7RNAP-tetA*	Protein overexpression strain. Tightly controlled *araBAD* promoted T7RNAP. Protease deficient. Tetracyclin resistant. From Stratagene.
BL21 Codonplus	*F– ompT hsdS(rB*$^{-}$ *mB*$^{-}$*) dcm+ Tetr gal endA Hte [argU ileY leuW Cam*^{r}	Protein overexpression strain. Rare tRNA gene expression. Protease deficient. pLysS encoded T7RNAP. Chloramphenicol resistant. From Stratagene.
T7Express	*fhuA2 lacZ::T7 gene1 [lon] ompT gal sulA11 R(mcr-73::miniTn10--***Tet**^{S}*)2 [dcm] R(zgb-210::Tn10--***Tet**^{S}*) endA1 Δ(mcrC-mrr)114::IS10*	Protein overexpression strain. Protease deficient. T7RNAP within *lac* operon. Enhanced BL21 derivative. From NEB.
SoluBL21	*F- ompT hsdSB (rB*$^{-}$ *mB*$^{-}$ *) gal dcm (DE3)*	Protein overexpression strain. T7RNAP compatible. Protease deficient. Uncharacterised mutations obtained through special selection criteria for soluble intact expression. From Genlantis
IIB924	*F*$^{-}$*, DE(araD-araB)567, lacZ4787(del)::rrnB-3, LAM*$^{-}$*, rph-1, DE(rhaD-rhaB)568, hsdR514 araB::T7RNAP-tetA*	Protein overexpression and genetic test strain. Inducible T7RNAP. Tetracycline resistant. From Ivana Ivancic-Bace
XL10-Gold	*F´ proAB lacIqZDM15 Tn10 (Tet*^{r}*) Amy Cam*^{r}	Cloning strain. High efficiency plasmid transformation. Tetracycline and Chloramphenicol resistant. Restriction, endonuclease and recombination deficient. From Stratagene.

2.1.2 Chemicals and enzymes

In this study all chemicals and solutions were purchased from either Sigma, Fisher or VWR and were used as detailed in the methods later in this section. All commercial enzymes used were from New England Biolabs (NEB).

2.1.3 Media

2.1.3.1 E.coli media

Mu Broth: 1% (w/v) tryptone (Bacto), 0.5% (w/v) yeast extract, 340 mM NaCl, 2 mM NaOH, pH 8.0.

Mu Agar: 300ml Mu Broth, 0.5% Agar (w/v).

All solutions were prepared and sterilised by autoclave. Broth was stored in the dark at room temperature until needed. Mu agar plates were stored in sealed bags to prevent desiccation at 4°C.

2.1.3.2 *E.coli media supplements*

Each supplement was dissolved in Sterile Distilled Water (SDW), with the exception of chloramphenicol which was dissolved in 100% ethanol.

Table 5. Summary of media supplements used and their final working concentrations.

Supplement	**Abbreviation**	**Final Concentration**
Ampicillin	Amp	50 µg/ml
Chloramphenicol	Cm	10 µg/ml
Kanamycin	Km	40 µg/ml
Tetracycline	Tet	10 µg/ml
Streptomycin	Str	100 µg/ml
Apramycin	Appr	40 µg/ml
Isopropyl β-D-1-thiogalactopyranoside	IPTG	0.5 mM
D-Arabinose	Ara	0.2 %

2.2 Methods

2.2.1 General microbiology

2.2.1.1 *Growth and Storage of Escherichia coli*

Cultures on solid media were grown overnight at 37°C in a static incubator (LEEC). Liquid cultures were grown at either small scale (5-10 ml) or large scale (200-2000ml) in a shaking water bath or orbital shaker at 150-180rpm and were grown overnight or until desired OD_{600}. Typically large-scale cultures were for over-expression of proteins; growth and induction temperatures varied from 22°C to 37°C. Plates, when stored, were refrigerated. For long-term storage, glycerol stocks were prepared. A final concentration of 20% (v/v) glycerol was added to an overnight culture, flash frozen and stored at -80°C.

2.2.1.2 *Preparation of chemically competent Escherichia coli*

E.coli strains were initially streaked out on Mu agar plates containing appropriate antibiotic for strain selection and grown overnight at stated above. 8ml of Mu broth containing the selective antibiotic was inoculated with a single colony from the streak plate and grown overnight in a shaking water bath. 1:100 inocula of the overnight were inoculated into 50ml Mu broth, again with antibiotic selection where appropriate. Cultures were incubated with shaking until they reached an OD_{600} between 0.4-0.8. Typically at OD_{600} = 0.6 the culture was centrifuged at 4 krpm for 10 minutes. Supernatant was discarded and cell pellet was resuspended in 5ml ice cold 0.1M Calcium chloride (CaCl) and kept on ice for a minimum of an hour. Cells were centrifuged and resuspended as before with fresh CaCl for a further 15 minutes. For storage, 2ml of ice cold 80% glycerol was added and mixed. Cells were then aliquoted and frozen at -80°C until required.

2.2.1.3 Transformation of chemically competent Escherichia coli

For all transformations the Heat shock procedure was used, summarised briefly. 500ng of plasmid DNA was added to 100µl of chemically competent *E.coli* cells and incubated on ice for 30 minutes. Then the DNA/cell mix was subjected to heat shock at 42°C for 1 minute. After 1 minute rest on ice, 750µl of Mu broth was added and the cells were then incubated at 37°C in a shaking water bath for 30-60 minutes. Cells were then centrifuged at 13.2 krpm for 1 minute and 750µl of supernatant removed. The resuspended cell pellet was then plated on Mu agar that contained the appropriate antibiotic to select for strain and plasmid. Plates once dried were incubated overnight at 37°C.

2.2.2 Molecular cloning and mutagenesis

2.2.2.1 PCR gene amplification and Quick change (QuCh) mutagenesis

Amplification of either genomic or plasmid DNA was carried out using either Vent or Q5 polymerases. These enzymes are both high fidelity enzymes that 'proof-read' to ensure accurate gene amplification by checking correct base pairing. Reaction conditions for both enzymes are shown below and either Biometra thermocycler basic or gradient were used. Typical PCR reaction is as follows:

Vent DNA Polymerase:
1 x Thermopol Buffer
10mM dNTPs
10µM forward and reverse Primers
10-100ng template DNA
0.5µl Vent polymerase
(Additional $MgSO_4$ if required)
Final reaction volume of 50µl

Q5 High Fidelity DNA Polymerase:
1 x Q5 Reaction Buffer
10mM dNTPs
10µM forward and reverse Primers
10-100ng template DNA
0.5µl Q5 polymerase
1x GC enhancer
Final reaction volume of 50µl

For annealing temperatures NEB T_m Calculator was used, which follows Howley *et al*'s equation, taking into account GC content, primer length and sequence homology. If this annealing temperature generated no product however, a gradient PCR was set up over the annealing temperatures to cover 7.5°C above and below the suggested T_m.

Table 6. Typical thermocycling programme for Polymerase Chain Reaction. Steps of primer based extension include denaturing template, annealing primers and polymerase extension, repeated 30 times.

Cycle	**Vent**	**Q5**	
Denaturation	5min 95°C	5min 95°C	
Denaturation	45sec 95°C	45sec 95°C	x 30
Annealing	1min 45-65°C	1min 45-65°C	
Extension	1min+1min per kb at 72°C	30sec+30sec per kb at 72°C	
Final extension	5min at 72°C	5min at 72°C	

For QuCh PCR's the same principle was followed. In each of these specific mono or dinucleotide mismatched primers were used for the forward and the reverse primer.

The template was always that of a sequenced plasmid encoding the gene of interest. A total of 13 cycles of the 'PCR' were used; thermocycling is the more accurate description here.

2.2.2.2 Restriction Endonuclease Digestion of oligonucleotides

All restriction digests were carried out as per the manufacturer's instructions (NEB). Where required digests were supplemented with the addition of 200 ng/µl BSA (NEB). When double digests were performed, NEB buffers were selected where both enzymes had 100% activity, otherwise a sequential digest was carried out. DNA was digested for between 30 minutes and 16 hours, when appropriate for cutting near the ends of PCR products. Longer digests were carried out in a sealed 37°C incubator, to prevent evaporation and star activity from increased enzyme concentration.

2.2.2.3 Dephosphorylation of digested vector DNA during cloning

Antarctic phosphatase was used to remove 5'phosphate groups of the linearized plasmids. Removal of the 5' phosphate prevents self ligation of the linear vector, thus reducing the chance of colonies containing only vector DNA. Vectors were incubated with 1 unit of Antarctic phosphatase/µg of DNA with the addition of Antarctic phosphatase buffer to a final 1X concentration for 30 minutes at 37°C. Antarctic phosphatase could be heat inactivated at 65°C for 5 minutes afterwards, often though this step was omitted because the enzyme did not affect subsequent modification of the DNA

2.2.2.4 Ligation of DNA

Ligations of gene insert and vector were performed using T4 DNA ligase. 5 units of DNA ligase were used plus T4 ligase buffer at a final 1X concentration for each reaction. Ligations contained a molar ratio of ~3:1 (insert:vector) DNA. Ligations were carried out at 4°C overnight.

2.2.2.5 Ethanol Precipitation of DNA

Ethanol precipitation was used to either concentrate DNA samples, exchange buffer or remove unwanted protein contaminants. 1 volume of 4M ammonium acetate and 4 volumes of 100% ethanol (stored at -20°C) were added to DNA samples and incubated at -20°C overnight. Samples were centrifuged at 4°C, 20,000 ×*g* for 30 minutes and the supernatant removed. Pellets were immediately washed in 100µl of 70% ethanol followed by centrifugation as before. The supernatant was removed and pellets allowed to air dry before resuspension in buffer (usually Sterile distilled water, SDW).

2.2.2.6 Nucleic acid Purification

DNA was purified by two main methods. Plasmid DNA was purified by Qiagen mini and maxi prep protocols. PCRs and restriction digests were purified using the Qiagen Gel Extraction kit. Plasmid extractions are based on the alkaline lysis method followed by adsorption of DNA onto a silica gel in high salt. Maxi preps include an ethanol precipitation step for higher quality DNA. Plasmid extractions were from *E.coli* overnight cultures. 8ml cultures in the miniprep method and 500ml for maxiprep. These cultures were always grown in the DH5α strain at 37°C in Mu broth with appropriate antibiotic supplement. The protocols were followed exactly as is the manufacturer's instructions directed, except DNA was eluted from mini preps and finally resuspended from maxi preps with SDW. This was the Case with gel extractions also, elution into SDW. Plasmid stocks generated were stored at -20°C.

2.2.2.7 DNA Sequencing

The Biopolymer Synthesis and Analysis Unit, University of Nottingham, or Source Biosciences (Cambridge UK) carried out all sequencing reactions and analysis.

Sequencing was carried out using the dideoxy chain termination method (Sanger et al., 1977).

2.2.2.8 Oligonucleotide Synthesis

Oligonucleotides were synthesised either by the Eurofins MWG (Germany), or Sigma-Aldrich (UK). Primer oligonucleotides were synthesis at 0.01µM scale and desalted. Substrate oligonucleotides were synthesised at 0.2µM scale and HPLC purified. All sequences were verified by MALDI-ToF.

2.2.2.9 Nucleic Acid Quantification

Plasmid preparation concentration was determined using the 260/280 nm absorbance ratio from spectrophotometer measurements (NanoDrop) (Beckman Coulter DU 530).

2.2.3 Gel electrophoresis

2.2.3.1 Agarose Gel Electrophoresis

Buffers and Solutions:

TBE (<u>T</u>ris <u>B</u>orate <u>E</u>DTA): 89 mM Tris.HCl, 89 mM boric acid, 2 mM EDTA.

Gel Loading Dye (5×): 50 mM Tris · HCl, 100 mM EDTA, 15% Ficoll (w/v), 0.25% Bromophenol Blue (w/v), 0.25% Xylene Cyanol FF (w/v).

Agarose gels were Cast using agarose powder (between 0.5-3%) and 1x TBE. Ethidium bromide was either added to a final concentration of 0.5 µg/ml or gels were washed after electrophoresis in TBE containing 0.5 µg/ml ethidium bromide to allow for visualisation of DNA. Samples were mixed with gel loading dye to a 1x final concentration and loaded alongside either a 1Kb or a 100bp (NEB) size markers. Standard TBE gels (10cm) were run at 100 V for 1 hour, other gels varied as detailed in later sections.

2.2.3.2 Agarose Gel Extraction and Purification of DNA

Purification of DNA from agarose gels was carried out by UV exposure using a UV transilluminator (UVP inc.). Desired DNA was extracted dependant on size and purified using the Qiagen Gel extraction kit, detailed earlier.

2.2.3.3 SDS-Polyacrylamide Gel Electrophoresis

Buffers and solutions

7.5, 10, 12 or 15% SDS-PAGE gel (resolving): 7.5, 10, 12 or 15% acrylamide/bisacrylamide, 0.37 M Tris (pH 8.8), 0.1% (w/v) SDS, 0.05% (w/v) APS (ammonium persulfate), 0.05% (v/v) TEMED (tetramethyleethylenediamine).

3.0% SDS-PAGE gel (stacking): 3% acrylamide/bisacrylamide, 0.25 M Tris (pH 6.8), 0.2% (w/v) SDS, 0.125% APS, (w/v) 0.125% (v/v) TEMED.

SDS-PAGE running buffer: 0.25 M Tris, 1.92 M glycine, 1% (w/v) SDS.

Laemmli buffer (4×): 50mM Tris pH 6.8, 100 mM DTT, 2% (w/v) SDS, 0.1% (w/v) bromophenol blue, 10% (v/v) glycerol.

Staining solution: 40% (v/v) methanol, 10% (v/v) glacial acetic acid, 0.85mM Coomassie Brilliant Blue G-250

Destaining solution: 20% (v/v) methanol, 7% (v/v) glacial acetic acid.

Protein samples were analysed using SDS-PAGE (sodium dodecyl sulphate polyacrylamide gel electrophoresis). Gels were made in self-assembly Cassettes (BioRad). A 7.5, 10, 12 or 15% resolving gel was poured with a layer of butanol saturated water on top, leaving a level surface. Once set, the butanol was removed and a 3% stacking gel poured and a comb inserted. Protein samples were mixed with Laemmli buffer to a final concentration of 1x. Samples were denatured by boiling at 95°C for 10 minutes and run alongside a PageRuler size ladder (Fermentas) or Colourplus marker (NEB). Gels were run for ~1 hour 10 minutes at 220V SDS-PAGE running buffer and then stained with Coomassie Brilliant blue staining solution for 10-30 minutes, followed by gel destaining to visualise proteins.

2.2.3.4 Western blot

Buffers and Solutions

Transfer Buffer: 25mM Tris, 190mM Glycine, 0.1% (w/v) SDS.

TBST (Tris Buffered Saline Tween): 20mM Tris-HCl pH7.5, 150mM NaCl, 0.1% (v/v) Tween 20.

WBB (Western Blocking Buffer): 20mM Tris-HCl pH7.5, 150mM NaCl, 0.1% (v/v) Tween 20, 5% (w/v) milk powder.

After SDS-PAGE, gels that were to be probed with antibodies rather than Coomassie stained were electroblotted onto PVDF (Polyvinylidene Fluoride) membrane. Gels were equilibrated in Transfer buffer for 10 minutes. During this time, PVDF was activated by exposure to 100% methanol for about 10 seconds and then also equilibrated in transfer buffer. For a wet blot, the Xcell SureLock (Thermo Fisher Scientific) was set up in the typical electroblotting 'sandwich', ensuring membrane was closer to the positive electrode than the gel and no air bubbles were present between any of the individual layers. Electroblot was then carried out at 100V for 1 hour or 15V for between 4 hours and overnight.

After blotting, membranes were blocked overnight at 4°C using the WBB with constant agitation to prevent nonspecific antibody binding. Primary antibodies, either anti-His (Sigma Aldrich), anti-MBP (NEB) or anti-strep (NEB) were washed to probe the membrane in WBB with a usual dilution 1:5000 of antibodies. This was incubated for 1 hour at room temperature with agitation. Unbound primary antibody was then removed by 3 wash steps. Washes were 5 minutes with fresh TBST buffer, and then discarded. Secondary antibodies were then used to detect the primary antibody depending on how the primary antibody was raised. Regularly primary antibodies are raised in mouse and rabbit. So secondary antibodies are anti-mouse or anti-rabbit and are HRP (Horse Radish Peroxidase) conjugated. Secondary was washed over membranes for 1 hour at room temperature at a dilution of between

1:1000 and 1:2500. Membranes were washed as before and the ready for development. Western blots were developed by ECL (Thermo Scientific) and visualised via FujiFilm Las-3000 chemoluminescence photography. Exposure times between 1 – 30 minutes.

2.2.4 Phenotyping of *Haloferax volcanii* by DNA-Damage Assays

In order to determine the extent of sensitivity to DNA damaging agents, assays were performed using a number of mutagens:

2.2.4.1 UV-Irradiation

Ultraviolet light is electromagnetic radiation outside of the visible spectrum. UV light is the most common cause of DNA radiation damage, creating thymidine dimers. 100µl of each *H.volcanii* strain was sequentially diluted in 5 ml of HV-YPC broth (+Thy if required) and grown at 45°C rotating at 20rpm, to an OD $_{650}$ ≈ 0.35-0.4. 1 ml of culture was pelleted and resuspended in 18% salt water and serial diluted x10^{-1} to x10^{-6}. Spot tests were then carried out; 20µl spots were spotted out in duplicate onto Hv-YPC agar (+Thy if required) and allowed to dry at room temperature. Plates were exposed to UV light (254 nm, 1 J/m^2/sec) and shielded from visible light to prevent photo-reactivation. Plates were incubated at 45°C for 4-7 days until no detectable growth was observed and colonies counted. Survival fractions were then calculated with an unirradiated control.

2.2.4.2 Mitomycin C

Mitomycin C (MMC), a DNA crosslinking agent often used in chemotherapies, sensitivity was analysed as a chronic exposure, with MMC in the agar plates. Cultures were grown as before and then cells were spotted directly onto Hv-YPC agar (+Thy if required) containing 0-0.03 µg/ml of MMC. The half-life of MMC is 4-6 days so plates were made fresh and only used within 5 days.

2.2.5 *Methanothermobacter thermautotrophicus* protein overexpression and purification

Buffers and Solutions

Buffer A: 20mM Tris-HCl pH 8.0, 150mM NaCl, 1mM DTT, 0.1mM PMSF.

Buffer B: 20mM Tris-HCl pH 8.0, 150mM NaCl, 1mM DTT, 0.1mM PMSF, 3mM Maltose.

His Binding Buffer: 20mM Tris-HCl pH8.0, 500mM NaCl.

His Charge Buffer: 20mM $NiCl_2$

His Wash Buffer: 20mM Tris-HCl pH8.0, 500mM NaCl, 10mM Imidazole.

His Elute Buffer: 20mM Tris-HCl pH8.0, 500mM NaCl, 500mM Imidazole.

His Strip Buffer: 20mM Tris-HCl pH8.0, 500mM NaCl, 50mM EDTA.

Buffer C: 20mM Tris-HCl pH 8.0, 1500mM NaCl, 1mM DTT, 0.1mM PMSF.

Dialysis Buffer 1: Same as Buffer A

Dialysis Buffer 2: 20mM Tris-HCl pH 8.0, 150mM NaCl, 1mM DTT, 0.1mM PMSF, 40% (v/v) Glycerol.

Dialysis Buffer 3: 20mM Tris-HCl pH 8.0, 500mM KoAc, 1mM DTT, 0.1mM PMSF, 40% (v/v) Glycerol

2.2.5.1 Protein induction and overexpression

Some of the proteins used in this study already had established purification procedures that could be followed; these are detailed later. An initial screen on the remaining *Mth* Cas proteins was carried out by overexpression and solubility profiling.

In these expression tests the plasmid constructs were transformed into BL21AI and BL21C+ strains. After overnight growth fresh 8ml of Mu was inoculated 1:50 with overnight culture and grown to an OD_{600}= 0.6 and 37°C. Samples were taken before induction and then protein expression was induced with 0.5mM IPTG and 0.2% arabinose for BL21AI. Cultures were sampled at time points of between 2-4 hours and overnight. 1ml samples were centrifuged at 13.2 krpm for 1 minute, supernatant removed and pellet resuspended in 300µl SDW and 100µl Laemmli buffer. These were then analysed via SDS-PAGE for protein expression.

If expression was observed, a solubility screen followed. For each strain fresh 8ml of Mu was inoculated with overnight cultures as before and then grown at 37°C to OD_{600}=0.6. Expression was then induced at temperatures between 22°C and 37°C. Time point samples were taken as before. Once samples were resuspended they were lysed via sonication, 30 seconds at 50% amplitude (intensity). Lysate was then clarified by centrifugation at 13.2 krpm for 10 minutes. A sample of both the pellet and supernatant were then taken for SDS-PAGE.

For some *Mth* Cas proteins expression was also attempted in SoluBL21 and T7express. *E.coli* Cas 1 and Cas2 were found to express in either T7Express or IIB994. The screening tests outlined above were followed by the lab until expression conditions were found that were appropriate for expression.

In each of the Cases below, upscaling of the protein expression for protein purification followed different procedures and they are outlined at the beginning of each subsequent section. All purification was carried out using the Amersham Pharmacia AKTA FPLC system unless stated. All pre-packed columns were from GE healthcare.

2.2.5.2 Purification of Mth Cas3

The plasmid pEB359 encoding the *Mth Cas3* gene (with no affinity tag) was transformed into BL21C+. 4 litre expression cultures were set up containing ampicillin and chloramphenicol and grown to OD_{600}=0.6 at 37°C. Once induced with IPTG at 0.5mM the cells were incubated at 30°C for a further 2-4 hours. Biomass was harvested and frozen as before.

Purification of untagged proteins can be difficult, but this 4 step chromatographic method worked well (as developed by Jamieson Howard). The first step was passing the clarified sonicated supernatant through 5ml Heparin FF; Cas3 did not bind but many contaminants were removed. The flowthrough from the heparin column was then used as the input for a 5ml Q Sepharose FF column (cation exchange). *Mth* Cas3 was eluted over a NaCl Gradient (150-1500mM) between 400-600mM NaCl. Pooled fractions were then loaded onto a 5ml Phenyl Sepharose FF (Hydrophobic interaction column, HIC) column, which was pre-equilibrated at 600mM NaCl. A gradient of NaCl (600-0mM) was used; Cas3 eluted at 0mM NaCl. Cas3 fractions were then loaded onto 7ml manually prepared DEAE (diethyl-aminoethyl) column, a second, weaker cation exchange column. Elution was over NaCl gradient (150-1500mM). Pooled fractions were dialysed into Dialysis Buffer 3 and stored at -80°C.

2.2.5.3 Purification of Mth His$_6$Cas8'

The original clone of Cas8' (pEB389) was used to provide N-terminally tagged His$_6$-Cas8'; this construct was used for site directed mutagenesis of Cas8', based on the Quick-change protocol. These mutations, after verification by DNA sequencing, were expressed and purified as with the wild-type protein (*144*).

Each His$_6$Cas8' gene was transformed into BL21C+, overnight culture used to inoculate cultures and grown to O.D$_{600}$=0.6, and expression induced for 2-4 hours at 30°C with IPTG. Harvested cells were resuspended in His Binding Buffer containing phenyl methyl sulfonyl fluoride (PMSF, 0.1 mM), and freeze-thawed prior to lysis by sonication, followed by centrifugation at 39,000 g for 20 min. Soluble proteins were loaded onto a 5ml HiTrap Chelating FF column charged with nickel chloride and equilibrated in His Binding Buffer. His$_6$Cas8' eluted into fractions within a gradient of 10-500mM imidazole and were pooled and loaded onto a HI Load Superdex 200

26/60 column equilibrated in Buffer B followed by elution in the same buffer in one column volume. His_6Cas8' fractions were pooled and loaded onto 5ml Heparin HP column equilibrated in buffer B. His_6Cas8' proteins eluted in a gradient of 150-1500mM NaCl, and fractions containing His_6Cas8' were pooled and dialysed into Dialysis Buffer 2 for storage in aliquots at -80°C.

2.2.5.4 Purification of Mth His_6Cas7

Cas7 was available as a clone that was known to overexpress and produce a soluble protein. pEB388 (His_6Cas7 in pET14b) was transformed into BL21C+, and expressed in the exact same manner as His_6Cas8' proteins. Soluble proteins were then loaded onto 5ml HiTrap Chelating HP Column and eluted within 10-500mM imidazole gradient. Fractions containing His_6Cas7 were pooled and loading onto 1ml Heparin HP equilibrated in Buffer B. Elution gradient of NaCl (150-1500mM) was used and fractions containing Cas7 were pooled and dialysed into Dialysis Buffer 3. Aliquots were stored at -80°C.

2.2.5.5 Co-purification of Mth MBPCas5 and Cas7

This method is summarised in Cass *et* al. and in the appendix (*144*); a full explanation of the method used is here. Cas5 protein could be overexpressed but not purified in isolation, the protein was insoluble; however, using MBP (maltose binding protein) and a strongly predicted interaction partner Cas7, co-expression of these 2 proteins generated soluble MBPCas5 and Cas7 proteins. MBPCas5 (pSDC25, Cas5 in pMal-C2x) and non-tagged Cas7 (pSDC38, Cas7 in pCDF-1b) were sequentially transformed into BL21C+ because co-transformation would not yield any colonies under any conditions. Plates and media contained ampicillin (for pSDC25) and streptomycin (for pSDC38) and cultures for overexpression were supplemented with 0.2% (w/v) glucose to reduce amylase production. Overexpression was carried out in the exact conditions as Cas8’ and Cas7. Cultures were grown from overnight cultures at 37°C, expression induced with 0.5mM IPTG at 30°C for 2-4 hours and then cells harvested and frozen.

Soluble proteins after sonication were loaded onto 7ml Amylose resin selfpour column. MBPCas5 and Cas7 co-eluted within the maltose gradient (0-3mM). Fractions containing MBPCas5 and Cas7 were pooled and then loaded onto 1ml Heparin HP and eluted within an NaCl gradient (150-1500mM). Fractions containing MBPCas5 and Cas7 were dialysed in Dialysis buffer 3, aliquoted and stored at -80°C.

2.2.5.6 Purification of Mth Cas5 and Cas6

Cas5, like Cas6, only expressed as insoluble proteins in *E.coli,* so after various expression conditions were tested, purification was attempted via refolding dialysis. Cas6 proteins were expressed in the same manner as previous *Mth* proteins. Inclusion bodies were then prepared. After sonication in a lysis buffer (50mM Tris-HCl pH 8.0, 100mM NaCl, 5mM EDTA, 0.1%NaN_3, 0.5% Triton X-100, 1mM DTT and 0.1mM PMSF) the insoluble fraction was retained. 10mM $MgSO_4$ was then added to chelate the EDTA, followed by DNaseI (0.01mg/ml) and lysozyme (0.1mg/ml) treatment for 20 minutes at room temperature. The mix was then centrifuged at 6 krpm for 15 minutes. The pellet was crushed with a spatula before resuspending in lysis buffer by sonication. DNaseI and lysozyme treatment was carried out again. This process was repeated three more times; the final resuspension step was in lysis buffer minus Triton X-100 followed by centrifugation, pelleting inclusion bodies. Inclusion bodies were diluted in 100mM Tris-HCl pH 8.0, 50mM Glycine and dispersed using sonication. To dissolve the suspension, it was added dropwise to vigorously mixing solubilisation buffer (100mM Tris-HCl pH 8.0, 50mM Glycine, 8.5M Urea). Dissolved inclusion bodies were then ready for refolding. Refolding was a week-long process of sequential dialysis steps in a refolding buffer (0.1M Tris, 0.4M L-Arginine) supplemented with decreasing urea (4M, 2M, 1M, 0M), adjusted to pH 8.0 and then 1mM EDTA and 0.1mM PMSF added immediately before use. Each dialysis step was for 24 hours. At different stages of this purification Cas5 and Cas6 precipitated and formed stubborn aggregates that were only soluble in reducing agents. So a similar method was approached with Cas6, for soluble expression in *E.coli*.

2.2.5.7 Purification of Mth MBPCas6

MBPCas6 (pSDC28, Cas6 in pMal-C2x) was transformed into BL21C+ and was expressed and purified in the exact same manner as MBPCas5 and Cas7. Expression at 30°C, purification by amylose affinity column and then heparin affinity/cation exchange and storage in Dialysis buffer 3 at -80°C.

2.2.6 *E.coli* Protein overexpression and purification

Buffers and Solutions

Buffer A: 20mM Tris-HCl pH 8.0, 150mM NaCl, 1mM DTT, 0.1mM PMSF.

Buffer B: 20mM Tris-HCl pH 8.0, 150mM NaCl, 1mM DTT, 0.1mM PMSF, 3mM Maltose.

His Binding Buffer: 20mM Tris-HCl pH8.0, 500mM NaCl.

His Charge Buffer: 20mM $NiCl_2$

His Wash Buffer: 20mM Tris-HCl pH8.0, 500mM NaCl, 10mM Imidazole.

His Elute Buffer: 20mM Tris-HCl pH8.0, 500mM NaCl, 500mM Imidazole.

His Strip Buffer: 20mM Tris-HCl pH8.0, 500mM NaCl, 50mM EDTA.

Buffer C: 20mM Tris-HCl pH 8.0, 1500mM NaCl, 1mM DTT, 0.1mM PMSF.

Dialysis Buffer 1: Same as Buffer A

Dialysis Buffer 2: 20mM Tris-HCl pH 8.0, 150mM NaCl, 1mM DTT, 0.1mM PMSF, 40% (v/v) Glycerol.

Dialysis Buffer 3: 20mM Tris-HCl pH 8.0, 500mM KoAc, 1mM DTT, 0.1mM PMSF, 40% (v/v) Glycerol

2.2.6.1 Purification of E.coli MBPCas3

E.coli MBPCas was overexpressed following the established protocol described in Howard *et al.* 2011 (*143*). DH5α was transformed with pAH1 (Cas3 in pMal-C2x). A fresh overnight culture was used to inoculate 4 litres of Mu broth containing ampicillin and 0.2% Glucose, to inhibit amylase production. 10 x 400ml of cultures were incubated at 37°C until OD_{600}=0.6. Protein expression was induced with 0.5mM IPTG for 30 minutes. Culture was then immediately transferred to chilled centrifuge bottles and centrifuged at 4 krpm for 20 minutes at 4°C in a pre-chilled centrifuge Avanti-G50. Supernatants were discarded and pellets resuspended in a minimum volume of Buffer A, 5-20ml, and flash frozen and stored at -80°C.

Biomass was thawed on ice and sonicated at 50% intensity for a total of 1 minute, (10 seconds on, 10 seconds off) to allow cooling. The lysate was clarified by centrifugation at 20 krpm for 20 minutes. The soluble fraction was then loaded directly onto a self-pour 7ml Amylose resin column, pre-equilibrated in Buffer A. An amylose gradient was then used to elute bound proteins (0-3mM maltose). Elution peaks were analysed and fractions containing MBPCas3 were pooled. These fractions were then loaded onto 5ml Heparin HP column. Bound proteins were eluted over a NaCl gradient (150-1500mM). One broad peak was observed on the UV trace, fractions were analysed as before and highest concentration fractions were pooled. A repeat of the previous Amylose column was carried out here. This is because of the contaminating MBP degradation products (previously shown by western blotting by Jamieson Howard). This column removed some of the degradation products and resulted in a higher purity of full length MBPCas3. After elution from this column pooled fractions were dialysed overnight into Dialysis buffer 2, aliquoted and stored at -80°C.

2.2.6.2 *Purification of E.coli His$_6$-Topoisomerase I*

pEB350 (TopoI in pET14b) was transformed into BL21AI. TopoI was expressed in a similar way as MBPCas3. 4 litres of Mu broth with ampicillin was inoculated with overnight culture and grown to OD$_{600}$=0.6. Expression was then induced with 0.2% arabinose and 0.5mM IPTG for 2-4 hours at 37°C. Cells were pelleted and resuspended in His Binding Buffer and frozen.

Biomass was thawed, sonicated and clarified as before. 5ml HiTrap Chelating column was prepared by charging with His Charge Buffer and the equilibrated in His Binding Buffer. Supernantant from clarification was loaded and after a wash step in His Wash Buffer an elution of imidazole (10-500mM) was used to elute bound proteins. Peak fractions were pooled and then dialysed for 3 hours or overnight into Dialysis Buffer 1. After this the sample was loaded onto a 1ml Heparin FF column. A NaCl gradient (150-1500mM) was used for elution, again peak fractions were analysed, pooled and the dialysed into Dialysis Buffer 2. Concentrated protein was then aliquoted and stored at -80°C.

2.2.6.3 *Purification of E.coli His$_6$Cas1*

Cas1 was generated in the lab with several different tagging methods. His$_6$Cas1 was found to be the easiest to handle and purify and is summarised in Rollie *et al*. (*150*) and Ivancic-Bace *et al* (*268*). His$_6$Cas1 (pEB549) was transformed into IIB942. Cells were grown from overnight cultures to OD$_{600}$= 0.5, before expression was induced with arabinose (arabinose inducible T7 promotor) and ITPG (0.5mM) for 3-6 hours. Cells were harvested by the normal centrifugation and resuspension step into His binding buffer.

After sonication and clarification, soluble proteins were loaded onto an equilibrated 5ml HiTrap Chelation FF column, charged with Ni^{2+} ions. His_6Cas1 eluted over an imidazole gradient (10-500mM). Fractions containing Cas1 were pooled and loaded directly onto a HI Load Superdex 200 26/60 column equilibrated in Buffer B. Fractions of His_6Cas1 were collected over an isocratic gradient of 1 CV in buffer B and pooled. These were the loaded onto a 1ml Heparin column and Cas1 eluted within the NaCl gradient (150-1500mM). Protein was pooled, dialysed against Dialysis buffer 2 and stored in aliquots at -80°C.

2.2.6.4 Purification of E.coli Cas1HisStrep

Cas1HisStrep was already available cloned (pASB11, Cas1 in pQE-His1). pASB11 was transformed into T7express cells for optimal overexpression. Overnights were used to inoculate up to 4 litres of growth media. Cultures were grown to exponential phase OD_{600} of 0.6 at 37°C and expression was induced with IPTG (0.5mM) for 3-6 hours at 30°C. Biomass was harvested and resuspended in Buffer A. After sonication and clarification soluble protein were loaded onto a 1ml Strep-Tactin SuperFlow Plus (Qiagen) column. Cas1HisStrep was eluted within a gradient of desthiobiotin (0-2.5mM). Fractions containing Cas1 were then loaded directly onto 1ml HisTrap FF column, charged with Ni^{2+}. Elution over imidazole gradient (10-500mM) liberated Cas1HisStrep. Pooled fractions containing Cas1 were then dialysed again Dialysis Buffer 2 overnight at 4°C, aliquoted and stored at -80°C.

2.2.7 Protein crystallisation

2.2.7.1 Minimal buffering conditions

Cas8' protein that had been expressed by the method outlined above, but in a 10 litre fermentation reaction vessel, with vigorous agitation and aeration. This provided a large biomass for bulk purification. Cas8' was expressed and purified this way. Minimum solubility parameters were then set up as the starting point for crystallisation. The storage buffer for the protein was Dialysis Buffer 3 (20mM Tris-HCl pH 8.0, 500mM KOAc, 1mM DTT, 0.1mM PMSF, 40% (v/v) Glycerol), so each constituent's concentration was altered to get the minimum salt, glycerol and DTT while maintaining soluble protein, the conditions attempted are highlighted below in Table 7.

Table 7. Buffer conditions trailed for *Mth* Cas8' crystallisation.

Buffer Condition	KOAc (mM)	DTT (mM)	Glycerol (% v/v)
1	275	1	40
2	275	0	5
3	500	1	22.5
4	50	0	22.5
5	500	1	5
6	50	0	40
7	500	0	40
8	50	1	5
9	275	0	22.5
10	275	1	22.5

2.2.7.2 Crystallisation

Once the minimal conditions were set, crystal tray preliminary screens were set up, thanks to Richard Rymer of Panos Soultanas' Lab in the Centre of Biomolecular Sciences (University of Nottingham) for assistance in the step up and analysis.

2.2.8 *In* vitro protein-protein interactions assay

2.2.8.1 In vitro pull down interactions between Mth proteins MBP-Cas5-Cas7 and His-Cas8' and E.coli proteins MBP-Cas3 and DNA Topoisomerase I.

Buffers and Solutions

Buffer A: 20mM Tris-HCl pH 8.0, 150mM NaCl, 1mM DTT, 0.1mM PMSF.

Dialysis Buffer 3: 20mM Tris-HCl pH 8.0, 500mM KoAc, 1mM DTT and 40% (w/v) glycerol.

Wash buffer (W): 20mM Tris HCl pH 8.0, 150mM NaCl, 1mM EDTA and 1% (v/v) Tween 20.

Western Blocking Buffer (WBB): 50mM Tris HCl pH 7.6, 150mM NaCl and 0.1% (v/v) Tween 20, supplemented with 5% (w/v) milk powder.

This method is summarised in Cass *et al.* (*144*) and in the appendix; a full explanation of the method used is here. The gene encoding *Cas5* (ORF *Mth*1087) was amplified from *Methanothermobacter thermautotrophicus* (*Mth*) ΔH genomic DNA by PCR, and the gene fragment cloned into pMal-C2x for expression of *Mth* Cas5 fused at its N-terminus to *E. coli* maltose binding protein (MBP-Cas5). MBP tagging of *Mth* Cas5 greatly improved its solubility and stability for expression in *E. coli*. *Cas7* (ORF *Mth*1088) was amplified similarly to *Cas5,* for cloning into pCDF-1b generating a non-tagged Cas7 protein. Co-expression of MBP-Cas5 and Cas7 in *E. coli* strain BL21 Codon Plus was in broth containing additional glucose (0.2 % w/v), protein expression being induced by addition of IPTG (0.5 mM) at OD_{600} between 0.4-0.6. Cas5-Cas7 was purified as a complex through multiple steps on an AKTA-FPLC, followed using SDS-PAGE as described in *Co-purification of MBPCas5 and Cas7*. Briefly, clarified soluble proteins were loaded into a column containing 5ml amylose sepharose. MBPCas5 and Cas7 co-eluted within a gradient of 0-5mM maltose in Buffer A and fractions containing MBPCas5-Cas7 were pooled and loaded onto 5ml Heparin HP column equilibrated in buffer A**.** MBPCas5-Cas7 co-eluted in a gradient of 150-1500mM NaCl, and fractions containing MBPCas5-Cas7 were pooled and dialysed into Dialysis Buffer 3 for storage in aliquots at -80^{o}C.

MBPCas5-Cas7 was used to test for physical interaction with Cas8'. 50µl of amylose resin slurry was equilibrated in 100µl of wash buffer (W) and centrifuged at 700g for 30 sec, supernatant removed and washing repeated five times. 20µg of MBPCas5-Cas7, His_6Cas8' or MBPCas5 and Cas7 and His_6Cas8b' we added to the resin to a final volume of 500µl and end-to-end mixed for 2-4 hours at 4°C. Resin was pelleted as before and washed 3 times as previously. SDS-PAGE disruption buffer was added to resin pellet and boiled. First wash and pellet were analysed via SDSPAGE.

Two identically loaded SDS-PAGE gels were used for Coomassie staining or electro blotting onto PVDF to detect the presence of MBPCas5 or His_6Cas8' proteins *via* their affinity tags. Membranes were incubated overnight at 4°C in Western Blocking Buffer (WBB), before probing each separately with monoclonal antibodies against MBP (NEB), or His_6 (Sigma). Washed membranes were then probed with HRP-conjugated anti-mouse antibody (against His_6) or anti-goat antibody (against MBP) to develop using an ECL detection kit and imaged using FujiFilm LAS300 machine. The exact same method was used for testing the interaction between *E.coli* MBPCas3 or MBP-ΔC-Cas3 and DNA Topoisomerase I.

2.2.8.2 Size exclusion chromatography

For each assay, an analytical Superose 6 size exclusion column was equilibrated in running buffer (20mM Tris-HCl, 150mM NaCl, 1mM DTT, 0.1mM PMSF, 10% (v/v) glycerol); typically three times the void volume of the column (column volumes, CV) was used to equilibrate. The column was then loaded with a diluted 250µl input of molecular weight standards (BioRad), consisting of 5 proteins of known molecular weights. Molecular weight standards generated a standard curve plotting elution volume (V_e) verses protein molecular. This standard is then used as a reference when analysing sample molecular weights.

SE was out on *Mth* Cas3 and the D347A mutant proteins. Each protein was incubated -/+ 10mM ATP and/or Mg^{2+}, and then loaded and eluted from the Superose 6 column. UV absorbance was analysed to predict elution volume and therefore a change in structure or oligomeric state.

This method was also attempted for the interaction of Cas3 from *E.coli* and an interaction observed previously between Cas3 and Topoisomerase I. Combinations of mixed and individual proteins, 100µg of MBPCas3, MBP or His_6TopoI were mixed for

30 minutes at room temperature before loading onto the SE column and eluted. Fractions were collected from this run and then analysed by SDS-PAGE, and then Western Blotting to detect a shift in elution of His$_6$TopoI only when incubated with MBPCas3. This was also carried out with a mutant MBPCas3, that had the uncharacterised C-terminal domain deleted, referred to as ΔCMBPCas3. This protein was treated in the same way as the wild-type protein.

This was used for the *Mth* proteins Cas5, Cas7 and Cas8'. Again combinations of the proteins were loaded onto a Superose 6 SE column and elution profiles were analysed relative to the molecular weight standards. The only difference here is that the buffer for elution was supplemented with 500mM KOAc rather than NaCl.

2.2.8.3 Cross linking

Proteins that are thought to interact were dialysed into a buffer suitable to carry out the cross linking experiment (20mM HEPES pH 8.0, rather than Tris-HCl). 20µg of each protein was then incubated in isolation and with predicted reaction partners with the addition of glutaraldehyde at increasing concentrations (0.0125, 0.125 and 1.25mM). Reaction proceeded for 5 minutes at room temperature and then samples were prepared for SDS-PAGE.

2.2.9 Nucleic acid substrates for *in vitro* biochemical assays

Buffers and solutions

Elution Buffer: 10mM Tris-HCl pH 7.5, 50mM NaCl

2.2.9.1 Radiolabelling of Nucleic acid substrates

Oligonucleotides were purchased from MWG and are listed in the appendix. Labelling of oligonucleotides and their annealing into substrates followed standard methods. Oligonucleotides were resuspended in nuclease-free water to a final concentration of 600ng/µl. Single stranded nucleic acids were 5' end-labelled with ^{32}P from γ[^{32}P]-ATP using T4 polynucleotide kinase (PNK) (NEB) in a standard reaction set up as follows. 300ng of substrate was incubated with a final concentration of 1 x T4 PNK buffer, γ[^{32}P]-ATP (Perkin-Elmer) (approximately equimolar to DNA used) and T4 PNK. Reaction was incubated at 37°C for 30-60 minutes. Labelled oligonucleotide was purified from unincorporated γ[^{32}P]-ATP in BioSpin6 columns (Bio-Rad), using the protocol provided. Prepared columns were loaded with 20-80µl of labelled nucleic acid substrate and separated by centrifugation at 1 x g for 2 minutes. Once purified from unincorporated γ[^{32}P]-ATP, if a single stranded substrate it was now ready for assay. Forming of double stranded of branched and therefore multiple stranded substrates annealing was as follows. 300ng of labelled nucleic acid strand was mixed with 900ng of unlabelled oligonucleotide and incubated in 10mM sodium citrate buffer at 95°C for 10 minutes. The reaction was then allowed to cool to room temperature overnight to facilitate efficient annealing on substrate strands.

2.2.9.2 Purification of labelled nucleic acid substrates

Substrates that consisted of two of more oligonucleotides were purified by electrophoresis through 10% polyacrylamide/ 1x TBE electrophoresis, for 3 hours at 120V and then excision of the appropriate band, detected on photographic film, and elution of DNA by diffusion into Elution Buffer at 4°C, for between 24-72 hours.

2.2.10 Electrophoretic mobility shift assays (EMSA)

Buffers and Solutions

Buffer HB: 100mM Tris-HCl pH 7.5, 10mM DTT, 500µg/ml BSA and 30% (v/v) glycerol.

TBE (Tris Borate EDTA): 89 mM Tris-HCl, 89 mM boric acid, 2 mM EDTA

EMSAs mixed protein(s) with substrate in buffer HB, typically incubated at 44.8°C for 10 min. Reactions were then mixed by pipetting and loaded directly into wells of a gel comprising 7% polyacrylamide in 1x Tris-Borate-EDTA (TBE) buffer. Protein-nucleic acid complexes were separated by electrophoresis at 105V for approximately 170 min in 1 x TBE running buffer, and detected by gel drying and phosphorimaging. Protein-nucleic acid complex formation was quantified compared to a no-protein control, using AIDA software to calculate the percentage of substrate bound, and plotting in Prism to determine binding affinity expressed as K_D. K_D is the disassociation constant for the reversible reaction *M* + *L* ⇌ *ML*, where *M* is macromolecule (protein), *L* is the free ligand (oligonucleotide) and *ML* is the macromolecule-ligand complex (held together by intermolecular interactions, not covalent bonding). K_D can therefore be described as:

$$K_D = \frac{[M]_{eq} \times [L]_{eq}}{[ML]_{eq}}$$

$[M]_{eq}$ can be referred to as [M free]. [M free] = $[M_{total} – ML]$
$[L]_{eq}$ can be referred to as [L free]. [L free] = $[L_{total} – ML]$

$$K_D = \frac{([M_{total} – ML]) \times ([L_{total} – ML])}{[ML]}$$

$$K_D \times [ML] = ([M_{total} – ML]) \times ([L_{total} – ML])$$
$$K_D \times [ML] = [M_{total}][L_{total}] – [L_{total}][ML] – [M_{total}][ML] + [ML]^2$$
$$0 = [ML]^2 – [L_{total}][ML] – [M_{total}][ML] - K_D[ML] + [M_{total}][L_{total}]$$
$$0 = [ML]^2 – ([L_{total}] – [M_{total}] - K_D) [ML] + [M_{total}][L_{total}]$$

Therefore input this into the quadratic equation.
If $ax^2 + bx + c = 0$ then $x = \frac{-b \pm \sqrt{(b^2 – 4ac)}}{2a}$

$$[ML] = \frac{([L_{total}] – [M_{total}] - K_D) \pm \sqrt{(} [L_{total}] – [M_{total}] - K_D)^2 – 4([M_{total}][L_{total}])}{2}$$

2.2.11 Nuclease assays

Buffers and Solutions

Buffer HB: 100mM Tris-HCl pH 7.5, 10mM DTT, 500µg/ml BSA and 30% (v/v) glycerol.

TBE (Tris Borate EDTA): 89 mM Tris-HCl, 89 mM boric acid, 2 mM EDTA.

Stop solution: 2.5% (w/v) SDS, 200mM EDTA and 10mg/ml proteinase K.

Nuclease assays are summarised in Cass *et al*. (*144*) for His$_6$-Cas8' and MBPCas5 and Cas7, and in Rollie *et al*. (*150*) and Ivancic-Bace *et al (268)* for His$_6$-Cas1, described in more detail here.

Mth His$_6$-Cas8' proteins and MBPCas5 and Cas7 were mixed with substrates (2 nM) in Buffer HB supplemented with either 10 mM $MgCl_2$, 5 mM EDTA or nothing and incubated at a range of temperatures between 44.8-65°C for 10 min. Reactions were terminated by addition of 3 µl stop solution and loaded into 10% acrylamide-TBE non-denaturing gels, or 15% polyacrylamide/urea denaturing gels. Gels were dried, imaged and analysed as for EMSAs. *E.coli* His$_6$-Cas1 proteins were assayed in the same manner, except incubation was at 37°C for 30 min.

2.2.12 HeliCase unwinding Assay

Buffers and Solutions

Buffer HB: 100mM Tris-HCl pH 8.5, 10mM DTT, 500µg/ml BSA and 30% (v/v) glycerol.

TBE (Tris Borate EDTA): 89 mM Tris-HCl, 89 mM boric acid, 2 mM EDTA.

Stop solution: 2.5% (w/v) SDS, 200mM EDTA and 10mg/ml proteinase K.

RecG or PriA proteins were incubated with various forked substrates (2nM) in buffer HB supplemented with 5 or 10mM $MgCl_2$ and ATP for 10 minutes at 37°C. Reactions were terminated by addition of 3µl stop solution and loaded into 10% acrylamide-TBE non-denaturing gels. Gels were dried, imaged and analysed as for EMSAs.

2.2.13 Strand Exchange Assay

Buffers and Solutions

Buffer HB: 100mM Tris-HCl pH 7.5, 10mM DTT, 500µg/ml BSA and 30% (v/v) glycerol.

TBE (Tris Borate EDTA): 89 mM Tris-HCl, 89 mM boric acid, 2 mM EDTA.

Stop solution: 2.5% (w/v) SDS, 200mM EDTA and 10mg/ml proteinase K.

MBPCas5 and Cas7 or His$_6$Cas7 were mixed with a DNA duplex with one strand labelled, and another oligonucleotide that if annealed would displace the labelled DNA strand. Protein(s) were mixed with the substrate in buffer HB and incubated for 30mins at 44.8°C. Reactions were terminated with 3µl Stop solution and analysed by 10% TBE non-denaturing gels. Gels were dried, imaged and analysed as for EMSAs.

2.2.14 ATPase assay

Buffers and Solutions

ATPase Buffer: 100mM Tris HCl pH 8.5, 10mM DTT, 500µg/ml BSA and 30% (v/v) glycerol, 10mM $MgCl_2$ and 10mM ATP.

Stop solution: 0.0324% (w/v) malachite green, 1% (w/v) ammonium molybdate and 1M NaCl (filtered).

Colour stability solution: 34% (w/v) Citric acid

Malachite green assays were used to measure ATP hydrolysis through the detection of liberated phosphate. 800µl reactions contained 10mM $MgCl_2$ and 10mM ATP and 0-1000nM Cas8b protein, and were incubated at 45^{o}C for 30min, supplemented with 200ng of either ssDNA (crDNA1) or ssRNA (crRNA1). 100µl stop solution was added and incubated for 5 mins followed by addition of 100µl colour stability solution and incubation at RT for 30 mins. Solutions were transferred to cuvettes and absorbance measured at 660nm and corrected against a zero protein blank. Phosphate liberation subsequently quantified, by comparison to a standard curve generated using 0-100µM K_2PO_4.

2.2.15 Transesterification Assay

Buffers and Solutions

Buffer HB: 100mM Tris HCl pH 8.5, 10mM DTT, 500µg/ml BSA and 30% (v/v) glycerol.

TBE (Tris Borate EDTA): 89 mM Tris.HCl, 89 mM boric acid, 2 mM EDTA.

Stop solution: 2.5% (w/v) SDS, 200mM EDTA and 10mg/ml proteinase K.

Initial Transesterification (TE) reactions were carried out on DNA flap substrates with His$_6$Cas1 and Cas1HisStrep. Protein was incubated with the flapped substrate in buffer HB (needed to be pH8.5 as pH7.5 produced no activity), for 30 minutes at 37°C. Reactions were stopped with 3µl of stop solution and loaded into 15% polyacrylamide/urea denaturing gels, 170V for 55 minutes.

This assay was then developed to include Cas2 in the reaction mixture with Cas1 and also different substrates were used. Other replication fork structures were assembled as described in Nucleic acid substrates for *in vitro* biochemical assays. Further development of the TE reactions involved preincubation of the DNA substrates with RecG, PriA or SSB proteins, these conditions were identical to those used in the heliCase unwinding assays, subsequently Cas1 was added to these mixtures and the typical TE reaction was continued. In each of these examples reactions were and loaded into 15% polyacrylamide/urea denaturing gels and electrophoresed for 170V for 55 minutes.

2.2.16 SpIN (Spacer integration) assay

Buffers and Solutions

Buffer SpIN: 100mM Tris HCl pH 8.5, 10mM $MgCl_2$, 1mM KCl, 10mM DTT, 500µg/ml BSA and 30% (v/v) glycerol.

TBE (Tris Borate EDTA): 89 mM Tris.HCl, 89 mM boric acid, 2 mM EDTA.

Stop solution: 2.5% (w/v) SDS, 200mM EDTA and 10mg/ml proteinase K.

Short radiolabelled duplex DNA substrates were created as described above. Substrates consisted of 29bp spacer sequence flanked by + or - PAM and AAM motifs. Cas1 and Cas2 (100 and 50nM respectively) were pre-incubated on ice for 15 minutes. Spacer substrate was then added to the protein mix and further incubated on ice for 30 minutes. pJRW2 (plasmid containing *E.coli* CRISPR array and leader sequence) was added to a final reaction volume of 20µl and incubated at 37°C for 5, 10, 15, 20, 25 and 30 minutes. Reactions were then terminated by addition of STOP solution and samples loaded onto 1% (w/v) agarose gels. 10cm gels were electrophoresed for 1 hour 30 minutes at 100V.

3 Chapter 3: Analysis of Cas proteins from *Methanothermobacter thermautotrophicus* and *Escherichia coli*

3.1 Overview

This chapter details the plasmid constructs, protein expression profiling, protein purification and biochemical assays of Cas proteins from *Methanothermobacter thermautotrophicus* and *Escherichia coli*. All the produced plasmids were confirmed by restriction digestion and direct DNA sequencing. Cloning for protein expression also included engineering N- or C-terminal tags as described by the following abbreviations: His_6 (hexahistidine), Strep (StrepTactin) and MBP (Maltose Binding Protein). Tags were utilised for affinity purification of Cas and other proteins using AKTA FPLC systems. Additional methods of purification were employed to remove remaining contaminants where possible. After purification, the identity of each protein was verified by mass spectrometry (thanks to David Tooth and the BSAU, Faculty of Medicine and Health Sciences, QMC, University of Nottingham).

Table 8. Nomenclatures used for the CRISPR/Cas genes and proteins of *Mth* and *E.coli*. other synonyms exist but these are not used in this investigation for clarity.

Gene Family name	*Mth* Gene and protein names	*E. coli* gene and protein names
Cas1	*Mth* 1084, Cas1	Cas1
Cas2	*Mth* 1083, Cas2	Cas2
Cas3	*Mth* 1086, Cas3	Cas3
Cas4	*Mth* 1085, Cas4	-
Cas5	*Mth* 1087, Cas5	CasD, Cas5
Cas6	*Mth* 1091, Cas6	CasE, Cas6e
Cas7	*Mth* 1088, Cas7	CasC, Cse4
Cas8	*Mth* 1090, Cas8’ and *Mth* 1089, Cas8’’	CasA, Cse1 and CasB, Cse2

3.2 Cloning of *Cas* genes from *Methanothermobacter thermautotrophicus* and purification of the encoded proteins

3.2.1 Molecular cloning of *Cas* genes

A summary of *Mth Cas* genes studied is presented diagrammatically within the *Mth* CRISPR-Cas system in Figure 3-1. *Mth* Cas genes are encoded by ORFs 1076-1091.ORFs 1076-1082 encode the Type III-A Csm complex proteins, ORFs 1083 (*Cas2*), 1084 (*Cas1*) and 1085 (*Cas4*) encode the adaptation machinery, ORF 1086 encodes the interference (*Cas3*) protein and ORFs 1087-1091 encode the Type I-H Cascade proteins (*Cas5*, *Cas7, Cas8''*, *Cas8'* and *Cas6*). Each ORF was cloned into one of the following vectors: pET14b, pET-Duet, pET-ACYC, pMal-C2x, pCDF-1b or pQE-HisStrep1. Details of all the relevant cloning carried out, including those of other members of the laboratory applicable to this study, are given in Table7. Specifically, plasmids titled pEB or pJLH were kind gifts from Edward Bolt and Jamieson Howard, respectively. All the primers used can be found in the appendix along with plasmid construction plans from PCR or plasmid excision and ligations.

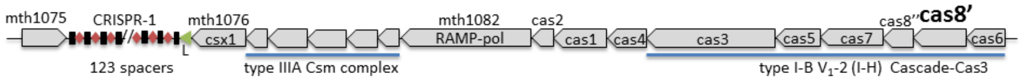


Figure 3-1. Schematic of the gene neighbourhood of the CRISPR-Cas system of *Methanothermobacter thermautotrophicus*, including *Cas* genes and CRISPR-1 array. This CRISPR system is a fusion of Type I-H and III-A found downstream of CRISPR-1 array encoding: Cas5, Cas7, Cas8', Cas8'' and Cas6 for the type I-H archaeal Cascade, Cas3 as the heliCase/nuclease that degrades invading DNA, Cas1 and Cas2 (and Cas4) as the adaptation machinery of both subtypes and *Mth* 1076-1082 encoding the Csm (Type III-A) interference complex (*144*).

Table 9. Summary of plasmid constructs generated and used from *Methanothermobacter thermautotrophicus* genes. pSDC (Simon David Cass), pEB (Edward Bolt), pJLH (Jamieson Leyland Howard).

Plasmid construct name	**Gene**	**ORF Number**	**Plasmid Vector**	**Affinity protein Tag**
pSDC25	Cas5	1087	pMal-C2x	N-MBP
pEB374	Cas5	1087	pET14-b	N-His$_6$
pSDC13	Cas6	1091	pET14-b	N-His$_6$
pSDC27	Cas6	1091	pMal-C2x	N-MBP
pEB388	Cas7	1088	pET14-b	N-His$_6$
pSDC29	Cas7	1088	pQE-HisStrep1	C-His$_6$-Strep
pSDC31	Cas7	1088	pCDF-1b	None
pEB367	Cas8′	1090	pT7-7	None
pEB389	Cas8′	1090	pET14-b	N-His$_6$
pEB403	Cas8′ (K68A)	1090	pT7-7	None
pSDC41	Cas8′ (K68A)	1090	pET14-b	N-His$_6$
pEB404	Cas8′ (K117A)	1090	pT7-7	None
pSDC22	Cas8′ K117A)	1090	pET14-b	N-His$_6$
pSDC39	Cas8′ (D151G)	1090	pET14-b	N-His$_6$
pSDC43	Cas8′ (N153A)	1090	pET14-b	N-His$_6$
pSDC40	Cas8′ (E155A)	1090	pET14-b	N-His$_6$
pSDC43	Cas8′ (N536A)	1090	pET14-b	N-His$_6$
pSDC38	Cas8′ (S540A)	1090	pET14-b	N-His$_6$
pSDC44	Cas8′ (A540G)	1090	pET14-b	N-His$_6$
pSDC45	Cas8′ (L542A)	1090	pET14-b	N-His$_6$
pEB383	Cas8′′	1089	pT7-7	None
pSDC5	1076		pETDuet-1	N-His$_6$
pEB359	Cas3	1086	pET22-b	None
pSDC15	Cas1	1084	pCDF-1b	N-His$_6$
pSDC16	Cas2	1083	pCDF-1b	N-His$_6$
pSDC17	Cas1 + Cas2	1084 + 1083	pCDF-1b	N-His$_6$
pJLH7	Cas1	1084	pET14-b	N-His$_6$
pJLH9	Cas2	1083	pET14-b	N-His$_6$

3.2.1.1.1 Summary

Optimisation of PCR, restriction digests and ligations were required for some of the constructs; this included annealing temperatures, extension times, digest length and ligation temperature and time. Integral for progress of this study was obtaining recombinant soluble archaeal Cascade proteins Cas5, Cas6, Cas7, Cas8' and Cas3.

3.2.2 Recombinant over-expression of *Mth* Cas proteins

Mth Cas proteins were expressed heterologously in *E.coli* host strains listed in Table 4. All the genes cloned from *Mth* were in vectors with T7 RNA polymerase promoters and Lac operators, therefore, strains of *E.coli* engineered to encode IPTG inducible T7 RNA polymerase for protein expression. This strategy was chosen as Cas3 and Cas8' have already been produced in our laboratory through this method (*143, 269*).

3.2.2.1 Pilot experiments for testing protein expression and solubility

Each *Mth* Cas protein was tested for over-expression and solubility in *E.coli* to find the optimal conditions for protein purification. For brevity, only the procedure and outcomes for Cas1 protein are shown, as a representative of the tests that were carried out for each Cas protein. Cas protein expression was trialled in *E. coli* strains BL21AI and BL21Codonplus (C+), followed by assessment of protein solubility. Cas1 protein over-expression was observed in both BL21AI and BL21C+ (Figure 3-2i and ii). BL21C+ was the expression strain carried forward to assess its solubility in this strain as slightly higher levels of protein were detected, the predicted Cas1 protein, indicated with an arrow, is clearly detectable in the S (soluble) 1 (Figure 3-2iii).

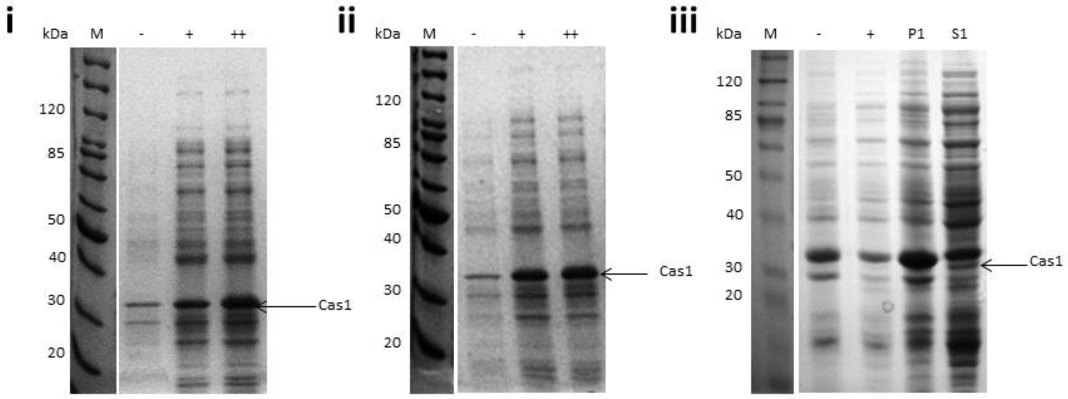


Figure 3-2. SDS-PAGE analysis of Cas1 protein (*Mth* ORF1084) in pilot protein over-expression and solubility of lysed cell extracts of *E. coli*. In panel i (strain Bl21 AI) and panel ii (strain C+) Cas1 over-expression is compared in un-induced cells (-) and IPTG induced cells grown for either 2 hours (+) or 4 hours post-induction (++) at 37°C. An arrow points to predicted Cas1 migrating at approximately 35 kDa (predicted MW= 33174 Da). In panel iii, BL21 C+ cells were lysed to compare soluble fraction (S1) with insoluble fraction (P1), identifying that Cas1 was present in S1, as indicated by the arrow.

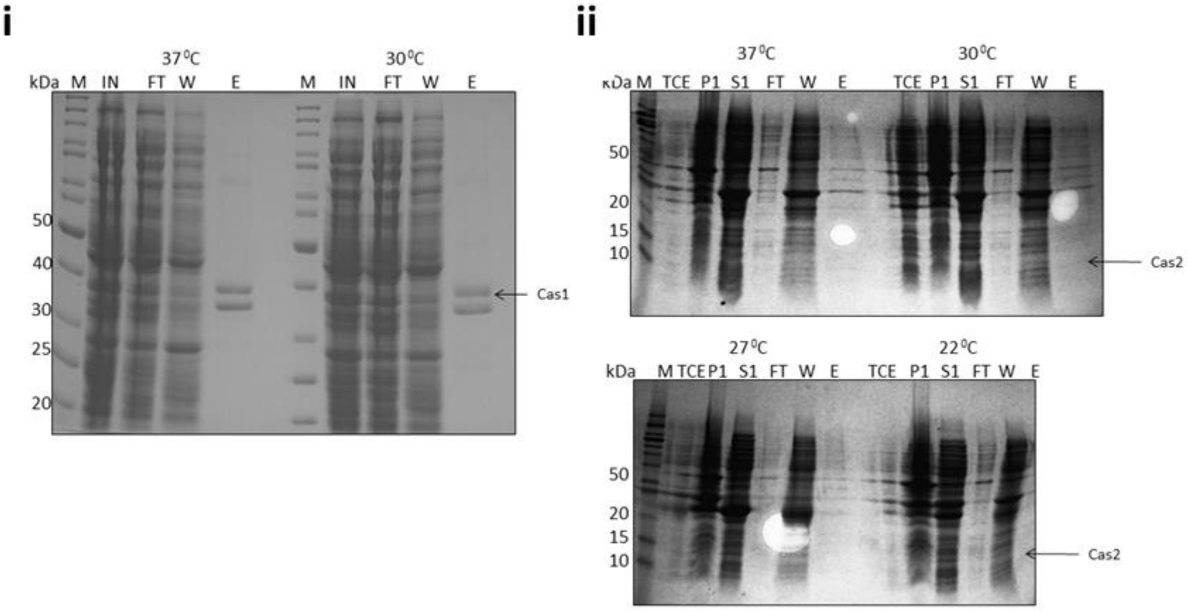

i
37°C
30°C
kDa M IN FT W E M IN FT W E
50
40
30
25
20
Cas1
ii
37°C
30°C
kDa M TCE P1 S1 FT W E TCE P1 S1 FT W E
50
20
15
10
Cas2
27°C
22°C
kDa M TCE P1 S1 FT W E TCE P1 S1 FT W E
50
20
15
10
Cas2

Figure 3-3. SDS-PAGE analysis of Cas1 (*Mth* ORF1084) and Cas2 (*Mth* ORF1083) proteins from pilot purifications by His Gravi-Trap. In panel i, Cas1 pilot purification is compared from biomass induced with IPTG at 30°C and 37°C. Lysed and centrifuged soluble fraction was loaded onto His Gravi-Trap (IN) and compared to flow through (FT), wash (W) and elution (E) for Cas1. An arrow points to Cas1 protein doublet detected in the elute migrating at approximately 35kDa. In panel ii, Cas2 pilot purification is compared from biomass induced with IPTG at 30°C, 37°C, 27°C and 22°C. Lysed and centrifuged soluble fraction was loaded onto His Gravi-Trap (TCE) and compared to pellet (P1), soluble material (S1), flow through (FT), wash (W) and elution (E) for Cas2. An arrow points to where Cas2 protein would be expected to elute of the column, migrating at approximately 9kDa.

After attaining soluble protein, different induction temperatures were tested to potentially increase yield. In particular, lower temperatures slow translation speed and so allow rare codon matching and allow correct disulphide bond formation. For Cas1 and Cas2 proteins, initial screens of expression were undertaken at 30°C and

37^{o}C, (Figure 3-3i and ii). Additional lowering of the incubation temperature of growing *E. coli* cells to 27^{o}C and 22^{o}C was also tested to help obtain soluble Cas2 protein. Different growth conditions were tested, even if there was no detectable soluble protein purification was carried out by His-Gravity Trap. Following expression cell biomass was collected by centrifugation. After clarification (sonication and centrifugation) soluble fractions were subjected to His-gravity Trap purification. First, the soluble fraction in 0mM Imidazole buffer was applied to the column whereby all unbound protein was collected as the flow through fraction (FT). Next, non-specifically binding proteins were released from the column with a low imidazole (5mM) wash (W). Finally, specifically bound proteins were eluted in high (500mM) imidazole buffer (E). A doublet was seen from Cas1 purification at 30°C and 37°C, this doublet likely being either a degradation product of Cas1, or a nickel rich contaminating protein such as SlyD (*270*), a known contaminant of His-Tagged protein purification following expression in bacteria. No Cas2 could be purified from biomass grown at any temperatures tested here, with arrows pointing to where purified Cas2 protein would have been expected (Figure 3-3ii).

3.2.2.2 Further overexpression and purification tests

The procedure detailed for Cas1 was also implemented on other Cas proteins that were to form part of this study. This is summarised for proteins Csx1, Csm2, Csm3, Csm4, Cas2, Cas1, Cas4, Cas5 and Cas7 (ORFs 1076, 1077, 1078, 1080, 1083, 1084, 1085, 1087 and 1088) in Figure 3-4. In each Case un-induced (-) and induced (+) total cell extracts were compared after cell lysis by sonication and centrifugation to ascertain the level of expression. Expression was detected in BL21C+ cell extracts of each protein except Csx1 and Cas2. The expressing ORFs of interest were further analysed to assess protein solubility. However, this revealed most of the proteins expressed in this screen were in fact insoluble, (Figure 3-5). Therefore, further investigations were focused on *Mth* ORFs 1086-1091 (*Cas3*-Cas6), encoding a type I-H CRISPR interference system, given they could be successfully purified, (see next section).

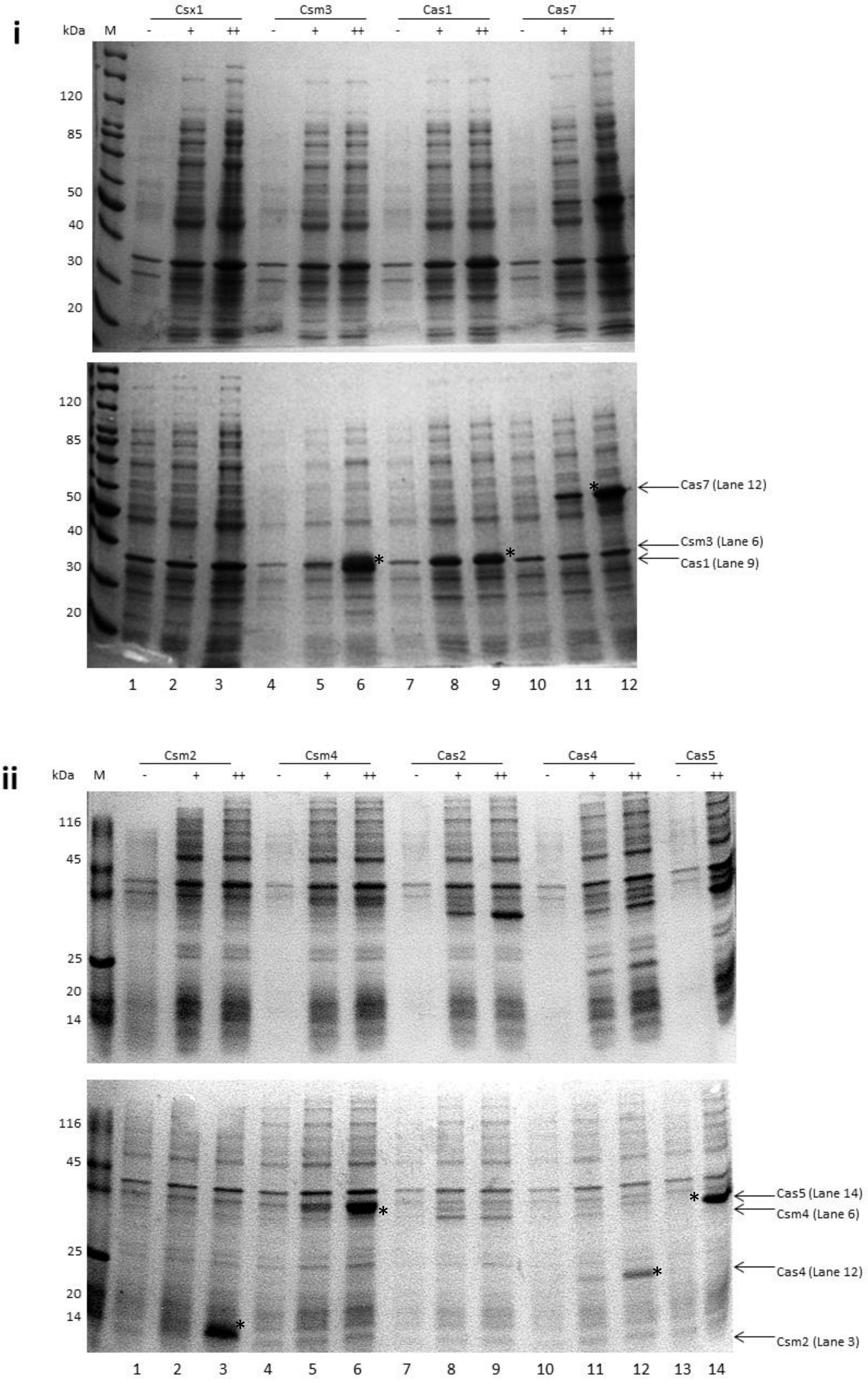


Figure 3-4. Composite SDS-PAGE analysis of *Mth* Cas proteins (*Mth* ORF1076-1091) in pilot protein over-expression studies. In panels i and ii each gene construct was expressed in BL21AI and BL21C+ (top and bottom panel respectively). Cas proteins over-expression is compared in un-induced cells (-) and IPTG induced cells grown for either 2 hours (+) or four hours post-induction (++). An arrow points to predicted Cas proteins migrating at approximate Relative Molecular Masses (RMMs). Proteins that expressed are identified and indicated with arrows and asterisks.

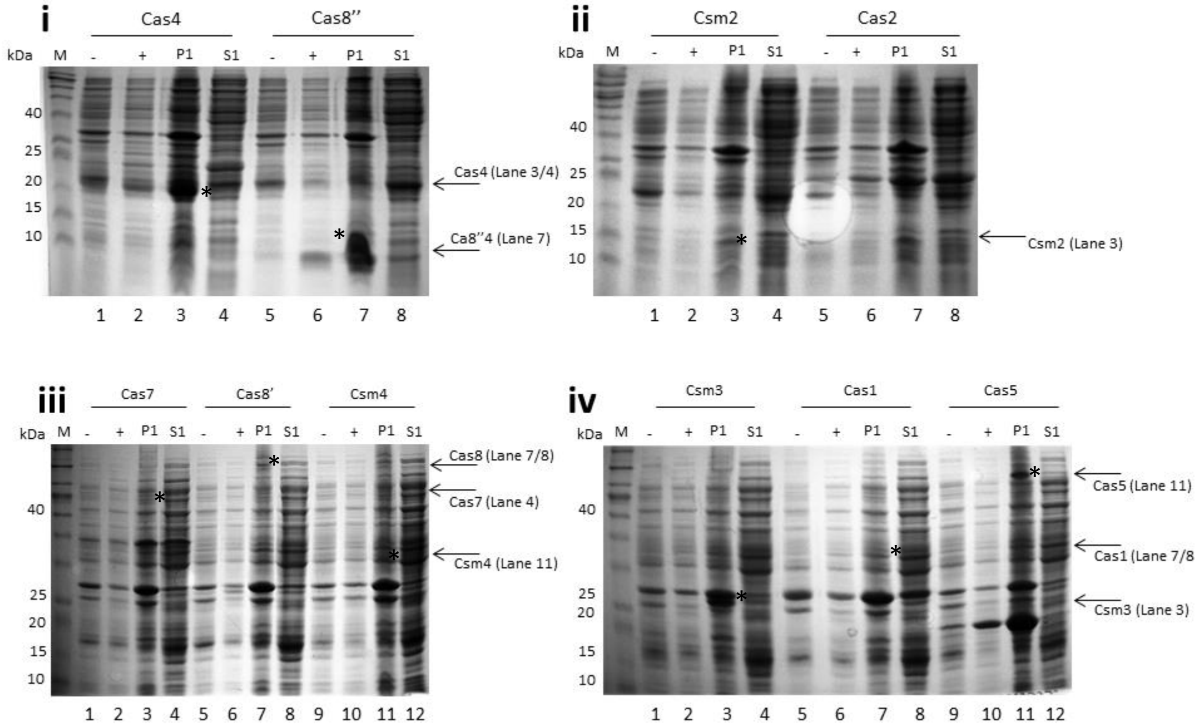


Figure 3-5. Composite SDS-PAGE analysis of *Mth* Cas proteins (*Mth* ORF1076-1091) in protein solubility screen. In each panel the indicated gene was expressed in BL21C+ and compared in un-induced cells (-), IPTG induced cells grown for 2 hours (+) and pellet (P1) and supernatant (S1) after sonication and centrifugation. An arrow points to predicted Cas proteins migrating at approximate RMMs. Proteins that expressed are identified and indicated with arrows and asterisks.

3.2.3 Purification of *Methanothermobacter thermautotrophicus* proteins

3.2.3.1 Cas8' and Cas7

Cas7 and Cas8' are predicted to form part of some Cascade complexes that are essential for interference during CRISPR-Cas immunity in bacteria and archaea. Cas8' (ORF 1089) and Cas7 (ORF 1088) have established protocols for expression and purification (*269*), which were adapted as described below.

His$_6$-Cas8' was purified *via* Ni^{2+} Chelating Column within an imidazole gradient (0-500mM), (Figure 3-6i). Peak fractions were analysed by SDS-PAGE and selected fractions dialysed to remove high imidazole and NaCl. The protein was further purified and concentrated *via* Heparin within a NaCl gradient (100-1500mM), (Figure 3-6ii). Peak fractions were processed and dialysed into storage buffer (containing 35-40% glycerol, which disrupts the ice crystal lattice [cryoprotectant] and depresses freezing temperatures, reducing protein degradation in storage). All Cas8' mutants generated from cloning were also expressed and purified using the same system outlined here. Fermentation growth of cells expressing Cas8’ was attempted to generate sufficient protein for crystal trials. Cas8’ was purified with an additional size exclusion chromatography (SEC) between the Chelating and Heparin columns. Once purified, each Cas8’ protein was analysed by SEC to confirm monomer peak elution, in agreement with previous methods (*269*), (Figure 3-6iii).

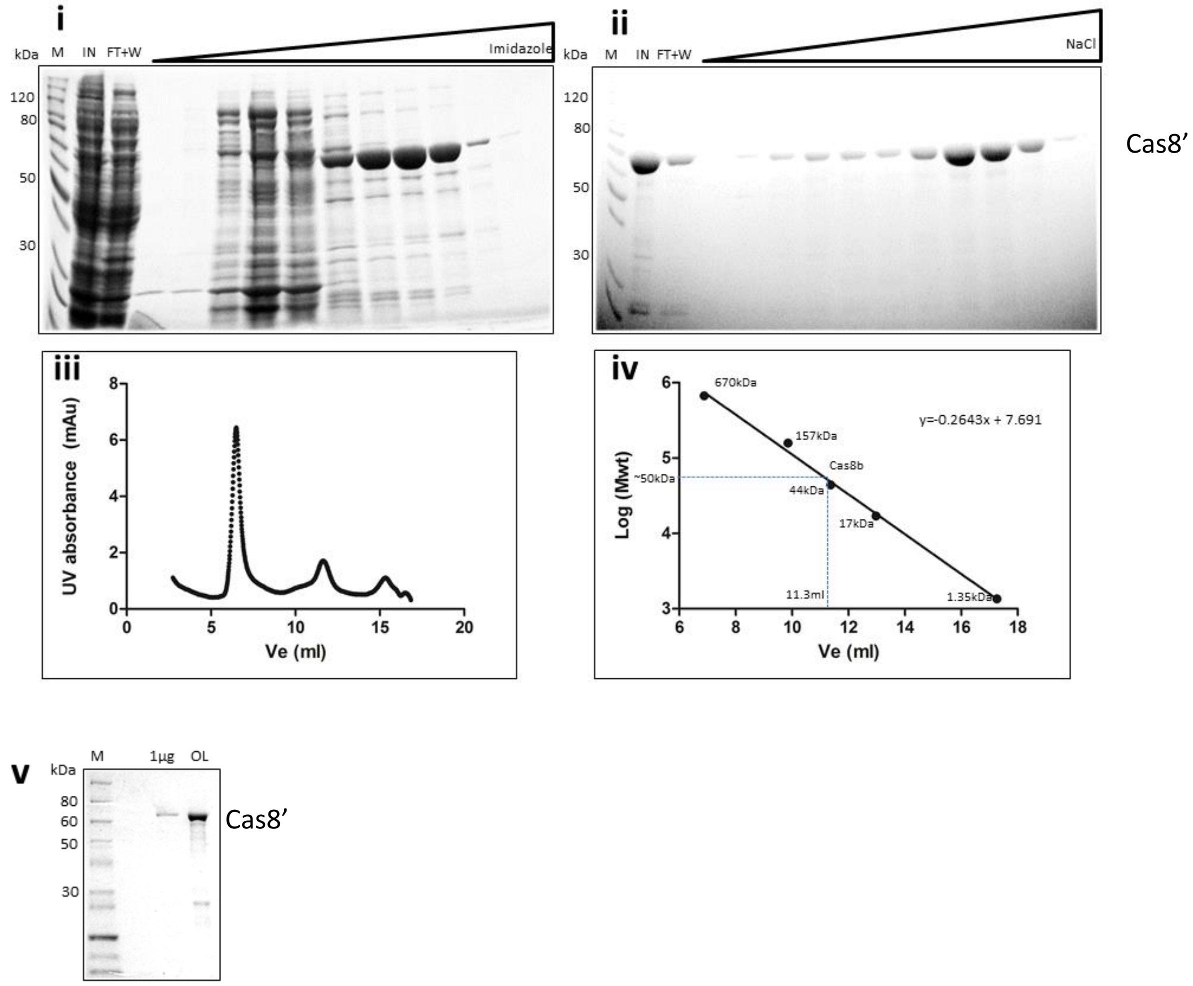


Figure 3-6. SDS-PAGE analysis of the purification of His$_6$Cas8’ expressed in *E.coli* BL21C+. (i) After sonication and centrifugation soluble material (IN) was subjected to Affinity chromatography (HisTrap) and fractions were compared for Cas8’ contents, flow through (FT), wash (W) and UV trace peaks during imidazole gradient (0-500mM). (ii) Cleanest fractions containing Cas8’ were pooled and dialysed to remove high salt before being subjected to a second affinity column (Heparin) and fractions were compared for Cas8’ contents; flow through (FT), wash (W) and UV trace peaks during NaCl gradient (150-1500mM). Pooled fractions were dialysed for storage and subjected to size exclusion chromatography (Superdex 260/60). (iii) UV-trace from analytical size exclusion chromatography plotting UV absorbance against elution volume during a 1column volume (~23ml) elution. Elution profile was compared to molecular weight standards to approximate the size of the purified protein shown in v.

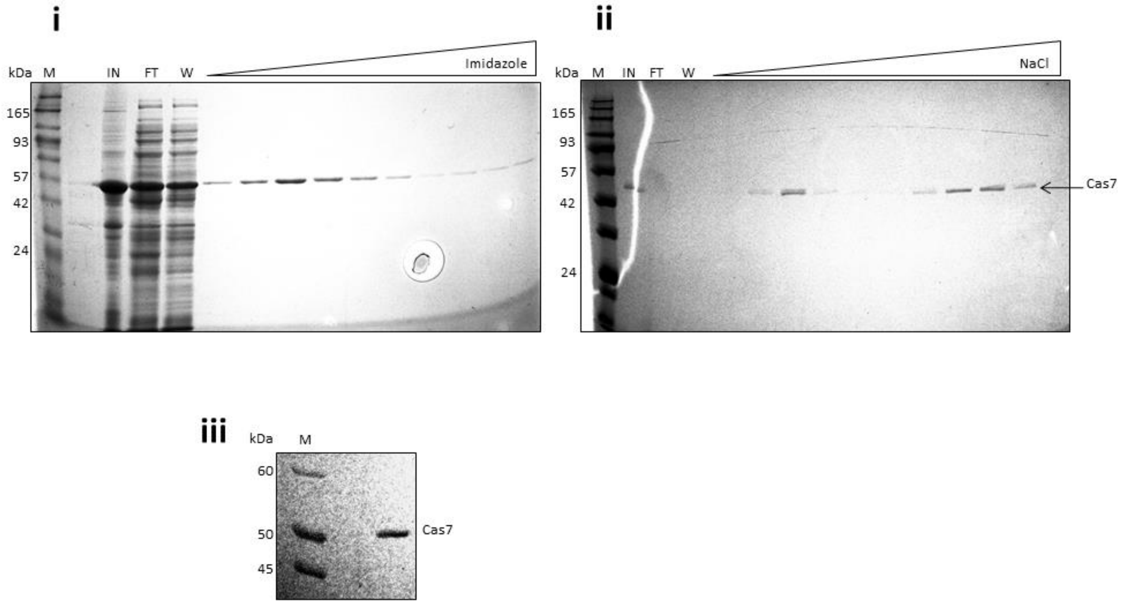


Figure 3-7. SDS-PAGE analysis of the purification of His$_6$Cas7 expressed in *E.coli* BL21C+. (i) After sonication and centrifugation soluble material (IN) was subjected to affinity chromatography (His Trap) and fractions were compared for Cas7 contents; flow through (FT), wash (W) and UV trace peaks during imidazole gradient (0-500mM). (ii) Cleanest fractions containing Cas7 were pooled and dialysed to remove high salt before being subjected to a second affinity column (Heparin) and fractions were compared for Cas7 contents, flow through (FT), wash (W) and UV trace peaks during NaCl gradient (150-1500mM). (iii) Final purified protein prepared for storage.

His$_6$Cas7 was overexpressed and purified in the same manner as Cas8' (Figure 3-7). However, elution of His$_6$Cas7 from the heparin column was detected in two peaks. This is a result of Cas7 forming two possible species with different affinities for the heparin matrix, possibly through formation of a multimer that alters affinity binding to the heparin matrix.

3.2.3.2 Cas5 and Cas6

Cas5 and Cas6 are the two remaining predicted elements of the *Mth* type I-H archaeal Cascade. Neither *Mth* Cas5 nor Cas6 proteins have been previously investigated although analogous proteins CasD and CasE in *E.coli* have been found to be essential for CRISPR interference.

Successful Cas5 and Cas6 over-expression was achieved from pET vectors (Figure 3-4), however, all protein were deemed insoluble. Although both Cas5 and Cas6 could be solubilised in 6M urea, re-folding (through iterative dialysis to remove urea) proved difficult due to each protein aggregating again into insoluble material at lower urea concentration. As an alternative MBP tagged Cas5 and Cas6 were tested, as it is a common method used to increase protein solubility and stability in *E. coli* expression systems. MBP-Cas6 expressed as a soluble protein and was purified via affinity purification through amylose sepharose, eluting within a maltose gradient (0-3.0 mM). Further purification through Heparin resin and elution within a NaCl gradient (150-1500mM) generated pure MBP-Cas6 (Figure 3-8). On the other hand, attaining soluble MBP-Cas5 required its co-expression with untagged Cas7, which presumably stabilised Cas5 against aggregation through Cas5-Cas7 complex formation. The simultaneous expression of MBP-Cas5 and Cas7 generated enough soluble material to successfully purify Cas5 and Cas7, (Figure 3-9).

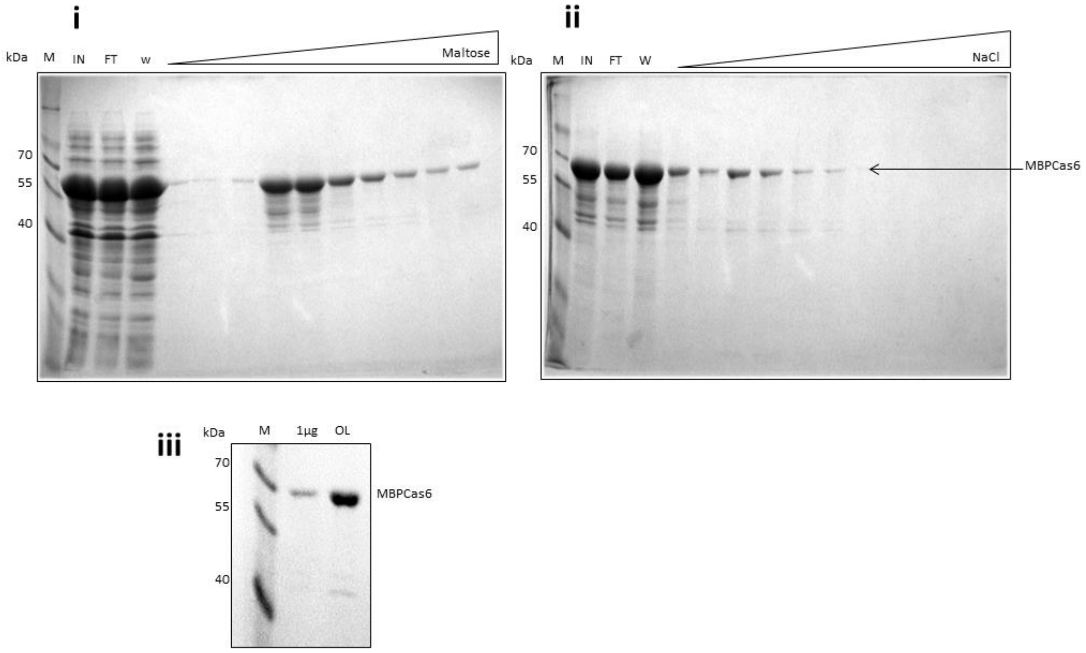


Figure 3-8. SDS-PAGE analysis of the purification of MBP-Cas6 expressed in *E.coli* BL21C+. (i) After sonication and centrifugation soluble material (IN) was subjected to affinity chromatography (Amylose) and fractions were compared for Cas6 contents, flow through (FT), wash (W) and UV trace peaks during maltose gradient (0-30mM). (ii) Cleanest fractions containing Cas6 were pooled and subjected to a second affinity column (Heparin) and fractions were compared for Cas6 contents, flow through (FT), wash (W) and UV trace peaks during NaCl gradient (150-1500mM). (iii) Final purified MBP-Cas6 protein prepared for storage.

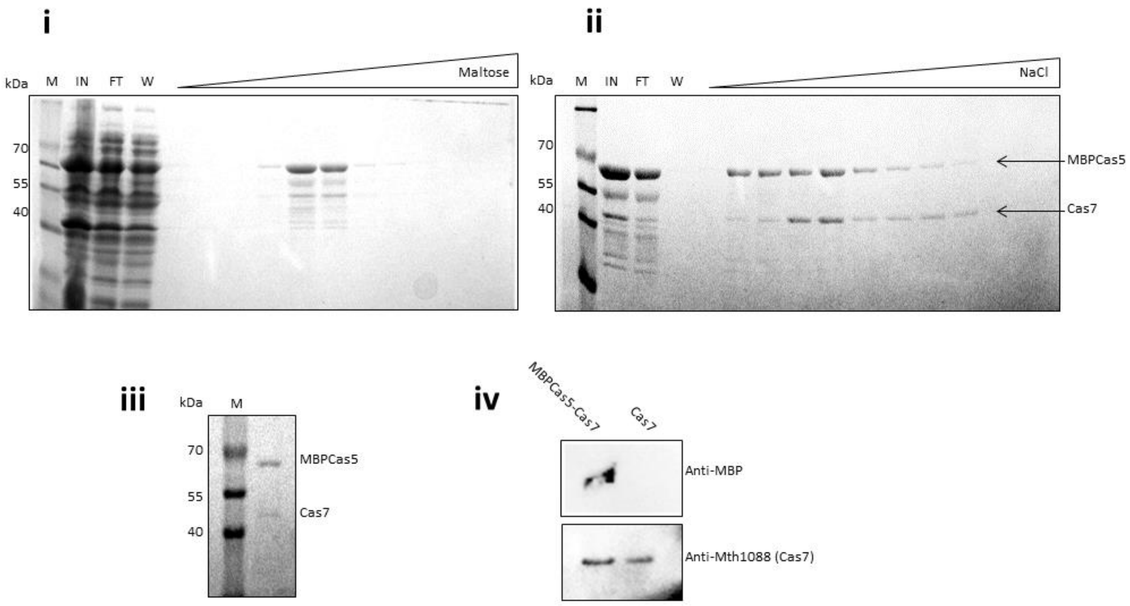


Figure 3-9. SDS-PAGE analysis of the purification of MBPCas5-Cas7 co-expressed in *E.coli* BL21C. (i) After sonication and centrifugation soluble material (IN) was subjected to affinity chromatography (Amylose) and fractions were compared for Cas5-Cas7 contents, flow through (FT), wash (W) and UV trace peaks during maltose gradient (0-30mM). (ii) Cleanest fractions containing Cas5-Cas7 were pooled and subjected to a second affinity column (Heparin) and fractions were compared for Cas5-Cas7 contents, flow through (FT), wash (W) and UV trace peaks during NaCl gradient (150-1500mM). (iii) Final purified MBP-Cas5 and Cas7 proteins prepared for storage. (iv) Western blot analysis of MBPCas5 and Cas7 compared to His$_6$Cas7, probed with anti-MBP and anti-Cas7 (*Mth* 1088) antibodies.

3.2.3.3 Purification of Mth Cas3 protein

Purification of Cas3 (ORF 1086) was modified from the method in Howard *et al*, 2011. Figure 3-10 follows the published method and Figure 3-11 details the outcome of the modified protocol. In both Cases, supernatant from lysed and clarified *E.coli* cells over-expressing Cas3 was passed through heparin to remove contaminating DNA interacting proteins (such as nucleases) followed by a series of hydrophobic and anion exchange columns (Q-sepharose, phenyl-sepharose and DEAE-sepharose). The optimised purification involved elongation of the Q-sepharose wash step, collection of smaller fractions and proceeding to the next step with only fractions containing Cas3 protein at optimal yield and homogeneity.

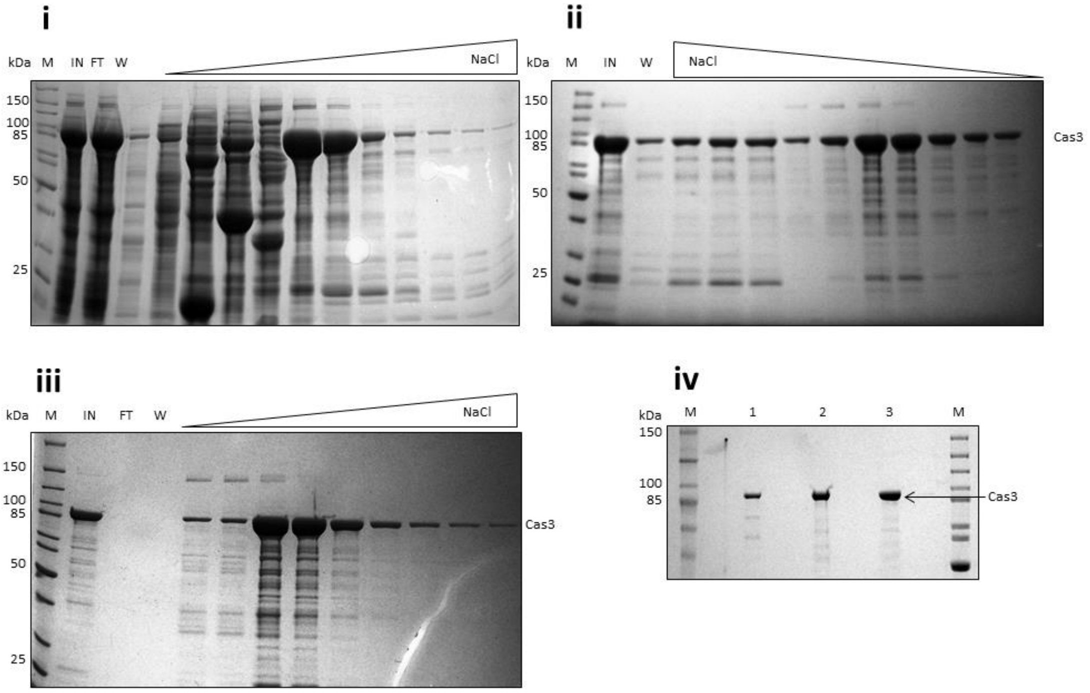


Figure 3-10. SDS-PAGE analysis of the purification of Cas3 expressed in *E.coli* BL21C+. (i) After sonication and centrifugation soluble material (IN) was subjected to anion exchange chromatography (Q-sepharose) and fractions were compared for Cas3 contents, flow through (FT), wash (W) and UV trace peaks during NaCl gradient (150-1500mM). (ii) Cleanest fractions containing Cas3 were pooled and subjected to hydrophobic interaction chromatography (HIC) (Phenyl-sepharose) and fractions were compared for Cas3 contents, flow through (FT), wash (W) and UV trace peaks during NaCl gradient (1500-0mM). (iii) Cleanest fractions containing Cas3 were pooled and subjected to a second anion exchange chromatography column (DEAE sepharose) and fractions were compared for Cas3 contents, flow through (FT), wash (W) and UV trace peaks during NaCl gradient (150-1500mM). Pooled fractions were dialysed for storage of Cas3 wild-type (iv, lane 1), H66G (iv, lane 2) and D347A (iv, lane 3).

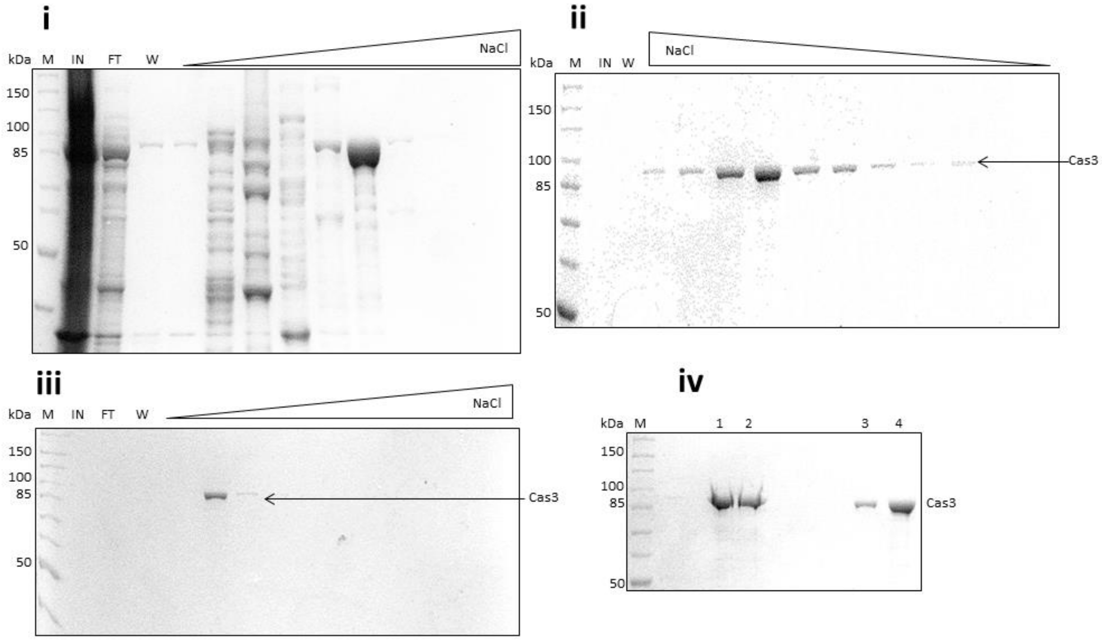


Figure 3-11. SDS-PAGE analysis of the optimised purification of Cas3 expressed in *E.coli* BL21C+. (i) After sonication and centrifugation, soluble material (IN) was subjected to anion exchange chromatography (Q-sepharose) and fractions were compared for Cas3 contents, flow through (FT), extended wash (W) and UV trace peaks during NaCl gradient (150-1500mM). (ii) Cleanest fractions containing Cas3 were pooled and directed subjected to hydrophobic interaction chromatography (HIC) (Phenyl-sepharose) and fractions were compared for Cas3 contents, flow through (FT), wash (W) and UV trace peaks during NaCl gradient (1500-0mM). (iii) Cleanest fractions containing Cas3 were pooled and directed subjected to a second anion exchange chromatography column (DEAE sepharose) and fractions were compared for Cas3 contents, flow through (FT), wash (W) and UV trace peaks during NaCl gradient (150-1500mM). Pooled fractions were dialysed for storage of Cas3 wild-type, 1µg and overload (iv, lanes 1 and 2), and D347A (iv, lanes 3 and 4).

3.2.3.3.1 Summary

Purification of *Mth* Cas5, Cas6, Cas7 and Cas3 proteins allowed the analysis of archaeal CRISPR-Cas interference reactions *in vitro*. The high yield of Cas8’ protein obtained enabled preliminary crystal trials, but at the time of writing these trials have been unsuccessful.

3.3 Biochemical analysis of *Methanothermobacter thermautotrophicus* Cascade proteins

3.3.1 Cas5, Cas6 and Cas7 preliminary activity tests

Cas7 (Csa2, *E.coli* equivalent) proteins perform helical ‘wrapping’ of crRNA forming the Cascade core (*213*). Indeed, Cas5 (CasD) atomic resolution structures revealed that essential contacts are made with the crRNA 5’ handle and the PAM region of the invaded duplex (*128*). Therefore, purified Cas7 was analysed by EMSA to determine functional parameters for RNA and DNA binding. His_6Cas7 was titrated into ssDNA, ssRNA and dsDNA in buffers with and without Mg^{2+}. The presence of Mg2+ marginally improved the binding of Cas7 as evidenced by the detection of specific in gel complexes at a lower Cas7 concentration (Figure 3.12i middle lanes). However, this observation was not reproduced in all experiments (Figure 3-12i right hand lanes) indicating that the effect of Mg^{2+} on Cas7 interaction with ssDNA is not a strong one. Importantly, the salt used during Cas7 purification provedto be a critical determinant in crRNA binding. Specifically, only Cas purified and assayed in Potassium acetate (KOAc) was able to form specific in gel complexes with ssRNA (right hand panel, Figure 3-12ii), when compared to reactions in Sodium chloride (NaCl) where none were detected (left hand panel, Figure 3-12ii). The importance of acetate salts was also apparent in dsDNA binding. Cas7 bound dsDNA 10x more effectively in potassium acetate (Figure 3-12iii, right panel) with the reaction saturating at 10nM rather than 100nM seen with Cas7 in NaCl (Figure 3-12iii, left hand panel).

Cascade complexes catalyse interference by invading crRNA into a homologous duplex DNA thus creating an R-loop. The stability of an interference R-loop is derived from RNA-DNA hybridisation over at least 30 base-pairs, the exact length depending on the length of a spacer sequence transcribed from a CRISPR locus. RNA-DNA hybridisation in the R-loop has the effect of displacing a single strand of the original duplex DNA, exposing it to the nuclease and ATP-dependent transloCase activity of Cas3. MBPCas5 and Cas7 were investigated for their ability to displace ssDNA from a duplex DNA with a short mismatched region. Displacement of one strand was observed at 100nM MBPCas5-Cas7, (Figure 3-13, bottom panel). This was not seen with Cas7 in isolation (top panel) at the same concentrations. The strand displacement activity required ATP (right hand lanes of bottom panel), though the highest concentrations of Cas5 and Cas7 actually manifested with decreased activity. Cas8' was also tested for ATP dependant strand displacement activity as previously reported (*269*), but no displacement was detected (data not shown).

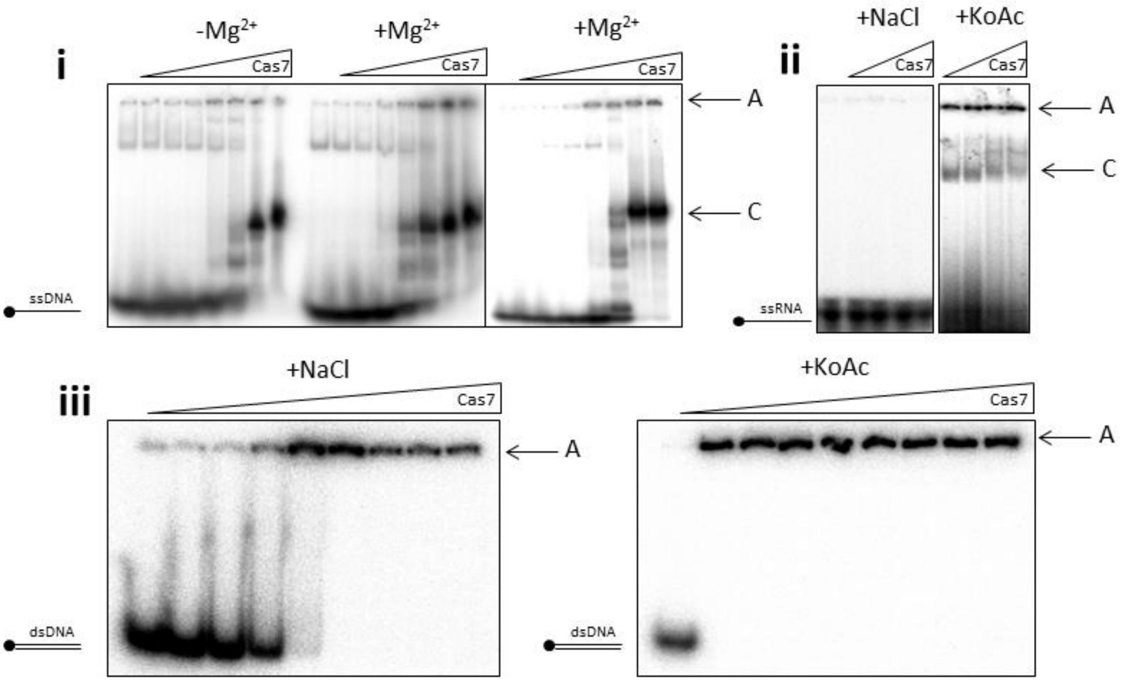


Figure 3-12.EMSA analysis of titrating Cas7 into ssDNA, ssRNA and dsDNA substrates in different salts. In panel i, Cas7 was used at 0, 25, 50, 100, 200, 400, 800 and 1000nM with 5nM ssDNA in reactions containing 125mM NaCl and either 10mM EDTA (-Mg^{2+}) or 10mM $MgCl_2$ (+Mg^{2+}). In ii, Cas7 was used at 0, 100, 200, 400 and 800nM with 5nM of ssRNA in a reaction containing 125mM NaCl or 125mM KOAc and 10mM $MgCl_2$. In iii, Cas7 was used at 0, 10, 20, 50, 100, 200, 400, 800 and 1000nM with 5nM dsDNA in a reaction containing 125mM NaCl or 125mM KOAc and 10mM $MgCl_2$. In reactions containing NaCl the protein was purified in a buffer containing 150mM NaCl and for KOAc protein buffer contained 500mM KOAc. Each panel shows phosphorimages of native TBE gels separating unbound substrate, in well aggregates (A) and specific in gel complexes (C) or each substrate end labelled with ^{32}P as indicated by (•). These gel images are representative of assays that have been carried out in duplicate.

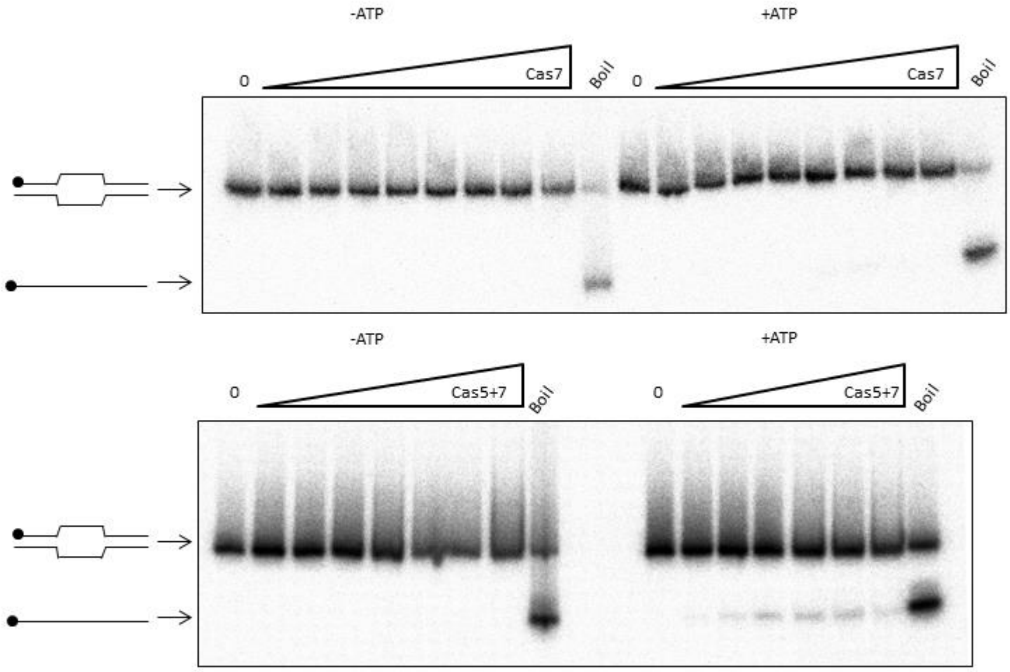


Figure 3-13. Native-PAGE separating assays of titrating Cas7 and MBPCas5-Cas7 into partial duplex substrates to identify strand displacement activity. In the top panel Cas7 was used at 0, 10, 20, 50, 100, 200, 400, 600 and 800nM with 5nM partial duplex in a reaction containing 125mM KOAc, 10mM $MgCl_2$ and -/+ 10mM ATP. In the bottom panel MBPCas5-Cas7 was used in the same conditions but at 0, 10, 20, 50, 100, 200 and 400nM. **'**Boil' lanes represent zero protein samples heated to 95°C for 5 mins before loading. Each panel shows phosphorimages of native TBE gels separating partial duplex and linear ssDNA as indication with each substrate end labelled with ^{32}P as indicated by (•). These gel images are representative of assays that have been carried out in duplicate.

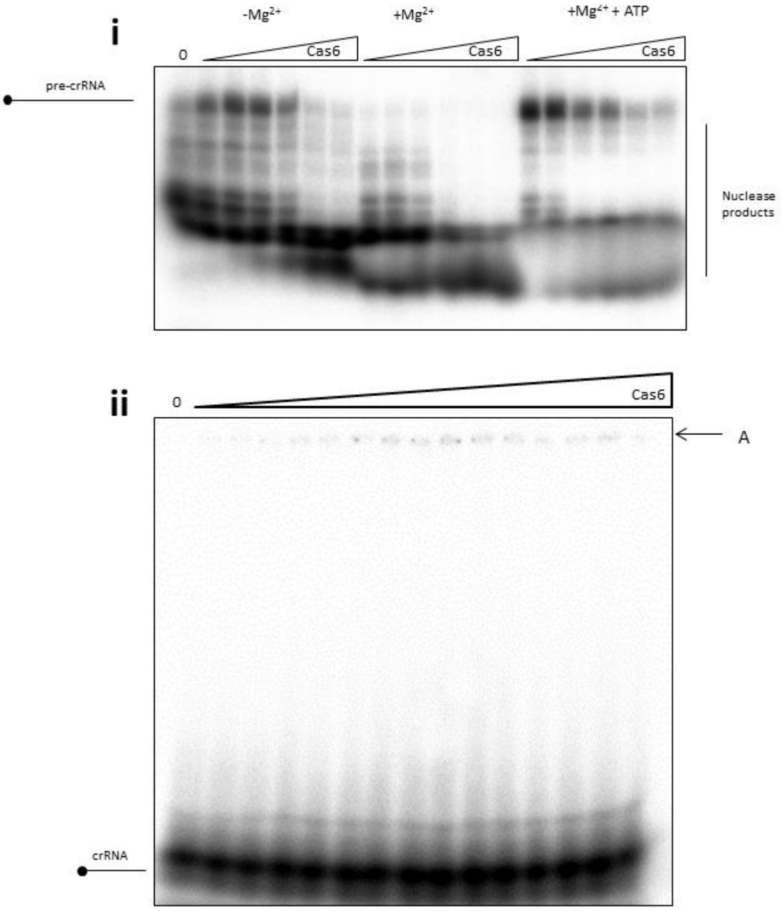


Figure 3-14. Nuclease assays and EMSA analysis of titrating Cas6 into long pre-crRNA and short mature crRNA respectively.. In panel i, Cas6 was used at 0, 10, 50, 100, 200, 400 and 800nM with 5nM of pre-crRNA in reactions containing 125mM KoAc and either 0 MgCl$_2$, 10 mM MgCl$_2$ or 10mM MgCl$_2$ and 10mM ATP. In ii, Cas6 was used at 0, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 254, 500, 750 and 1000nM with 5nM mature crRNA in reactions containing 125mM KoAc and 10mM MgCl$_2$. Panels i and iii, show phosphorimages of native TBE gels separating Cas6 nuclease products of pre-crRNA (i) and binding shifts of crRNA (ii), in well aggregates labelled (A). Each RNA end was labelled with ^{32}P as indicated by (•).These gel images are representative of single assays.

Cas6 family enzymes are nucleases that target pre-crRNA for maturation. Mature crRNA is assembled into Cascade interference complexes. In *E.coli*, Cas6 matures pre-crRNA by cleavage at stem loop structures formed from the repeat sequence of transcribed CRISPRs. Cas6 family enzymes show variation in their catalytic mechanism reflecting the diversity of CRISPR systems. *E.coli* Cas6 (Cas6e or CasE) is an integral part of the Cascade complex as it processes crRNA and forms a 3’ cap. However, other Cas6 proteins, including those of some CRISPR type I-B systems process pre-crRNA into crRNA and then transfer crRNA to Cas7 and are not part of the Cascade complex.

Because of the varied activities of Cas6 proteins, purified *Mth* MBPCas6 was tested for nuclease activity on pre-crRNA and binding to mature crRNA. MBPCas6 showed increased nuclease activity in an Mg^{2+} dependent manner (Figure 3-14I, left versus middle lanes). On the other hand, addition of ATP decreased RNase activity (right hand lanes), possibly through ATP sequestering Mg^{2+} away from Cas6. Moreover, EMSAs were carried out to ascertain Cas6 binding to short synthetic RNA molecules. No specific in-gel shifts were observed when Cas6 was titrated into RNA, only in-well protein-RNA aggregates (A), (Figure 3-14ii). Therefore *Mth* Cas6 cleaves pre-crRNA into mature crRNA but does not bind to this mature product.

3.3.2 Analysis of the oligomeric state of an archaeal Cas3

Cas3 has 3' to 5' ATP dependent ssDNA translocating activity that is required for interference in type I CRISPR systems. The oligomeric state of DNA transloCases/heliCases has for some years been recognized as an important factor in determining their mechanism. For example, Rep-UvrD family heliCases are ATP dependent transloCases (*271-273*). HeliCases such as HerA function as hexameric rings while others like UvrD have reported activity as monomers or dimers (*274-276*). The *E.coli* NER heliCase UvrD acts as a homodimer that enables ATP hydrolysis and substrate specific binding. Untagged *Mth* Cas3 was purified and analysed to discern its oligomeric state using Superose-6 analytical gel filtration in various conditions. Additionally, ATPase defective D347A (Walker B motif mutation) Cas3 was purified and analysed for comparison with wild-type Cas3. Following gel filtration, plotting the retention volume (V_o) (x-axis of Figure 3-15i) against the molecular weight standards (BioRad) gave a prediction of the molecular weight of Cas3 as ~125kDa in conditions containing no additives (Figure 3-15ii). Peak shifts were observed following the addition of ssDNA (pink and orange traces). These shifts corresponded to a doubling in the molecular weight (from ~125kDa to ~245kDa) suggesting Cas3 either formed a dimer or two separate Cas3 molecules bound ssDNA. Interestingly, identical results were observed for both the wild-type protein and the D347A mutant. Additionally, *Mth* Cas3 D347A elutions were analysed by SDS-PAGE. Whereas in the Cas3 protein-only gel the elution peak was in lane 10 only (Figure 3-16i), this peak shifted to include also lanes 6 and 7 when Cas3 was pre-incubated with ssDNA (EB44) (Figure 3-16v and vii). This suggested a 1ml difference in Cas3 elution supporting the UV peak shift that was observed. With the addition of ATP there was no obvious difference in elution profiles, though Cas3 did diffuse over several fractions (Figure 3-16ii, iv, vi and viii). When incubated with a plasmid (pUC19, Figure 3-16iii and iv), some Cas3 formed unresolvable aggregates (light blue and purple traces) at 6ml (Figure 3-15i), while the remaining peak was more diffuse. In several gels additional proteins were present, where were most likely degradation products of Cas3 itself as they had been present both at the beginning and at the end of the experimental assays.

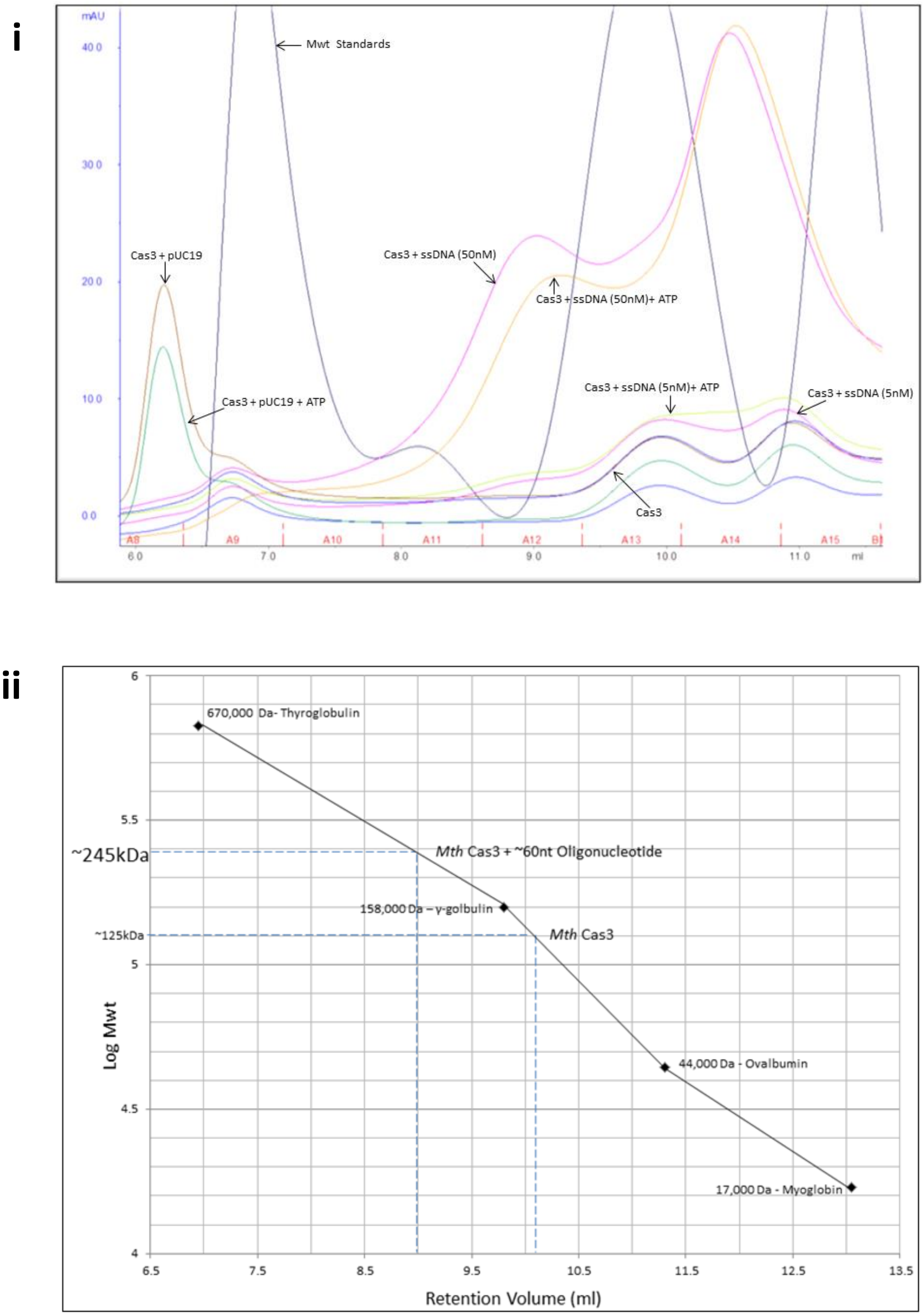


Figure 3-15. Analytical size exclusion chromatography profiling of Cas3. (i) 50µg of Cas3 was pre-incubated with the indicated substrates, -/+ 5mM ATP, 2mg pUC19 and -/+ 5mM ATP, 5nM ssDNA (EB44) and -/+ 5mM ATP and finally, 50nM EB44 and -/+ 10mM ATP (vii and viii). Each mixtures was loaded in a volume of 300µl onto a Superose 6 size exclusion column and eluted over an isocratic gradient of 1CV (22ml) and UV traces compiled.(ii) Peaks observed for Cas3 and Cas3 plus ssDNA were then compared for difference in elution volume. Log (Mwt) of the standards were plotted against the elution peaks of Cas3 to approximate the change in kDa of the eluting protein from 125kDa to 245kDa.

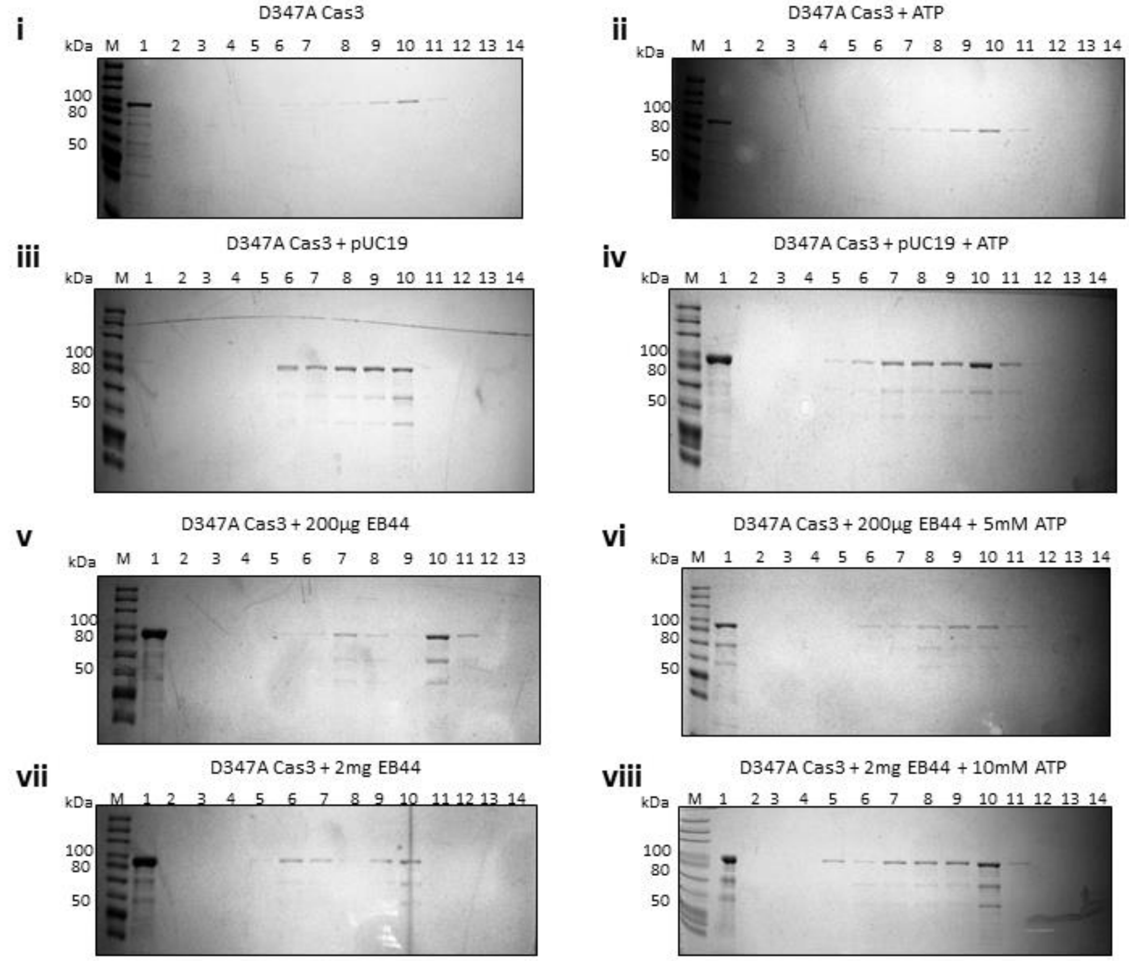

i
D347A Cas3
kDa M 1 2 3 4 5 6 7 8 9 10 11 12 13 14
100
80
50
ii
D347A Cas3 + ATP
kDa M 1 2 3 4 5 6 7 8 9 10 11 12 13 14
100
80
50
iii
D347A Cas3 + pUC19
kDa M 1 2 3 4 5 6 7 8 9 10 11 12 13 14
100
80
50
iv
D347A Cas3 + pUC19 + ATP
kDa M 1 2 3 4 5 6 7 8 9 10 11 12 13 14
100
80
50
v
D347A Cas3 + 200µg EB44
kDa M 1 2 3 4 5 6 7 8 9 10 11 12 13
100
80
50
vi
D347A Cas3 + 200µg EB44 + 5mM ATP
kDa M 1 2 3 4 5 6 7 8 9 10 11 12 13 14
100
80
50
vii
D347A Cas3 + 2mg EB44
kDa M 1 2 3 4 5 6 7 8 9 10 11 12 13 14
100
80
50
viii
D347A Cas3 + 2mg EB44 + 10mM ATP
kDa M 1 2 3 4 5 6 7 8 9 10 11 12 13 14
100
80
50

Figure 3-16. SDS-PAGE analysis of Cas3 D347A protein detected in analytical size exclusion chromatography. In each size exclusion elution 50µg of Cas3 was pre-incubated with the indicated substrates, -/+ 5mM ATP (i and ii), 2mg pUC19 and -/+ 5mM ATP (iii and iv), 5nM EB44 and -/+ 5mM ATP (v and vi) and finally), 50nM EB44 and -/+ 10mM ATP (vii and viii). Each mixture was loaded and eluted over an isocratic gradient of 1CV (22ml). Proteins from each fraction were then precipitated by TCA precipitation and analysed by SDS-PAGE. Cas3 protein alone elutes in fraction 10 (i) and becomes more diffuse with ATP (ii). With addition of ssDNA (EB44) a distinct second peak is observed in lane 7 (v and vii).

3.4 Cloning of *E.coli Cas* genes and purification of the encoded proteins

E.coli K-12 MG1655 contains a CRISPR Type I-E system, which has 8 Cas genes: Cas*1, Cas2, Cas3, CasA, CasB, CasC, CasD* and *CasE*, located adjacent to a CRISPR locus (CRISPR-1). *E.coli* Cascade is similar to archaeal Cascade (*213*), and has similarly named proteins: Cas5 (CasD), Cas7 (CasC) and Cse2 (CasA/Cas8 family) (*128, 129, 264, 277*). AS in Archaea, Cas3 is not an integral part of Cascade but is recruited by

Cascade to degrade the targeted DNA. Cas1 and Cas2 proteins are required for adaptation, and are functionally linked to Cascade-Cas3 reactions by an unknown mechanism during 'primed' adaptation (*126*). As part of the research included in this thesis plasmid constructs were generated to over-express *E. coli* Cascade, Cas1, Cas2 and Cas3 proteins with the aim of analysing primed adaptation *in vitro* and in genetic studies. The majority of this molecular cloning was carried out by Edward Bolt, Alex Hughes and Anna-Sophie Brinkmann who kindly provided plasmids labelled EB, AH and ASB, respectively and are summarised in Table 10. Whereas the production of the *E. coli* Cas proteins, including the study of Cas3, is reported in this chapter, the major results of this thesis relating to activities of *E. coli* Cas1 and Cas2 are presented in Chapter 5.

Table 10. Summary of plasmid constructs generated and used from *E.coli* genes. pSDC (Simon David Cass), pEB (Edward Bolt), pAH (Alex Hughes), pASB (Anna-Sophie Brinkmann).

Plasmid Construct name	Gene	Plasmid Vector	Affinity protein Tag
pEB499	CasC	pMal-C2x	N-MBP
pSDC3	CasD	pET14-b	N-His_6
pSDC4	CasE	pET14-b	N-His_6
pAH1	Cas3	pMal-C2x	N-MBP
pEB358	ΔC-Cas3	pMal-C2x	N-MBP
pEB505	Cas1	pET14-b	N-His_6
pASB8	Cas2	pQE-HisStrep1	C-His_8-Strep
pASB11	Cas1	pET14-b	N-His_6
pASB12	Cas2	pQE-HisStrep1	C-His_8-Strep
pEB488	TopoI	pET14-b	N-His_6

3.4.1 Overexpression

For MBPCas3 and MBPCasC, previously established expression and purification protocols were followed (in house). Briefly, DH5α cultures expressing either MBPCas3 or MBPCasC were grown to OD_{600}=0.6 and expression induced for only 30mins at 37°C with 0.5mM IPTG, generating soluble proteins for purification. On the other hand, CasD and CasE were expressed in BL21C+ cells, which resulted in soluble

CasE but insoluble CasD (Figure 3-18). Insoluble CasD was dissolved in 6M urea for refolding and purification.

Work of PhD student colleague (Sophie Brinkmann) had provided the optimised over-expression method for *E.coli* Cas1 and Cas2 proteins. Cas1 was produced as a soluble protein with either hexa-histidine tag (His_6-Cas1) in the strain IIB964 or with tandem C-terminal octa-histidine and StrepTactin tags (Cas1-Strep-His_8) in the strain T7express. Expression was optimal when cells were grown at 30°C following induction of expression. Hexa-histidine tagged *E. coli* Cas2 (His_6-Cas2) also expressed as a soluble protein in the strain T7express.

3.4.2 *Purification of E.coli* Cas proteins

3.4.2.1 Purification of E.coli Cas3

MBP-Cas3 was expressed for 30 minutes only due to the toxic effects and undesirable proteolysis of Cas3 associated with expression times >1 hour. Nonetheless, this still yielded sufficient intact MBPCas3 for purification via AKTA FPLC. A four step purification was performed to enhance the purity of the final protein preparation (Figure 3-17). Firstly, an affinity chromatography step employing amylose Sepharose was used to bind the MBP-tagged protein, eluting it over a maltose gradient (0-3mM) (Figure 3-17i). This was followed by ion exchange chromatography employing DEAE sepharose (anion exchanger) and elution within a NaCl gradient (150-1500mM) (Figure 3-17ii), though lower molecular weight contaminants were still observed throughout this step. Finally, affinity chromatography using amylose sepharose was again employed to further remove contaminants and degradation products (Figure 3-17iii). Indeed, it has been shown that the main contaminating protein of approximately 60kDa is a truncated or degraded MBPCas3 protein (as identified by western blot analysis with anti-MBP primary antibody) (*143*).

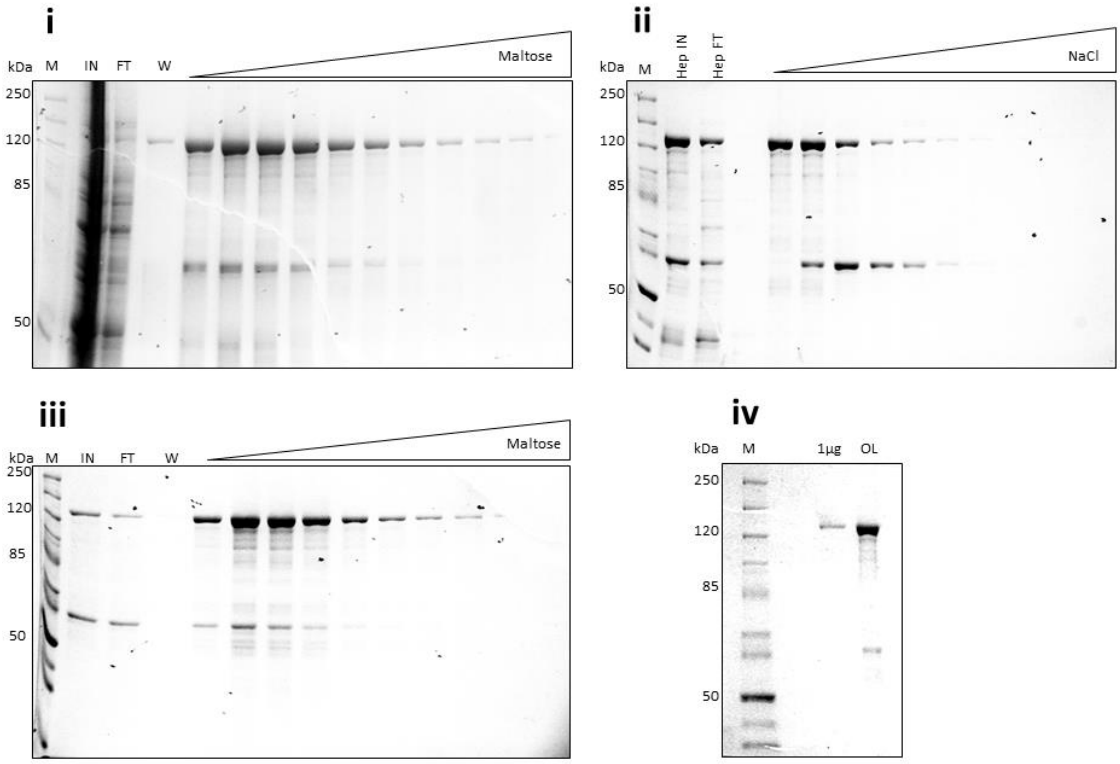


Figure 3-17. Purification of *E.coli* MBPCas3. (i) After sonication and centrifugation soluble material (IN) was subjected to Affinity chromatography (Amylose) and fractions were compared for Cas3 contents, flow through (FT), wash (W) and UV trace peaks during maltose gradient (0-30mM). (ii) Cleanest fractions containing Cas3 were pooled and directly subjected to a second affinity column (Heparin), flow through containing Cas3 was collected and then subjected to anion exchange chromatography (DEAE sepharose) and fractions were compared for Cas3 contents, flow through (FT) and UV trace peaks during NaCl gradient (150-1500mM). (iii) Cleanest fractions containing full length Cas3 were pooled and directly subjected to a third affinity column (Amylose) and fractions compared as before. Pooled fractions were dialysed for storage and 1µg and overload analysed for protein quality before storage.

3.4.2.2 Purification of E. coli CasD and CasE proteins

Insoluble His$_6$-CasD was refolded after dissolution in 6M urea and subsequently purified by His Gravi-Trap nickel-binding chromatography (Figure 3-18I and ii). The methodology used was identical to the Chelating column detailed for *Mth* Cas1 protein (Figure 3-3). Soluble His$_6$-CasE was also purified through His Gravi-Trap nickel-binding chromatography (Figure 3-18i).

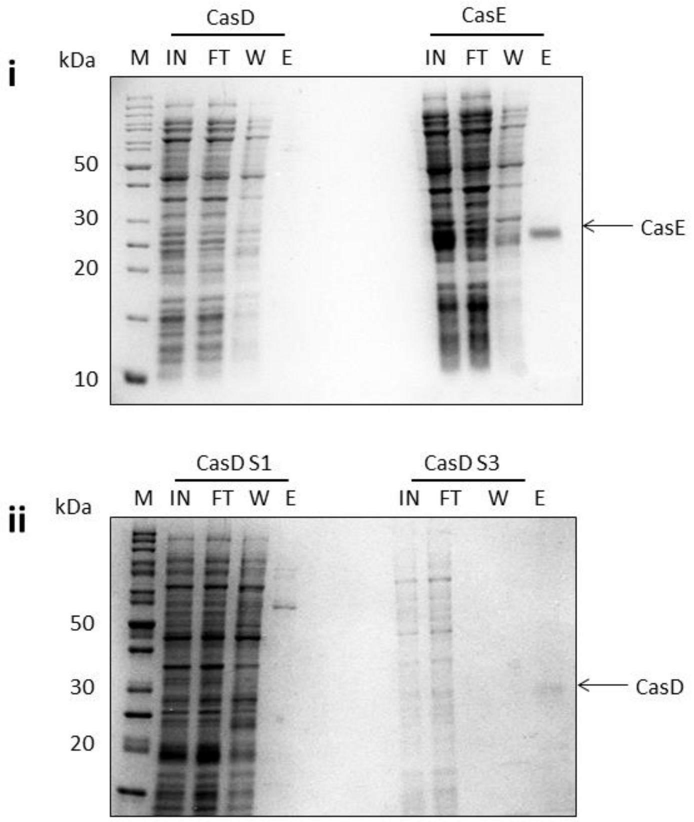


Figure 3-18. SDS-PAGE analysis of *E.coli* CasD and CasE proteins from pilot purification by His Gravi-Trap. In panel i, CasD and CasE pilot purifications of lysed and centrifuged soluble material loaded onto His Gravi-Trap (IN) and compared to flow through (FT), wash (W) and elute (E) fractions. An arrow points to CasE protein detected in the elute migrating at approximately 26kDa. (RMM=22293). No CasD was detected in the eluate. In panel ii, CasD pilot purification of soluble material (S1) and re-solubilised pellet material in 6M urea (S3) loaded onto His Gravi-Trap (IN) and compared to flow through (FT), wash (W) and elute (E) fractions. An arrow points to CasD protein detected in the elution of the S3 fraction migrating at approximately 30kDa (RMM=25209).

3.4.2.3 Purification of E. coli Cas1 and Cas2 proteins

Cas1 and Cas2 purified proteins were generously provided by Anna-Sophie Brinkmann for further study.

3.4.2.3.1 Summary

Successful purification of *E. coli* Cascade, Cas1 and Cas2 allowed further analysis of *E. coli* CRISPR immunity *in vitro*, with emphasis on studies of Cas1-Cas2 presented in Chapter 5. Cas3 was briefly analysed as described below in cooperation with a fellow PhD student (Jamieson Howard).

3.5 *In vitro* pull down interactions of *E.coli* Cas3 and DNA Topoisomerase I

A bacterial-2-hybrid screen of *E.coli* Cas3, carried out by Jamieson Howard (as detailed in his thesis: (*278*), identified DNA Topoisomerase I (TopoI) as a candidate Cas3 interacting partner. DNA Topoisomerase I alters the topomeric (number of coils) or supercoiled DNA by cleaving a single strand hence relaxing the super-coil. This is important during DNA replication and transcription where the unwinding of DNA by heliCases at the replication fork causes the DNA ahead of it to become over coiled Removal of this coiling is essential as otherwise a build-up of torsional tension would eventually halt DNA and RNA polymerase action. TopoI interaction with Cas3 was tested with *in vitro* analytical size exclusion chromatography though no robust data was generated. Therefore, an alternative strategy where the affinity tags of each protein were exploited was employed. MBPCas3 immobilised on amylose resin was mixed with His_6TopoI. His_6TopoI co-eluted with MBPCas3 in stoichiometric amounts when MBPCas3 was stripped from the amylose resin (Figure 3-19ii, lane E). Although some His_6TopoI was present in the wash (W) fraction (lane W); including in controls where His_6TopoI was mixed with the MBP tag only, His_6TopoI was absent from the elution (E) lanes in these controls. Therefore, it could be concluded that MBPCas3 formed a stable binding contact between the amylose resin and His_6TopoI. In the previous bacterial-2-hybrid data (*278*), An additional mutant Cas3 protein (ΔC-Cas3) was employed in the aforementioned bacterial-2-hybrid screen (*278*). ΔC-Cas3 is an *E.coli Cas3* gene that is truncated by approximately 500 bp at the C-terminus, so the protein lacked the at the time mysterious C-terminal domain. Interestingly, this Cas3 mutant no longer showed an interaction with TopoI in the screen. Similarly, this thesis corroborated that ΔC-MBPCas3 ΔC-MBPCas3 did not support this interaction as TopoI was only detected in flow-through and wash fractions (Figure 3-19iv, FT and

W). Reciprocal binding experiments proved unreliable as MBPCas3 bound significantly to nickel resin also in the absence of His$_6$TopoI.

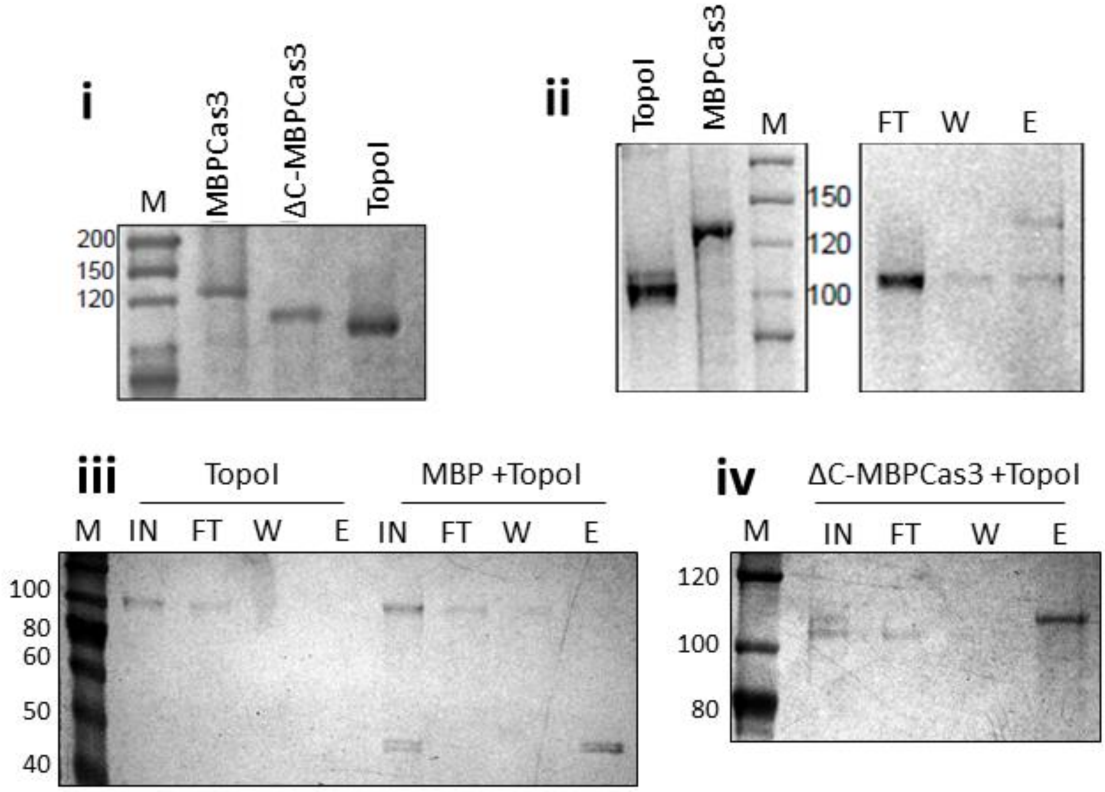


Figure 3-19. Pull down interaction assays between MBPCas3, His$_6$TopoI and ΔC-MBPCas3. (i) MBPCas3 and ΔC-MBPCas3 were purified by affinity chromatography as summarised in the section: *Purification of E.coli Cas3*, His$_6$TopoI was purified the same way as *Mth* His$_6$Cas7. Approximately 1µg of each protein is shown for reference. (ii) 100µg of MBPCas3 and His$_6$TopoI were premixed at room temperature for 30mins with equilibrated amylose resin. Slurry mix was added to a disposable self-pour gravity column. Fractions were collected of flow-through (FT), wash (W) with amylose column binding buffer and eluted (E) with 3mM maltose. (iii) Control pull down assays using 100µg of His$_6$TopoI or His$_6$TopoI and MBP (maltose binding protein) and fractions collected. (iv) 100µg of ΔC-MBPCas3 and His$_6$TopoI were premixed as before and fractions collected. Samples of each fraction were analysed by SDS-PAGE and stained with Coomassie Brilliant Blue.

3.6 Discussion

This chapter focused on the molecular cloning, expression, purification and preliminary characterisation of Cas proteins from *Mth* and *E.coli*. Of particular interest are the CRISPR Type I interference and adaptation proteins from both organisms. Similar methods were utilised for the optimisation of over-expression of both *Mth* and *E.coli* proteins, this included temperature and duration of expressions.

Once soluble protein was detected affinity purification was carried out to generate pure recombinant proteins for *in vitro* analysis.

Expression and purification of *Mth* proteins in *E.coli* came with several caveats: insolubility of various archaeal proteins was the most difficult to overcome. Cas5 and Cas6 proteins precipitated and were insoluble except in high urea concentrations. Generation of MBP fusion constructs and co-expression of MBP-Cas5 with Cas7 alleviated the solubility problems (Figure 3-9). Recombinant proteins of the archaeal Type I-H Cascade complex were now available for *in vitro* reconstitution and analysis (Figure 3-6, Figure 3-7, Figure 3-8 and Figure 3-9). Recombinant *Mth* Cas8', Cas5 and Cas7 are further studied in Chapter 4. *E.coli* Cas1 and Cas2 are investigated in Chapter 5.

3.6.1 *Mth* Cas5, Cas6 and Cas7 initial biochemical tests

The CRISPR-Cas mechanism requires Cas6 to process pre-crRNA into mature crRNA either as part of the Cascade complex in *E.coli* or as a separate processing unit in *Sso*. *Mth* Cas6 is a metal dependent endonuclease like other Cas6 proteins (*153, 154, 199-201*), generating distinct nuclease products. In this study *Mth* Cas6 does not interact with processed crRNA but does cleave pre-crRNA (Figure 3-14). Therefore the role of Cas6 may be to cleave and then transfer crRNA to the *Mth* archaeal Cascade complex.

Archaeal Cas5 and Cas7 proteins, having been investigated in *Sso*, are known to form ribonucleoprotein filaments (*213*). Cas5 acts as the anchor for Cas7 helical extension along the RNA molecule. Whereas Cas5 is important in the seeding of R-loop formation, Cas7 forms the structural core of the Cascade complex (*139, 159, 279*). This study found that Cas7 binds to ssDNA, ssRNA and dsDNA (Figure 3-12I, ii, iii). Importantly, the choice of the salt component in buffers was found to have significant impact on *in vitro* biochemical assays of the different Cas proteins. Potassium acetate (KOAc) is the physiological salt for *Mth* proteins. There was substantial 10 x increase in binding affinity to dsDNA between the proteins purified and assayed in KOAc, versus NaCl (Figure 3-12iii). It has been reported previously that Cas8' (Nar71) has strand displacement activity (*269*). This activity was detected in NaCl, in the assays presented here KOAc was used where no displacement activity was detected by Cas8'. This chapter did show that MBPCas5 in association with Cas7 carried out strand displacement, known to be essential for R-loop formation in the presence of both Mg^{2+} and ATP (Figure 3-13). (*162, 226*). The ATP dependence observed in these assays was consistent with the same requirements seen with other stand invading proteins such as RecA. The lack of strand displacement activity in Cas8' is in line with the recent identification of CasA's (analogous to Cas8') role in CRISPR interference which suggests that Cas8' identifies PAM and stimulates Cas3 degradation of ssDNA. Therefore, Cas8', together with Cas5 and Cas7, which have demonstrable strand displacement activity, act in unison to seed strand displacement and stabilise R-loop formation

3.6.2 Cas3 oligomeric state and interaction partners

CRISPR Type I systems recruit the heliCase/nuclease Cas3 to the Cascade complexes after stable R-loop formation. In *E.coli* CasA identifies the PAM sequence and stimulates Cas3 mediated degradation of the invading genetic element (*162, 163*). However, from the available co-crystal structure of the C-terminal domain of Cas3 with Cascade only limited mechanistic conclusions can be drawn because only part of the Cas3 protein is present. Therefore, the present study employed size exclusion chromatography to analyse the mode of Cas3 binding to DNA molecules. Both Wild type and D347A proteins elute at approximately double the molecular weight of Cas3 with DNA present. Two possibilities explain this observation; two Cas3 monomers are binding independently to the short DNA molecule, or like other heliCase/nucleases two Cas3 proteins surround the DNA molecule in a ring-like conformation facilitating translocation.

The interference proteins of CRISPR immunity are expressed upon infection (*146, 169*). When this infection is novel to the cell, disruption of invading DNA metabolism or viral lytic pathways allows acquisition of a novel spacer for targeted degradation. It has been postulated that Cas proteins are involved in this 'facilitation' stage of CRISPR immunity (*266*). This idea combines is further supported by previous work in our laboratory which showed the Cas3 and DNA topoisomerase I interaction through Bacteria-2-Hybrid and the positive effect of Cas3 on TopoI activity in converting supercoiled plasmid to relaxed (unpublished work carried out by Jamieson Howard). In the context of viral infections, this would be significant in restricting invader replication as phage O and P proteins require negatively supercoiled DNA for binding and replication initiation (*97-99*). Indeed, this study confirmed Cas3 could interact with TopoI both *in vivo* and *in vitro*. Moreover, the aforementioned Bacteria-2-Hybrid screen revealed a C-terminal region truncated Cas3 mutant (ΔC-Cas3) could no longer support this interaction. Indeed, the results from this chapter showed ΔC-Cas3 did not interact with TopoI and predictably did not stimulate the relaxation activity of TopoI. Therefore, it is the C-terminal region of Cas3 which mediates Cas3-TppoI interaction. The C-terminal region of Cas3 is possibly the location for several protein-protein interactions, supporting multiple interactions with host factors and other Cas proteins (CasA) important for CRISPR immunity, host processes or selfish protection of the CRISPR MGE (*162, 265*).

4 Biochemical analysis of archaeal Cas8 in CRISPR interference

4.1 Summary

The work presented in this chapter forms the core of a research paper on the role of *Methanothermobacter thermautotrophicus* Cas8' protein in CRISPR interference, which can be found attached at the end of this thesis (*144*). The work presented in section 4.2.5.4 and Figure 4-23 (panel ii) was performed and analysed by Karina Haas and Britta Stoll from the collaborating Dr. Anita Marchfelder's research group at the University of Ulm, Germany.

4.1.1 Cas8' – Current understanding of the signature gene of the *Methanothermobacter thermautotrophicus* Type I-H CRISPR-Cas immune system

Cas8 is essential for CRISPR interference in *Haloferax volcanii*, a discovery made when sequencing CRISPR-Cas loci from *Hvo* clones that had lost CRISPR interference (*280, 281*). Sequencing of CRISPR-Cas regions of these clones identified that either mutations in Cas3 or Cas8 genes, chromosomal deletions or reshuffling, in each Case abolished interference. This provided insight for the importance of Cas8 proteins for achieving effective CRISPR immunity. *Mth* ORF 1090 had been previously investigated in the Bolt laboratory following its identification in biochemical screens for novel DNA heliCases in archaea (*269*). The protein product of ORF 1090 was named Nar71, later defined as the CRISPR protein Cas8b, and now called Cas8'. In preliminary experiments in our laboratory, Cas8' was found to be a weak nuclease on DNA flayed duplex and 3'flap structures. Additionally, Cas8' displayed ATP-dependent strand displacement activity, that was abolished when poorly hydrolysable ATPγS was used. However, Cas8' nuclease activity on RNA substrates had not been assayed(*269*).

4.1.2 Aims

The aim of this part of the project was to corroborate *in vivo* data stemming from research of our collaborators at the University of Ulm, identifying the importance of *Hvo* Cas8 in interference with *in vitro* biochemical characterisation of the *Mth* Cas8' protein . To achieve this, Cas8 was characterised through mutational analysis to determine its significance/ role in the CRISPR interference in *Mth* and *Hvo* archaeal Cascades.

4.2 Results

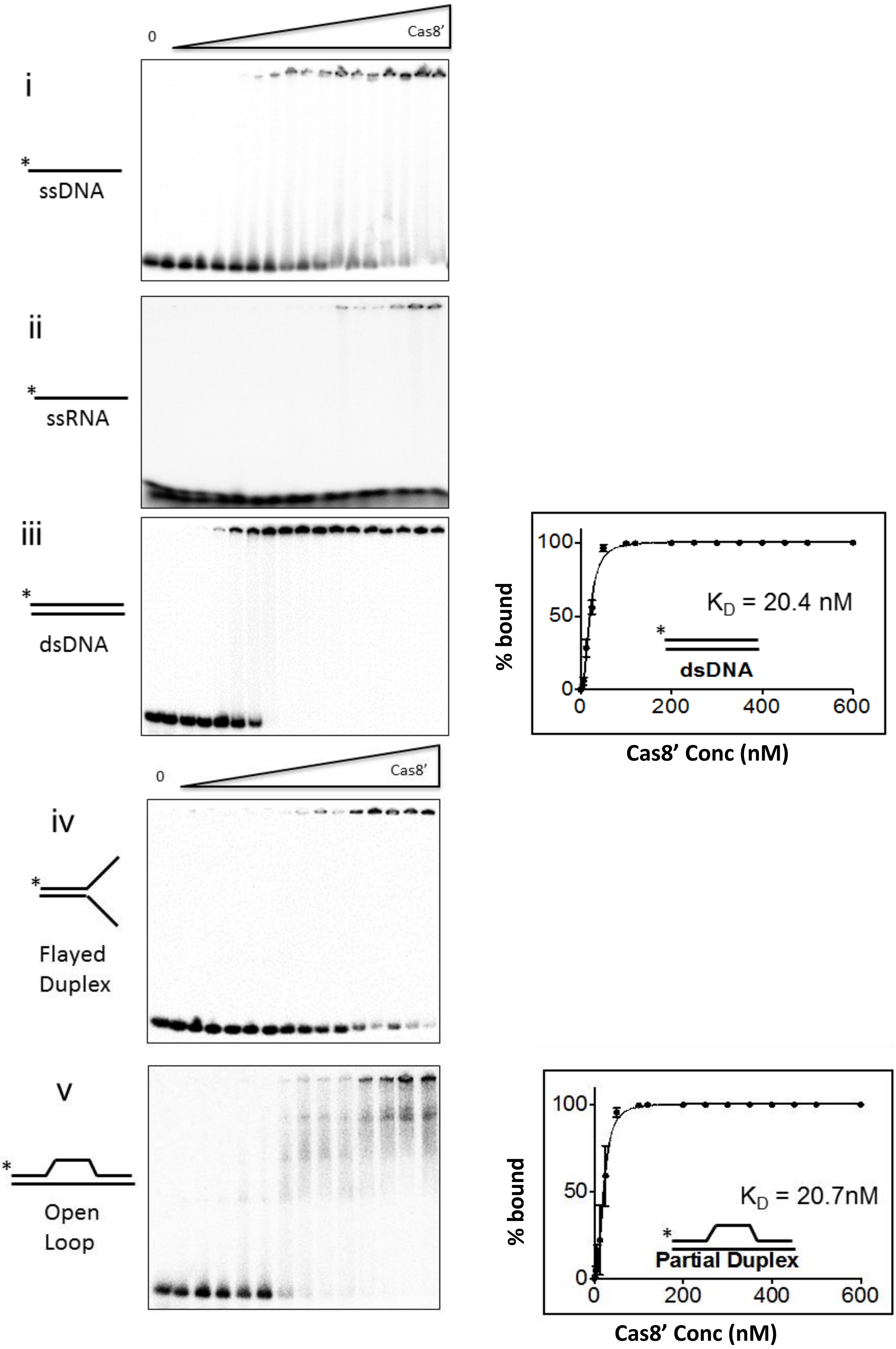
4.2.1 Substrate structure specificity of Cas8'

Previous analysis of Cas8' nucleic acid processing activity had been limited to branched DNA molecules given these were the substrates employed in the screening procedure to identify novel proteins (*269*). The importance of Cas8' in CRISPR interference, identified by genetic analysis by our collaborators, led us to re-examine the nucleic acid binding and nuclease activities of purified Cas8' using RNA and other DNA substrates. To this end, EMSAs titrating Cas8' into various substrates, ranging from simple single stranded to complex duplex or triplex molecules, were performed in triplicate. Representative gels of EMSAs with each substrate and the corresponding quantitative analyses can be seen in Figure 4-1. A summary in Table 11 shows the dissociation constants (K_D values) attained from these assays using Y = Bmax*X/(Kd + X), where Bmax is the substrate concentration at which highest binding efficiency is observed. These assays revealed Cas8' interacted very weakly with ssDNA, ssRNA and flayed duplex (Figure 4-2,panels i, ii and iv, respectively), which precluded the determination of a valid K_D in these instances.. However, for the purpose of completion, aproximate K_Ds which were in the range of 269-400nM, have been included in Table 11 to allow comparison with the K_D values of other substrates.

On the other hand, Cas8' bound with substantially higher affinity to simple duplex and flapped structures. Indeed, K_D values between 20-50 nM were obtained for duplex, partial duplex and various flapped RNA and DNA substrates (Figure 4-2, panel iii, v and xiii, ix and respectively). Interestingly, Cas8' bound more complex structures with even higher affinity, as shown by the K_Ds between 4 and 10 nM. Moreover, the pattern of binding to both D-loop and R-loop also differed from the other substrates (Figure, panels vii and viii). Specific distinct in-gel complexes were apparent with these triplex substrates indicating stable binding, whereas other simpler substrates only gave in-well aggregates, typical of non-specific protein-DNA interactions.

Table 11. Summary of Cas8' binding affinities to nucleic acid substrates quantified from EMSAs. K_D values were calculated from quantification of triplicate assays.

Substrate	**K_D**
ssDNA	300 ± 6.3nM
ssRNA	>400nM
Linear duplex	20.4 ± 4.4nM
Flayed duplex	269 ± 6.7nM
3'DNA/DNA flap	19.8 ± 0.9nM
3'RNA/RNA flap	46.8 ± 2.8nM
3'RNA/DNA flap	49.5 ± 3.4nM
Open loop	20.7 ± 2.2nM
D-loop	9.0 ± 2.2nM
R-loop	4.7 ± 0.35nM


0
Cas8'
i
*
ssDNA
ii
*
ssRNA
iii
*
dsDNA
100
50
0
0
200
400
600
% bound
K_D = 20.4 nM
*
dsDNA
Cas8' Conc (nM)
0
Cas8'
iv
*
Flayed Duplex
v
*
Open Loop
100
50
0
0
200
400
600
% bound
K_D = 20.7nM
*
Partial Duplex
Cas8' Conc (nM)

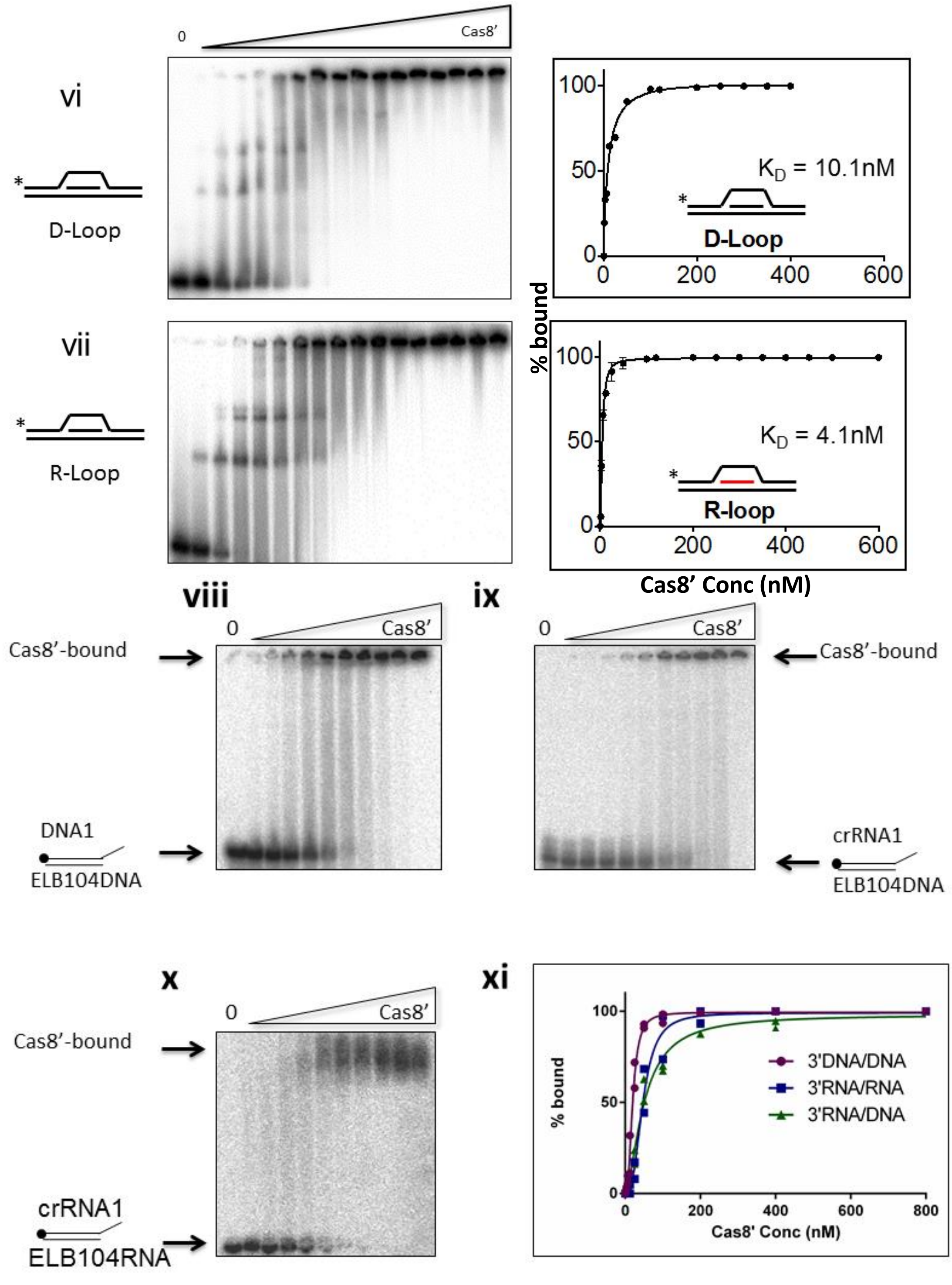


Figure 4-1. EMSAs of Cas8′ titrated into various generic nucleic acid substrates. Cas8′ was titrated at 0, 1.56, 3.125, 6.25, 12.5, 25.0, 50.0, 100, 150, 200, 250, 300, 350, 400, 450, 500 600nM to 5nM substrate in reactions containing 10mM EDTA. Representative phosphorimages of native TBE gels and resulting quantified binding isotherm plots, displaying percentage substrate bound against Cas8′ concentration, are shown. Cas8′ formed distinct complexes with Open loop (v), D-loop (vi) and R-loop (vii), compared to the remaining DNA substrates with which it only formed aggregates (i-iv). Binding affinities for each substrate were calculated from three independent EMSAs plotted on a single graph with standard error. In panels viii, ix and x, Cas8′ was used at 0, 1.56, 3.125, 6.25, 12.5, 25.0, 50.0, 100, 200, 400 and 800nM in reactions containing 10mM EDTA with 5nM substrate which was either (viii) DNA-DNA flap,(ix) RNA-DNA hybrid flap and (x) RNA-RNA flap. The calculates K$_D$ values are summarised in Table 1 . In each Case the ^{32}P labelled strand is marked by • or *.

4.2.1.1 Summary

Mth Cas8’ bound branched DNA and RNA substrates with highest affinity. Lowest K_Ds were observed for D- and R-loops, which are structures similar to those involved in CRISPR interference. Interestingly, Cas8’ binding preference was to substrates containing RNA specifically in duplex for rather than ssRNA only.

4.2.2 PAM sensitivity of Cas8’

Archaeal Cas8’ has been predicted to be able to recognize invader DNA, analogously to Cse1 (CasA) in *E. coli* Cascade. When the invader DNA sequence is complementary to a crRNA molecule an R-loop is formed in association with a Cascade complex. Therefore, this hypothesis was to be tested by assaying the binding of *Mth* Cas8’ to R-loop substrates with different PAM sequences. To this end, candidate *Mth* PAMs for incorporation into these R-loop substrates were selected via bioinformatics analysis of the 123 spacers found in the *Mth* genome. First, the online database, CRISPRfinder, containing all sequenced prokaryotic genomes and all CRISPR-like repeat regions with associated genes, was used to identify said spacer sequences from *Mth*. Next, each spacer was compared to all known nucleotide sequences by BLAST analysis. Perfect and near perfect sequence alignments of spacers with known nucleic acid sequences were used to determine the source of the spacer and therefore the DNA sequence flanking it, from this flanking sequence PAM sequences were identified. A full list of the perfect matches with their associated sequences is given in Therefore the sequence 5’-CCC-3’ was selected as the PAM to be incorporated into R-loop substrates and conversely, incorporation of the sequence 5’-AAA-3’ would denote substrates lacking PAM. The intermediate PAM sequences 5’-CTC-3’ and 5’-TTG-3 were also generated.

Table 12, whereas spacers which had 1– 9 mismatches are listed in the appendix. Interestingly, even though the BLAST searches performed were against the whole nucleotide library, all perfectly matched sequences were originating from phage or prophage sequences. Finally, collating the information from seven identical and 104 mismatched (1-9 mismatched bp) hits from 123 spacers gave a predicted PAM of 5'-CCN (Figure 4-2).

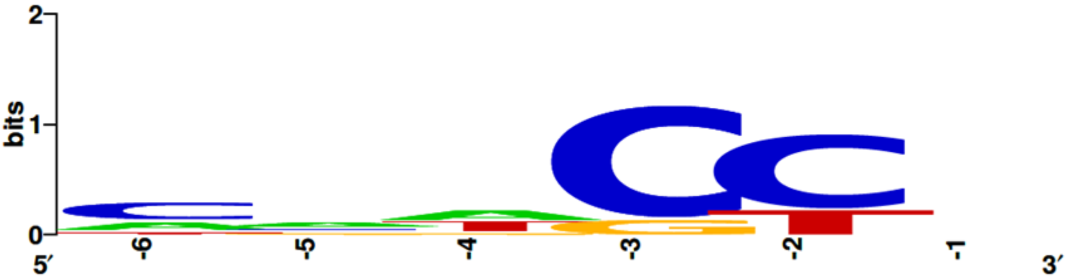


Figure 4-2. **Bioinformatics analysis of *Methanothermobacter thermautotrophicus (Mth)* CRISPR-Cas PAM sequence from genealogical context.** Plasmid/phage sequence matching the spacers with up to 9 mismatches allowed identification of a putative PAM for *Mth* (5'CCN) which was found to be in agreement with a previous study.

Therefore the sequence 5'-CCC-3' was selected as the PAM to be incorporated into R-loop substrates and conversely, incorporation of the sequence 5'-AAA-3' would denote substrates lacking PAM. The intermediate PAM sequences 5'-CTC-3' and 5'-TTG-3 were also generated.

Table 12. Spacers that perfectly matched known nucleotide sequences by nBlast of *Methanothermobacter thermautotrophicus* CRISPR-1 sequences. Genomic context was identified and putative PAM sequences shown 5' – 3'.

Spacer No	Sequence	nBlast match	PAM
6	AAGCGCCGGGCAGACAGCACACAT ACAAGACTTCACAA	*Methanothermobacter wolfe*ii prophage psiM100	CCC
8	TGATGTTGGGAAGGTTTGGCCATC TGAATGATTTGA	*Methanothermobacter wolfeii* prophage psiM100	GTA
17	TATCATCACGCTTGAAGAGTATAAT AAAGTTGTTAAGA	*Methanothermobacter wolfeii* prophage psiM100	GTA
22	AGTATGTGCAGTATCCTCTCTATGT CCCCTTCATTC	*Methanobacterium* phage psiM2	CCT
24	AATATTGAAACGTTCAAGGACATG TTGAAGAGGTATG	*Methanobacterium* phage psiM2	CCT
28	AGTATGTGCAGTATCCTCTCTATGT CCCCTTCATTC	*Methanobacterium* phage psiM2	CCT

74	CGGGGAGAGTCTGACATTCTTGAA GTAGAACCTCCCC	*Methanothermobacter wolfeii* prophage psiM100	CCC

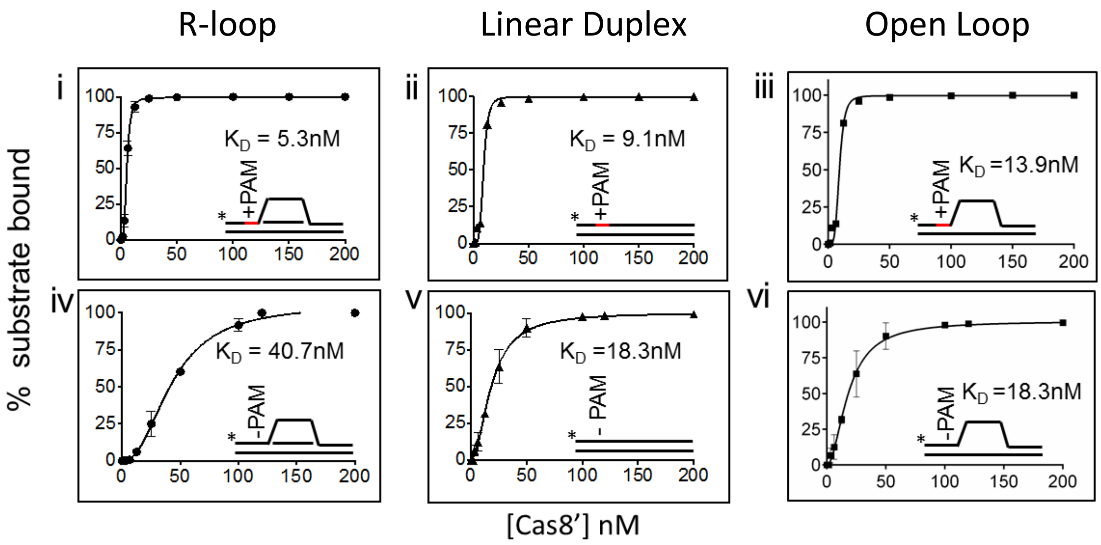


Figure 4-3. Quantified EMSA data of Cas8′ titrated into various substrates with or without PAM. Cas8′ was used at 0, 1.56, 3.125, 6.25, 12.5, 25.0, 50.0, 100, 120, 200, 250, 300, 350, 400, 450, 500 and 600nM with 5nM substrate (i and iv = R-loop, ii and v = linear duplex and iii and vi = open loop) in reactions containing 125mM KOAc and 10mM EDTA. Graphs were plotted as percentage of substrate shifted compared to 0 protein control (as quantified by 2D densitometry) against up to 200nM concentration of Cas8′. Standard error (error bars) and K_D were calculated from duplicate gel quantifications.

Of note, only structures physiologically relevant to CRISPR inference intermediates such as dsDNA, open duplex and R-loops were synthesised. Subsequently, PAM-sequence specific binding by wild-type Cas8′ protein could be tested by EMSA analysis. The K_D values of binding to the said substrates with and without selected PAM were determined (Figure 4-3). The simplest substrate, duplex DNA, manifested with a twofold decrease in K_D values, going from K_D of 18.3 ± 0.9nM without PAM (Figure 4-4, panel v) to K_D of 9.1 ± 1.2nM with PAM (Figure 4-4, panel ii). Similarly, Cas8′ binding of open loops showed a preference for sequences including PAM

(Figure 4-4, panel iii) over those lacking PAM (Figure 4-4, panel vi) with the corresponding K_D values of 13.9 ± 0.2nM and 18.3 ± 0.8nM, respectively. However, the most profound difference was observed for Synthetic R-loops with PAM which displayed an 8-fold increase in binding affinity to Cas8' compared to R-loops without PAM, K_D of 5.3 ± 0.6nM (panel i) versus K_D of 40.7 ±1nM (panel iv), respectively. Next, intermediate PAM sequences were used to detect subtle differences in binding of Cas8' to R-loops (Figure 4-4). The R-loop with the PAM sequence 5'-TTG-3' (Figure 4-5, panel iii) showed a higher K_D of 9.8 ± 0.4nM, which was relatively close to that obtained for the main PAM (CCC) substrate (K_D of 4.9 ± 0.6nM, Figure 4-3 panel i). On the other hand, for the R-loop with the 5'-CTC-3 sequence', a K_D of 29.7 ± 1.0nM was calculated which corresponded to a binding affinity intermediate between the R-loops with (CCC) and without (AAA) PAMs (Figure 4-5, panel iv).

The effect of PAM on Cas8' binding affinity was also tested on a physiological invading substrate, a plasmid, where an R-loop forms in CRISPR interference. Therefore, pUC19 derivatives containing a spacer (that would facilitate the synthetic crRNA annealing) and with or without PAM were created. Cas8' was tested for differences in binding affinity to the different plasmids. However, similar affinities were detected between these plasmid substrates (data not shown).

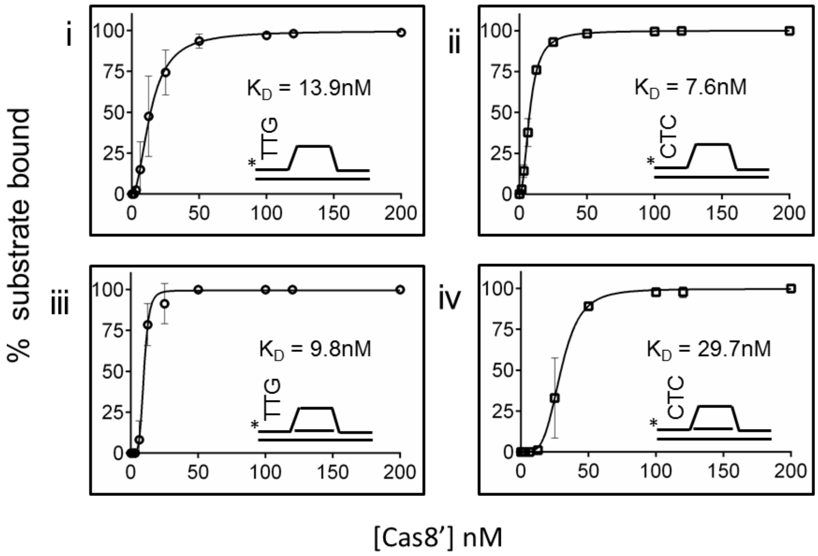


Figure 4-4 Binding isotherms from quantified EMSA assays with Cas8' titrated into alternate PAM substrates. Cas8' was used at 0, 1.56, 3.125, 6.25, 12.5, 25.0, 50.0, 100, 120, 200, 250, 300, 350, 400, 450, 500 and 600nM with 5nM substrate (I and ii = open loop, iii and iv = R-loop) in reactions containing 125mM KOAc and 10mM EDTA. Graphs were plotted as percentage of substrate shifted compared to 0 protein control against up to 200nM Cas8' concentration. Standard error (error bars) and K$_D$ were calculated from triplicate gel quantification.

4.2.2.1.1 Summary

Recombinant *Mth* Cas8' assayed *in vitro,* responded to the presence of a PAM sequence via altered binding affinity to all substrates tested, though with varying degree. Indeed, this effect on Cas8' affinity was particularly important with branched substrates such as the open duplex and R-loops. However, the most notable effect of the PAM sequence was observed in Cas8' binding to the R-loop substrates. Finally, it could be ascertained that different PAM sequences also affected the binding properties of Cas8' and are discussed later in greater detail.

4.2.3 Nuclease activity of Cas8'

For the biochemical analysis of Cas8' enzymatic mechanisms were investigated.

These included but were not limited to ligase, primer extension, strand displacement and nuclease activities. Cas8' had been previously shown to cleave DNA flaps in a

Magnesium-dependent manner (*269*). Therefore, this was the starting point of the enzymatic characterisation of Cas8'.

4.2.3.1 Optimisation of nuclease reactions in vitro

In agreement with a previously published study, Cas8' nuclease activity was observed on the 3' DNA flap region of a flayed duplex substrate, this activity was dependent on the presence of magnesium also, (Figure 4-5, panels i and ii). Interestingly, when the substrate was labelled on the RNA strand hybridised to a DNA strand the nuclease activity was enhanced, again in a magnesium dependent manner (Figure 4-5, panel iii and iv).

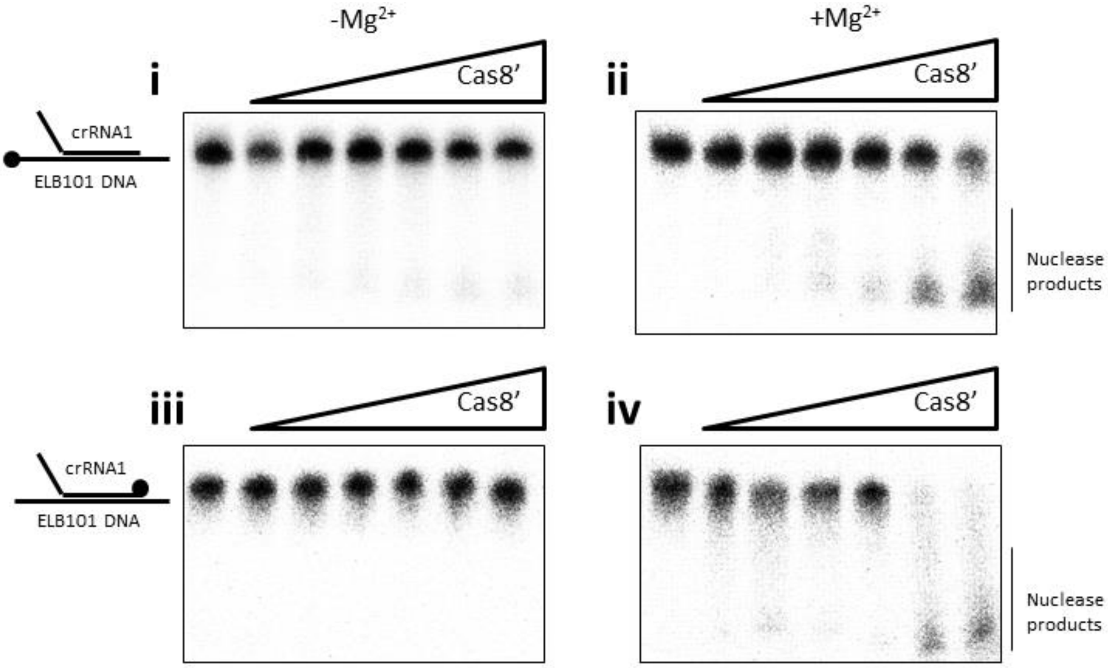


Figure 4-5. Nuclease assays titrating Cas8' into RNA flap substrates, testing metal dependency of the nuclease activity. In panels i and iii Cas8' was used at 0, 50, 100, 200, 400, 600, 800 and 1000nM with 5nM of substrate. In panels ii and iv reactions additionally contained 10 mM $MgCl_2$. Each reaction was incubated at 45°C for 30 mins. All panels show phosphorimages of denaturing urea gels separating Cas8' nuclease products of the 3'RNA flap substrate. DNA was end labelled with ^{32}P as indicated by (•) in panels i and ii, RNA was labelled in panels iii and iv.

Cas8' nuclease activity was compared using ssDNA, ssRNA and linear duplexes. ssRNA nuclease activity of Cas8' peaked at 600nM whereby all the primary substrate was degraded (Figure 4-6, panel ii), whereas ssDNA was poorly degraded even at the highest concentrations used (Figure 4-6, panel i). No similar nuclease activity was observed on linear duplex or plasmid DNA (both single stranded [panel v] and duplex plasmid [panel iv]). Plasmid duplex DNA shifted from relaxed (R) to super coiled (SC), indicating Cas8’ may act as a ligase converting nicked relaxed plasmid into SC. Duplex DNA also decreased in a Cas8’ -dependent manner (panel iii), which was indicative of degradation occurring at the labelled end of the substrate. In time-course analysis, nucleolytic activity of Cas8’ was strongest on RNA substrate (Figure 4-7, panel ii) over the DNA 3’ flapped substrate (Figure 4-7, panel i).

As *Mth* is a moderate thermophile with an intracellular acetate concentration of approximately 700mM, an essential precursor for methanogenesis, the impact of varying salt and temperature in the nuclease reaction conditions on Cas8’ nuclease activity was investigated. Initially, Cas8’ was purified and dialysed into a buffer containing 150mM sodium chloride. Nuclease assays were performed on 3’ flapped RNA and DNA substrates in 125mM potassium acetate. The result was reduction of Cas8’ nuclease activity, comparing lane 7 of Figure 4-7, panels i and ii with lane 3 or Figure 4-8, panels i and iii. Nuclease cleavage assays were then carried out on proteins both purified and assays in potassium acetate at 65°C, RNA flapped substrate nuclease activity was enhanced while DNA activity remained inhibited (Figure 4-8, panels ii and iv). 50nM Cas8’ completely degraded the 3’RNA flap initial substrate whereas DNA nuclease activity was still <50% at 1000nM.

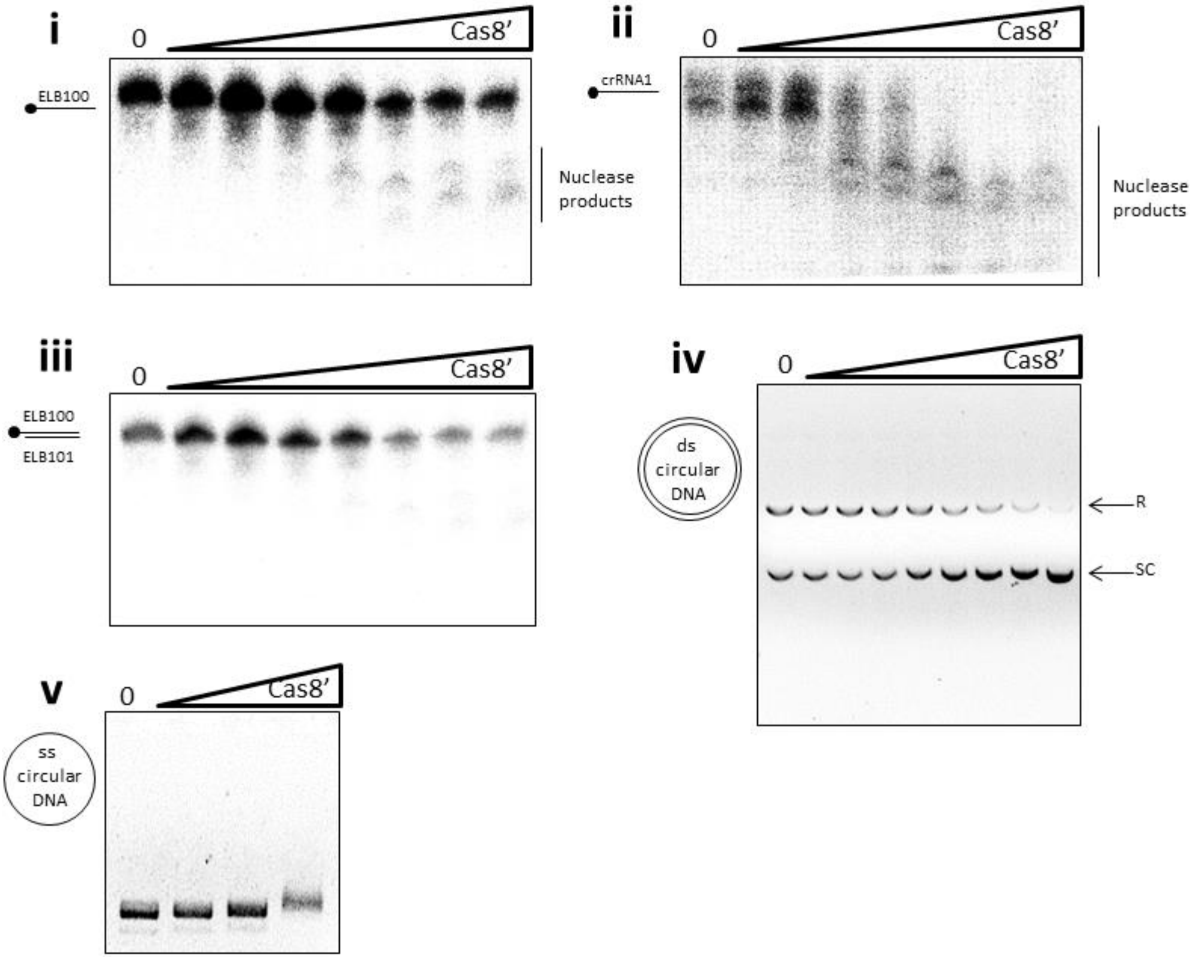


Figure 4-6. Nuclease assays titrating Cas8' into RNA and DNA substrates. In panels i-iv Cas8' was used at 0, 50, 100, 200, 400, 600, 800 and 1000nM with 5nM of DNA in reactions containing 10 mM $MgCl_2$. In panel v, Cas8' was used at 10, 100, 1000nM. Panels i, ii, and iii show phosphorimages of urea gels separating Cas8' nuclease products of, respectively, ssDNA, ssRNA and duplex DNA, each end labelled with ^{32}P as indicated by (•). Panels iv and v are images of ethidium bromide stained agarose gels to separate products of Cas8' activity on duplex or ssDNA plasmid.

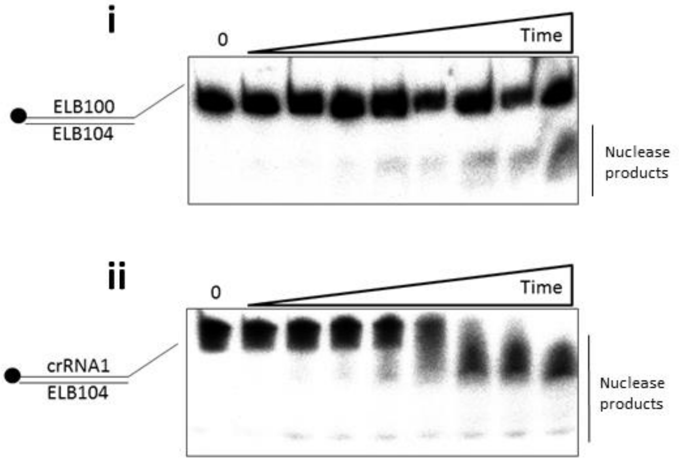


Figure 4-7. Nuclease time course assay of Cas8’ on 3’ flapped DNA and RNA substrates. Cas8’ was used at 100nM in reactions containing 10mM $MgCl_2$ on 5nM 3’DNA (i) and RNA (ii) flapped substrates. Reactions were incubated at 44.8°C and time points were evaluated at 0, 30, 60, 90, 150, 300, 600, 1200 and 1800 seconds. Panels show phosphorimages of urea gels separating Cas8’ nuclease products of ssDNA and ssRNA respectively. Each substrate was labelled with ^{32}P as indicated (•).

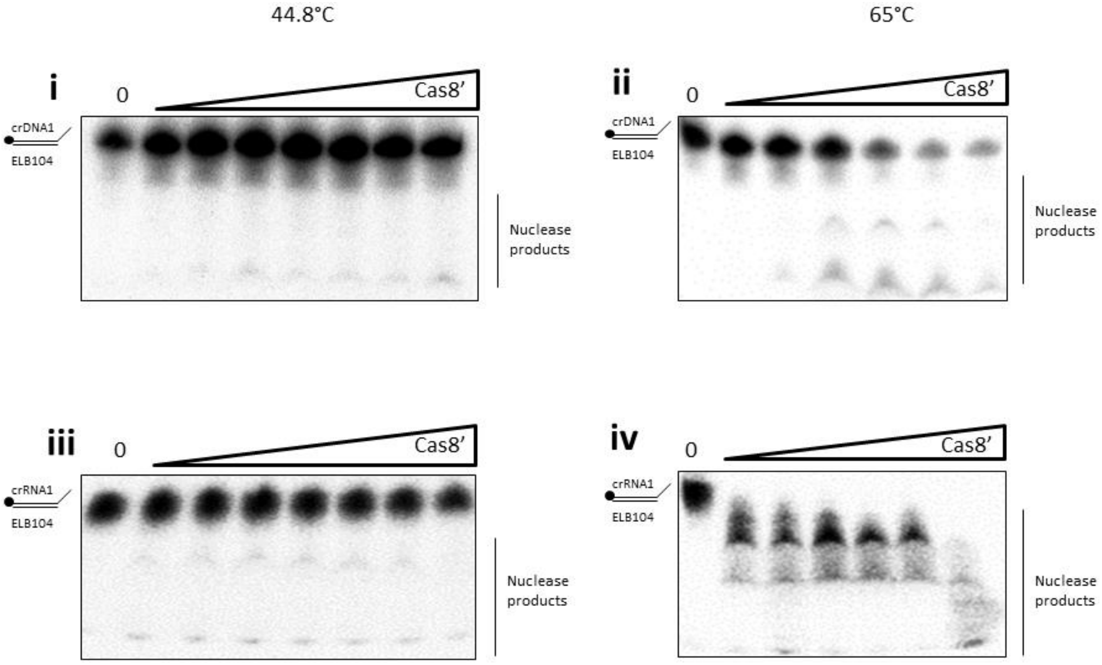


Figure 4-8. Nuclease assays of Cas8’activity in varied salt and temperature reaction conditions. Newly purified Cas8’ protein that was stored in 500mM KOAc and tested for nuclease activity on 3’ DNA and RNA flapped substrates at two temperatures: 44.8° and 65°C. In panels i-iv Cas8’ was used at 0, 50, 100, 200, 400, 600, 800 and 1000nM with 5nM of DNA in reactions containing 10 mM $MgCl_2$ and 125mM KOAc. Phosphorimages of urea gels separating Cas8’ nuclease products of 3’DNA flap (I and ii) and 3’RNA flap (iii and iv) are show. Each substrate was end labelled with ^{32}P as indicated by (•).

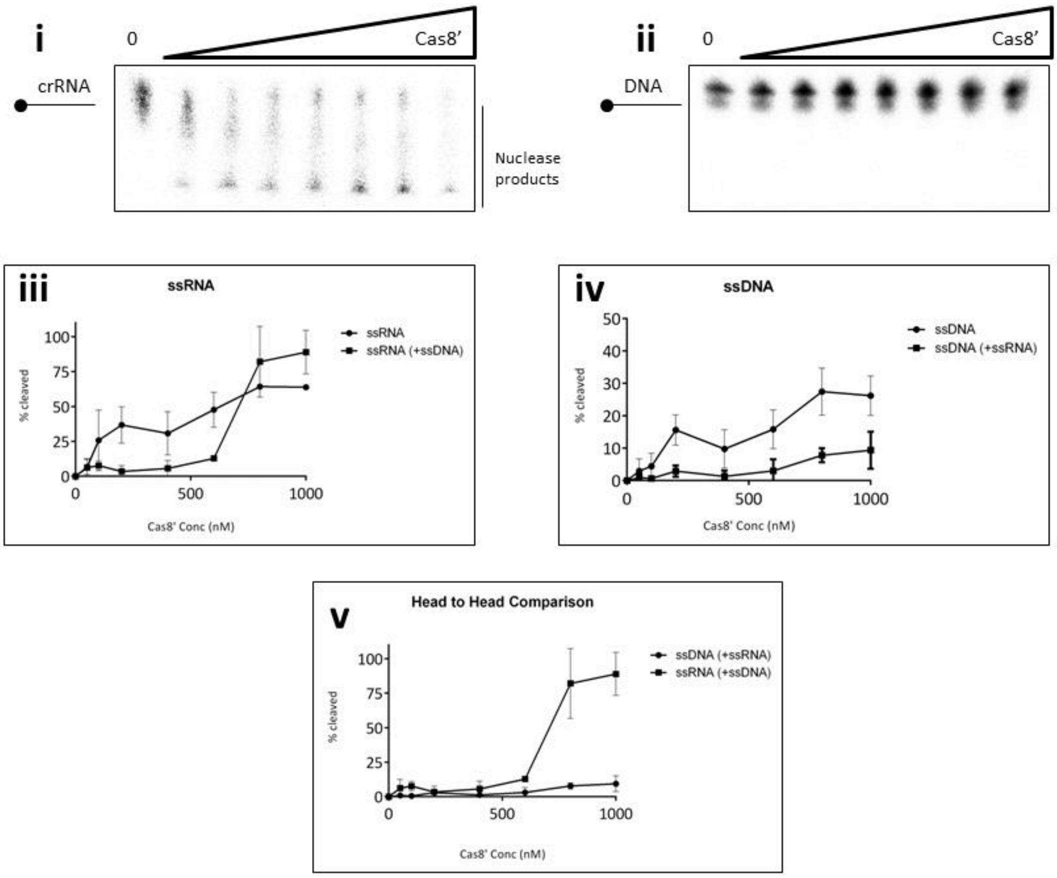


Figure 4-9. Head-to-head assays titrating Cas8' into labelled ssRNA and ssDNA or *vice-versa*. In panels I and ii, Cas8' was used at 0, 50, 100, 200, 400, 600, 800 and 1000nM with 5nM each of labelled and unlabelled substrate in reactions containing 10 mM $MgCl_2$ and 125mM KOAc. Panels i and ii show phosphorimages of urea gels separating Cas8' nuclease products of ssDNA (crDNA1) and ssRNA (crRNA1), respectively. The substrate end was labelled with ^{32}P as indicated by (•). Comparisons of cleavage activity of labelled ssRNA with non-labelled ssDNA (iii) and *vice versa* (iv). (v) Final comparison showing that ssRNA was degraded with ssDNA but cleavage of ssDNA was inhibited when ssRNA was present.

It seemed significant that despite the varying assay conditions used to test Cas8' nuclease activity, RNase function was in each Case the affinity and nuclease activity were higher for RNA than DNA. To test the relative activities of Cas8' on RNA (crRN1) and DNA (crDNA1) a head-to-head competition assay was used incubating, radiolabelled ssRNA (Figure 4-9, panels I and iii) with equimolar amounts of identical unlabelled DNA substrate, and *vice versa (*Figure 4-9, panels ii and iv). In these assays there were similar levels of RNA degradation, but when the DNA was labelled, the nuclease activity was lower, indicating that Cas8' preferentially binds and processes RNA (Figure 4-9, panel v).

4.2.3.2 Substrate specificity of Cas8'-mediated nucleolytic cleavage

In EMSAs, Cas8' was found to bind to 3'flapped structures of all RNA and DNA variations (Figure 4-1, panels viii, ix and x). Therefore, its nuclease activity was tested to identify a structural preference, if any, for cleaving specific substrates. Indeed, there were distinct differences in Cas8' nuclease activity when it was titrated into various flapped substrates. Cas8' showed ~5% cleavage activity on 5' and 3'DNA flaps at 1000nM (Figure 4-10, panels i and iii), and no activity on 3'RNA/RNA flaps at any of the concentrations tested (Figure 4-10, panel iv). However, there was 100% cleavage of the initial 3'RNA flaps already at 50nM Cas8' (Figure 4-10, panel ii). Importantly, Cas8' generated discrete cleavage products with both DNA and RNA substrates in these reactions. For DNA three and RNA four distinct cleavage products are seen. In previous published work on Cas8', DNA endonuclease activity was lost with the addition of ATP. Therefore, the nuclease activity of Cas8' on RNA flaps was also tested in the presence of ATP. The addition of ATP enhanced endonucleolytic processing of the RNA to the final degraded product, even at low (50nM) concentrations only one product is observed (Figure 4-11). Whereas, in previous assays without ATP this only occurred at 800 or 1000nM concentrations of Cas8' (Figure 4-10 panel ii).

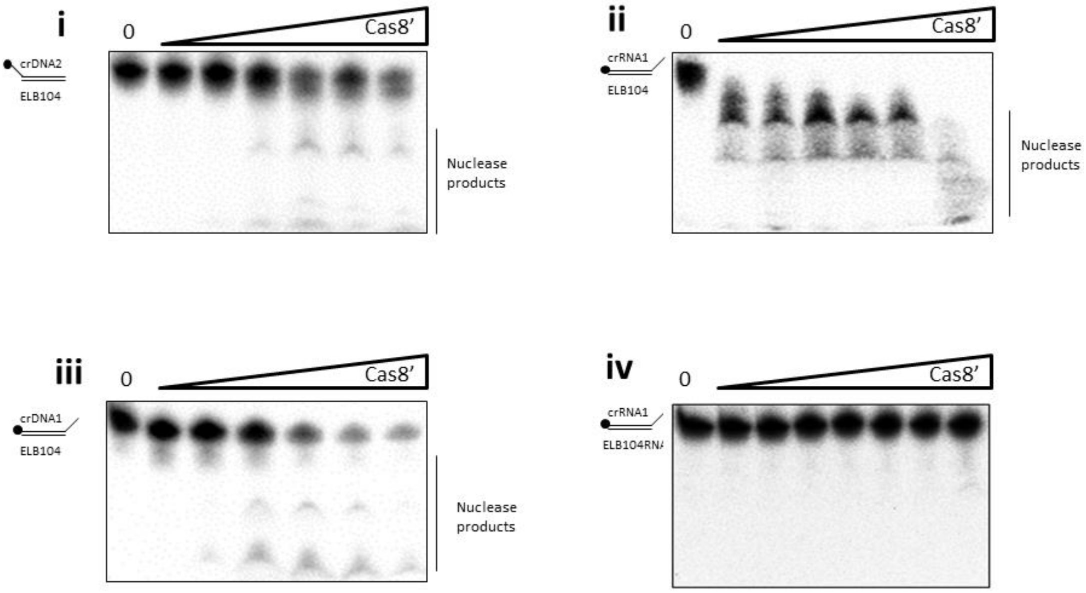


Figure 4-10. Nuclease assays titrating Cas8' into various flapped substrates. In panel's i-iv Cas8' was used at 0, 50, 100, 200, 400, 800 and 1000nM with 5nM of DNA in reactions containing 10 mM $MgCl_2$ and 125mM KoAc. Panels i-iv show phosphorimages of urea gels separating Cas8' nuclease products of 5'DNA flap (i), 3'RNA flap (ii), 3'DNA flap (iii) and 3'RNA/RNA flap (iv), respectively, each end labelled with ^{32}P as indicated by (•)

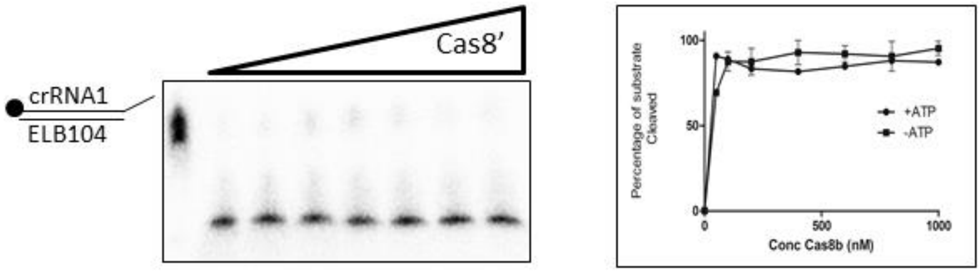


Figure 4-11. Nuclease assay comparison of Cas8' activity + or -ATP. Cas8' was used at 0, 50, 100, 200, 400, 600, 800 and 1000nM with 5nM of DNA in reactions containing 10 mM $MgCl_2$, 10mM ATP and 125mM KOAc, showing a phosphorimage of a urea gel separating Cas8' nuclease products of 3'RNA flap, end labelled with ^{32}P as indicated by (•). Comparison of quantified overall cleavage from initial 3'RNA substrate +/- ATP, using data from **Figure 4-10** with standard error.

4.2.3.2.1 Summary

The *in vitro* analysis of Cas8 nuclease activity suggested a role in processing of single stranded RNA (Figure 4-6, panel ii and Figure 4-10, panel ii). Moreover, assessment of different reaction conditions revealed acetate salts and a temperature of 65°C (Figure 4-8, panels ii and iv) stimulated Cas8' activity, in line with expected physiological conditions of *Mth*.

4.2.4 ATPase activity of Cas8’

ATPase assays were carried out using the Malachite Green colorimetric analysis method, detecting phosphate group liberation from ATP (*282*). Cas8’ was titrated into the reaction mixture and hydrolysed ATP was quantified against a standard curve of known phosphate concentrations. ATPase activity was detected in a Cas8’ concentration dependant manner, Figure 4-12i. A time course analysis of Cas8’ ATPase activity was carried out with a low concentration of Cas8’, as the reaction would not saturate. 50nM of Cas8’ was used to understand the kinetics of Cas8’ ATP hydrolysis, (Figure 4-12ii). This time course analysis showed that ATPase assays have many variables that make time course analysis difficult. Therefore, end point assays were carried out to compare different assay conditions with Cas8’. The relative ATPase activity of Cas8’ was compared with the addition of 5nM ssDNA or ssRNA. ATPase activity was enhanced in the presence of ssDNA but not ssRNA, (Figure 4-12iii).

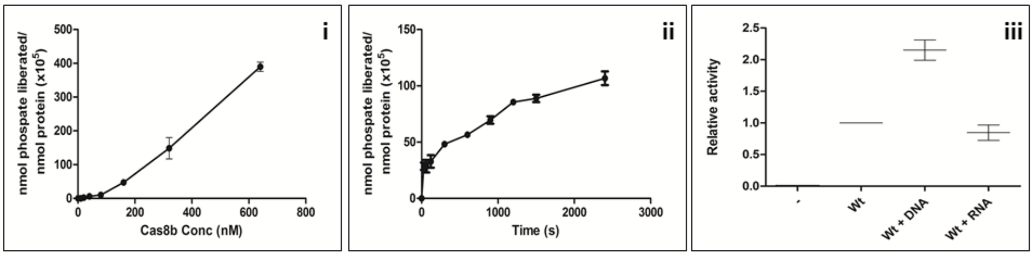


Figure 4-12. ATPase assays of wild-type Cas8’ activity with or without ssDNA or ssRNA. Cas8’ ATPase activity was analysed by Malachite Green colorimetric Assay (*282*). (i) A concentration titration was initially carried out containing 0, 5, 10, 20, 40, 80, 160, 320 and 640nM Cas8’. Reactions were terminated by the addition of Malachite Green solution after 10 minutes and OD measured at 595nm. (ii) A single concentration of Cas8’ (50nM) was used in a time course analysis over 45 minutes. Time points were taken at 0, 30, 60, 150, 300, 600, 900, 1200, 1500 and 2400 seconds. All assay points show a mean value from duplicate assays with standard error. (iii) Cas8’ ATPase hydrolysis when in the presence of 5nM ssDNA or ssRNA. Each reaction was incubated for 1500 seconds with 50nM Cas8’. Assays were carried out in duplicate and calculated as an activity relative to Cas8’ wild-type protein without oligonucleotide.

4.2.5 Targeted mutagenesis of Cas8’ to investigate its catalytic activity

RNA nuclease activity of *Mth* Cas8’ was investigated in more detail by generating Cas8’ mutations that we predicted may abrogate catalytic function *in vitro*. At this stage, work from the Marchfelder laboratory had demonstrated by genetics analysis that *Hvo* Cas8 was essential for CRISPR interference (*280*). A collaboration was set up between Dr. Edward Bolt, Dr. Anita Marchfelder, Karina Kaas and Simon Cass to assess the roles of Cas8 in more detail, including introduction of conserved amino acid residue mutations in *Mth* and *Hvo* Cas8 and analysis of the effect of said mutations.

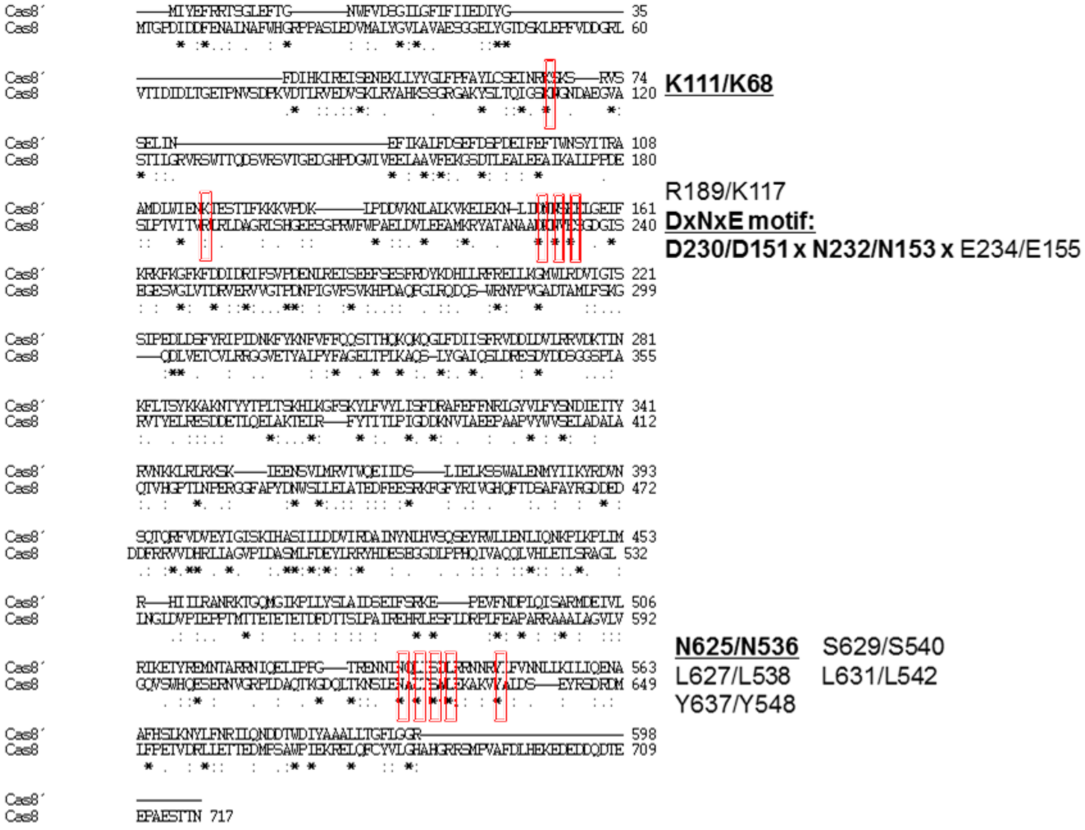


Figure 4-13. Clustal X alignment of amino acid sequences of *Haloferax volcanii* (*Hvo*) Cas8 and *Methanothermobacter thermautotrophicus* (*Mth*) Cas8'. Amino acids highlighted to the right of the alignment were conserved in both sequences and were subject to genetic analysis in *Haloferax* and those underlined in bold were studied both genetically *in vivo* and biochemically *in vitro*.

Cas8 family proteins are diverse, leading to low overall sequence conservation and to lack of any recognizably conserved common domain when assessed by presently available sequence databases. This is typified by the Cas8 proteins from *Mth* and *Hvo* which share only 30% amino acid identity. However, pairwise alignment of *Hvo* and *Mth* Cas8, did reveal conserved amino acid motifs which could be targeted by mutagenesis (Figure 4-13).From the sequence homology a region of conservation between *Mth* Cas8' αα 15-155 and 530-550 were the targets of mutants. Specifically, conserved residues within *Mth* Cas8' amino acid sequences 151-155 and 530-550 were selected: D151(G), N153(A), E155(A), N536(A), S540(A), L542(A) and Y548(A) to be mutated to alanine(A) or glycine (G), used together with the previously available K68A and K117A point mutants (*269*).

All the Cas8' point mutations were successfully generated except the Y548A mutation, which could not be made despite multiple attempts, for reasons unknown. The point mutants were then overexpressed as soluble proteins, except the S540A mutant. In situations where no protein expression was detected the alanine encoding codon was subsequently mutated to encode glycine. The D151G mutant generated soluble protein that could be purified whereas mutation of S540A to A540G yielded no protein expression for unknown reasons. The remaining *Mth* Cas8' proteins K68A, K117A, D151G, N153A, E155A, N536A and L542A were expressed and purified using the same method as for wild-type Cas8' (Figure 4-14, see section 3.2.3).

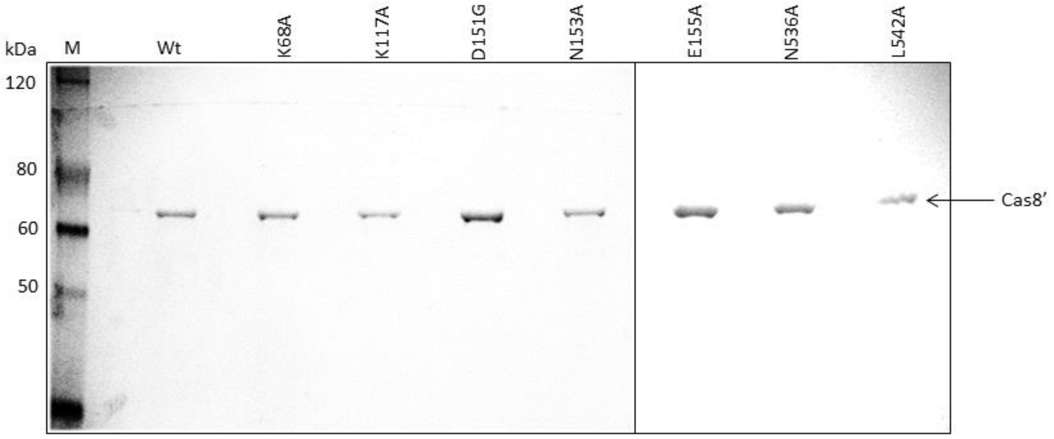


Figure 4-14. SDS-PAGE analysis of *Mth* Cas8' wild-type and mutant purified proteins. Following identically performed purifications for all proteins 1-2µg of each His_6-Cas8' protein was assessed by SDS-PAGE and stained by Coomassie Brilliant Blue.

4.2.5.1 Analysis of ATPase activity of Cas8' wild-type and mutant proteins

The ATPase activities of the purified Cas8' mutant proteins were determined and expressed relative to the ATPase activity obtained for the wild-type Cas8'protein which was set to 1 (Figure 4-15). K117A and D151A both appeared to have severely reduced ATPase activity with approximately 0.2 of the relative activity of the Cas8' wild-type protein. K68A, N153A, E155A, N536A and L542A mutants all showed similar activity to the wild-type protein varying from 0.78 (L542A) to 1.4 (N153A).

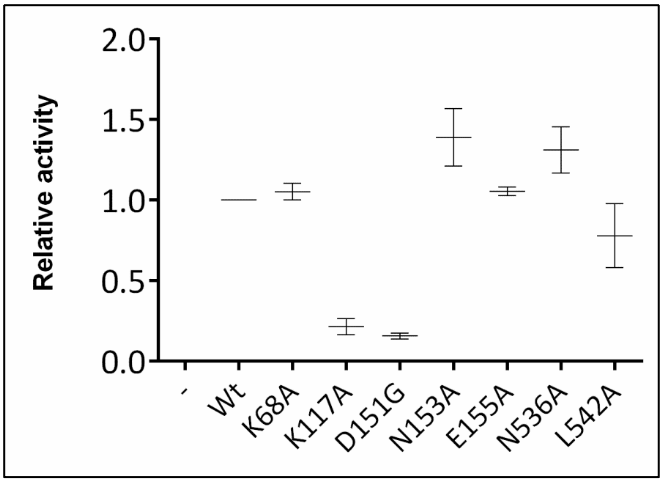
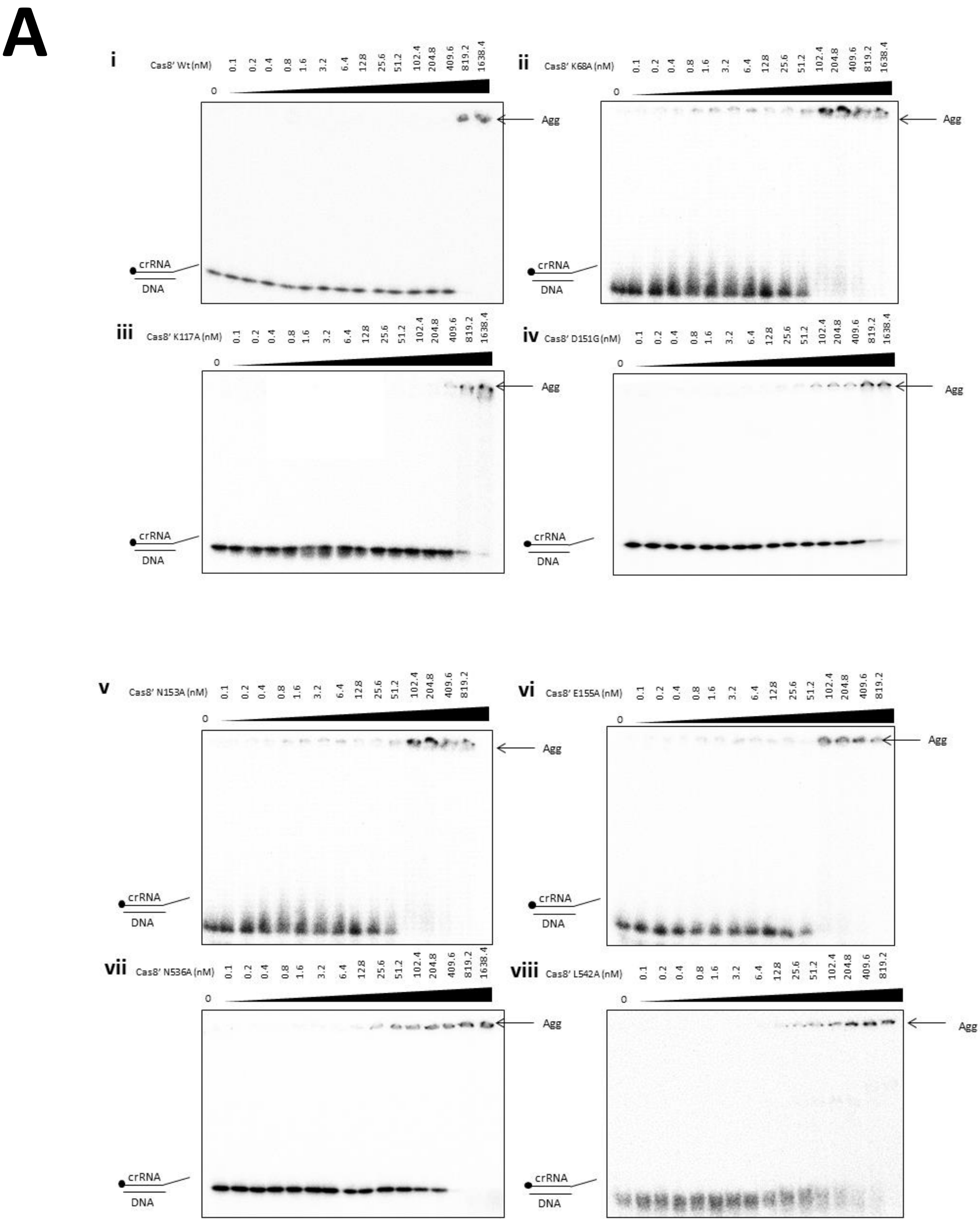


Figure 4-15. ATPase assay of Cas8' wild-type protein and Cas8' mutant proteins. 50nM of Cas8' proteins were tested for ATP hydrolysis activity by Malachite Green colorimetric assay as in **Figure 4-12**iii. Relative activity of mutants was calculated using the wild-type protein ATPase activity as reference. Each mutant was assayed in duplicate and whiskers show standard deviation from the mean.

4.2.5.2 Binding of mutant Cas8' proteins to 3'RNA/DNA Flap substrate

Binding assays with wild-type Cas8' showed preferential binding to substrates with flapped regions, particularly flapped structures and D- and R-loops. Cas8' also demonstrated nuclease activity, of particular interest was the RNase activity on a 3' RNA/DNA flap. To this end, wild-type Cas8' and mutant proteins were tested for binding the 3'RNA/DNA flap substrate. Interestingly, the binding characteristics of mutant proteins differed to wild-type. The amount of substrate bound was expressed in percentage and plotted against concentration of Cas8' protein (Figure 4-16) with the resulting calculated K_D values tabulated (Table 13. K_D values for Cas8' proteins binding to 3’RNA/DNA flap substrate from EMSA analysis. **K_Ds were determined from triplicate EMSAs quantified by 2D densitometry.**). These assays were carried out prior to the identification of the optimal salt condition. Therefore, the reactions are in NaCl containing buffers rather than the optimised KOAc buffers. As a result there are differences seen between the wild-type Cas8’ K_D values calculated here when compared to previous values, shown in Figure 4-1. As expected following the mutation of conserved residues mutant Cas8’ proteins manifested with varied binding affinities to the substrate when compared to the wild-type protein. The mutant proteins K117A (Figure 4-16iii) and D151G (Figure 4-16iv) bound the substrate most similarly to wild-type, though K_D values showed slightly reduced binding affinity for K117A, and even more so for D151G. In contrast, all other Cas8’ mutant proteins presented with improved binding to the 3’RNA/DNA substrate. K_D values indicated 3-fold increase in binding affinity of the N536A mutant, with other mutants following in the order of increasing affinities, L542A, E155A and K68A (Figure 4-16viii, vi and ii). The highest, 17-fold increase in binding affinity was observed for the mutant N153A (Figure 4-16v).


A
i
Cas8' Wt (nM)
0
0.1 0.2 0.4 0.8 1.6 3.2 6.4 12.8 25.6 51.2 102.4 204.8 409.6 819.2 1638.4
Agg
crRNA
DNA
ii
Cas8' K68A (nM)
0
0.1 0.2 0.4 0.8 1.6 3.2 6.4 12.8 25.6 51.2 102.4 204.8 409.6 819.2 1638.4
Agg
crRNA
DNA
iii
Cas8' K117A (nM)
0
0.1 0.2 0.4 0.8 1.6 3.2 6.4 12.8 25.6 51.2 102.4 204.8 409.6 819.2 1638.4
Agg
crRNA
DNA
iv
Cas8' D151G (nM)
0
0.1 0.2 0.4 0.8 1.6 3.2 6.4 12.8 25.6 51.2 102.4 204.8 409.6 819.2 1638.4
Agg
crRNA
DNA
v
Cas8' N153A (nM)
0
0.1 0.2 0.4 0.8 1.6 3.2 6.4 12.8 25.6 51.2 102.4 204.8 409.6 819.2
Agg
crRNA
DNA
vi
Cas8' E155A (nM)
0
0.1 0.2 0.4 0.8 1.6 3.2 6.4 12.8 25.6 51.2 102.4 204.8 409.6 819.2
Agg
crRNA
DNA
vii
Cas8' N536A (nM)
0
0.1 0.2 0.4 0.8 1.6 3.2 6.4 12.8 25.6 51.2 102.4 204.8 409.6 819.2 1638.4
Agg
crRNA
DNA
viii
Cas8' L542A (nM)
0
0.1 0.2 0.4 0.8 1.6 3.2 6.4 12.8 25.6 51.2 102.4 204.8 409.6 819.2
Agg
crRNA
DNA

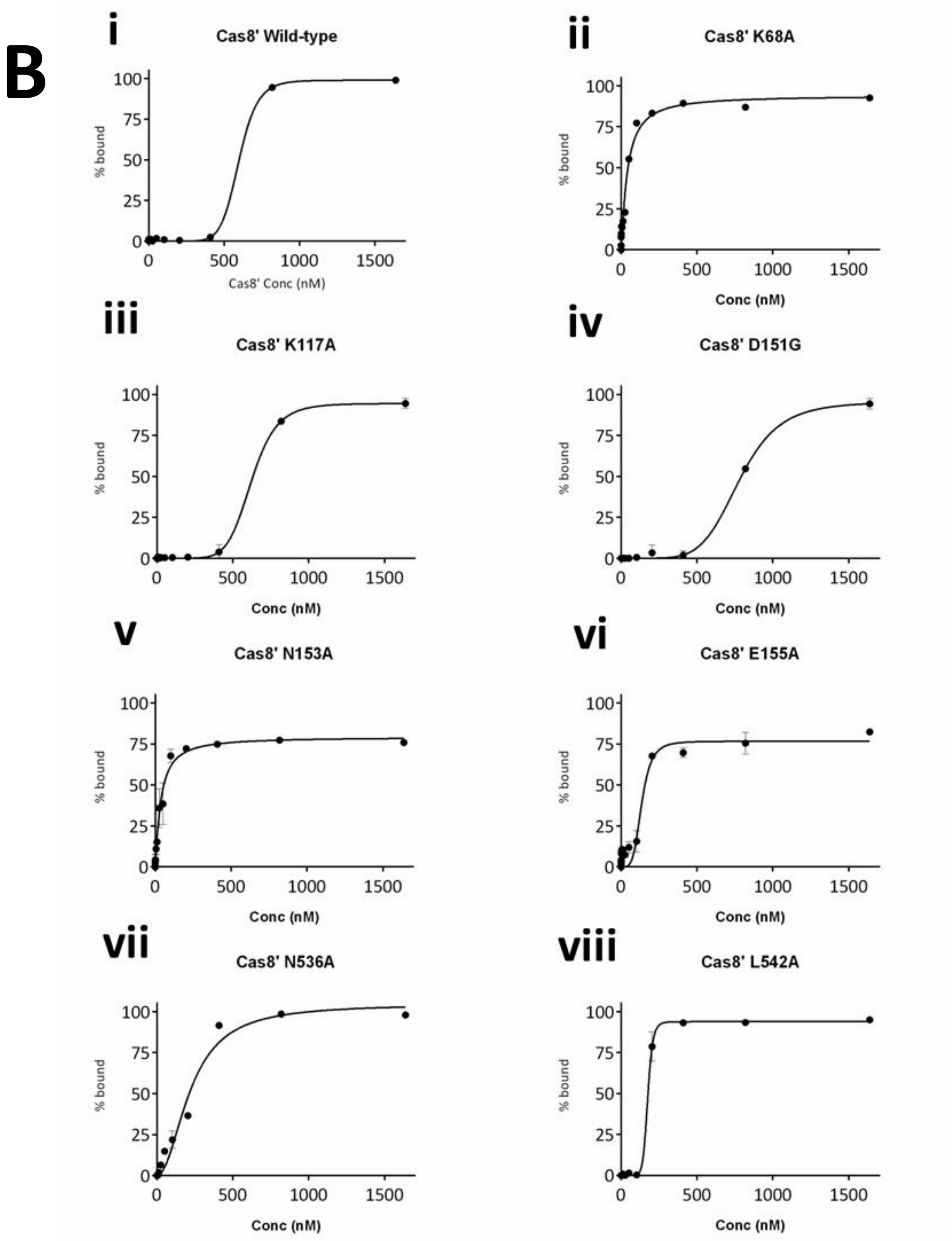


Figure 4-16. EMSA analysis titrating Cas8’ into 3’RNA/DNA substrate. Cas8’ was used at 0, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2, 6.4, 12.8, 25.6, 51.2, 102.4, 204.8, 409.6, 819.2 and 1638.4nM with 5nM substrate in reactions containing 10mM EDTA. Each protein was assayed in at least duplicate (usually triplicate): Wild-type (i), K68A (ii), K117A (iii), D151G (iv), N153A (v), E155A (vi), N536A (vii) and L542A (viii). Section A shows phosphorimages of native TBE gels separating bound and unbound substrate species and aggregates and in gel bands respectively. Section B shows binding isotherms with percentage of substrate shifted, as quantified 2D densitometry plotted against Cas8’ concentration. Binding isotherms were plotted with standard error and the calculated K_D values summarised in Table 13.

.

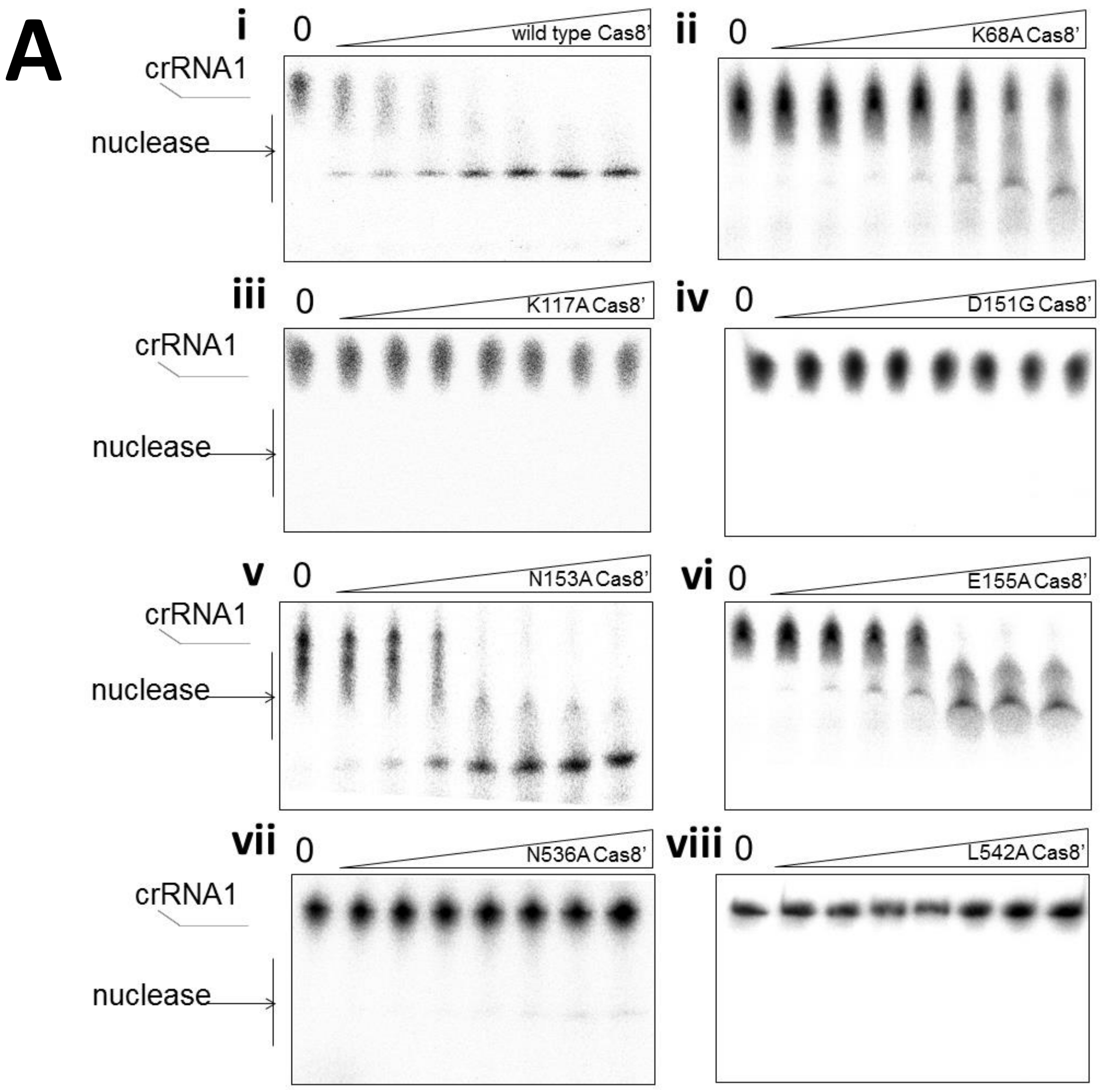
Table 13. K_D values for Cas8' proteins binding to 3'RNA/DNA flap substrate from EMSA analysis. K_Ds were determined from triplicate EMSAs quantified by 2D densitometry.

Cas8'	**K_D**
Wild Type	598.7 ± 12.1 nM
K68A	94.1 ± 7.2 nM
K117A	623.6 ± 9.4 nM
D151G	779.7 ± 7.5 nM
N153A	36.3 ± 4.3 nM
E155A	135.4 ± 10.9 nM
N536A	222.2 ± 16.7 nM
L542A	175.2 ± 15.0 nM

4.2.5.3 RNA nuclease activity of Cas8' proteins

As each mutant protein either maintained or enhanced its binding affinity for the substrate compared to the wild type Cas8' protein, the next step was to examine the impact of the mutations on the nuclease activity of the mutant proteins. Specifically, RNA nuclease activity was assayed using the same 3'RNA/DNA substrate as in the prior binding affinity assays (EMSAs). Previous experiments showed that Cas8' nuclease activity was dependent on the inclusion of magnesium in the reaction conditions (see Figure 4.5). Therefore, also in these nuclease assays, EDTA was occluded and reactions were supplemented with magnesium chloride to allow the magnesium-dependent nuclease activity of Cas8' to be analysed. Given that at the time of these assays it had been determined by other experiments that potassium acetate salts potentiated the nuclease activity of Cas8', these tests were also carried out in potassium acetate instead of sodium chloride.

Here, representative denaturing urea gels showing the labelled RNA strand of the 3'RNA/DNA flap substrate and the resultant smaller-sized products of the nuclease activity of each mutant Cas8' protein are presented (Figure 4-17). Despite most of the mutants showed enhanced binding to the substrate, their nuclease activity was attenuated compared to the wild-type protein. The two mutants K117A and D151G which bound the substrate with lower affinity than the wild-type Cas' protein lacked nuclease activity entirely (Figure 4-17, panels iii and iv, respectively). On the other hand, the L542A mutant which had higher binding affinity was also nuclease defective (Figure 4-17viii). The remaining mutant proteins, arranged in the order of increasing binding affinities, N536A, E155A, K68A and N153A, all manifested with detectable nuclease activity. However, quantitative analysis revealed all activities were clearly lower than the wild-type protein, though to varying degrees. For example, the mutants N153A and E155A still successfully cleaved approximately 75% of the substrate, K68A approximately 50% whereas the activity N536A was almost abolished. This indicates that each of these residues plays a different role in maintaining the appropriate conformation of the Cas8' protein to correctly interact with and process the substrate.


A
i
0
wild type Cas8'
crRNA1
nuclease
ii
0
K68A Cas8'
iii
0
K117A Cas8'
crRNA1
nuclease
iv
0
D151G Cas8'
v
0
N153A Cas8'
crRNA1
nuclease
vi
0
E155A Cas8'
vii
0
N536A Cas8'
crRNA1
nuclease
viii
0
L542A Cas8'

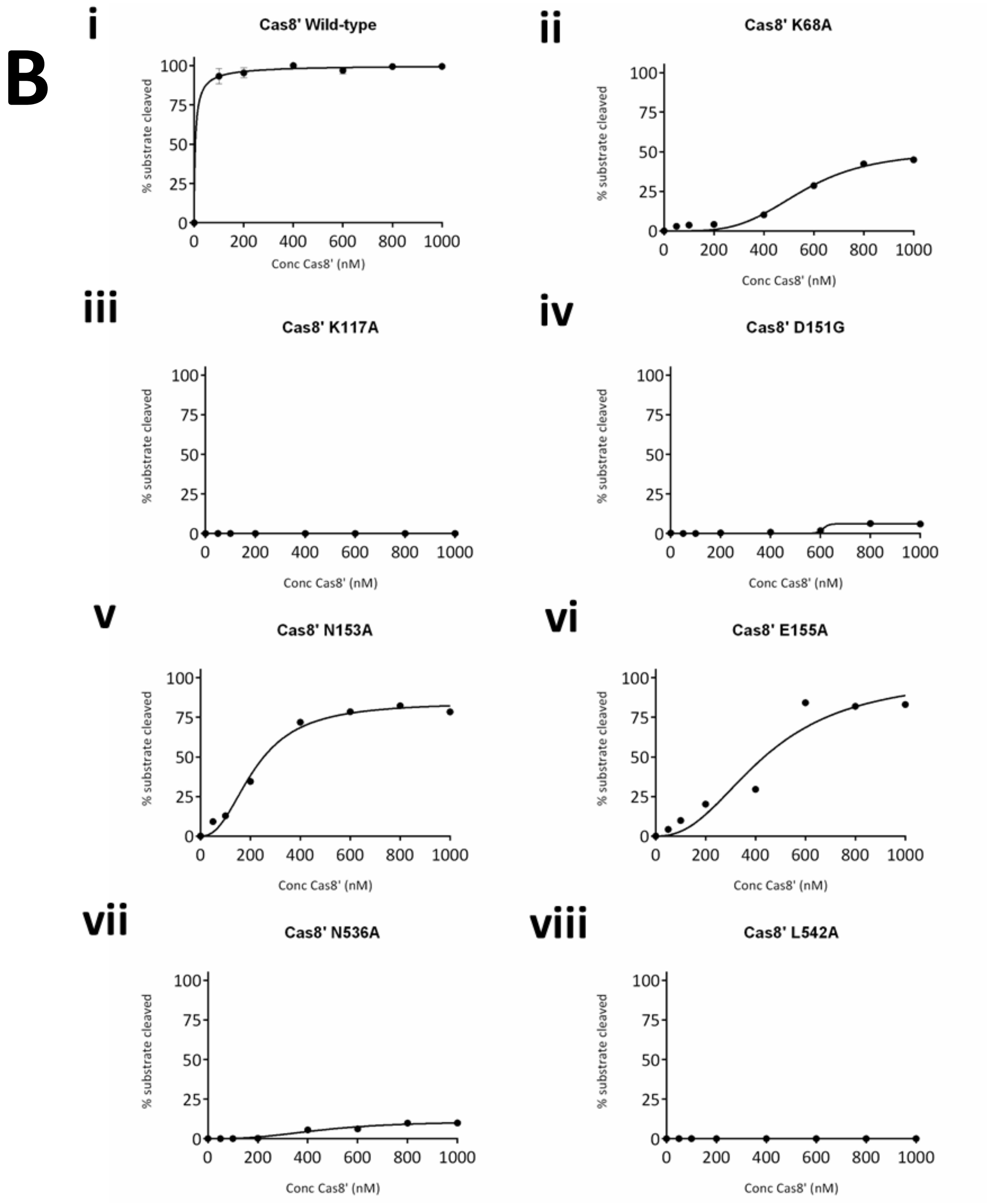
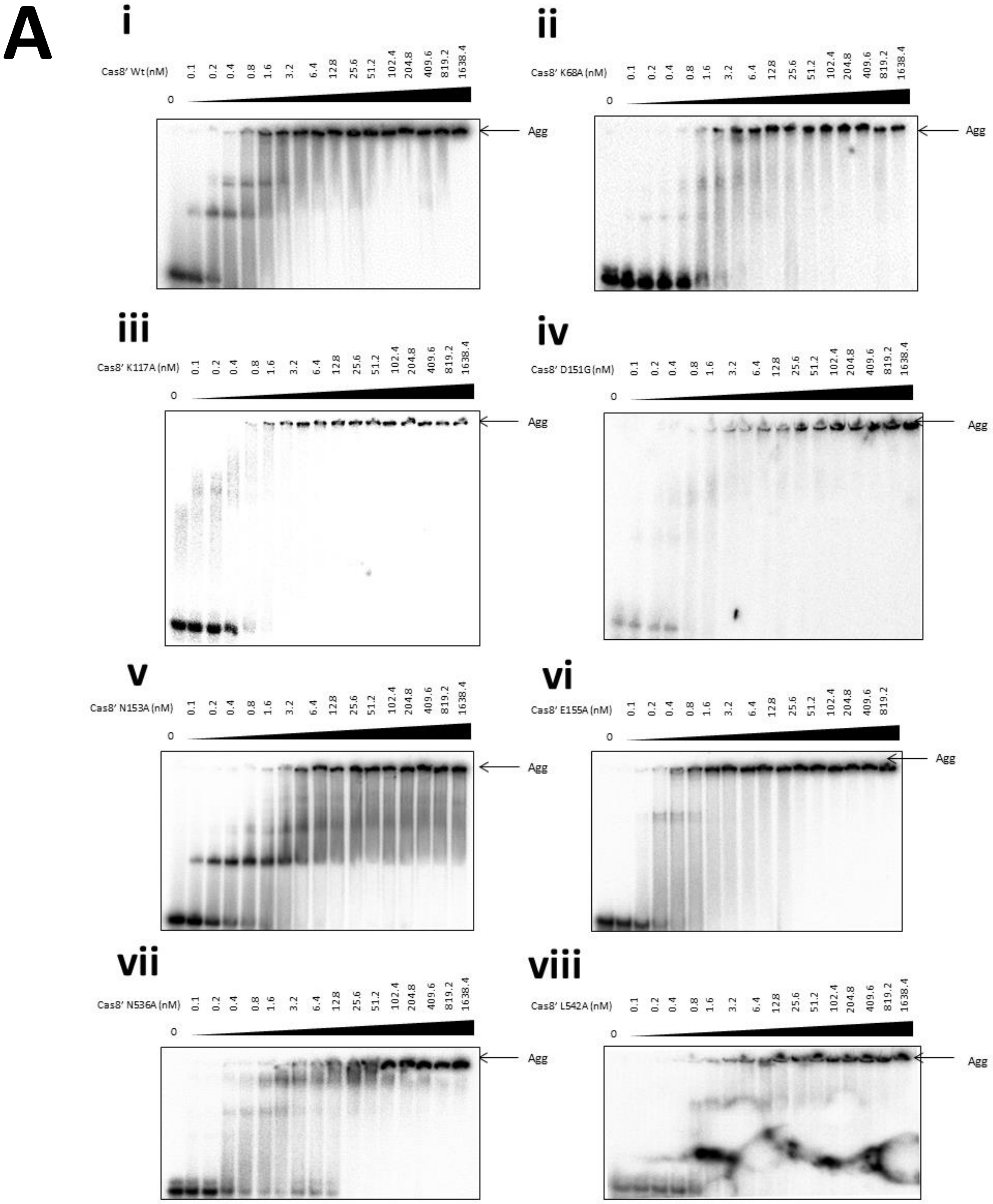


Figure 4-17. Nuclease assays titrating Cas8' wild-type and mutant proteins into 3'RNA/DNA flap substrate. Cas8' proteins were used at 0, 50, 100, 200, 400, 600, 800 and 1000nM with 5nM substrate in reactions containing 125mM KOAc and 10mM $MgCl_2$. Each protein was assayed in at least duplicate (usually triplicate). Section A displays representative gels of nuclease assays with: wild-type (i), K68A (ii), K117A (iii), D151G (iv), N153A (v), E155A (vi), N536A (vii) and L542A (viii). These are phosphorimages of urea gels separating Cas8' nuclease products of 3'RNA/DNA flap, each end labelled with ^{32}P as indicated by (•). Section B shows percentage of substrate degraded as determined by quantification using 2D densitometry with corresponding standard error.

4.2.5.4 *In vitro EMSA analysis of Mth Cas8' proteins binding to R-loops*

Previously, the recombinant *Mth* Cas8' mutant proteins were assessed by EMSA assays for binding to 3'RNA/DNA flap substrate (see section 4.2.5.3). Therefore, here these assays were recapitulated to evaluate the binding behaviour of the Cas8' mutant proteins to synthetic R-loops. Of note, the R-loop substrate used was inclusive of a PAM sequence. Representative EMSA gels and the resulting quantified binding isotherms can be seen in Figure 4-18A&B. Wild-type Cas8' bound the R-loop substrate with the affinity corresponding to K_D of 4.7 ± 0.6nM (Figure 4-1). All Cas8' mutants also showed binding to the generic R-loop substrate, albeit with lower affinity than that observed for the wild-type protein. The proteins that showed the largest decreases in binding affinity were K68A, E155A, K117A and L542A. The calculated K_D of these mutants gave an 8 to 10 fold change in binding affinity: 38.1 ± 1.5nM, 41.4 ± 1.2nM, 44.3 ± 1.4nM and 53.3 ± 9.2nM, respectively. These apparent disassociation constants were still relatively high and still generated some substrate-specific shift patterns typical of the wild-type protein. The remaining mutants bound the R-loop substrate with intermediate binding affinity, showing an approximate 2-3 fold decrease in their disassociation constants. The nuclease-defective D151G Cas8' mutant bound the R-loop substrate with a K_D of 16 ± 1.4nM, while the remaining two nuclease-active proteins had a slightly higher level of binding. Specifically, the mutants N153A had a K_D of 14.8 ± 6.6nM, and N536A a K_D of11.4 ± 1.3nM. Similarly to the binding to 3'RNA/DNA flap substrate, single point mutations of the *Mth* Cas8' protein affected the binding to synthetic R-loops. Nonetheless, all of the mutants could still interact with the substrate, albeit with increased apparent disassociation constants.


A
i
Cas8' Wt (nM)
0
0.1 0.2 0.4 0.8 1.6 3.2 6.4 12.8 25.6 51.2 102.4 204.8 409.6 819.2 1638.4
Agg
ii
Cas8' K68A (nM)
0
0.1 0.2 0.4 0.8 1.6 3.2 6.4 12.8 25.6 51.2 102.4 204.8 409.6 819.2 1638.4
Agg
iii
Cas8' K117A (nM)
0
0.1 0.2 0.4 0.8 1.6 3.2 6.4 12.8 25.6 51.2 102.4 204.8 409.6 819.2 1638.4
Agg
iv
Cas8' D151G (nM)
0
0.1 0.2 0.4 0.8 1.6 3.2 6.4 12.8 25.6 51.2 102.4 204.8 409.6 819.2 1638.4
Agg
v
Cas8' N153A (nM)
0
0.1 0.2 0.4 0.8 1.6 3.2 6.4 12.8 25.6 51.2 102.4 204.8 409.6 819.2 1638.4
Agg
vi
Cas8' E155A (nM)
0
0.1 0.2 0.4 0.8 1.6 3.2 6.4 12.8 25.6 51.2 102.4 204.8 409.6 819.2
Agg
vii
Cas8' N536A (nM)
0
0.1 0.2 0.4 0.8 1.6 3.2 6.4 12.8 25.6 51.2 102.4 204.8 409.6 819.2 1638.4
Agg
viii
Cas8' L542A (nM)
0
0.1 0.2 0.4 0.8 1.6 3.2 6.4 12.8 25.6 51.2 102.4 204.8 409.6 819.2 1638.4
Agg

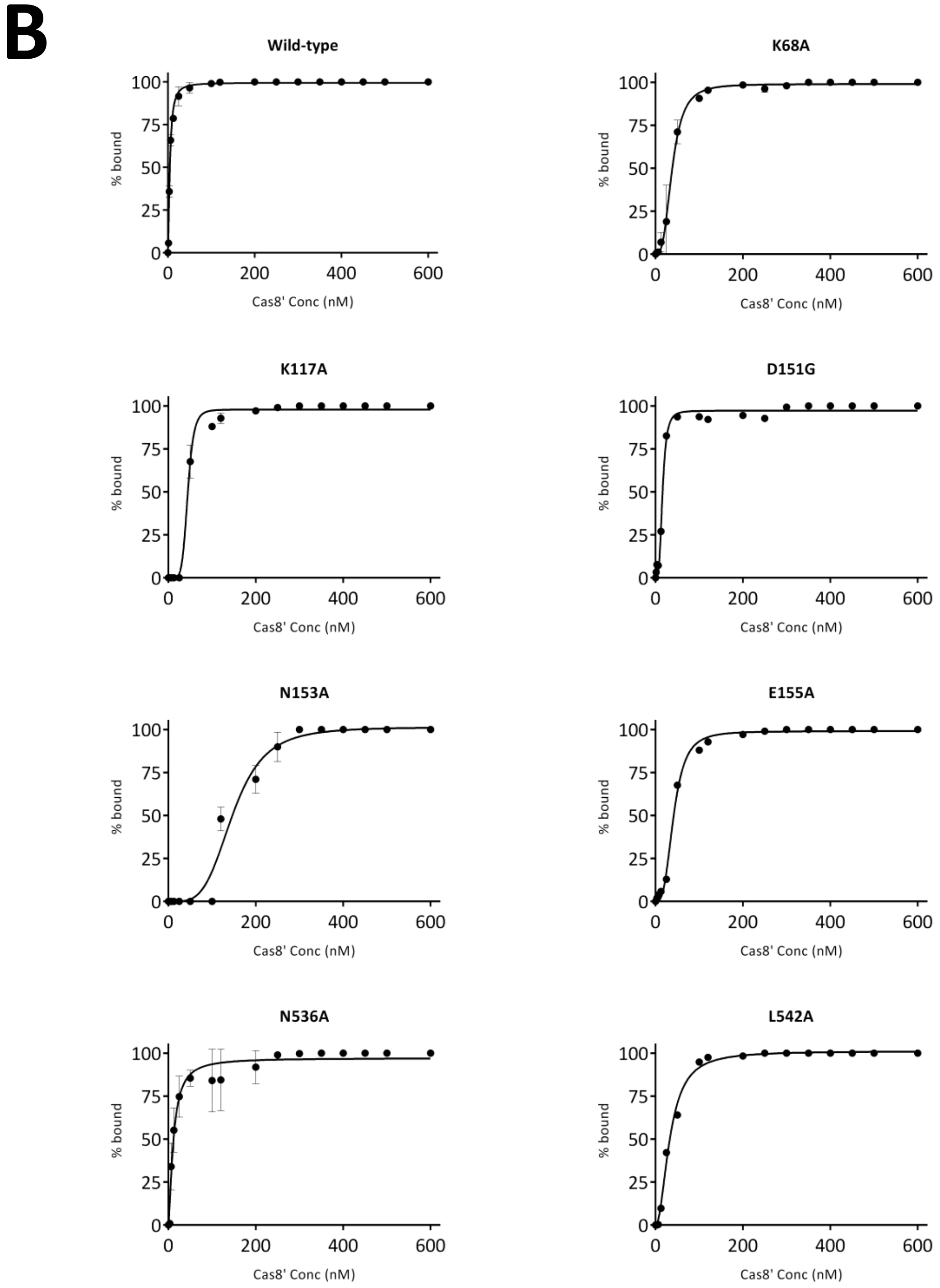


Figure 4-18. EMSA analysis titrating Cas8’ wild-type and the various Cas8’ mutants into generic R-loop substrate. Cas8’ was used at 0, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2, 6.4, 12.8, 25.6, 51.2, 102.4, 204.8, 409.6, 819.2 and 1638.4nM with 5nM substrate in reactions containing 125mM KOAc and 10mM EDTA. Each protein was assayed in at least duplicate (usually triplicate): wild-type (i), K68A (ii), K117A (iii), D151G (iv), N153A (v), E155A (vi), N536A (vii) and L542A (viii). Section A shows phosphorimages of native TBE gels separating bound and unbound substrate species and aggregates and in gel bands respectively. Section B shows binding isotherms with percentage of substrate shifted, as quantified by 2D densiometry (AIDA), with corresponding error bars plotted against Cas8’ protein concentration. The calculated K_D values are summarised in Table 14.

Table 14. K_D values for Cas8' proteins binding to R-loop substrate obtained from EMSA analysis. K_Ds were determined from at least duplicate EMSAs quantified by 2D densitometry.

Cas8'	K_D
Wild Type	4.7 ± 0.2nM
K68A	38.1 ± 1.6nM
K117A	44.3 ± 9.4nM
D151G	16.0 ± 1.4nM
N153A	14.8 ± 6.6nM
E155A	41.4 ± 1.2nM
N536A	11.4 ± 1.3nM
L542A	33.1 ± 1.3nM

Previously, the recombinant *Mth* Cas8' mutant proteins were assessed by EMSA assays for binding to 3'RNA/DNA flap substrate (see section 4.2.5.3). Therefore, here these assays were recapitulated to evaluate the binding behaviour of the Cas8' mutant proteins to synthetic R-loops. Of note, the R-loop substrate used was inclusive of a PAM sequence. Representative EMSA gels and the resulting quantified binding isotherms can be seen in Figure 4-18A&B. Wild-type Cas8' bound the R-loop substrate with the affinity corresponding to KD of 4.7 ± 0.6nM (Figure 4-1). All Cas8' mutants also showed binding to the generic R-loop substrate, albeit with lower affinity than that observed for the wild-type protein. The proteins that showed the largest decreases in binding affinity were K68A, E155A, K117A and L542A. The calculated KD of these mutants gave an 8 to 10 fold change in binding affinity: 38.1 ± 1.5nM, 41.4 ± 1.2nM, 44.3 ± 1.4nM and 53.3 ± 9.2nM, respectively. These apparent disassociation constants were still relatively high and still generated some substrate-specific shift patterns typical of the wild-type protein. The remaining mutants bound the R-loop substrate with intermediate binding affinity, showing an approximate 2-3 fold decrease in their disassociation constants. The nuclease-defective D151G Cas8' mutant bound the R-loop substrate with a KD of 16 ± 1.4nM, while the remaining two nuclease-active proteins had a slightly higher level of binding. Specifically, the mutants N153A had a KD of 14.8 ± 6.6nM, and N536A a KD of11.4 ± 1.3nM. Similarly to the binding to 3'RNA/DNA flap substrate, single point mutations of the *Mth* Cas8' protein affected the binding to synthetic R-loops. Nonetheless, all of the mutants could still interact with the substrate, albeit with increased apparent disassociation constants.

4.2.5.5 In vivo genetic analysis of conserved Cas8 point mutations in Haloferax volcanii

Parallel *in vivo* studies on the function of Cas8 in *Haloferax volcanii* CRISPR interference were carried out entirely by Karina Haas from the collaborating research group of Dr. Anita Marchfelder at the University Of Ulm, Germany. In Karina's study, point mutations in conserved residues, analogous to the mutations in *Mth* Cas8' generated as part of this study, were made in Hvo Cas8. To assess the competence of the mutant Cas8 proteins to adequately activate CRISPR interference, their functionality was tested via heterologous expression in a stable strain lacking wild-type Cas8 protein. In brief, Δ*Cas8* cells were first transformed with a plasmid (pTA927) containing either wild type *Cas8* (Cas8^{+}), or a mutant *Cas8*. Next, to provoke an immune reaction, strains were transformed with a second plasmid (pTA352) carrying a protospacer flanked by either PAM3 (TTC) or PAM9 (ACT), hence mimicking an invasive element. Only Cas8 proteins with sufficient functionality were then able to organize a functional Cascade targeting Cas3 for degradation of the invading plasmid. After a set time of incubation, the cultures were then plated onto media that selected for a marker contained on the second "invading" plasmid. If the CRISPR system of interference was active then this plasmid would have been destroyed, resulting in no colonies being observed on the plate. Based on the transformation efficiency of such cultures, the interference capability of each Cas8 mutant was then scored as either "reduced interference", "interference" or "no interference". Each assay was repeated in triplicate and only 100-fold differences in transformation efficiency were considered to be ineffective interference. The resulting data were collected and analysed fully by Karina Haas and can be seen presented in Table 15. All Cas8 mutant proteins presented with high transformation rates which suggested that the second "invading" plasmid was not being degraded by the CRISPR immune response due the reduced the efficacy of interference.

Of note, the results of this assay need to be interpreted with caution as spontaneously occurring mutations in essential DNA sequences might give false positive interference results. Indeed, the key role of Cas8 proteins in interference was discovered by genome sequencing of one such mutation (*280*).

Table 15. Effect of Cas8 mutations on plasmid interference in *Haloferax volcanii*. Δ*Cas8* cells were first transformed with plasmid (pTA927) containing either wild type *Cas8* (Cas8$^+$), or a mutant *Cas8* that encodes the amino acid substitution listed. To provoke an immune reaction, strains were transformed with a second plasmid (pTA352) carrying a protospacer flanked by either PAM3 (TTC) or PAM9 (ACT). Reduction of transformation rate was determined relative to transformation with a plasmid lacking protospacer-PAM sequence. Reduction in efficiency of transformation in the range 0.05-0.5 was classified as "reduced interference" and reduction by less than 0.04 as "interference". "No interference" indicates that transformation efficiency was high, comparable to plasmid lacking PAM-protospacer. This work is exclusively the work of Karina Haas, from Dr. Anita Marchfelder's research group at the University of Ulm, Germany.

First transformation plasmid	**Reduction of transformation rate with invader plasmid by factor**	
	pTA352-PAM3	**pTA352-PAM9**
---	no interference	no interference
pTA927-*Cas8b*	0.0003 (interference)	0.0006 (interference)
pTA927-K111A	no interference	no interference
pTA927-R126A	no interference	no interference
pTA927-R189A	0.007 (interference)	0.01 (interference)
pTA927-D230A	no interference	no interference
pTA927-N232A	0.2 (reduced interference)	0.004 (interference)
pTA927-E234A	0.005 (interference)	0.004 (interference)
pTA927-N625A	no interference	no interference
pTA927-L627A	no interference	no interference

pTA927-S629A	0.02 (interference)	0.01 (interference)
pTA927-L631A	0.001 (interference)	0.007 (interference)
pTA927-Y637A	no interference	no interference

4.2.5.6 *Characterisation of Cas8' N153A binding to generic substrates and sensitivity to PAM*

In vivo analysis of *Hvo* Cas8 mutants in interference reactions, using two "invader" plasmid versions carrying a protospacer flanked by different PAM sequences identified that mutant Cas8 N232A could support interference against one "invader "plasmid but not the other. Plasmids used in these assays had one of six possible *Hvo* PAMs, which indicated that Cas8 N232A could discern only some PAM sequences, but not others, which would have then reflected in variable interference that was observed. Therefore, to understand whether this residue played a similar role in PAM recognition in *Mth* Cas8', the equivalent mutant N153A protein was analysed for its ability to bind different DNA PAMs. These were determined from bioinformatics analysis of *Mth* spacers (see Figure 4-2) and here the main PAM (CCC) and two others (PAM2-CTC and PAM3-TTG) were tested in binding assays (EMSAs). However, first the substrate for which the N153A mutant had the highest affinity had to be determined, which was then to be carried forward for PAM sensitivity testing. Thus, the N153A mutant was tested for binding to generic RNA and DNA substrates (linear duplex, flapped duplex and an R-loop) and compared to the wild-type Cas8' protein. The binding isotherms of either wild-type Cas8' or N153A mutant with the individual substrates can be seen in Figure 4-19 with the resulting calculated K_D values summarized in Table 15. These binding assays showed that the N153 mutant bound to the duplex substrate with lower affinity than the wild type protein (81.8 ± 2.8nM versus 20.4 ± 0.6nM) (Figure 4-19 panels i and ii). Next, given that the best-fit line of binding isotherms to flayed duplex was linear for both proteins, the K_D values could not be calculated in this instance and the binding affinities could not be established and subsequently compared (Figure 4-19, panels iii and iv). Finally, N153A had a high affinity for binding to R-loops which was also comparable to that of the wild-type Cas8’ (5.3 ± 0.3nM versus 4.7 ± 0.2nM) making it the ideal candidate substrate (Figure 4-19, panels v and vi). Therefore, the ability of N153A to recognize different PAM sequences was examined using R-loop substrate variations either lacking PAM or including one of three different selected PAM sequences.

Wild-type Cas8’ bound to R-loops with (CCC) or without (AAA) PAM with apparent dissociation constants of 5.3nM and 40.7nM, respectively, demonstrating an 8-fold increase in binding affinity in the presence of a PAM sequence (See Figure 4-3). However, the N153A mutant Cas8’ displayed opposing behaviour, binding to R-loops with PAM with lower affinity (K_D = 31.6nM) than to R-loops without a PAM sequence (K_D= 12nM) (Figure 4-20i and ii, Table 16). Thus, Cas8’ N153A binding to R-loop substrates incorporating PAM2 (CTC) or PAM3 (TTG) were also assessed and compared to the pattern of binding affinity previously obtained for the wild type Cas8’ protein (see Figure 4-4 iii and iv) . From these assays it became evident that the N153A mutation completely altered the affinity at which the Cas8’ protein bound the different R-loops. The preferential PAM binding of the wild type protein was determined to be PAM (CCC) > PAM3 (TTG) > PAM2 (CTC) > no PAM (AAA), whereas the N153A Cas8’ trend was shown to be no PAM (AAA) > PAM2 (CTC) > PAM (CCC) > PAM3 (TTG), as deduced by comparing the respective K_D values (Table 16).

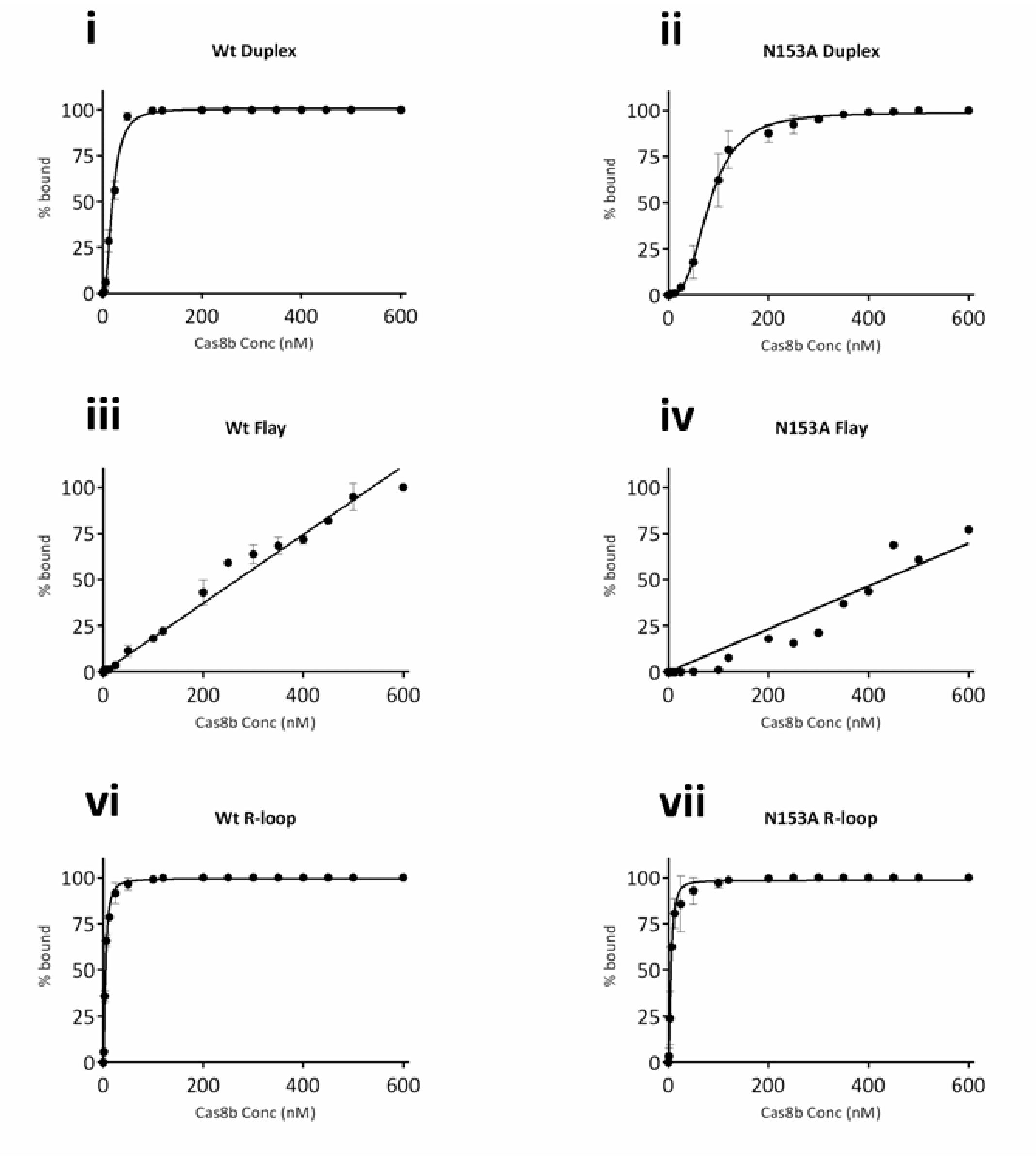


Figure 4-19. EMSA analysis of titrating wild-type Cas8' and point mutant N153A to generic DNA and RNA substrates. Cas8' proteins were used at 0, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2, 6.4, 12.8, 25.6, 51.2, 102.4, 204.8, 409.6, 819.2 and 1638.4nM with 5nM substrate in reactions containing 125mM KOAc and 10mM EDTA. Each pair of binding isotherms compares Cas8' wild-type to the N153A point mutant as a percentage of substrate shifted quantified by 2D densitometry. Binding isotherms allowed determination of with standard error and K_Ds. A summary of K_Ds determined can be seen in Table 16.

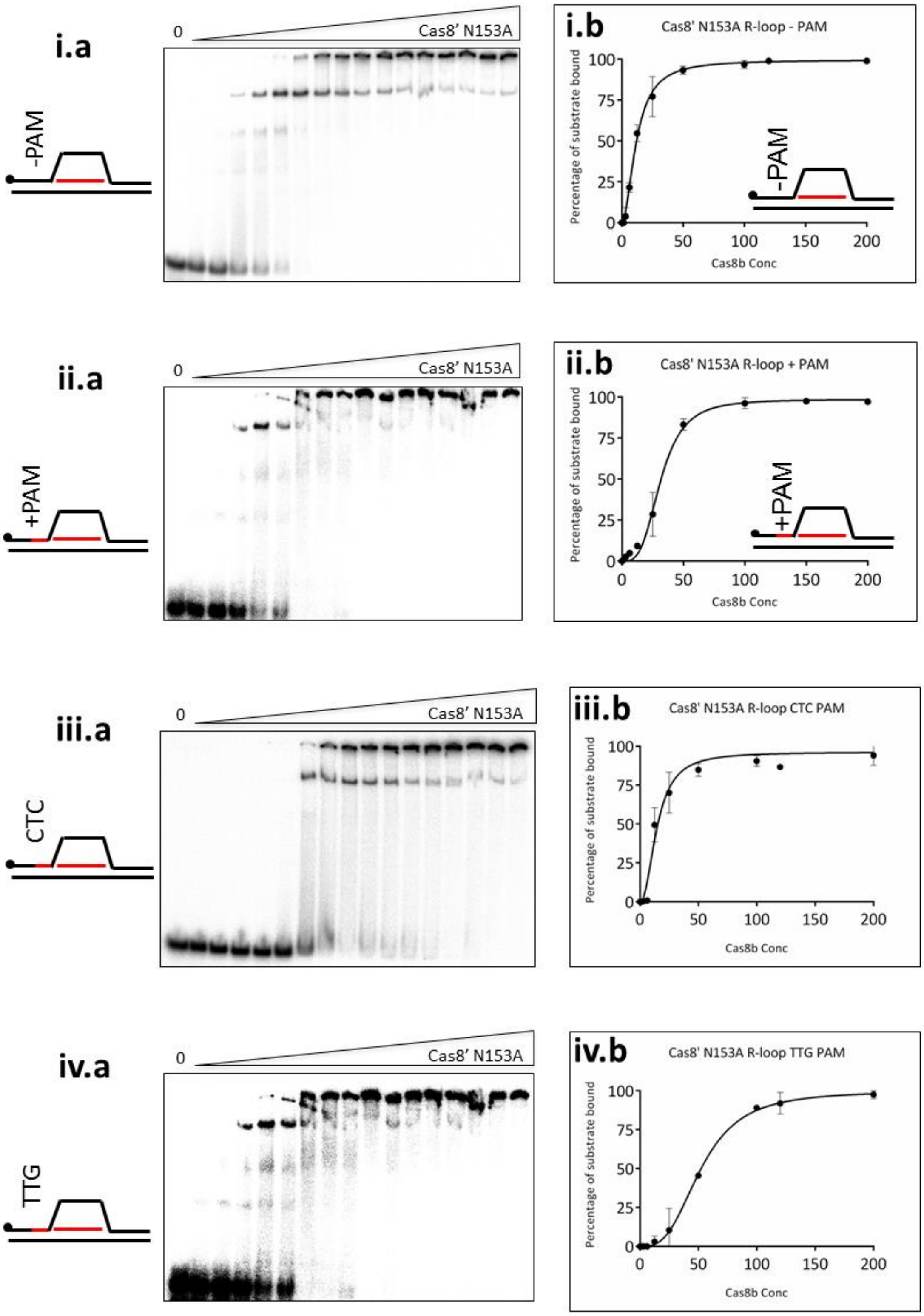


Figure 4-20. EMSA analysis of titrating Cas8’ N153A into R-loop substrates containing various PAM sequences. In all panels part (.a) Cas8’ N153A was used at 0, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2, 6.4, 12.8, 25.6, 51.2, 102.4, 204.8, 409.6, 819.2 and 1638.4nM with 5nM substrate in reactions containing 125mM KOAc and 10mM EDTA. Each phosphorimage of native TBE gel shows separation of unbound, specifically shifted and in well aggregated substrate. Assays were carried out in at least duplicate (usually triplicate), representative gels are shown for: N153A – PAM (i.a), + PAM (ii.a), PAM2 (iii.a) and PAM3 (iv.a). Percentage of substrate shifted was quantified by 2D densitometry (AIDA) to determine K_Ds with standard error (.b). A summary can be seen in

Table 16. Summary of K_Ds for Cas8' wild-type and Cas8' N153A proteins binding to generic DNA and RNA substrates. K_Ds were determined from duplicate EMSAs quantified by 2D densitometry.

Substrate	Cas8'	Cas8'N153A
Duplex	20.4 ± 0.6 nM	81.8 ± 2.8 nM
Flayed Duplex	-	-
R-loop	4.7 ± 0.2 nM	5.3 ± 0.3 nM

Table 17. Summary of K_D values for Cas8' N153A binding to R-loops with various PAM sequences. K_Ds were determined from at least duplicate EMSAs quantified by 2D densitometry.

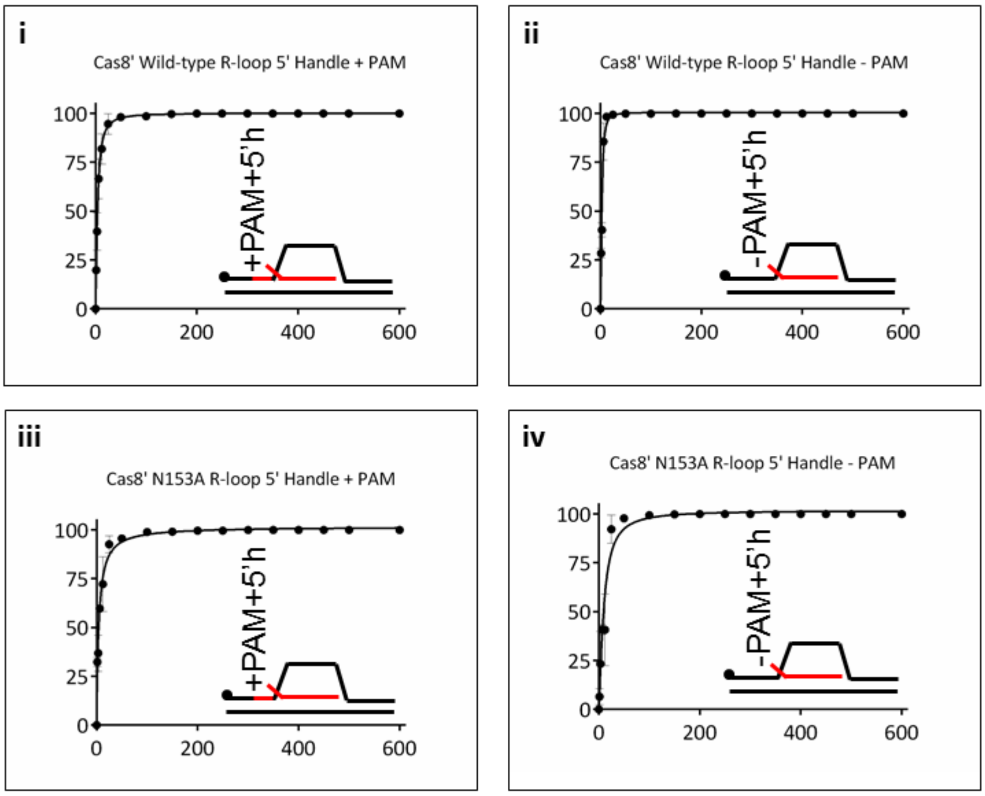
PAM	K_D
AAA (-)	12.0 ± 0.4 nM
CCC (+)	31.6 ± 0.9 nM
CTC	14.8 ± 1.0 nM
TTG	52.3 ± 1.4 nM

4.2.5.6.1 Summary

A compilation of selected mutations in several conserved residues made in both the *Mth* and *Hvo* Cas8 genes allowed the *in vitro* biochemical characterization of *Mth* Cas8' protein and *in vivo* genetic analysis of *Hvo* Cas8. The *Mth* Cas8 mutants all showed decreased binding affinity to all of the oligonucleotide substrates tested (Figure 4-16 and Figure 4-18). Similarly, there were detectable changes in their nuclease and ATPase activity (Figure 4-15 and Figure 4-17). On the other hand, the *in vivo* analysis of *Hvo* Cas8 raised N232A as a PAM insensitive mutant (Table 13). This was corroborated with the analogous *Mth* Cas8 mutant, N153A, which displayed abnormal binding preference to R-loops with different PAM sequences when compared to the wild-type protein (Figure 4-20).

4.2.6 Binding affinity of Cas8' wild type and N153A proteins to 5' Handle R-loop

The R-loops used previously contained perfectly matched RNA that completely complemented one of the DNA strands to form an R-loop. This is not the complete physiological CRISPR/Cas interference Cascade R-loop. The crRNA consists of a 5' handle, complement region and normally a 3’ tail containing a stem-loop structure. So a more relevant structure was tested, this time containing the typical 8 nucleotide 5’ handle sequence from the repeat sequence from *Mth* CRISPR locus. This structure was generated both with and without PAM sequences. When these substrates were assayed by EMSA analysis for Cas8’ wild-type and the Cas8’ mutant N153A binding the PAM sequence proved to have a lesser effect on binding affinity (Figure 4-21). Indeed, wild-type Cas8’ had similar binding to both 5’ Handle R-loop derivatives with corresponding K_D values of 4.1 ± 0.1nM with PAM and 3.1 ± 0.1nM without PAM (Figure 4-21, panels i and ii, Table 17). These apparent dissociation constant values were consistent with the K_D value obtained in previous experiments with classical R-loop with PAM (K_D of 5.3nM, see Figure 4-1). Of note, the R-loop substrates, used here and in previous experiments, had the same PAM sequence though differed structurally with the 5’ Handle R-loop substrate containing an additional 5’ RNA flap region at the PAM branch point. Therefore, this suggested that



in physiological conditions, the correct confirmation/structure of the R-loop substrate, i.e. the presence of a 5' RNA flap, seems to be more important for establishing a high binding affinity between Cas8' protein and the substrate than the PAM sequence. Thus, *in vivo*, PAM sequence most likely fulfils other roles, as explored later.

Likewise, the Cas8' N153A mutant responded to the 5' handle structure inclusive of PAM in a very similar manner to the wild-type protein with a K_D of 4.4 ± 0.3nM versus the wild-type protein's K_D of 4.1 ± 0.1nM (Figure 4-21iii, Table 17). However, the 5' handle substrate lacking PAM showed a decreased binding affinity with a K_D of 9.8 ± 1.0nM (Figure 4-21iv, Table 17). This K_D reflects a similar value to the simple R-loop substrate, 12.0 ± 0.4nM (Table 17). Therefore, with the physiological substrate, the N153A mutation only had a subtle effect on the binding potential of the protein which became apparent solely in the absence of a PAM sequence. While it abolished the ability of the protein to appropriately discern the different PAM sequences in a complete R-loop, obtaining the highest binding affinity in the absence of PAM (Table 18), in the presence of a 5' handle on the substrate, the PAM sequence had a positive effect on the binding affinity which became similar to that of the wild-type protein. On the other hand, the wild-type protein bound the 5' handle substrate with the same affinity irrespective of PAM.

Table 18. Comparative K_D affinities of Cas8' Wild-type protein to N153A point mutant to 5'handle R-loop substrates + and – PAM. K_Ds were determined from at least duplicate EMSAs quantified by 2D densitometry.

Cas8'	**+PAM+5'handle R-loop**	**-PAM+5'handle R-loop**
Wild-type	4.1 ± 0.1 nM	3.1 ± 0.1 nM
N153A	4.4 ± 0.4 nM	9.8 ± 1.0 nM

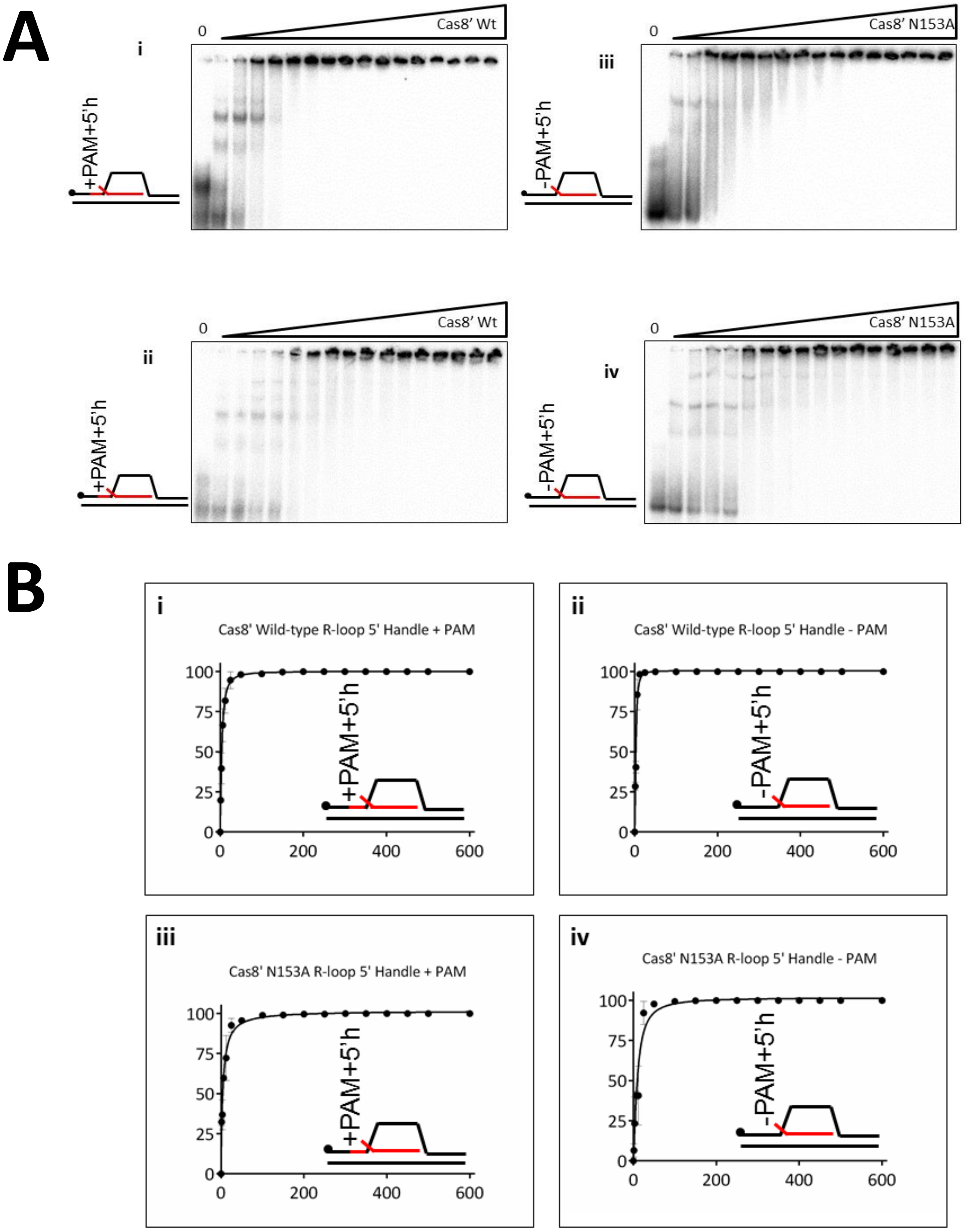


Figure 4-21. EMSA analysis of titrating wild-type Cas8' and the point mutant N153A into R-loop substrates with (CCC) or without (AAA) PAM and a 5'handle. Cas8' proteins were used at 0, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2, 6.4, 12.8, 25.6, 51.2, 102.4, 204.8, 409.6, 819.2 and 1638.4nM with 5nM substrate in reactions containing 125mM KOAc and 10mM EDTA. Each phosphorimage of native TBE gel shows separation of unbound, specifically shifted and in-well aggregated substrate (A). Assays were carried out in at least duplicate (usually triplicate), representative gels are shown for: Wild-type R-loop + PAM + 5'handle (i), Wild-type R-loop - PAM + 5'handle (ii), N153A R-loop + PAM + 5'handle (iii), N153A R-loop - PAM + 5'handle (iv). Percentage of substrate shifted was quantified by 2D densitometry (AIDA) and plotted against Cas8' concentrations to determine standard error and K_Ds (B). A summary of the calculated K_D can be seen in Table 18.

4.2.6.1 Nuclease degradation of crRNA in R-loops

Previous experiments established that the preferentially bound substrate of Cas8' was RNA, either a 3' or 5' flap (see Figure 4-1). Therefore, it was desirable to determine whether Cas8' was cleaving at this site of the substrate. Specifically, the physiologically relevant 5' handle R-loop was tested to see if Cas8' could be expected to act/cleave on the 5' flap of the crRNA *in vivo*. To this end, Cas8' was titrated into this R-loop in acetate salt and analysed as in the previous nuclease assays (Figure 4-17). However, this assay revealed that even with increasing Cas8' concentrations the substrate was not being degraded (lanes 6-9, Figure 4-22). Thus, in the next step the nuclease activity of Cas8' was tested on a modified version of the 5' handle R-loop substrate where the strand of DNA that had been displaced to form the R-loop was removed, therefore mimicking the Cas3-mediated degradation of this strand which occurs *in vivo*. Here, clear increasing degradation was observed in Cas8' concentration-dependent manner (lanes 2-4, Figure 4-22). To ensure that the observed cleavage was specifically due to the activity of the protein, a previously characterized nuclease deficient mutant was used at the highest concentration of wild-type Cas8' as control. Indeed, the Cas8' D151G mutant showed no nuclease activity as would have been expected, (lane 5, Figure 4-22).

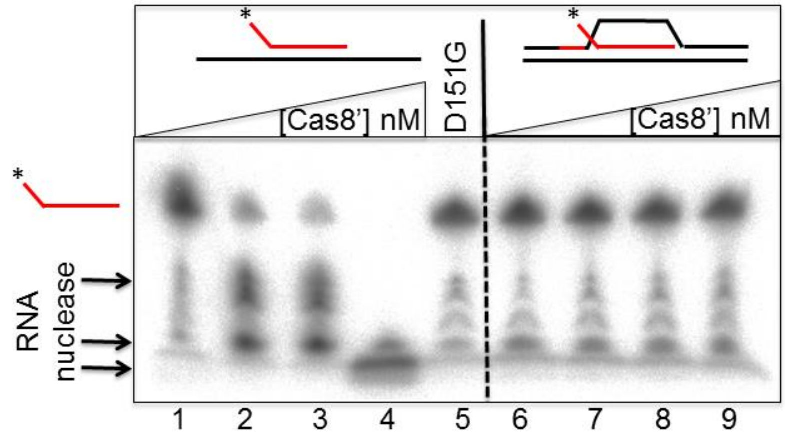


Figure 4-22. Nuclease assay titrating Cas8' into RNA substrates physiologically relevant to CRISPR interference. Wild-type Cas8' was used at 0, 50, 100 and 200nM and Cas8' D151G at 200nM with 5nM substrate in reactions containing 125mM KOAc and 10mM $MgCl_2$. Phosphorimage of urea gel separating Cas8' nuclease products of the 5' handle R-loop substrates. The RNA strand radiolabelled with ^{32}P as denoted by * is in red. Wild-type Cas8' had RNA nuclease activity on 5' RNA handle in a flap substrate (lanes 1-4), whereas the nuclease defective Cas8' D151G did not (lane 5). Wild type Cas8' was inactive as a nuclease on the same RNA strand within an R-loop substrate (lanes 6-9).

4.2.6.1.1 Summary

R-loops with a 5' handle, the physiological substrate for Cas8', stimulate the binding of the wild-type protein (Figure 4-21i). However, the N153A mutant protein is sensitive to the 5' handle substrate, as seen by lower disassociation constants (Figure 4-21iii) and less sensitive to PAM (Table 17). This supported the idea that Cas8' had structural and sequence specificity. This study found that Cas8' could only degrade crRNA from a synthetic R-loop when a strand of DNA has been displaced or degraded (by Cas3) (Figure 4-22), which is discussed later.

4.2.7 Physical association of Cas8' with archaeal Cascade proteins Cas5 and Cas7

Cas5 and Cas7 are integral to bacterial and archaeal Cascade complexes and have been shown to function with Cas8 during CRISPR interference (*128*). In crystal structures of the 'large' subunit (CasA/Cas8) shows interactions with Cas5, Cas7 and the crRNA 5' handle (*162*). Through the collaboration with Anita Marchfelder's research group at the University of Ulm, Germany and work carried out by Britta Stoll, FLAG-tagged Cas7 expressed in *Hvo* was used as bait to identify interacting partners. Therefore, a physical interaction between Cas8’ and Cas5-Cas7 was tested *in vitro*. First, *Mth* Cas7 was co-purified with Cas5, an N-terminal fusion to *E. coli* maltose binding protein (MBP) after being overexpressed recombinantly in E. coli BL21C+ via affinity chromatography (Figure 4-23i). Similarly, *Mth* Cas8’ was also purified with an N-terminal His tag (Figure 3-6). *In vitro* pull down assays, using such purified MBPCas5-Cas7 immobilised on amylose resin as bait with His_6Cas8’ as prey, were then performed to ascertain whether there was a direct interaction. Indeed, immunoblotting of the fraction eluted off this amylose with anti-MBP and anti-His antibodies clearly indicated that MBPCas5-Cas7 interacted physically with His_6Cas8’ (Figure 4-23iiI, top panel, lane 8). This interaction was specific as in the control pull-down where amylose resin was pre-incubated with bovine serum albumin, no Cas8’ was detected in the elution fraction (Figure 4-23iii, top panel, lane 6). Of note, an identical result was obtained either with or without pre-incubation of Cas5-Cas7 with crRNA1 (Figure 4-23iv). However, this pull-down assay indicated only a weak physical interaction of Cas8’ with Cas5-Cas7 as a maximum of approximately 10% of Cas8’ input could be detected as binding to MBPCas5-Cas7 under the conditions used. Moreover, the reciprocal pull-down assay using Ni^{2+}-NTA immobilised Cas8’ as bait with MBPCas5-Cas7 in solution could not be performed as MBPCas5-Cas7 was non-specifically interacting with the Ni^{2+}-NTA resin itself. Next, the ability to directly interact with MBPCas5-Cas7 was also tested for the Cas8’ mutant proteins which have showed loss of function phenotype in previous experiments. The mutants examined included the nuclease inactive Cas8’ proteins (D151G and N536A), as assessed *in vitro*, and the CRISPR interference defective Cas8 protein, (also the PAM insensitive mutant, N153A), initially identified by *in vivo* genetic analysis in *Hvo*. However, all the selected mutant Cas8’ proteins bound to MBPCas5-Cas7 comparably to the wild-type Cas8’ protein (Figure 4-23iv).

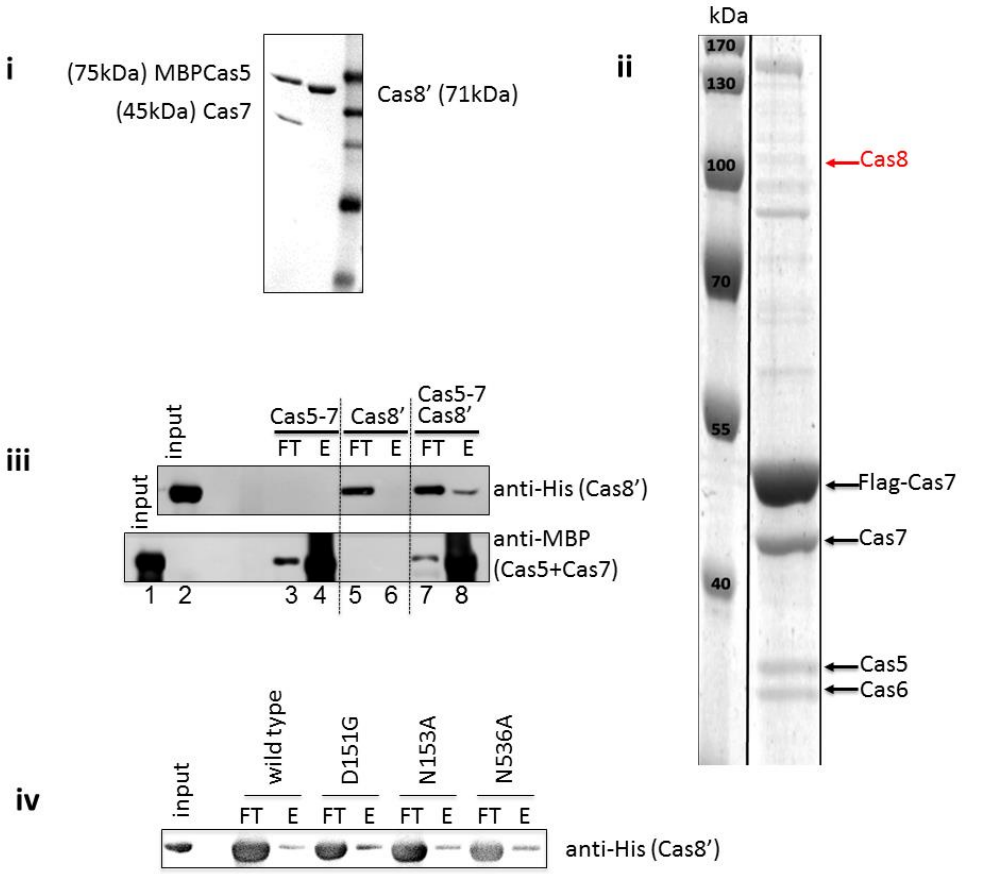
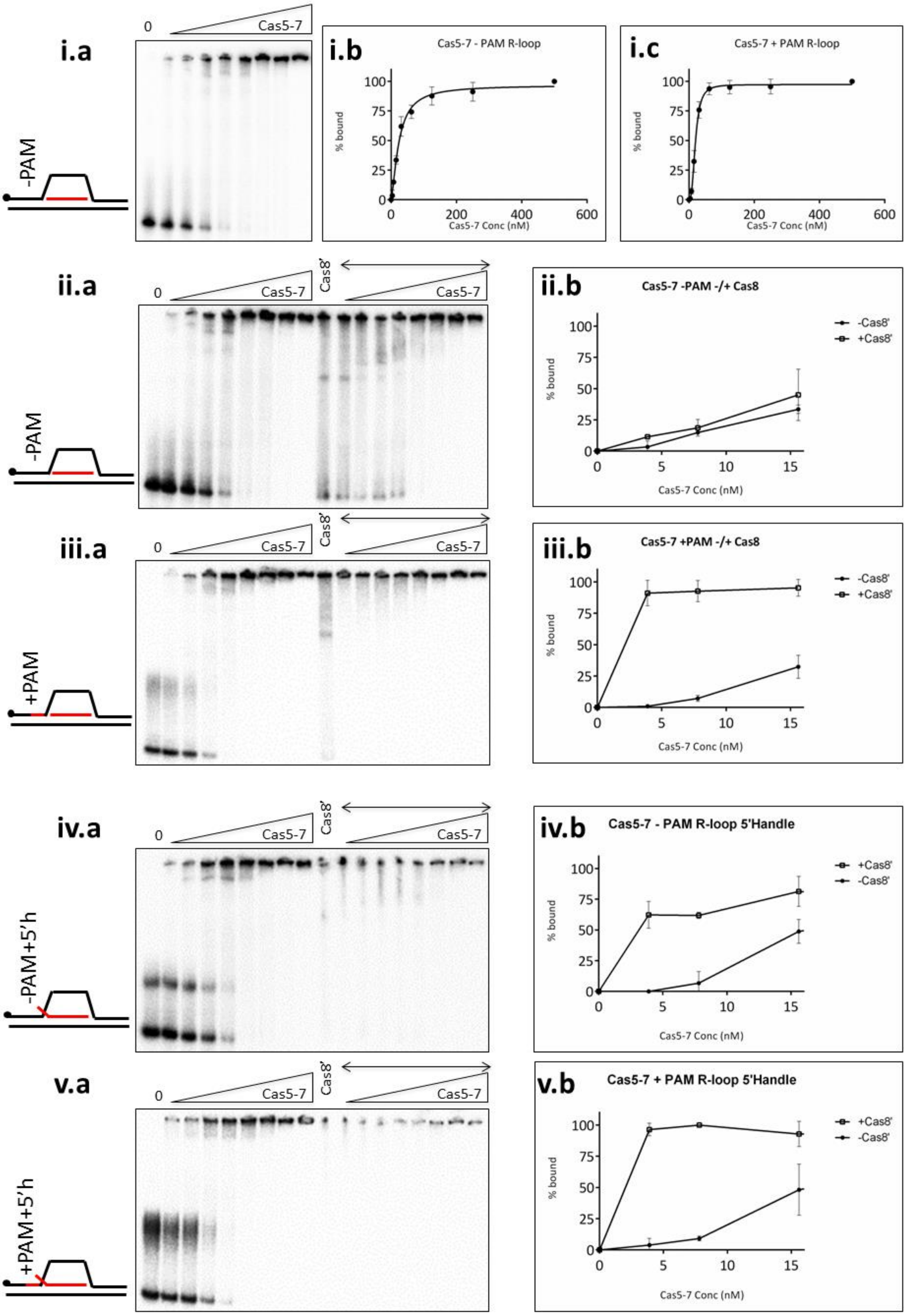


Figure 4-23. Analysis of Cas8 protein interaction with Cas5-Cas7 from *Hvo* and *Mth.* (i) Coomassie stained SDS-PAGE gel of approximately 2µg of recombinant *Mth* MBPCas5-Cas7 and His$_6$Cas8, used in subsequent pull down assays. (ii) Coomassie stained SDS-PAGE profile of proteins co-purifying with Flag-Tagged Cas7 expressed in *Haloferax* cells. Cas8 was detected by mass spectrometry. This work was carried out by Britta Stoll of Anita Marchfelder's research group at the University of Ulm, Germany. (iii) Reconstitution of physical interaction between purified *Mth* Cas8' (20µg) with purified complex of affinity tagged *Methanothermobacter* Cas5-Cas7 (20µg). Upper panel shows Western blot using anti-(His)$_6$ antibody to detect (His)$_6$Cas8', and the lower panel used anti-MBP antibody to detect MBP in MBP-Cas5-7. "Input" is a duplicate loading of total amount of used Cas8' (upper panel) or Cas5-7 (lower panel). Cas8' was detected in the elution (E) after binding to amylose-MBPCas5-7 (lane 8) but was absent from the elution of amylose pre-bound with bovine serum albumin (BSA, lane6). (iv) Repeat of (iii) lanes 7 and 8 but with preincubation of 100ng crRNA1. Additionally, Cas8' mutant proteins D151G, N153A, N536A were also assessed for binding to MBPCas5-Cas7 and where found to interact similarly to wild-type Cas8'.

Following the verification of interaction of Cas8' with Cas5-7, the impact of this interaction on the binding affinity of Cas8' to various R-loop structures with or without PAM was then tested by EMSAs. First, Cas5-Cas7 complex was assessed for its binding characteristics to R-loops with or without PAM. MBPCas5-Cas7 bound the R-loop structure and formed in well aggregates. Approximately all R-loop was bound at 100nM Cas5-Cas7 (Figure 4-24i). This allowed the determination of a baseline binding affinity of Cas5-Cas7 to the R-loop substrates. Next, a set concentration of

Cas8’ was added to the reaction mixtures containing increasing concentrations of Cas5-Cas7 and either standard R-loop substrate with/without PAM or 5’ handle R-loop substrate also with/without PAM. Representative phosphorimages of TBE gels of EMSAs are shown in panels of part A of Figure 4-24 with their corresponding quantified binding isotherms displaying percentage of substrate bound plotted against Cas5-Cas7 concentration in part B. Using standard R-loop substrate, Cas8’ stimulated total substrate binding in the presence of a PAM sequence though had little effect on binding to the R-loop lacking PAM (Figure 4-24ii and iii). On the other hand, Cas8’ enhanced the binding affinity of Cas5-Cas7 to 5’handle R-loops irrespective of the PAM sequence, although a more significant difference in binding was observed with the 5’ handle R-loop substrate containing PAM (Figure 4-24iv and v). These same assays, though with standard R-loop substrate only, were then also repeated with the Cas8’ point mutant N153A, which has been shown to have an altered PAM sensitivity (see Table 17). Similarly to the results obtained with the wild-type Cas8’, the Cas8’ N153A mutant had limited effect on MBPCas5-Cas7 binding to R-loop substrate without PAM (Figure 4-24vi). However, the binding to the PAM R-loop substrate was enhanced even in the presence of this PAM insensitive mutant which was comparable to the increase seen with the wild-type protein, though only at higher concentrations of Cas5-7 (comparing Figure 4-24v [wild-type] to Figure 4-24vii [N153A]).

Next, the binding characteristics of the Cas8’ and Cas5-Cas7 complexes with the R-loop was analysed to identify any difference in in gel species observed. Significantly, when identical EMSAs where Cas5-Cas7 and Cas8’ were mixed, analysis of these EMSAs was carried out either using the normal radiographic development method or via Western blot following transfer onto PVDV membrane to identify MBPCas5-Cas7 with anti MBP antibody. This analysis revealed Cas5-Cas7 at both, 10nM and 100nM, formed in-well protein-DNA aggregates with 8% and 67% of substrate, respectively (labelled A, lanes 1 and 2 Figure 4-25). Similarly, Cas8’ alone (5nM) bound 55% of the substrate in distinct complexes (labelled B in lane 3 Figure 4-25). However, a new complex (complex C) arose when Cas8’ (5nM) was pre-mixed with Cas5-7 (10 or 100nM), binding 90-100% of the substrate (lanes 4 and 5, Figure 4-25). This confirmed that Cas5-Cas7 formed a distinct complex in EMSAs that was not an aggregate and was dependent on Cas8’.


i.a
0
Cas5-7
-PAM
i.b
Cas5-7 - PAM R-loop
% bound
Cas5-7 Conc (nM)
i.c
Cas5-7 + PAM R-loop
% bound
Cas5-7 Conc (nM)
ii.a
0
Cas5-7
Cas8'
Cas5-7
-PAM
ii.b
Cas5-7 -PAM -/+ Cas8
-Cas8'
+Cas8'
% bound
Cas5-7 Conc (nM)
iii.a
0
Cas5-7
Cas8'
Cas5-7
+PAM
iii.b
Cas5-7 +PAM -/+ Cas8
-Cas8'
+Cas8'
% bound
Cas5-7 Conc (nM)
iv.a
0
Cas5-7
Cas8'
Cas5-7
-PAM+5'h
iv.b
Cas5-7 - PAM R-loop 5'Handle
+Cas8'
-Cas8'
% bound
Cas5-7 Conc (nM)
v.a
0
Cas5-7
Cas8'
Cas5-7
+PAM+5'h
v.b
Cas5-7 + PAM R-loop 5'Handle
+Cas8'
-Cas8'
% bound
Cas5-7 Conc (nM)

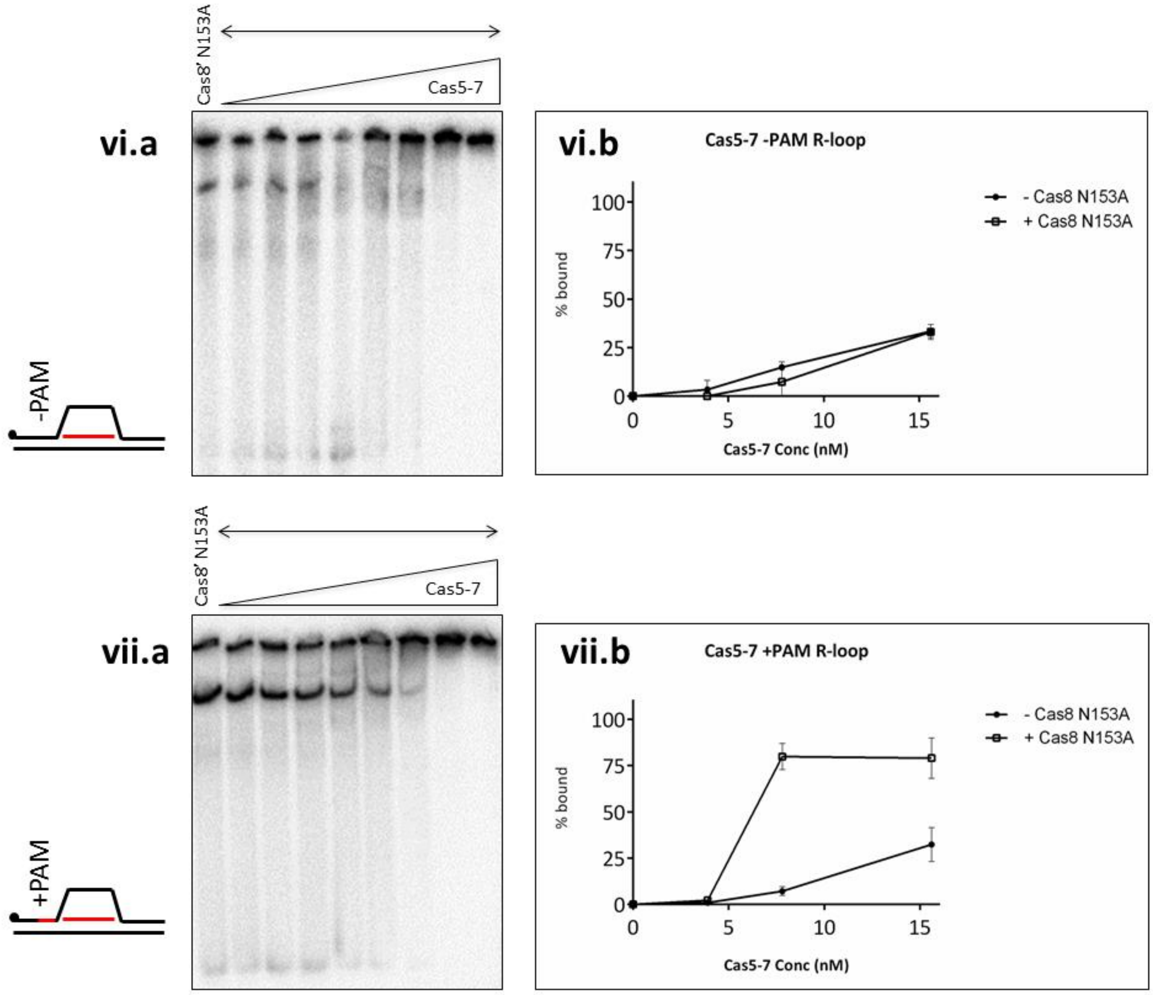


Figure 4-24. EMSA analysis of titrating MBPCas5-Cas7 into wild-type Cas8' and the Cas8' point mutant N153A into various R-loop substrates. MBPCas5-Cas7 was used at 0, 3.9 7.8, 15.6, 31.25, 62.5, 125, 250 and 500nM with 5nM substrate in reactions containing 125mM KOAc and 10mM EDTA. (i.a) Representative phosphorimage of native TBE gel of MBPCas5-Cas7 binding to R-loops forming only in well aggregates, quantified by 2D densitometry for – (i.b) and + PAM (i.c), to determine binding isotherms. MBPCas5-Cas7 titrations were then premixed with or without Cas8' (2.5nM) and equivalent titrations were carried out – PAM (ii), + PAM R-loop (iii), -PAM +5'handle (iv) and +PAM +5'handle R-loop (v Relative increases in binding were compared – and + Cas8', seen in (.b) graphs with standard error.). MBPCas5-Cas7 titrations were then also carried out premixed with or without Cas8' (2.5nM) on the same – PAM (vi) and + PAM (vii) R-loop substrates.

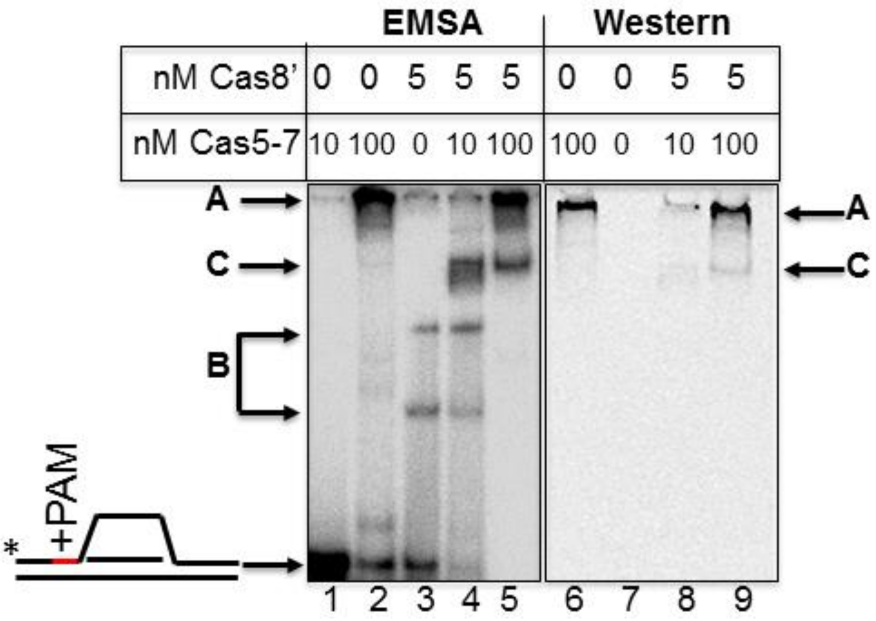

EMSA
Western
nM Cas8' 0 0 5 5 5 0 0 5 5
nM Cas5-7 10 100 0 10 100 100 0 10 100
A
C
B
A
C
*
+PAM
1 2 3 4 5 6 7 8 9

Figure 4-25. EMSA and Western blots for detection of Cas5-7 in a Cas8' dependent in-gel complex. Lanes 1-5 (left panel) show phosphorimaged EMSA complexes arising from reactions binding of Cas5-7 (complex A) or Cas8' (B complexes). A new complex C was observed when Cas5-7 and Cas8' are present. Western blotting detected MBPCas5-7 (anti-MBP) in complex C (lanes 8 and 9), as well as complex A (lanes 6, 8 and 9).

4.3 Discussion

Cas8 proteins and other 'large' subunit proteins of Cascade complexes have been implicated as the guide module of Type I interference complexes (*162, 163, 226, 227, 264*). Indeed, the *E. coli* 'large' subunit CasA identifies correct PAM sequence and activates Cas3-mediated degradation of the invading nucleic acid. Therefore, in this study, the role of Cas8 in *Mth* CRISPR interference was studied in greater detail in parallel to a similar study in *Hvo* being carried out by a collaborating research group led by Dr. Anita Marchfelder at the University Of Ulm, Germany. This parallel study generated different point mutations in conserved residues of Cas8 and using *in vivo* genetic analysis identified which of these resulted in loss of CRISPR interference in *Haloferax volcanii (Hvo)* (see Table 15). Similarly, conserved residue mutational analysis of *Mth* Cas8 was performed in the scope of this study whereby nucleic acid binding and processing by Cas8' was examined. *Mth* Cas8' has binding specificity for branched nucleic acid substrates, specifically those containing a single RNA strand.

This is in line with the predicted role in CRISPR interference and the various DNA and DNA/RNA structures that would be expected to form in an effective immune response (see Figure 4-1). Cas8’ was also shown to be sensitive to PAM sequence flanking an R-loop, important for a complete interference reaction (see Figure 4-3, Figure 4-4). However, the binding affinity of Cas8’ was less sensitive to PAM sequence than the 5’ handle R-loop substrate, (Figure 4-19). An interesting ssRNA nuclease activity was detected *in vitro* (Figure 4-5), which was disrupted following mutation of K117, D151, N536 and L542 residues (Figure 4-17).

When tested on generic substrates, Cas8’ could bind all DNA and RNA structures. However, the majority of these substrates resulted in in-well protein-DNA aggregation (Figure 4-1). The native intracellular conditions (acetate salts) of *Mth* may cause *in vitro* proteins to become non-specifically adhesive. Only when generic D- and R-loops were used as substrates for binding in EMSAs did specific species become apparent (Figure 4-1vi and vii). This indicates that Cas8’ has a binding specificity dependent on the substrate structure, such that these branched DNA structures are preferentially bound, in concurrence with previously published work (*269*). Further analysis revealed that the bioinformatically deduced PAM 5’-CCN-3’ (Figure 4-2 and Table 11) could stimulate binding to both duplex and R-loop substrates (Figure 4-3). Binding assays revealed specific nucleic acid binding and PAM sensing by Cas8’, both in isolation (Figure 4-3) and when mixed with Cas5-Cas7 (Figure 4-24). In each Case apparent binding affinity increased in substrates with PAM over those without (Figure 4-3 and Figure 4-4). Following mutation of a conserved asparagine: N153 in *Mth* and N232 in *Hvo*, subtle PAM-induced behaviours were discovered from both *in vivo* genetic interference reactions (Table 15) and EMSAs (Figure 4-20). Furthermore, these EMSAs demonstrated a perturbed interaction of this Cas8 mutant with PAM containing substrates either in isolation or when in complex with Cas5-Cas7 in a reconstituted archaeal Type I Cascade (Figure 4-24vi and vii). Indeed, binding to R-loop substrates inclusive of PAM was enhanced in the presence of the Cas8’-Cas5-Cas7 complex when compared to either Cas5-Cas7 or Cas8’ alone (Figure 4-24i, iii and v). However, such increase in binding affinity was not seen with R-loop substrates without PAM, total amounts of substrate bound remained similar (Figure 4-24i, ii and iv). Interestingly, in these assays Cas8’ appeared responsible for the conversion of Cas5-Cas7 protein aggregates into distinct binding complexes, as shown by the specific species C identified by western blot analysis (Figure 4-25). This suggests that Cas8’ coordinates the assembly of Cas5-Cas7 precisely on the substrate, which would explain the improved specific binding seen when the Cas8’-Cas5-Cas7 complex is assessed in binding assays. Analogously to the positioning of *E.coli* CasA relative to CasD (Cas5e), within the *E.coli* Cascade structure which bind at the branch point of the R-loop and detect PAM sequence for the specific stabilisation of the R-loop and recruitment of Cas3. It is therefore likely that the interaction between Cas8’ and Cas5-Cas7 is important for PAM sensing, activating precise Cascade nucleic acid binding, stable R-loop formation and target degradation (*162, 176, 280, 283*). It is now thought that the Cas8’ N153 residue may act in a similar way to that of the asparagine residue that is located in L1 loop of the CasA protein from *E.coli (227)*.

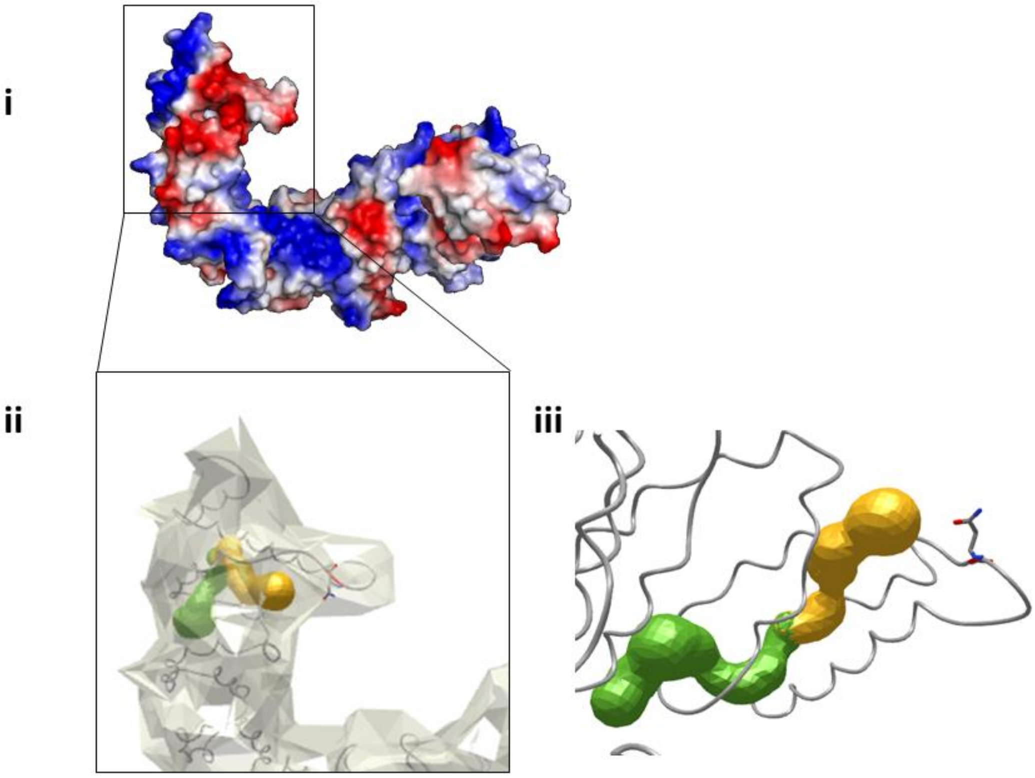


Figure 4-26. Structural predictions of *Mth* Cas8' from amino acid sequences. Electrostatic surface charge of CPHmodel structure of Cas8' (<0.001% confidence) (i). The generated PDB file of this model was used to identify possible DNA binding channels in the Cas8' protein using MOLE. An example of two linked channels found in this Cas8' structure which are located close to key conserved residues (ii). For example, *Mth* N153 residue is shown as a stick amino acid as it is relevant to establishing substrate PAM sensitivity (iii).

As at the time of writing this thesis, no *Mth* Cas8 structures have been yet published, different homology modelling servers were employed to obtain a plausible structural model. However, only CPHmodel successfully generated at least a partial predicted structure (Figure 4-26i). Next, MOLE software was used to predict possible binding tunnels or surface ligand binding regions. The most interesting hit from this analysis, shown in Figure 4-26ii, identified a tunnel through which ssDNA/RNA could potentially traverse the Cas8 protein. Interestingly, The N153 PAM sensitive residue, highlighted as a stick amino acid in Figure 4-26iii, is located at the entrance of this tunnel, and may therefore affect the binding of Cas8' to R-loop structures. Indeed, mutation of the analogous conserved residue, N131A, in *E. coli* perturbed the interaction of Cascade with Cas3 and disrupted the degradation of the target *(227)*.

The *in vivo* genetic analysis of Cas8 mutations in *Hvo,* carried out by the collaborating research group in Germany, identified residues Asp 230 and Asn 625 as important to CRISPR interference given their mutation abolished interference activity (Table 15). In agreement with this observation, the corresponding *Mth* Cas8' residues, Asp 151 and Asn 536, abolished the previously detected ssRNA nuclease activity (Figure 4-17I, iv and vii). Loss of nuclease function was apparent also with the *Mth* K117A mutant, however this residue is not conserved in *Hvo* which precluded a comparison between the two organisms. Importantly, Cas8' degraded both 3' and 5' ends of ssRNA, exhibiting endonuclease activity because of the generation of discrete cleavage patterns rather than a ladder effect of the labelled strand being cleaved at specific intervals (Figure 4-10). It was not possible to ascertain RNase activity for the *Hvo* Cas8 in *in vitro* reactions. A wide range of conditions were tested here, including high and low salt, alternate metal ions and anions. The absence of activity is possibly because of the instability of the purified *Hvo* protein on storage, delivered as a gift from Dr. Anita Marchfelder's research group. It cannot therefore be concluded that this loss of RNase activity in *Mth* Cas8' D151G and N536A can be directly linked to *Hvo* Cas8 D230A and N625A loss of genetic interference. However, this correlation is at least intriguing. It has been shown previously that deletion of Cas8 from *Hvo* did not affect the maturation of pre-crRNA into mature crRNA, as assessed by northern blotting for crRNA (*129, 280*). Indeed, following correspondence with Dr. Anita Marchfelder at the University of Ulm, Germany and Dr. Thorsten Allers at the University of Nottingham about other predicted nucleases in *Hvo*, potential candidates that could further mature the 3' terminus of the pre-crRNA were recognized . This maturation would occur after Cas6- mediated cleavage of the initial full length transcript, independently of Cas8 in *Hvo*. Moreover, given that 5' handles are essential for interference it seems highly unlikely that these structures are the target of the Cas8' cleavage activity during crRNA maturation. Therefore, the functional significance of Cas8 RNase activity in the context of CRISPR immunity remains elusive, though two possible roles are proposed here. Firstly, there is the potential of a recycling role. Cas8' failed to cleave a 5' handled RNA when in an R-loop (Figure 4-22ii), though following the removal of the displaced strand of DNA forming the R-loop, RNA cleavage proceeded as expected (Figure 4-22i). Therefore, following successful recruitment of Cas3 and transfer of the correct information necessary to target invading nucleic acids, this RNA degradation may initiate the clearing of the Cascade complex. Secondly, this RNase activity of Cas8 may be required in some other aspect of cellular RNA metabolism in these organisms that is indirectly important for CRISPR Type I systems, or a cellular role. This was briefly investigated here by DNA damage assays. *Hvo* strains, a wild-type and ΔCas8 were subjected to DNA damage from UV irradiation and DNA chemical cross-linking by Mitomycin C. Both of these however resulted in no alteration to DNA damage sensitivity (data not shown). There was also to possibility that the predicted 'resolvase' activity reported previously could have had a cellular benefit (*269*) as some DNA repair pathways include RNA molecules that act as primers.

5 Cas1 and Cas2 facilitation of primed and naïve adaptation in CRISPR-Cas immunity

5.1 Summary

In this chapter, the ongoing work in characterisation of the biochemical activities of Cas1 and Cas2 proteins in CRISPR adaptation was described together with the collaborative effort to identify and analyse host factors that may contribute to this mechanism. At the time of this study, no mechanistic detail was known about CRISPR adaptation. However, the fact that CRISPR-Cas immune system of prokaryotes is built on the capture and integration of invading genetic elements into CRISPR loci by Cas1 and Cas2 proteins had been widely accepted (*67, 126, 148, 173, 184, 267*). Adaptation can be stimulated by inefficient interference ('primed'), or can act independently ('naïve'). 'Primed' adaptation re-establishes immunity against the invader that had previously evaded interference through mutation in PAM or protospacer sequences. Primed adaptation requires Cas*1*, Cas*2*, *Cas3*, *Cascade* and a spacer that is complementary or closely matches the protospacer sequence, as has been shown genetically (*268*). On the other hand, naïve adaptation is catalysed by Cas1 and Cas2, independently of Cascade both *in vivo* and *in vitro (118, 148, 184).*

E.coli adaptation requires the functional complex of catalytically active Cas1, and Cas2 ([Cas1]$_4$-[Cas2]$_2$) (*148*). This complex generates a new spacer-repeat pair at the zero position of the chromosomal CRISPR locus. It is proposed that a series of targeted transesterification reactions inserts a protospacer at cruciform structures formed by CRISPR repeats (*150, 184*). Though the mechanism of protospacer capture remains elusive, the major source of new spacer DNA in naïve adaptation is the *ter* and *Chi* sites, likely at stalled replication forks (*67*). This discovery also identified the role of RecBCD complex in naïve adaptation as a mechanism of targeting non-self DNA. RecBCD degrades linear DNA until reaching a *Chi* site. Prokaryotic genomes

have a high density of *Chi* sites which restricts the amount of free DNA that can be used in spacer acquisition. As most plasmids and viral DNA enter cells in a linear unprotected form, immediate degradation of the DNA upon cell entry is possible. Moreover, plasmid and viral DNA will be extensively replicated therefore increasing to likelihood of stalled replication forks. Stalled replication forks arise when the progression of the replication machinery is halted. This can be from a DNA lesion/ mutation or DNA bound protein, such as RNA polymerase. RecBCD is required for the remodelling of replication forks and along with RecG and PriA restart replication once the blockage has been removed or resolved.

5.1.1 Aims

To understand CRISPR adaptation the catalytic specificity of target capture and protospacer integration involving Cas1 and Cas2 needed to be determined. Understanding Cas1 nuclease ‘nicking’ and/or transesterification activity and its importance for processing blocked or broken replication forks was the main aim of this section of research. Further, the requirements of Cas2 and other host factor effects on Cas1 were also studied for these reactions. The resulting findings were then to be used for the proposal of a model of naïve and primed adaptation.

5.2 Results

5.2.1 Binding specificity of Cas1 and Cas1-Cas2

Previously, Cas1 has been shown to directly interact with forked substrates and Holliday junctions (*118, 148*). However, these results were obtained with 20-500 fold excess of protein over DNA substrates (ranging from 100nM - 2µM). Therefore, as part of this study, purified His$_6$Cas1 protein from *E.coli* was tested for DNA fork and Holliday junction binding at more physiological concentrations (0-25nM). Cas1 monomer, used at 0.1-25nM, bound DNA forks containing 25 nucleotide (nt) of ssDNA (‘Fork-1 or Fork-2’) forming a stable complex (X) (lanes 1-12, Figure 5-1i). However, less than 5% of the equivalent fork substrates, either fully base-paired or so-called Holliday junctions (Fork-3), were bound (lanes 13-24, Figure 5-1i). Fork-1 and Fork-2 have opposite polarities, Fork-1 with a 5’ ssDNA flap and Fork-2 with a 3’ ssDNA region. Next, the binding of His_6Cas2 alone and in conjunction with Cas1 was tested. Whereas Cas2 alone did not bind fork substrates, pre-incubation of Cas1 with Cas2 before addition to forked DNA gave a super-shifted complex in EMSAs (Complex Y) (lane 3, Figure 5-1ii). As a control for the specificity of the binding complexes observed, the putative DNA binding mutant Cas1 R84A was also included in the study (*148*). Indeed, this mutant could not bind to the Fork-1 substrate nor form super-shifted complex Y in the presence of Cas2 (lane 6, Figure 5-1ii). Complex Y formation represents a stable binding of Cas1 and Cas2 with the DNA, supporting the need for a Cas1-Cas2 complex for CRISPR adaptation.

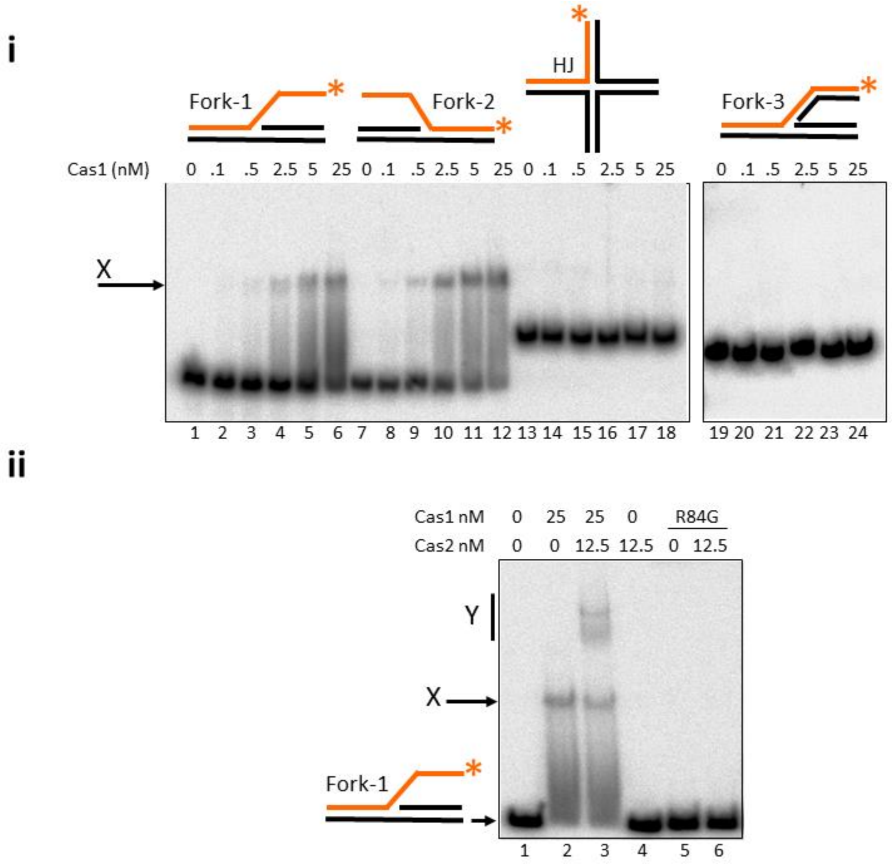


Figure 5-1. DNA binding by *E. coli* Cas1 and Cas2. (i). EMSAs of Cas1 binding to DNA substrates end labelled with ^{32}P (*). Cas1 monomer protein concentrations are indicated above the panels, in reactions containing 6nM of DNA. 'X' indicates stable Cas1-DNA complex observed with forks containing ssDNA regions. (ii). EMSAs of Cas1, Cas2 and Cas1 pre-mixed with Cas2, binding to fork-1, at protein monomer concentrations indicated above the panel, in reactions containing 6nM of end labelled DNA (*). 'X' marks defined Cas1-DNA complex, and 'Y' indicates a second complex requiring the presence of Cas2 for Cas1 fork binding. Lanes 5 and 6 show that when R84G Cas1 mutant protein (25nM), unable to bind to fork DNA was pre-incubated with Cas2 complex Y is not formed. R84G Cas1 mutant protein (25nM) was pre-incubated with Cas2.

5.2.2 Nuclease and transesterification activity of Cas1

Reported catalytic activities of Cas1 were investigated on DNA forked substrates *(207)*. Cas1 proteins were assessed with both N-terminal hexahistidine and C-terminal hexahistidine-StrepTactin tags, in Case the process of tagging the proteins at one or other terminus affected the biochemical properties of Cas1. Cas1 nuclease (nicking) and transesterase activity was detected via urea denaturing gels to show nuclease degradation (Figure 5-2I, left panel) or transesterification elongation of the labelled DNA strands (Figure 5-2i, right panel). The location of the histidine tag on Cas1 was found to have critical impact on the activity of protein. Specifically, the catalytic activity of His$_6$Cas1, in terms of it functioning both as a nuclease and a transesterase, on the Fork-1 substrate was at least 10 fold higher than Cas1-His-Strep (Figure 5-2I, left and right panels). Therefore, all further nuclease assays with the different DNA fork substrate variations were performed with His$_6$Cas1, hereafter referred to simply as Cas1. Cas1 nicking was observed on Fork-1 strand containing a 25nt ssDNA region (Figure 5-2i and iii) and also on Fork-1a, a derivative of Fork-1 with only a 4nt ssDNA region at the branch point (lane 1, Figure 5-2iv). Of note, Cas1 nicking activity on fork-1 and fork-1a yielded different nuclease products of 18nt and 26nt, respectively, from the ssDNA region of each labelled strand (lanes 1 and 6, Figure 5-2iv). In contrast, Cas1 showed no activity on Fork-2 (Figure 5-2iii, right panel). Though the ssDNA region of the Fork-2 substrate was identical sequence-wise to Fork-1, it had opposite polarity which suggested substrate polarity was an essential criterion for substrate cleavage by Cas1. Similarly, no nuclease degradation of Holliday junction substrates was observed. However, there was a possibility cleavage simply occurred on a different strand. Therefore, each strand was labelled in turn and reassessed for Cas1 nuclease activity with each. Regardless of which strand labelled simply no nicking activity was observed for this kind of substrate, fully base paired fork structure or Holliday junction strands (Figure 5-2v).

Next, as Cas1 and Cas2 are known to form a complex in *E.coli* and were indeed found to form a stable complex with DNA substrates *in vitro* (Figure 5-1), Cas1 nicking nuclease activity was also tested in the presence of Cas2. First, Cas2 was assayed on its own with either Fork-1 or Fork-2 but showed no nuclease activity which was consistent with its lack of binding (lanes 4 and 8, Figure 5-2iii and Figure 5-1). Next, Cas1 was pre-incubated with Cas2, yet no stimulation of nuclease activity was detected with either of the same substrates (lanes 3 and 7, Figure 5-2iii).

Transesterification was also tested on fork-1, 1a and 2 substrates and other replication forks with varied leading or lagging strand gaps (2 and 4nts) in time-course assays (Figure 5-3). Urea gels were then used to resolve the extension of the labelled DNA strand to a 50mer. Cas1 transesterified fork-1 and fork-1a (Figure 5-3i and iv and Figure 5-2ii, right panel). All other assessed substrates did not generate any extension products. As a control for the reaction, catalytically inactive Cas1 D218A mutant was also included in the assay and, as expected, yielded no products on fork-1 (Figure 5-3 viii). Cas1 nuclease and transesterification activity within the ssDNA region of fork-1 and of replication fork with a 4nt lagging strand gap (Fork-1a) showed that a free 3' –OH end was required for the transesterification reaction to occur and no DNA end was required to load Cas1 for the equivalent nuclease activity.

The nuclease and transesterification activity observed for Cas1 was consistent with the known *in vivo* role of Cas1 and Cas2 in spacer acquisition. Indeed, an *in vitro* assay of spacer acquisition (SpIN) showed that in the presence of Cas1, a short piece of labelled duplex DNA was integrated into a plasmid (pJRW2) which contained the chromosomal CRISPR leader and an entire locus from *E. coli* (Figure 5-4). Moreover, upon introduction of Cas2, Cas1 was stimulated such that this *in vitro* spacer acquisition became more efficient as assessed over the time-course shown by comparing integrated plasmid product at the top of the left panel to the right panel (Figure 5-4).

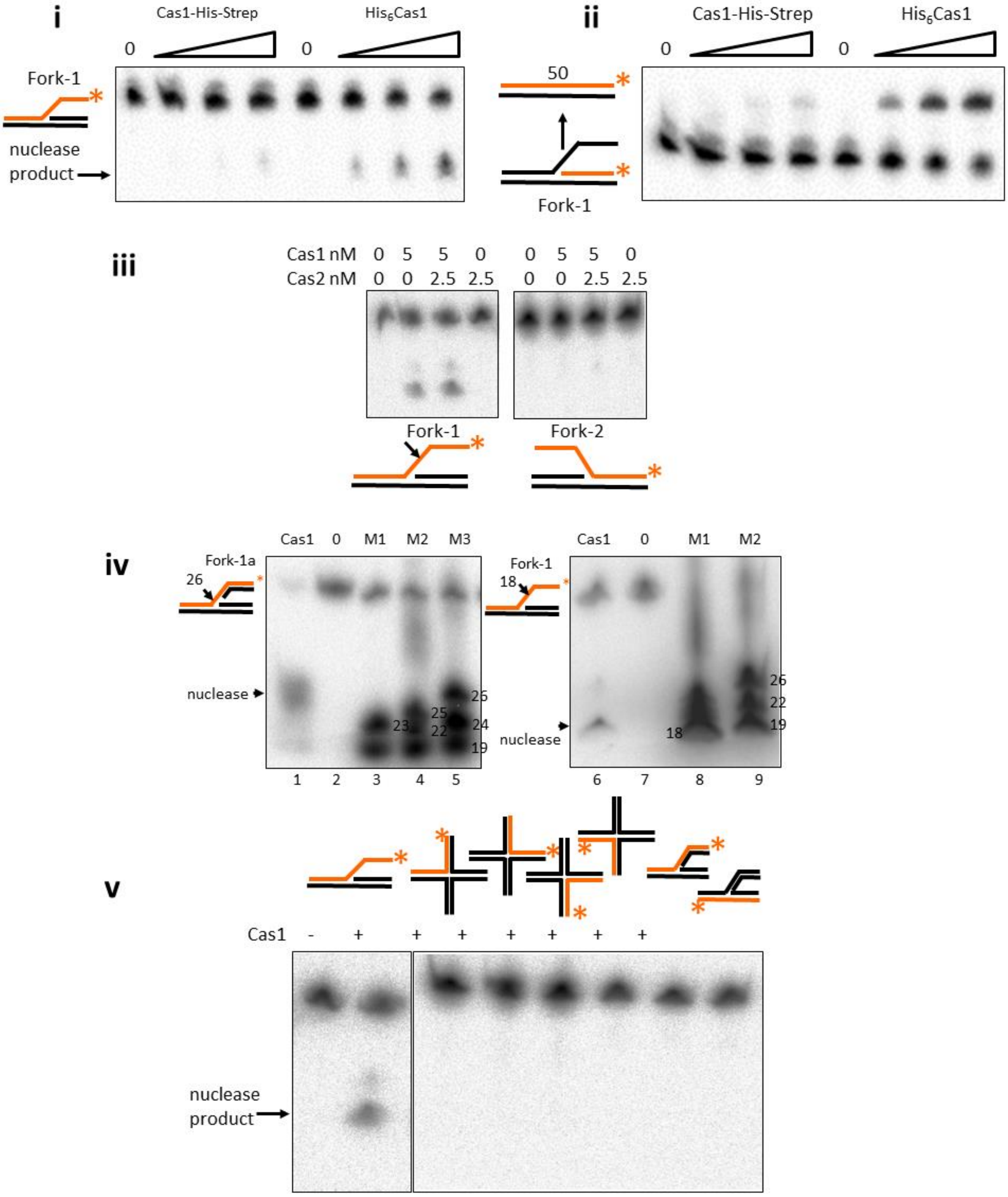


Figure 5-2. DNA nuclease and transesterification activity of Cas1 and Cas2. (i) Products from fork DNA (6 nM) cleavage by Cas1 proteins (0, 2.5, 12.5 and 25nM) shown in phosphorimages of denaturing urea gels. Cas1 activity on the labelled strand (*) yielded 18nt length products from fork-1. (ii) Products from transesterification of the leading strand generating a 50nt product, extending the labelled strand. (iii) Products from Cas1 cutting fork-1 and fork-2 after pre-incubation with Cas2. (iv) Mapping of the nuclease products from Fork-1 and -1a. (v) Nuclease products of Cas1 fork nicking activity not detected of Holliday junction of fully base paired forks.

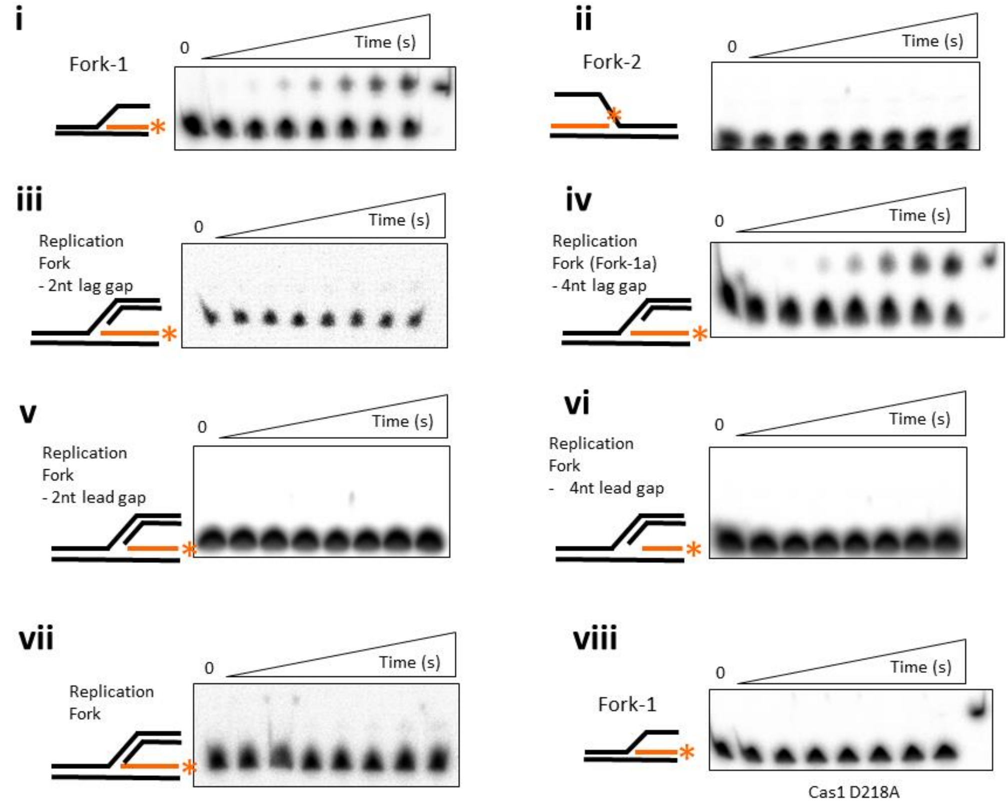


Figure 5-3. Time course analysis of Cas1 transesterification on replication fork-like substrates. Each panel shows phosphorimages of denaturing urea gels of reactions containing 25nM Cas1 with 6nM DNA substrate in 125mM NaCl and 10mM $MgCl_2$. A master reaction was set up in each Case from which samples were removed and terminated at time points 0, 30, 60, 120, 300, 600. 900 and 1800 seconds. In each Case, the leading strand is labelled (*) and transesterification products are seen in i and iv; all other panels including that of the catalytically inactive Cas1 mutant D218A (viii) show no transesterification.

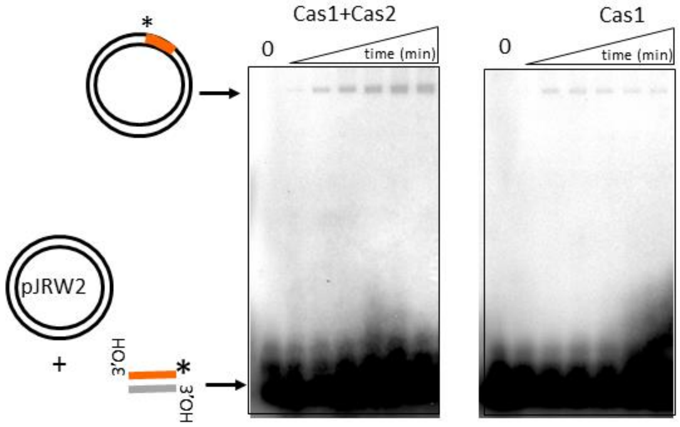


Figure 5-4. Spacer integration (SpIN) assay of short duplex DNA into plasmid. Phosphorimages of agarose gels to detect integration of an end labelled (*) 35 base pair duplex DNA into plasmid pJRW2 catalysed by Cas1 (100 nM) or Cas1 and Cas2 (100 and 50 nM respectively) incubated for 5, 10, 15, 20, 25 and 30 minutes.

5.2.3 RecG and PriA genetic interactions and attenuation of Cas1 activities.

The experimental work in this section were generated as a result of genetic analysis carried out by Ivana Ivancic-Bace at the University of Zagreb, Croatia, who investigated adaption in *E.coli* strains deleted for genome stability and maintenance encoding genes to identify which of these were possible factors contributing to the adaptation process. Here is a short summary of the data generated and presented in a collaborative publication.

These genetic tests identified the differential requirements for both primed and naïve adaptation mechanisms. Primed adaptation was tested with chromosomally inducible Cas1, Cas2, Cas3, Cascade and a CRISPR spacer (spT3) that targets the essential R gene of virulent lambda phage. Naïve adaptation only expressed Cas1 and Cas2 proteins from inducible plamids. Most deletions of associated DNA repair and metabolism genes had no observable effect on CRISPR array expansion (spacer acquisition). However, this could be restored with plasmid complementation. Interestingly, though heliCase inactive PriA K230R restored the normal CRISPR expansion phenotype, the heliCase inactive RecG Q640R could not. RecG rescues stalled replication forks and dissociates R-loops, promoting genome stability in most bacteria. PriA restarts arrested DNA replication forks by both heliCase-dependent and independent pathways that interact with RecG. RecG and PriA act as antagonists to balance replication and prevent pathological excessive replication. DNA polymerase I is a gap-filling DNA repair enzyme. Δ*polA* abolished expansion but ΔrecG was able to acquire spacers normally. Δ*priA* could not be tested as PriA was required for plasmid propagation. Additionally, naïve adaptation was found to be sensitive to deletion of RecB (Δ*recB*). RecB is part of the RecBCD complex that initiates homologous recombination at DNA breaks, creating a substrate for RecA to generate D-loops.

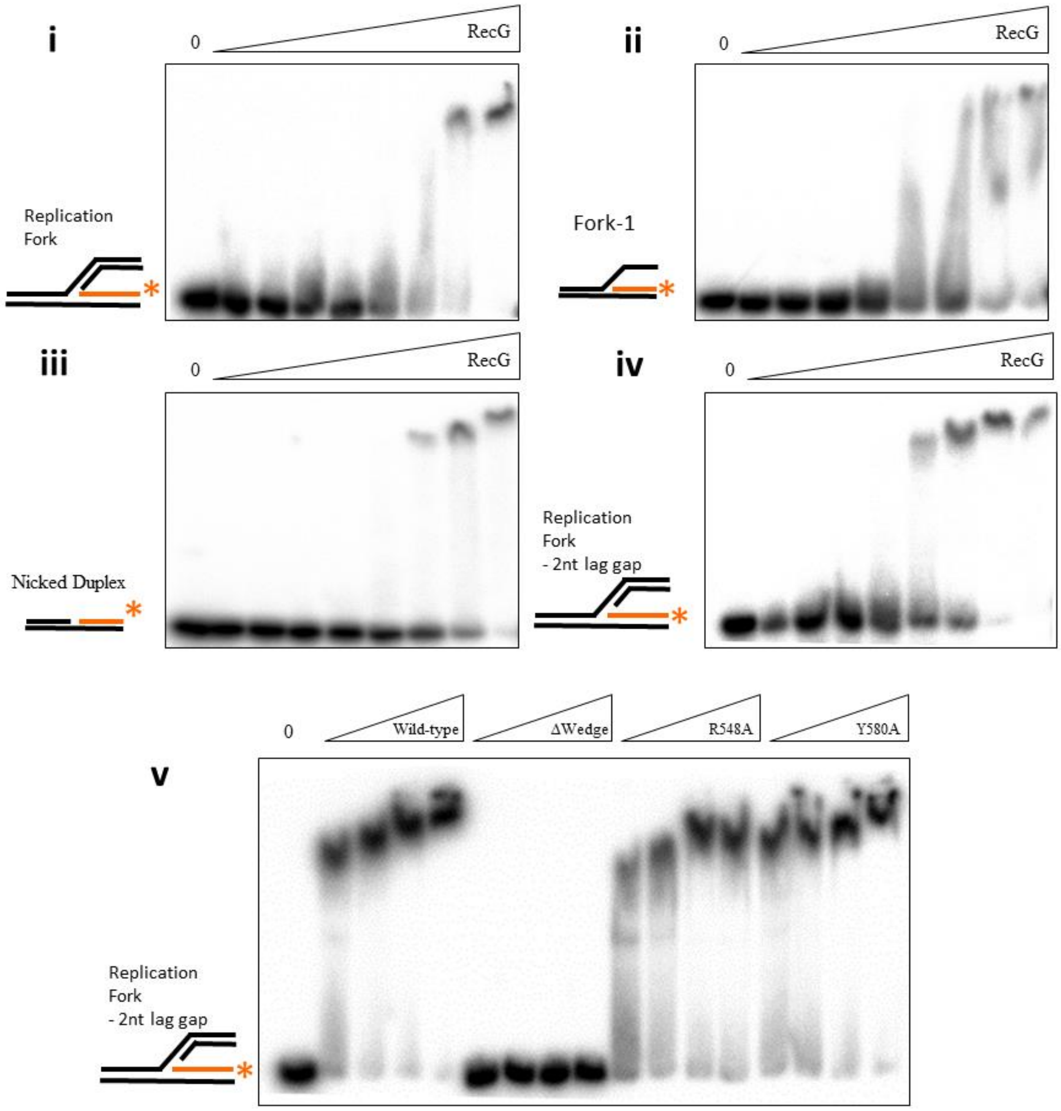


Figure 5-5. EMSA analysis of titrating RecG into various DNA substrates. In panels i – iv RecG was used at ; 0, 3.9, 7.8, 15.625, 31.25, 62.5, 125, 250 and 500nM with 6nM Replication fork (2nt lag gap) in reactions containing 125mM NaCl and 10mM EDTA. (v) RecG proteins were used at 0, 100, 200, 400 and 800nM 500nM with 5nM DNA substrate in reactions containing 125mM NaCl and 10mM EDTA. Each panel is a phosphorimage of native-PAGE (TBE) gel with specific in gel complexes formed with substrates labelled with ^{32}P (*).

From the genetic analysis carried out the biochemical importance of RecG and PriA in primed adaptation was therefore investigated for attenuation or activation of Cas1 catalytic activities. The role of RecB is currently thought to act in a model discussed later and was not the focus here. Historical protein stocks of RecG, PriA and various characterised mutants were available as a generous gift from the Robert Lloyd research group, University of Nottingham (*284*). Activity of these proteins was checked through EMSA analysis of RecG proteins (Figure 5-5) and heliCase unwinding assays on both RecG and PriA (Figure 5-6). In binding assays, RecG formed specific in-gel complexes with all the substrates tested (Figure 5-5). Specifically, 50% of 2nt lagging gap substrate was bound at 125nM RecG (Figure 5-5, panel iv) which was two-fold higher than the binding to the other substrates (replication fork, fork-1 and nicked duplex) (Figure 5-5i-iii). Subsequent EMSAs with RecG mutants (Δwedge, R548A and Y580A) on 2nt lagging gap substrate detected similar binding to the wild-type protein except for the Δwedge mutant, previously characterised as having abolished DNA binding activity, which failed to bind the substrate (Figure 5-5v).

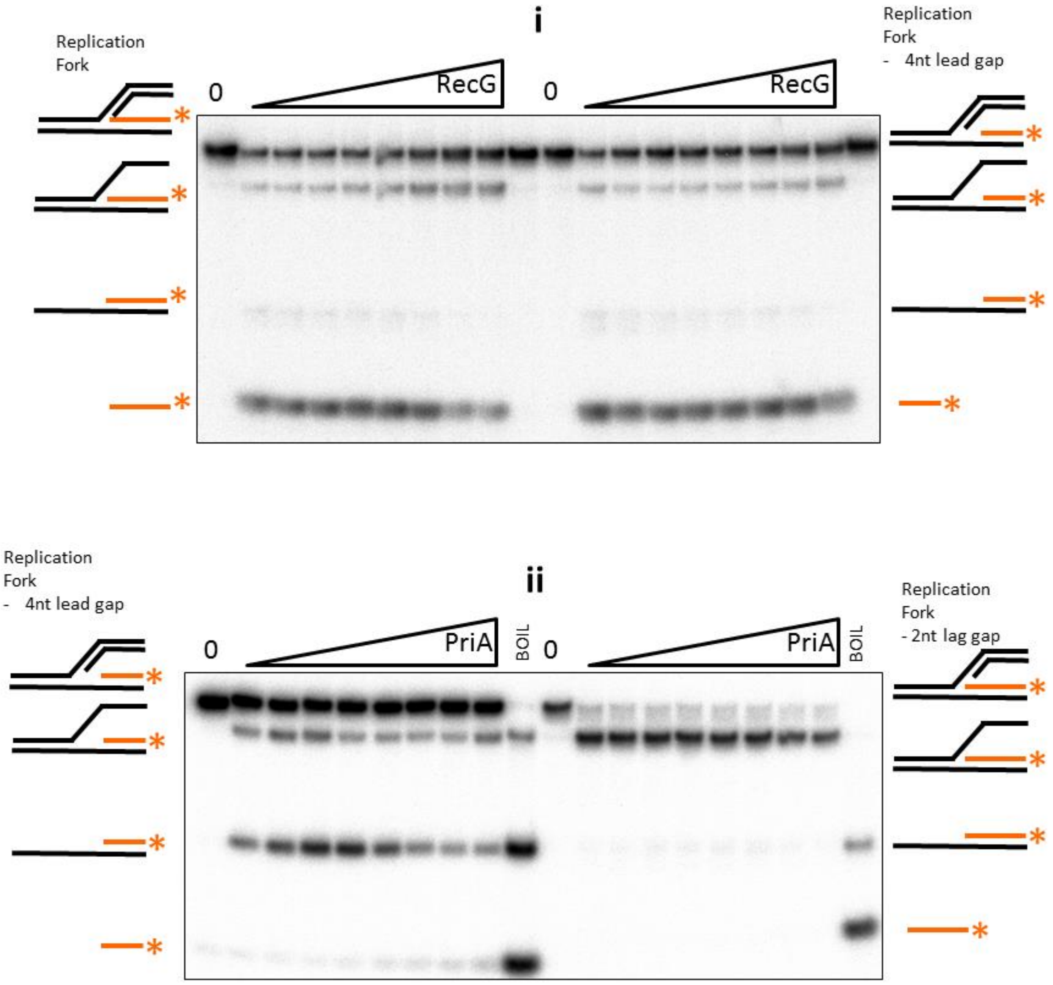

i
Replication Fork
0
RecG
0
RecG
Replication Fork - 4nt lead gap
ii
Replication Fork - 4nt lead gap
0
PriA
BOIL
0
PriA
BOIL
Replication Fork - 2nt lag gap

Figure 5-6. DNA heliCase/unwinding assays titrating RecG or PriA into varied DNA replication fork-like substrates. In panels i and ii RecG or PriA protein was used at 0, 3.9, 7.8, 15.625, 31.25, 62.5, 125, 250 and 500nM with 6nM DNA in reactions containing 125mM NaCl, 10mM $MgCl_2$ and 10mM ATP.

Reactions were terminated by addition of 10mg/ml proteinase K and 10mM EDTA. Each panel shows phosphorimages of native-PAGE (TBE) gels and the species of DNA substrate generated were identified by size and ^{32}P end labelled (*) DNA.

Next, unwinding activity assays showed RecG had 5’ – 3’ heliCase activity, forcing complex DNA structures completely apart, as the major product observed was linear ssDNA (Figure 5-6i). On the other hand, PriA has a 3’ DNA terminus binding pocket that loads on DNA for 3’ – 5’ unwinding, which was consistent with the partial duplex product being formed from the leading strand gap replication fork substrate but not the lagging strand substrate (Figure 5-6ii, left hand side and right hand side titrations, respectively). Therefore, both RecG and PriA proteins could separate strands of different replication fork-like complexes (Figure 5-6). While RecG completely unwinds the substrates to minimal ssDNA, PriA 3’ – 5’ heliCase activity results in a partial duplex or 5’ flapped substrate, which allows fork resetting and replication restart *in vivo*.

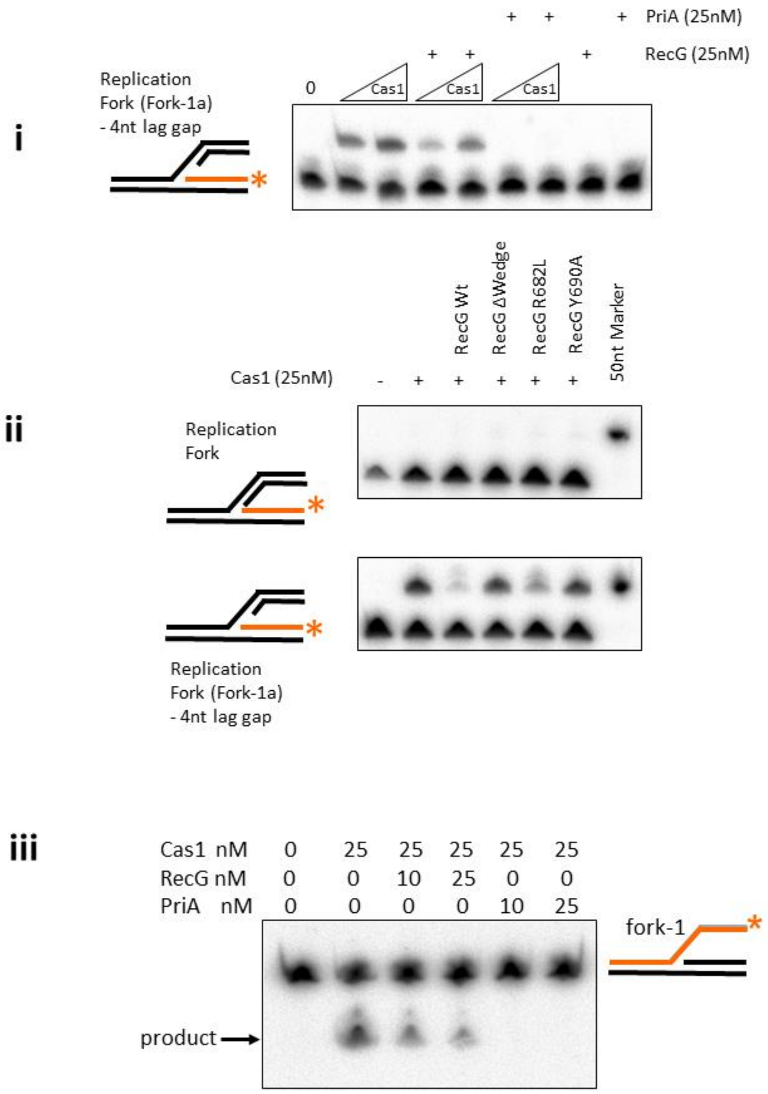


Figure 5-7. The effect of RecG and PriA on the catalytic activity of Cas1. Each panel shows a phosphorimage of a denaturing urea gel separating different species of the ^{32}P labelled (*) DNA strand. Panels i and ii show transesterification products and iii shows nuclease products. (i) Cas1 was used at 0, 12.5 and 25nM with or without RecG or PriA at 25nM. (ii) Cas1 was used at 25nM with 25nM of each RecG protein (wild-type, ΔWedge, R682L, Y690A). (iii) Cas1 was used at 25nM and RecG or PriA were titrated in at 10 and 25nM. Whereas RecG decreased overall activity of Cas1 but PriA completely abolished activity of Cas1. Each reaction contains 6nM DNA substrate in reactions containing 125mM NaCl, 10mM $MgCl_2$ and 10mM ATP.

RecG and PriA are involved in replication fork stability and overall replication rate, therefore, either protein could promote Cas1 nuclease nicking or transesterification activity. First, the latter was tested by adding either RecG or PriA to the replication fork with 4nt lagging strand gap (Fork-1a) which had been previously determined to be the preferred substrate for Cas1 mediated transesterification(see Figure 5-3). RecG reduced Cas1 transesterase activity by ~50% (lanes 4 and 5, Figure 5-7i) whereas PriA completely abolished transesterification (lanes 6 and 7, Figure 5-7i).

PriA had the very same effect also on Cas1 nuclease activity. Whereas RecG only reduced Cas1 nuclease activity (lanes 3 and 4, Figure 5-7iii), PriA abrogated it altogether (lanes 5 and 6, Figure 5-7iii). Loss of Cas1 activity in the presence of PriA was surprising, as the majority of the PriA unwinding assay products, which were identical to fork-1, would have been expected to undergo transesterification and nicking by Cas1. As PriA acts as a molecular scaffold to which other proteins are recruited so it may bind to the 3'-OH required by Cas1.

In addition to the wild-type RecG, Cas1 transesterase activity on fork-1a was also assessed with the different RecG mutants (Figure 5-7ii, bottom panel). RecG R682L mutant reduced Cas1 activity by approximately 50%, similarly to the wild-type protein (lanes 3 and 5, Figure 5-7ii). Interestingly, RecG Y690A had no impact on Cas1 activity (lane 6, Figure 5-7ii). Similarly, RecG Δwedge, which is known to be unable to bind the substrate, did not interfere with Cas1 (lane 4, Figure 5-7ii). Next, the ability of RecG and the different RecG mutants to create a suitable substrate for transesterification from a fully base paired replication fork was examined(Figure 5-7ii, top panel). However, no extension of the labelled strand was detected with either of the RecG proteins.

5.3 Discussion

The aim of this section of work was to identify the enzymatic processes of Cas1 and Cas2 that may be essential to spacer acquisition in the CRISPR immune response and to generate a model for the mechanism. Adaptation is achieved by two distinct routes: naïve, which requires only Cas1 and Cas2 and is independent of CRISPR interference, and primed, which is stimulated by Cascade-Cas3 incomplete interference reactions. Genetic analysis carried out in Croatia by Ivana Inancic-Bace identified differential requirements of host non-Cas proteins essential for CRISPR adaptation. Deletions of *recB* or *polA* were found to abolish naïve adaptation in corroboration with published data that suggests the host repair complex RecBCD is needed for naïve adaptation in *E.coli* to generate the protospacer substrate for integration in the CRISPR array (*67*). On the other hand, the host requirements identified for primed adaptation were: PolA, RecG and PriA (*268*). This indicated these were two distinct pathways, requiring different host proteins for CRISPR adaptation.

A model is presented in Figure 5-8 which illustrates the coordination of naïve and primed adaptation via different enzymatic requirements for transesterification and nuclease cleavage. The model relies upon stalled replication fork remodelling and repair, in the process of which a protospacer substrate for Cas1-Cas2 is generated. This is because RecG, PriA, RecBCD and PolA are all involved in genome repair and stability in *E.coli*, which occurs at stalled replication forks. The model consists of three parts: (a) primed protospacer generation, Cascade R-loop complexes stall replication forks. RecG and PriA stimulate fork remodelling and blockage removal, exposing DNA for Cas1 mediated capture of the protospacer DNA (Figure 5-6). Cas3 nuclease activity, essential for primed adaptation, may generate the protospacer substrate as with RecBCD in naïve acquisition (*65, 67*). (b) In the second possible pathway RecBCD processes DNA ends at sites of DNA damage and collapsed replication forks. Cas1 nicking/nuclease activity described as part of this study could orchestrate the collapse of stalled replication forks themselves and with other nucleases generate the protospacer (Figure 5-2i). (c) In both primed and naïve adaptation, a conserved route is followed after the protospacer-Cas1-Cas2 complex is formed. A series of transesterification reactions insert the protospacer at R1 of the CRISPR array (Figure 5-2ii and Figure 5-4), followed by PolA gap-filling, as in nucleotide excision repair (*285*).

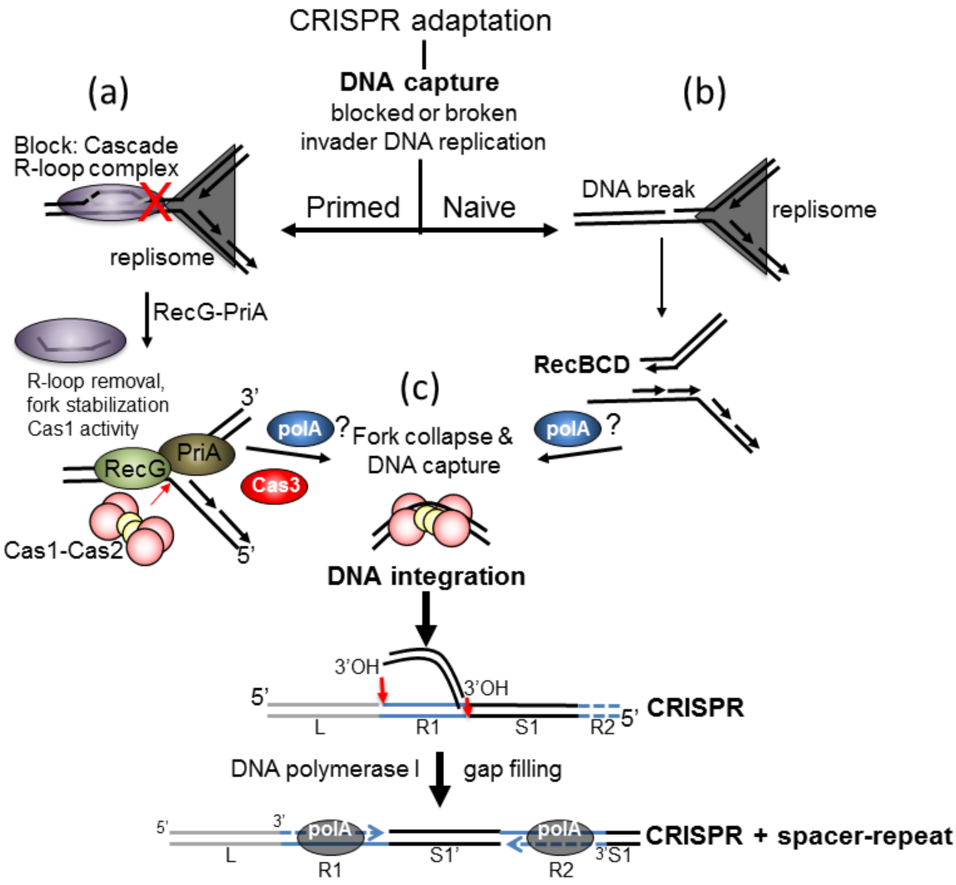


Figure 5-8. Model for spacer acquisition by capture and integration of invader DNA by targeting of blocked or collapsed DNA replication forks in *E. coli*. (A) In primed adaptation Cascade R-loop complexes that cannot stimulate invader degradation arrest advancing replication forks. Fork re-activation is triggered and controlled by RecG and PriA, removing the R-loops and remodeling the blocked fork. Cas1 can then access the invader fork substrate for DNA capture, collapsing the invader replication fork. Further nucleolytic processing of DNA, possibly by Cas1 cutting a fork more than once, or by other nucleases (Cas3) may be required to liberate DNA for Cas1-Cas2 capture. Specific targeting of invader DNA by Cascade ensures non-host targeting. (B) Naïve adaptation relies upon invader DNA damage and hijacked host DNA repair enzymes (RecBCD) stimulating repair. Degradation by RecBCD liberates DNA for Cas1-Cas2 capture. (C) Cas1-Cas2 and protospacer DNA then follow the same method of insertion by a series of transesterification reactions. DNA polymerase I (PolA) is required for both naïve and primed adaptation. PolA comes into play after new spacer integration is successfully catalysed by Cas1 (S1') at structures formed in CRISPR repeats (e.g. R1). Integration leaves DNA gaps flanking the new spacer, requiring synthesis of new repeat DNA, which is catalysed by PolA in a reaction similar to 'gap filling' DNA repair.

The exact mechanism of Cas1 nuclease and transesterification activity in CRISPR spacer acquisition is still unknown. For spacer integration a free 3'-OH at the CRISPR R1 must be present. Cas1 has a distinct preference for binding to replication-fork like structures with branched ssDNA regions (Figure 5-2v), displaying robust catalytic activity on these substrates, (Figure 5-2 and Figure 5-3). This supports the model for primed protospacer capture, whereby Cas1 facilitates spontaneous integration into a CRISPR array, as seen in the SpIN assay (Figure 5-4). However, *in vitro* conditions do not allow further understanding of Cas1 activity in protospacer capture due to the close similarities between substrates. Order of addition assays were carried out with RecG, PriA and single stranded binding protein (SSB) to identify interactions that affected Cas1 activity. Results observed showed no difference in Cas1 activity to the data presented in Figure 5-7. PriA complete inhibition of Cas1 activities is surprising given that PriA is required genetically for primed adaptation (*268*). Again limitations of *in vitro* analysis on small substrates means that PriA is most likely bound to the 3' end of DNA at the branch point, out-competing Cas1 and blocking access for Cas1 catalytic transesterification or nuclease activity. Further experiments to ascertain if the primed model is correct would require *in vitro* reconstitution of the *E.coli* replisome, induced Cascade-Cas3 R-loop blockages and identification of spacer acquisition requiring RecG and/or PriA for replication fork remodelling properties (*286-290*).

6 Discussion

This study focused on the generation of recombinant purified proteins for analysis in *in vitro* biochemical assays to better understand the mechanistic details of the CRISPR/Cas adaptive immune system of prokaryotes. The rapidly evolving CRISPR research field has led to the application of this understanding for technological advances in genome engineering and regulation, CRISPR/Cas9. On the other hand, many of the remaining unknown mechanisms of CRISPR immunity have been overlooked. One example is the analysis of archaeal Cascade complexes and comparing these to the well-defined model of *E.coli*. In coordination with this study, the large Cascade protein Cas8 was identified as essential for *in vivo* CRISPR interference (work carried out in Dr. Anita Marchfelder's research group at The University of Ulm, Germany), and now a model of this mechanism is presented. Through *in vitro* analysis Cas8 proteins bound with a high specificity to R-loop structures (analogous to those present in CRISPR interference), and of particular importance, this apparent binding affinity is influenced by PAM sequence. Cas8 proteins also function in cooperation with other Cascade proteins suspected to form the *Mth* archaeal Cascade complex; Cas5 and Cas7. These discoveries correlate with data that was published at a similar time for the interaction partners and PAM recognition mechanism of CasA, the large Cascade protein from *E.coli*. Interestingly, the asparagine mutant from *Mth* and *Hvo* tested showed PAM sensitivity, similarly, an asparagine residue of CasA is essential for the stabilisation of CasA, CasD (Cas5) and CasC (Cas7) interactions and the stimulation of Cas3 ssDNA degradation. Cas3, the targeted ssDNA endo- and exonuclease identifies a new class of programmed nucleases. Unlike the well-known restriction endonucleases commonly used for molecular cloning that cleave in a sequence specific manner, Cas3 identifies a correctly formed stabilised R-loop with crRNA and a Cascade complex to initiate cleavage. The developed understanding of both bacterial and archaeal CRISPR interference mechanisms strongly implies diversity in protein sequences and structures with overal process conservation. This indicates a convergent evolution of CRISPR immune systems that regulate horizontal gene transfer. The restriction of the transfer of oligonucleotide information (DNA and RNA) has to share some mechanistic details as the fundamental reactions are the same.

One line of reasoning is that CRISPR evolved from a selfish transposon-like element containing Cas1-Cas2. Cas1 and Cas2 are common between all active CRISPR systems and are predicted to have a conserved spacer acquisition mechanism. One possibility for the origin of this is genome rearrangement and integration of novel gene sequences into the transposable element of the Casposon. As a result, investigations of spacer acquisition allow understanding of mobile and selfish genetic element evolution. With predicted models and the one suggested in this study, these elements utilise host DNA replication machinery as either a binding scaffold or provide the initial substrate for spacer integration. This integration event is thought to have a similar mechanism to retroviral integrases, but with a slightly different goal: CRISPR systems create a library of recently exposed DNA sequences whereas viral integrates rely upon integration to avoid detection and replicate. These mechanisms are similar as they both require host factors for stabilisation into the host genome. With a full CRISPR immune system the benefits of spacer integration and inference of invading genetic elements is apparent, a viability advantage for controlling genetic material size within a cell or preventing phage-induced lysis. This is an example of a mutualistic relationship between CRISPR and the host. Whereas, viral integration is selfish and parasitic as the only goal is its own proliferation.

With the presented model from both the archaeal interference and bacterial adaptation, and the recent literature more fully explaining the mechanism of naïve spacer acquisition. The next step would be experimental design to understand the precise requirements of the host proteins RecG, PriA and PolA. The ideal way to set this investigation up would be the creation of an *in vitro* rolling circle replication assay wherein the *E.coli* replication machinery continuously replicates a plasmid. This assay has been set up by other research groups and could then be adapted and targeted by Cascade or Cas9 with an associated crRNA molecule to create stalled replication forks. The addition of RecG and PriA proteins with Cas1 and Cas2 would then allow analysis of the DNA to characterise the mechanism of primed spacer acquisition. Primed acquisition could also be further investigated in archaea by *in vitro* reconstitution of the *Mth* Cascade complex and blocking a replication fork. Similarities between the archaeal and bacterial adaptation mechanisms could therefore be further compared to identify the essential reactions between Cas1 and Cas2, host factors and Cascade proteins. Understanding of these mechanisms opens new possibilities of molecular tools for novel applications downstream.

7 References

1. U. Pul *et al.*, Identification and characterization of *E. coli* CRISPR-Cas promoters and their silencing by H-NS. *Mol Microbiol* **75**, 1495-1512 (2010).
2. D. Song, J. J. Loparo, Building bridges within the bacterial chromosome. *Trends Genet* **31**, 164-173 (2015).
3. N. Delihas, Impact of small repeat sequences on bacterial genome evolution. *Genome Biol Evol* **3**, 959-973 (2011).
4. R. Moxon, C. Bayliss, D. Hood, Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. *Annu Rev Genet* **40**, 307-333 (2006).
5. K. Dybvig, DNA rearrangements and phenotypic switching in prokaryotes. *Mol Microbiol* **10**, 465-471 (1993).
6. M. W. van der Woude, A. J. Baumler, Phase and antigenic variation in bacteria. *Clin Microbiol Rev* **17**, 581-611, table of contents (2004).
7. F. Wisniewski-Dye, L. Vial, Phase and antigenic variation mediated by genome modifications. *Antonie Van Leeuwenhoek* **94**, 493-515 (2008).
8. C. R. Woese, Interpreting the universal phylogenetic tree. *Proc Natl Acad Sci U S A* **97**, 8392-8396 (2000).
9. O. G. Berg, C. G. Kurland, Evolution of microbial genomes: sequence acquisition and loss. *Mol Biol Evol* **19**, 2265-2276 (2002).
10. A. R. Francis, M. M. Tanaka, Evolution of variation in presence and absence of genes in bacterial pathways. *BMC Evol Biol* **12**, 55 (2012).
11. F. Baumdicker, W. R. Hess, P. Pfaffelhuber, The infinitely many genes model for the distributed genome of bacteria. *Genome Biol Evol* **4**, 443-456 (2012).
12. T. Davidsen, T. Tonjum, Meningococcal genome dynamics. *Nat Rev Microbiol* **4**, 11-22 (2006).
13. J. Lederberg, The transformation of genetics by DNA: an anniversary celebration of Avery, MacLeod and McCarty (1944). *Genetics* **136**, 423-426 (1994).
14. T. J. Beveridge, Structures of gram-negative cell walls and their derived membrane vesicles. *J Bacteriol* **181**, 4725-4733 (1999).
15. M. Touchon *et al.*, Organised Genome Dynamics in the Escherichia coli Species Results in Highly Diverse Adaptive Pathsin *PLoS Genet* 5, (2009).
16. S. N. Cohen, A. C. Y. Chang, L. Hsu, Nonchromosomal Antibiotic Resistance in Bacteria: Genetic Transformation of *Escherichia coli* by R-Factor DNA*. *Proc Natl Acad Sci U S A* **69**, 2110-2114 (1972).
17. D. Hanahan, Studies on transformation of *Escherichia coli* with plasmids. *J Mol Biol* **166**, 557-580 (1983).
18. K. Willi, H. Sandmeier, E. M. Kulik, J. Meyer, Transduction of antibiotic resistance markers among *Actinobacillus actinomycetemcomitans* strains by temperate bacteriophages Aa phi 23. *Cell Mol Life Sci* **53**, 904-910 (1997).
19. O. Popa, T. Dagan, Trends and barriers to lateral gene transfer in prokaryotes. *Curr Opin Microbiol* **14**, 615-623 (2011).
20. L. D. McDaniel *et al.*, High frequency of horizontal gene transfer in the oceans. *Science* **330**, 50 (2010).
21. M. Solioz, H. C. Yen, B. Marris, Release and uptake of gene transfer agent by *Rhodopseudomonas capsulata*. *J Bacteriol* **123**, 651-657 (1975).
22. H. Ochman, J. G. Lawrence, E. A. Groisman, Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299-304 (2000).
23. S. Garcia-Vallve, A. Romeu, J. Palau, Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res* **10**, 1719-1725 (2000).

24. J. G. Lawrence, H. Ochman, Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci U S A* **95**, 9413-9417 (1998).

25. J. A. Heinemann, G. F. Sprague, Jr., Bacterial conjugative plasmids mobilize DNA transfer between bacteria and yeast. *Nature* **340**, 205-209 (1989).

26. Y. G. Yoon, M. D. Koob, Transformation of isolated mammalian mitochondria by bacterial conjugation. *Nucleic Acids Res* **33**, e139 (2005).

27. P. Forterre, Darwin's goldmine is still open: variation and selection run the world. *Front Cell Infect Microbiol* **2**, 106 (2012).

28. P. Forterre, D. Prangishvili, The major role of viruses in cellular evolution: facts and hypotheses. *Curr Opin Virol* **3**, 558-565 (2013).

29. M. Krupovic, P. Forterre, Single-stranded DNA viruses employ a variety of mechanisms for integration into host genomes. *Ann N Y Acad Sci* **1341**, 41-53 (2015).

30. S. A. Lujan, L. M. Guogas, H. Ragonese, S. W. Matson, M. R. Redinbo, Disrupting antibiotic resistance propagation by inhibiting the conjugative DNA relaxase. *Proc Natl Acad Sci U S A* **104**, 12282-12287 (2007).

31. A. Ilangovan, S. Connery, G. Waksman, Structural biology of the Gram-negative bacterial conjugation systems. *Trends Microbiol* **23**, 301-310 (2015).

32. K. M. Derbyshire, T. A. Gray, Distributive Conjugal Transfer: New Insights into Horizontal Gene Transfer and Genetic Exchange in Mycobacteria. *Microbiol Spectr* **2**, (2014).

33. I. R. Henderson, P. Owen, J. P. Nataro, Molecular switches--the ON and OFF of bacterial phase variation. *Mol Microbiol* **33**, 919-932 (1999).

34. M. W. van der Woude, Phase variation: how to create and coordinate population diversity. *Curr Opin Microbiol* **14**, 205-211 (2011).

35. J. E. Samson, A. H. Magadan, M. Sabri, S. Moineau, Revenge of the phages: defeating bacterial defences. *Nat Rev Microbiol* **11**, 675-687 (2013).

36. F. Hayes, Toxins-antitoxins: plasmid maintenance, programmed cell death, and cell cycle arrest. *Science* **301**, 1496-1499 (2003).

37. K. Gerdes, S. K. Christensen, A. Lobner-Olesen, Prokaryotic toxin-antitoxin stress response loci. *Nat Rev Microbiol* **3**, 371-382 (2005).

38. D. P. Pandey, K. Gerdes, Toxin-antitoxin loci are highly abundant in free-living but lost from host-associated prokaryotes. *Nucleic Acids Res* **33**, 966-976 (2005).

39. Q. Tan, N. Awano, M. Inouye, YeeV is an *Escherichia coli* toxin that inhibits cell division by targeting the cytoskeleton proteins, FtsZ and MreB. *Mol Microbiol* **79**, 109-118 (2011).

40. C. Unoson, E. G. Wagner, A small SOS-induced toxin is targeted against the inner membrane in *Escherichia coli*. *Mol Microbiol* **70**, 258-270 (2008).

41. J. Robson, J. L. McKenzie, R. Cursons, G. M. Cook, V. L. Arcus, The vapBC operon from *Mycobacterium smegmatis* is an autoregulated toxin-antitoxin module that controls growth via inhibition of translation. *J Mol Biol* **390**, 353-367 (2009).

42. T. R. Blower *et al.*, Mutagenesis and functional characterization of the RNA and protein components of the toxIN abortive infection and toxin-antitoxin locus of *Erwinia*. *J Bacteriol* **191**, 6029-6039 (2009).

43. B. C. Ramisetty, B. Natarajan, R. S. Santhosh, mazEF-mediated programmed cell death in bacteria: "what is this?". *Crit Rev Microbiol* **41**, 89-100 (2015).

44. X. Wang, T. K. Wood, Toxin-antitoxin systems influence biofilm and persister cell formation and the general stress response. *Appl Environ Microbiol* **77**, 5577-5583 (2011).

45. P. C. Fineran *et al.*, The phage abortive infection system, ToxIN, functions as a protein-RNA toxin-antitoxin pair. *Proc Natl Acad Sci U S A* **106**, 894-899 (2009).

46. I. Kobayashi, A. Nobusato, N. Kobayashi-Takahashi, I. Uchiyama, Shaping the genome--restriction-modification systems as mobile genetic elements. *Curr Opin Genet Dev* **9**, 649-656 (1999).

47. I. Kobayashi, Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Res* **29**, 3742-3756 (2001).

48. K. Vasu, V. Nagaraja, Diverse functions of restriction-modification systems in addition to cellular defense. *Microbiol Mol Biol Rev* **77**, 53-72 (2013).

49. Y. Nakayama, I. Kobayashi, Restriction-modification gene complexes as selfish gene entities: roles of a regulatory system in their establishment, maintenance, and apoptotic mutual exclusion. *Proc Natl Acad Sci U S A* **95**, 6442-6447 (1998).

50. L. I. Glatman, A. F. Moroz, M. B. Yablokova, B. A. Rebentish, G. V. Kcholmina, A novel plasmid-mediated DNA restriction-modification system in *E. coli*. *Plasmid* **4**, 350-351 (1980).

51. R. Vaisvila, G. Vilkaitis, A. Janulaitis, Identification of a gene encoding a DNA invertase-like enzyme adjacent to the *PaeR7I* restriction-modification system. *Gene* **157**, 81-84 (1995).

52. K. Ishikawa, E. Fukuda, I. Kobayashi, Conflicts targeting epigenetic systems and their resolution by cell death: novel concepts for methyl-specific and other restriction systems. *DNA Res* **17**, 325-342 (2010).

53. J. Reeks, J. H. Naismith, M. F. White, CRISPR interference: a structural perspective. *Biochem J* **453**, 155-166 (2013).

54. J. van der Oost, M. M. Jore, E. R. Westra, M. Lundgren, S. J. Brouns, CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends Biochem Sci* **34**, 401-407 (2009).

55. R. T. Dame, The role of nucleoid-associated proteins in the organization and compaction of bacterial chromatin. *Mol Microbiol* **56**, 858-870 (2005).

56. C. J. Dorman, H-NS: a universal regulator for a dynamic genome. *Nat Rev Microbiol* **2**, 391-400 (2004).

57. R. T. Dame, C. Wyman, N. Goosen, Structural basis for preferential binding of H-NS to curved DNA. *Biochimie* **83**, 231-234 (2001).

58. M. Barth, C. Marschall, A. Muffler, D. Fischer, R. Hengge-Aronis, Role for the histone-like protein H-NS in growth phase-dependent and osmotic regulation of sigma S and many sigma S-dependent genes in Escherichia coli. *J Bacteriol* **177**, 3455-3464 (1995).

59. C. J. Dorman, H-NS, the genome sentinel. *Nat Rev Microbiol* **5**, 157-161 (2007).

60. C. Ueguchi, T. Mizuno, The *Escherichia coli* nucleoid protein H-NS functions directly as a transcriptional repressor. *Embo j* **12**, 1039-1046 (1993).

61. A. Spassky, S. Rimsky, H. Garreau, H. Buc, H1a, an *E. coli* DNA-binding protein which accumulates in stationary phase, strongly compacts DNA in vitro. *Nucleic Acids Res* **12**, 5321-5340 (1984).

62. R. T. Dame, C. Wyman, N. Goosen, H-NS mediated compaction of DNA visualised by atomic force microscopy. *Nucleic Acids Res* **28**, 3504-3510 (2000).

63. S. Keeney, Spo11 and the Formation of DNA Double-Strand Breaks in Meiosis. *Genome Dyn Stab* **2**, 81-123 (2008).

64. B. Michel, H. Boubakri, Z. Baharoglu, M. LeMasson, R. Lestini, Recombination proteins and rescue of arrested replication forks. *DNA Repair (Amst)* **6**, 967-980 (2007).
65. M. S. Dillingham, S. C. Kowalczykowski, RecBCD enzyme and the repair of double-stranded DNA breaks. *Microbiol Mol Biol Rev* **72**, 642-671, Table of Contents (2008).
66. G. R. Smith, Homologous recombination in *E. coli*: multiple pathways for multiple reasons. *Cell* **58**, 807-809 (1989).
67. A. Levy *et al.*, CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* **520**, 505-510 (2015).
68. A. Reymer, S. Babik, M. Takahashi, B. Norden, T. Beke-Somfai, ATP Hydrolysis in the RecA-DNA Filament Promotes Structural Changes at the Protein-DNA Interface. *Biochemistry* **54**, 4579-4582 (2015).
69. W. J. t. Graham, M. L. Rolfsmeier, C. A. Haseltine, An archaeal RadA paralog influences presynaptic filament formation. *DNA Repair (Amst)* **12**, 403-413 (2013).
70. A. N. Osipov *et al.*, Activation of homologous recombination DNA repair in human skin fibroblasts continuously exposed to X-ray radiation. *Oncotarget*, (2015).
71. H. Sun *et al.*, Application of the Cre/loxP Site-Specific Recombination System for Gene Transformation in *Aurantiochytrium limacinum*. *Molecules* **20**, 10110-10121 (2015).
72. J. S. Mawer, D. R. Leach, Branch migration prevents DNA loss during double-strand break repair. *PLoS Genet* **10**, e1004485 (2014).
73. M. Le Masson, Z. Baharoglu, B. Michel, ruvA and ruvB mutants specifically impaired for replication fork reversal. *Mol Microbiol* **70**, 537-548 (2008).
74. H. George *et al.*, RuvAB-mediated branch migration does not involve extensive DNA opening within the RuvB hexamer. *Curr Biol* **10**, 103-106 (2000).
75. K. Ishioka, H. Iwasaki, H. Shinagawa, Roles of the recG gene product of *Escherichia coli* in recombination repair: effects of the delta recG mutation on cell division and chromosome partition. *Genes Genet Syst* **72**, 91-99 (1997).
76. T. R. Meddows, A. P. Savory, R. G. Lloyd, RecG heliCase promotes DNA double-strand break repair. *Mol Microbiol* **52**, 119-132 (2004).
77. A. V. Gregg, P. McGlynn, R. P. Jaktaji, R. G. Lloyd, Direct rescue of stalled DNA replication forks via the combined action of PriA and RecG heliCase activities. *Mol Cell* **9**, 241-251 (2002).
78. E. L. Bolt, G. J. Sharples, R. G. Lloyd, Analysis of conserved basic residues associated with DNA binding (Arg69) and catalysis (Lys76) by the RusA holliday junction resolvase. *J Mol Biol* **304**, 165-176 (2000).
79. L. Wardrope, E. Okely, D. Leach, Resolution of joint molecules by RuvABC and RecG following cleavage of the *Escherichia coli* chromosome by EcoKI. *PLoS One* **4**, e6542 (2009).
80. J. Garcia-Luis, F. Machin, Mus81-Mms4 and Yen1 resolve a novel anaphase bridge formed by noncanonical Holliday junctions. *Nat Commun* **5**, 5652 (2014).
81. N. C. Fonville, M. D. Blankschien, D. B. Magner, S. M. Rosenberg, RecQ-dependent death-by-recombination in cells lacking RecG and UvrD. *DNA Repair (Amst)* **9**, 403-413 (2010).
82. P. Swuec, A. Costa, Molecular mechanism of double Holliday junction dissolution. *Cell Biosci* **4**, 36 (2014).

83. A. H. Bizard, I. D. Hickson, The dissolution of double Holliday junctions. *Cold Spring Harb Perspect Biol* **6**, a016477 (2014).

84. N. Bocquet *et al.*, Structural and mechanistic insight into Holliday-junction dissolution by topoisomerase IIIalpha and RMI1. *Nat Struct Mol Biol* **21**, 261-268 (2014).

85. J. H. Sutherland, Y. C. Tse-Dinh, Analysis of RuvABC and RecG involvement in the *escherichia coli* response to the covalent topoisomerase-DNA complex. *J Bacteriol* **192**, 4445-4451 (2010).

86. S. Sarbajna, S. C. West, Holliday junction processing enzymes as guardians of genome stability. *Trends Biochem Sci* **39**, 409-419 (2014).

87. J. L. Smith, A. D. Grossman, In Vitro Whole Genome DNA Binding Analysis of the Bacterial Replication Initiator and Transcription Factor DnaA. *PLoS Genet* **11**, e1005258 (2015).

88. M. Wolanski, R. Donczew, A. Zawilak-Pawlik, J. Zakrzewska-Czerwinska, oriC-encoded instructions for the initiation of bacterial chromosome replication. *Front Microbiol* **5**, 735 (2014).

89. C. J. Rudolph, A. L. Upton, R. G. Lloyd, Replication fork collisions cause pathological chromosomal amplification in cells lacking RecG DNA translocase. *Mol Microbiol* **74**, 940-955 (2009).

90. C. J. Rudolph, A. A. Mahdi, A. L. Upton, R. G. Lloyd, RecG protein and single-strand DNA exonucleases avoid cell lethality associated with PriA helicase activity in *Escherichia coli*. *Genetics* **186**, 473-492 (2010).

91. C. J. Rudolph, A. L. Upton, G. S. Briggs, R. G. Lloyd, Is RecG a general guardian of the bacterial genome? *DNA Repair (Amst)* **9**, 210-223 (2010).

92. C. J. Rudolph, A. L. Upton, L. Harris, R. G. Lloyd, Pathological replication in cells lacking RecG DNA translocase. *Mol Microbiol* **73**, 352-366 (2009).

93. R. P. Jaktaji, R. G. Lloyd, PriA supports two distinct pathways for replication restart in UV-irradiated *Escherichia coli* cells. *Mol Microbiol* **47**, 1091-1100 (2003).

94. W. Choi, S. Jang, R. M. Harshey, Mu transpososome and RecBCD nuclease collaborate in the repair of simple Mu insertions. *Proc Natl Acad Sci U S A* **111**, 14112-14117 (2014).

95. A. O. Paatero *et al.*, *Bacteriophage Mu* integration in yeast and mammalian genomes. *Nucleic Acids Res* **36**, e148 (2008).

96. I. Konieczny, J. Marszalek, The requirement for molecular chaperones in lambda DNA replication is reduced by the mutation pi in lambda P gene, which weakens the interaction between lambda P protein and DnaB helicase. *J Biol Chem* **270**, 9792-9799 (1995).

97. A. Szambowska, M. Pierechod, G. Wegrzyn, M. Glinkowska, Coupling of transcription and replication machineries in lambda DNA replication initiation: evidence for direct interaction of *Escherichia coli* RNA polymerase and the lambdaO protein. *Nucleic Acids Res* **39**, 168-177 (2011).

98. I. Datta, S. Sau, A. K. Sil, N. C. Mandal, The bacteriophage lambda DNA replication protein P inhibits the oriC DNA- and ATP-binding functions of the DNA replication initiator protein DnaA of *Escherichia coli*. *J Biochem Mol Biol* **38**, 97-103 (2005).

99. R. Odegrip, S. Schoen, E. Haggard-Ljungquist, K. Park, D. K. Chattoraj, The interaction of *bacteriophage P2* B protein with *Escherichia coli* DnaB heliCase. *J Virol* **74**, 4057-4063 (2000).

100. E. L. Bolt, R. G. Lloyd, G. J. Sharples, Genetic analysis of an archaeal Holliday junction resolvase in *Escherichia coli*. *J Mol Biol* **310**, 577-589 (2001).

101. C. S. Bond, M. Kvaratskhelia, D. Richard, M. F. White, W. N. Hunter, Structure of Hjc, a Holliday junction resolvase, from *Sulfolobus solfataricus*. *Proc Natl Acad Sci U S A* **98**, 5509-5514 (2001).

102. J. M. Mason *et al.*, RAD54 family translocases counter genotoxic effects of RAD51 in human tumor cells. *Nucleic Acids Res* **43**, 3180-3196 (2015).

103. W. D. Wright, W. D. Heyer, Rad54 functions as a heteroduplex DNA pump modulated by its DNA substrates and Rad51 during D loop formation. *Mol Cell* **53**, 420-432 (2014).

104. R. Zahra, J. K. Blackwood, J. Sales, D. R. Leach, Proofreading and secondary structure processing determine the orientation dependence of CAG x CTG trinucleotide repeat instability in *Escherichia coli*. *Genetics* **176**, 27-41 (2007).

105. X. Bi, L. F. Liu, A replicational model for DNA recombination between direct repeats. *J Mol Biol* **256**, 849-858 (1996).

106. M. Bichara, J. Wagner, I. B. Lambert, Mechanisms of tandem repeat instability in bacteria. *Mutat Res* **598**, 144-163 (2006).

107. T. Q. Trinh, R. R. Sinden, Preferential DNA secondary structure mutagenesis in the lagging strand of replication in *E. coli*. *Nature* **352**, 544-547 (1991).

108. D. J. Pinder, C. E. Blake, J. C. Lindsey, D. R. Leach, Replication strand preference for deletions associated with DNA palindromes. *Mol Microbiol* **28**, 719-727 (1998).

109. D. R. Leach, E. A. Okely, D. J. Pinder, Repair by recombination of DNA containing a palindromic sequence. *Mol Microbiol* **26**, 597-606 (1997).

110. J. K. Eykelenboom, J. K. Blackwood, E. Okely, D. R. Leach, SbcCD causes a double-strand break at a DNA palindrome in the *Escherichia coli* chromosome. *Mol Cell* **29**, 644-651 (2008).

111. R. Prakash *et al.*, Yeast Mph1 helicase dissociates Rad51-made D-loops: implications for crossover control in mitotic recombination. *Genes Dev* **23**, 67-79 (2009).

112. D. K. Nag, S. J. Cavallo, Effects of mutations in SGS1 and in genes functionally related to SGS1 on inverted repeat-stimulated spontaneous unequal sister-chromatid exchange in yeast. *BMC Mol Biol* **8**, 120 (2007).

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

114. D. H. Haft, J. Selengut, E. F. Mongodin, K. E. Nelson, A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput Biol* **1**, e60 (2005).

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

116. N. Beloglazova *et al.*, A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. *J Biol Chem* **283**, 20361-20371 (2008).

117. J. S. Godde, A. Bickerton, The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. *J Mol Evol* **62**, 718-729 (2006).

118. M. Babu *et al.*, A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. *Mol Microbiol* **79**, 484-502 (2011).

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

120. C. Rousseau, M. Gonnet, M. Le Romancer, J. Nicolas, CRISPI: a CRISPR interactive database. *Bioinformatics* **25**, 3317-3318 (2009).

121. L. A. Marraffini, E. J. Sontheimer, CRISPR interference limits horizontal gene transfer in *staphylococci* by targeting DNA. *Science* **322**, 1843-1845 (2008).

122. A. B. Hickman, F. Dyda, CRISPR-Cas immunity and mobile DNA: a new superfamily of DNA transposons encoding a Cas1 endonuclease. *Mob DNA* **5**, 23 (2014).

123. M. Krupovic, K. S. Makarova, P. Forterre, D. Prangishvili, E. V. Koonin, Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. *BMC Biol* **12**, 36 (2014).

124. K. S. Makarova, L. Aravind, Y. I. Wolf, E. V. Koonin, Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biol Direct* **6**, 38 (2011).

125. K. S. Makarova *et al.*, Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol* **9**, 467-477 (2011).

126. K. A. Datsenko *et al.*, Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat Commun* **3**, 945 (2012).

127. A. Plagens, H. Richter, E. Charpentier, L. Randau, DNA and RNA interference mechanisms by CRISPR-Cas surveillance complexes. *FEMS Microbiol Rev* **39**, 442-463 (2015).

128. N. Beloglazova *et al.*, CRISPR RNA binding and DNA target recognition by purified Cascade complexes from *Escherichia coli*. *Nucleic Acids Res* **43**, 530-543 (2015).

129. J. Brendel *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* **289**, 7164-7177 (2014).

130. M. Jinek *et al.*, Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* **343**, 1247997 (2014).

131. H. Nishimasu *et al.*, Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**, 935-949 (2014).

132. E. Charpentier, L. A. Marraffini, Harnessing CRISPR-Cas9 immunity for genetic engineering. *Curr Opin Microbiol* **19**, 114-119 (2014).

133. W. Jiang, L. A. Marraffini, CRISPR-Cas: New Tools for Genetic Manipulations from Bacterial Immunity Systems. *Annu Rev Microbiol*, (2015).

134. W. T. Hendriks, X. Jiang, L. Daheron, C. A. Cowan, TALEN- and CRISPR/Cas9-Mediated Gene Editing in Human Pluripotent Stem Cells Using Lipid-Based Transfection. *Curr Protoc Stem Cell Biol* **34**, 5b.3.1-5b.3.25 (2015).

135. K. Kaur, H. Tandon, A. K. Gupta, M. Kumar, CrisprGE: a central hub of CRISPR/Cas-based genome editing. *Database (Oxford)* **2015**, bav055 (2015).

136. H. Liu *et al.*, CRISPR-ERA: a comprehensive design tool for CRISPR-mediated gene editing, repression and activation. *Bioinformatics*, (2015).

137. C. Rouillon *et al.*, Structure of the CRISPR interference complex CSM reveals key similarities with Cascade. *Mol Cell* **52**, 124-134 (2013).

138. S. Bailey, The Cmr complex: an RNA-guided endoribonuclease. *Biochem Soc Trans* **41**, 1464-1467 (2013).

139. J. van der Oost, E. R. Westra, R. N. Jackson, B. Wiedenheft, Unravelling the structural and mechanistic basis of CRISPR-Cas systems. *Nat Rev Microbiol* **12**, 479-492 (2014).

140. J. Reeks *et al.*, Structure of the archaeal Cascade subunit Csa5: relating the small subunits of CRISPR effector complexes. *RNA Biol* **10**, 762-769 (2013).

141. R. A. Garrett *et al.*, CRISPR-Cas Adaptive Immune Systems of the *Sulfolobales*: Unravelling Their Complexity and Diversity. *Life (Basel)* **5**, 783-817 (2015).

142. G. Vestergaard, R. A. Garrett, S. A. Shah, CRISPR adaptive immune systems of Archaea. *RNA Biol* **11**, 156-167 (2014).

143. J. A. Howard, S. Delmas, I. Ivancic-Bace, E. L. Bolt, Helicase dissociation and annealing of RNA-DNA hybrids by *Escherichia coli* Cas3 protein. *Biochem J* **439**, 85-95 (2011).

144. S. D. Cass *et al.*, The role of Cas8 in type I CRISPR interference. *Biosci Rep* **35**, (2015).

145. D. Bikard, L. A. Marraffini, Control of gene expression by CRISPR-Cas systems. *F1000Prime Rep* **5**, 47 (2013).

146. R. Perez-Rodriguez *et al.*, Envelope stress is a trigger of CRISPR RNA-mediated DNA silencing in *Escherichia coli*. *Mol Microbiol* **79**, 584-599 (2011).

147. I. Yosef, M. G. Goren, U. Qimron, Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res*, (2012).

148. J. K. Nunez *et al.*, Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nat Struct Mol Biol* **21**, 528-534 (2014).

149. R. Kiro, M. G. Goren, I. Yosef, U. Qimron, CRISPR adaptation in *Escherichia coli* subtypeI-E system. *Biochem Soc Trans* **41**, 1412-1415 (2013).

150. C. Rollie, S. Schneider, A. S. Brinkmann, E. L. Bolt, M. F. White, Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition. *Elife* **4**, (2015).

151. E. Savitskaya, E. Semenova, V. Dedkov, A. Metlitskaya, K. Severinov, High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in *E. coli*. *RNA Biol* **10**, 716-725 (2013).

152. D. C. Swarts, C. Mosterd, M. W. van Passel, S. J. Brouns, CRISPR interference directs strand specific spacer acquisition. *PLoS One* **7**, e35888 (2012).

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

154. J. Carte, N. T. Pfister, M. M. Compton, R. M. Terns, M. P. Terns, Binding and cleavage of CRISPR RNA by Cas6. *RNA* **16**, 2181-2188 (2010).

155. E. Deltcheva *et al.*, CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602-607 (2011).

156. E. Semenova *et al.*, The Cas6e ribonuclease is not required for interference and adaptation by the *E. coli* type I-E CRISPR-Cas system. *Nucleic Acids Res* **43**, 6049-6061 (2015).

157. R. Cencic *et al.*, Protospacer adjacent motif (PAM)-distal sequences engage CRISPR Cas9 DNA target cleavage. *PLoS One* **9**, e109213 (2014).

158. M. D. Szczelkun *et al.*, Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. *Proc Natl Acad Sci U S A* **111**, 9798-9803 (2014).

159. I. Ivancic-Bace, J. A. Howard, E. L. Bolt, Tuning in to interference: R-loops and Cascade complexes in CRISPR immunity. *J Mol Biol* **422**, 607-616 (2012).

160. G. Cannone, M. Webber-Birungi, L. Spagnolo, Electron microscopy studies of Type III CRISPR machines in *Sulfolobus solfataricus*. *Biochem Soc Trans* **41**, 1427-1430 (2013).

161. N. Beloglazova *et al.*, Structure and activity of the Cas3 HD nuclease MJ0384, an effector enzyme of the CRISPR interference. *Embo j* **30**, 4616-4627 (2011).

162. M. L. Hochstrasser *et al.*, CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. *Proc Natl Acad Sci U S A* **111**, 6618-6623 (2014).

163. Y. Huo *et al.*, Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. *Nat Struct Mol Biol* **21**, 771-777 (2014).

164. S. Mulepati, S. Bailey, Structural and biochemical analysis of nuclease domain of clustered regularly interspaced short palindromic repeat (CRISPR)-associated protein 3 (Cas3). *J Biol Chem* **286**, 31896-31903 (2011).

165. T. Sinkunas *et al.*, Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *EMBO J* **30**, 1335-1342 (2011).

166. T. Sinkunas, G. Gasiunas, V. Siksnys, Cas3 nuclease-helicase activity assays. *Methods Mol Biol* **1311**, 277-291 (2015).

167. E. R. Westra *et al.*, CRISPR Immunity Relies on the Consecutive Binding and Degradation of Negatively Supercoiled Invader DNA by Cascade and Cas3. *Mol Cell*, (2012).

168. T. Karvelis *et al.*, crRNA and tracrRNA guide Cas9-mediated DNA interference in *Streptococcus thermophilus*. *RNA Biol* **10**, 841-851 (2013).

169. E. R. Westra *et al.*, H-NS-mediated repression of CRISPR-based immunity in *Escherichia coli* K12 can be relieved by the transcription activator *LeuO*. *Mol Microbiol* **77**, 1380-1393 (2010).

170. R. P. Cordeiro, D. O. Krause, J. H. Doria, R. A. Holley, Role of the BaeSR two-component regulatory system in resistance of *Escherichia coli* O157:H7 to allyl isothiocyanate. *Food Microbiol* **42**, 136-141 (2014).

171. L. Imamovic, A. Martinez-Castillo, C. Benavides, M. Muniesa, BaeSR, involved in envelope stress response, protects against lysogenic conversion by Shiga toxin 2-encoding phages. *Infect Immun* **83**, 1451-1457 (2015).

172. A. G. Patterson, J. T. Chang, C. Taylor, P. C. Fineran, Regulation of the Type I-F CRISPR-Cas system by CRP-cAMP and GalM controls spacer acquisition and interference. *Nucleic Acids Res* **43**, 6038-6048 (2015).

173. Z. Arslan, V. Hermanns, R. Wurm, R. Wagner, U. Pul, Detection and characterization of spacer integration intermediates in type I-E CRISPR-Cas system. *Nucleic Acids Res* **42**, 7884-7893 (2014).

174. T. Liu *et al.*, Transcriptional regulator-mediated activation of adaptation genes triggers CRISPR *de novo* spacer acquisition. *Nucleic Acids Res* **43**, 1044-1055 (2015).

175. A. Manica, Z. Zebec, J. Steinkellner, C. Schleper, Unexpectedly broad target recognition of the CRISPR-mediated virus defence system in the archaeon *Sulfolobus solfataricus*. *Nucleic Acids Res* **41**, 10509-10517 (2013).

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

177. C. Anders, O. Niewoehner, A. Duerst, M. Jinek, Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **513**, 569-573 (2014).

178. W. Peng *et al.*, Genetic determinants of PAM-dependent DNA targeting and pre-crRNA processing in *Sulfolobus islandicus*. *RNA Biol* **10**, 738-748 (2013).

179. E. R. Westra *et al.*, Type I-E CRISPR-Cas systems discriminate target from non-target DNA through base pairing-independent PAM recognition. *PLoS Genet* **9**, e1003742 (2013).

180. P. C. Fineran *et al.*, Degenerate target sites mediate rapid primed CRISPR adaptation. *Proc Natl Acad Sci U S A* **111**, E1629-1638 (2014).

181. C. Richter *et al.*, Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. *Nucleic Acids Res* **42**, 8516-8526 (2014).

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

183. A. P. Hynes, M. Villion, S. Moineau, Adaptation in bacterial CRISPR-Cas immunity can be driven by defective phages. *Nat Commun* **5**, 4399 (2014).

184. J. K. Nunez, A. S. Lee, A. Engelman, J. A. Doudna, Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature* **519**, 193-198 (2015).

185. N. Beloglazova, S. Lemak, R. Flick, A. F. Yakunin, Analysis of nuclease activity of Cas1 proteins against complex DNA substrates. *Methods Mol Biol* **1311**, 251-264 (2015).

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

187. J. Demeulemeester, J. De Rijck, R. Gijsbers, Z. Debyser, Retroviral integration: Site matters: Mechanisms and consequences of retroviral integration site selection. *Bioessays*, (2015).

188. A. Engelman, P. Cherepanov, Retroviral Integrase Structure and DNA Recombination Mechanism. *Microbiol Spectr* **2**, (2014).

189. D. P. Maskell *et al.*, Structural basis for retroviral integration into nucleosomes. *Nature* **523**, 366-369 (2015).

190. L. Krishnan, A. Engelman, Retroviral integrase proteins and HIV-1 DNA integration. *J Biol Chem* **287**, 40858-40866 (2012).

191. D. Ka, D. Kim, G. Baek, E. Bae, Structural and functional characterization of Streptococcus pyogenes Cas2 protein under different pH conditions. *Biochem Biophys Res Commun* **451**, 152-157 (2014).

192. K. H. Nam *et al.*, Double-stranded endonuclease activity in *Bacillus halodurans* clustered regularly interspaced short palindromic repeats (CRISPR)-associated Cas2 protein. *J Biol Chem* **287**, 35943-35952 (2012).

193. S. Lemak *et al.*, Toroidal structure and DNA cleavage by the CRISPR-associated [4Fe-4S] cluster containing Cas4 nuclease SSO0001 from *Sulfolobus solfataricus*. *J Am Chem Soc* **135**, 17476-17487 (2013).

194. J. Zhang, T. Kasciukovic, M. F. White, The CRISPR associated protein Cas4 Is a 5' to 3' DNA exonuclease with an iron-sulfur cluster. *PLoS One* **7**, e47232 (2012).

195. A. Plagens *et al.*, In vitro assembly and activity of an archaeal CRISPR-Cas type I-A Cascade interference complex. *Nucleic Acids Res* **42**, 5125-5138 (2014).

196. K. H. Nam, I. Kurinov, A. Ke, Crystal structure of clustered regularly interspaced short palindromic repeats (CRISPR)-associated Csn2 protein revealed Ca2+-dependent double-stranded DNA binding activity. *J Biol Chem* **286**, 30759-30768 (2011).

197. J. Reeks *et al.*, Structure of a dimeric crenarchaeal Cas6 enzyme with an atypical active site for CRISPR RNA processing. *Biochem J* **452**, 223-230 (2013).

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

199. R. D. Sokolowski, S. Graham, M. F. White, Cas6 specificity and CRISPR RNA loading in a complex CRISPR-Cas system. *Nucleic Acids Res* **42**, 6532-6541 (2014).

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

201. R. Wang, H. Zheng, G. Preamplume, Y. Shao, H. Li, The impact of CRISPR repeat sequence on structures of a Cas6 protein-RNA complex. *Protein Sci* **21**, 405-417 (2012).

202. L. Deng, R. A. Garrett, S. A. Shah, X. Peng, Q. She, A novel interference mechanism by a type IIIB CRISPR-Cmr module in *Sulfolobus*. *Mol Microbiol* **87**, 1088-1099 (2013).

203. C. Estarellas *et al.*, Molecular dynamic simulations of protein/RNA complexes: CRISPR/Csy4 endoribonuclease. *Biochim Biophys Acta* **1850**, 1072-1090 (2015).

204. I. Scholz, S. J. Lange, S. Hein, W. R. Hess, R. Backofen, CRISPR-Cas systems in the *cyanobacterium Synechocystis* sp. PCC6803 exhibit distinct processing pathways involving at least two Cas6 and a Cmr2 protein. *PLoS One* **8**, e56470 (2013).

205. E. Charpentier, H. Richter, J. van der Oost, M. F. White, Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. *FEMS Microbiol Rev* **39**, 428-441 (2015).

206. A. A. Su, V. Tripp, L. Randau, RNA-Seq analyses reveal the order of tRNA processing events and the maturation of C/D box and CRISPR RNAs in the hyperthermophile *Methanopyrus kandleri*. *Nucleic Acids Res* **41**, 6250-6258 (2013).

207. K. Chylinski, A. Le Rhun, E. Charpentier, The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems. *RNA Biol* **10**, 726-737 (2013).

208. D. L. Court *et al.*, RNase III: Genetics and function; structure and mechanism. *Annu Rev Genet* **47**, 405-431 (2013).

209. G. Gasiunas, R. Barrangou, P. Horvath, V. Siksnys, Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc Natl Acad Sci U S A* **109**, E2579-2586 (2012).

210. L. Joshua-Tor, The Argonautes. *Cold Spring Harb Symp Quant Biol* **71**, 67-72 (2006).

211. K. Burger, M. Gullerova, Swiss army knives: non-canonical functions of nuclear Drosha and Dicer. *Nat Rev Mol Cell Biol* **16**, 417-430 (2015).

212. A. Kurzynska-Kokorniak *et al.*, The many faces of Dicer: the complexity of the mechanisms regulating Dicer gene expression and enzyme activities. *Nucleic Acids Res* **43**, 4365-4380 (2015).

213. N. G. Lintner *et al.*, Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE). *J Biol Chem* **286**, 21643-21656 (2011).

214. B. Wiedenheft *et al.*, Structures of the RNA-guided surveillance complex from a bacterial immune system. *Nature* **477**, 486-489 (2011).

215. B. Wiedenheft *et al.*, RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proc Natl Acad Sci U S A* **108**, 10092-10097 (2011).

216. L. A. Kira S Makarova, Yuri I Wolf, Eugene V Koonin, Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biology Direct* **6**, 38 (2011).

217. E. L. Garside *et al.*, Cas5d processes pre-crRNA and is a member of a larger family of CRISPR RNA endonucleases. *Rna* **18**, 2020-2028 (2012).

218. A. Punetha, R. Sivathanu, B. Anand, Active site plasticity enables metal-dependent tuning of Cas5d nuclease activity in CRISPR-Cas type I-C system. *Nucleic Acids Res* **42**, 3846-3856 (2014).

219. K. H. Nam *et al.*, Cas5d protein processes pre-crRNA and assembles into a Cascade-like interference complex in subtype I-C/Dvulg CRISPR-Cas system. *Structure* **20**, 1574-1584 (2012).

220. K. H. Nam, Q. Huang, A. Ke, Nucleic acid binding surface and dimer interface revealed by CRISPR-associated CasB protein structures. *FEBS Lett* **586**, 3956-3961 (2012).

221. Y. Agari, S. Yokoyama, S. Kuramitsu, A. Shinkai, X-ray crystal structure of a CRISPR-associated protein, Cse2, from *Thermus thermophilus* HB8. *Proteins* **73**, 1063-1067 (2008).

222. K. Sakamoto *et al.*, X-ray crystal structure of a CRISPR-associated RAMP module [corrected] Cmr5 protein [corrected] from *Thermus thermophilus* HB8. *Proteins* **75**, 528-532 (2009).

223. N. V. G. Kira S Makarova, Svetlana A Shabalina, Yuri I Wolf, Eugene V Koonin, A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biology Direct* **1**, 7 (2011).

224. J. Zhang *et al.*, Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. *Mol Cell* **45**, 303-313 (2012).

225. Y. Shao *et al.*, Structure of the Cmr2-Cmr3 subcomplex of the Cmr RNA silencing complex. *Structure* **21**, 376-384 (2013).

226. S. Mulepati, A. Orr, S. Bailey, Crystal structure of the largest subunit of a bacterial RNA-guided immune complex and its role in DNA target binding. *J Biol Chem* **287**, 22445-22449 (2012).

227. D. G. Sashital, B. Wiedenheft, J. A. Doudna, Mechanism of Foreign DNA Selection in a Bacterial Adaptive Immune System. *Mol Cell*, (2012).

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

229. E. van Duijn *et al.*, Native tandem and ion mobility mass spectrometry highlight structural and modular similarities in clustered-regularly-interspaced shot-palindromic-repeats (CRISPR)-associated protein complexes from *Escherichia coli* and *Pseudomonas aeruginosa*. *Mol Cell Proteomics* **11**, 1430-1441 (2012).

230. D. W. Taylor *et al.*, Structural biology. Structures of the CRISPR-Cmr complex reveal mode of RNA target positioning. *Science* **348**, 581-585 (2015).

231. A. I. Cocozaki *et al.*, Structure of the Cmr2 subunit of the CRISPR-Cas RNA silencing complex. *Structure* **20**, 545-553 (2012).

232. M. Jinek *et al.*, A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821 (2012).

233. C. R. Hale *et al.*, RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* **139**, 945-956 (2009).

234. A. R. Bassett, C. Tibbit, C. P. Ponting, J. L. Liu, Highly Efficient Targeted Mutagenesis of *Drosophila* with the CRISPR/Cas9 System. *Cell Rep* **4**, 220-228 (2013).

235. M. Crispo *et al.*, Efficient Generation of Myostatin Knock-Out Sheep Using CRISPR/Cas9 Technology and Microinjection into Zygotes. *PLoS One* **10**, e0136690 (2015).

236. L. G. Lowder *et al.*, A CRISPR/Cas9 toolbox for multiplexed plant genome editing and transcriptional regulation. *Plant Physiol*, (2015).

237. M. Boettcher, M. T. McManus, Choosing the Right Tool for the Job: RNAi, TALEN, or CRISPR. *Mol Cell* **58**, 575-585 (2015).

238. T. Gaj, C. A. Gersbach, C. F. Barbas, 3rd, ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol* **31**, 397-405 (2013).

239. T. Sprink, J. Metje, F. Hartung, Plant genome editing by novel tools: TALEN and other sequence specific nucleases. *Curr Opin Biotechnol* **32**, 47-53 (2015).

240. Q. Ul Ain, J. Y. Chung, Y. H. Kim, Current and future delivery systems for engineered nucleases: ZFN, TALEN and RGEN. *J Control Release* **205**, 120-127 (2015).

241. D. A. Wright, T. Li, B. Yang, M. H. Spalding, TALEN-mediated genome editing: prospects and perspectives. *Biochem J* **462**, 15-24 (2014).

242. J. Boch, U. Bonas, Xanthomonas AvrBs3 family-type III effectors: discovery and function. *Annu Rev Phytopathol* **48**, 419-436 (2010).

243. M. M. Mahfouz *et al.*, *De novo*-engineered transcription activator-like effector (TALE) hybrid nuclease with novel DNA binding specificity creates double-strand breaks. (2011).

244. F. W. Alt, Y. Zhang, F. L. Meng, C. Guo, B. Schwer, Mechanisms of programmed DNA lesions and genomic instability in the immune system. *Cell* **152**, 417-429 (2013).

245. F. Jiang, K. Zhou, L. Ma, S. Gressel, J. A. Doudna, STRUCTURAL BIOLOGY. A Cas9-guide RNA complex preorganized for target DNA recognition. *Science* **348**, 1477-1481 (2015).

246. X. Liang *et al.*, Rapid and highly efficient mammalian cell engineering via Cas9 protein transfection. *J Biotechnol* **208**, 44-53 (2015).

247. D. Bikard *et al.*, Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic Acids Res* **41**, 7429-7437 (2013).

248. R. Barrangou, RNA-mediated programmable DNA cleavage. *Nature Biotechnology* **30**, 836-838 (2012).

249. X. Li, J. L. Manley, Cotranscriptional processes and their influence on genome stability. *Genes Dev* **20**, 1838-1847 (2006).

250. T. Itoh, J. Tomizawa, Formation of an RNA primer for initiation of replication of ColE1 DNA by ribonuclease H. *Proc Natl Acad Sci U S A* **77**, 2450-2454 (1980).

251. A. Aguilera, T. Garcia-Muse, R loops: from transcription byproducts to threats to genome stability. *Mol Cell* **46**, 115-124 (2012).

252. P. M. Vertino, P. A. Wade, R loops: lassoing DNA methylation at CpGi. *Mol Cell* **45**, 708-709 (2012).

253. D. Y. Lee, D. A. Clayton, RNase mitochondrial RNA processing correctly cleaves a novel R loop at the mitochondrial DNA leading-strand origin of replication. *Genes Dev* **11**, 582-592 (1997).

254. J. Stavnezer, J. E. Guikema, C. E. Schrader, Mechanism and regulation of class switch recombination. *Annu Rev Immunol* **26**, 261-292 (2008).

255. S. Feuerhahn, N. Iglesias, A. Panza, A. Porro, J. Lingner, TERRA biogenesis, turnover and implications for function. *FEBS Lett* **584**, 3812-3818 (2010).

256. W. Yang, Nucleases: diversity of structure, function and mechanism. *Q Rev Biophys* **44**, 1-93 (2011).

257. S. Tuduri *et al.*, Topoisomerase I suppresses genomic instability by preventing interference between replication and transcription. *Nat Cell Biol* **11**, 1315-1324 (2009).

258. S. D. Vincent, A. A. Mahdi, R. G. Lloyd, The RecG branch migration protein of *Escherichia coli* dissociates R-loops. *J Mol Biol* **264**, 713-721 (1996).

259. J. B. Boule, V. A. Zakian, The yeast Pif1p DNA helicase preferentially unwinds RNA DNA substrates. *Nucleic Acids Res* **35**, 5809-5818 (2007).

260. H. E. Mischo *et al.*, Yeast Sen1 helicase protects the genome from transcription-associated instability. *Mol Cell* **41**, 21-32 (2011).

261. R. N. Jackson, M. Lavin, J. Carter, B. Wiedenheft, Fitting CRISPR-associated Cas3 into the helicase family tree. *Curr Opin Struct Biol* **24**, 106-114 (2014).

262. K. Buttner, S. Nehring, K. P. Hopfner, Structural basis for DNA duplex separation by a superfamily-2 helicase. *Nat Struct Mol Biol* **14**, 647-652 (2007).

263. D. Luo *et al.*, Crystal structure of the NS3 protease-helicase from *dengue* virus. *J Virol* **82**, 173-183 (2008).

264. M. M. Jore *et al.*, Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nat Struct Mol Biol* **18**, 529-536 (2011).

265. B. Gong *et al.*, Molecular insights into DNA interference by CRISPR-associated nuclease-helicase Cas3. *Proc Natl Acad Sci U S A* **111**, 16359-16364 (2014).

266. S. T. Abedon, Facilitation of CRISPR adaptation. *Bacteriophage* **1**, 179-181 (2011).

267. E. R. Westra, S. J. Brouns, The rise and fall of CRISPRs--dynamics of spacer acquisition and loss. *Mol Microbiol* **85**, 1021-1025 (2012).

268. I. Ivancic-Bace, S. D. Cass, S. J. Wearne, E. L. Bolt, Different genome stability proteins underpin primed and naive adaptation in *E. coli* CRISPR-Cas immunity. *Nucleic Acids Res* **43**, 10821-10830 (2015).

269. C. P. Guy, A. I. Majernik, J. P. Chong, E. L. Bolt, A novel nuclease-ATPase (Nar71) from archaea is part of a proposed thermophilic DNA repair system. *Nucleic Acids Res* **32**, 6176-6186 (2004).

270. V. M. Bolanos-Garcia, O. R. Davies, Structural analysis and classification of native proteins from *E. coli* commonly co-purified by immobilised metal affinity chromatography. *Biochim Biophys Acta* **1760**, 1304-1313 (2006).

271. X. Veaute *et al.*, UvrD helicase, unlike Rep helicase, dismantles RecA nucleoprotein filaments in *Escherichia coli*. *Embo j* **24**, 180-189 (2005).

272. B. K. Washburn, S. R. Kushner, Construction and analysis of deletions in the structural gene (uvrD) for DNA helicase II of *Escherichia coli*. *J Bacteriol* **173**, 2569-2575 (1991).

273. I. Wong, M. Amaratunga, T. M. Lohman, Heterodimer formation between *Escherichia coli* Rep and UvrD proteins. *J Biol Chem* **268**, 20386-20391 (1993).

274. J. K. Blackwood *et al.*, Structural and functional insights into DNA-end processing by the archaeal HerA helicase-NurA nuclease complex. *Nucleic Acids Res* **40**, 3183-3196 (2012).

275. Q. Huang *et al.*, Efficient 5'-3' DNA end resection by HerA and NurA is essential for cell viability in the crenarchaeon *Sulfolobus islandicus*. *BMC Mol Biol* **16**, 2 (2015).

276. N. J. Rzechorzek *et al.*, Structure of the hexameric HerA ATPase reveals a mechanism of translocation-coupled DNA-end processing in archaea. *Nat Commun* **5**, 5506 (2014).

277. R. N. Jackson *et al.*, Structural biology. Crystal structure of the CRISPR RNA-guided surveillance complex from *Escherichia coli*. *Science* **345**, 1473-1479 (2014).

278. J. A. L. Howard, *Cas3 : R-loop formation as a step in CRISPR acquired immunity*. (2013).

279. E. Semenova *et al.*, Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc Natl Acad Sci U S A* **108**, 10098-10103 (2011).

280. L. K. Maier *et al.*, Essential requirements for the detection and degradation of invaders by the *Haloferax volcanii* CRISPR/Cas system I-B. *RNA Biol* **10**, 865-874 (2013).

281. B. Stoll *et al.*, Requirements for a successful defence reaction by the CRISPR-Cas subtype I-B system. *Biochem Soc Trans* **41**, 1444-1448 (2013).

282. G. S. Bird, H. Takemura, O. Thastrup, J. W. Putney, Jr., F. S. Menniti, Mechanisms of activated Ca2+ entry in the rat pancreatoma cell line, AR4-2J. *Cell Calcium* **13**, 49-58 (1992).

283. S. Fischer *et al.*, An archaeal immune system can detect multiple protospacer adjacent motifs (PAMs) to target invader DNA. *J Biol Chem* **287**, 33351-33363 (2012).

284. J. Zhang, A. A. Mahdi, G. S. Briggs, R. G. Lloyd, in *Genetics*. (2010), vol. 185, pp. 23-37.

285. E. C. Friedberg, The eureka enzyme: the discovery of DNA polymerase. *Nat Rev Mol Cell Biol* **7**, 143-147 (2006).

286. M. K. Gupta *et al.*, Protein-DNA complexes are the primary sources of replication fork pausing in *Escherichia coli*. *Proc Natl Acad Sci U S A* **110**, 7252-7257 (2013).

287. P. McGlynn, R. G. Lloyd, Rescue of stalled replication forks by RecG: simultaneous translocation on the leading and lagging strand templates supports an active DNA unwinding model of fork reversal and Holliday junction formation. *Proc Natl Acad Sci U S A* **98**, 8227-8234 (2001).

288. C. J. Rudolph, P. Dhillon, T. Moore, R. G. Lloyd, Avoiding and resolving conflicts between DNA replication and transcription. *DNA Repair (Amst)* **6**, 981-993 (2007).

289. W. Gan *et al.*, R-loop-mediated genomic instability is caused by impairment of replication fork progression. *Genes Dev* **25**, 2041-2056 (2011).

290. A. Fukuoh, H. Iwasaki, K. Ishioka, H. Shinagawa, ATP-dependent resolution of R-loops at the ColE1 replication origin by *Escherichia coli* RecG protein, a Holliday junction-specific helicase. *EMBO J* **16**, 203-209 (1997).

8 Appendix

8.1 Primer List

All primers used in this study at listed here. Most primers were used for the construction of plasmids containing coding sequences for Cas genes from either *E.coli* or *Methanothermobacter thermautotrphicus* into vectors to add an affinity tag. Hexahistidine (His), Streptactin (Strep) and Maltose binding protein (MBP) tags were added either N- or C-terminally to the coding sequence.

Table 19. Primers used for construction and manipulation of plasmids vectors

Primer Name	Sequence	Used for
Mth1076-A	CCG**GAATTC**CGATAGAGTACTCTTCATG	N-terminal His-Tag *Mth*1076 in pET-Duet-1
Mth1076-B	CCC**AAGCTT**TTATTATAAGCGGGGAAAG	N-terminal His-Tag *Mth*1076 in pET-Duet-1
Mth1077-A	CGC**GGATCC**CAAAAAAAGGAACTCCCCTG	N-terminal His-Tag *Mth*1077 in pET-Duet-1
Mth1077-B	CCG**GAATTC**CTACTATCCATTGAAGATCACC	N-terminal His-Tag *Mth*1077 in pET-Duet-1
Mth1078-A	CCG**GAATTC**CCAATGCACCCTTGAGGTTATAAC	N-terminal His-Tag *Mth*1078 in pET-Duet-1
Mth1078-B	ATAAGAAT**GCGGCCGC**CTACTAGGGGAGTTCCTTTTTTTC	N-terminal His-Tag *Mth*1078 in pET-Duet-1
Mth1079-A	CGC**GGATCC**CCTCGTATACCTCAAACCC	N-terminal His-Tag *Mth*1079 in pET-Duet-1
Mth1079-B	CCC**AAGCTT**TTATTAACCACCACCAATC	N-terminal His-Tag *Mth*1079 in pET-Duet-1
Mth1080-A	CCG**GAATTC**CAGGTTCCAGAAAAATTAT	N-terminal His-Tag *Mth*1080 in pET-Duet-1
Mth1080-B	CCC**AAGCTT**CTCATCACTCTGCAGTTGTTGGG	N-terminal His-Tag *Mth*1080 in pET-Duet-1
Mth1081-A	CGC**GGATCC**CGTGATATCCATGAGTGATTTAAC	N-terminal His-Tag *Mth*1081 in pET-Duet-1
Mth1081-B	CCG**GAATTC**TTATTAGCTTCTTGGGTTGTAG	N-terminal His-Tag *Mth*1081 in pET-Duet-1
Mth1082-A	CCG**GAATTC**CCCGGTGTTCTATATGAATC	N-terminal His-Tag *Mth*1082 in pET-Duet-1
Mth1082-B	CCC**AAGCTT**TCATCACTCATGGATATCACC	N-terminal His-Tag *Mth*1082 in pET-Duet-1
Mth1083-A	CGC**GGATCC**CTGGTGGTAACTGTGTACCTTC	N-terminal His-Tag *Mth*1083 in pET-Duet-1
Mth1083-B	CCC**AAGCTT**TCATCAGAGAATAACATCAAGTG	N-terminal His-Tag *Mth*1083 in pET-Duet-1
Mth1084-A	CGC**GGATCC**CAACTCTGCTGGGCTTGAGC	N-terminal His-Tag *Mth*1084 in pET-Duet-1
Mth1084-B	CCG**GAATTC**TTATTACCACCACATCACTAATG	N-terminal His-Tag *Mth*1084 in pET-Duet-1
Mth1085-A	CGCGGATC**CCGCGG**TTTTCTTGATAATCG	N-terminal His-Tag *Mth*1085 in pET-Duet-1
Mth1085-B	CCG**GAATTC**TCATCAAGCCCAGCAGAGTTC	N-terminal His-Tag *Mth*1085 in pET-Duet-1
Mth1087-A	CCG**GAATTC**CGAAACCCTTGCAGTGGAG	N-terminal His-Tag *Mth*1087 in pET-Duet-1
Mth1087-B	CCC**AAGCTT**TCATCATAGCCACCCGAAATC	N-terminal His-Tag *Mth*1087 in pET-Duet-1
Mth1089-A	CGC**GGATCC**CTATAAGAAAATGAAACTC	N-terminal His-Tag *Mth*1089 in pET-Duet-1
Mth1089-B	CGC**GGATCC**CTATAAGAAAATGAAACTC	N-terminal His-Tag *Mth*1089 in pET-Duet-1
Mth1091-A	CC**CAAGCTT**GAGGGATTCATCGACAGGC	N-terminal His-Tag *Mth*1091 in pET-Duet-1
Mth1091-B	GGAATTC**CATATG**TCATCATGATGTAACCACCATACC	N-terminal His-Tag *Mth*1091 in pET-Duet-1

Mth1624-A	CGC**GGATCC**CAGATCCCTGAGTGGTAAG	N-terminal His-Tag *Mth*1624 in pET-Duet-1
Mth1624-B	CCG**GAATTC**TCATCAACCCTCGGGGACCTC	N-terminal His-Tag *Mth*1624 in pET-Duet-1
Cas7pQE1-A	CG**GAATTC**ATGAAAATGTCAAGGTAC	C-terminal Strep-tag *Mth* Cas7 in pQE-His1
Cas7pQE1-B	GG**GGTACC**GATCCCCTGAACCTTTCCGTAACTC	C-terminal Strep-tag *Mth* Cas7 in pQE-His1
S-Cas5A	G**GAATTC**ATGGAAACCCTTGCAGTGGAG	C-terminal Strep-tag *Mth* Cas5 in pQE-His1
S-Cas5B	CCC**AAGCTT**TAGCCACCCGAAATCTGCAGG	C-terminal Strep-tag *Mth* Cas5 in pQE-His1
S-Cas6A	G**GAATTC**ATGGAGGGATTCATCGACAG	C-terminal Strep-tag *Mth* Cas6 in pQE-His1
S-Cas6B	CCC**AAGCTT**TGATGTAACCACCATACC	C-terminal Strep-tag *Mth* Cas6 in pQE-His1
S-Cas5’A	GAA**GCTTCG**AAACCCTTGCAGTGGAG	N-terminal Strep-tag *Mth* Cas5 in pQE-His2
S-Cas5’B	CCC**AAGCTT**TAGCCACCCGAAATCTGCAGG	N-terminal Strep-tag *Mth* Cas5 in pQE-His2
S-Cas6’A	CCC**AAGCTT**GAGGGATTCATCGACAGG	N-terminal Strep-tag *Mth* Cas6 in pQE-His2
S-Cas6’B	CCG**CTCGAG**TGATGTAACCACCATACC	N-terminal Strep-tag *Mth* Cas6 in pQE-His2
MthCRISPR’-A	CCC**AAGCTT**CGCTATCTCCAGTTTACTTC	*Mth* CRISPR into pUC19
MthCRISPR’-B	TGC**TCTAGA**GATAAAATAGGAGTGGTCC	*Mth* CRISPR into pUC19
MthCRISPRT7	TTGTAATACGACTCACTATAGGGCGCTATCTCCAGTTTACTTC	In vitro transcription of CRISPR
MthCas-1A	ACGC**GTCGAC**ATGGGCTATGATGGAATTAATCG	Cas3-Cas5-Cas7-Cas8-Cas8b into pCDF-1b
MthCas-1B	ATTT**GCGGCCGC**TTATCTGCCTCCTAAAAATCC	Cas3-Cas5-Cas7-Cas8-Cas8b into pCDF-1b
MthCas-2A	ACGC**GTCGAC**ATGGGCTATGATGGAATTAATCG	Cas3-Cas5-Cas7-Cas8-Cas8b-Cas6 In pCDF-1b
MthCas-2B	ATTT**GCGGCCGC**TCATGATGTAACCACCATACC	Cas3-Cas5-Cas7-Cas8-Cas8b-Cas6 In pCDF-1b
***E.Coli* Cas1+2-A**	CATG**CCATGG**ATGACCTGGCTTCCCC	*E.coli* Cas1 and Cas2 into pCDF-1b
***E.Coli* Cas1+2-B**	ATAAGAAT**GCGGCCGC**TCATCAAACAGGTAAAAAAGAC	*E.coli* Cas1 and Cas2 into pCDF-1b
***E.Coli* Cas1-A**	CATG**CCATGG**ATGACCTGGCTTCCCC	*E.coli* Cas1 into pCDF-1b
***E.Coli* Cas1-B**	ATAAGAAT**GCGGCCGC**TCATCAGCTACTCCGATGGCC	*E.coli* Cas1 into pCDF-1b
***E.Coli* Cas2-A**	CATG**CCATGG**ATGAGTATGTTGGTCG	*E.coli* Cas2 into pCDF-1b
***E.Coli* Cas2-B**	ATAAGAAT**GCGGCCGC**TCATCAAACAGGTAAAAAAGAC	*E.coli* Cas2 into pCDF-1b
Mth Cas1+2-A	CATG**CCATGG**ATGAACTCTGCTGGGCTTG	*Mth* Cas1 and Cas2 into pCDF-1b
Mth Cas1+2-B	ATAAGAAT**GCGGCCGC**TCATCAGAGAATAACATC	*Mth* Cas1 and Cas2 into pCDF-1b
Mth Cas1-A	CATG**CCATGG**ATGAACTCTGCTGGGCTTG	*Mth* Cas1 into pCDF-1b
Mth Cas1-B	ATAAGAAT**GCGGCCGC**TCATCACCACCACATCAC	*Mth* Cas1 into pCDF-1b
Mth Cas2-A	CATG**CCATGG**ATGGTGGTAACTGTGTACC	*Mth* Cas2 into pCDF-1b
Mth Cas2-B	ATAAGAAT**GCGGCCGC**TCATCAGAGAATAACATC	*Mth* Cas2 into pCDF-1b
CRISPRamp-1	ATTTTGCGTTTCGTTCAGGT	BL21AI CRISPR amplification to detect spacer acquisition
CRISPRamp-2	TGGATGTGTTGTTTGTGTG	BL21AI CRISPR amplification to detect spacer acquisition
D151G-A	CGAGAAAAATTTAATTGGTAATAATTCAGAGGAAC	Cas8’ QuCh of pEB389
D151G-B	GTTCCTCTGAATTATTACCAATTAAATTTTTCTCG	Cas8’ QuCh of pEB389
N153A-A	ATTTAATTGATAATGCTTCAGAGGAACTGG	Cas8’ QuCh of pEB389
N153A-B	CCAGTTCCTCTGAAGCATTATCAATTAAAT	Cas8’ QuCh of pEB389
E155A-A	TTGATAATAATTCAGCTGAACTGGGAGAGATT	Cas8’ QuCh of pEB389
E155A-B	AATCTCTCCCAGTTCAGCTGAATTATTATCAA	Cas8’ QuCh of pEB389
N536G-A	CCAGAGAGAACAACATAGGTCAGCTAATATCAATCC	Cas8’ QuCh of pEB389

N536G-B	GGATTGATATTAGCTGACCTATGTTGTTCTCTCTGG	Cas8' QuCh of pEB389
S540A-A	ACATAAATCAGCTAATAGCAATCCTTAGGAGGAAC	Cas8' QuCh of pEB389
S540A-B	GTTCCTCCTAAGGATTGCTATTAGCTGATTTATGT	Cas8' QuCh of pEB389
S540A540G	CATAAATCAGCTAATAGGAATCCTTAGGAGGAAC	Cas8' QuCh of pEB389
S540A540G	GTTCCTCCTAAGGATTCCTATTAGCTGATTTATG	Cas8' QuCh of pEB389
Y548F-A	GGAGGAACAACAGGTTCCTCTTCGTTAACAA	Cas8' QuCh of pEB389
Y548F-B	TTGTTAACGAAGAGGAACCTGTTGTTCCTCC	Cas8' QuCh of pEB389
Y548A-A	TAGGAGGAACAACAGGGGCCTCTTCGTTAACAACCTC	Cas8' QuCh of pEB389
Y548A-B	GAGGTTGTTAACGAAGAGGCCCCTGTTGTTCCTCCTA	Cas8' QuCh of pEB389
L542A-A	CAGCTAATATCAATCGCTAGGAGGAACAACAG	Cas8' QuCh of pEB389
L542A-B	CTGTTGTTCCTCCTAGCGATTGATATTAGCTG	Cas8' QuCh of pEB389
Cas5MBP-A	CCG**GAATTC**GAAACCCTTGCAGTGGAGATATTC	N-terminal MBP-tag *Mth* Cas5 in pMal-C2x
Cas5MBP-B	GC**TCTAGA**TCATAGCCACCCGAAATCTGC	N-terminal MBP-tag *Mth* Cas5 in pMal-C2x
Cas6MBP-A	GC**TCTAGA**GAGGGATTCATCGACAGGCCAG	N-terminal MBP-tag *Mth* Cas6 in pMal-C2x
Cas6MBP-B	GG**AAGCTT**TCATGATGTAACCACCATACC	N-terminal MBP-tag *Mth* Cas6 in pMal-C2x
Cas8bpQE-A	AA**CTGCAG**ATGATCTACGAATTCAGACGC	C-terminal Strep-tag *Mth* Cas8' in pQE-His1
Cas8bpQE-B	TTT**GCGGCCGC**TCTGCCTCCTAAAAATCCAGTTAG	C-terminal Strep-tag *Mth* Cas8' in pQE-His1

8.1.1 Cloning

Initially all the CRISPR genes from *Mth* were cloned into suitable plasmid vectors, ORF1076-1091. This includes genes from the Type-IG and Type-IIIA CRISPR subsets. Each gene was cloned individually into one of the following: pET14b, pET-Duet, pET-ACYC, pMal-C2x, pCDF-1b and pBad-A. Some plasmids were already available for ORF genes: 1083 (pJLH9) and 1084 (pJLH7) from Jamieson Howard and: 1085 (pEB377), 1086 (pEB359), 1087 (pEB374), 1088 (pEB388), 1089 (pEB383) and 1090 (pEB389) from Edward Bolt. All cloned into vectors with N-terminal His-tags, except pEB359 which had no tag. Some of these proteins have been expressed and purified previously, such as ORF1090 in Bolt *et al 2004*. To clone the remaining 8 genes, along with some previously cloned genes for expression in different plasmid vectors associated to the CRISPR locus of interest techniques described in Chapter 2 for all stages of molecular cloning. *Mth* aCASCADE consists of Cas5, Cas7, Cas8a2, Cas8' and

Cas6, encoded by ORFs 1087-1091 respectively. aCASCADE is expanded upon in chapter 4.

8.1.1.1 Cas5

Cas5 was originally cloned into pET14b (pEB374). Here Cas5 was sub-cloned into pMal-C2x with primers MBPCas5-1 and MBPCas5-2, introducing restriction sites EcoRI and XbaI. The PCR product was digested and the vector similarly. After restriction clean up the insert was ligated into the vector overnight and then transformed into DH5α and plated onto Amp. DNA was isolated from colonies and analytically digested, and sequenced confirming generation of pSDC25, Figure 8-1.

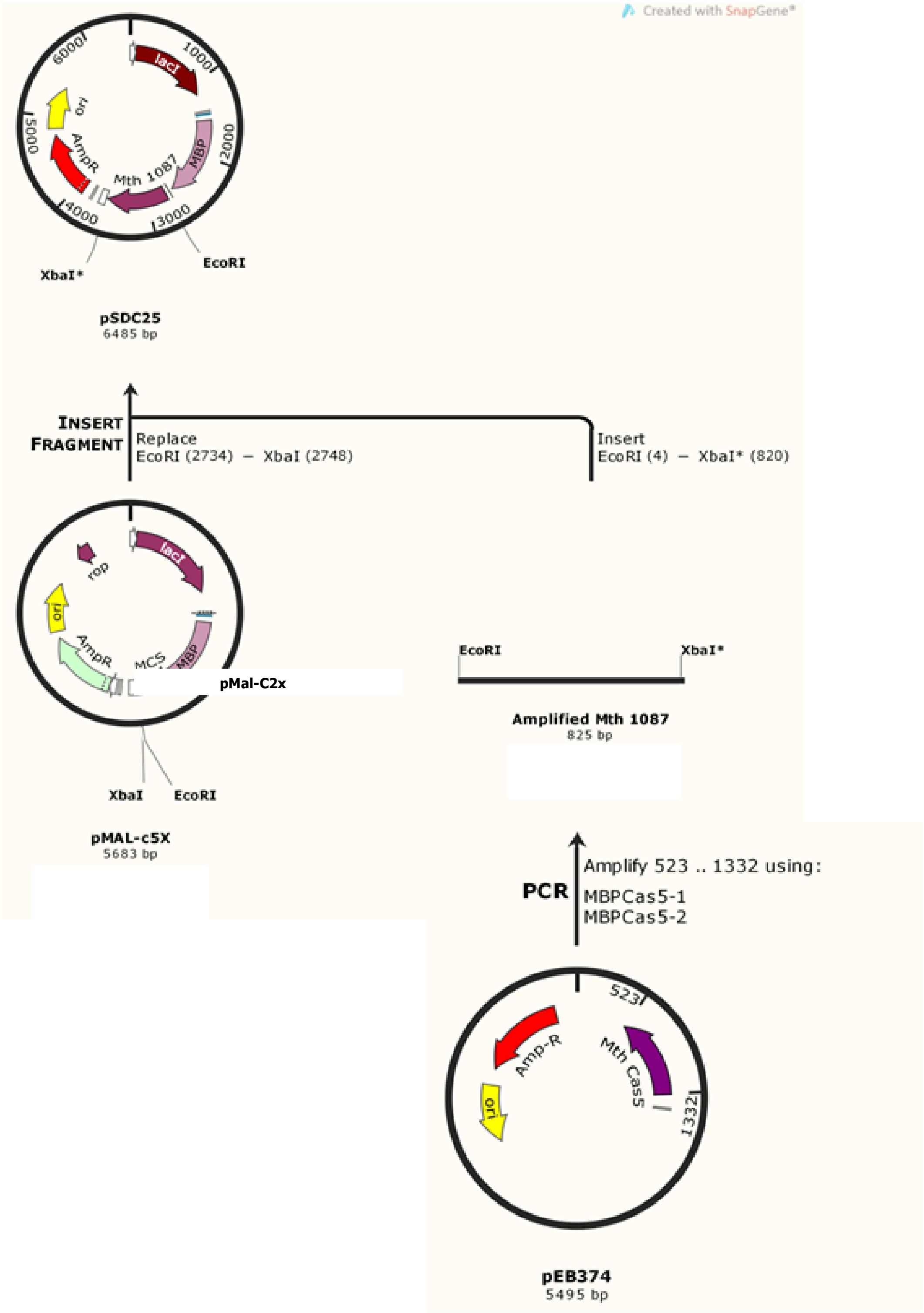


Figure 8-1 Cloning strategy for MBPCas5. Cloning procedure followed for the generation of N-terminally MBP tagged Cas5 from *Mth* (ORF1087), displaying PCR step followed by insert and vector digestion and the resultant construct pSDC25.

8.1.1.2 Cas6

Cas6 (ORF1091) was not available from a previously constructed vector. Plasmids were constructed to include N-terminal His$_6$ and MBP tags on Cas6. These were generated by PCR amplification of available genomic *Mth* DNA (Dr. James Chong, University of York) with HisCas6-1+2 and MBPCas61+2 for respective tag introduction. This introduced NdeI/HindIII and XbaI/HindIII restriction sites respectfully. When amplification was optimised, DNA was excised; restriction digested and cleaned up, followed by ligation and transformation. Transformants were then analysed by DNA extraction and confirmation of cloning. Generating pSDC13 (His-tag, Figure 8-2) and pSDC27 (MBP-tag, Figure 8-3).

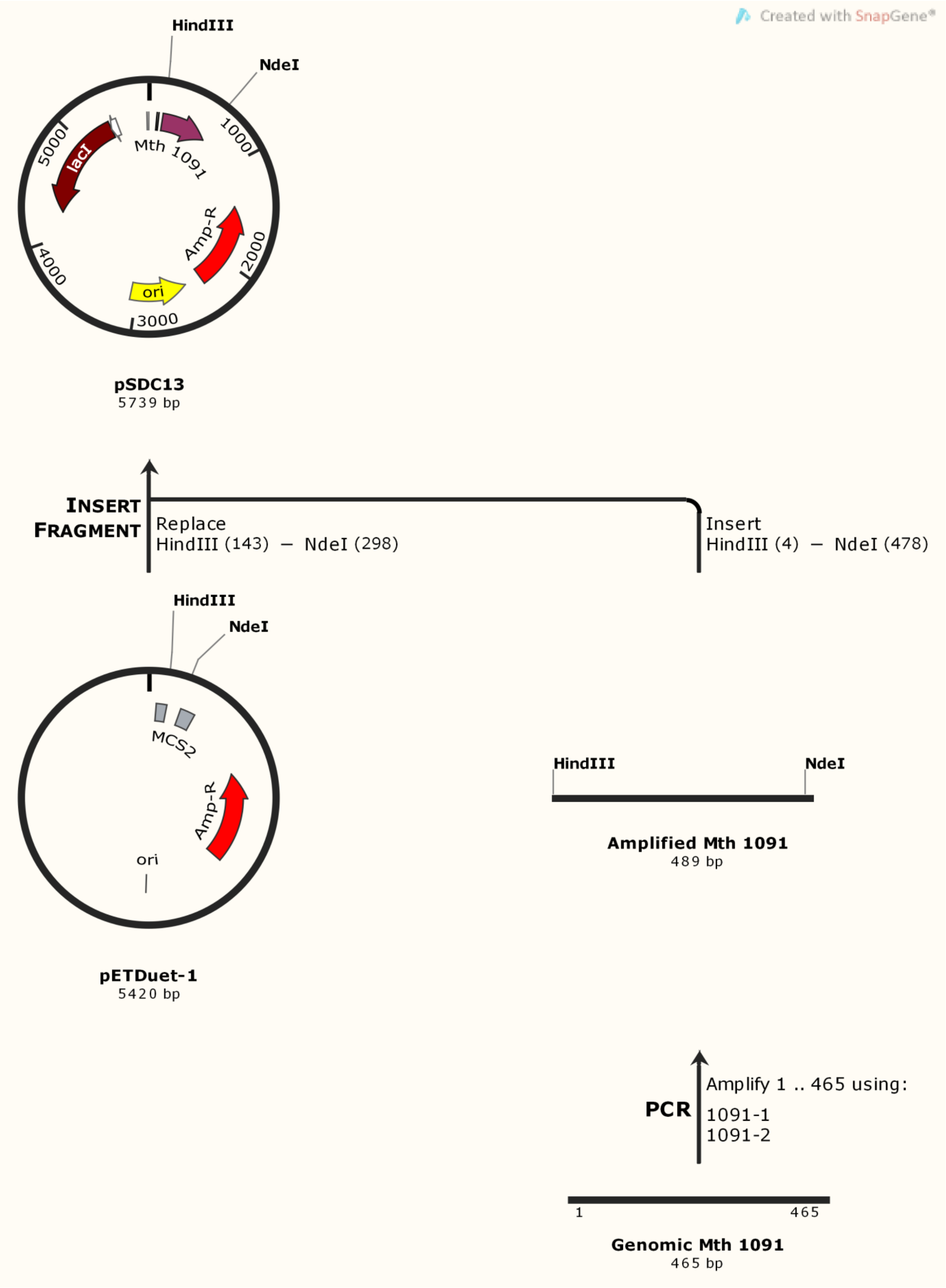


Figure 8-2 Cloning strategy for His$_6$Cas6. Cloning procedure followed for the generation of N-terminally hexahistidine tagged Cas6 from *Mth* (ORF1091), displaying PCR step from genomic DNA and addition of restriction endonuclease sites. Followed by insert and vector digestion and the resultant construct pSDC13.

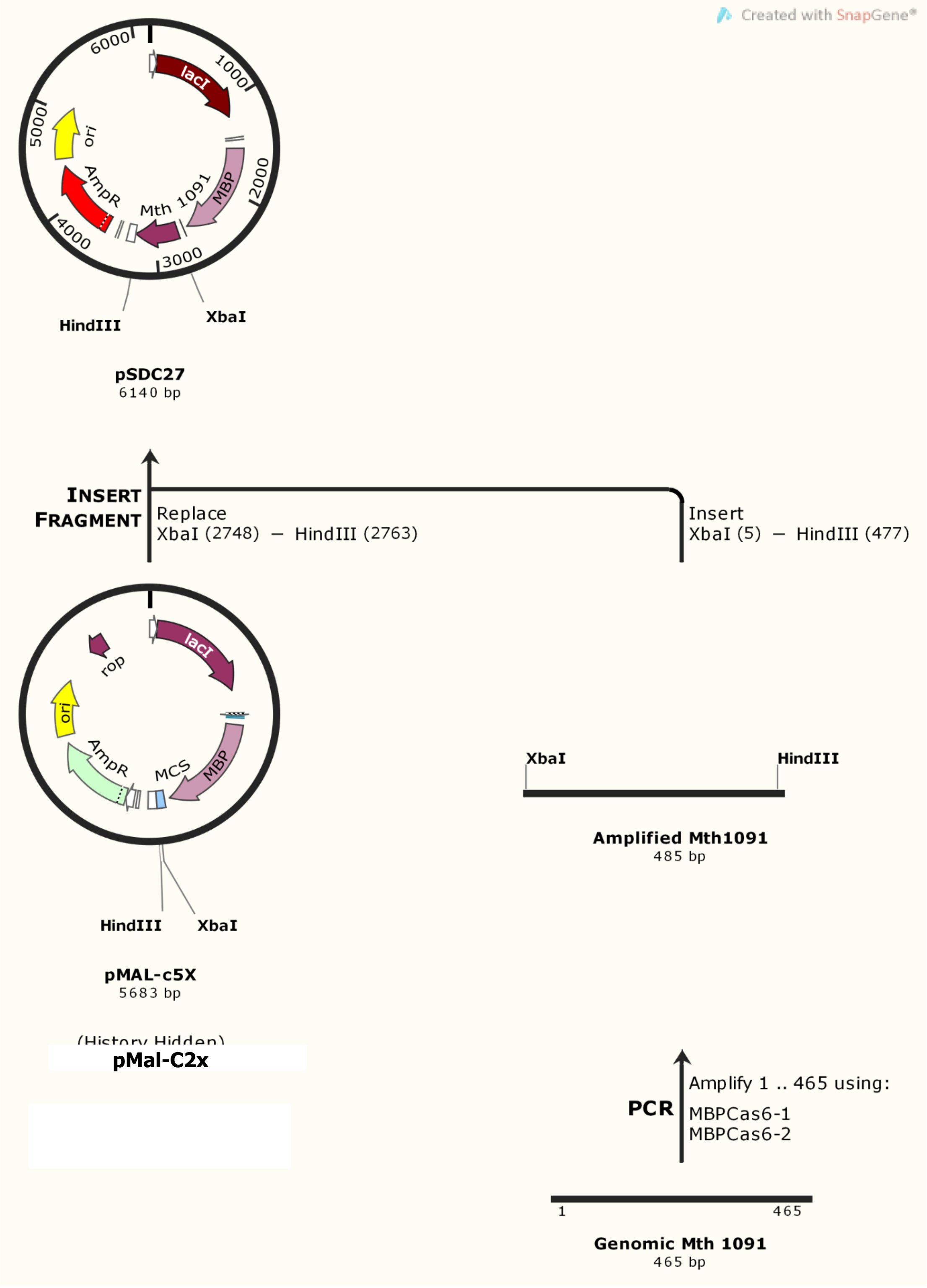


Figure 8-3 Cloning strategy for MBP Cas6. Cloning procedure followed for the generation of N-terminally MBP tagged Cas6 from *Mth* (ORF1091), displaying PCR step from genomic DNA and addition of restriction endonuclease sites. Followed by insert and vector digestion and the resultant construct pSDC27.

8.1.1.3 Cas7

Cas7 was available cloned into pET14b (pEB388) Figure 8-4, but to facilitate co-expression with Cas5 it was sub-cloned into pCDF-1b. pCDF-1b contains Strp^R gene and CDF origin making this plasmid compatible with pET and pMal constructs for co-expression. Cas7 Was first cloned in pQE-His1 from genomic DNA creating pSDC29, Figure 8-5. Then Cas7 was sub-cloned with no tag into pCDF-1b, a simple restriction digest of pSDC29 by NcoI and KpnI liberated the Cas7 gene. This was excised and ligated directly into pCDF-1b that was similarly digested. Colony DNA was purified and confirmed.

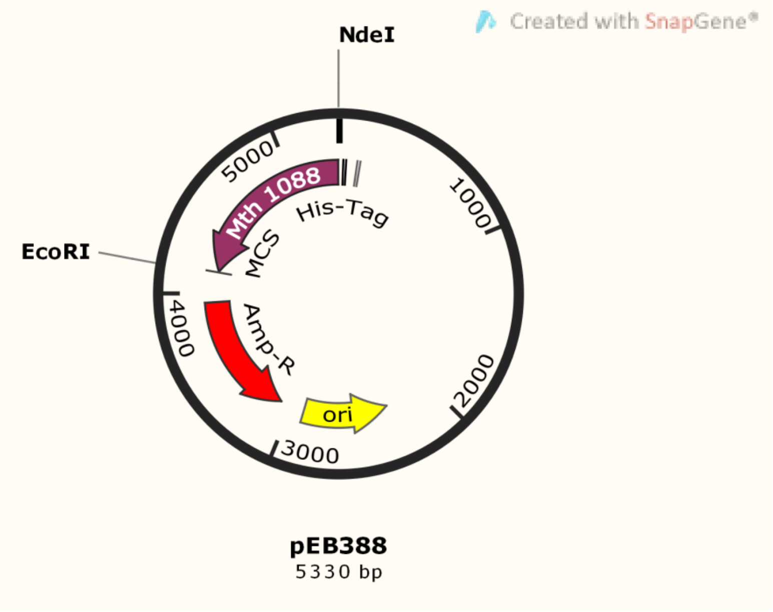


Figure 8-4 Schematic representation of available *Mth* Cas7 clone. pEB388 allowed expression of MtCas7 (ORF 1088) on a pET14-b plasmid for generation of N-terminally hexahistidine tagged Cas7.

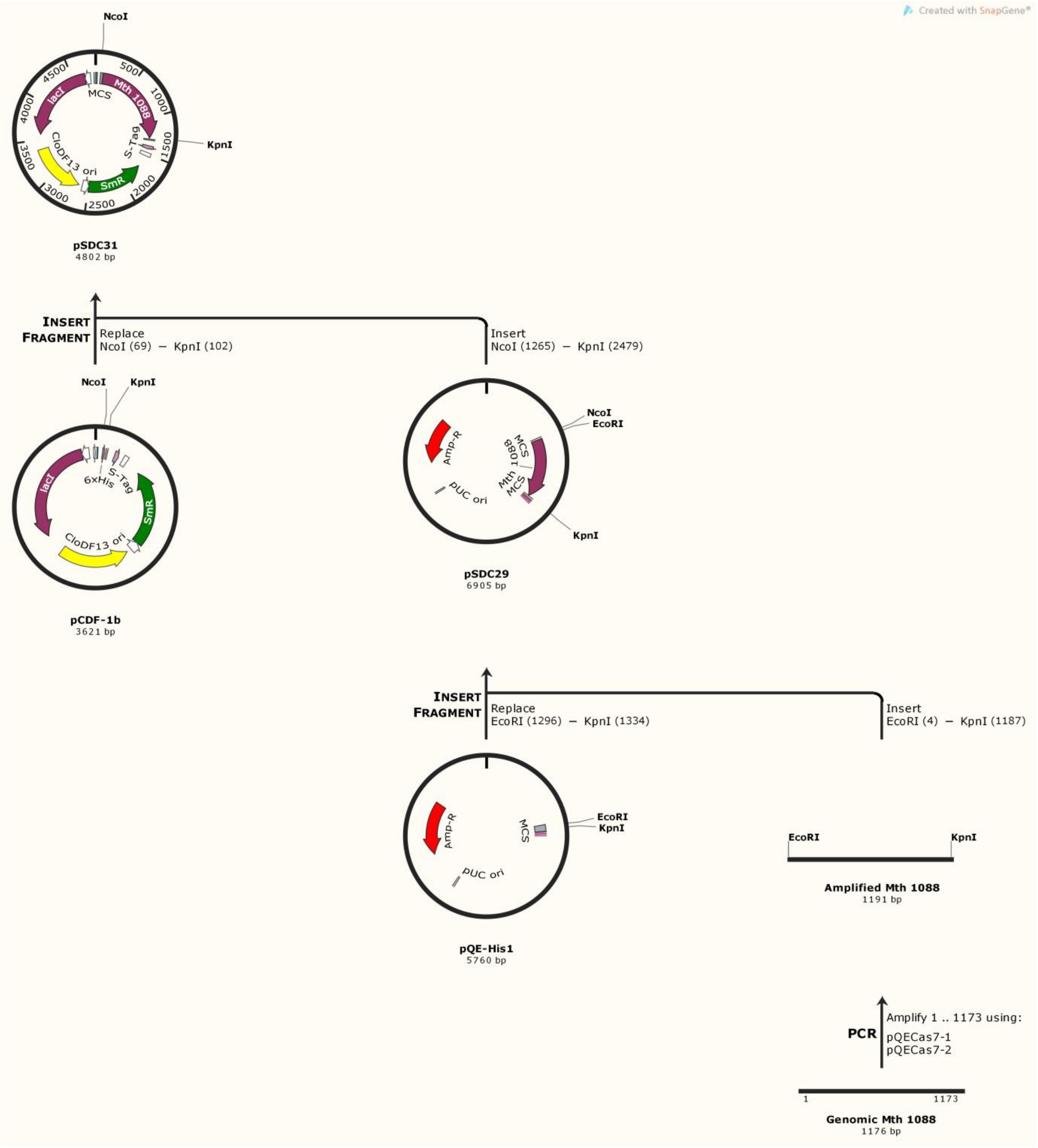


Figure 8-5 Cloning and subcloning of Cas7 to facilitate co-expression with Cas5. *Mth* Cas7 (ORF 1088) was initially cloned into pQE-His1 from genomic DNA. Generating pSDC29. pSDC29 was used as the precursor for the generation of pSDC31. A non-tagged Cas7 protein could be produced from this plasmid (pCDF-1b) that is compatible with pSDC25 (MBPCas5).

8.1.1.4 Cas8'

Cas8' was already available (pEB389, Figure 8-6) and was used in a previous study where it was referred to as Nar71 (nucleic acid resolvase, molecular weight 71kDa) in Guy *et al* 2004. Cas8' is the main focus in Chapter 4, biochemical characterisation. As such several site directed mutants were created. Residues of interest were targeted and mutagenized using the Quick Change Site directed mutagenesis procedure. In summary, this involves generating two overlapping primers that have a mono- or di-nucleotide mismatch. Whole plasmid thermoclycing (akin to PCR) is the carried out with a high fidelity polymerase eg. Vent (NEB). Nicked plasmids are created, and original template DNA is removed by DpnI digestion of *E.coli* biosynthesised methylated DNA. Mutants already available from this method are K68A and K117A. Each of these were sub-cloned from the original pT7-7 vector into pET14b using NdeI and ClaI restriction digest, ligation and confirmation for N-terminal His$_6$ tag, to allow purification (Figure 8-7 and Figure 8-8). Other mutants created for this study include D151G (Figure 8-9), N153A, E155A (Figure 8-10), N536A, S540A (Figure 8-11), L542A. two other mutants were attempted, Y548A and L564A but no successful transformations were observed after multiple attempts. In each Case primers were designed and used in the thermocycling reaction eg. D151G-1 and D151G-2 and processed as summarised in chapter 2. Clones were then sequenced and compared to wild-type nucleic acid sequences to identify if the mutation had been introduced.

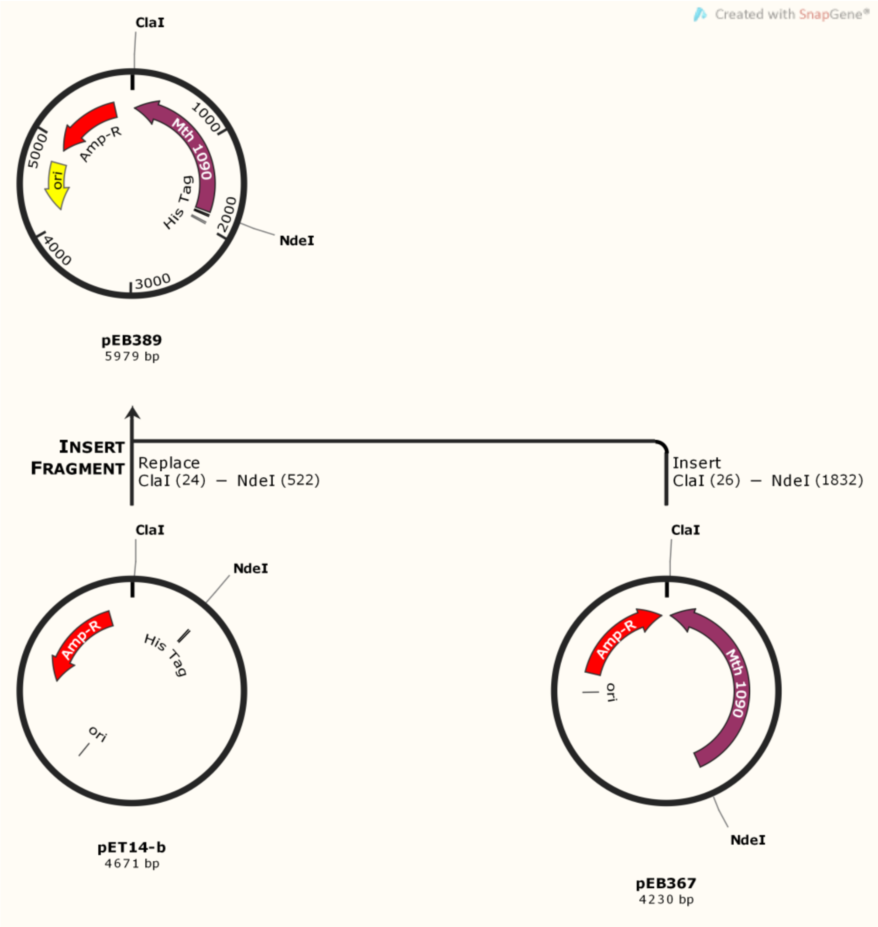


Figure 8-6 The original vectors used for all subsequent Cas8' constructs. pEB367 is the *Mth* Cas8' gene (ORF1090) cloned into the pT7-7 vector. Cas8' was then subcloned into pET14-b to generate a N-terminally hexahistidine tagged Cas8' protein from pEB389.

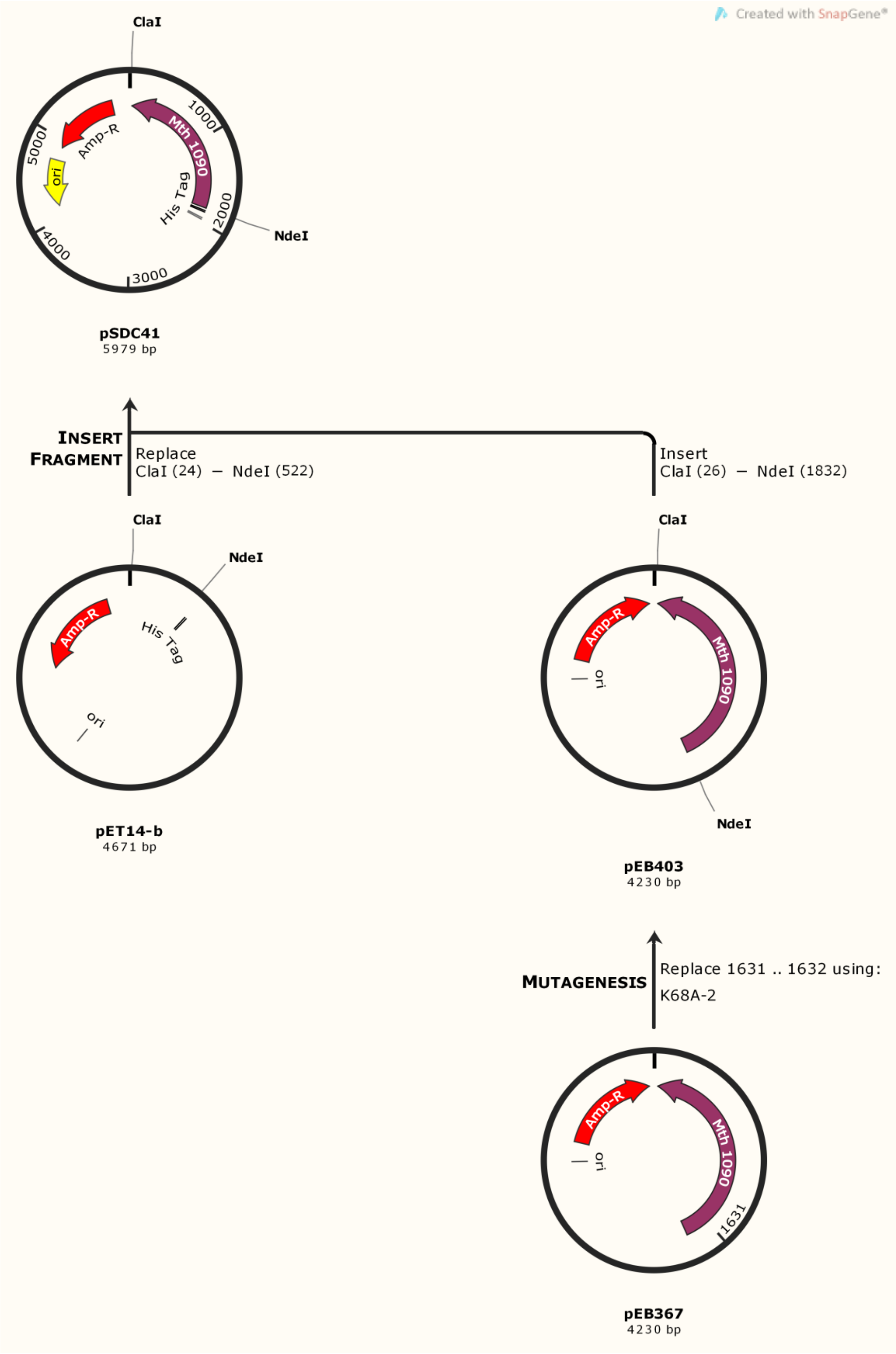


Figure 8-7 Cloning strategy for quick Change K68A mutant Cas8'. Sub cloning procedure followed for generation of an N-terminally hexahistidine tagged Cas8' mutant protein, as with wild-type gene to create pSDC41.

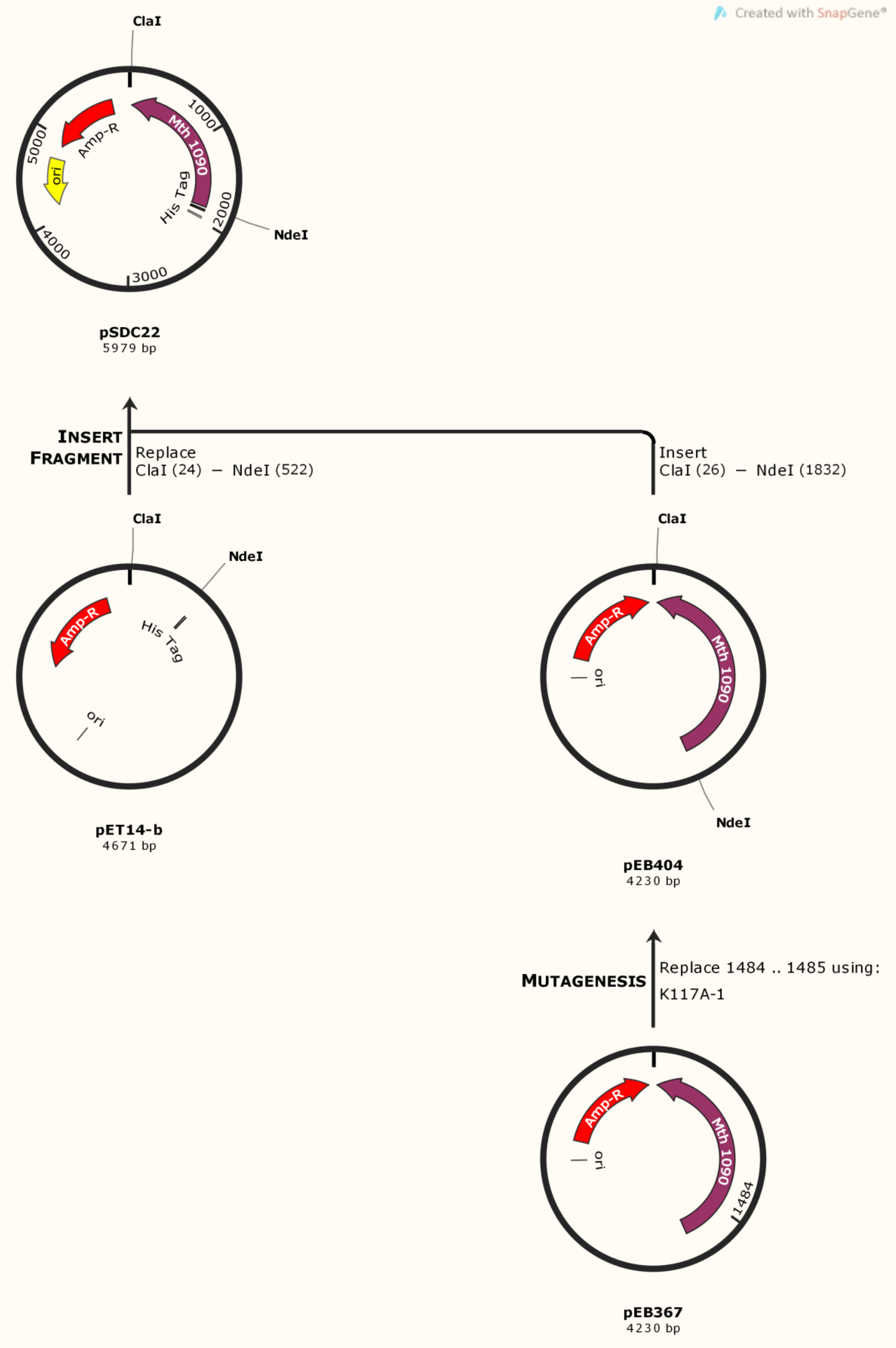


Figure 8-8 Cloning strategy for quick Change K117A mutant Cas8'. Sub cloning procedure followed for generation of an N-terminally hexahistidine tagged Cas8' mutant protein, as with wild-type gene to create pSDC22.

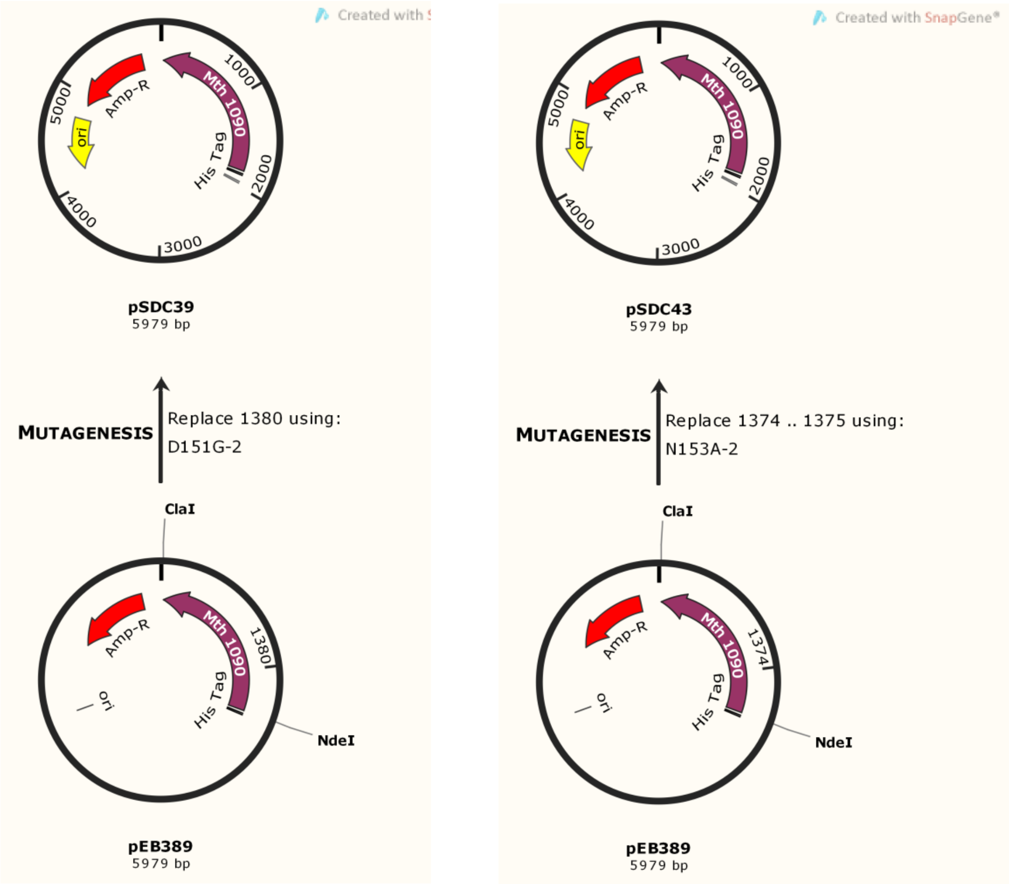


Figure 8-9 Cloning strategy for quick change mutants D151G and N153A of Cas8'. Primers D151G-1 and -2 were used to generate pSDC39, a vector that encodes N-terminally hexahistidine tagged mutant Cas8' protein. Similarly to create pSDC43 primer N153A-1 and -2 were used in the generation of this construct.

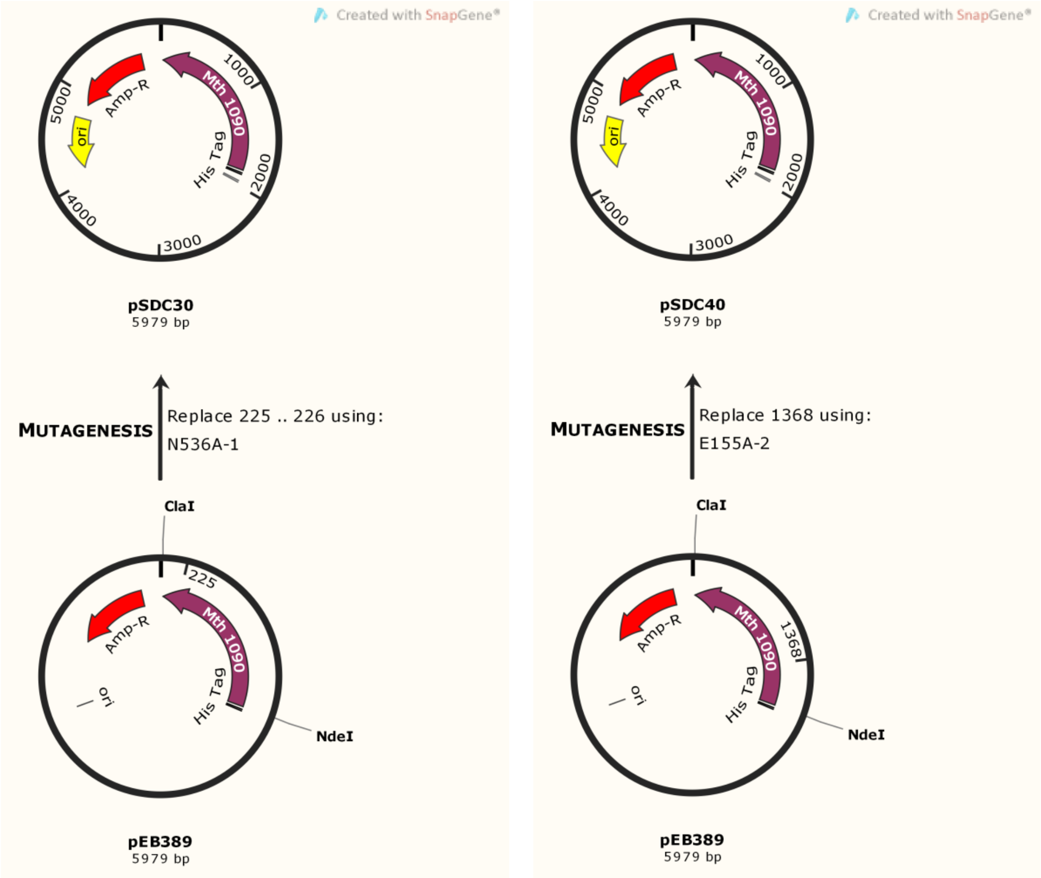
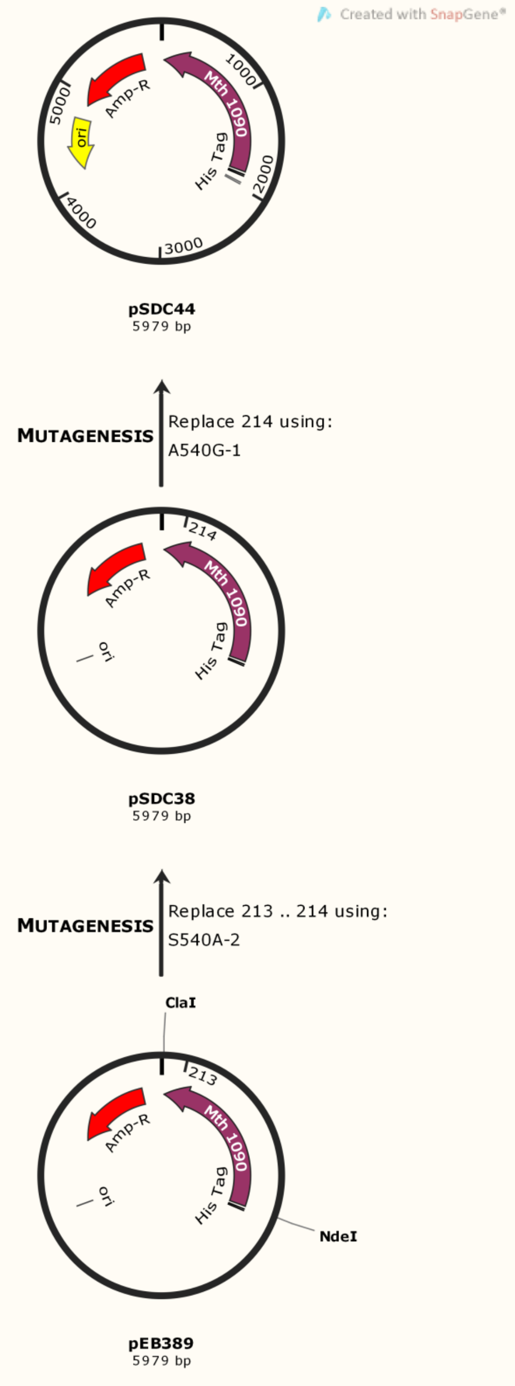


Figure 8-10 Cloning strategy for quick change mutants E155A and N536A of Cas8'. Primers E155A-1 and -2 were used to generate pSDC40, a vector that encodes N-terminally hexahistidine tagged mutant Cas8' protein. Similarly to create pSDC30 primer N536A-1 and -2 were used in the generation of this construct.



Created with SnapGene®
1000
2000
3000
4000
5000
Amp-R
ori
Mth 1090
His Tag
pSDC44
5979 bp
MUTAGENESIS
Replace 214 using:
A540G-1
214
Amp-R
ori
Mth 1090
His Tag
pSDC38
5979 bp
MUTAGENESIS
Replace 213 .. 214 using:
S540A-2
ClaI
213
Amp-R
ori
Mth 1090
His Tag
NdeI
pEB389
5979 bp

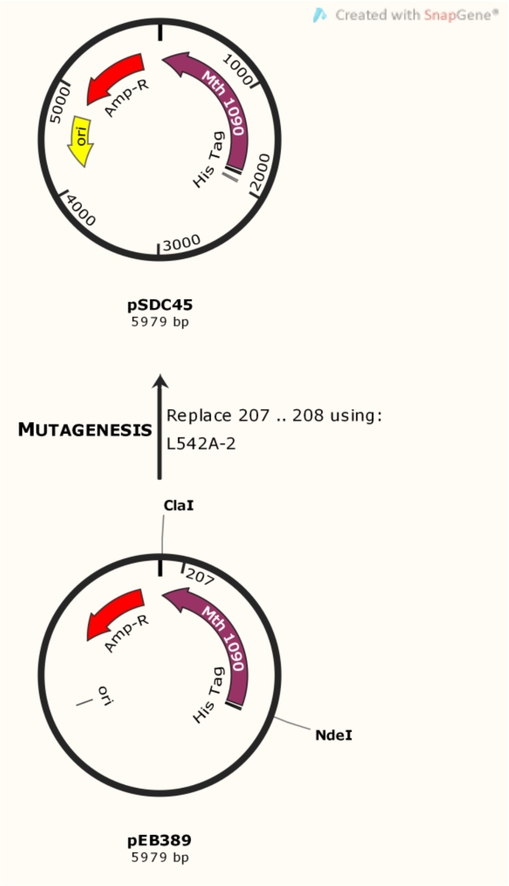


Figure 8-11 Cloning strategy for quick change mutants S540A to A540G and L542A of Cas8'. Primers L542A-1 and -2 were used to generate pSDC45, a vector that encodes N-terminally hexahistidine tagged mutant Cas8' protein. Similarly to create pSDC38 and pSDC44 primer sets S540A-1 and -2 and A540G-1 and -2 were used in the generation of these constructs.

8.1.1.5 Cas8a2

While a clone was already available (pEB383), this was excluded from this study because of previous work and difficulties described later.

8.1.1.6 ORFs 1076-1082

These ORFs form the Csm complex of type IIIA CRISPR immune system, and were cloned also, and pursued as far as solubility screening, but not followed up further in this study. A summary of the cloning can be found in appendix table X.X, detailing constructs and restriction enzymes used. Below is one example of the cloning procedure followed in all circumstances, Figure 8-12.

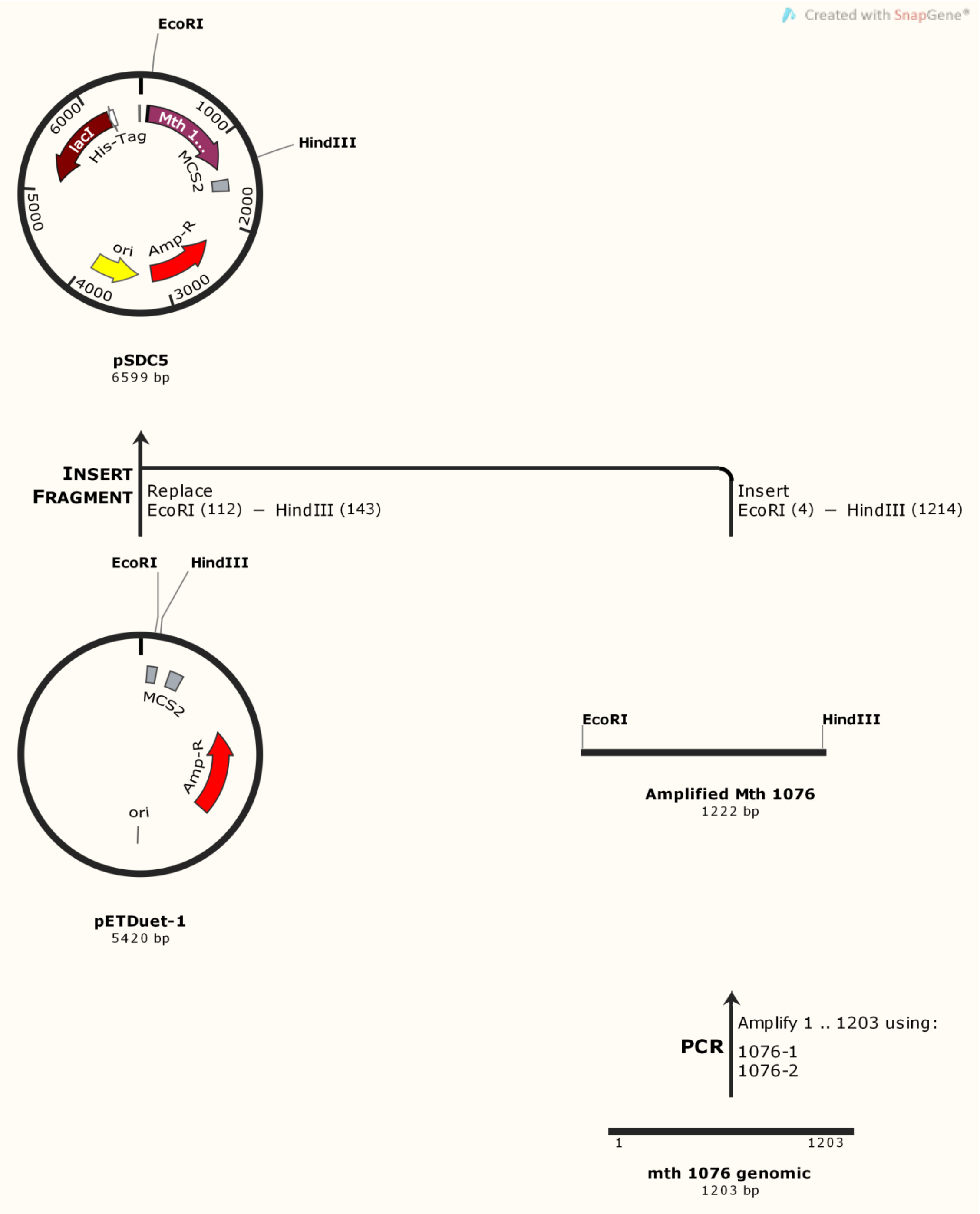


Figure 8-12 Cloning procedure used for construction of pSDC5. *Mth* ORf 1076, encoding Csm1 was PCR amplified with primers 1076-1 and 1076-2 to introduce restriction sites EcoRI and HindIII. This amplified product and pETDuet-1 were digested with the restriction endonucleases EcoRI and HindIII and ligated to produce pSDC5.

8.1.1.7 Cas3

Cas3 while not being directly associated is the essential link to the interference stage of CRISPR immunity in the systems focused on in this study, CASCADE mediated immunity. This was available (pEB359, Figure 8-13) and has been characterised in Howard *et al* 2011.

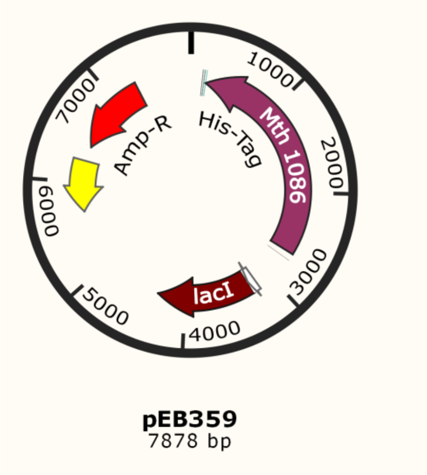


Figure 8-13 Construct available for Cas3. Expression vector for non-tagged Cas3 protein expression.

8.1.1.8 Cas1 and Cas2

Cas1 and Cas2 were no the original focus of this study, but came into use later on when the lab was joined by a second PhD student, Sophie Brinkmann. Then this happened there was a joint effort in generating Cas1 and Cas2 proteins for biochemical analysis *in vitro* of both the *E.coli* and *Mth* proteins combined with genetic analysis *in vivo* in *E.coli*. As such, the cloning was split between the two of us. *Mth* Cas1 and Cas2 were cloned individually and as a continuous operon. Creating pSDC15 (Figure 8-14), 16 (Figure 8-15) and 17 (Figure 8-16) respectively.

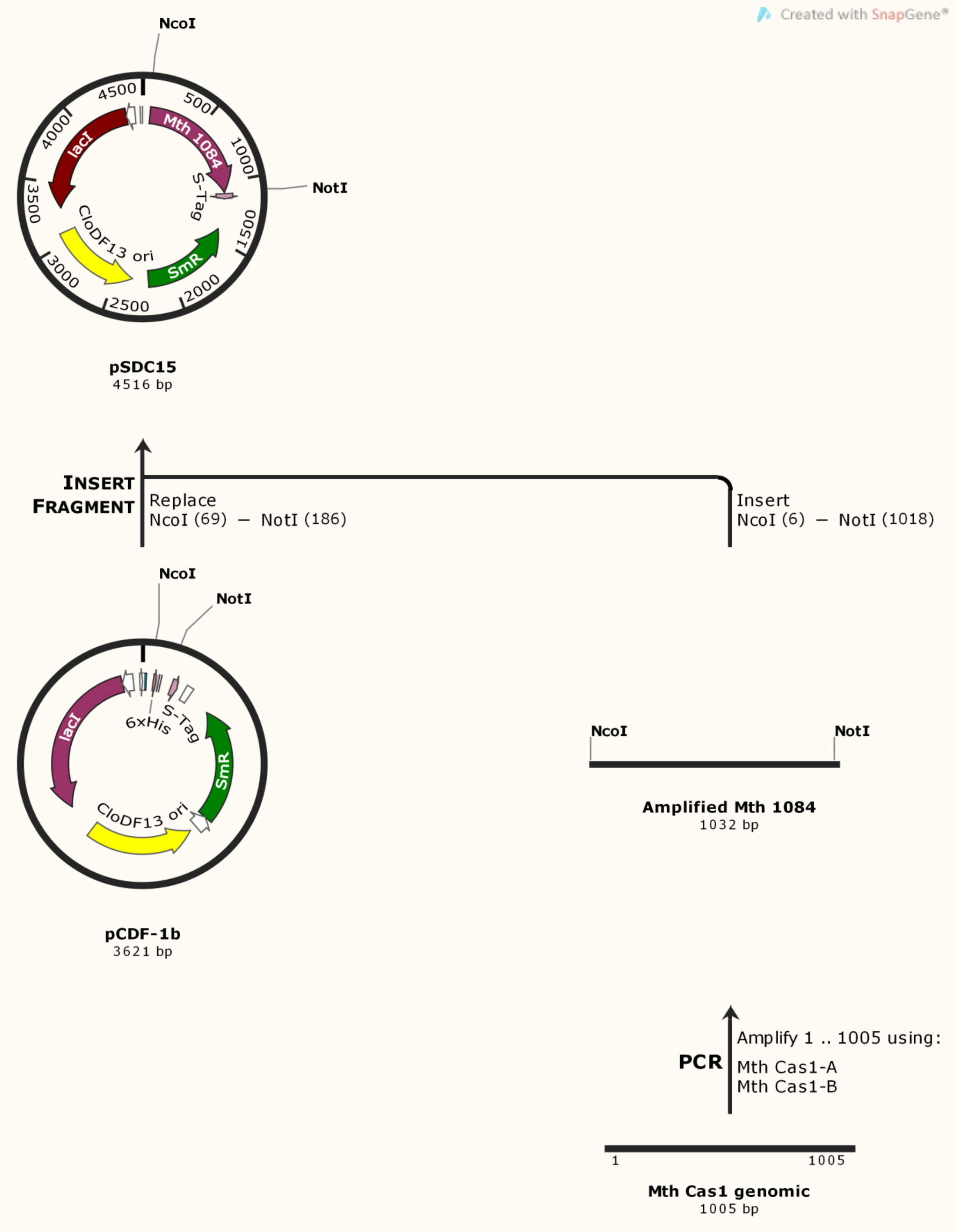


Figure 8-14 Cloning strategy for Cas1 cloning into compatible co expression vector. Cas1 was amplified from genomic DNA using *Mth* Cas1-A and –B primers, introducing NcoI and NotI restriction sites. This allowed restriction cloning into pCDF-1b and creation of pSDC15, N-terminally hexahistidine tagged Cas1.

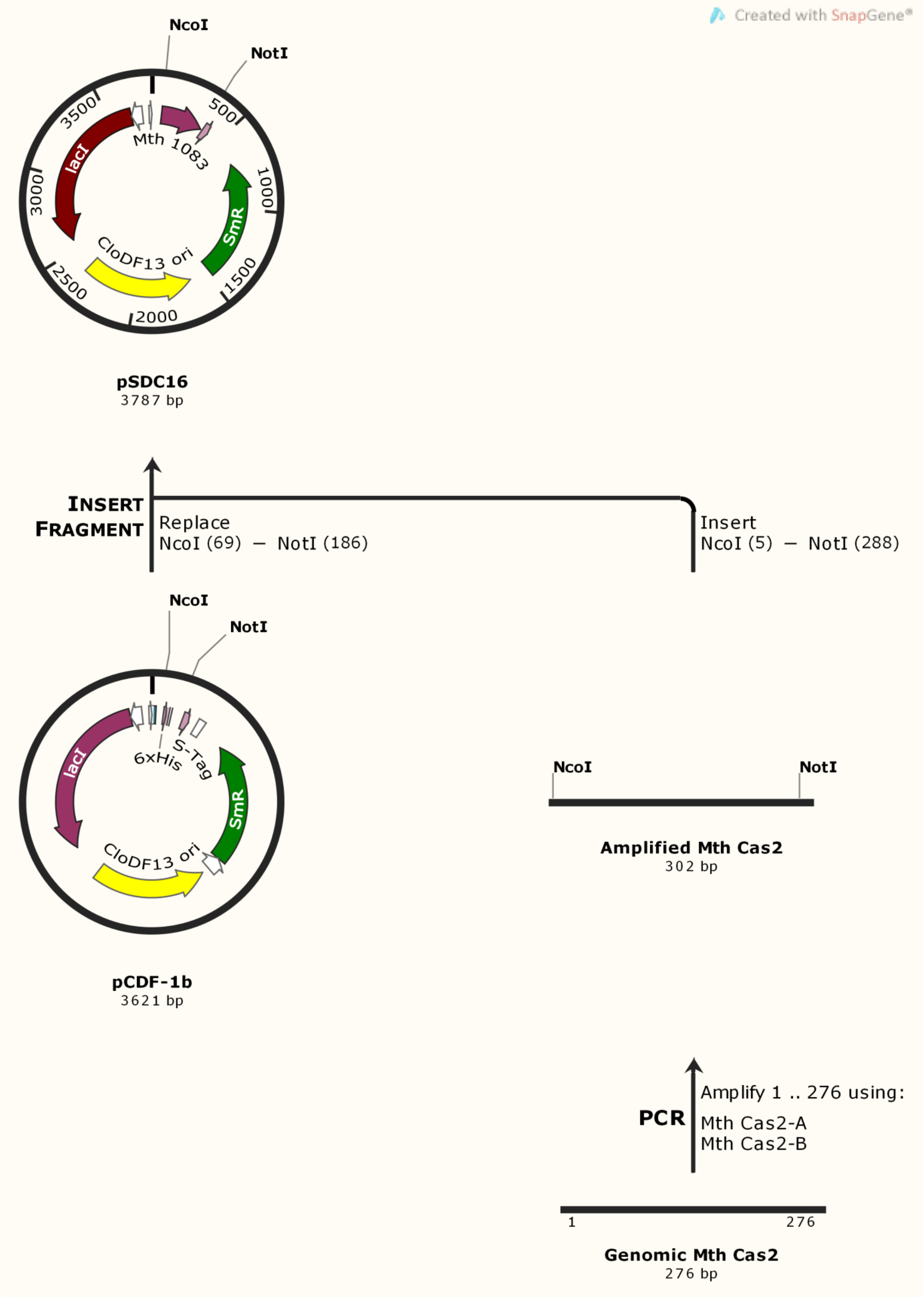
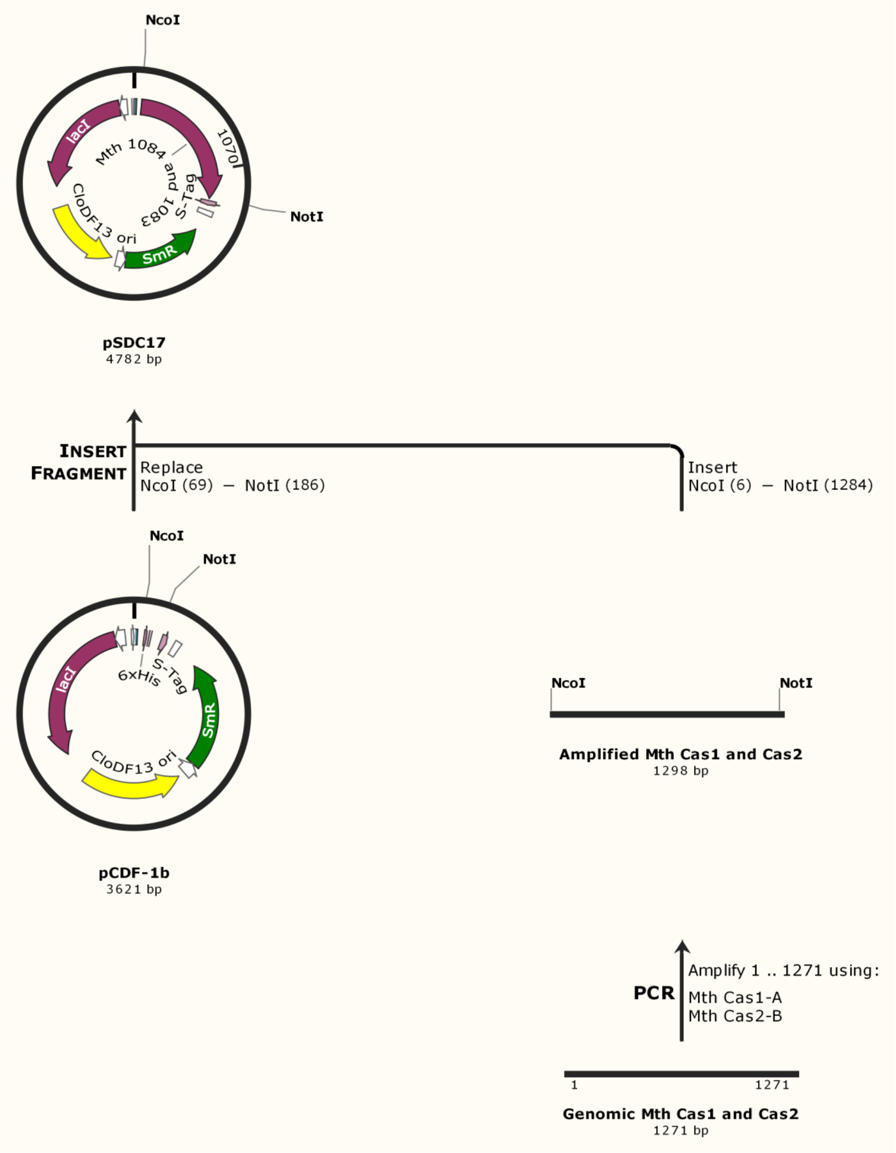


Figure 8-15 Cloning strategy for Cas2 cloning into compatible co expression vector. Cas2 was amplified from genomic DNA using *Mth* Cas2-A and –B primers, introducing NcoI and NotI restriction sites. This allowed restriction cloning into pCDF-1b and creation of pSDC16, N-terminally hexahistidine tagged Cas1.



8.1.1.8.1

Figure 8-16 Cloning strategy for Cas1 and Cas2 cloning for co-expression. Cas1 and Cas2 operon was amplified from genomic DNA using *Mth* Cas1-A and Cas2–B primers, introducing NcoI and NotI restriction sites. This allowed restriction cloning into pCDF-1b and creation of pSDC17, N-terminally hexahistidine tagged Cas1 and non-tagged Cas2.

8.2 Oligonucleotides

8.2.1 The following oligos were used to form substrates shown in Chapter 4.

ssDNA

RGL16: 5’-ATCGATAGTCTCTAGACAGCATGTCCTAGCAAGCCAGAATTCGGCAGCGT-3’

dsDNA: RGL16 annealed to ELB37

ELB37: 5’-ACGCTGCCGAATTCTGGCTTGCTAGGACATGCTGTCTAGAGACTATCGAT-3’

Flayed Duplex: RGL16 partially annealed to ELB20

ELB20: 5’-GACGCTGCCGAATTCTGGCTTGCTATGTAACTCTTTGCCCACGTTGACCC-3’

Partial fork DNA: RGL16 partially annealed to ELB20 and partially annealed to PM2

PM2: 5’-GGACATGCTGTCTAGAGACTATCGAT-3’

Loop duplex DNA: RGL19 annealed to PM4

RGL19: 5’-GACGCTGCCGAATTCTACCAGTGCCTTGCTAGGACATCTTTGCCCACCTGCAGGTTCACCC-3’

PM4: 5’-GGGTGAACCTGCAGGTGGGCGGCTGCTCATCGTAGGTTAGTTGGTAGAATTCGGCAGCGTC-3’

D-Loop: RGL19 annealed to PM4 and PM5

PM5: 5’-AAAGATGTCCTAGCAAGGCAC-3’

R-Loop1: RGL19 annealed to PM4 and ELBRNA1

ELBRNA1: 5’-AAAGAUGUCCUAGCAAGGAC-3’

R-Loop2 substrates were constructed by annealing of the following strands

-PAM (AA) R-loops: ELB103 was annealed to ELB103-B and either crRNA2 (no flap) or crRNA3 (5’handle)

ELB103: 5’-GATAAGCTTAAAATAACATCAACCACCTACAATCCAAATGTGTGGTATGGTTTTTACGGATCCTGG-3’

ELB103-B: 5’-CCA*GGATCC*GTAAAAAAACGCAACACACGGGTTCGGTTAGGTGGTTGATGTTATTTT*AAGCTT*ATC-3’

crRNA2: 5’-CCAUACCACACAUUUGGAUUG-3’

crRNA3: 5’-ATTGAAATCCAUACCACACAUUUGGAUUG-3’

+PAM (CC) R-loops: ELB108 was annealed to ELB108-B and either crRNA2 (no flap) or crRNA3 (5’handle)

ELB108: 5’-GAT*AAGCTT*ACCCTAACATCAACCACCTACAATCCAAATGTGTGGTATGGGGGTTAC*GGATCC*TGG-3’

ELB108-B: 5’-CCA*GGATCC*GTAACCCAACGCAACACACGGGTTCGGTTAGGTGGTTGATGTTAGGGT*AAGCTT*ATC-3’

Other R-loops; PAM TG with ELB101 and ELB101-B, PAM TC with ELB106 and ELB106-B, and either crRNA2 (no flap) or crRNA3 (5’handle)

ELB101: 5’-GATAAGCTTACCTTAACATCAACCACCTACAATCCAAATGTGTGGTATGGCAATTACGGATCCTGG-3’

ELB101-B: 5’-CCAGGATCCGTAATTGAACGCAACACACGGGTTCGGTTAGGTGGTTGATGTTAAGGTAAGCTTATC-3’

ELB106: 5’-GATAAGCTTACTCTAACATCAACCACCTACAATCCAAATGTGTGGTATGGGAGTTACGGATCCTGG-3’

ELB106-B: 5’-CCAGGATCCGTAACTCAACGCAACACACGGGTTCGGTTAGGTGGTTGATGTTAGAGTAAGCTTATC-3’

Partial duplex structures were constructed as above with ELB10X and ELB10X-B for all permutations.

Complete duplex structures constructed as follows:

Duplex 1 (PAM TG): ELB100 annealed to ELB101:

ELB100: 5’-CCAGGATCCGTAATTGCCATACCACACATTTGGATTGTAGGTGGTTGATGTTAAGGTAAGCTTATC-3’

Duplex 2 (PAM – AA): ELB 102 annealed to ELB103:

ELB102: 5’-CCAGGATCCGTAAAAACCATACCACACATTTGGATTGTAGGTGGTTGATGTTATTTTAAGCTTATC-3’

Duplex 3 (PAM TC): ELB 105 annealed to ELB106:

ELB105: 5’-CCAGGATCCGTAACTCCCATACCACACATTTGGATTGTAGGTGGTTGATGTTAGAGTAAGCTTATC-3’

D-Loop (PAMs) substrates were created as above with the addition of crDNA2 annealed.

crDNA2: 5’-GTTTTACCCTAACTTTACCATACCACACATTTGGATTGTAGGTGGTTGATGTTA-3’

The following oligos were used to form 3’ and 5’ flap structures for nuclease activity analysis:

3’ Flaps: Either ELB104 or ELB104RNA were annealed to crRNA1 or crDNA1 for each flap.

ELB104: 5’-TAACATCAACCACCTACAATCCAAATGTGTGGTATGG-3’

ELB104RNA: 5’-UAACAUCAACCACCUACAAUCCAAAUGUGUGGUAUGG-3’ crRNA1: 5’-CCAUACCACACAUUUGGAUUGUAGGUGGUUGAUGUUAAUUUCAAUCCCAUUUUG-3’

crDNA1: 5’-CATACCACACATTTGGATTGTAGGTGGTTGATGTTAATTTCAATCCCATTTTG-3’

5’ Flaps: Either ELB107 was annealed to crRNA1 or crDNA2 for each flap.

ELB107: 5’-CAAAATGGGATTGAAATTAACATCAACCACCTACAAT-3’

8.2.2 The following oligos were used to form substrates shown in Chapter 5.

DNA oligonucleotides for Cas1-Cas2 assays *in vitro*

Oligonucleotides annealed to generate substrates for EMSA and catalytic assays are given below.

Fork-1: Nucleotides forming ssDNA from the fork branch point are underlined, and the branch point cytosine/guanine is in bold. The 5’ ended ssDNA is from oligonucleotide MW14.

‘MW12-mod’ 5’-TCGGATCCTCTAGACAGCTCCAT**G**ATCGTTACATTAGCAGATACTGCAAC-3’

‘MW14’ 5’-CAACGTCATAGACGATTACATTGCTA**C**ATGGAGCTGTCTAGAGGATCCGA-3’

‘PM16’ 5’-TGCCGAATTCTACCAGTGCCAGTGAT-3’

Fork 1a: As Fork-1 but with addition of ‘PM21’ to anneal to ‘MW14’ giving only a 4 nt ssDNA gap of sequence 5’-GCTA.

‘PM21’ 5’-AATGTAATCGTCTATGACGTT-3’

Fork 1b: As Fork-1 but with addition of ‘ELB201’ to anneal to ‘MW14’ giving only a 2 nt ssDNA gap

‘ELB201’ 5’-AATGTAATCGTCTATGACGTT-3’

Fork-2: As fork-1 for ‘MW14’, which is annealed to ‘PM17’ and the 3’ ended ssDNA is from oligo ‘MW12-mod’

‘PM17’ 5’-TAGCAATGTAATCGTCTATGACGTT-3’

‘MW14-mod’ 5’-TCGGATCCTCTAGACAGCTCCATG**A**TCGTTACATTAGCAGATACTGCAAC-3’

Fork-3: As fork-1 with addition of PM17 for full base pairing throughout

Holliday junction:

‘MW14’ and ‘MW12 annealed to;

‘MW13’ 5’-TGCCGAATTCTACCAGTGCCAGTGATGGACATCTTTGCCCACGTTGACCC-3’

‘RGL-14’ 5’-TGGGTCAACGTGGGCAAAGATGTCCTAGCAATGTAATCGTCTATGACGTT-3’

8.3 Bioinformatic analysis of *Methanothermobacter thermautotrophicus* spacer sequences

To determine the PAM sequence for the substrates used in Chapter 5 each spacer from *Mth* was subjected to nBLAST analysis compared to all known nucleotide sequences. Each spacer was compared with 0 to 9 mismatches to any homologous sequences. The positions of the spacer were the correlated to the virus or plasmid DNA. From this positioning data the PAM sequence could be determined flanking the spacer that would be targeted in an interference reaction (as described in Chapters 1,3 and 4). These PAM sequences were then compiled (Table 20) and represented graphically in Chapter 4, Figure 4-2.

Table 20. Bioinformatic analysis of *Mth* spacer sequences. Matched spacer and up to 9 mismatched sequences compared to whole known nucleotide sequence database. Genomic location of each match was determined to identify PAM sequence as summarised in Figure 4-2.

Number of mismatches	Spacer length	Spacer name	Virus/Plasmid	S.posit	E.posit
0	36	AE000666_C1_spacer22	psiM2	3480	3515
0	36	AE000666_C1_spacer28	psiM2	3480	3515
0	36	AE000666_C1_spacer74	psiM100	3831	3866
0	36	AE000666_C1_spacer8	psiM100	24022	24057
0	37	AE000666_C1_spacer24	psiM2	22580	22616
0	37	CP001710_C1_spacer17	psiM100	4444	4480
0	38	AE000666_C1_spacer6	psiM100	3611	3648
1	37	AE000666_C1_spacer110	psiM2	12197	12233
1	38	AE000666_C1_spacer45	psiM2	3835	3872
1	38	AE000666_C1_spacer7	psiM100	3430	3467
2	36	AE000666_C1_spacer40	psiM2	3536	3570
4	37	AE000666_C1_spacer122	psiM100	4025	4060
5	29	CP000102_C3_Spacer61	pFZ1	2039	2062
5	32	AE000666_C1_spacer100	psiM2	2616	2647
5	37	AE000666_C1_spacer5	pME2001	330	357
5	37	AE000666_C1_spacer5	pME2200	322	349
5	37	AE000666_C1_spacer5	pMTBMA4	24	51

6	36	CP000678_C1_spacer25	pME2200	2149	2183
6	36	CP001719_C5_spacer11	psiM2	9055	9088
6	37	CP000102_C2_spacer10	pFZ1	7780	7815
6	38	CP000102_C2_spacer42	pFZ1	8184	8221
7	29	CP000102_C3_Spacer61	pFV1	4378	4402
7	29	CP000102_C3_Spacer61	pFV1	3218	3246
7	34	CP001719_C5_spacer2	psiM2	23907	23938
7	35	CP001710_C1_spacer26	pFV1	8082	8114
7	35	CP001719_C4_spacer1	pFZ1	9540	9573
7	35	CP001719_C5_spacer20	psiM2	25364	25393
7	35	CP002772_C2_spacer69	psiM100	30506	30538
7	36	AE000666_C1_spacer36	psiM100	1128	1164
7	36	CP000102_C1_spacer7	psiM2	19096	19130
7	36	CP000102_C2_spacer33	psiM100	7063	7094
7	36	CP000678_C1_spacer4	pFZ1	8598	8633
7	36	CP001719_C3_spacer46	pFV1	12655	12680
7	36	CP002772_C2_spacer9	psiM2	2509	2543
7	37	AE000666_C1_spacer42	psiM2	25855	25888
7	37	CP000102_C1_spacer16	psiM2	25944	25980
7	37	CP000102_C2_spacer55	pFZ1	7788	7823
8	25	CP001710_C2_spacer1	psiM100	1083	1106
8	25	CP001710_C2_spacer1	pME2001	843	864
8	25	CP001710_C2_spacer1	pME2200	835	856
8	25	CP001710_C2_spacer1	pMTBMA4	537	558
8	30	CP000102_C2_spacer7	pFZ1	10716	10739
8	32	CP001719_C3_spacer13	pFV1	10037	10064
8	34	AE000666_C1_spacer101	psiM100	2821	2852
8	34	AE000666_C2_spacer44	pFV1	3100	3129
8	34	CP000678_C1_spacer9	psiM2	16271	16300
8	35	AE000666_C1_spacer98	pFV1	1559	1587
8	35	CP000678_C1_spacer28	psiM100	20935	20967
8	35	CP001719_C3_spacer51	psiM100	14402	14434
8	35	CP001719_C5_spacer20	pFV1	13214	13246
8	35	CP001719_C5_spacer20	pFZ1	10714	10746
8	35	CP002772_C2_spacer72	pFZ1	8908	8938
8	36	AE000666_C1_spacer34	pME2001	3736	3767
8	36	AE000666_C1_spacer34	pME2200	4861	4892
8	36	AE000666_C1_spacer34	pMTBMA4	3431	3462
8	36	AE000666_C1_spacer44	psiM100	28400	28435
8	36	CP000102_C1_spacer39	pFZ1	10441	10477
8	36	CP000678_C1_spacer34	psiM100	29236	29268
8	36	CP001719_C5_spacer27	psiM100	3164	3193
8	37	AE000666_C1_spacer47	pME2001	3727	3756
8	37	AE000666_C1_spacer47	pME2200	4852	4881
8	37	AE000666_C1_spacer47	pMTBMA4	3422	3451
8	37	AE000666_C2_spacer42	pME2200	4768	4804

8	37	CP000102_C1_spacer12	psiM100	7555	7591
8	37	CP000102_C1_spacer37	pFV1	10838	10875
9	25	CP001710_C2_spacer1	psiM2	12810	12831
9	27	CP001710_C2_spacer3	psiM100	13089	13114
9	29	CP000102_C3_Spacer61	psiM100	29225	29252
9	32	AE000666_C1_spacer100	psiM100	19199	19231
9	32	AE000666_C1_spacer100	pFZ1	9203	9229
9	34	AE000666_C1_spacer101	pFZ1	9597	9628
9	34	AE000666_C1_spacer91	pFV1	559	585
9	34	AE000666_C1_spacer91	pFZ1	559	585
9	34	AE000666_C2_spacer41	psiM2	12439	12470
9	35	AE000666_C1_spacer9	psiM100	24849	24883
9	35	AE000666_C1_spacer9	psiM2	18429	18463
9	35	AE000666_C1_spacer98	pFZ1	2149	2177
9	35	AE000666_C2_spacer19	psiM100	18961	18992
9	35	CP000678_C1_spacer32	psiM2	12588	12614
9	35	CP001719_C4_spacer10	psiM100	15477	15507
9	35	CP001719_C4_spacer1	pFV1	4291	4323
9	35	CP001719_C5_spacer14	pME2001	3578	3608
9	35	CP001719_C5_spacer14	pME2200	3562	3592
9	35	CP001719_C5_spacer14	pMTBMA4	3273	3303
9	35	CP001719_C5_spacer14	pFV1	3111	3142
9	36	AE000666_C1_spacer116	psiM2	234	269
9	36	AE000666_C1_spacer93	psiM100	13020	13048
9	36	CP000102_C1_spacer17	psiM100	2817	2852
9	36	CP000102_C2_spacer24	psiM100	8485	8514
9	36	CP000102_C2_spacer54	psiM100	16338	16376
9	36	CP000678_C1_spacer13	psiM100	7822	7854
9	36	CP000678_C1_spacer22	pFZ1	10048	10084
9	36	CP001719_C3_spacer19	pFZ1	2022	2056
9	36	CP001719_C3_spacer23	pFZ1	7402	7433
9	36	CP001719_C3_spacer44	psiM100	9509	9540
9	37	AE000666_C1_spacer103	pME2200	45	80
9	37	AE000666_C1_spacer112	psiM100	3190	3227
9	37	AE000666_C1_spacer114	pME2001	3736	3770
9	37	AE000666_C1_spacer114	pME2200	4861	4895
9	37	AE000666_C1_spacer114	pMTBMA4	3431	3465
9	37	AE000666_C1_spacer13	pFV1	8768	8800
9	37	AE000666_C1_spacer13	pFV1	12614	12640
9	37	AE000666_C1_spacer56	pFZ1	9958	9990
9	37	AE000666_C1_spacer62	pFZ1	9958	9990
9	37	AE000666_C2_spacer21	pFV1	10862	10895
9	37	AE000666_C2_spacer26	pFV1	10862	10895
9	37	CP000102_C1_spacer41	psiM100	29498	29535
9	37	CP000102_C1_spacer43	psiM100	29498	29535
9	37	CP000102_C1_spacer9	psiM100	1495	1531
9	37	CP000102_C2_spacer10	psiM2	23608	23641
9	37	CP000102_C2_spacer34	pFZ1	9582	9613
9	38	AE000666_C1_spacer18	psiM2	6326	6359
9	38	AE000666_C2_spacer43	psiM100	3180	3217
9	39	AE000666_C2_spacer5	pFV1	11488	11526