Investigation of CRISPR-Cas Adaptation Mechanism Through Exploration of Cas4-1 Fusions

A thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy

Emily Walker, Bsc (Hons), MRes

September 2018

Abstract

CRISPR-Cas is an adaptive immune system present in bacteria and archaea. It involves two linked stages: adaptation and interference. Adaptation generates a database of mobile genetic element sequences within a CRISPR locus through 'capture' of DNA fragments by Cas1-Cas2 and integration into the CRISPR locus. The CRISPR locus is transcribed and separated into individual units, called crRNA. During interference crRNA bound within 'interference' ribonucleoprotein complexes targets DNA from Mobile genetic elements (MGE) through complementary base pairing between crRNA and the MGE before degradation of the target.

Adaptation is catalysed by Cas1 and Cas2 proteins in three contexts: naïve, targeted and primed adaptation. The pathway taken is dependent on whether the organism has encountered the MGE previously. A crucial stage of adaptation is the generation of fragments for 'capture', however the mechanism is unknown. In bacteria several proteins have been hypothesised to be involved in supporting adaptation including RecB, RecG, PriA and Cas4.

Cas4-1 fusions are a naturally occurring fusion of Cas4 and Cas1. Cas4-1 proteins from *Methanosaeta harundinacea* and *Pyrinomonas methylaliphatogenes* were selected for investigating the role of Cas4-1 in adaptation and to establish a single *in vitro* reaction for naïve adaptation. *M. harundinacea* Cas4-1 did not produce tractable protein, however a high yield of active Cas4-1 was obtained from *P. methylaliphatogenes* along with potential interacting partners (Cas2 and HPS). Cas4-1 was shown to contain active sites from both Cas1 and Cas4 through sequence alignment and *in vitro* biochemistry demonstrated Cas4-1 existed as a dimer in solution. Cas4-1 bound DNA with ssDNA regions and cleavage both ssDNA and dsDNA. No physical interaction was observed between Cas4-1 and Cas2 or HPS, but Cas2 was shown to be activator of Cas4-1 nuclease activity. This work demonstrates the activities of Cas4-1 and proposes a role for the protein in adaptation.

Table of Contents

List of figures

List of Tables

List of abbreviations

AGF	Analytical gel filtration
AI	Arabinose induced
ATP	Adenosine triphosphate
BN-PAGE	Blue native polyacrylamide gel electrophoresis
BSA	Bovine serum albumin
C+	Codon plus
Cas	CRISPR-associated proteins
Cascade	CRISPR-associated complex for antiviral defence
CIP	Calf intestinal phosphatase
Cmr	CRISPR RAMP module
CRISPR	Clustered regularly interspaced short palindromic repeats
crRNA	CRISPR RNA
Csy	CRISPR RNA guided surveillance complex
CV	Column volume
D-loop	Deoxyribonucleic acid loop
DNA	Deoxyribonucleic acid
ds	Double-stranded
DSBR	Double-strand break repair
EDTA	Ethylenediaminetetraacetic acid
EM	Electron microscopy
EMSA	Electrophoretic mobility shift assay
ERIC	Enterobacterial repetitive intergenic consensus
HJ	Holliday junction
H-NS	Heat stable nucleotide structuring protein
HPL	Hypothetical protein large
HPS	Hypothetical protein small
HR	Homologous recombination
HRP	Horseradish peroxidase
HTH	Helix-turn-helix
IAS	Integrase anchoring site
IBS	Integration host factor binding site
IHF	Integration host factor
IPTG	Isopropyl β-D-I-thiogalactopyranoside

K_{av}	Gel phase distribution coefficient
KEGG	Kyoto encyclopaedia of genes and genomes
LB	Luria broth
MGE	Mobile genetic element
Mha	*Methanosaeta harundinacea*
nt	Nucleotide
ORF	Open reading frame
PAM	Protospacer adjacent motif
PAS	Primosome assembly site
PCR	Polymerase chain reaction
Pme	*Pyrinomonas methylaliphatogenes*
PMSF	Phenylmethylsulfonyl fluoride
Phyre2	Protein homology/analogy recognition enzyme version 2
RE	Restriction enzyme
REP	Repetitive extragenic palindromic sequence
R-loop	Ribonucleic acid loop
RNA	Ribonucleic acid
RNAi	RNA interference
RT	Room temperature
SAM	Spacer acquisition motif
SDM	Site directed mutagenesis
SDS	Sodium dodecyl sulphate
SDS-PAGE	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
SDW	Sterile distilled water
ss	Single-stranded
SSB	Single-stranded DNA binding protein
TBE	Trisaminomethane borate ethylenediaminetetraacetic acid
TBS	Tris buffered saline
TBST	Tris buffered saline and Tween 20
TCEP	Tris(2-carboxyethyl)phosphine
TES	Transesterification
TG	Tris glycine
TIM	Target interference motif
TIR	Terminal inverted repeats
tracrRNA	Trans-activating RNA

TSD	Target site duplication
UniProt	Universal protein resource
UV	Ultraviolet
WT	Wildtype

1. Introduction

1.1 Initial discovery of CRISPR-Cas

CRISPR (clustered regularly interspaced short palindromic repeats) were first discovered during the analysis of the *iap* gene of *Escherichia coli*. A motif at the 3' end of the gene contained five 29nt (nucleotide) palindromic repeats spaced by non-homologous sequences of 35nt (Ishino *et al.*, 1987). This series of repeats was analogous, but not identical to repeat families studied at the time such as Rep (repetitive extragenic palindromic) sequence and ERIC (enterobacterial repetitive intergenic consensus) sequence (Stern *et al.*, 1984; Hulton *et al*, 1991). Both these repeats contain palindromic regions that form stem-loop structures. Rep is a 35nt sequence that can cluster as inverted repeats, whereas ERIC is 126nt sequence with 14 conserved regions that form inverted repeats. Therefore, both repeats can contain multiple stem-loop structures. These repeats were located within non-coding transcribed regions usually at the 5' or 3' end of an open reading frame (ORF) or operon (Higgins *et al.*, 1982; Gilson *et al.*, 1984; Sharples and Lloyd, 1990; Hulton *et al*, 1991; Lupski and Weinstock, 1992). The CRISPR repeats were analogous to these repeats due to their palindromic nature, ability to form stem-loop structures and location at the 3' end of the *iap* gene. On the other hand, the repeats were spaced with a consistent length of sequence that had not been observed for other repeats and the repeats were not inverted in relation to the adjacent repeats. These differences prevented CRISPR from being included in any repeat family, and potentially established CRISPR as part of its own unique family.

CRISPR was discovered in a number of diverse organisms including *Mycobacterium tuberculosis*, *Mycobacterium bovis, Streptococcus, Anabaena* and *Haloferax mediterranei* (Hermans *et al.*, 1991; Groenen *et al.*, 1993; Mojica *et al* 1993; Mojica *et al.*, 1995; Masepohl *et al* 1996; Hoe *et al.*, 1999). The presence of CRISPR across a diverse set of organisms was suggestive of a function. This was compounded by the detection of four ORFs adjacent to the CRISPR locus suggesting a combined function between the repeats and adjacent proteins. The four genes named *cas1, cas2, cas3* and *cas4*, were separated from the CRISPR locus by around 200nt of sequence (Figure 1). This 200nt sequence was later established as a promoter called the leader sequence. The *cas1* gene was associated with all CRISPR

loci examined in the research (Jansen *et al.*, 2002) which suggested a linked function between *cas1* and CRISPR.

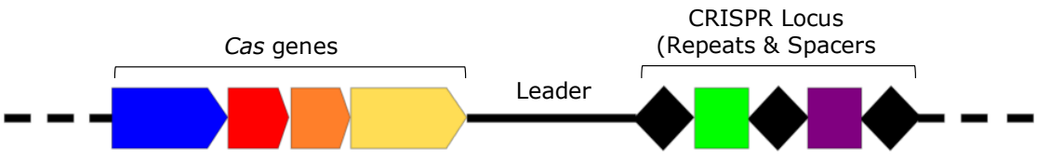

Figure 1: Example of CRISPR locus with neighbouring *cas* ORFs. The CRISPR locus contains repeats 'spaced' by unique sequences. The ORFs for the *cas* genes are separated from the CRISPR locus by around 200nt of sequence, and can be found upstream or downstream of the CRISPR locus. The make-up of each CRISPR loci and *cas* genes differs across organisms.

Spacer sequences were found to match phage genomes and plasmid sequences (Bolotin *et al.*, 2005) and examination of strains resistance to phage showed that the number of phage matching spacers negatively correlated to phage sensitivity (Bolotin *et al.*, 2005; Barrangou *et al.*, 2007). When an organism contained a spacer that matched a phage sequence, the phage would be less likely to infect this organism. The more spacers available that matched the phage genome the greater the reduction in infection of that organism. The removal of spacers removed phage resistance (Barrangou *et al.*, 2007), so the phage resistance was dependent on matching spacer sequence.

The CRISPR locus was shown to be transcribed after northern blots of extracted RNA showed RNA complementary to the repeat sequences (Tang *et al.*, 2002, 2005). The CRISPR transcript was called pre-CRISPR RNA (pre-crRNA) and was a transcript of the whole CRISPR locus. An initial hypothesis after this discovery was that CRISPR-Cas may function in a RNA interference (RNAi) like mechanism. The premise was that the crRNA would bind to the complementary RNA sequences preventing translation (Bolotin *et al.*, 2005; Makarova *et al.*, 2006). However, the crRNA were complementary to both coding and non-coding regions, meaning that many targets would not be present as RNA. The true mechanism also involved complementary binding, but the target was DNA.

Research in *E. coli* discovered a multi-protein complex called Cascade (CRISPR-associated complex for antiviral defence) consisting of five proteins in various oligomeric states that was shown to process pre-crRNA into single crRNA units. This processing involved the Cas6e subunit of Cascade and the phage resistance

provided by the system was dependent on Cas3 (Brouns *et al.* 2008). This research led to the establishment of a mechanism for CRISPR-Cas function (van der Oost *et al.*, 2009).

1.2 The CRISPR-Cas mechanism

CRISPR-Cas has been found in 87% of archaea and 45% of bacteria studied (Grissa *et al*, 2007). There is a vast diversity in the components (e.g. *cas* genes) of each system (see section 1.4 for more details), but there is a general consensus that the mechanism contains two linked stages: Adaptation and Interference (Figure 2). Adaptation is the 'capture' of MGE DNA fragments by Cas1-Cas2 and integration of the new spacers into the CRISPR locus (Barrangou *et al.*, 2007). As the spacers are inserted into the genome they can be passed on to the next generation creating a heritable system. In interference, following CRISPR locus transcription, crRNA in combination with a Cas nuclease protein targets MGEs and degrades them (van der Oost *et al.*, 2009).

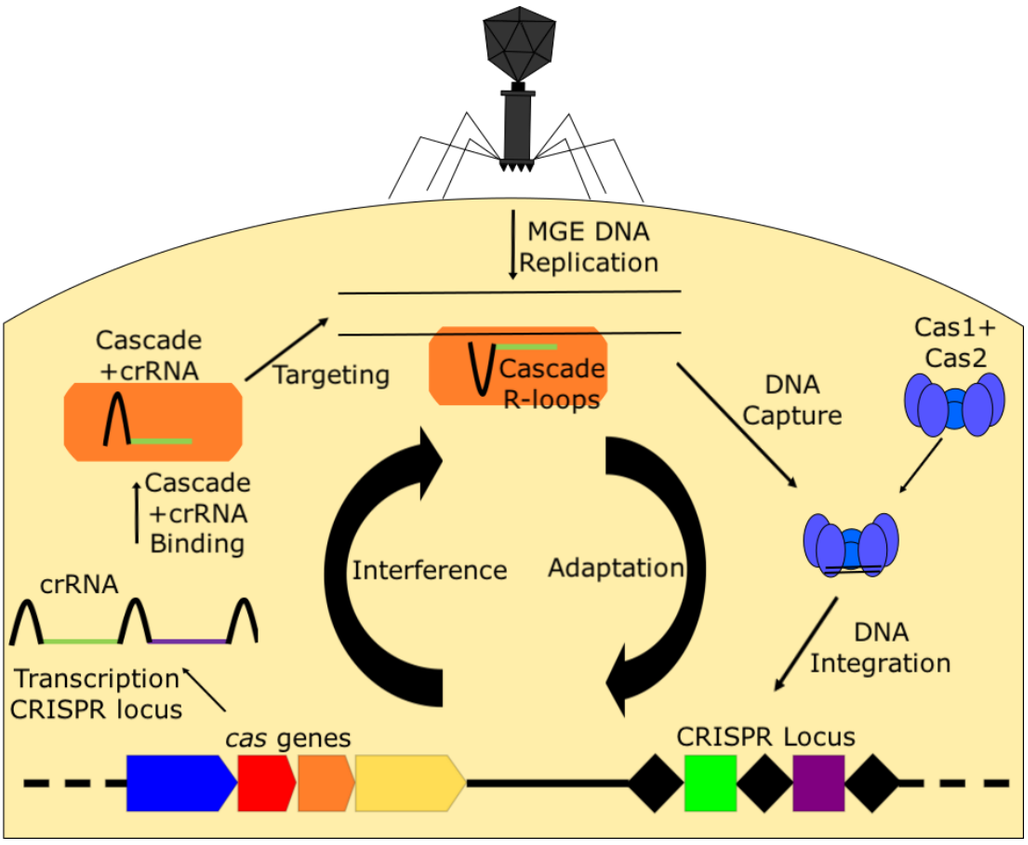

Figure 2: CRISPR-Cas immunity summary mechanism. (1) Adaptation: capture of MGE DNA fragments by the Cas1-Cas2 complex followed by integration into the CRISPR locus. This integration generates a database of MGE DNA sequences. (2) Interference requires the transcription and processing of the CRISPR locus to create singular crRNA units comprising one spacer and one repeat sequence. These are combined with a CRISPR-associated nuclease (Cascade complex from *E. coli* is given as an example). Cascade with bound crRNA binds to complementary MGE DNA via an R-loop. This allows targeted degradation by Cas3 at that location.

Adaptation is the capture of MGE DNA fragments, known as protospacers, which requires Cas1 with Cas2 or an analogous protein (Arslan *et al.*, 2014; Ivancic-Bace *et al.*, 2015). In *E. coli* capture requires a Cas1-Cas2 complex with a stoichiometry $Cas1_4$-$Cas2_2$. A crystal structure with a bound spacer is shown in Figure 3 (Nuñez *et al.*, 2014). Cas1 is required for capture, but the complexes involved in capture differ dependent on the organism, for example in *Sulfolobus solfataricus* the stoichiometry is $Cas1_2$-Cas2-3_4 (Fagerlund *et al.*, 2017). As *E. coli* is the most researched model of CRISPR-Cas, the information included in this section will focus on that model. It is important to note that each CRISPR-Cas system is slightly different and will differ from the *E. coli* model.

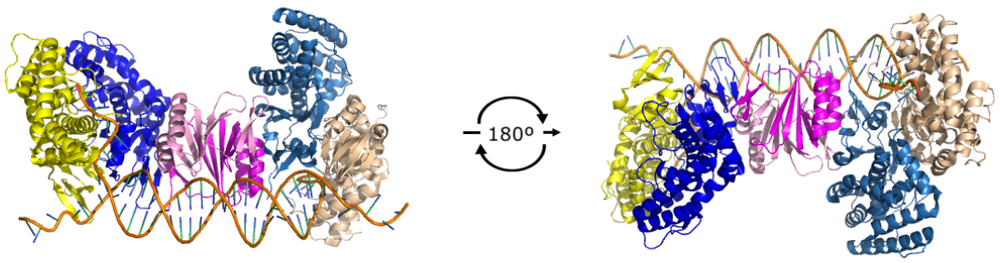

Figure 3: Crystal structure of *E. coli* Cas1-Cas2 complex bound to dual forked DNA. The *E. coli* Cas1-Cas2 complex contains a Cas2 dimer (pinks) sandwiched between two Cas1 dimers (blues and yellows). This complex was crystallised whilst bound to a dual forked DNA containing 23nt of double-stranded DNA with 6nt 5' overhangs and 10nt 3' overhangs. The DNA is bound across the flat region of the complex, with the 3' overhangs descending into the top Cas1 molecules (yellows).

Cas1 is an integrase which can nick DNA and exists as a homodimer in solution. Cas1 nuclease activity is used for integration of spacers into the CRISPR locus not for cleavage of the MGE DNA to generate protospacers. (Babu *et al.*, 2011; Jore *et al.*, 2011; Kim *et al.*, 2013). Cas2 can possess RNase activity against ssRNA shown in *S. solfataricus* and *Archaeoglobus fulgidus* (Beloglazova *et al.*, 2008). However, this RNase activity is not universal and many Cas2 proteins lack active site residues required (Samai *et al*, 2010). The role of this RNase activity is unknown, but due to the lack of activity in some organism in cannot be essential to all systems. Though the involvement of the RNase activity of Cas2 is not resolved, Cas2 has been shown to bind protospacer DNA when complexed with Cas1 (Wang *et al.*, 2015; Fagerlund *et al.*, 2017) but the residues involved in binding vary between organisms. Cas1-Cas2 complex is required for adaptation, but it cannot generate protospacers. However, there is a consensus that DNA replication and DNA repair proteins are important for adaptation, potentially to generate DNA fragments for integration (Ivancic-Bace *et al.*, 2015; Levy *et al.*, 2015; Killelea and Bolt, 2017; Liu *et al.*, 2017).

Sequence alignment of the protospacers and surrounded sequences from the original MGEs revealed a 2-5nt sequence consistently adjacent to all protospacers. This sequence was called the PAM (protospacer adjacent motif). There are two different types of PAM the spacer acquisition motif (SAM) used for identifying protospacers for acquisition and the target interference motif (TIM) used for identifying complementary MGE sequences for degradation which will be discussed

later (Shah *et al.*, 2013). SAM sequences differ across organisms (Mojica *et al.*, 2009; Datsenko *et al.*, 2012; Díez-Villaseñor *et al.*, 2013) and cleavage occurs adjacent to or within the SAM. When cleavage occurs within the SAM, part of the SAM is incorporated into the spacer (Swarts *et al.*, 2012). SAMs are distributed throughout MGEs, but hotspots for spacer acquisition have been found at sites of homologous recombination (HR), ds (double-strand) breaks and at Ter sites (Levy *et al.*, 2015; Shiimori *et al.*, 2017). Protospacers obtained from sites of HR and ds breaks are acquired from the free DNA ends generated through these processes. This represents a potential self/non-self-mechanism as genomic DNA is generally maintained in a circular form, whereas bacteriophage DNA is transferred into cells in a linear form with free DNA ends for targeting. Another protospacer hotspot was Ter sites. Ter sites are replication terminus sites where replication forks meet from opposing directions. The approaching replication forks are stalled at ter sites to allow replication to be terminated simultaneously. The stalled replication fork is the target of acquisition as opposed to the Ter site as inducement of a replication stall site creates an artificial hotspot. The protospacer hotspot at the Ter site was bordered on one end by the stalled replication fork site and by a Chi site at the other (Levy *et al.*, 2015).

The significance of chi sites bordering the protospacer hotspot, is the interaction between Chi and RecBCD. RecBCD is involved in the HR pathway and has been implicated in CRISPR-Cas (see section 1.3.1). RecBCD will be discussed in more detail later, but in brief RecBCD binds to ds breaks in DNA and cleaves both DNA strands until it reaches a Chi site. At the Chi site, the activity for RecBCD is altered preventing cleavage at the 3’ ended DNA strand creating a long 3’ ssDNA tail. This 3’ ssDNA tail can be used for HR (Bianco and Kowalczykowski, 1997; Spies *et al.*, 2003; Wong *et al.*, 2006). The presence of Chi at the border of the protospacer hotspot suggests that RecBCD degradation to the Chi site may form DNA fragments for ‘capture’ by Cas1-Cas2. This hypothesis is also supported by evidence that inducing a ds break creates an artificial hotspot and RecBCD null cells have decreased acquisition (Levy *et al.*, 2015). As will be discussed later, despite research into the link between RecBCD and CRISPR-Cas the role of RecBCD in CRISPR-Cas is still not fully understood.

Once a protospacer is generated it is bound by the Cas1-Cas2 complex and processed to create a spacer of the correct length and with overhangs to aid insertion into the CRISPR locus. This cleavage can be carried out by Cas1 (Wang *et al.*, 2015). As shown in Figure 4 in order to integrate the new spacer the CRISPR locus is nicked by Cas1 at the leader-repeat junction before Cas1 catalyses a transesterification reaction joining the 3′ end of the spacer to the 5′ end of the repeat (Arslan *et al.*, 2014; Rollie *et al.*, 2015). Next the repeat-first spacer junction is nicked and transesterification joins the other 3′ end of the spacer to the 5′ end of the first repeat (Arslan *et al.*, 2014; Rollie *et al.*, 2015). After the spacer is inserted the repeat is duplicated, using the first repeat as a template and all gaps filled (Yosef *et al*, 2012). Evidence supports PolA as the gap-filling enzyme in *E. coli* (Ivancic-Bace *et al.*, 2015).

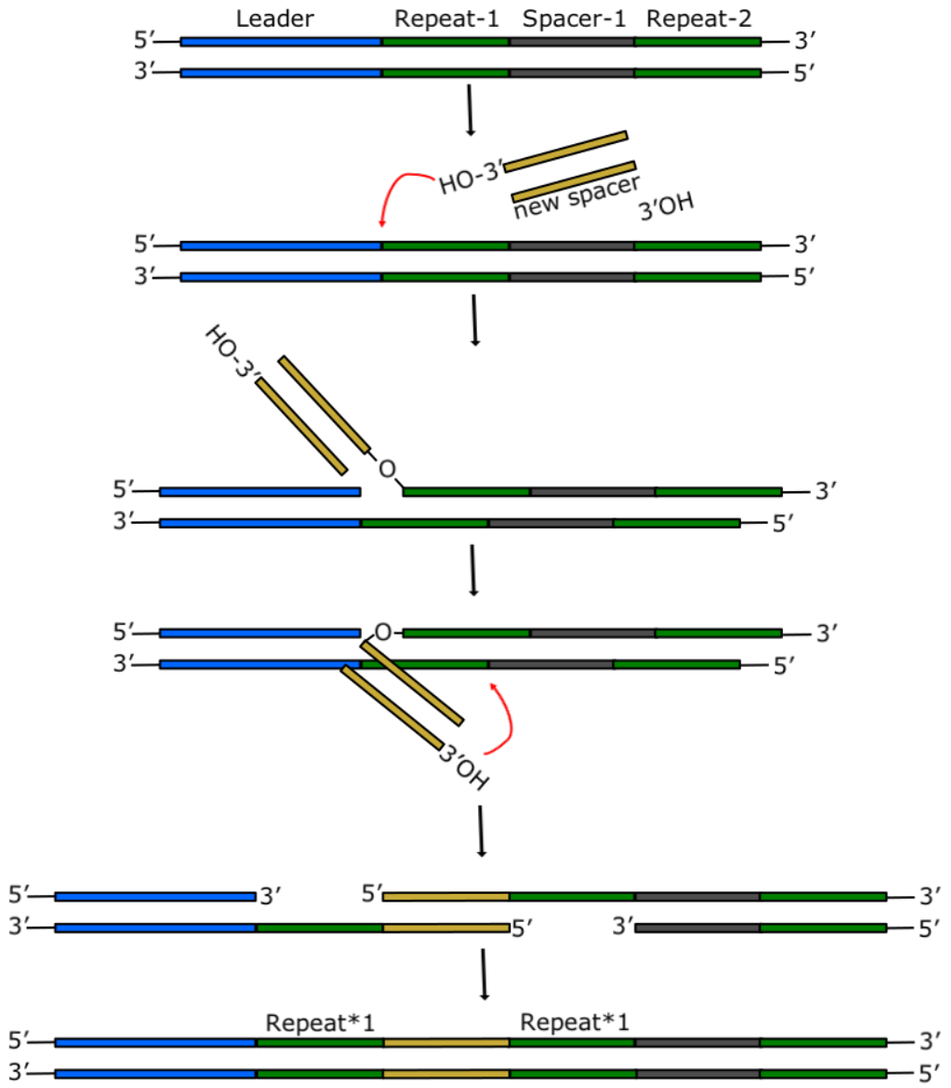

Figure 4: Proposed mechanism for integration of new spacer DNA into a CRISPR locus in *E. coli*. Cas1 nicks the leader-first repeat junction and catalyses a transesterification (TES) reaction joining the 3′ end of the incoming spacer to the 5′ end of the first repeat. Another TES reaction is carried out by Cas1 at the first repeat-first spacer junction, joining the 3′ end of the incoming spacer to the other 5′ end of the first repeat. After spacer integration, duplication of the first repeat and gap filling occurs. Adapted from Nuñez *et al*. 2016 that used data reported from Rollie *et al*. 2015.

To facilitate interference the CRISPR locus is transcribed, initiated from the leader sequence. Transcription is not always complete leading to a higher abundance of spacers from the 5' end of the CRISPR locus than the 3' end. As new spacers are always integrated at the leader end of the CRISPR locus (i.e. 5' end), new spacers will be less affected by incomplete transcription. This leads to higher transcription of newer spacers nearest to the leader and a lower transcription of older spacers at the 3' end of the locus. Transcription produces a transcript of the CRISPR locus called pre-CRISPR RNA (pre-crRNA), which in *E. coli* is 950nt long. This 950nt transcript is cleaved into individual crRNAs of 61nt consisting of a single spacer with part of the adjacent repeat (Tang *et al.*, 2002; Lillestøl *et al.*, 2006) by Cascade in *E. coli* but other proteins are involved in other systems (e.g. Cmr/Csy complex, RNase III, Cas9, Cas6) (Cady and O'Toole, 2011; Benda *et al.*, 2014). To facilitate cleavage, pre-crRNA contain secondary structure in the form of stem loops where cleavage occurs to separate individual crRNA units (Jore *et al.*, 2011). The crRNA and Cascade search DNA to find the complementary match to the crRNA. Cascade binds to DNA at the TIM and opens the DNA to allow binding by the crRNA. If complementary binding occurs, the target will be cleaved at the TIM. The TIM sequence is similar to the SAM sequence, and in some cases, they are the same sequence. The TIM sequence often incorporates the SAM sequence, but may be more stringent in tolerated sequence or may contain additional nucleotides (Garneau *et al.*, 2010; Gudbergsdottir *et al.*, 2011; Almendros *et al.*, 2012; Swarts *et al.*, 2012; Shah *et al.*, 2013) To prevent self-recognition of the spacers in the CRISPR locus two methods of self/non-self-recognition are employed. Firstly the TIM site is not present in the CRISPR locus, and as cleavage requires a TIM self-cleavage will not occur (Marraffini and Sontheimer, 2010; Li *et al.*, 2014). Secondly part of the repeat sequence is present in the spacer which does not occur in the original MGE sequence. Therefore, if the repeat sequence binds the binding site is identified as the CRISPR locus and nucleic degradation does not occur (Marraffini and Sontheimer, 2010).

The complementary binding at the MGE results in the formation of an R-loop. The crRNA invades the DNA duplex and base-pairs with its complementary sequence displacing the other DNA strand (Ivančić-Baće *et al*, 2012). R-loop formation begins at the TIM and extends towards the opposite end of the protospacer (Rutkauskas *et al.*, 2015). Upon stable R-loop formation Cas3 is recruited and cleavage of the

MGE DNA takes place (Garneau *et al.*, 2010; Howard *et al.*, 2011; Sapranauskas *et al.*, 2011; Benda *et al.*, 2014).

CRISPR-Cas functions as an adaptive immune system, but the system is not perfect. Despite a non-self-mechanism, self-targeting can still occur leading to cell death, inactivation of CRISPR-Cas or reduced expression of proteins. For example, self-targeting in *Pelobacter carbinolicus* has led to lowered histidine content of proteins due to targeting of the histidyl-tRNA synthetase (Aklujkar and Lovley, 2010). MGE targeting is also not robust because bacteriophages are constantly evolving to avoid being targeted through the CRISPR-Cas system. The CRISPR-Cas system relies on complementary binding of crRNA and the target sequence before cleavage occurs. Therefore, mutations within the target sequence would prevent cleavage, and allow the bacteriophage to propagate. These mutants are known as 'escape mutants' as they allow the bacteriophage to 'escape' targeted degradation. (Datsenko *et al.*, 2012). Instances of anti-CRISPR proteins which prevent CRISPR targeting (Bondy-Denomy *et al.*, 2013) and a bacteriophage with its own CRISPR-Cas system targeting the host genome have been found (Seed *et al.*, 2013). Although research has not determined how an organism combats anti-CRISPR, the CRISPR-Cas system does have a way of retargeting 'escape' mutants.

1.2.1 Adaptation occurs in three differing processes: Naïve, Targeted and Primed

Adaptation establishes a new spacer into a CRISPR locus and can be catalysed in three contexts: naïve, targeted and primed (Figure 5). Upon encountering an MGE for the first-time naïve adaptation takes place. This type of adaptation is inefficient and relies on a yet unknown mechanism to generate fragments from the MGE for capture by Cas1-Cas2. Targeted and primed adaptations occur when a spacer already exists (Datsenko *et al.*, 2012). In the case of targeted adaptation, the spacer has a perfect match meaning no mutation has occurred in the target DNA. This spacer targets the MGE and degrades it/recruits other proteins for degradation potentially creating products for capture by Cas1-Cas2. The DNA fragments produced and captured are from the primed strand (i.e. the strand where the crRNA binds) (Savitskaya *et al.*, 2013). Primed adaptation occurs when there is a mismatch between the spacer and MGE sequence (Richter *et al.*, 2014). Cleavage

does not generally occur but the crRNA can still bind with its protein partner. Following this Cas1-Cas2 could be recruited for 'capture' or another protein may be recruited to create DNA fragments for capture.

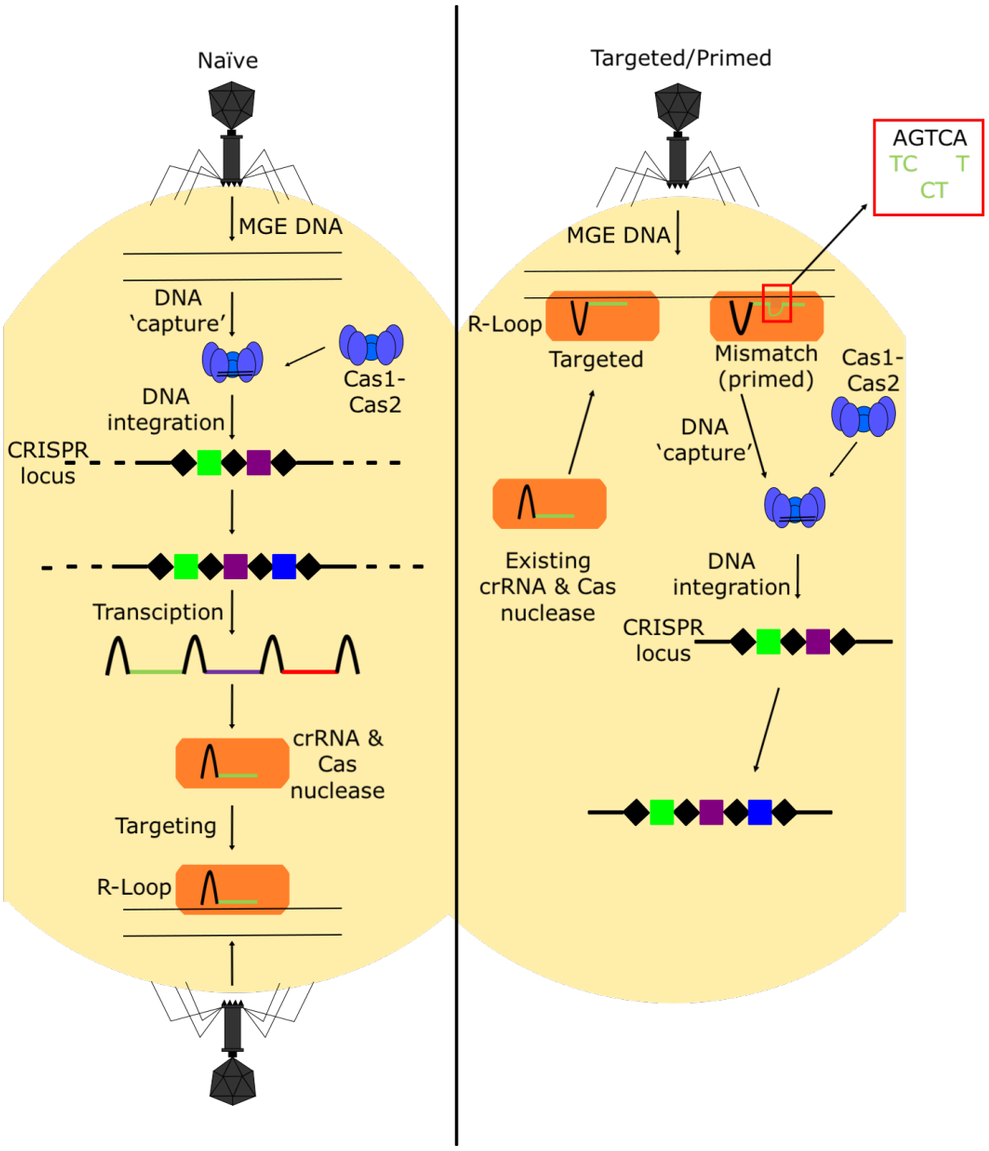

Figure 5: Summary of naïve, targeted and primed adaptation. Naïve adaptation (shown left) is the capture of spacers by Cas1-Cas2 from an unknown MGE. DNA fragments, generated from MGE DNA by unknown mechanism, are captured by Cas1-Cas2. Targeted or primed adaptation (shown right) occur via a similar mechanism, however a spacer against the MGE already exists in the CRISPR locus. Targeted adaptation uses a spacer with a perfect match which leads to degradation of the target, potentially creating fragments for capture by Cas1-Cas2. In primed adaptation, the spacer targets a mutated target sequence that does not allow cleavage. The spacer can still bind and may recruit Cas1-Cas2 for capture or recruit another protein that will generate fragments for Cas1-Cas2 capture.

Targeted and primed adaptations are more efficient than naïve adaptation, as spacer uptake was demonstrated in 4.3% of naïve cells, but 77% of primed/targeted cells (Datsenko *et al.*, 2012). Targeted adaptation is the dominant adaptation as when a 'perfect' target and a mismatched target were both available spacers were preferentially taken from the untargeted strand opposite the 'perfect' target. However, targeted adaptation is less likely to lead to multiple spacer

integrations due to the target being degraded by the interference machinery before further acquisition. (Datsenko *et al.*, 2012; Semenova *et al.*, 2016). All types of adaptation are required for a fully functional CRISPR-Cas system as naïve adaptation allows for resistance against MGEs not previously encountered, and targeted and primed adaptation allow the cell to keep/regain the resistance against a target. Though there is a system reported where only targeted and primed adaptation function (Li *et al.*, 2014) this was in a laboratory strain, meaning the cells may have lost the ability to carry out naïve adaptation due to a lack of bacteriophages in their environment. Targeted and primed adaptations involve both interference proteins and Cas1-Cas2. However, the interference machinery in *E. coli* and other Cascade/Cas3 based systems produces single-stranded products. Cas1-Cas2 require dsDNA spacers for integration into the CRISPR locus (Fagerlund *et al.*, 2017). As discussed earlier RecBCD, which has been implicated in naïve adaptation, also produces ss products. RecBCD and Cas3 products may be converted to suitable ds products through the same mechanism.

1.3 The process of DNA repair and its link with CRISPR-Cas immunity

The least understood, but most interesting parts of CRISPR-Cas immunity is the emerging involvement of DNA repair proteins required for 'pre-processing' DNA prior to DNA capture. Studies in *E. coli* and *Sulfolobus islandicus* have demonstrated the importance of DNA repair and DNA replication proteins for adaptation (Ivancic-Bace *et al.*, 2015; Levy *et al.*, 2015; Killelea and Bolt, 2017; Liu *et al.*, 2017). In *E. coli* RecG, RecBCD, PriA and DNA polymerase I have all been shown to be essential for adaptation, as knockout strains have reduced or prevented adaptation (Ivancic-Bace *et al.*, 2015; Levy *et al.*, 2015). In *S. islandicus* though proteins have not been shown to be essential for adaptation, the Csa3a transcriptional regulator has been shown to induce expression of *herA, nurA, DNA polymerase II* and *DNA polymerase beta* (Liu *et al.*, 2017). It is proposed that HerA and NurA are involved in adaptation in archaea, though more research is required. This section will highlight what is known about the DNA repair enzymes that have been identified as promoting or allowing adaptation in bacteria, particularly RecBCD, RecG and PriA.

1.3.1 DNA repair by Homologous Recombination requires RecBCD

DNA repair is required to maintain DNA sequence and structure following DNA anomalies including but not limited to ds breaks, nucleotide mismatch and oxidation of bases. HR is involved in the repair of ds breaks (Szostak *et al.*, 1983) and follows a similar mechanism across the three domains of life, though the proteins involved differ (reviewed in Blackwood et al. 2013). The crucial initiation of HR is end processing (resection) of DNA by nuclease-helicase enzymes e.g. RecBCD, AddAB, Mre-Rad50. In many bacteria, including *E. coli* RecBCD carries out this role.

RecBCD is a heterotrimer consisting of RecB (134kDa), Rec C (129kDa) and RecD (67kDa) (Amundsen *et al.*, 1986; Biek and Cohen, 1986; Finch *et al.*, 1986; Gorbalenya and Koonin, 1993; Aravind *et* al, 2000). RecB is a helicase-nuclease with a 3'→5' polarity (Bianco and Kowalczykowski, 2000) and both endo and exonuclease activities (Sun *et al.* 2006; Wang *et al.* 2000; M. Yu *et al*. 1998; Yu *et al.* 1998). RecB also contains an arm domain which contacts DNA, and is predicted to bind duplex DNA ahead of the complex. Through this binding the 'arm' could then pull the DNA in the opposite direction to the complex creating tension in the DNA opening it up (Singleton *et al.*, 2004; Krajewski *et al.*, 2014). RecC contains the Chi recognition site, a octamer sequence that modulates activity of RecBCD (Handa *et al.*, 2012). RecC contains three channels, two for each strand of the duplex DNA and one that interacts with RecB (Singleton *et al.*, 2004). In-between the two DNA channels is a 'wedge' domain which separates the duplex strands as they enter the complex (Singleton *et al.*, 2004). RecD is a helicase with 5'→3' polarity. The opposite polarities of RecB and RecD allows the proteins to bind opposite strands of dsDNA and track in the same net direction (Dillingham, Spies and Kowalczykowski, 2003). Both helicase domains require ATP hydrolysis to function (Roman and Kowalczykowski, 1989b, 1989a; Korangy and Julin, 1994) and the nuclease activity requires Mg^{2+} (Wright *et al*, 1971; Rosamond *et al*, 1979; Sun *et al*, 2006).

HR involves four stages: Initiation, homologous pairing and DNA strand exchange, DNA heteroduplex extension and resolution (Resnick, 1976; Szostak *et al.*, 1983). HR via double-strand break repair (DSBR) in *E. coli* as shown in Figure 6, occurs via the RecBCD pathway 95-99% of the time (Howard-Flanders and Theriot, 1966;

Emmerson, 1968; Willetts and Clark, 1969). When a double strand break occurs RecBCD binds at a blunt dsDNA end in a sequence independent manner (Farah and Smith, 1997; Bianco and Kowalczykowski, 2000) opening the duplex DNA by 5/6nt without the requirement for ATP (Farah and Smith, 1997; Dohoney and Gelles, 2001). Translocation occurs in an ATP-dependent fashion (Handa *et al.*, 2012) where RecD acts as the lead motor with the RecB motor translocating behind at a slower rate (Taylor and Smith, 2003). Both strands of DNA are degraded during translocation (Spies *et al.*, 2003) by the nuclease domain of RecB in an asymmetrical fashion. The cleavage rate is dependent on ATP: Mg^{2+} ratios, but under physiological conditions the 3' strand is cleaved every 10-100nt while the 5' strand is cleaved every 1000(+)nt (Dixon and Kowalczykowski, 1995). This activity is altered following recognition of the recombination site, Chi, by RecC. The Chi site (5'-GCTGGTGG-3') is approached from the 3' side (Bianco and Kowalczykowski, 1997) and recognition by RecC occurs through specific residues in its central channel (Handa *et al.*, 2012). The complex briefly pauses at Chi allowing for the slower RecB motor to catch up. Translocation after Chi is at a slower rate as RecB becomes the lead motor following an unknown conformational change inactivating the RecD helicase. (Taylor and Smith, 2003). Chi also causes an attenuation of the nuclease activity on the 3' strand, whilst degradation of the 5' strand increases (Wang *et al.* 2000). Chi remains bound in the RecC recognition site during translocation leading to a long recombinogenic 3' ssDNA tail (Wong *et al.*, 2006).

RecA is loaded onto the 3'ssDNA tail by the RecB subunit (Spies and Kowalczykowski, 2006). RecA is a DNA binding protein with ATPase activity (Shibata, Dasgupta and Cunningham, 1979). It was first discovered from mutations causing sensitivity to UV and deficiencies in recombination (Clark and Margulies, 1965). RecA is loaded onto ssDNA, displacing any single stranded DNA binding protein (SSB), (Cox and Lehman, 1982; Soltis and Lehman, 1983), creating a nucleoprotein filament. (Cox and Lehman, 1981b, 1982). This process requires ATP binding, but not ATP hydrolysis (Kowalczykowski, 1991). Following the formation of the nucleoprotein filament, the filament makes a number of simultaneous contacts along its length with duplex DNA looking for microhomology (Forget and Kowalczykowski, 2012). 8nt patches are searched for homology and if homology is present (Hsieh *et al*, 1992) the DNA is invaded and complementary binding occurs (McEntee *et al*, 1979; Shibata, Dasgupta and Cunningham, 1979; Weinstock *et al*,

1979; Cox and Lehman, 1981a) creating a D-loop. The invading DNA acts as a primer for DNA synthesis from the 3’ end by DNA Polymerase I (or another polymerase) (Kowalczykowski *et al.*, 1994; Berg, Tymoczko and Stryer, 2002; Hastings *et al.*, 2010). Branch migration by RecG or RuvAB migrates the branch point to allow full replication of the break point (Azeroglu and Leach, 2017). This creates a Holliday junction (HJ) which is then resolved through cleavage by RuvC.

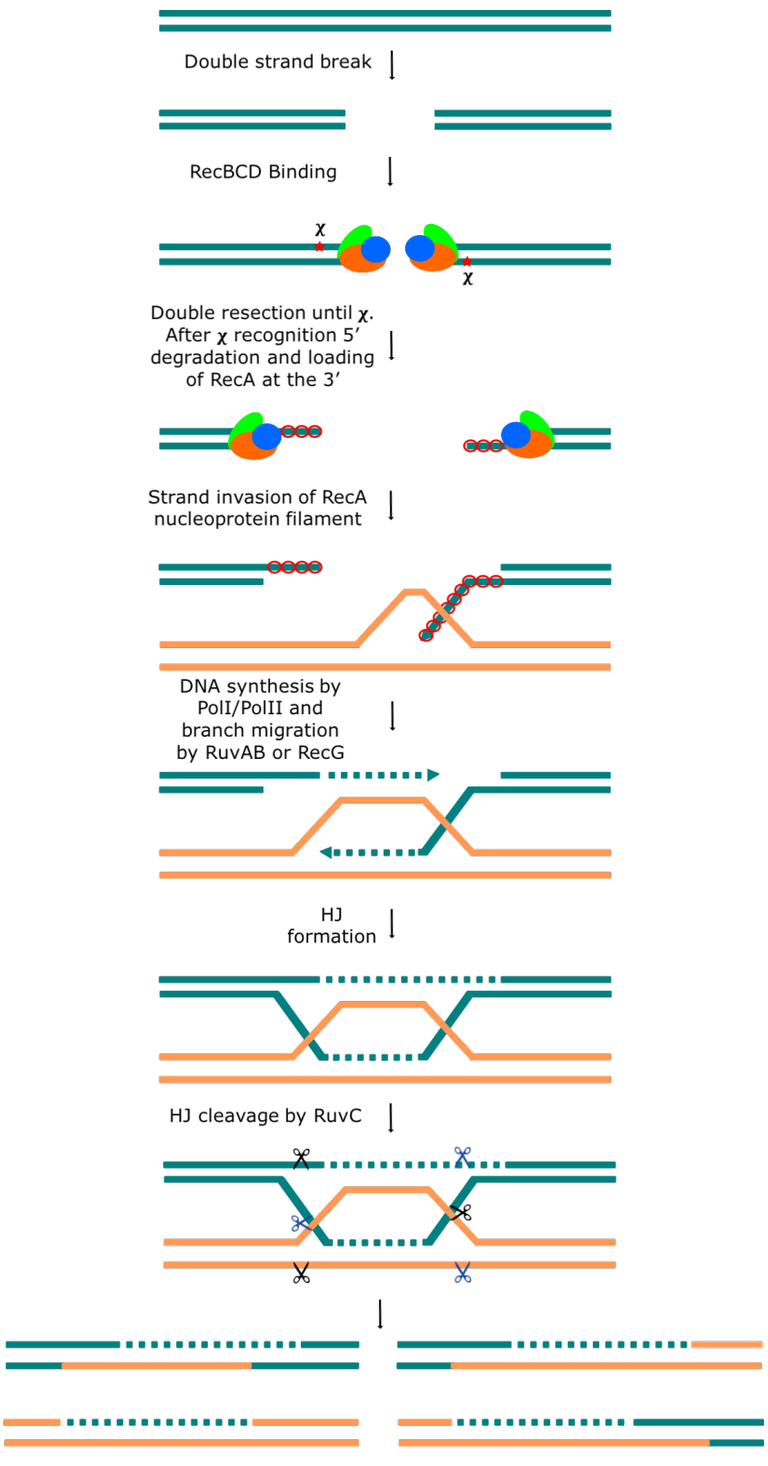

Figure 6: DSBR via RecBCD mechanism. A simplified mechanism of HR via RecBCD. Following DNA damage resulting in a ds break, RecBCD binds to the blunt end and resects the DNA until reaching a Chi site. Chi recognition alters activity attenuating 3’ cleavage creating 3’ ssDNA. RecB loads RecA onto the 3’ ssDNA creating a nucleoprotein filament which invades the homologous duplex. DNA synthesis and branch migration occur generating a HJ which can then be cleaved by RuvC.

RecBCD involvement in CRISPR-Cas adaptation has been tested through genetic studies. Ivancic-Bace *et* al demonstrated the requirement of RecB during naïve adaptation, as deletion of RecB removed the ability of E*. coli* to carry out naïve adaptation. RecC and RecD deletion strains were not tested. Levy *et al* also tested the effects of deletions on adaptation and showed that deletions of RecB, RecC and RecD all reduce naïve adaptation. These results would suggest that RecBCD is required for naïve adaptation, however latest research suggests that RecD is not involved and the phenotype is a consequence of hyperactive RecA loading (Unpublished data, Bolt & Ivancic-Bace). Previous data led to suggestions that RecBCD degradation of DNA generated DNA fragments for 'capture', but research has been unable to show these fragments or generate spacers from RecBCD degradation.

1.3.2 RecG is a branch migration protein involved in CRISPR-Cas immunity

As seen in Figure 6 a critical stage after RecBCD mediated resection and RecA mediated recombination is D-loop formation and migration. This can occur by one of several alternative mechanisms that are beyond the scope of this thesis. RecG is involved in one such mechanism and has been shown to be required for CRISPR-Cas immunity.

RecG, as discussed is involved in branch migration, but the function and role of RecG within cells is not fully understood. This is due to contrary experimental results and redundancy, meaning other proteins can carry out the function of RecG. Despite this RecG is present in almost all sequenced bacteria, suggesting an conserved function (Sharples *et al*, 1999; Rocha *et al*, 2005).

RecG has been established as a monomeric dsDNA translocase that targets HJ, three-strand junctions, D-loops and R-loops (Fukuoh *et al.* 1997; McGlynn *et al.* 1997; McGlynn *et al.* 2000; Singleton *et al.* 2001; Vincent *et al.* 1996; Whitby & Lloyd 1998). RecG contains three structural domains, as shown in the crystal structure in Figure 7. These are not true domains that function independently, but rather three regions that are structurally separated. The N-terminal domain contains the main DNA binding region where the DNA junction is bound (Singleton

et al, 2001) and mutation or deletion of the N-terminal removes DNA binding activity (Mahdi *et al.* 1997). The C-terminal domains contain the helicase region which has a 3'-5' polarity (Whitby *et al.* 1994; Singleton *et al.* 2001) and mutation or deletion of the C-terminal removes helicase activity (McGlynn *et al*. 2000). Suprisingly even deletion of the final residue at the C-terminal removes helicase activity (Upton *et al.*, 2014). The C-terminal also contains the protein-protein interaction site for which the residues R682 and W683 are important (Upton *et al.*, 2014).

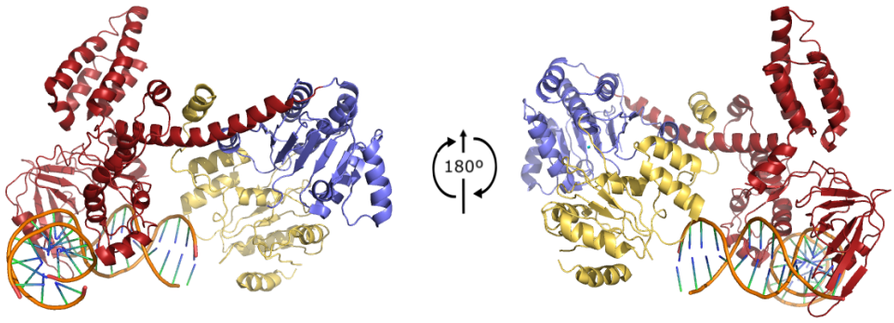

Figure 7: RecG crystal structure. RecG crystal structure from *T. maritima* contains three structural domains. Domain 1 (red) is located at the N-terminal of the protein and contains the main DNA binding site as shown by the bound DNA. The C-terminal contains domain 2 & 3 (blue and yellow) where the helicase region is located. Taken from Singleton *et al*. 2001

Mutational studies in RecG null cells and experiments with DNA damaging agents have provided most of the experimental evidence for the role of RecG. As mentioned the results are often contrary showing that RecG both increases and decreases recombination. RecG null cells have been shown to have deficiency in recombination (Storm *et al.*, 1971), particularly in high frequency recombination cells (Lloyd and Buckman, 1991). But it has been also shown that RecG null cells have increased deletion of repeats through the RecBCD pathway and therefore increased recombination (Lovett *et al.*, 1993; Lovett, 2006). The resolution of these experiments is that RecG increases and decreases recombination through different pathways. RecG may also have a role in DNA repair as RecG null cells have a loss of viability after UV exposure (Ishioka *et al*, 1997; Rudolph *et al.*, 2009).

RecG has been shown *in vitro* to carry out branch migration in either direction between a forked substrate and a Holliday junction (McGlynn and Lloyd, 2000, 2001; McGlynn *et al*, 2001). This activity requires localisation at the branched substrate which RecG achieves through interactions with SSB (Buss *et al.* 2008;

Lecointe *et al.* 2007; Upton *et al.* 2014; Zhang *et al.* 2010). Once localised RecG can translocate along the DNA separating the two strands and reannealing them again to carry out branch migration (Bianco and Lyubchenko, 2017). RecG can carry out branch migration *in vitro,* but *in vivo* evidence is lacking. RecG has been shown to carry out branch migration *in vivo*, but only in the absence of RuvAB (Mahdi *et al.*, 1996) as the two have overlapping function (Lloyd, 1991). RecG has also been suggested to be involved in fork reversal following DNA damage, to bypass the lesion and allow for replication restart. This will be discussed in section 1.3.3

RecG is required for primed adaptation as deletion of RecG in *E. coli* and *Pseudomonas aeruginosa* has been shown to reduce or remove primed adaptation (Ivancic-Bace *et al.*, 2015; Heussler *et al.*, 2016). Mutation of helicase and localisation activities prevents primed adaptation; however mutation of DNA repair activities allows primed adaptation. Therefore, the role of RecG in DNA repair is not required. RNase HI, which degrades R-loops, rescued primed adaptation activity when added to RecG null cells (Ivancic-Bace *et al.*, 2015). RecG can remove R-loops through branch migration and it was hypothesised that RecG removes roadblocks to replication caused by Cascade binding. Recent research has shown that RecG can remove these Cascade roadblocks by displacing the R-loop (Killelea *et al.*, 2018). This process of removing the Cascade roadblock may create DNA fragments for capture by Cas1-Cas2. However, these DNA fragment have no yet been experimentally shown.

1.3.3 PriA is a replication restart protein involved in CRISPR-Cas immunity

Another branching mechanism from DSBR is replication restart, which involves PriA. PriA is a DNA binding protein with 3’-5’ helicase activity (Tanaka *et al.*, 2002; Chen *et al*, 2004; Lopper *et al.*, 2007). PriA has 6 subdomains as shown in Figure 8: 3’ DNA-binding domain, winged helix, two helicase lobes, Cys-rich region and C-terminal domain. The 3’ DNA-binding and winged helix domains are the main sites of DNA binding and bind branched DNA structures with ssDNA present (Lee & Marians 1989; McGlynn *et al*. 1997). The two helicase lobes have 3’-5’ DNA polarity (Lee and Marians, 1989) and can unwind D-loops and replication forks (McGlynn *et*

al., 1997). Within lobe 2 is a Cys-rich region which coordinates two Zn^{2+} ions required for helicase activity (Bhattacharyya *et al.*, 2014). The final subdomain is C-terminal domain that is also involved in DNA binding and is important for PriA activity, as its removal effects activity (Jaktaji and Lloyd, 2003; Bhattacharyya *et al.*, 2014).

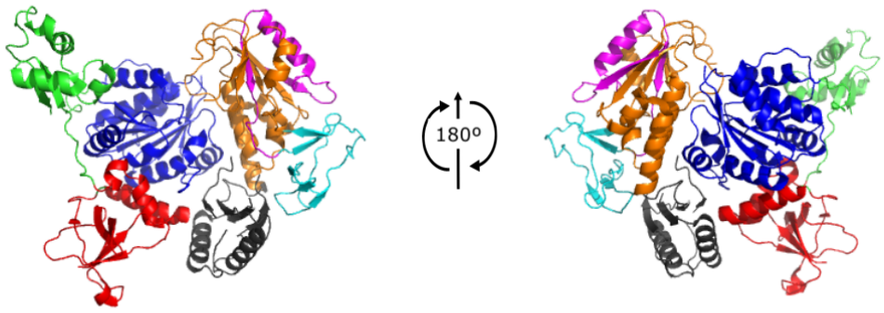

Figure 8: PriA (*Klebsiella pneumoniae*) crystal structure. PriA is split into 6 subdomains: 3' DNA binding domain (red), winged helix (green), helicase lobe 1 (blue), helicase lobe 2 (orange), Cys-rich region (light blue) and C-terminal domain (black). The 3' DNA binding domain and winged helix coordinate DNA binding. The helicase activity is contained within the two helicase lobes, though the Cys-rich region coordinates the Zn^{2+} ions required for helicase activity. The C-terminal appears to also be involved in DNA binding and general activity, as its removal is detrimental.

The main pathway involving PriA is replication restart. PriA is directed to a suitable substrate for replication restart, such as a stalled replication fork or D-loop (McGlynn *et al.*, 1997; Nurse *et al*, 1999; Tanaka *et al.*, 2007) through interactions with SSB (Cadman & McGlynn 2004; McGlynn *et al.* 1997; Kozlov *et al.* 2010; Tanaka *et al.* 2003) and the PAS (primosome assembly site) (Shlomai and Kornberg, 1980). PriA can then unwind the branched substrate using the 3'-5' helicase activity and remove SSB to create a suitable substrate for primosome loading (Gabbai & Marians 2010; McGlynn *et al.* 1997). PriA can then recruit proteins for replication restart including PriB and DnaT before loading of the replisome. While PriA helicase activity is required to create a suitable substrate for replication restart, it is not required for replisome loading (Heller & Marians 2006; Liu *et al*. 1996; Liu & Marians 1999; Liu *et al.* 1999).

As mentioned earlier, RecG may have role in replication restart by converting a Holliday junction to a substrate suitable for PriA interaction and loading of the replisome. PriA and RecG deletions are poorly viable, meaning they may interact

(McCool and Sandler, 2001; Gregg *et al.*, 2002). Also, deletion of RuvABC leads to another mechanism to resolve Holliday junctions which requires RecG and PriA. The hypothesis is that RecG branch migrates the Holliday junction in the opposite direction creating a forked structure upon which PriA can bind and load the replisome (Al-Deib *et al.* 1996; Gregg *et al*. 2002; Jaktaji & Lloyd 2003; McGlynn & Lloyd 2000). However, there is no evidence that RecG or PriA interact together or that RecG is involved in replication restart (McGlynn *et al.* 1997; Rudolph *et al.* 2010).

PriA is also required for primed adaptation, however in a way that does not require the helicase activity or primosome loading capacity of PriA (Ivancic-Bace *et al.*, 2015). One hypothesis is that RecG removes Cascade roadblocks (see section 1.3.2.) and PriA then binds the resulting branched substrate. This prevents the branched DNA being converted into a substrate for HR and potentially creates a substrate for capture. RecG and PriA may interact in similar fashion as in replication restart (if the interaction can be experimentally proven).

1.3.4 PolA is a DNA polymerase protein involved in CRISPR-Cas immunity

PolA encodes DNA polymerase I, a Family A DNA polymerase (Ito and Braithwaite, 1991; Garcia-Diaz, 2007). All DNA polymerases catalyse the addition of deoxyribonucleotides to a DNA chain in a 5'-3' direction (Kornberg, 1969; Berg, Tymoczko and Stryer, 2002) and the reaction requires all four deoxyribonucleotides and two metal ions (typically Mg^{2+}) (Berg, Tymoczko and Stryer, 2002).

DNA polymerase I was the first polymerase identified and was initially isolated from *E. coli* as a 109kDa protein (Lehman *et al.*, 1958; Jovin *et al*, 1969; Jovin *et al,* 1969). It has three active regions: 3'-5' exonuclease, polymerase and, 5'-3' exonuclease (Lehman and Richardson, 1964; Klett *et al*, 1968; Kelly *et al.*, 1969). The 3'-5' exonuclease region provides a proofreading function. After insertion of an incorrect base DNA synthesis is stalled. The 3'-5' exonuclease activity of DNA polymerase I can excise this incorrect base before synthesis continues (Richardson *et al.*, 1964; Brutlag and Kornberg, 1972). The polymerase region contains three conserved motifs, with Motif A and C being catalytic sites and Motif B the dNTP

binding site (Albà, 2001). The 5’-3’ exonuclease region is involved in excision of primer sequences before synthesis of DNA to complete replication (Westergaard, Brutlag and Kornberg, 1973; Grossman *et al.*, 1975; Berg, Tymoczko and Stryer, 2002).

Due its poor proccessivity where only ~20 nucleotides are added per synthesis reaction, DNA Polymerase I is not the main replicative polymerases. DNA polymerase I is instead involved in gap-filling (which is required during DNA replication) and DNA repair (Konrad and Lehman, 1974; Grossman *et al.*, 1975; Savic, Jankovic and Kostic, 1990). In DNA replication (Figure 9), the two strands (leading and lagging strands) are synthesised differently due to the 5’-3’ directionality of DNA polymerases. The leading strand is synthesised continuously, whereas the lagging strand is synthesised discontinuously as okazaki fragments (Okazaki, 1968; Painter and Scaefer, 1969; Savic, Jankovic and Kostic, 1990). Both strands require RNA primers for synthesis to occur, but as the lagging strand is synthesised discontinuously multiple primers are required (Sugino *et al*, 1972; Wagar and Huberman, 1973). The gaps between the okazaki fragments must be filled and the primers removed. The RNA primer is removed by the 5’-3’ exonuclease activity of DNA polymerase I, before gap-filling by the polymerase region (Konrad and Lehman, 1974) and joining of the backbone by DNA ligase (Sugimoto, Okazaki and Okazaki, 1968; Gefter, 1975; Cooper, 2000).

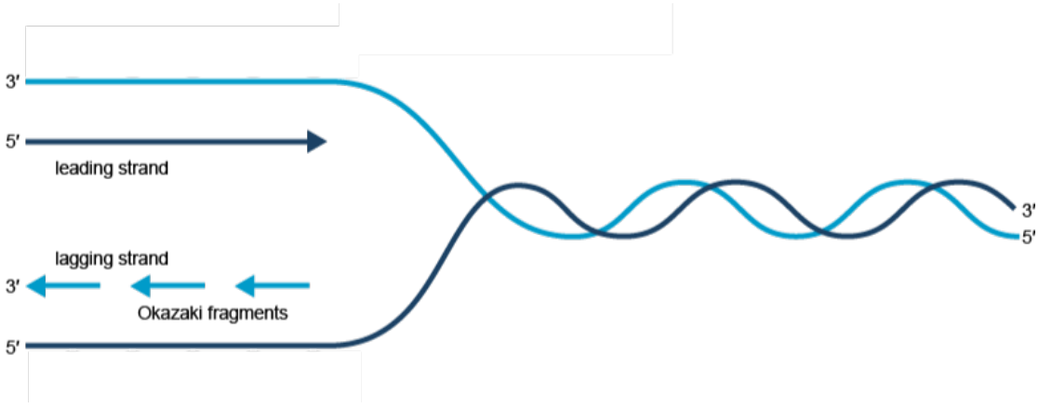

Figure 9: Simplified diagram of DNA replication. The leading strand and lagging strand are synthesised continuously and discontinuously respectively as a result of polymerase activity occurring only in the 5’-3’ direction.

In DSBR repair (Figure 6), DNA synthesis is required to replace the DNA resected by RecBCD. After invasion of duplex DNA by the 3’ invading strand and the creation of a D-loop, DNA polymerase I can synthesise DNA using the 3’ invading strand as a primer (Cooper and Hanawalt, 1972; Holmes and Haber, 1999; Hastings *et al.*,

2010). This in combination with branch migration allows the resected DNA to be synthesised (Cooper and Hanawalt, 1972).

DNA polymerase I is required for both naïve and primed adaptation. However, only the polymerase activity is required, as a mutant *polA* possessing both exonuclease activities does not support naïve or primed adaptation in *E. coli.* DNA polymerase I, as discussed above, is involved in gap-filling and gap-filling is required after spacer integration to duplicate repeat sequences. It is highly likely that *polA* is required for repeat duplication after the initial integration of a new spacer (Ivančić-Bace *et al.*, 2015).

Genetic work within *E. coli* has provided the majority of evidence for roles for RecBCD, RecG, PriA and PolA within CRISPR-Cas immunity (Ivancic-Bace *et al.*, 2015; Levy *et al.*, 2015). Not all CRISPR-Cas systems will utilise these proteins and some systems will be present in organisms lacking these proteins. CRISPR-Cas has been linked to DNA repair, programmed cell death, signal transduction and, horizontal gene transfer across different organisms (Faure, Makarova and Koonin, 2019). CRISPR-Cas therefore have diverse mechanisms involving different Cas proteins and interacting partners (i.e. RecBCD). For example, the *cas* gene *cas4* has an unknown role in CRISPR-Cas immunity, but has some similarities to RecB and may carry out a similar role. Cas4 and its CRISPR-Cas system are discussed in more detail in section 1.4.3.

1.4 Molecular Genetics and phylogeny of CRISPR-Cas

Numerous and diverse Cas proteins are involved in CRISPR-Cas immunity, though Cas1 and Cas2 are found in most known systems (Makarova, *et al*. 2015). This diversity is likely due to the competitive evolution between viruses and CRISPR-Cas, creating rapid evolution in *cas* genes leading to diverse gene architectures (Haft *et al.*, 2005; Takeuchi *et al.*, 2012). A classification system was devised to allow characterisation of CRISPR-Cas systems based on their *cas* gene content shown in Figure 10 (Class 1) and Figure 11 (Class 2). This figure shows the complexity and variety of *cas* genes within the CRISPR-Cas systems. The classification system is not all-encompassing, so each classification shows genes commonly found within that class but variations and additions to those common

genes are tolerated. Not shown in this classification is the casposon system, a transposable element containing a *cas1* homolog. Due to differences in mechanism and gene content, it cannot be classified as a CRISPR-Cas system, but will be discussed in more detail in section 1.4.4. Also, discussed below are classes of importance to CRISPR-Cas (*E. coli* and Cas9 systems) and classes relating to this research (Type I-U).

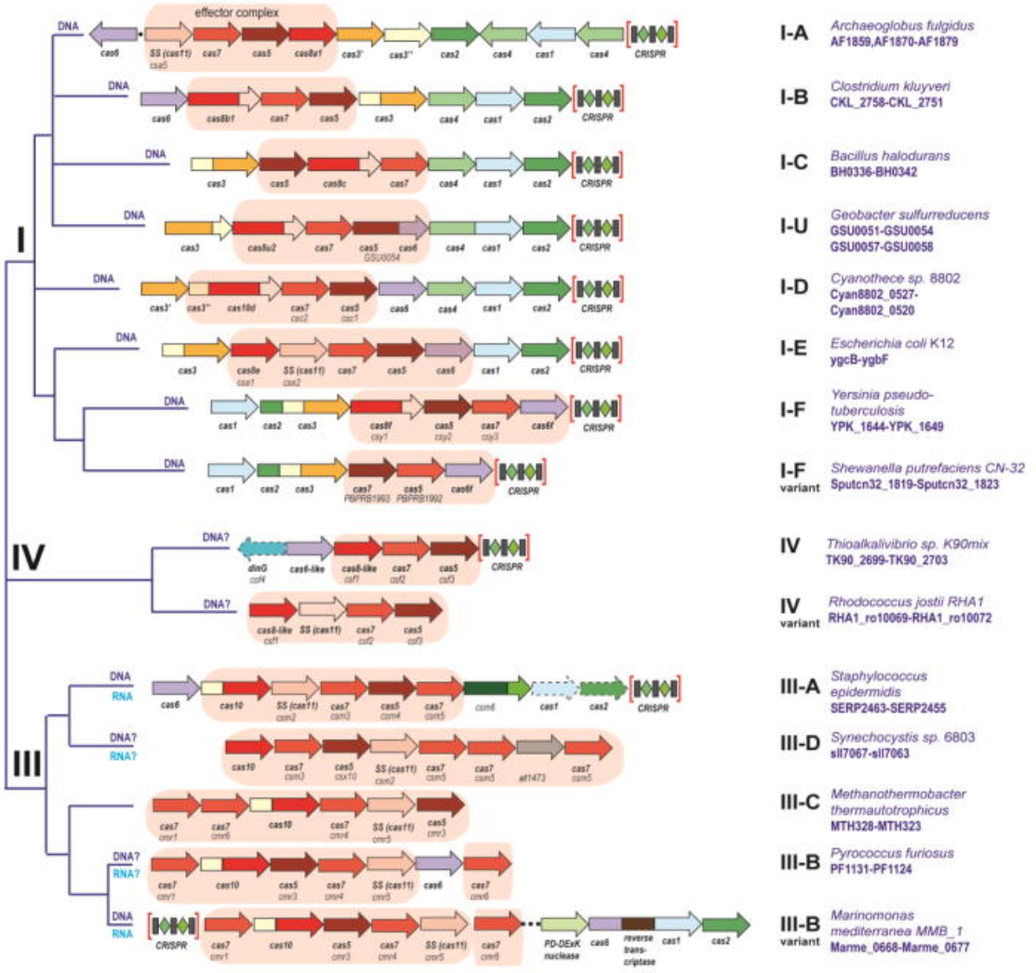

Figure 10: Classification of Class 1 CRISPR-Cas systems. The domain architecture of each sub-type is shown as well as the predicted target. The genes involved in the effector subunits for Class 1 are shaded. Taken from Koonin, Makarova and Zhang, 2017

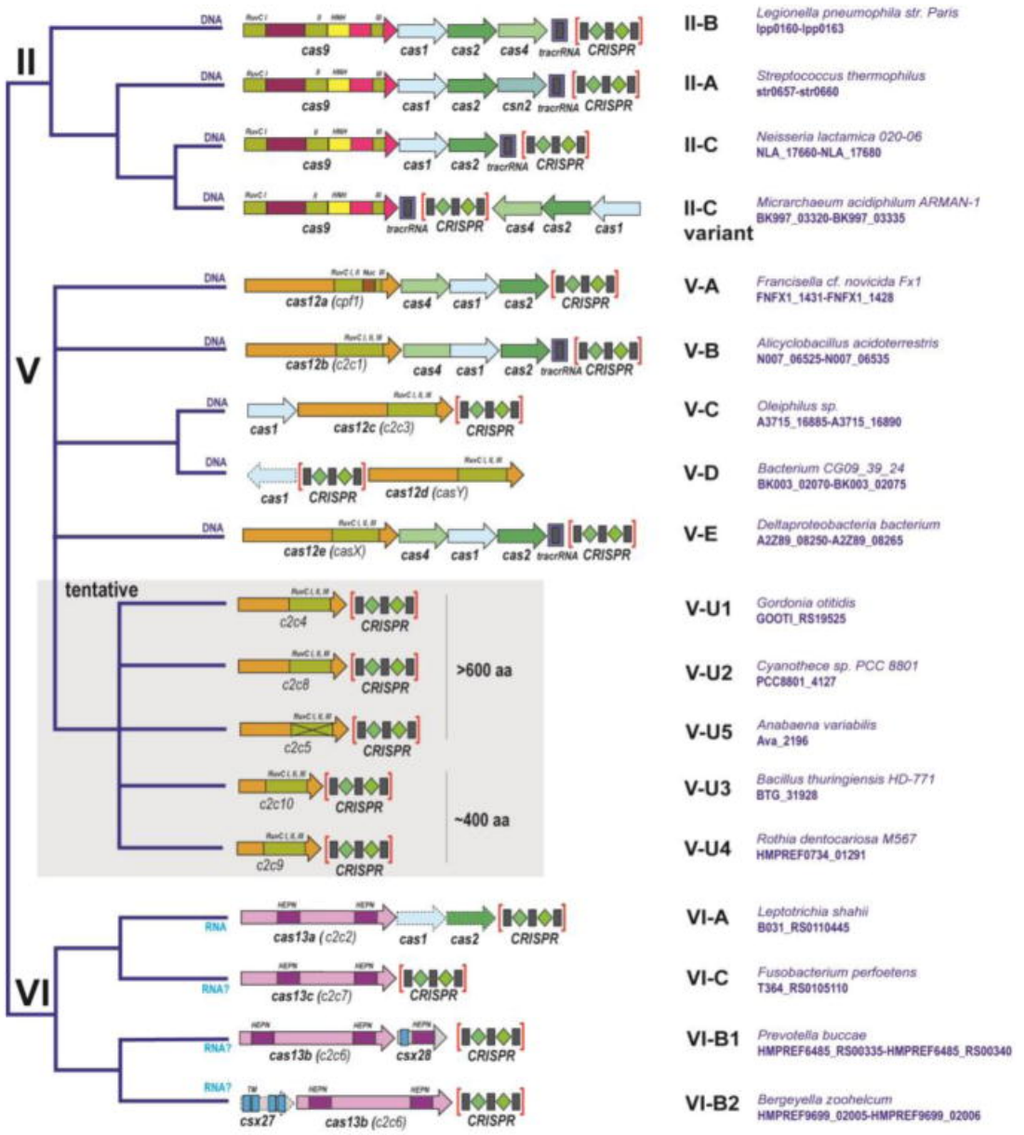

Figure 11: Classification of Class 2 CRISPR-Cas systems. The domain architecture of each sub-type is shown as well as the predicted target. The genes involved in the effector subunits for Class 2 are shaded. Taken from Koonin, Makarova and Zhang, 2017

1.4.1 The *E. coli* type I-E system

The *E. coli* type I-E system is the model system for CRISPR-Cas, due to the large amount of research conducted within this system. Cas1 and Cas2 carry out adaptation as described in section 1.2. Interference in *E. coli* involves Cascade and Cas3. Cascade is a protein complex made of 5 subunits in the ratio 1:Cse1, 2:Cse2, 5:Cas7, 1:Cas5 and 1:Cas6e. It binds and processes crRNA. Cas3 contains both a HD nuclease domain and a SF2 helicase domain (Westra *et al.*, 2012; Gong *et al.*, 2014).

Despite *E. coli* being the model system, the native system is silenced by H-NS meaning CRISPR-Cas immunity is not functional in wild-type cells. H-NS binds to the sense strand of the CRISPR locus preventing RNA polymerase binding and

therefore transcription. (Pul *et al.*, 2010; Westra *et al.*, 2010). As a result, CRISPR-Cas is not conserved across all *E. coli* phylogeny. The B2 phylogenetic group no longer possesses any Cas genes, though the CRISPR locus is still present. *E. coli* that have retained CRISPR-Cas have little difference in CRISPR sequences, showing the system is not active. Conservation of Cas genes in these *E. coli* may be due to alternative roles of Cas proteins (Touchon and Rocha, 2010; Touchon *et al.*, 2011). To research the system, the Cas proteins must be expressed through an inducible system. So, whilst *E. coli* is the most researched, it may not be the best model system.

In the *E. coli* CRISPR-Cas system Cas1-Cas2 capture a 32nt spacer, which includes the last nucleotide of the SAM (Goren *et al.*, 2012), and integrates it into the CRISPR locus (Arslan *et al.*, 2014). The type I-E system along with other Type I systems requires the integration host factor (IHF) to bind to the leader sequence and bend the DNA before integration. IHF is made up of two subunits IHFα and IHFβ which complex together as a heterodimer (Friedman, 1988). IHF binds to its recognition sequence WATCAANNNNTTR (where W is A/T, R is A or G and N is any nucleotide) located in the leader sequence at the minor grove of DNA (Yang and Nash, 1989; Goodrich *et al*, 1990; Aeling *et al.*, 2006). This binding forces a 180° bend in the DNA (Rice *et al.*, 1996) opening the DNA allowing Cas1 to nick the locus and integrate the spacer. The binding of IHF inhibits the disintegration reaction (Rollie *et al.*, 2015; Yoganand *et al.*, 2017), the opposite reaction to integration that Cas1-Cas2 can also carry out, promoting integration of spacers. Integration can occur in the absence of IHF but spacers are not inserted adjacent to the leader (Nuñez *et al.*, 2016). Another binding site has been found called the integrase anchoring site (IAS) and is believed to be the Cas1-Cas2 binding site in the leader sequence. In type I-E systems IHF binding site (IBS) and IAS are conserved within each system. The IAS is required for adaptation with its deletion preventing acquisition, though it has yet to be conclusively proven as the Cas1-Cas2 binding site (Yoganand *et al.*, 2017).

Following transcription of the CRISPR locus, Cascade binds to the pre-crRNA and the Cas6e subunit cleaves a single crRNA of 61nt which remains complexed with Cascade (Brouns *et al.*, 2008). Cascade samples and opens up ds target DNA through interactions between lysine residues in the Cas7 subunits and the

phosphate backbone (Xue *et al.*, 2017). The Cse1 subunit recognises any TIM sequences present (Sashital *et al*, 2012; Hayes *et al.*, 2016) and the crRNA binds through complementary binding creating an R-loop and a bulge in the DNA (Jore *et al.*, 2011). Following R-loop formation Cse1 undergoes a conformational change allowing the recruitment of Cas3 (Rutkauskas *et al.*, 2015; Brown *et al.*, 2017; Xiao *et al.*, 2017). Cas3 then cleaves the target DNA using its HD domain, before unwinding and degrading DNA in 3'-5' direction (Westra *et al.*, 2012).

1.4.2 Cas9 is found in type II systems

Type II CRISPR-Cas systems are characterised by the *cas9* gene (Makarova, *et al.* 2015). Cas9 is best known for its manipulation as a genome editing tool. However, it was discovered originally as part of a CRISPR-Cas system. Cas9 was initially discovered in *Streptococcus thermophilus,* an organism important in the dairy industry. *S. thermophilus* contains four CRISPR systems, of which two contain Cas9 (Horvath and Barrangou, 2010; Makarova *et al.*, 2011). It was in *S. thermophilus* that the CRISPR locus was first shown to take up new spacer sequences in response to phage exposure (Barrangou *et al.*, 2007; Deveau *et al.*, 2008; Garneau *et al.*, 2010). This was of importance as it could help provide protection to *S. thermophilus* dairy cultures. In Type II systems Cas1-Cas2 complex capture DNA and integrate the DNA into the CRISPR locus as described in section 1.2. Transcription of the CRISPR locus in type II systems occurs in two directions generating a pre-crRNA transcript and a trans-activated RNA (tracrRNA) transcript transcribed from the adjacent DNA strand in the opposite direction (Jinek *et al.*, 2012). This tracrRNA contains a sequence match for the repeat sequence and the two long unprocessed transcripts of crRNA and tracrRNA associate. This binding provides a RNA with a secondary structure which can be recognised and cleaved by RNase III. RNase III is not contained within the CRISPR gene neighbourhood, but is required to process type II crRNA and tracrRNA correctly due to a lack of a RNA processing Cas protein (Deltcheva *et al.*, 2011). Following cleavage, the crRNA and tracrRNA bind to Cas9, which is then directed to the target site where the crRNA binds via an R-loop. The cleavage is two-fold carried out on opposite strands by the two active sites of Cas9, HNH and RuvC, just upstream of the TIM site (Chen *et al*. 2014; Chylinski *et al*. 2013; Gasiunas *et a*l. 2012).

For genome editing, the crRNA and tracrRNA are fused together to create a single guide RNA (sgRNA) (Anders and Jinek, 2014). Cas9 can be directed to a specific DNA site to cut the target DNA to allow insertion of a new sequence or to create indels (insertions or deletions). The technology is currently held back due to editing often occurring at a low-levels (Cho *et al.*, 2013) and editing off-target (Fu *et al.*, 2013). Low-level editing means most cells are unaffected. For the mutation to take hold in a population or organism the majority of cells need to be affected. This can be remedied through optimisation and editing of the mechanism so whilst it is an issue now, it could be solved for the future. Off-target effects occurs because mismatches in crRNA can be tolerated to an extend during binding, resulting in Cas9 targeting to a mismatch sequence. Cas9 will therefore cause a mutation away from the target (Mali *et al.*, 2013). Off-targets effects are more pronounced in guide sequences with a high GC content (Lin *et al.*, 2014). To improve this bioinformatic tools have been developed to predict off-target events and help avoid them (see Cui *et al.* 2018 for a review). Cas9 technology has been suggested/used for treating bacterial infections (Bikard *et al.*, 2014), treatment of genetic diseases (Mali *et al.*, 2013; Yin *et al.*, 2014), changing gene expression (Bikard *et al.*, 2013) and creation of knock-outs. It therefore has much potential for the future.

1.4.3 Cas4-1 fusion is found in type I-U systems

The type I-U system is similar to the Type I-E system (Makarova, *et al.* 2015), but is characterised by a Cas4-1 fusion. This system is little studied but it is predicted that Cas4-1 with Cas2 can carry out adaptation. The Cas4 region of the fusion has an unknown function, but is involved in adaptation (Plagens *et al.*, 2012; Li *et al.*, 2014; Liu *et al.*, 2017).

1.4.3.1 CRISPR-Cas Protein Cas4

Cas4 is part of the PD-(D/E)XK family, a diverse family containing DNA restriction enzymes, DNA recombination enzymes and DNA polymerases (Steczkiewicz *et al.*, 2012). Cas4 contains a RecB like nuclease domain (Jansen *et al.*, 2002; Makarova *et al.*, 2006) and crystal structures from *S. solfataricus* and *Pyrobaculum calidifontis* have shown similar folds to AddB (shown in Figure 12) (Lemak *et al.* 2013; Lemak *et al.* 2014). Both AddB and RecB are involved in DNA repair where they resect DNA from a ds break to create recombinogenic 3’ ssDNA tails (more details can be found section 1.3.1). Cas4 also contains an iron sulfur cluster

(4Fe:4S/2Fe:2S) coordinated by four cysteines (Lemak *et al.* 2013; Lemak *et al.* 2014). Iron sulfur clusters are usually involved in DNA binding and Cas4 has been shown to bind DNA with single-stranded regions and cleave ssDNA and dsDNA (at a slower rate) in a 5'-3' direction (Lemak *et al.*, 2013; Lemak *et al.*, 2014).

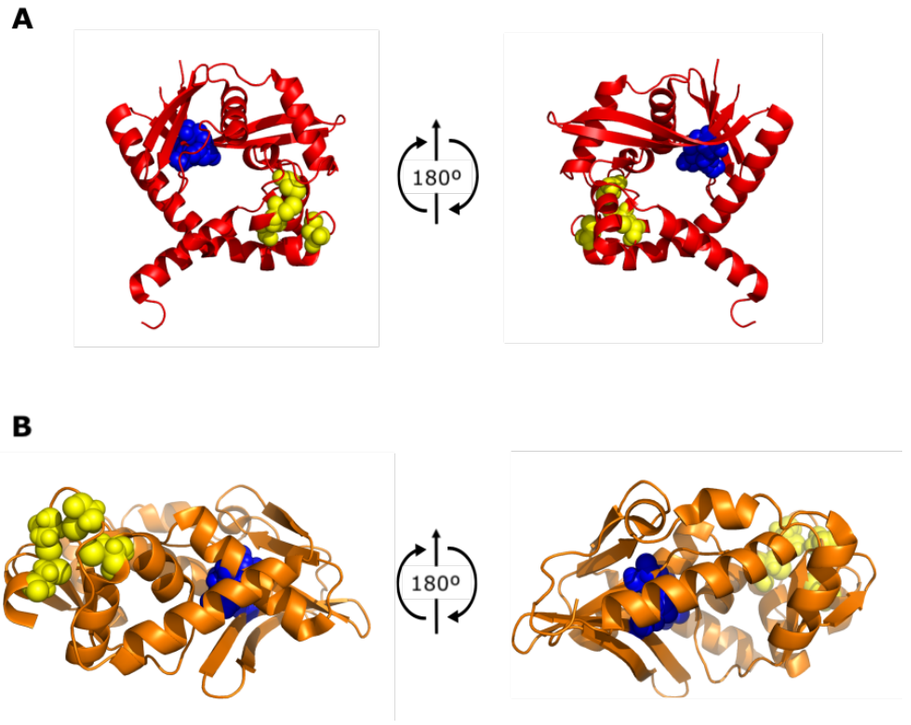

Figure 12: Monomeric structures of Cas4 from *S. solfataricus* and *P. calidifontis.* Crystal structures of Cas4 monomers from *S. solfataricus* (A) and *P. calidifontis* (B). Both proteins contain a RecB nuclease site (blue) and a iron sulfur cluster (yellow).

Cas4 is required for adaptation in *S. islandicus* and *Haloarcula hispanica* (Li *et al.*, 2014; Liu *et al.*, 2017) and can form a complex with Cas1 and Cas2 *in vitro* (Plagens *et al.*, 2012). Therefore it is hypothesised that Cas4 nuclease activity may be involved in generating protospacers or for trimming protospacer to a suitable length for integration (Zhang, Kasciukovic and White, 2012; Lemak *et al.*, 2013; S Lemak *et al.*, 2014)

1.4.3.2 Interest in Cas4-1 fusion

The role and function of Cas4-1 fusions has not been explored. However, research of this fusion could help understand the role of Cas4 in adaptation. In adaptation Cas4 may create DNA fragments for capture by the linked Cas1. As Cas4 contains a RecB-like nuclease domain this may provide information on a possible mechanism for the role of RecB in naïve adaptation. If Cas4 can generate fragments for

integration, then it may be possible to experimentally create a single *in vitro* adaptation reaction.

1.4.4 Casposon system is a self-transposable element containing a *cas1* gene

Transposable elements are DNA elements that can be excised from the genome and integrated elsewhere. This excision can be carried out by an enzyme encoded in the transposable element or in the genome. Transposable elements are known as 'selfish DNA' as they are self-serving and have no function within the genome (Hickman and Dyda, 2016). The casposon is a transposable element which contains a *cas1* gene, usually referred to as casposase. Other associated proteins/protein motifs are Family B DNA polymerase, HNH nuclease, helix-turn-helix (HTH) and occasionally Cas4 (Koonin and Krupovic, 2015; Hudaiberdiev *et al.*, 2017). The role of most of these proteins within the casposase is unknown (Hickman and Dyda, 2015). These genes are flanked by terminal inverted repeats (TIR) and target site duplication (TSD), where the operon can be excised by the casposase and integrated elsewhere in the genome rather like a transposon (Krupovic *et al.*, 2014). This ability of the casposase to integrate DNA has been shown *in vitro* by the integration of duplex DNA into a plasmid in a sequence specific manner (Hickman and Dyda, 2015; Béguin *et al.*, 2016; Krupovic *et al.*, 2016).

The casposon system may represent a separate evolution step, or the origin of the CRISPR-Cas system. A possible evolution pathway is shown in Figure 13. The theory is that the ancestral casposon system was inserted into the genome at its target site adjacent to an effector module containing interference *cas* genes. Over time the polymerase, HNH and NTH were lost along with the TIRs and the *cas2* gene was recruited through an unknown mechanism. One TSD is repeated creating the CRISPR locus, whilst the other becomes the leader sequence (Krupovic, Béguin and Koonin, 2017). This theory is supported by the sequence and secondary structure similarities between Cas1 and casposase which leaves little doubt to their relatedness. However, there is not enough evidence currently to fully support this hypothesis.

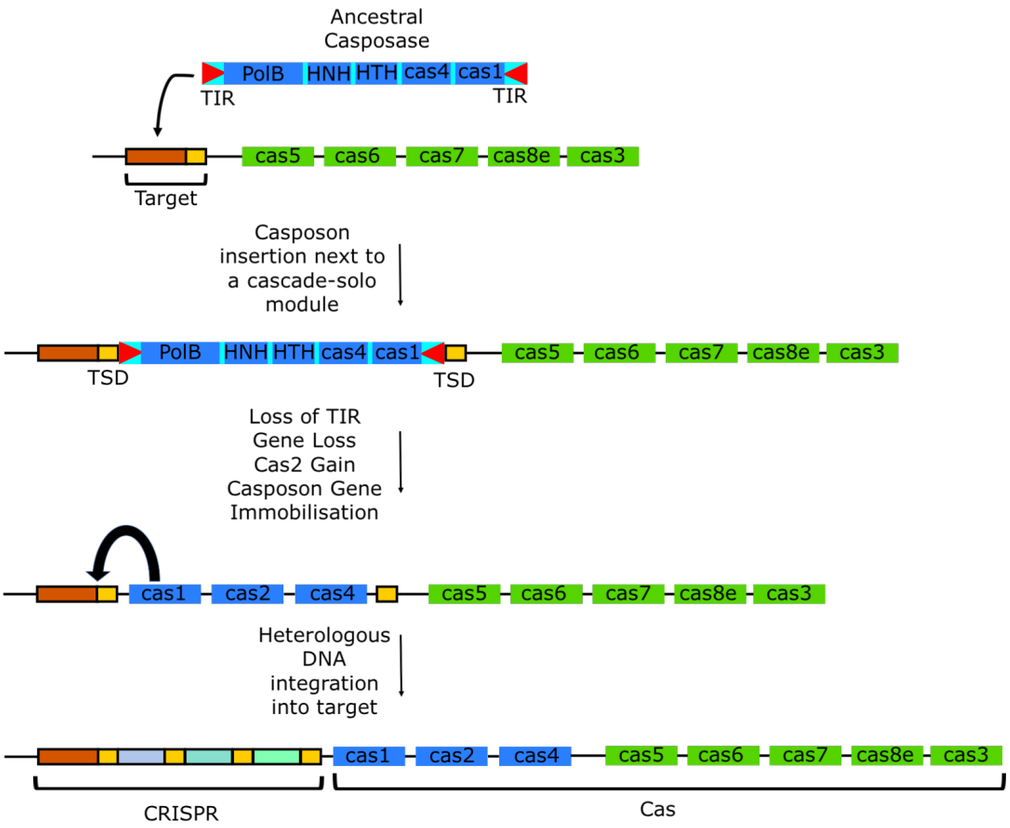

Figure 13: Proposed evolutionary mechanism of casposon to CRISPR-cas. The casposon integrates adjacent to a solo-effector operon, in this case a cascade module. Over time the majority of casposon genes are lost leaving Cas4 and Cas1. Cas2 is recruited from an unknown location and the TSD at one end are duplicated. This gives a CRISPR operon like those found today. Adapted from Krupovic *et al*. 2017.

1.5 Summary

CRISPR-Cas is an adaptive immune system with two main stages: adaptation and interference. Adaptation is the 'capture' of MGE DNA and its integration into the CRISPR locus as a spacer. The CRISPR locus is transcribed and separated into single units of the locus, this crRNA while bound to interference protein/complex allows targeted degradation. There are three types of adaptation: naïve, targeted and primed. Naïve is the 'capture' of spacers from MGE that has not been encountered before. Targeted is the 'capture' of spacers from MGE that are targeted by a perfectly matched crRNA. Primed is 'capture' of spacers from MGE targeted by a mismatched crRNA. Targeted and primed adaptation show there is a link between adaptation and interference, but the mechanism of this is unknown.

RecBCD, RecG and PriA are involved in HR and branching pathways, and have been shown to be required for adaptation. The function of these proteins within CRISPR-Cas immunity is not known, but hypotheses have been made. These proteins are

not conserved across all organisms containing CRISPR-Cas systems. Therefore, other protein partners or *cas* genes of unknown function may carry out these roles in other organisms. The Cas4-1 fusion from type I-U CRISPR-Cas systems has a RecB-like nuclease domain and may therefore carry out the same role as RecB.

1.5.1 Research Aims

The broad aim at the outset of this project was to investigate the properties of Cas4-1 fusion enzymes as a potential route to establishing a single *in vitro* reaction for naïve adaptation. This would comprise DNA pre-processing, DNA capture and DNA integration in a single tube. The aim was to use Cas4-1 fusions from an archaeon (*Methanosaeta harundinacea)* and a bacterium (*Pyrinomonas methylaliphatogenes)*, but as detailed below, only the bacterial system produced manageable protein during this study.

Chapter 2: Materials and Methods

2.1 Chemicals

All chemicals were supplied by Merck or ThermoFisher Scientific unless stated otherwise.

2.1.1 Antibiotics

Table 1: Antibiotics used in this work. The working concentrations are the concentrations used for experiments**.**

Antibiotic	**Stock Concentration mg/ml**	**Working Concentration μg/ml**
Ampicillin	1	50
Kanamycin (Supplied by PanReac AppliChem)	0.5	25
Chloramphenicol	0.5	25

2.2 *E. coli* strains

Table 2: *E. coli* strains and genotypes as referred to in the main text.

***E. coli* strain**	**Genotype**
DH5α	*F- Φ80lacZΔM15 Δ(lacZYA-argF) U169 recA1 endA1 hsdR17(rk-, mk+) phoA supE44 thi-1 gyrA96 relA1 λ-*
BL21 AI	*F-ompT hsdSB (rB- mB-) gal dcm araB:T7RNAP-tetA*
BL21 CodonPlus(DE3)-RIL strain	*E. coli B F– ompT hsdS(rB– mB–) dcm+ Tetr gal λ(DE3) endA Hte [argU ile YleuW Camr]* Extra tRNAs available: argU (AGA, AGG), ileY (AUA), leuW (CUA)

2.3 Standard Buffers and Media

Table 3. Composition of general buffers

Buffer or reagent	Composition
10x TBE	1M Tris, 1M Boric Acid, 20mM EDTA (pH8.0)
4x SDS-PAGE loading Dye	80mM Tris-HCl (pH 6.8), 2.7% SDS, 15% glycerol (v/v), 3mg/ml Bromophenol blue
10x SDS-PAGE running buffer	250mM Tris, 1.92 Glycine, 1% (v/v) SDS
Coomassie blue stain	10% (v/v) Acetic Acid, 40% (v/v) Methanol, 1g/L Bromophenol Blue
SDS-PAGE Destain	10% (v/v) Acetic Acid, 40% (v/v) Methanol
10x TG buffer	250mM Tris, 0.5M Glycine
Transfer buffer	1x TG, 20% (v/v) Methanol
10x TBS	200mM Tris, 1.4M NaCl, pH7.6
TBST	1x TBS, 0.2% Tween 20
Blocking buffer	TBST + 5% (w/v) milk powder
10x Annealing buffer	100mM Tris-HCl (pH7.5), 500mM NaCl, 10mM EDTA (pH8)
Oligonucleotide Elution buffer	4mM Tris-HCl (pH8.0), 10mM NaCl
10x Binding buffer	7mM Tris-HCl (pH7.0), 9% (v/v) Glycerol, 50mM NaCl, 100μg/ml BSA, 5mM EDTA (pH8), Orange G to colour
EMSA Loading Dye	80% Glycerol + Orange G to colour
10x Nuclease buffer	7mM Tris-HCl (pH7.0), 9% (v/v) Glycerol, 50mM NaCl, 100μg/ml BSA, 5mM Mg^{2+}
STOP Buffer	1 mg/ml proteinase K, 2.5% w/v SDS

Table 4: Media used during research. All media was sterilised by autoclaving before use.

Media	**Composition**
LB	10g/L Tryptone, 10g/L NaCl, 5g/L Yeast Extract (Becton, Dickinson & Company)
Agar	LB, 15g/L Agar (VWR)

2.4 Commercial Enzymes

Table 5: Commercial enzymes used in DNA manipulation and molecular cloning. All enzymes were purchased from New England Biolabs (NEB) and were used as stated in the text.

Enzyme	**Cognate DNA sequence:_**indicates a site of phosphodiester bond cleavage
BamHI restriction endonuclease (RE)	5′ G_GATCC 3′ 3′ CCATG_G 5′
NdeI RE	5′ CA_TATG 3′ 3′ GTAT_AC 5′
EcoRI RE	5′ G_AATTC 3′ 3′ CTTAA_G 5′
HindIII RE	5′ A_AGCTT 3′ 3′ TTCGA_A 5′
SacI RE	5′ GAGCT_C 3′ 3′ C_TCGAG 5′
XbaI RE	5′ T_CTAGA 3′ 3′ AGATC_T 5′
KpnI RE	5′ GGTAC_C 3′ 3′ C_CATGG 5′
XhoI RE	5′ C_TCGAG 3′ 3′ GAGCT_C 5′
DpnI RE	5′ GA(CH_3)_TC 3′ 3′ CT_(CH_3)AG 3′
Vent DNA polymerase	
Q5 DNA polymerase	
Calf Intestinal Phosphatase (CIP)	
T4 Ligase	

2.5 Databases and Programs used to Obtain Sequences and Analyse Proteins and Gels.

2.5.1 Analysis of DNA and protein sequence information

Genome maps for *M. harundinacea* and *P. methylaliphatogenes* were accessed from Kyoto Encyclopedia of Genes and Genomes (KEGG) (https://www.genome.jp/kegg/) a collection of databases including genomes, gene, human diseases and orthology. These genome maps were used to obtain the local gene composition surrounding the CRISPR locus. The protein sequence for genes of interest surrounding the CRISPR locus were obtained from the Universal Protein Resource (UniProt) (https://www.uniprot.org/). Uniprot is a protein sequence database with annotations and protein amino acid sequences were obtained in a FASTA format. This represents the single letter amino acid code of the proteins in a text based format. The FASTA sequences were aligned using Clustal OMEGA where global alignment identified conserved regions in amino acid sequence.

Plasmid maps were generated using Snapgene (AddGene) by importing plasmid and gene sequence and simulating cloning steps to generate the end plasmid. Sequencing data was imported and aligned with plasmid maps to confirm successful cloning.

2.5.2 3D analysis of proteins

FASTA protein sequences were used with the Phyre2 molecular modelling program (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index) to create folding predictions for the tertiary structure of proteins. Protein alignment identified homologous protein with known structures. The structures of the homologous proteins in conserved regions were used to generate the model. Intensive modelling mode was used, which means any areas lacking conserved sequence and therefore structure models can be modelled using mathematical based simulated folding to generate a complete model. Structures were downloaded as PDB files and rendered in MacPymol.

Phyre2 model of Cas4-1 was entered into Galaxy Gemini (http://galaxy.seoklab.org/cgi-bin/submit.cgi?type=GEMINI) (Lee *et al.*, 2013), to generate a oligomeric structure. As with Phyre2, the Galaxy Gemini model is based

on homology. The structure is detected by HHsearch and compared to exisiting oligomeric states of proteins with a similar monomeric structure. Results were downloaded as PDB files and rendered using MacPymol.

2.5.3 Analysis of gels using ImageJ

EMSA and nuclease agarose gels were analysed using ImageJ. Bands are detected and intensity measured. These intensity values can then be compared to calculate binding and cleavage percentages.

2.5.8 Analysis of Phylogeny using MacVector

FASTA sequences for casposase and Cas1 sequences were entered into MacVector. Amino acid sequences were used instead of DNA sequences as the amino acid sequence is less diverged. As sequences were obtained from both archaea and bacteria, DNA sequences will naturally be divergent due to codon bias. Codon bias is the preference for a tRNA recognising a particular DNA triplet (codon). The use of the amino acid sequence eliminates the effects of codon bias. Amino acid sequences are also under more selective pressure compared to DNA sequences. A change in a single nucleotide can result in a silent mutation that does not change the amino acid sequence. Whereas the change of a single amino acid has the potential to be more detrimental. These factors allow us to align and analyse these divergent sequences more accurately.

After entry into MacVector, the sequences are aligned using the ClustalW alignment algorithm. This sequence alignment was used to generate a phylogenetic tree using neighbour-joining (Saitou and Nei, 1987) and with bootstrap replicates (Efron, 1979; Felsenstein, 1985) at 1000. Neighbour-joining calculates the distances between each sequence i.e. the number of changes between each sequence and the sequences with the smallest distances are paired together. The distance between each pair is then compared and the smallest distance pairs are also connected. The distance between nodes is based on the distance (sequence difference) between sequences. Once a phylogenetic tree has been constructed bootstrap replicates test the strength of each pair. For each bootstrap replicate a random sample of sequence is taken and this sample is used to create a new tree using neighbour joining. This is carried 1000 times, and nodes are given a

percentage value based on the percentage of times the node appeared in that position for the 1000 samples.

2.6 Oligonucleotides

Oligonucleotides used as primers for PCR or for preparation into model DNA substrates were sourced from Eurofins or Sigma. Oligonucleotides were supplied lyophilised and diluted in sterile distilled water (SDW) to generate a stock solution of 100μM, with 10μM used as the working stock.

2.6.1 Oligonucleotides for Polymerase Chain Reaction

The list of primers used in polymerase chain reaction (PCR) can be seen in Table 6. These primers were used to amplify DNA to clone ORFs into over-expression vectors, to confirm DNA sequences of clones and mutated genes and for use within NEB Q5 kit for mutagenesis via NEB-Base-changer. The primers are listed in pairs where 1 and 2 are the forward and reverse primers respectively. The table also states the gene name and the name of the plasmid created after cloning.

Table 6: Oligonucleotides for PCR. Oligonucleotide sequences for PCR are detailed along with the amplified gene and the resulting plasmid made from these genes after a cloning step.

Gene/Plasmid Name	**Oligonucleotide sequence 5'→3'**
cas4-1 (Mha)/ pEW7	1. GGGCATATGCTCGGAGTGCACGACCACGAAG 2. CCCGGATCCTCACCTCGTCAAAAAGG
polA/ pEW8	1. GCGAATTGGCGGAAGGCCGTCAA 2. GGAAAGCGGGCAGTGGCGGAAG
cas2 (Mha)/ pEW11	1. GCGGATCCAGAGGACGAAACTGCTATG 2. GCAAGCTTTCATACGATGATCGCATTC
polA/ pEW14	1. ATTGAGCTCAACCAAAAAAGTTCTGAG 2. ATTCTAGACTTAACTACGACCAATTGCAC
cas4-1 (Pme)/ pEW16	1. AATGGATCCTATGGCTGACGCGATCGCC 2. CCCGAATTCTCATCGTGTGCAGAAAGG
cas4-1 (Mha)/ pEW17	1. AATGAATTCATGGAGCCACCCGCAGTTCGAAAAGA TGCTGGGTGTTCATGATC 2. AATGAGCTCCCCATGAGGCCCAGGTCGAG
cas2 (Pme)/ pEW21	1. TAGGATCCGATGCGCAATCGTTACATT 2. GCCGCGAATTCTTAAACTATAATGGCGAT
Mutagenic C20S/ pEW23	1. TTCGCCTATTCCCCCCGGCTC 2. CTCGTTGAGCATGCGAGC

Mutagenic C212S/ pEW24	1. GTCGGCATCTCTCTACCGGATG 2. TAAAGAGCAACGCGGACA
Mutagenic E392A/ pEW25	1. CTCGGCGTCGCGGGCACAGCG 2. CAGCGTTGCCAGAGAGATTGCC
Mutagenic H462A/ pEW26	1. TGGCTTCTACGCTCAACCGAAATACG 2. AGATAAGGATCGAAGCCAAC
Mutagenic C203S/ pEW27	1. TCGCCAAAGTCTCCGCGTTGCTCTTTAGTCGG 2. AGCGACGAGCGGCGGCGG
HPS/ pEW28	1. CGGGTACCATGATTGAAGTAAAAGC 2. ATAGGATCCTTAGGGTTGTCGGGCT
Mutagenic K115A/ pEW29	1. CGTGGACTACGCACGCGGCTCG 2. GGGACGAGTTTTCCGCCA
Mutagenic D113A/ pEW30	1. GTCCCCGTGGCCTACAAACGC 2. GAGTTTTCCGCCATCGGATTC
Mutagenic E477A/ pEW31	1. GACTTGATGGCAGAATTTCGTCCGCTCATAGCAG 2. GAGCGCCAAAGCCGGGCG
HPL	1. AGGTACCATGCTGGAAGGCTGTCGAGAA 2. ATAGCTCGAGATTGCTGCATTCTTCTCTCCCC
Mutagenic H46A	1. AAAGATTCGTCGCCCGGCGCGTAG 2. CCTTCGAGCGTATCAACG
Mutagenic D100A	1. CGCGCGCATAGCTCTCATCGAATCCG 2. GATGGCGCCGAGGCGTTC
Mutagenic C206S	1. TGTCCGCGTTCCTCTTTAGTCG 2. CTTTGGCGAAGCGACGAG

2.6.2 Oligonucleotides for Creating Fluorescently-Labelled Substrates

Fluorescent labelled substrates were used to test the DNA binding and nuclease activity of purified protein. All substrates were labelled on the MW14 oligonucleotide on the 3’ or 5’ end with Cy5 as detailed within the work. Fluorescent labels absorb and emit energy and Cy5 the dye used in this research absorbs at 649nm and emits at 670nm. Substrates and the oligonucleotides used to create them are detailed in Table 7.

Table 7: Oligonucleotides used to create DNA substrates.

Substrate	Oligonucleotide name	Oligonucleotide sequence 5' → 3'
ssDNA	MW14	CAACGTCATAGACGATTACATTGCTACATGGAGC TGTCTAGAGGATCCGA
Linear Duplex	MW14	CAACGTCATAGACGATTACATTGCTACATGGAGC TGTCTAGAGGATCCGA
	EW3	TCGGATCCTCTAGACAGCTCCATGTAGCAATGTA ATCGTCTATGACGTTG
3' Overhang duplex	MW14	CAACGTCATAGACGATTACATTGCTACATGGAGC TGTCTAGAGGATCCGA
	EW2	TAGCAATGTAATCGTCTATGACGTTG
5' Overhang duplex	MW14	CAACGTCATAGACGATTACATTGCTACATGGAGC TGTCTAGAGGATCCGA
	EW1	TCGGATCCTCTAGACAGCTCCATG
Flayed Duplex	MW14	CAACGTCATAGACGATTACATTGCTACATGGAGC TGTCTAGAGGATCCGA
	MW12	TCGGATCCTCTAGACAGCTCCATGATCACTGGCA CTGGTAGAATTCGGC
Leading Strand Fork	MW14	CAACGTCATAGACGATTACATTGCTACATGGAGC TGTCTAGAGGATCCGA
	MW12	TCGGATCCTCTAGACAGCTCCATGATCACTGGCA CTGGTAGAATTCGGC
	PM16	TGCCGAATTCTACCAGTGCCAGTGAT
Lagging Strand Fork	MW14	CAACGTCATAGACGATTACATTGCTACATGGAGC TGTCTAGAGGATCCGA
	MW12	TCGGATCCTCTAGACAGCTCCATGATCACTGGCA CTGGTAGAATTCGGC
	PM17	TAGCAATGTAATCGTCTATGACGTTG
Fully Base Paired Fork	MW14	CAACGTCATAGACGATTACATTGCTACATGGAGC TGTCTAGAGGATCCGA
	MW12	TCGGATCCTCTAGACAGCTCCATGATCACTGGCA CTGGTAGAATTCGGC
	PM16	TGCCGAATTCTACCAGTGCCAGTGAT
	PM17	TAGCAATGTAATCGTCTATGACGTTG

2.6.3 Annealing of Oligonucleotides to Create DNA Substrates

Annealing reactions (50μl) containing 1x Annealing buffer, 5μM fluorescently labelled oligonucleotide and 6μM unlabelled oligonucleotide(s) were incubated at 95°C for 10 minutes and allowed to cool to 37°C before addition of 10μl EMSA loading dye. Substrates were loaded onto an appropriate percentage acrylamide

gel within a Protean II tank (Bio-Rad) and migrated through the gel for 3 hours at 120V. Fluorescent bands visible by eye were cut out of the gel and placed in oligonucleotide elution buffer. Substrates were left to diffuse out of the gel slice into the buffer for 48 hours. Elution buffer containing substrates was removed from the tube and concentrated using a speed-vac at 30°C until a volume of 50μl was reached. The concentration was calculated using the following equation:

$$C = \frac{A_{260}}{\varepsilon L}$$

Where C is the concentration of the DNA in M, A_{260} is the absorbance measured at 260nm using a NanoDrop™ spectrophotometer (Thermo Fisher Scientific), L is the path length (readings are taken at two path lengths: 1mm and 2mm, to obtain absorbance values normalised with a path length of 10mm) and ε is the extinction coefficient of the DNA. The extinction coefficient was calculated using the DNA sequence and an oligo analyser (http://www.idtdna.com/calc/analyzer).

2.7 Analysis of DNA and proteins by electrophoresis

2.7.1 Agarose gel electrophoresis.

DNA was visualised using agarose gels. 1 or 2% agarose was dissolved in 1x TBE with the addition of 600ng/ml ethidium bromide for DNA visualisation. Electrophoresis was carried out at 120V in 1x TBE using a Fisherbrand™ Maxi horizontal gel system. DNA was visualised using a Syngene U: genius 3.

2.7.2 SDS-PAGE

Coomassie blue stained SDS-PAGE gels were used to check expression of proteins and examine the presence of proteins in fractions during protein purification. Proteins were loaded onto a hand-cast SDS-PAGE gel and migrated in a Bio-Rad mini-PROTEAN apparatus at 120V in 1x SDS running buffer. Samples were prepared in 1x SDS-PAGE loading dye by boiling at 95°C for 10 minutes. Proteins were visualised by the addition of coomassie blue stain, followed by destaining with a SDS destain solution. Gels were imaged by camera on a white background.

2.7.3 Western Blot

Western blots were used to confirm that expressed or purified proteins were the proteins expected. Western blotting first separates proteins by their molecular weight, then provides specific identification by antibody binding (Burnette, 1981). SDS-PAGE separates proteins (see section 2.7.2) before transfer to a nitrocellulose membrane. This membrane is treated with antibodies creating a sensitive method for precise visualisation of individual proteins in a protein mixture.

Samples are treated as previously and run on an appropriate percentage acrylamide gel. This gel is then equilibrated in transfer buffer along with whatman paper, nitrocellulose membrane and sponges. These are all assembled together with the acrylamide gel and nitrocellulose membrane sandwiched adjacent one another within a holder. This is placed inside a trans-blot insert and run in a mini-protean tank with an ice block in transfer buffer at 60V for 2 hours. The nitrocellulose membrane is removed and placed in blocking buffer overnight at 4°C. Blocking buffer is removed through sequential washing by TBST before incubation in blocking buffer with 10μg/μl primary antibody (anti-histidine, invitrogen) for 2 hours at 4°C. Excess primary antibody is removed through sequential washing by TBST before incubation at 4°C in blocking buffer with 5μg/μl secondary antibody (anti-rabbit, invitrogen) for 2 hours. Final wash steps with TBST removed excess secondary antibody. The secondary antibody has horseradish peroxidase (HRP) attached to it, which when treated with Pierce™ ECL western blotting substrate (Thermo Scientific) according to manufacturer's instructions produces a chemiluminescent effect. This is then detected using a Fujifilm Las3000 mini.

2.7.4 Blue Native PAGE

Blue native PAGE (BN-PAGE) was used to investigate the oligomeric state of Cas4-1. BN-PAGE does not use SDS, instead the negative charge is provided by coomassie blue G-250 which binds to proteins non-specifically. The removal of SDS permits proteins to migrate in their native state. 3-12% NativePAGE™ Bis-Tris gels (Life Technologies) were used to visualise the protein. For buffer composition see manufacturer's instructions. Cas4-1 was mixed with 1x NativePAGE sample buffer and loaded onto a 3-12% gel inserted into a XCell SureLock™ Mini-Cell Electrophoresis System. Buffer containing coomassie blue G-250 was added to the

inner chamber with a non-coomassie blue G-250 buffer in the outer chamber. The gel was migrated for 95 minutes at 150V. The gel was destained overnight using SDS destain solution. The destained gel was imaged on a white background using a camera.

2.8 Molecular Cloning

Genomic DNA or GeneArt® synthesised plasmids, detailed below, were used to create several gene constructs detailed in this thesis. General protocols are discussed first, followed by the cloning protocol.

2.8.1 Source of Genomic DNA to clone relevant genes studied.

Genomic DNA was purchased from DSMZ (German collection of microorganisms and cell cultures) under DSM numbers 25857 and 17206 for *P. methylaliphatogenes* and *M. harundinacea* respectively. The genomic DNA was delivered in a buffer solution and concentration was calculated upon delivery using a Nanodrop.

2.8.2 GeneArt® customised DNA synthesis

GeneArt® custom DNA (Thermo Fisher Scientific) services (available at https://www.thermofisher.com/uk/en/home/life-science/cloning/gene-synthesis/geneart-gene-synthesis.html) is a service that synthesises genes or gene fragments which are difficult to clone or require codon optimisation. Cas4-1 and PolA from *M. harundinacea* were codon optimised to improve protein overexpression in *E. coli.* Codon optimised genes were supplied in a pMA-RQ plasmid.

2.8.3 Polymerase chain reaction to generate open reading frames (ORF) with restriction enzyme sites for cloning into a expression vector

To generate all plasmids, except the Site-direct mutants, 50μl PCR reaction mix was set up containing approximately 100ng of template DNA, 1x ThermoPol buffer, 200μM dNTPs, 0.4μM forward and reverse primers (see Table 6) and 0.02 U/μl Vent DNA polymerase. General PCR conditions are seen in Table 8, and specific

information about annealing temperatures and extension times are seen in Table 9.

Table 8: PCR cycling conditions. Steps in bold are cycled 25-30 times.

PCR stage	Temperature/°C	Time
Initial denaturation	95	90 seconds
Denaturation	**95**	**45 seconds**
Annealing	**50-70**	**60 seconds**
Extension	**72**	**60 seconds/kB**
Final extension	72	180 seconds

Table 9: Specific Annealing Temperatures and Extension Times for PCRs.

Gene/Plasmid	**Annealing Temperature/°C**	**Extension Time**
cas4-1 (Mha)/pEW7	61	180 seconds
polA/pEW8	66	120 seconds
cas2 (Mha)/pEW12	62	30 seconds
cas4-1(Pme)/pEW16	63	120 seconds
cas2 (Pme)/pEW21	57	30 seconds
HPS/pEW28	55	30 seconds
HPL	63	210 seconds

2.8.4 Site-direct mutagenesis (SDM)

Site-Directed Mutagenesis was carried out using Q5® Site-Directed Mutagenesis Kit (NEB). Primers were initially designed using NEBasechanger™ software to create the primers seen in Table 6. These primers were used with the mutagenesis kit according to manufacturer's instructions, with specific annealing temperatures and extension times detailed in Table 10. The final product was transformed into DH5α.

Table 10: Annealing temperatures and extensions times for SDM PCR.

Mutation/plasmid	**Annealing Temperature/°C**	**Extension Time/minutes**
Insert Strep Tag/pEW7	64	3.5
C20S/pEW16	67	3.5
C212S/pEW16	64	3.5
E392A/pEW16	72	3.5
H462A/pEW16	59	3.5
C203S/pEW16	72	3.5
K115A/pEW16	68	3.5
D113A/pEW16	67	3.5
E477A/pEW16	72	3.5
H46A/pEW16	63	3.5
D100A/pEW16	70	3.5
C206S/pEW16	65	3.5

2.8.5 Preparation of Chemically-competent *E. coli*

A 5ml overnight starter culture was grown in appropriate antibiotics, after inoculation of LB with *E. coli* grown on an agar plate. The overnight culture was used to inoculate 50ml of LB in a 1:100 dilution. The culture was grown with shaking at 37°C until an OD of 0.4-0.6 was reached. The cells were pelleted at 7,870*xg* using a 5430 Centrifuge (Eppendorf) for 10 minutes and the supernatant discarded. The pellet was resuspended in a $1/8^{th}$ volume of 0.1M $CaCl_{2,}$ before incubation on ice for 3 hours. The cells were pelleted again at 7,870*xg* for 10 minutes and the supernatant discarded. Pellet resuspension occurred with the same volume of 0.1M $CaCl_2$. Glycerol was added to 30% v/v before flash freezing in dry ice and storage at -80°C.

2.8.6 Transformation of chemical-competent *E. coli*

Plasmid DNA (~250ng) was added to 100μl of chemical-competent *E. coli* and incubated on ice for 30 minutes. The cells were heat-shocked at 42°C for 2 minutes, before incubation on ice for 2 minutes. LB was added to 1ml before incubation at 37°C with shaking. Cells were spun down at 10,000*xg* for 1 minute using mySPIN 12 centrifuge (ThermoFisher Scientific) and the supernatant discarded. The pellet was resuspended in 800μl of LB and between 200-500μl plated out on an agar plate containing the appropriate antibiotics.

2.8.7 Plasmid Purification

An overnight culture was set up by inoculating 6ml LB containing the appropriate antibiotics with a colony from an agar plate. The plasmid was then extracted from the culture using the QIAprep Spin Miniprep Kit (Qiagen) or Wizard® Minipreps DNA Purification System (Promega) according to manufacturer's instructions. Purified plasmids were stored at -20°C.

2.8.8 DNA sequencing

Sanger sequencing was used via SourceBioscience overnight service to verify gene sequences. Sequencing was carried out using standard primers for the plasmid. Sequencing data was uploaded into Snapgene and aligned with plasmid maps to confirm successful cloning or mutation.

2.8.9 Cloning of Recombinant DNA

ORFs were amplified from genomic DNA or GeneArt plasmids as specified in the main text. 10μl PCR product or plasmid backbone was digested in a 20μl reaction with 0.5U/μl of each restriction enzyme, 1x CutSmart buffer and 0.5U/μl of CIP for the plasmid backbone. The restriction enzymes and plasmid backbone for each cloning reaction are stated in Table 11. The addition of CIP dephosphorylates the plasmid backbone to prevent re-ligation of the plasmid. 2μl of 5x gel loading dye was added to each restriction enzyme reaction and loaded onto an appropriate percentage agarose gel. For information about gel electrophoresis, please see section 2.7.1. The bands were visualised using a fluorescent box and cut out with a scalpel.

DNA was extracted from the gel slices using QIAquick gel extraction kit (Qiagen) according to manufacturer's instructions. The digested plasmid and PCR product were ligated together in a 1:4 ratio along with 1x T4 ligase buffer and 0.01U/μl T4 ligase in a 30μl reaction overnight at 16°C. The ligated product was transformed into DH5α cells and plated out on agar plates with appropriate antibiotics. Resulting colonies were used to set up overnight cultures, and plasmids were extracted, see section 2.8.7. Restriction digests were carried out as above to check for correct digestion patterns indicating a combination of plasmid and PCR product. Plasmids with the correct digestion patterns were sent for sequencing, see section 2.8.8. Plasmids confirmed as correct by sequencing, were stored at -20°C ready for use.

Table 11: Plasmid backbone and restriction enzymes used to create plasmids.

Gene/Plasmid name	Plasmid Backbone	Restriction Enzymes
cas4-1/pEW7	pET14b	BamHI, NdeI
polA/pEW8	pETDUET	BamHI, EcoRI
cas2/pEW12	pACYCDUET	BamHI, HindIII
polA/pEW14	pHTN HaloTag® CMV-neo	SacI, XbaI
cas4-1/pEW16	pETDUET	BamHI, EcoRI
cas2/pEW21	pACYCDUET	BamHI, EcoRI
HPS/pEW28	pRSF1-B	Kpn1, BamHI
HPL	pETDUET	Kpn1, XhoI

2.9 Protein Expression

Expression of proteins was initially tested in a pilot overexpression using a 50ml culture. This allowed testing of two different expression strains BL21AI and BL21C+. BL21AI (arabinose inducible) contains a T7 RNA polymerase under control of the araBad promoter. Expression of T7 RNA polymerase is induced upon the addition of L-arabinose. This helps reduce leaky expression i.e. expression of proteins without addition of inducers. BL21C+ (codon plus) are also based upon BL21, but contain a plasmid that encodes tRNAs rare in *E. coli*. This increases expression of recombinant genes in *E. coli* that are limited by codon bias. All proteins expressed in this research are from organisms other than *E. coli* and may require rare tRNAs for expression. The strain that best expressed the protein tested was used for scaled-up overexpression. The procedures to overexpress each protein are described individually below. Resuspension buffers contained either PMSF or a EDTA-free protease cocktail inhibitor (cOmplete™ EDTA-free, Roche) and resuspension buffer composition are stated in Table 12.

2.9.1 His-tagged Cas4-1 (*Methanosaeta harundinacea*)

To over-express N-terminally hexahistidine tagged Cas4-1, pEW7 encoding Cas4-1 was transformed into chemically competent BL21C+ *E. coli.* A single colony was taken to inoculate an overnight starter culture containing 50μg/ml ampicillin and 25μg/ml of chloramphenicol. 6L of LB containing 50μg/ml ampicillin and 25μg/ml of chloramphenicol was inoculated 1:100 from the starter culture and grown at 37°C with shaking until reaching an OD (at λ600) of 0.5-0.7. Protein expression was induced using 0.5mM IPTG and was expressed for 18 hours at 18°C with shaking. Cells were harvested at 7680*xg* using Sorvall RC3C plus centrifuge with H6000 rotor for 35 minutes. The supernatant was removed and the pelleted cells were resuspended in Buffer A. Resuspended cells were stored at -80°C.

2.9.2 His-tagged DNA Polymerase I (*Methanosaeta harundinacea)*

To over-express N-terminally hexahistidine tagged PolA, pEW8 encoding PolA was transformed into chemically competent BL21AI *E. coli.* A single colony was used to inoculate an overnight starter culture containing 50μg/ml ampicillin. This starter culture was used to inoculate 1:100, 4L LB containing 50μg/ml ampicillin. This culture was grown at 37°C until OD 0.5-0.7 and then incubated at 18°C for 18

hours. The protein expressed without inducers, due to leaky expression. Expression is usually induced by the binding of inducers, but RNA polymerases can still bind to promoters in the plasmid and begin to transcribe the ORF. This generally leads to low expression of protein before inducing, but here enough protein was expressed for purification. Cells were harvested as detailed above and resuspended in buffer B, before storage at -80°C.

2.9.3 His-tagged Cas2 (*Methanosaeta harundinacea*)

Overexpression of Cas2 with a N-terminally hexahistidine tag required the transformation of pEW11 (encodes Cas2) into chemically competent B21AI *E. coli.* An overnight starter culture containing 50μg/ml ampicillin was inoculated with a single colony. 2L LB containing 50μg/ml ampicillin was inoculated 1:100 by the starer culture and grown at 37°C until OD 0.5-0.7. Protein overexpression was induced by the addition of 0.2% arabinose and 0.5mM IPTG. The culture was left to express protein at 37°C for 3 hours. Cells were harvested as stated above and resuspended in buffer B before storage at -80°C.

2.9.4 Streptavidin-tagged Cas4-1 (*Methanosaeta harundinacea*)

Cas4-1 (Step) overexpression started with the transformation of pEW17 into chemically competent BL21C+ *E. coli.* An overnight startedr culture containing 50μg/ml ampicillin and 25μg/ml chloramphenicol was inoculated with a single colony. 6L LB containing 50μg/ml ampicillin and 25μg/ml chloramphenicol was inoculated 1:100 by the starter culture and grown at 37°C until OD 0.5-0.7. Protein overexpression was induced by 0.5mM IPTG. The culture was left to express protein at 18°C for 18 hours. Cells were harvested as stated above and resuspended in Buffer C before storage at -80°C

2.9.5 Halo-tagged PolA (*Methanosaeta harundinacea)*

PolA (Halo) expressing pEW14 was transformed into chemically competent BL21C+ *E. coli.* A single colony was used to inoculate an overnight starter culture containing 50μg/ml ampicillin and 25μg/ml chloramphenicol. This overnight culture was used to inoculate 1:100 4L LB containing 50μg/ml ampicillin and 25μg/ml chloramphenicol which was grown at 37°C until OD 0.5-0.7. Protein overexpression was induced by 0.5mM IPTG. The culture was incubated to express protein at 18°C

for 18 hours. Cells were harvested as stated above and resuspended in Buffer D before storage at -80°C.

2.9.6 His-tagged Cas4-1 (*Pyrinomonas methylaliphatogenes*)

To over-express N-terminally hexahistidine tagged Cas4-1, pEW16 encoding Cas4-1 was transformed into chemically competent B21AI *E. coli.* An overnight culture containing 50μg/ml ampicillin was inoculated with a single colony. 6L of LB containing 50μg/ml ampicillin was inoculated 1:100 from the starter culture and grown at 37°C with shaking until reaching an OD (at λ600) of 0.5-0.7. Protein expression was induced using 0.2% L-arabinose and 0.5mM IPTG before expression for 18 hours at 18°C with shaking. Cells were harvested as stated above and resuspended in buffer E. Resuspended cells were stored at -80°C.

2.9.7 His-tagged Cas2 (*Pyrinomonas methylaliphatogenes*)

Overexpression of Cas2 with a N-terminally hexahistidine tag required the transformation of pEW21 (encodes Cas2) into chemically competent B21AI *E. coli.* An overnight starter culture containing 25μg/ml chloramphenicol was inoculated with a single colony. 6L LB containing 25μg/ml chloramphenicol was inoculated 1:100 by the starter culture and grown at 37°C until OD 0.5-0.7. Protein overexpression was induced by the addition of 0.2% arabinose and 0.5mM IPTG. The culture was left to express protein at 18°C for 18 hours. Cells were harvested as stated above and resuspended in buffer B before storage at -80°C.

2.9.8 His-tagged HPS (*Pyrinomonas methylaliphatogenes)*

HPS expressing pEW28 was transformed into chemically competent BL21AI *E. coli.* A single colony was used to inoculate an overnight starter culture containing 25μg/ml kanamycin. This overnight culture was used to inoculate 1:100 4L LB containing 25μg/ml kanamycin which was grown at 37°C until OD 0.5-0.7. Protein overexpression was induced by 0.2% L-arabinose and 0.5mM IPTG. The culture was incubated to express protein at 18°C for 18 hours. Cells were harvested as stated above and resuspended in Buffer B before storage at -80°C.

Table 12: Composition of resuspension buffers used for protein expression.

Resuspension Buffer	Composition	Proteins Used For
A	20mM Tris pH7.5, 20mM Imidazole, 0.5M NaCl, EDTA-free protease cocktail inhibitor	Cas4-1
B	20mM Tris pH7.5, 20mM Imidazole, 0.5M NaCl, PMSF	Cas2, PolA, Cas2 (Pme), HPS
C	100mM Tris pH7.5, 150ml NaCl, EDTA-free protease cocktail inhibitor	Cas4-1 (Strep)
D	50mM HEPES pH 7.5 and 150mM NaCl, PMSF	PolA (Halo)
E	20mM Tris pH7.5, 20mM Imidazole, 150mM NaCl, EDTA-free protease cocktail inhibitor	Cas4-1 (Pme)

2.10 Protein purification

Protein purifications were carried out at room temperature and unless stated otherwise columns were obtained from GE healthcare. All buffers (see Table 13 for composition) were filtered and degassed before use in protein purification. Purifications were carried out on an AKTA start system, unless otherwise stated. Prior to chromatography, biomass for protein was removed from the -80°C, thawed and lysed by sonication at 80% amplitude in 10 second pulses using a Vibra Cell sonicator (Jencons). The lysate was clarified by centrifugation at 22005*xg* for 45 minutes using Beckman coulter Avanti J-25 with the JA25.50 rotor.

2.10.1 Histidine-tagged Proteins (minus Cas4-1, *Pyrinomonas methylaliphatogenes*)

All histidine tagged proteins (excepting *P. methylaliphatogenes* Cas4-1) were all purified with the same initial column. Clarified lysate was loaded onto a 5ml HiTrap chelating HP column pre-equilibrated in Buffer A. The column was thoroughly washed with buffer A to remove unbound proteins. Proteins were eluted over 12 column volumes (CV) using an imidazole gradient to 100% Buffer B. Fractions containing Cas4-1 were dialysed overnight at 4°C in dialysis buffer. The remaining purification procedures are detailed below.

2.10.1.1 Cas4-1 (*Methanosaeta harundinacea*)

Dialysed protein was loaded onto a cibacron blue gravity column pre-equilibrated in Buffer C. Bound proteins were thoroughly washed with Buffer C. Proteins were eluted by addition of 5 CVs of Buffer D. Fractions containing Cas4-1 were dialysed overnight at 4°C. Protein concentration was calculated using Bradford's reagent (see section 2.11), before aliquoting and flash freezing in dry ice and storage at -80°C.

2.10.1.2 Cas2 (*Methanosaeta harundinacea*)

Dialysed protein was loaded onto a 1ml HiTrap SP sepharose FF pre-equilibrated by Buffer C. Cas2 was eluted through washing by Buffer C. The wash solution containing Cas2 was dialysed overnight at 4°C in dialysis buffer. Concentration of protein after dialysis was calculated using a Bradford assay, before aliquoting, flash freezing in dry ice and storage at -80°C.

2.10.1.3 DNA polymerase I (*Methanosaeta harundinacea*)

Dialysed proteins were loaded onto a HiTrap Q FF column pre-equilibrated in Buffer C. Bound proteins were washed with Buffer C and eluted over 10 CVs via a salt gradient to 100% buffer D. PolA containing fractions were dialysed for 2 hours at 4°C in dialysis buffer. The concentration of dialysed PolA was calculated using a Bradford's assay. PolA was then aliquoted, flash frozen in dry ice and stored at -80°C

2.10.1.4 Cas2 (*Pyrinomonas methylaliphatogenes*)

The concentration of dialysed protein was calculated using a Bradford assay, before aliquoting, flash freezing in dry ice and storage at -80°C.

2.10.1.5 HPS (*Pyrinomonas methylaliphatogenes)*

The concentration of dialysed protein was calculated using a Bradford assay, before aliquoting, flash freezing in dry ice and storage at -80°C.

2.10.2 Streptavidin-tagged Cas4-1 (*Methanosaeta harundinacea*)

Clarified lysate was loaded onto a 5ml StrepTrap HP column pre-equilibrated in Buffer E. Bound proteins were thoroughly washed with Buffer E before elution at

100% Buffer F for 5 CVs. Fractions containing Cas4-1 were dialysed overnight at 4°C in dialysis buffer. Dialysed protein was loaded onto a 1ml HiTrap Heparin HP column pre-equilibrated in Buffer C. Bound proteins were washed thoroughly with Buffer C before elution over 10 CVs via a salt gradient to 100% buffer C. Cas4-1 containing fractions were dialysed overnight at 4°C in dialysis buffer. The concentration of dialysed Cas4-1 was calculated using a Bradford's assay. Cas4-1 was then aliquoted, flash frozen in dry ice and stored at -80°C

2.10.3 Halo-tagged DNA Polymerase I (*Methanosaeta harundinacea)*

Two similar purifications were attempted to purify PolA using the HaloTag® Protein Purification System (Promega). Clarified lysate was added to a gravity column containing HaloLink™ resin pre-equilibrated with Buffer G. For the first purification, the mixture was inverted a few times and left to incubate at 4°C for 10 minutes. The unbound lysate solution was then allowed to flow through the column and bound proteins washed with Buffer G. Buffer H was added and the mixture was inverted 3 times to mix and incubated for 10 minutes at 4°C. The now unbound protein was then allowed to flow through the column.

During the second purification, the mixture was left to mix on a rotating wheel for 2 hours at 4°C. The unbound lysate was flowed through the column and bound proteins washed with Buffer G. Resin and bound proteins were mixed with Buffer H for 2 hours at 4°C to elute proteins which were then allowed to flow through. PolA was not eluted successfully with either method.

2.10.4 His-tagged Cas4-1 (*Pyrinomonas methylaliphatogenes*) and Associated Mutants

In the initial purification clarified lysate was loaded onto a 5ml HiTrap chelating HP column pre-equilibrated in Buffer A. The column was thoroughly washed with buffer A to remove unbound proteins. Proteins were eluted over 12 CVs (column volumes) using an imidazole gradient to 100% Buffer B. Fractions containing Cas4-1 were dialysed overnight at 4°C in dialysis buffer. Dialysed protein was loaded onto a 1ml HiTrap Heparin HP column pre-equilibrated in Buffer C. Bound proteins were thoroughly washed with Buffer C. Proteins were eluted by addition of 5 CVs of Buffer D. Fractions containing Cas4-1 were dialysed overnight at 4°C. Protein

concentration was calculated using Bradford's reagent, before aliquoting and flash freezing in dry ice and storage at -80°C.

This original purification was achieved over 3 days, but during this time Cas4-1 protein degraded so the protocol was optimised to allow purification within 1 day. For the optimised purification clarified lysate was loaded onto a 5ml HiTrap chelating HP column pre-equilibrated in Buffer I. The column was thoroughly washed with Buffer I to remove unbound proteins. Proteins were eluted over 12 CVs using an imidazole gradient to 100% Buffer J. Fractions containing Cas4-1 were loaded onto a 1ml HiTrap Heparin HP column pre-equilibrated in Buffer J. Bound proteins were thoroughly washed with Buffer J. Proteins were eluted by addition of 5 CVs of Buffer K. Buffer conditions of pooled fractions from the heparin column were exchanged using a PD10 column pre-equilibrated with dialysis buffer. The concentration of the buffer exchanged Cas4-1 was calculated using a Bradford before aliquoting and storage at -80°C.

Table 13: Purification Buffer Composition.

Buffer	Composition
A	20mM Tris pH7.5, 20mM Imidazole, 0.5M NaCl
B	20mM Tris pH7.5, 1M Imidazole, 0.5M NaCl
Dialysis	20mM Tris pH 7.5, 150mM NaCl, 35% Glycerol
C	20mM Tris pH7.5, 150mM NaCl
D	20mM Tris pH7.5, 1M NaCl
E	100mM Tris pH7.5, 150ml NaCl
F	100mM Tris pH7.5, 150ml NaCl, 25mM Desthiobiotin
G	50mM HEPES pH 7.5 and 150mM NaCl.
H	50mM HEPES pH 7.5, 150mM NaCl, 1μM TEV protease
I	20mM Tris pH7.5, 20mM Imidazole, 150mM NaCl
J	20mM Tris pH7.5 0.5M Imidazole, 150mM NaCl
K	20mM Tris pH7.5, 0.5M Imidazole, 1MNaCl

2.11 Bradford's Assay for estimation of protein concentrations

A standard curve was calculated using BSA (Bovine Serum Albumin). Concentrations ranging from 1mg/ml to 15mg/ml of BSA were diluted in Bradford's reagent. This mix was incubated at RT for 20 minutes before measuring the absorbance at 595nm. Absorbance values were plotted against concentration to

create a standard curve. Between 2-20μl of protein sample was added to Bradford's reagent and incubated at RT for 20 minutes. Absorbance was measured at 595nm and plotted on the standard curve to calculate concentration.

2.12 Electromobility Shift Assay (EMSA) for in-gel analysis of protein-DNA binding

An appropriate percentage acrylamide gel was hand cast and loaded into a Protean II tank (Bio-Rad). 25nM fluorescently labelled substrate, 1xbinding buffer and a range of protein concentrations were mixed and incubated at 37°C for 10 minutes. EMSA loading dye was added and the samples were loaded onto the acrylamide gel. Gels were migrated for 3 hours at 120V before imaging on FLA 3000 (Fujifilm). Gel images were analysed in ImageJ.

2.13 Nuclease Assays

250ng M13 or 200ng pUC18 was incubated with 1x nuclease buffer and a range of protein concentrations at 37°C for 2 hours. The reaction was stopped by the addition of STOP buffer and incubated for a further 10 minutes. EMSA loading dye was added and the samples were loaded and run on a 1% agarose gel for 1 hour at 120V. The gel was stained with 600ng/ml ethidium bromide in 1x TBE for 30 minutes before imaging on Syngene U: genius 3. Gels were analysed by ImageJ and graphs created using Prism. For Cas4-1 and Cas2 nuclease assays, Cas4-1 and Cas2 were incubated on ice together for 30 minutes before the experiment.

2.14 Analytical Gel Filtration for assessment of protein oligomeric state

Analytical gel filtration (AGF) uses size exclusion chromatography through a bed of densely packed porous beads to assess protein oligomeric state. Small proteins/complexes diffuse into the beads causing them to flow through the column slowly, whereas large proteins/complex cannot diffuse into the beads or diffuse less causing them to flow quickly through the column. This separates the proteins based on size.

To assess the protein oligomeric state, a 1mg protein sample was injected onto a Superose 200 increase column that pre-equilibrated in 20mM Tris pH7.5 containing 150mM NaCl using an AKTA Start (GE Healthcare). GE Healthcare HMW calibration standards (28403841) were used to create a standard line, to calculate the sample molecular weight.

For the standard line the gel phase distribution coefficient (K_{av}) was plotted against the log of molecular weight. The K_{av} is the relationship between the void volume, elution volume and column volume. Using the K_{av} and log of molecular weight creates a more reliable standard line than simply plotting the elution volumes against the molecular weight. The K_{av} is calculated using the following equation:

$$\mathrm{Kav} = \frac{\mathrm{Ve}-\mathrm{Vo}}{\mathrm{Vc}-\mathrm{Vo}}.$$

Ve is the elution volume of each standard or sample. Vo is the void volume. Each sample takes a certain amount of time to pass through the column and the void volume is the volume at which the first protein could elute. Vc is column volume, which in this case was 120ml.

2.15 Spacer Integration (spIN) assay

spIN assay was used in conjunction with *P. methylaliphatogenes* protein and CRISPR locus. Proteins were incubated on ice for 30 minutes in 1x nuclease buffer (Cas4-1, Cas2 and HPS) as stated for each experiment. 200ng of pUC18 and pCRISPR (GeneArt plasmid containing the CRISPR locus) were added to each of the protein mixtures and incubated at 37°C for 2 hours. Reactions were stopped by the addition of STOP buffer and incubated for a further 10 minutes. EMSA loading dye was added and the samples were loaded and run on a 1% agarose gel for 1 hour at 120V. The gel was stained with 600ng/ml ethidium bromide in 1XTBE for 30 minutes before imaging on Syngene U: genius 3.

Chapter 3: Bioinformatics, Molecular Cloning and Purification of *Methanosaeta harundinacea* adaptation proteins (Cas4-1, Cas2) and associated DNA polymerase I.

3.1 Introduction

In chapter one, the CRISPR-Cas mechanism was described as two linked parts: adaptation and interference. Interference is the better understood stage, and while the general mechanism of adaptation is understood the process to generate protospacers is still unknown. Protospacers can be generated in a naïve way potentially involving RecB and Cas1-Cas2 or in a primed/targeted way involving interference proteins and potentially RecG. But the actual mechanism of protospacer generation is not known, and most proteins with predicted involvement generate ssDNA fragments whereas a spacer is required to be ds.

Cas4-1 is a natural fusion of Cas4 and Cas1 and is implicated in adaptation. Cas4 has been shown to be required for adaptation in both *S. islandicus* and *H. hispanica* (Li *et al.*, 2014; Liu *et al.*, 2017) and can form a complex with Cas1 and Cas2 (Plagens *et al.*, 2012). This along with the fact Cas4 is often found fused to Cas1 in numerous organisms (Koonin *et al*. 2017; Makarova *et al*. 2015) and Cas4 is always found encoded in the same locus as Cas1 (Hudaiberdiev *et al.*, 2017) shows Cas4 is involved in adaptation with a connecting function with Cas1. Cas4 can cleave ssDNA in a 5'-3' direction and can cleave dsDNA with 5' overhangs. This has led to a hypothesised role for Cas4 in adaptation in generating protospacers or in cleaving protospacers to remove or create overhangs for integration (Zhang, Kasciukovic and White, 2012; Lemak *et al.*, 2013; Sofia Lemak *et al.*, 2014).

*Methanosaeta harundinace*a is a archaeal methogen, identified in Beijing, China (Ma, Liu and Dong, 2006). Sequencing of the genome (Zhu *et al.*, 2012) showed a CRISPR-Cas system with three proteins predicted to be involved in adaptation Cas4-1, Cas2 and PolA. This system was selected to try to establish a single *in vitro* reaction for naïve adaptation using the linked activities of these different enzymes.

To achieve this goal, purification of the protein was required. Here, bioinformatic analysis, molecular cloning, protein-overexpression and purification of adaptation proteins are described in conjunction with biochemical assays.

3.2 The gene neighbourhood of *Methanosaeta harundinacea* CRISPR-Cas

The *M. harundinacea* CRISPR-Cas gene neighbourhood contains a *cas4-1* fusion located adjacent to a *polA* ORF. This arrangement was comparable to the arrangement within the casposon system that contains both Cas1 and a polymerase. This was also unusual as *polA* is a rare gene in archaea and a BLAST search shows only two other archaea with this gene present: *Methanosaeta concilii* and *Methanomethylophilus altus.* Due to the close proximity of *polA* and *cas1, polA* may be involved in gap filling after integration. To study the constitution of the full CRISPR neighbourhood the genome map was accessed through Kegg (accessed at https://www.genome.jp/kegg/). As shown in Figure 14, *cas* genes are positioned adjacent to each end of the CRISPR locus. Genes associated with interference are located downstream, whereas upstream are the adaption ORFs of *cas4-1*, *cas2* and *polA*.

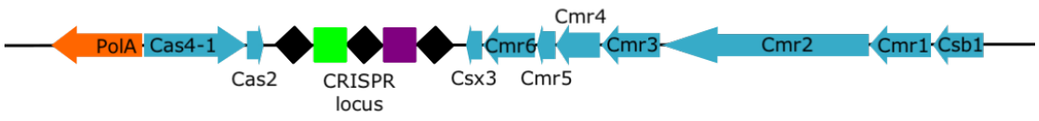

Figure 14: *M. harundinacea* CRISPR locus and *cas* gene neighbourhood. A central CRISPR locus separates *cas* genes hypothesized to be involved in adaptation (*cas4-1, cas2* and *polA*) from the *cas* gene predicted to be involved in interference (*csx3, cmr6, cmr5,cmr4, cmr3, cmr2, cmr1* and *csb1*).

Having established the local gene neighbourhood, the adaptation proteins (Cas4-1, Cas2 and DNA polymerase I) where investigated further using sequence alignments and fold predictions, to study conserved active site residues and folds. This would confirm that *M. harundinacea* proteins were correctly annotated and had the capacity to function as predicted.

3.3 Bioinformatic analysis of *M. harundinacea* ORFs

3.3.1 Identification of Conserved Residues Though Sequence Alignment Using Clustal Omega

3.31.1 Sequence alignment of Cas4-1

Cas4-1 as a fusion protein contains structural and functional aspects of both Cas1 and Cas4, but the extent of the homology of each region was unknown. Figure 15 shows alignment of the amino acid sequence of *M. harundinacea* Cas4-1 with Cas4 sequences (*S. solfataricus* and *P. calidifontis)* and Cas1 sequences (*E. coli, S. islandicus* and *M. tuberculosis). S. solfataricus* Cas4 proteins (SSO0001 and SSO1391) along with *P. calidifontis* Cas4 were selected as the 4Fe:4S cluster cysteine residues and active site residues were identified through research. These known active site residues could be used to find conserved active site residues within Cas4-1. Cas1 protein sequences from *E. coli, S. islandicus* and *M. tuberculosis* were also selected as research had also established active site residues within these proteins. In the same way as Cas4, the Cas4-1 active site residues could be found through looking for conservation of these resides from Cas1.

Cas4 homology was restricted to the N-terminal region with Cas1 homology at the C-terminal region. The N-terminal Cas4 region had conserved residues corresponding to two functional sites: a RecB-like nuclease site (H121, D158, E171 and K173) and a 4Fe:4S cluster (C95, C258, C261 and C267). The C-terminal Cas1 had conserved residues for a Cas1 active site (E485, H552 and E567). Between the two regions was a short sequence with no homology to Cas4 or Cas1. This sequence was composed of 33% glycine residues. As glycine residues only have a hydrogen atom as an R-group, they provide flexibility in a chain. Natural linkers include glycine in combination with other amino acid residues and the percentage of glycine present dictates the flexibility of the linker (Yan and Sun, 1996; Van Rosmalen *et al*, 2017). This sequence is glycine rich (33%) so has the potential to form a flexible linker that could allow the Cas4 and Cas1 active sites to function independently apart or come together to function. To confirm this region as a flexible linker mutation of the residues to large more rigid residues (tryptophan, phenylalanine or proline) or removal of the linker altogether could be carried out to test the ability of Cas4-1 to function. At the outset of this PhD project, it was hypothesised that Cas4-1 can carry out the functions of both Cas4 and Cas1 independently.

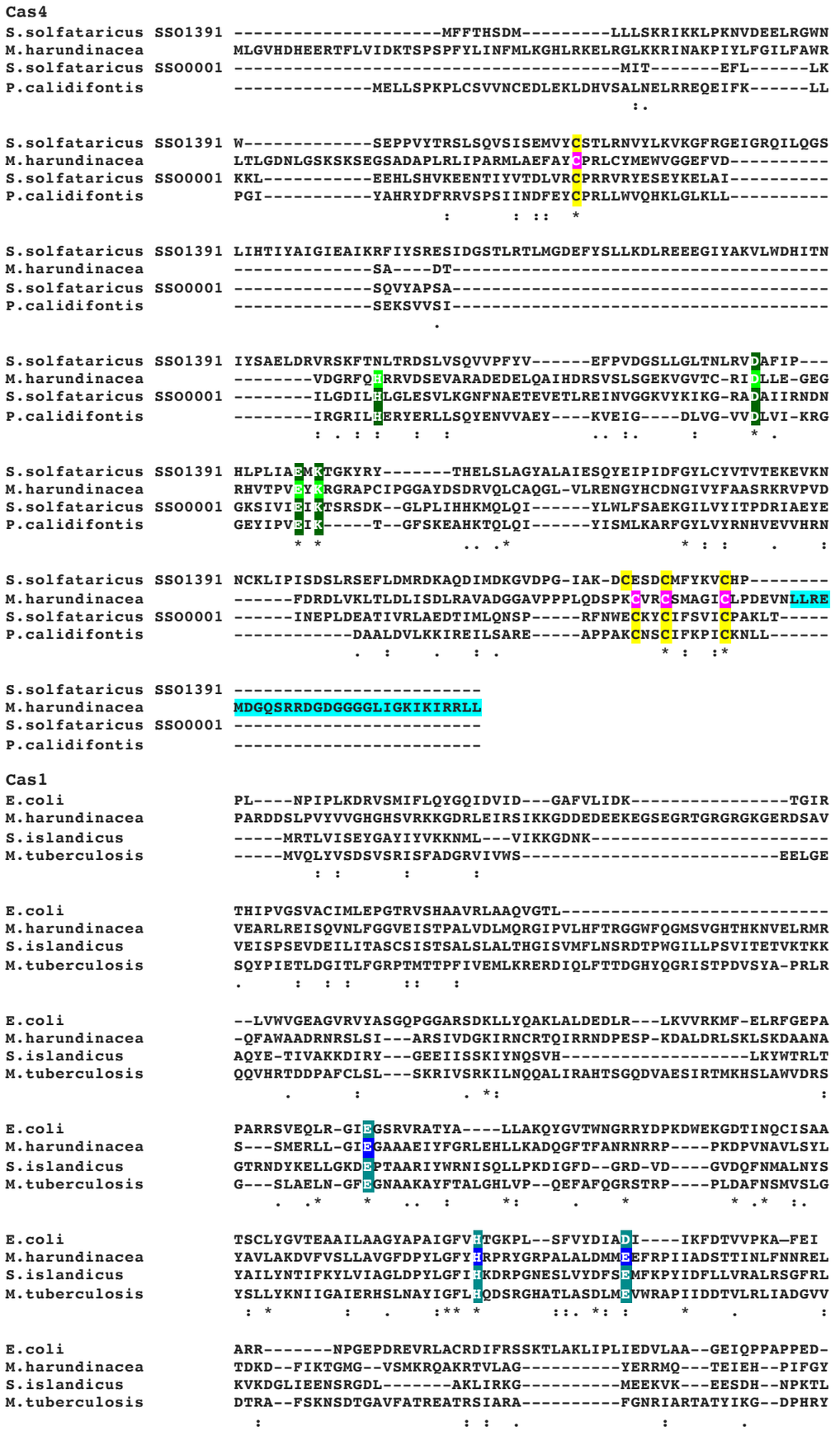

Figure 15: Homology of amino acid sequences for Cas4-1, Cas4 and Cas1. Cas4 homology was shown with alignment with Cas4 sequences from *S. solfataricus* and *P. calidifontis*. This homology was exclusively at the N-terminal with conserved residues for the Cas4 active site (green) and 4Fe:4S cluster (yellow and purple). Cas1 was aligned with *E. coli, S. islandicus* and *M. tuberculosis* and showed homology solely at the C-terminal and contained conserved residues corresponding to the Cas1 active site (dark blue). No homology to Cas1 or Cas4 was established for a short glycine rich sequence between the two regions (light blue) which could represent a linker region.

3.3.1.2 Cas2 sequence alignment

M. harundinacea Cas2 was aligned with Cas2 amino acid sequences from *E. coli, S. solfataricus* and *D. vulgaris* (Figure 16). *E. coli* Cas2 was chosen as it has been extensively studied. *S. solfataricus* Cas2 was chosen as it has been shown to have RNase activity (Beloglazova *et al.*, 2008) and the active side residues could be used to look for conserved residues in *M. harundinacea* Cas2. The final Cas2 sequence used was from *Desulfovibrio vulgaris*, which has been shown to lack RNase activity due to the lack of a conserved arginine (R17 in *S. solfataricus* highlighted in purple*)*. If *M. harundinacea* also lacked this residue, it would be unlikely to have RNase activity. As can be seen this residue is missing from *M. hardunaincaea's* Cas2.

```
E.coli          ----MSMLVVVTEN------VPPRLRGRLAIWL-----LEVRAGVYVGDVSAKIREMI--
S.solfataricus  ----MAMLYLIFYDITDDN-----LRNRVAEFLKKKGLDRIQYSVFMGDLNSSRLKDVEA
M.harundinacea  MRGRNCY--VVSYD-----IMEPRRLQKVHKMM-KGFGDPVHYSVFRCDLTPKGRVEMIA
D.vulgaris      MYGNDAMLVLISYDVSFEDPGGQRRLRRIAKAC-QDYGQRVQYSVFECVVDPAQWAKLKH
                     .   ::  :             ::           :: .*:   :       :

E.coli          ----------------------------WE------QIAGLAEEGNVVMAWATNTETGFE
S.solfataricus  GLKIIGNRKKLQEDERFFILIVPITENQFRERIVI-GYSGSEREEKSNVVW---------
M.harundinacea  ALT--GLIKH-DEDRVMIIDLGP-VEGMAEDRIEFLGVHSPKEKENAIIV----------
D.vulgaris      RLL--SEMDK-EKDCLRFYYLGA-N---WRNKVEHVGAKPAYDPEGPLIL----------
                                          .                 :

E.coli          FQTFGLNRRTPVDLDGLRLVSFLPV
S.solfataricus  -------------------------
M.harundinacea  -------------------------
D.vulgaris      -------------------------
```

Figure 16: Sequence alignment of *M. harundiancea* Cas2 amino acid sequences with *E. coli* and *S. solfactaricus*. Residues coordinating RNase activity in S*. solfataricus* are distinguished in yellow. Four of these residues are conserved in the *M. harduninacea* sequence: Y11, D12, D30 and F37. The crucial missing residue from *D. vulgaris* is also missing in *M. harundinacea.*

3.3.1.3 Sequence alignment of DNA polymerase I (*PolA*)

DNA polymerase I is a Family A polymerase a family not typically found in archaea, usual families found in archaea are Family B and Family D (Cann *et al.*, 1998; Cann and Ishino, 1999). Homology was therefore determined with bacterial sequences (*Helicobacter pylori, Thermus thermophilus, E. coli* and *Haemophilus influenzae)* with established active site residues, shown in Figure 17. DNA polymerase I typically contains three active regions a 5'-3' exonuclease, 3-5' exonuclease and polymerase. The exonuclease regions are excluded from the sequence alignment diagram due to length of sequence and lack of homology in the 5'-3' exonuclease region. There was homology to the 3'-5' exonuclease region (not shown for conciseness), meaning the polymerase may have the capacity to proofread but not to remove primers. The alignment for the polymerase region (red text) is shown in Figure 17. Though *M. harundinacea* DNA polymerase I aligns with the 3'-5'

exonuclease and polymerase regions that does not guarantee functionality. Polymerases contain three highly conserved regions: motif A (blue highlighted), motif B (teal highlighted) and motif C (grey highlighted). Motif A and C are the catalytic sites and Motif B is the dNTP binding site (Albà, 2001) Within each motif are conserved residues (green), conserved between family A polymerases and conserved in DNA polymerase I. Conservation indicates that DNA polymerase I is likely to comprise a 3'-5' exonuclease and a polymerase.

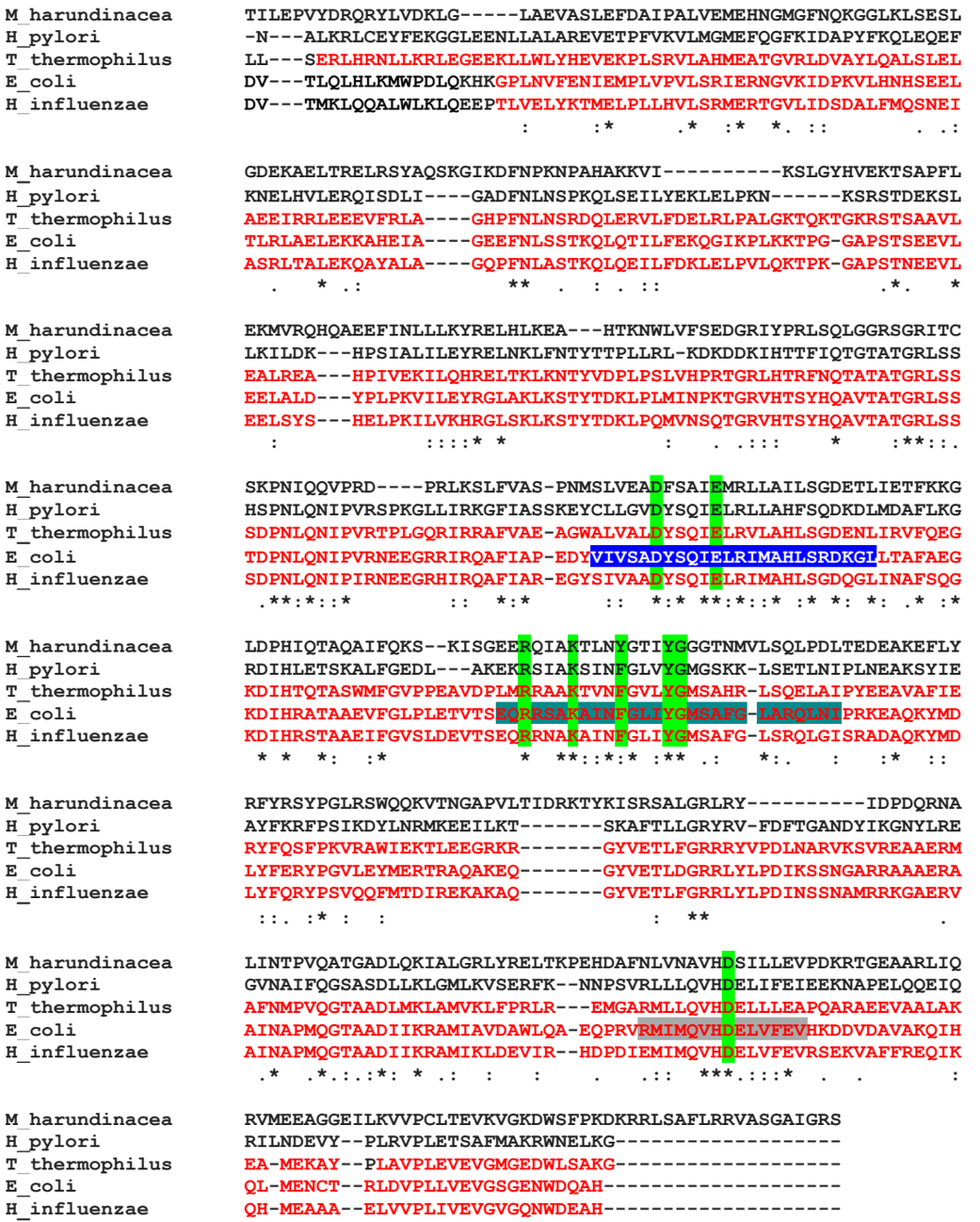

Figure 17: Amino acid sequence homology between DNA polymerase I and sequences from of *M. harundinacea, E. coli, H. pylori, T. thermophilus* and *H. influenzae.* This sequence alignment shows the alignment of the polymerase regions of the four polymerases with defined polymerase regions shown in red. Three polymerase motifs are highlighted in blue, teal and grey, within these regions are highly conserved residues (green highlight). *M. harundinacea* has conserved active residues within all three polymerase motifs.

To look further at the relationship between *M. harundinacea* DNA polymerase I with other Family A polymerases, a phylogenetic tree was made (Figure 18). This phylogenetic tree looked at the similarities between Family A DNA polymerases from eukarya, bacteria and archaea. Four archaeal species were found with Family A polymerases. All of these archaea are methanogenic, so it is possible that Family A polymerases are confined to methanogenic archaea. Despite this only two of these sequences clustered together (*M. harundinacea* and *Candidatus methanomethylophilus alvus*). The archaea sequences cluster with eukarya and bacteria, so there is no clear divide between bacteria, archaea and eukarya. It is possible that the sequences are diverged within methanogenic archaea or that the sample size does not show a complete picture.

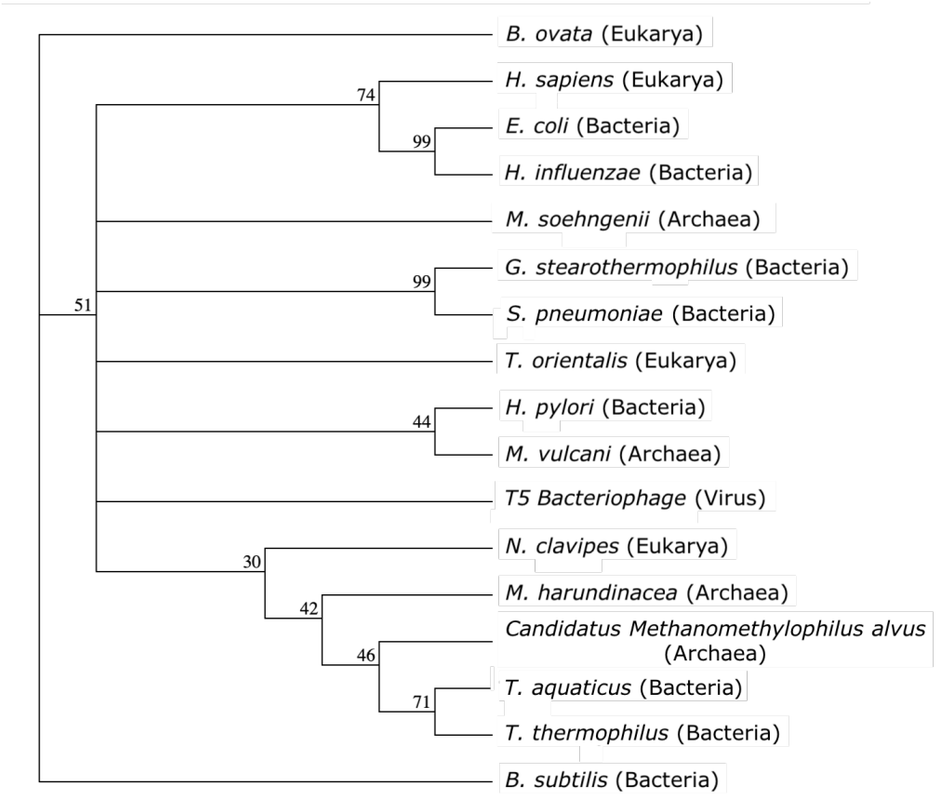

Figure 18: Phylogenetic tree examining the relationship between Family A DNA polymerases. *M. harundinacea* DNA polymerase I clusters with another methanogenic archaea as well as eukaryal and bacterial DNA polymerase I. There appears to be no clear separation of eukarya, bacteria and archaea

3.3.2 Tertiary Fold Prediction Using Phyre2

Molecular models for *M. harundinacea* were obtained using an online molecular modelling program, Phyre2. The FASTA file of the amino acid sequence was entered into Phyre2 and a model generated using the intensive modelling mode.

3.3.2.1 Cas4-1 predicted fold model

A molecular model was generated for Cas4-1 using Phyre2 (accessed at http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index) (Figure 19). To generate this model sequence alignment was carried out to identify homologous proteins with known structures. The structural regions that correspond to the homologous sequences are used to construct a model. The top homologous proteins of known structure used to create the model are shown in Table 14. The top 12 hits of homologous proteins are from Cas1 proteins, with the following 2 hits from Cas4. This is due to Cas4 covering a smaller proportion of the enzyme (42%). All Cas1 and Cas4 sequences have 100% confidence. The confidence in homologous proteins drops after this point when proteins other than Cas1 and Cas4 were used for modelling. The model is consistent with sequence alignment in that Cas4 is located at the N-terminal and Cas1 at the C-terminal.

Table 14: Highest ranked templates for Cas4-1 model. Five highest ranked templates used for the creation of the Cas4-1 model by Phyre2. Confidence has a value in percentage and percentage values over 90% are considered applicable.

Rank	PDB template	Confidence	% sequence identity	Protein
1	c4n06A	100	26	Cas1 from *Archaeoglobus fulgidus*
2	c3lfxE	100	28	Cas1 from *Thermotoga maritima*
3	c2yzsB	100	23	Cas1 from *Aquifex aeolicus*
4	c3pv9D	100	25	Cas1 from *Pyrococcus horikoshii*
5	c4w8kB	100	17	Cas1 from *Vibrio phage icp1*

No sequence alignment was possible for the first 83 residues of Cas4-1 which was largely unstructured in the Phyre2 model. These 83 residues at the N-terminal of Cas4 may be diverged from the Cas4 proteins used for this model. Only two Cas4 structures are available for comparison, so a lack of Cas4 structures available for comparison may have caused this instead of sequence divergence. No sequence alignment was present for the glycine rich region between the Cas4 and Cas1 regions. As only Cas4 and Cas1 sequences were used to generate this model, they would lack this linker region and therefore be unable to align. There was also no sequence alignment towards the start of the Cas1 region, as this is where Cas1 is joined to Cas4 via a glycine rich region it is likely that this sequence has diverged from non-fused Cas1 sequences due to the fusion.

The regions lacking sequence alignment, shown purple in the structure (Figure 19C), were predominantly unstructured regions. As stated previously these regions

have no sequence alignment, so no existing structure could be used for modelling. Though mathematical models are used for modelling, they are not reliable. Therefore, secondary structure may exist in these unstructured regions it just cannot be modelled.

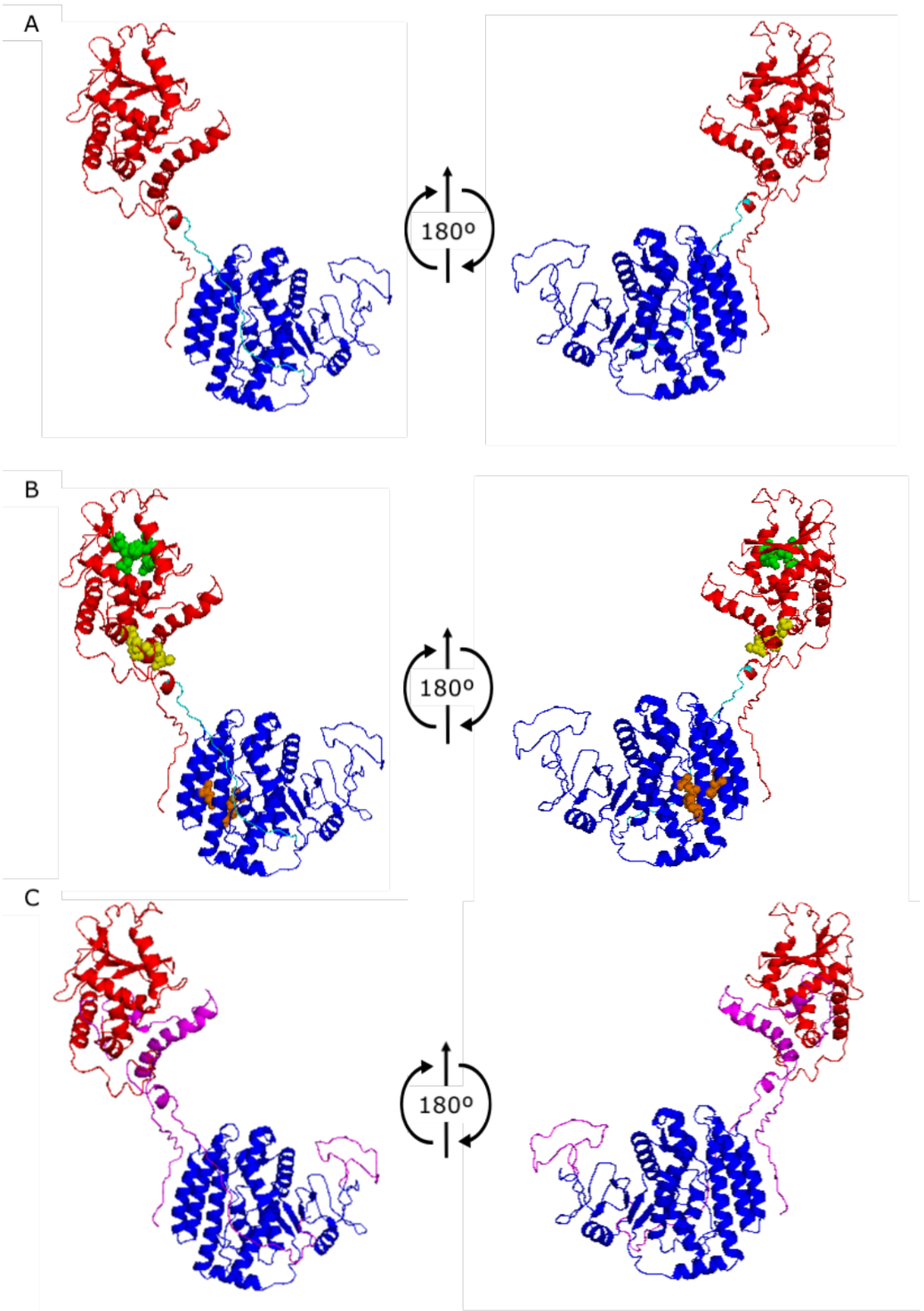

Figure 19: Cas4-1 Phyre2 model. A) Cas4-1 modelling in line with sequence alignment shows a N-terminal Cas4 (red) and C-terminal Cas1 (blue) separated by a linker (light blue). **B)** Predicted active sites from sequence alignment are shown in green for Cas4 and orange for Cas1 in addition to a predicted 4Fe:4S cluster in yellow. **C)** Areas with no sequence alignment for modelling purposes are highlighted in in purple. Predominantly these are unstructured regions.

3.3.2.2 Predicted fold model Cas2

The molecular model generated for Cas2 can be seen in Figure 20, and the top hits for protein homologs used for modelling are shown in Table 15. The top five hits

for homologous proteins are all Cas2 proteins, in fact all the proteins used with a confidence over 90% were Cas2 structures. All the top sequences were bacterial, but this is likely because bacterial systems are the best studied and have more crystal structures available. This modelling backs up the sequence alignment in that this gene encodes a Cas2 protein. For this model, only a single residue, the first residue had no sequence alignment.

Table 15: Cas2 model top homology hits. Top hits used for alignment and creation of Cas2 molecular model by Phyre2.

Rank	PDB template	Confidence	% sequence identity	Protein
1	c3oq2A	100	23	Cas2 from *Desulfovibrio vulgaris*
2	d1zpwx1	100	29	Cas2 from *Thermus thermophilus*
3	c5h1pB	99.9	31	Cas2 from *Xanthomonas albilineans*
4	c4es2A	99.9	28	Cas2 from *Bacillus halodurans*
5	c4qr1B	99.9	32	Cas2 from *Streptococcus pyogenes*

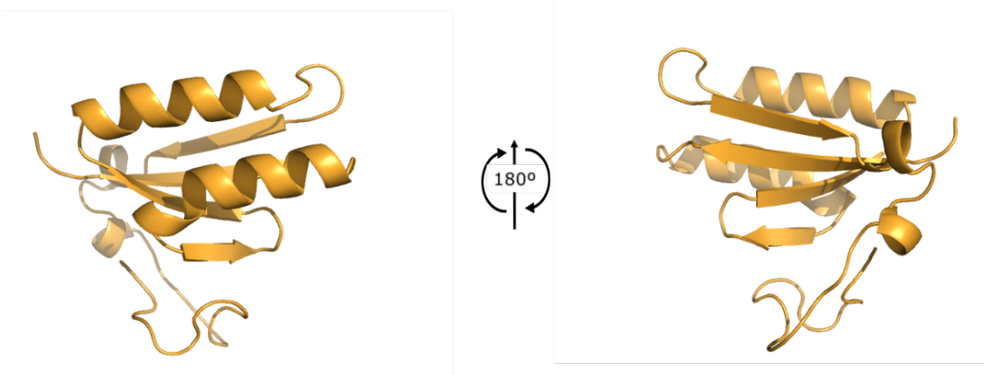

Figure 20: Cas2 Phyre2 molecule model. The model of Cas2 shows similar structural folds to other Cas2 structures.

3.3.2.3 DNA polymerase 1 predicted fold model

The molecular model generated by Phyre2 for DNA polymerase I is seen in Figure 21 and the top protein homologs used for modelling are seen in Table 16. The top five hits were all DNA polymerases from bacteria or eukarya. As discussed sections 1.3.4 and 3.3.1.3 DNA polymerase I is a Family A polymerase, a family more commonly found in eukarya and bacteria than archaea. As these domains are more commonly studied it is not surprising that these sequences/folds were used for the model. Most >90% confidence proteins were DNA polymerases or exonucleases. There were no missing sequence alignments for this model.

Table 16: Top protein homologs for DNA polymerase I model. Top hits of protein homologs utilised in the creation of the PolA molecular model by Phyre2

Rank	PDB template	Confidence	% sequence identity	Protein
1	c4xviA	100	22	DNA polymerase ν from *Homo sapiens*
2	c1njzA	100	25	DNA polymerase I from *Geobacillus stearothermophilus*
3	c4x0pB	100	24	DNA polymerase ϕ from *Homo sapiens*
4	c2kzzA	100	27	DNA polymerase from *Bacillus halodurans*
5	c4ktqA	100	29	DNA polymerase from *Escherichia coli*

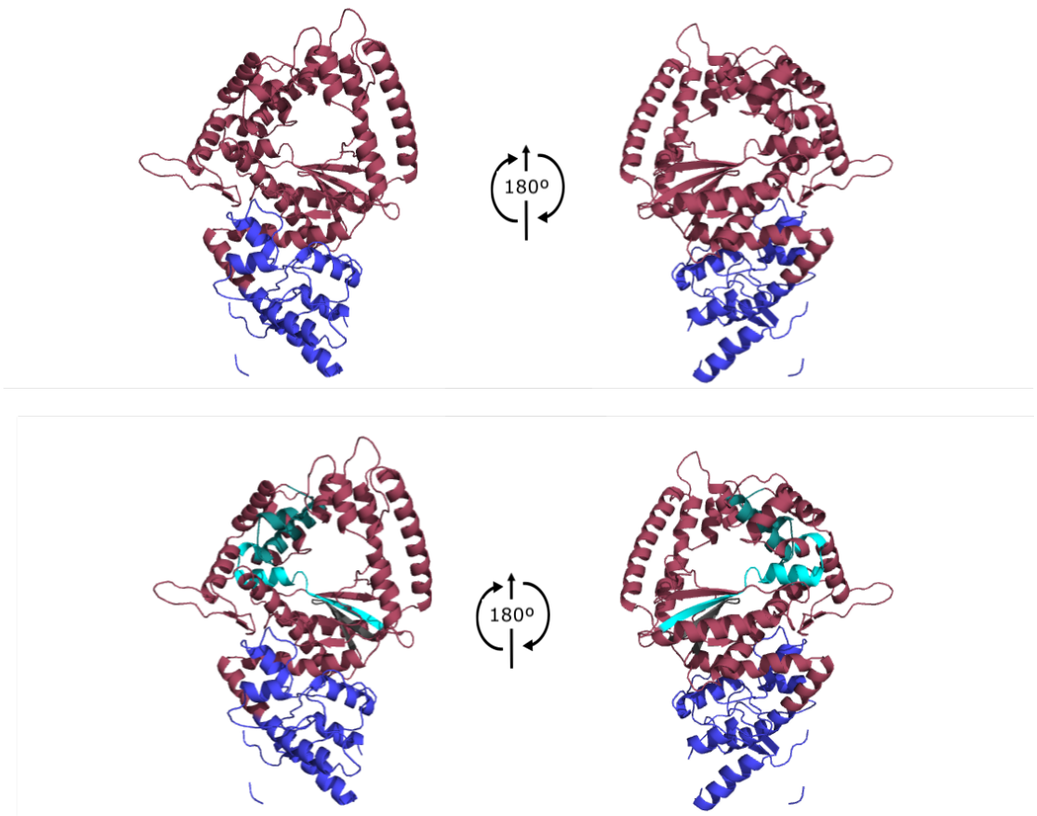

Figure 21: Phyre2 model of DNA polymerase I. A) Hypothetical division into 3'-5' exonuclease (blue) and polymerase (pink) **B)** Conserved motifs are shown in blue (Motif A), teal (Motif B) and grey (Motif C).

3.3.3 Phylogenetic analysis of Cas1 domain of Cas4-1 using MacVector

To examine the evolutionary relationship between Casposase and Cas1 a phylogenetic tree was creating using MacVector. The tree was constructed using a group of amino acid sequences taken from archaea and bacteria for Casposase and Cas1. The sequences used were a small percentage of available sequences to give a general idea of the evolutionary relationship. The tree (Figure 22) shows separate groupings of Cas1 and Casposase proteins. The tree was left unrooted.

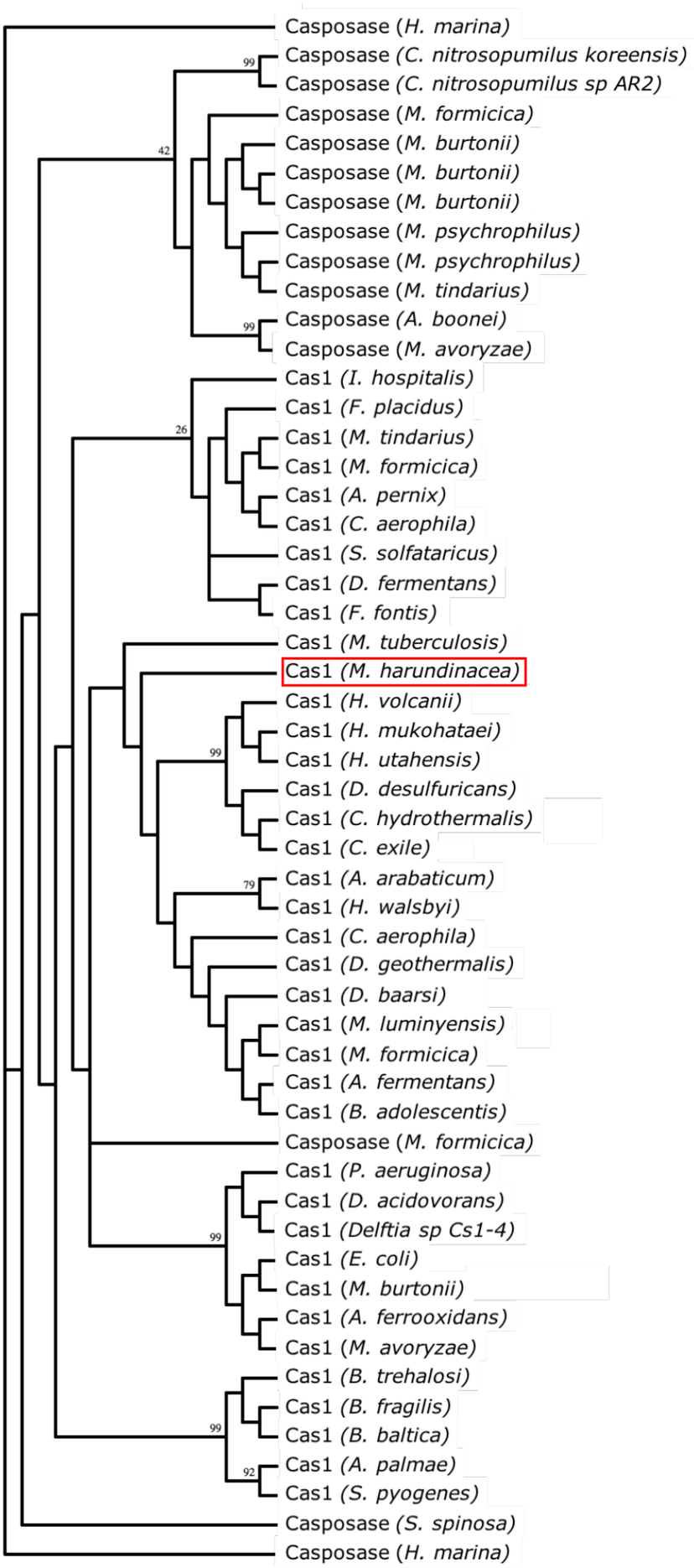

Figure 22: Phylogenetic tree examining the relationship between Casposase and Cas1 sequences. Casposase and Cas1 sequences were aligned with the Cas1 region of Cas4-1. Casposase and Cas1 sequences remained in separate clade groupings, with only one casposase sequence (*Methanoregula formicica*) clustering within Cas1 sequences. The Cas1 region of Cas4-1 clustered with Cas1 sequences. Bootstrap values for the majority of branches are not shown as they are below 20%, therefore there is little confidence in this tree.

The casposase sequences, bar one from *Methanoregula formicica*, cluster separately from Cas1 sequences. *M. harundinacea* Cas1 region was clustered with Cas1 sequences within a clade. This would suggest that *M. harundinacea's* Cas1 is more closely related to Cas1 sequences than Casposase. However, the bootstrap values for the tree are low. Only values over 20% are shown, meaning most of the tree has low confidence and Cas4-1 was in a region with low confidence. This phylogenetic tree provides an idea of the grouping, but a larger sample of sequences may provide a better insight and create a tree with better bootstrap values.

3.4 Molecular cloning of *M. harundinacea* ORFs

GeneArt® custom DNA was purchased from Thermo Fisher Scientific (see section 2.8.2 for details) for *E. coli* codon optimised *polA* and *cas4-1,* to improve expression. Each amino acid is encoded for by a nucleotide triplet, but amino acids can be coded for by multiple triplet sequences. Some species have a preference for a particular codon to code for an amino acid, called "codon bias". This "codon bias" means that codons found in the *M. harundinacea* ORF may pair with rare tRNAs in *E. coli* creating sub-optimal translation resulting in issues for protein over-expression. Optimisation allows the "codon bias" issue to be bypassed, improving protein expression.

Despite this, issues were found with the GeneArt® plasmids as they had been designed incorrectly leading to missing or duplicated restriction sites and frame-shifts upon sub-cloning. PCR was utilised to change restriction sites. For *polA* primers were designed to amplify the ORF from the GeneArt® construct. But for *cas4-1* the codon optimisation had introduced restriction sites that made cloning difficult so primers were designed to amplify the ORF from genomic DNA. *cas2* was not purchased in a codon-optimised form as the small size of the protein was thought to be less affected by codon-bias. Instead primers were designed for amplification of the *cas2* gene from genomic DNA.

PCR reactions were carried out as stated in section 2.8.3, and the resulting products run on agarose gels, seen in Figure 23. All products were the size expected and

were cloned into their respective expression vectors detailed in 2.8.9. Cloning was confirmed through restriction digestion and sequencing.

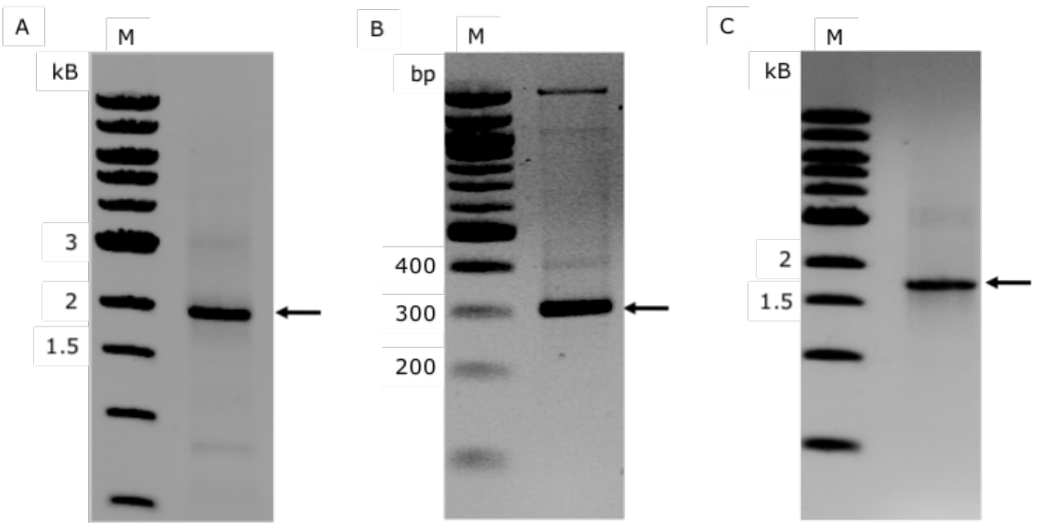

Figure 23: PCR amplification of *M. harundinacea* ORFs. A) *cas4-1* PCR amplifies an expected band of 1.9kB with BamHI and NdeI restriction sites. **B)** PCR amplification of *cas2* ORF with BamHI and HindIII restriction sites at 300bp consistent with its expected size. **C)** An expected band of 1.7kB was amplified by PCR with BamHI and EcoRI restriction sites for *polA*.

3.5 Protein Over-expression in *E. coli*

An *E. coli* protein expression system was used to express the proteins as they were derived from a single-celled organism so more complex expression (such as insect cells or mammalian cells) for post-translation modification was not required. *E. coli* overexpression is also relatively quick and cost effective.

3.5.1 Expression of Cas4-1

Expression of soluble Cas4-1 with a N-terminal hexahistidine tag was achieved in BL21C+ cells at 18°C after 18 hours post induction. Expression at 18°C was required for soluble expression, possibly because expression at 18°C slows hydrophobic interactions favouring folding over aggregation (Baldwin, 1986), preventing insolubility seen at 37°C (Data not shown). Cas4-1 expression was visualised by SDS-PAGE (Figure 24A), and both empty vector and uninduced controls contained a band the expected size of Cas4-1. Therefore, to confirm expression a western blot was under-taken (Figure 24B), following the method as stated in section 2.7.3. It confirmed the presence of a hexa-histidine tagged protein at a size consistent with Cas4-1, but no tagged protein in the control samples. Confirming that the protein overexpressed was Cas4-1.

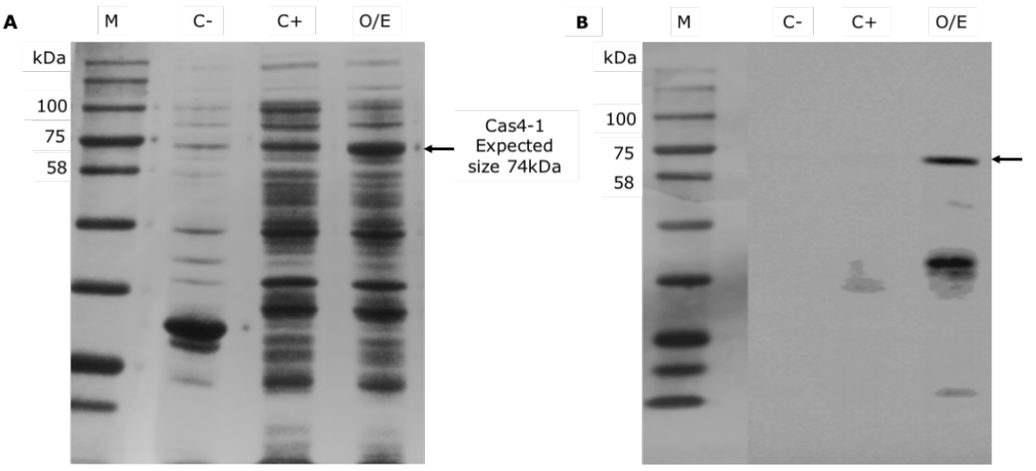

Figure 24: Overexpression of Cas4-1. A) Overexpression of Cas4-1 shown on a 12.5% acrylamide gel. The protein expected at 74kDa migrates around the 75kDa marker. Both the empty plasmid vector control and uninduced control showed a band of the same size as Cas4-1. **B)** To confirm the overexpression band present in the induced sample was Cas4-1 a western blot was carried out. Both control samples and overexpression sample were tested using anti-histidine antibodies which detected a his-tagged protein at the size expected for Cas4-1 in the overexpression sample. The image itself is a normal light image of ladder and chemiluminescent image of samples laid side by side.

3.5.2 Overexpression Cas2

Soluble expression of N-terminally hexahistidine tagged Cas2 was achieved in BL21AI *E. coli* cells at 37°C after three hours post induction, tested by SDS-PAGE (Figure 25). Cas2 has a predicted mass of 11kDa, but the overexpressed protein migrates just above the 11kDa marker.

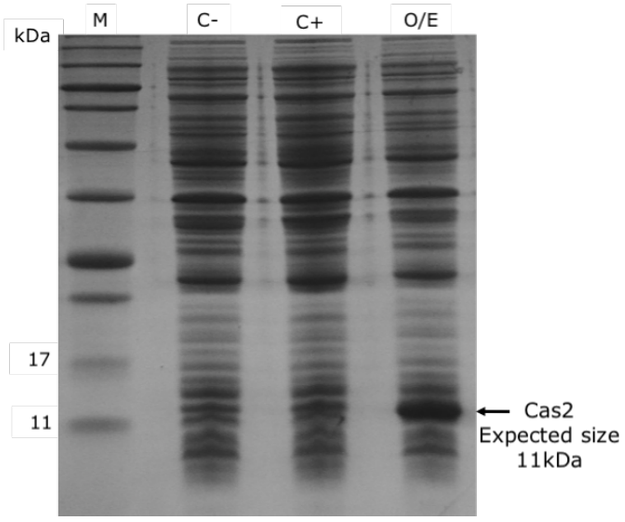

Figure 25: Overexpression of Cas2 protein. Cas2 overexpression shown on a 15% acrylamide. The protein is expected at 11kDa, but the protein band had migrated slightly slower than expected.

3.5.3 Expression DNA polymerase I

Soluble expression of DNA polymerase I required no inducing due to "leaky" expression. Expression from plasmid vectors is induced using as L-arabinose or

IPTG. However, if the promoter is not tightly regulated then expression can occur before the addition of inducers. This known as 'leaky' expression. As DNA polymerase I expression is leaky instead of inducing the plasmid upon reaching OD, the culture was simply incubated at 18°C for 18 hours. This expression can be seen in Figure 26 and DNA Polymerase I runs slightly quicker than expected.

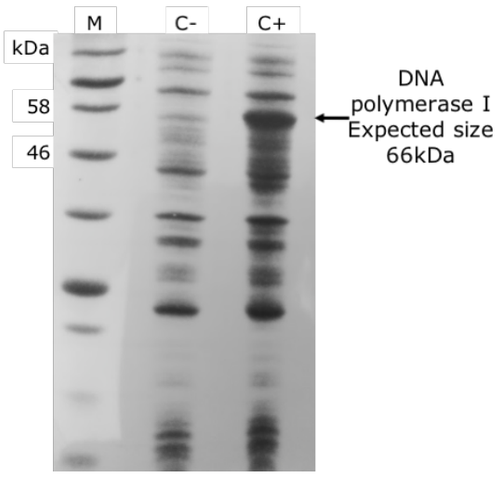

Figure 26: Overexpression of PolA. 12.5% acrylamide gel was used to visualise DNA Polymerase I overexpression. The band for DNA Polymerase I migrates quicker through the gel than expected for a 66kDa protein.

3.6 Purification of *M. harundinacea* Proteins

3.6.1 Purification of His-tagged Cas4-1

Cas4-1 purification was attempted using NiNTA affinity column with a FPLC system and cibacron blue gravity column as described in section 2.9.1. Protein was purified at 1.3μM and sent for mass spectrometry (University of Leciester) for verification. Unfortunately, the protein identified was an *E. coli* protein (ArnA) not Cas4-1. In order to purify Cas4-1 a new strategy was required.

3.6.2 Purification of Cas2

Cas2 was purified to homogeneity using NiNTA affinity and cationic exchange chromatography via an FPLC system. Gels detailing the purification are shown in Figure 28. Although Cas2 did not bind to the SP sepharose column, its impurities did bind and were removed. Cas2 was purified at 10μM and protein identity was tested via western blotting (Figure 27), where antibodies detected a His-tagged protein at the size expected.

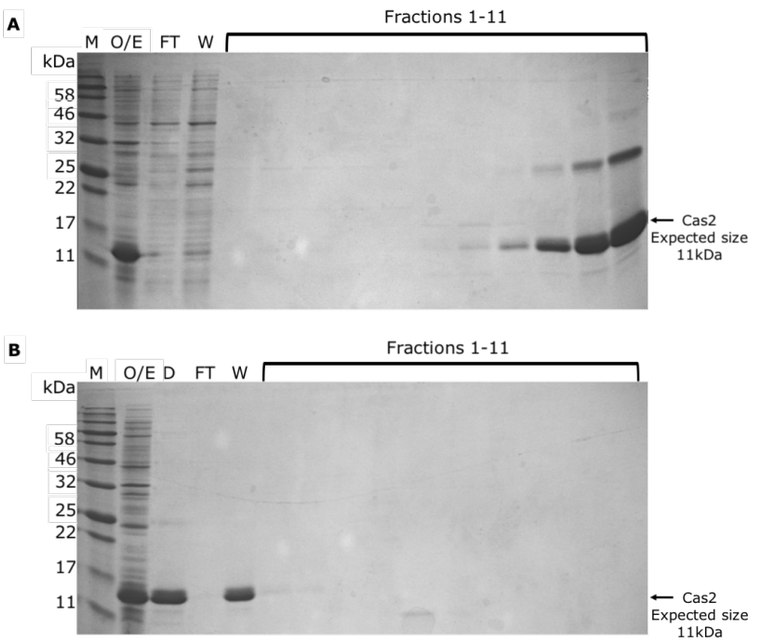

Figure 28: Cas2 Purification Gels. A) Affinity purification from clarified lysate. Fractions 9-11 were pooled for dialysis. **B)** Cationic purification from dialysed sample. The wash sample was dialysed.

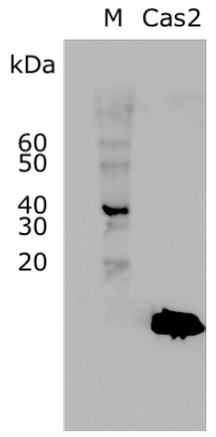

Figure 27: Cas2 Western Blot. Purified protein was treated with anti-His antibodies to confirm the presence of His-tagged Cas2. A his-tagged protein was observed at the size expected for Cas2.

3.6.3 Purification of His-tagged DNA polymerase I

DNA polymerase I purification was attempted by NiNTA affinity and anionic exchange columns via an FPLC system. Protein was purified at 2μM and was sent for mass spectrometry (University of Leciester) for verification. As with Cas4-1 the protein identified was an *E. coli* protein (Chaperonin 1). Therefore, a different method was required to purify the protein correctly.

3.7 New Strategies for Cas4-1 and DNA Polymerase I Cloning and Purification

As described above his-tagged versions of Cas4-1 and DNA polymerase I failed to purify successfully, instead producing native *E. coli* proteins. To separate Cas4-1 and PolA from these native proteins alternative tags were utilised. His-tags were initially used as they are small tags that do not generally interfere with protein activity and purification is relatively simple. However, proteins with a high histidine content can binds to the HiTrap chelating column. Two different tags were utilised for purification: Halo-tag and Strep-tag. A Halo-tag was used for *polA* as the tag is both highly specific and adds solubility. The majority of PolA was insoluble after expression and the use of the Halo-tag would hypothetically increase soluble expression of PolA and specific purification. For Cas4-1 solubility was of lower concern, therefore a more cost effective specific tag was used in the form of a Strep-tag. *polA* and *cas4-1* were cloned into new expression vectors for production of proteins with a Halo-tag and Strep-tag respectively.

3.7.1 Molecular Cloning of *polA* (*M. harundinacea)*

New primers were designed to amplify *polA* from the GeneArt® plasmid to add new restriction enzyme sites for cloning into a HaloTag vector. The correct sized PCR product was seen when run on an agarose gel (Figure 29). This product was cloned in a HaloTag vector which was confirmed by restriction digest and sequencing.

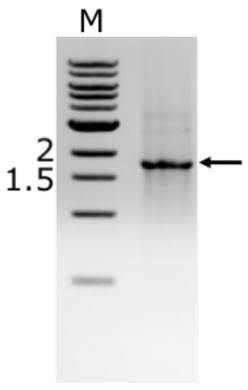

Figure 29: PCR amplification of *polA.* *polA* PCR amplifies an expected band of 1.7kB with SacI and XbaI restriction sites.

3.7.2 Site-Directed Mutagenesis Cas4-1

A different strategy was employed to add an alternative tag to Cas4-1, instead of cloning into a new vector pEW7 was mutagenised to insert a streptavidin tag. Site-directed mutagenesis was carried out as detailed in section 2.8.4. Mutation was confirmed by DNA sequencing.

3.7.3 Protein Over-expression in *E. coli*

Expression of N-terminal streptavidin tagged Cas4-1 and N-terminal Halo tagged DNA polymerase I was achieved in BL21C+ cells at 18°C, 18 hours post induction. Expression gels are shown in Figure 30.

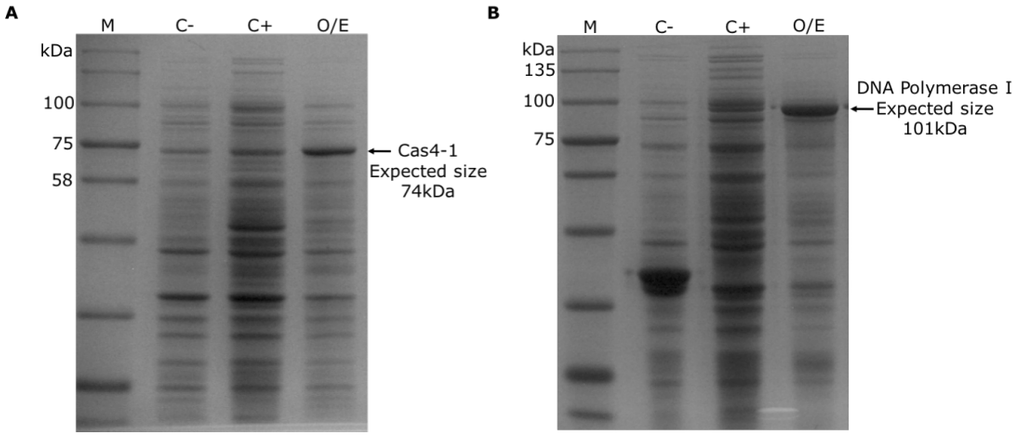

Figure 30: Overexpression of Streptavidin tagged Cas4-1 and Halo tagged PolA. A) Cas4-1 expression is seen at the expected size, just below the 74kDa marker**. B)** PolA expression with the addition of the Halo tag brings the size to 101kDa around the 100kDa marker.

3.7.4 Protein Purification of Cas4-1 and DNA polymerase I

3.7.4.1 Purification of strep-tagged Cas4-1

Cas4-1 was purified to homogeneity using Streptavidin and heparin affinity chromatography via an FPLC system. Gels detailing the purification are shown in Figure 31. Cas4-1 was purified to 2.6μM and sent for mass spectrometry (University of Leicester), which identified the protein as Strep-tagged Cas4-1.

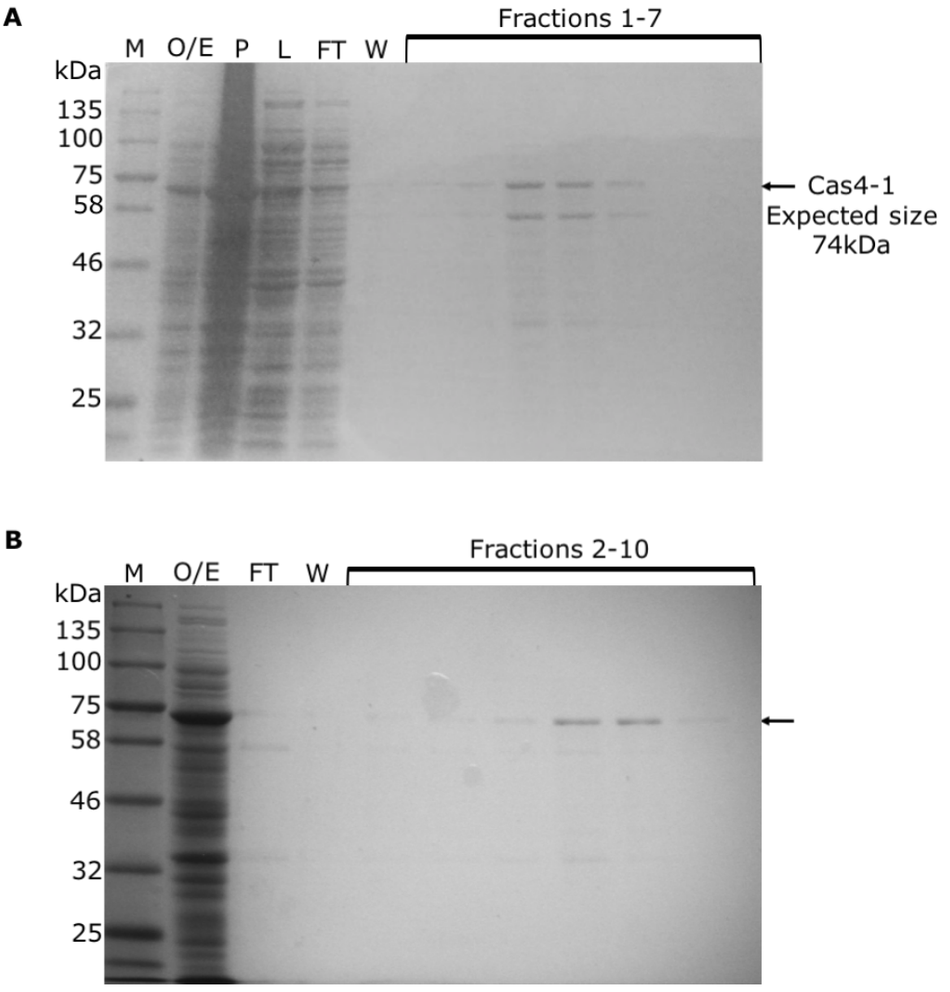

Figure 31: Purification of Streptavidin tagged Cas4-1. A) Clarified lysate was initially separate by a streptavidin column purification. Fraction 2-5 were pooled for dialysis. **B)** Dialysed fractions were separated by a heparin column. Fraction 6-8 were pooled for dialysis and storage at -80°C

3.7.4.2 Purification of halo-tagged DNA polymerase I

Purification of Halo-tagged DNA polymerase I was attempted using the HaloTag® Purification system (Promega). Purification was only attempted twice due to limited resources and the methods are detailed in section 2.9.5. The first purification was unsuccessful as the resin was not washed thoroughly enough resulting in contaminating proteins in the elution. Therefore, in the second purification binding and washing steps were carried out for longer, the result was that no proteins bound to the resin and eluted. PolA was mostly insoluble and what was soluble appeared unable to bind to the HaloLink™ resin.

3.8 Analysis of Cas4-1

Following successful purification of Cas4-1, the DNA binding and nuclease activities were tested.

3.8.1 Analysis of DNA binding by Cas4-1 protein using EMSAs

Cas4-1 capacity to bind fluorescently labelled DNA substrates was tested via EMSAs (Electrophoretic mobility shift assays). Before a full range of substrates was tested, binding of a flayed duplex (a likely binding substrate for Cas4-1) was tested at four different pH (pH6.5-7.5) conditions to find the best pH. No binding of a flayed duplex was seen at any concentration of Cas4-1 at any pH (Figure 32). Dragging of the fluorescent band was seen, and aggregation occurred in the wells, though neither of these are considered true binding.

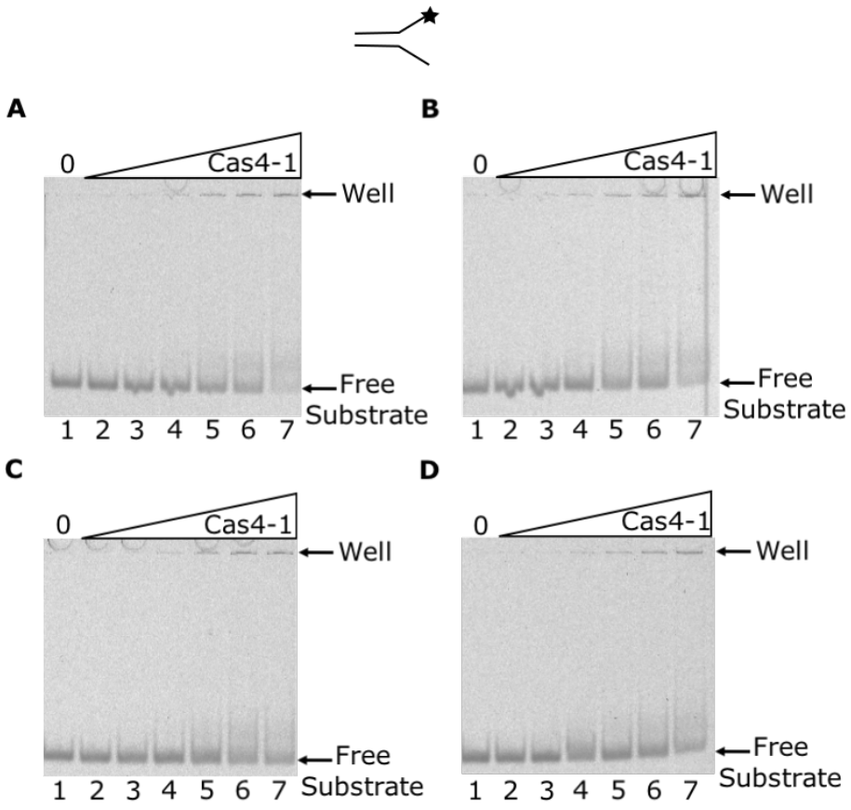

Figure 32: Cas4-1 binding of Flayed Duplex. Binding of Cas4-1 with a flayed duplex was tested at four different pHs 6-(A), 6.5-(B), 7-(C) and 7.5-(D). No binding was seen at any concentration or pH.

Towards the end of the concentration range, the fluorescent substrate leaves a streak up the gel as though a complex was almost forming. Therefore, a higher concentration range was used for the EMSAs and the gels were migrated for longer. Despite using a higher concentration there was still no true binding band (Figure 33). As concentration increased, so did the aggregation.

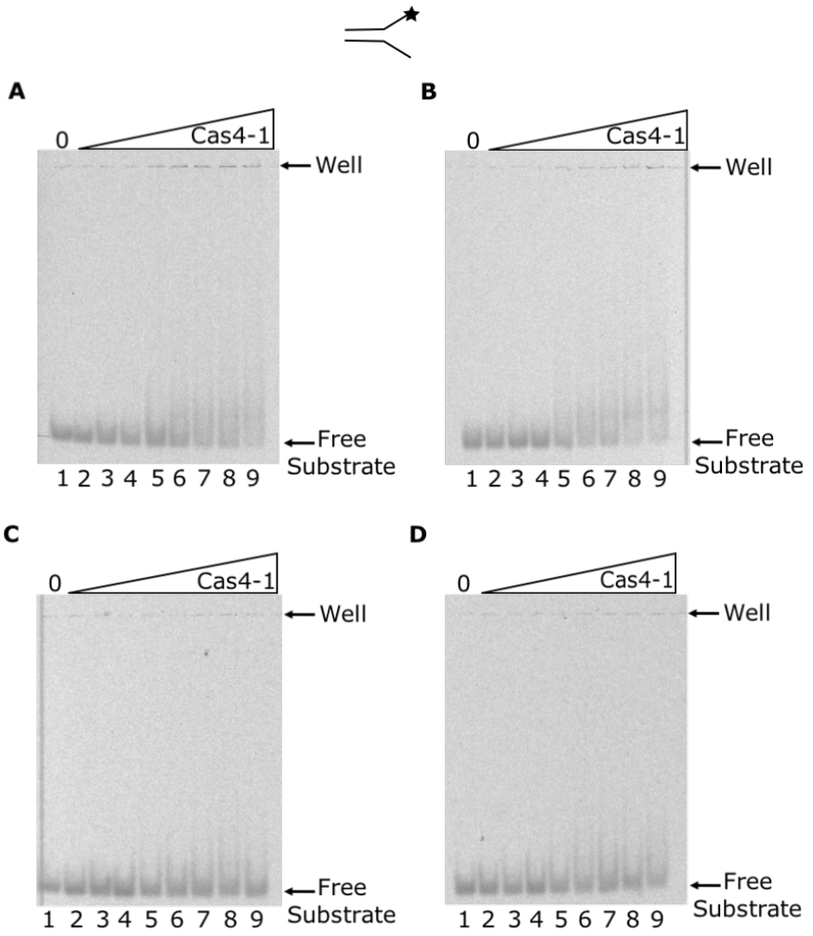

Figure 33: Higher concentrations of Cas4-1 binding of Flayed Duplex. Binding of Cas4-1 with a flayed duplex was tested at four different pHs 6-(A), 6.5-(B), 7-(C) and 7.5-(D) at a higher concentration range. No binding was seen at any concentration or any pH, only aggregation in the wells.

To help prevent aggregation occurring between Cas4-1 and the flayed duplex, 0.2% triton was added to the reactions. The addition of triton should help prevent non-specific aggregation of Cas4-1 and the flayed duplex. However, as seen in Figure 34 this only led to more aggregation in the wells.

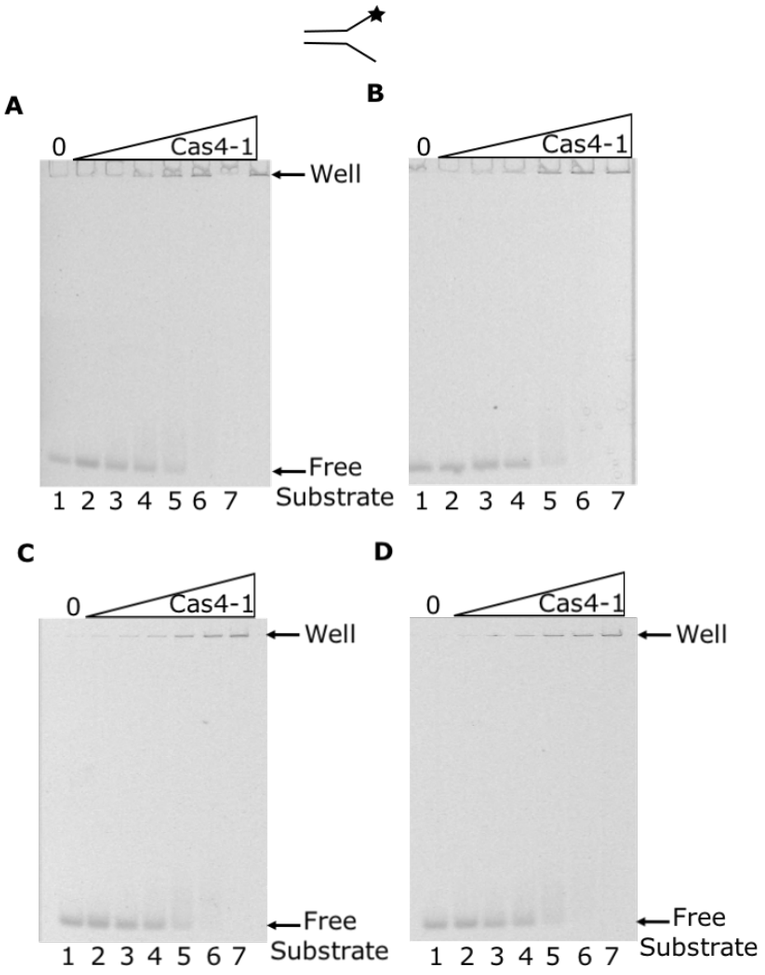

Figure 34: Cas4-1 binding of Flayed Duplex with 0.2% triton. Binding of Cas4-1 with a flayed duplex was tested at four different pHs 6-(A), 6.5-(B), 7-(C) and 7.5-(D) with the addition of 0.2% triton.

The addition of triton had failed to change amount of aggregation, so it may have been that the complex forming between Cas4-1 and the flayed duplex was too large to migrate into the gel. Therefore, a lower percentage acrylamide gel was made with a low percentage stacking gel on top. A larger complex should be able to migrate within that gel. Despite this, still no shifted bands appeared within the gel (Figure 35).

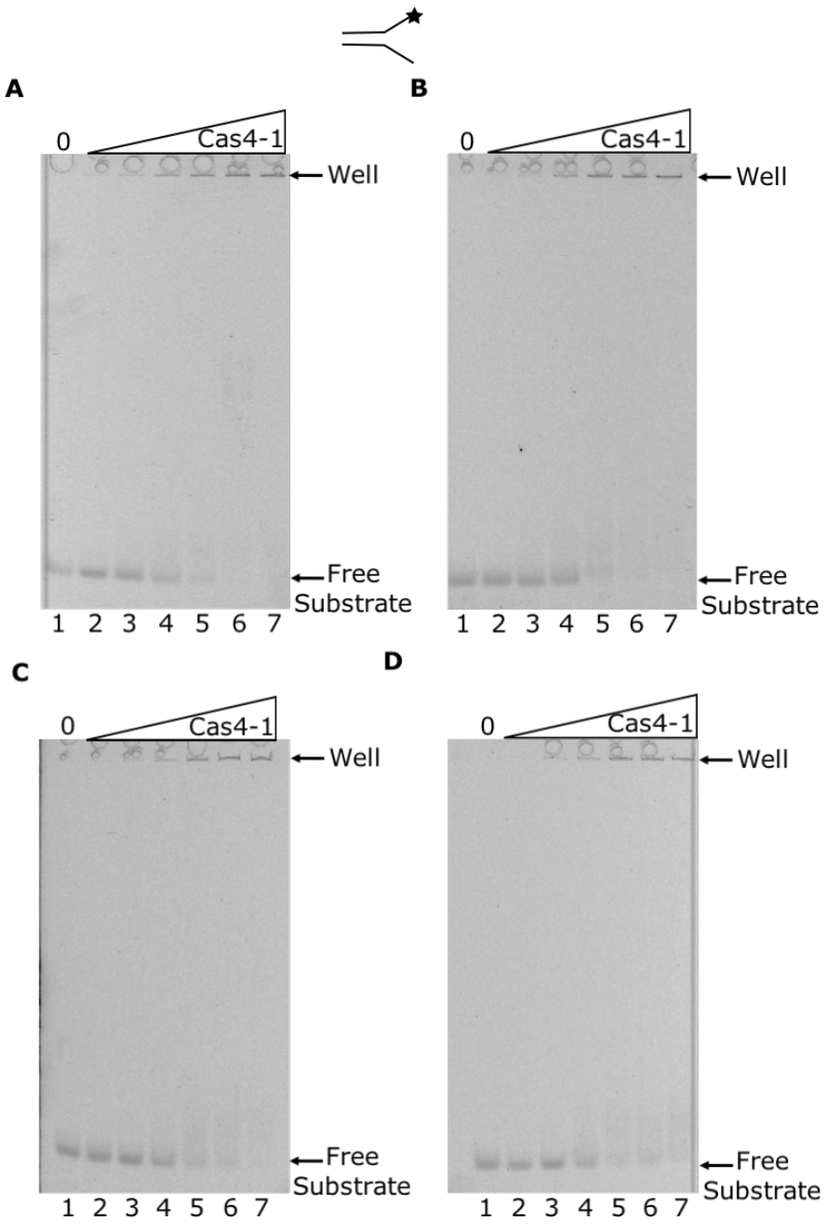

Figure 35: Cas4-1 binding of Flayed Duplex with 0.2% triton and lower percentage gel. Binding of Cas4-1 with a flayed duplex was tested at four different pHs 6-(A), 6.5-(B), 7-(C) and 7.5-(D) with the addition of 0.2% triton and migration on a lower percentage gel. No binding was seen at any pH at any concentration.

Due to diminishing protein, it was decided to test the nuclease activity of Cas4-1. If Cas4-1 only transiently binds DNA before cleavage then this will not be detected via EMSA.

3.8.2 Exploring Nuclease Activity of Cas4-1 against M13 ssDNA

Cas4 has been shown to cleave M13 ssDNA circular DNA, so Cas4-1 was also tested on this substrate (Figure 36). Cleavage does occur in small amount indicated by the smearing of the DNA towards the higher concentrations of Cas4-1. Unfortunately, higher concentrations of Cas4-1 cannot be tested to see if cleavage increases as the purified stock has a concentration of 2.6μM. It is also unclear

whether this cleavage is due to Cas4-1, or potential contaminating nucleases. In order to fully test this, Cas4-1 would need to be mutated to remove active site residues.

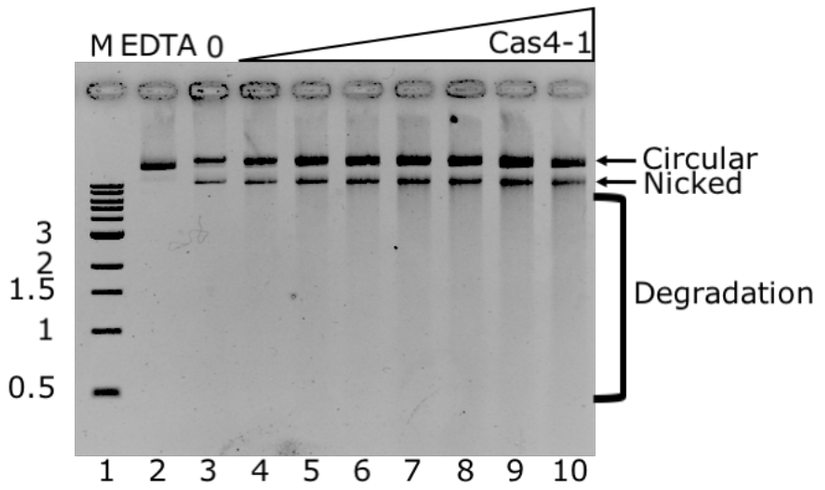

Figure 36: Cas4-1 cleavage of M13 circular ssDNA. M13 was incubated with increasing concentrations of Cas4-1 from 0.25μM to 1.75μM. A small amount of cleavage is seen in the form of smearing.

At this point the current stock of purified Cas4-1 ran out and as will be discussed in Chapter 4, work with another Cas4-1 from another organism (*Pyrinomonas methylaliphatogenes)* was proving more fruitful.

3.9 Discussion

Cas4-1 is implicated in 'capture' of new spacers within CRISPR-Cas (Lemak *et al*. 2013; Lemak *et al.* 2014; Liu *et al*. 2017; Plagens *et al*. 2012). Cas4-1 was obtained from *M. harundinacea* an methogenic archaea isolated in Beijing, China (Ma, Liu and Dong, 2006; Zhu *et al.*, 2012). *M. harundinacea* also contains a Cas2 and DNA polymerase I adjacent the Cas4-1 ORF. The location of the gene for DNA polymerase is interesting because: DNA polymerase I is a rare protein within archaea and casposon systems have polymerases located next to the casposon '*cas1'* gene (Hickman and Dyda, 2015).

Initial bioinformatics revealed that Cas4-1 and PolA contained conserved active site residues, whereas Cas2 lacked an active site residue required for RNase activity. Cas4-1 sequence alignment revealed that the N-terminal aligns with Cas4 and the C-terminal with Cas1. Both Cas4 and Cas1 homologous regions contain conserved active site residues including 4 conserved cysteines in the Cas4 region. Between

these two regions of homology was a glycine rich region which could act as a flexible linker. Structural fold predications also showed that the Cas4 and Cas1 regions fold into two separate structural regions linked by the glycine rich region. This bioinformatics shows that Cas4-1 likely contains Cas4 and Cas1 domains that have the functions of Cas4 and Cas1.

Cas2 sequence shows homology with Cas2 sequences and the fold prediction was based upon Cas2 sequences. Whilst some Cas2 proteins have been shown to exhibit RNase activity, *M. harundinacea* Cas2 lacks an arginine residues involved in the active site. Previous work on another Cas2 which lacked these two residues could find no evidence of RNase activity. Therefore, it is unlikely that this Cas2 has this activity, so the RNase activity was not tested.

DNA polymerase I had conserved active site residues within its polymerase region. It lacked any sequence identity to the 5‘-3’ exonuclease region. The 5’-3’ exonuclease in DNA polymerase I is required for nick translation activities which are involved in DNA repair. However, the DNA repair pathway of archaea is not well understood and many redundancies exist. Therefore, it could be that nick translation within archaea proceeds through another mechanism which does not requires DNA polymerase I and as a result the 5’-3 exonuclease activity was lost. A BLAST search for homology reveals little identity with other DNA polymerase I amino acid sequences, with the best match having only 33% identity (*Cyanobacterium aponinum*). *M. harundinacea*’s DNA polymerase I appears to be quite divergent from others. Despite an apparent loss of the 5’-3’ exonuclease region of the protein, the remainder of the protein shows homology to the 3’-5’ exonuclease and DNA polymerase region. For the DNA polymerase region conserved active residues are present, so whilst it may lack the 5’-3’ exonuclease it is likely to have other activities associated with DNA polymerase I.

Protein purification proved difficult in the cases of Cas4-1 and DNA polymerase I. Initial purification that resulted in the purification of native *E. coli* proteins included many months of trying different conditions and columns. But for simplicity only the final purification method and gels were included. Despite re-cloning efforts, it proved too difficult to purify DNA polymerase I. The protein was mostly insoluble and what was produced in a soluble-state only purified into highly contaminated

fractions or wash steps. For future pursuits, it may be possible to purify aggregated protein from inclusion bodies and re-fold the protein.

In the case of Cas4-1, the protein was produced but most likely without a stable 4Fe:4S cluster. The 4Fe:4S cluster results in the production of a protein that is brown in solution, however Cas4-1 produced a clear protein solution. The reasoning for this could be that the bacterial and archaeal systems are too different to allow proper folding within the *E. coli* expression system. Secondly the archaea from which this Cas4-1 is derived is an anaerobic organism and 4Fe:4S clusters are sensitive to oxidation. So, within the *E. coli* expression system it is highly likely that the 4Fe:4S cluster becomes oxidised and loses its brown colour. No DNA binding could be observed within an EMSA under different pH conditions, which could mean that Cas4-1 only interacts transiently with DNA as this cannot be detected via EMSA or that oxidisation of the 4Fe:4S cluster effects DNA binding. Slight potential nuclease activity occurs with M13 circular ssDNA as some smearing can be observed on the gel. The concentration of Cas4-1 is already quite high (1.75μM) so either cleavage is from background nucleases within the purified sample or Cas4-1 is not a very active nuclease. The nuclease activity may also be effected by an oxidised 4Fe:4S cluster. To conclusively conclude whether this weak activity is from Cas4-1, active site mutants would need to be tested to see whether they have nuclease activity.

Due to the various issues with purification and the conclusive lack of DNA binding or nuclease activities as well as more promising results arising from an alternative Cas4-1 (discussed in the next chapter) the work here was halted.

Chapter 4: Identification, Molecular Cloning, Purification and Analysis of Alternative Cas4-1 and Associated Proteins

4.1 Introduction

As discussed in chapter 3, adaptation proteins in *M. harundinacea* were intractable with purification only successful for Cas4-1 and Cas2. Cas2 was purified relatively easily, but Cas4-1 purification required multiple attempts and optimisation. Despite purification, Cas4-1 lacked the brown colour indicative of a 4Fe:4S cluster and did not bind DNA in an EMSA. Although cleavage of ss M13 DNA was demonstrated, due to the difficulties in purification and lack of DNA binding it was decided that an alternative Cas4-1 would be identified and experimentally explored. Through BLAST searching a different Cas4-1 was identified from *Pyrinomonas methylaliphatogenes.*

P. methylaliphatogenes is an aerobic acidophilic bacterium isolated in geothermally warmed soil in New Zealand (Crowe *et al.*, 2014) with a CRISPR gene neighbourhood comprising a Cas4-1 fusion, Cas2 and two hypothetical proteins. These proteins were selected to establish a single *in vitro* reaction for naïve adaptation. To accomplish this goal these proteins were analysed through bioinformatics, cloned into expression vectors, overexpressed, purified and then biochemically assayed.

4.2 Identifying an Alternative Cas4-1 Protein Using BLAST

Homologous proteins to *M. harundinacea* Cas4-1 amino acid sequence were searched for using BLAST. Several Cas4-1 sequences were identified, but an aerobic acidophilic bacterium was selected. *P. methylaliphatogenes* was the 23rd hit for the BLAST search and had 47% sequence identity to Cas4-1 from *M. harundinacea.* Two factors were hypothesised to produce a Cas4-1 with improved function from *P. methylaliphatogenes*. The first was the fact *P. methylaliphatogenes* was a bacterium. As the previous proteins were derived from archaea the

differences between archaea and bacteria may have led to issues in production and purification. Expressing bacterial proteins within a bacterial system will limit the effects of codon bias. In addition, the expression machinery will less diverged and therefore will produce proteins with correct folding and post-translational modification. The second factor was the aerobic nature of *P. methylaliphatogenes.* 4Fe:4S clusters are susceptible to oxidation, and oxidation of 4Fe:4S clusters can effect activity and cause significant conformation changes of proteins (Johnson, 1998; Sticht and Rösch, 1998; Talib *et al.*, 2014). It is a reasonable suggestion that a 4Fe:4S cluster protein from an anaerobic system (such as *M. harundinacea)* is more susceptible to oxidation due to a lack of adaptation to oxygen containing environments. *P. methylaliphatogenes* aerobic nature means that Cas4-1 will be exposed to atmospheric oxygen, so may be better suited to deal with oxidation. *M. harundinacea* Cas4-1 lacked the brown colouration associated with a properly functioning 4Fe:4S cluster, so was likely oxidised. *P. methylaliphatogenes* may be easier to produce with a stable reduced 4Fe:4S cluster.

The layout of *P. methylaliphatogenes* CRISPR locus and gene neighbourhood (Figure 37) was gathered through the online database, Kegg. *cas4-1* and *cas2* ORFs are located upstream of the CRISPR locus alongside two different hypothetical genes PYK22_01529 termed *hps* (Hypothetical protein small) in this thesis and PYK22_01526 termed *hpl* (Hypothetical protein large) in this thesis. BLAST of *hps* and *hpl* showed no significant similarity with any other proteins, so function was unknown. Further upstream, were a variety of interference proteins including Cas3, Cas6 and Cas10 (not shown).

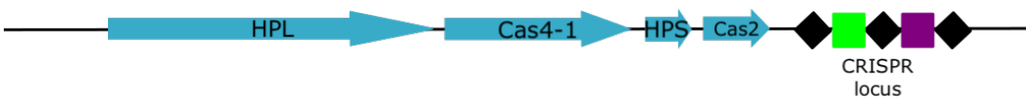

Figure 37: Organisation of *Pyrinomonas methylaliphatogenes* CRISPR gene neighbourhood. *cas4-1, cas2* and two hypothetical proteins are located upstream of the CRISPR locus. Associated interference genes were located upstream of HPL, but are not shown in this diagram.

4.3 Bioinformatic Analysis of *P. methylaliphatogenes* Genes

4.3.1 Identification of Conserved Residues Though Sequence Alignment Using Clustal Omega

4.3.1.1 Sequence Alignment of Cas4-1

The homology of Cas4-1 to Cas4 (*S. solfataricus* and *P. calidifontis)* and Cas1 (*E. coli, S. islandicus* and *M. tuberculosis)* amino acid sequences were tested using Clustal Omega (Figure 38). The reasoning for choosing these sequences was as described previously (section 3.3.1.1). Similarly, to *M. harundinacea* Cas4-1, Cas4 homology was restricted to the N-terminal end and Cas1 to the C-terminal end. However, unlike *M. harundinacea* there was no a clear definition for the end of Cas4 region and the start of Cas1 region due to an overlap in the two regions of homology. Despite this both regions have conserved active site residues. In the N-terminal Cas4 region there was a conserved RecB-like nuclease site (H46, D100, D113 and K115) and a 4Fe:4S cluster (C20, C203, C206 and C212). In the Cas1 C-terminal region there were conserved residues for the Cas1 active site (E392, H462 and E477). It was therefore hypothesised that both Cas4 and Cas1 regions will have activities associated with Cas4 and Cas1, but the absence of a linker region may affect the ability of the two proteins to function independently.

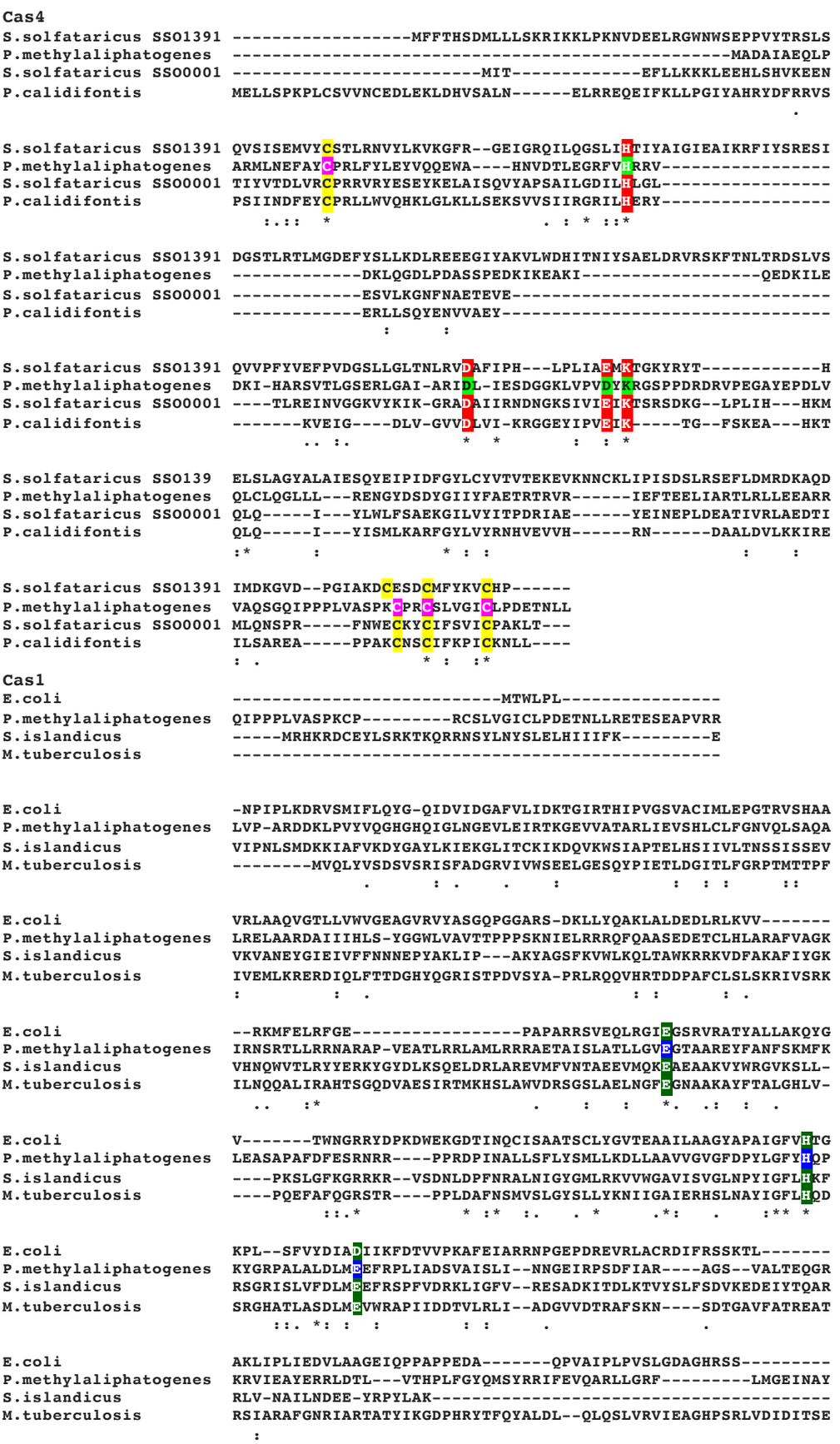

Figure 38: Homology of amino acid sequences for Cas4-1, Cas4 and Cas1. The N-terminal of Cas4-1 carries homology exclusively for Cas4, as shown by alignment of Cas4 sequences from *S. marinus, S. solfataricus* and *P. calidifontis*. This homologous region also contains conserved residues for a Cas4 active site (red and green) and 4Fe:4S cluster (yellow and pink). The C-terminal of Cas4-1 on the other hand had homology solely to Cas1, shown by alignment with Cas1 sequences from *E. coli, S. islandicus* and *M. tuberculosis*. This Cas1 homologous region also contains conserved Cas1 active site residues (green and blue).

4.3.1.2 Cas2 Sequence Alignment

Homology of *P. methylaliphatogenes* Cas2 amino acid was investigated using Clustal Omega with Cas2 amino acid sequences from *D. vulgaris, E. coli* and *S. solfataricus* (Figure 39). *E. coli , S. solfataricus* and *D. vulgaris* were chosen for the same reason described in section 3.3.1.2. Similarly to *M. harduninacea* Cas2, *P. methylaliphatogenes* Cas2 lacks an essential arginine residue (R17 in *S. solfataricus* highlighted in purple*)* and was therefore unlikely to exhibit the Cas2 RNase activity seen in *S. solfataricus*. Therefore the capacity of this Cas2 to carry out RNase activity was not tested.

```
E. coli                    ----MSMLVVVTENVP------PRLRGRLAIWLL-----EVRAGVYVGDVSAKIREMIWE
D. vulgaris                MYGNDAMLVLISYDVSFEDPGGQRRLRRIAKACQDY-GQRVQYSVFECVVDPAQWAKLKH
P. methylaliphatogenes     ----MRNRYIVSYDIS-----DPRRWRRVYRTMRGY-GDPIQYSVFQCDLLPAERIMMIE
S. solfataricus            ----MAMLYLIFYDIT-----DDNLRNRVAEFLKKKGLDRIQYSVFMGDLNSSR-LKDVE
                                     ::  ::        .   *:           :: .*:   :         .

E. coli                    QIAGLAEEG-----------NVVMAWATNTETGFEFQTFGLNR----RTPVDLDGLRLVS
D. vulgaris                RLLSEMDKEKD---CLRFY-YLGANW----RN--KVEHVGAKPAYDPEGPLIL-------
P. methylaliphatogenes     ALTGIIDHR----EDRVMLIDVGPAD---GRGRWSIETLGRAIKHEERIAIIV-------
S. solfataricus            AGLKIIGNRKKLQEDERFFILIVPITENQFRERIVIGYSGSER--EEKSNVVW-------
                                  .            :        .    .   *       .  :

E. coli                    FLPV
D. vulgaris                ----
P. methylaliphatogenes     ----
S. solfataricus            ----
```

Figure 39: *P. methylaliphatogenes* Cas2 homology with other Cas2 sequences. *P. methylaliphatogenes* Cas2 amino acid sequence was aligned with Cas2 sequences from *D. vulgaris, E. coli* and *S. solfataricus*. Cas2 active site residues for RNase activity are highlighted in yellow as determined in *S. solfataricus. P. methylaliphatogenes* Cas2 lacks two arginine residues, one of which is essential for RNase activity.

4.3.2 Tertiary Fold Prediction Using Phyre2

Molecular models were created using the online molecular modelling program, Phyre2. All models were created using the intensive modelling mode. Further information on Phyre2 is detailed in sections 2.5.4 and 3.3.2.

4.3.2.1 Predicted Tertiary Fold of Cas4-1

The Cas4-1 molecular model (Figure 40) was generated using several homologous proteins, as determined by sequence alignment. The structures of top 5 homologs used to create this Phyre2 model are detailed in Table 17. As with *M. harundinacea* Cas4-1, the top 12 hits were from Cas1 proteins, with the next two from Cas4. The top 5 Cas1 and Cas4 folds were used to create this model as the *M. harundinacea* model. In the *M. harundinacea* the only proteins with significant confidence were Cas1 and Cas4 sequences. However, with *P. methylaliphatogenes* significant confidence was present for AddB and RecB sequences. AddB and RecB sequences

were used for the generation model, creating some of the differences between the two Cas4-1 models.

Table 17: Top homologous protein sequences for Cas4-1 model. The top protein homologs used to create the Phyre2 Cas4-1 model. Confidence has a value in percentage and all proteins used here have a high confidence.

Rank	PDB template	Confidence	% sequence identity	Protein
1	c4n06A	100	28	Cas1 from *Archaeoglobus fulgidus*
2	c3lfxE	100	26	Cas1 from *Thermotoga maritima*
3	c2yzsB	100	25	Cas1 from *Aquiferex aeolicus*
4	c3p9D	100	28	Cas1 from *Pyrococcus horikoshii*
5	c4w8kB	100	15	Cas1 from *Vibrio phage icp1*

Although large parts of the *M. harundinacea* Cas4-1 model lacked sequence alignment, the *P. methylaliphatogenes* model only lacked sequence alignment for the first 6 residues and 3 residues in-between conserved cysteines. Despite the lack of a linker region, the two regions appeared separate in the model. Therefore, the Cas1 and Cas4 regions may be able to function independently. Some of the model lacks tertiary structure, but part of this was around the conserved cysteines that make up the 4Fe:4S cluster (Figure 40B). It is likely that this area is more structured due to the presence of a 4Fe:4S cluster, which is not modelled.

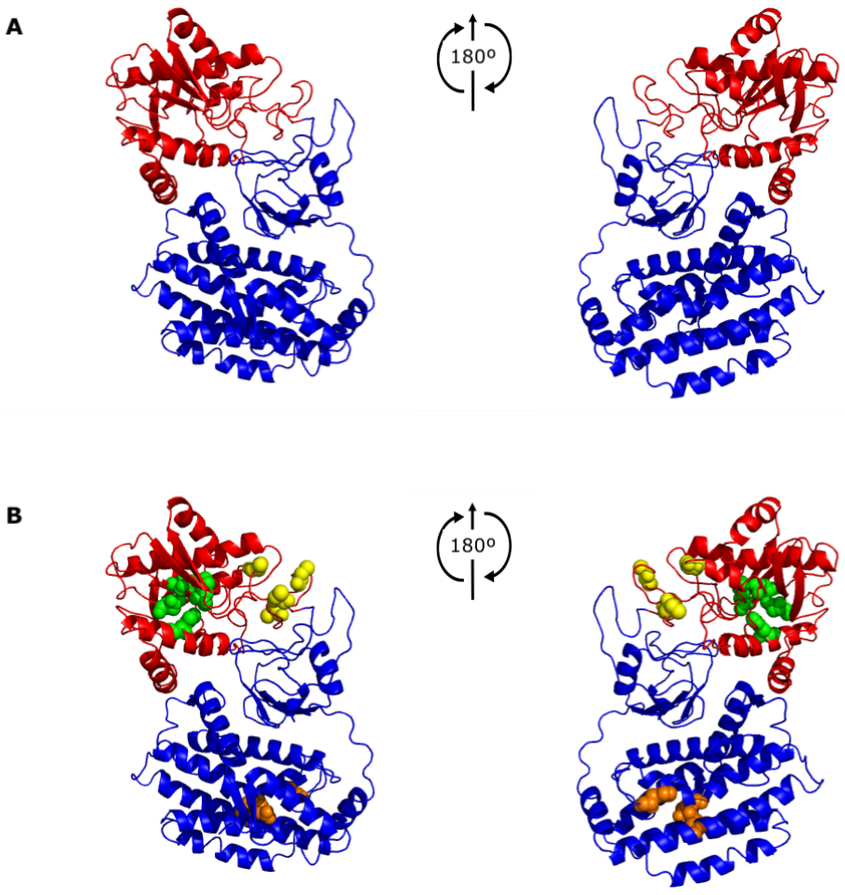

Figure 40: Predicted Model of Cas4-1. A) The rough divide of the N-terminal Cas4 region and C-terminal Cas1 region are shown in red and blue respectively. **B)** Predicted active sites identified from sequence alignment are represented by coloured spheres: Green- Cas4 active site, Yellow- 4Fe:4S cluster, Orange- Cas1 active site.

4.3.2.2 Cas2 Predicted Tertiary Fold

P. methylaliphatogenes Cas2 Phyre2 model (Figure 41) top five hits for homology used to create the model were the same as *M. harundinacea* Cas2 model (Table 18). The two models as expected were similar. As seen previously all the Cas2 homologs used had a confidence over 90% and only the residue that was not aligned was the first.

Table 18: Top Protein Homologs Utilised to Create Phyre2 model. Cas2 homologs were used for sequence alignment and production of Phyre2 model.

Rank	PDB Template	Confidence	% Sequence Identity	Protein
1	c3oq2A	100	27	Cas2 from *Desulfovibrio vulgaris*
2	d1zpwx1	99.9	27	Cas2 from *Thermus thermophilus*
3	c4es2A	99.9	28	Cas2 from *Bacillus halodurans*
4	c5hlpB	99.9	28	Cas2 from *Xanthomonas albilineans*
5	c4qr1B	99.9	34	Cas2 from *Streptococcus pyogenes*

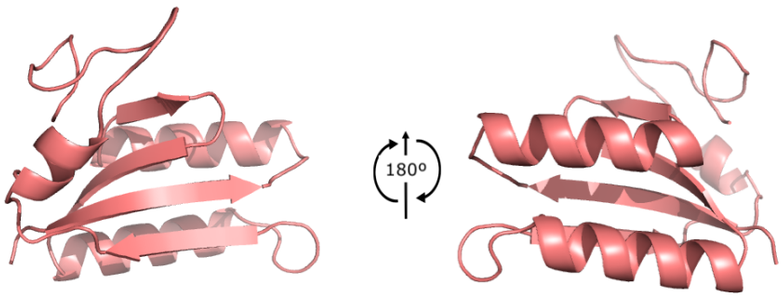

Figure 41: Cas2 Phyre2 Predicted Model. *P. methylaliphatogenes* Cas2 resembles other Cas2 structures, unsurprising as the predicted model was created using other Cas2 structures

4.3.2.3 HPS Tertiary Fold Prediction

As discussed in section 4.2 a BLAST search revealed no sequence homologs for HPS using either DNA or amino acid sequences. Therefore, it was unsurprising that the Phyre2 model found no sequence homologs which could be modelled with over 90% confidence. While a model was created (Figure 42A) there was little confidence in the model. Top hits were from a variety of proteins (Table 19), but the eukaryotic translation initiation factor 4e was used to generate the most reliable part of the model, highlighted in red in Figure 42B. The remainder of the model was generated using *ab inito* modelling, where the energetics involved in folding are predicted and the lowest free energy structure is generated. This does generate a model, but an unreliable one.

Table 19: HPS Top Homology Hits Phyre2. Top hits of protein homologs used to create HPS model.

Rank	PDB Template	Confidence	% Sequence Identity	Protein
1	d1tdja2	10.3	33	Ferredoxin like allosteric threonine Deaminase C-terminal domain.
2	c4ue9B	10	64	Eukaryotic translation initiation factor 4e
3	c4xgrH	7.7	47	Antitoxin VapB 30
4	c4xgrF	7.5	47	Antitoxin VapB 30
5	c4xgqB	6.7	47	Antitoxin VapB 30

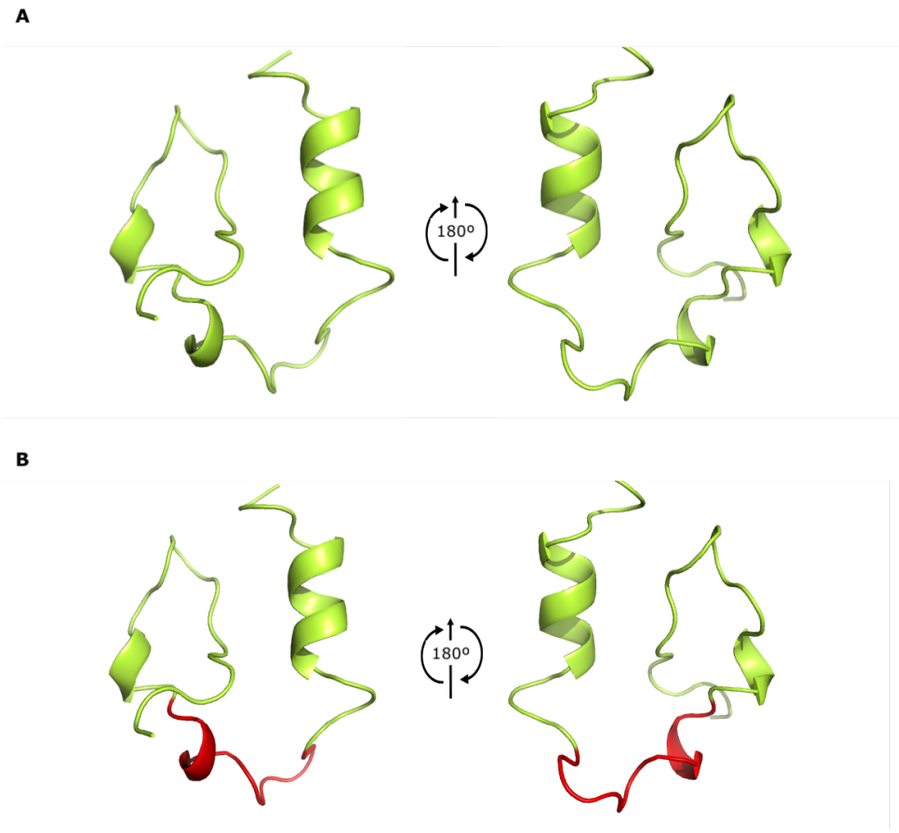

Figure 42: HPS Phyre2 model. A) The fold prediction for HPS appears to have little tertiary structure which could be due to low homology available for structure modelling. **B)** The red highlighted region shows the area modelled using the eukaryotic translation initiation factor 4e structure. The remainder was modelled using ab inito modelling.

4.3.2.4 Tertiary Fold Prediction of HPL

Much like HPS, HPL had no homology found through a BLAST search. Despite this the confidence for top hits in sequence homology for the HPL model (Table 20) were much higher than HPS, though still not at a reliable level. The homologs used were from a variety of different proteins, which failed to suggest what role this protein may have. The model showed large parts with no tertiary structure (Figure 43).

Table 20: Top Homology Hits Utilised for HPL Model. Phyre2 top hits for homology used to generate the HPL model. None had a confidence over 90%.

Rank	PDB template	Confidence	% Sequence Identity	Protein
1	c6f42V	82.6	14	RNA polymerase iii from *Saccharomyces cerevisiae*
2	d2f15aI	76.6	19	5' AMP activated glycogen protein kinase (AMPK)-beta binding domain
3	c4cffB	73	20	AMPK from *Homo sapiens*
4	c2ostC	71.5	34	Endonuclease from *Escherichia coli*
5	c3qraB	70.3	16	5-hydroxyisourate hydrolase from *Klebseilla pneumoniae*

Large parts of the model had no sequence alignment. These are shown in Figure 43C in green, these unaligned parts were interspersed by small sections of tertiary structure. This model is highly likely to be inaccurate, which is not surprising as the protein appears to be different to any other protein previously studied. However, whilst this could be an ORF, we have no proof that it produces a functional protein.

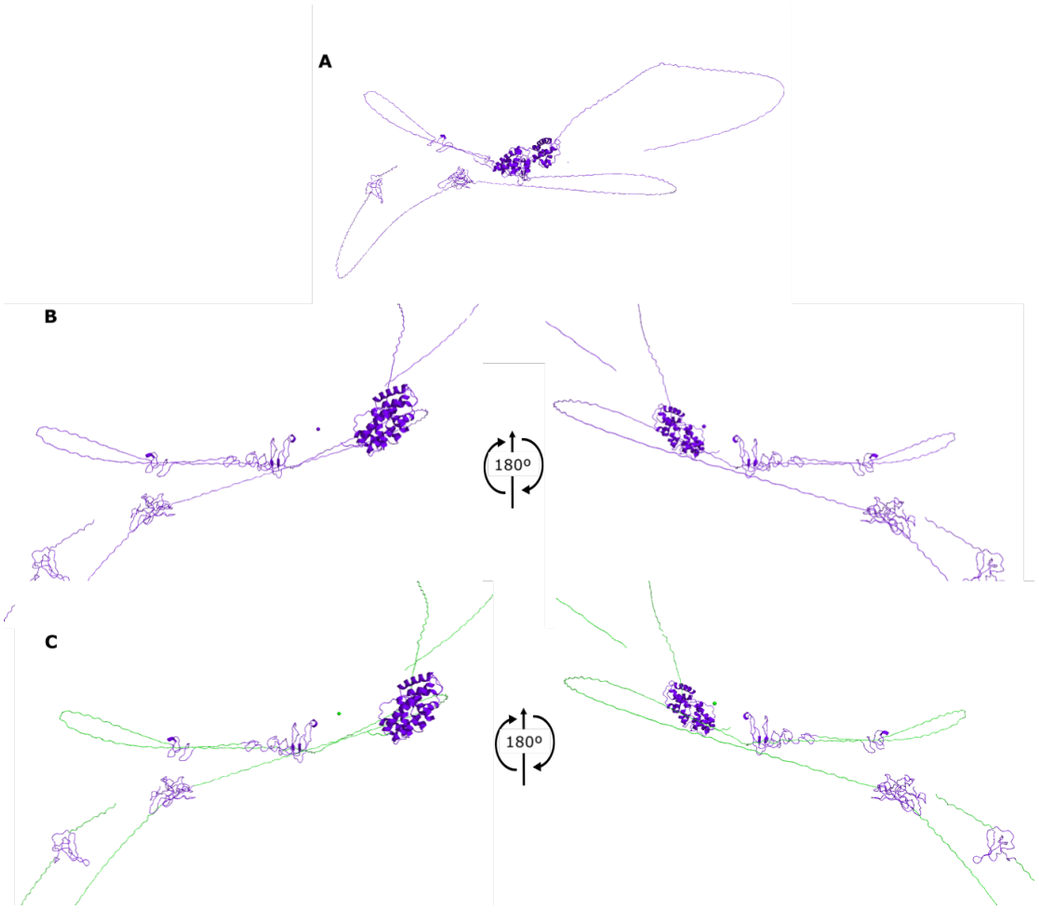

Figure 43: HPL model generated via Phyre2. A) The Phyre2 model predicts a large structure lacking tertiary structure. This is due to a lack of structural homologs from which to construct this model **B)** An enlarged view of the model focusing on the tertiary structures. **C)** Unaligned sequences are highlighted in green on the enlarged view of the model. Unaligned sequences are found in the large unstructured sections.

4.3 Molecular cloning of *P. methylaliphatogenes* ORFs

Primers were designed to amplify *P. methylaliphatogenes* genes from genomic DNA with restriction sites at each end to allow cloning into selected vectors. Codon optimised genes were not used for *P. methylaliphatogenes* it was believed that codon bias would have less of an effect and therefore for cost effectiveness codon optimised genes were not purchased. PCR reactions were performed as stated in section 2.8.3, and the resulting PCR products were imaged by electrophoresis on an agarose gel, seen in Figure 44. All amplified genes were at the sizes expected,

and cloning was attempted into expression vectors as detailed in section 2.8.9. Cloning was confirmed for *cas4-1, cas2* and *HPS* by restriction digest and sequencing. However, despite numerous attempts of cloning *HPL* proved too difficult to clone. This was potentially due to multiple bands present in the PCR making isolation of the *HPL* gene difficult. New primers sequences were attempted, but did not improve the PCR.

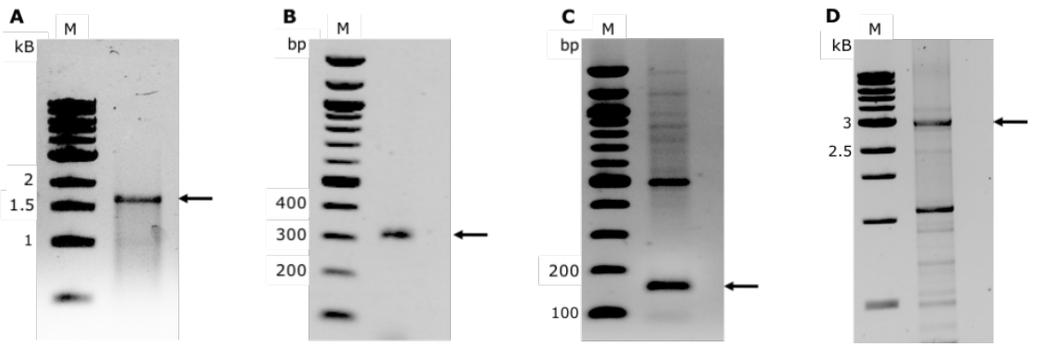

Figure 44: Amplification of *P. methylaliphatogenes* ORFs by PCR. A) *cas4-1* PCR amplifies an expected band of 1.7kB with BamHI and EcoRI restriction sites. **B)** PCR amplification of *cas2* with BamHI and EcoRI restriction sites at ≈300bp. **C)** An expected band of 144bp was amplified by PCR with KpnI and BamHI restriction sites for *HPS* **D)** Restriction sites KpnI and XhoI were amplified at the end of *HPL* gene by PCR at 3kB.

4.4 Protein Over-expression in *E. coli*

Soluble expression of *P. methylaliphatogenes* proteins with N-terminal hexahistdine tags were achieved in BL21AI cells at 18°C after 18 hours post induction. All protein expressions were visualised by SDS-PAGE, with empty vector and uninduced controls in Figure 45.

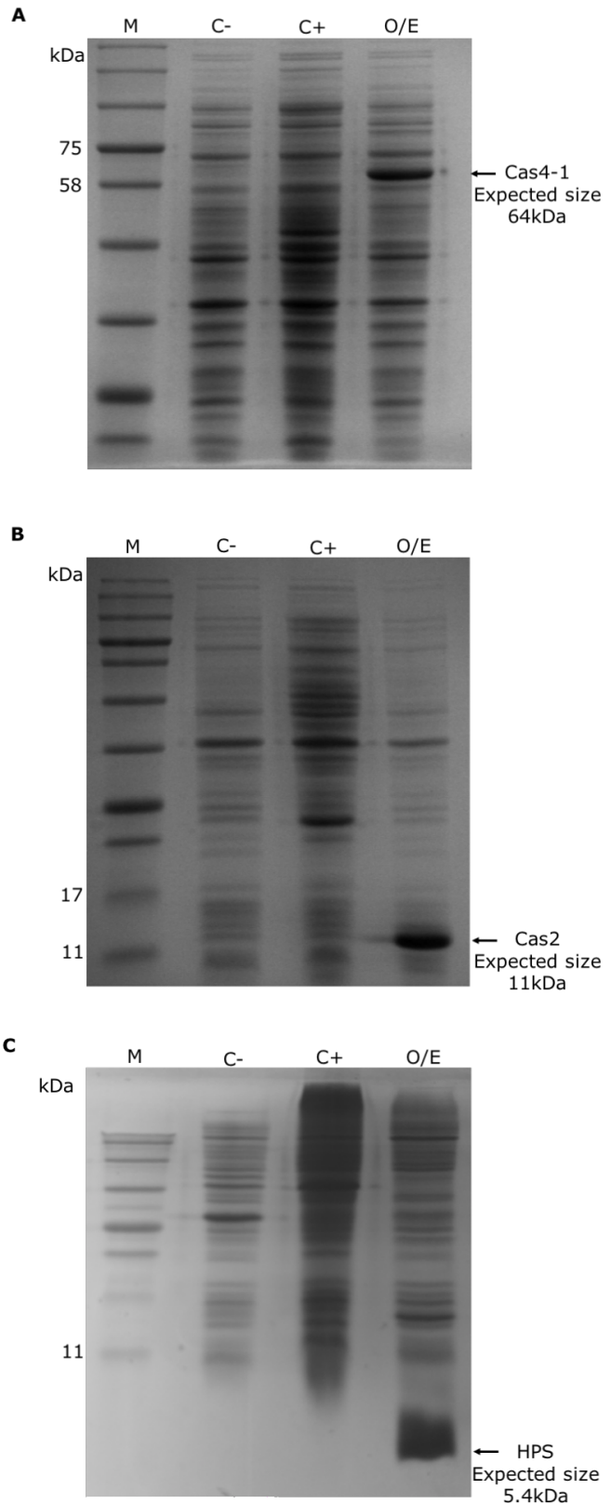

Figure 45: Overexpression of *P. methylaliphatogenes* proteins. A) Overexpression of Cas4-1 as shown on a 10% acrylamide gel. The protein expected at 64kDa runs between the 58 and 75 markers. **B)** Cas2 overexpression shown on a 15% acrylamide gel, a band was expected at 11kDa but the protein runs a little high. **C)** 20% gel was used to visualise HPS overexpression, which gives a band below all markers.

4.5 Protein Purification

4.5.1 Purification of Cas4-1

Cas4-1 was purified to homogeneity by affinity chromatography through Ni^{2+}-NTA and heparin resin as described in section 2.10.6. Purification analysis by SDS-PAGE

is shown in Figure 46A&B. This purification protocol was also used to purify mutant Cas4-1 proteins. The initial purification produced 2μM of protein and the purified solution was brown in colour indicating the presence of a 4Fe:4S cluster (Figure 46C). A spectrum of the protein to show the 4Fe:4S cluster was unsuccessful, with no conclusive shoulder at 400nm. In comparison to other published Cas4 purifications, the brown colour is rather light (Lemak *et al.* 2014; Zhang *et al*. 2012). This could signify that there ae two populations present within the purified sample: those with a 4Fe:4S cluster and those with oxidised 4Fe:4S cluster. A large amount of oxidised 4Fe:4S cluster may have interfered with the spectra results. Cas4-1 purification was confirmed via western blotting. The western blot shown in Figure 46D identified a His-tagged protein at the expected size of Cas4-1. Purified Cas4-1 was sent for mass spectrometry (University of Leicester) for verification and confirmed the purified protein was Cas4-1. Subsequent purifications of Cas4-1 and mutant proteins detailed later were purified using a similar, but optimised purification protocol as described in section 2.10.6.

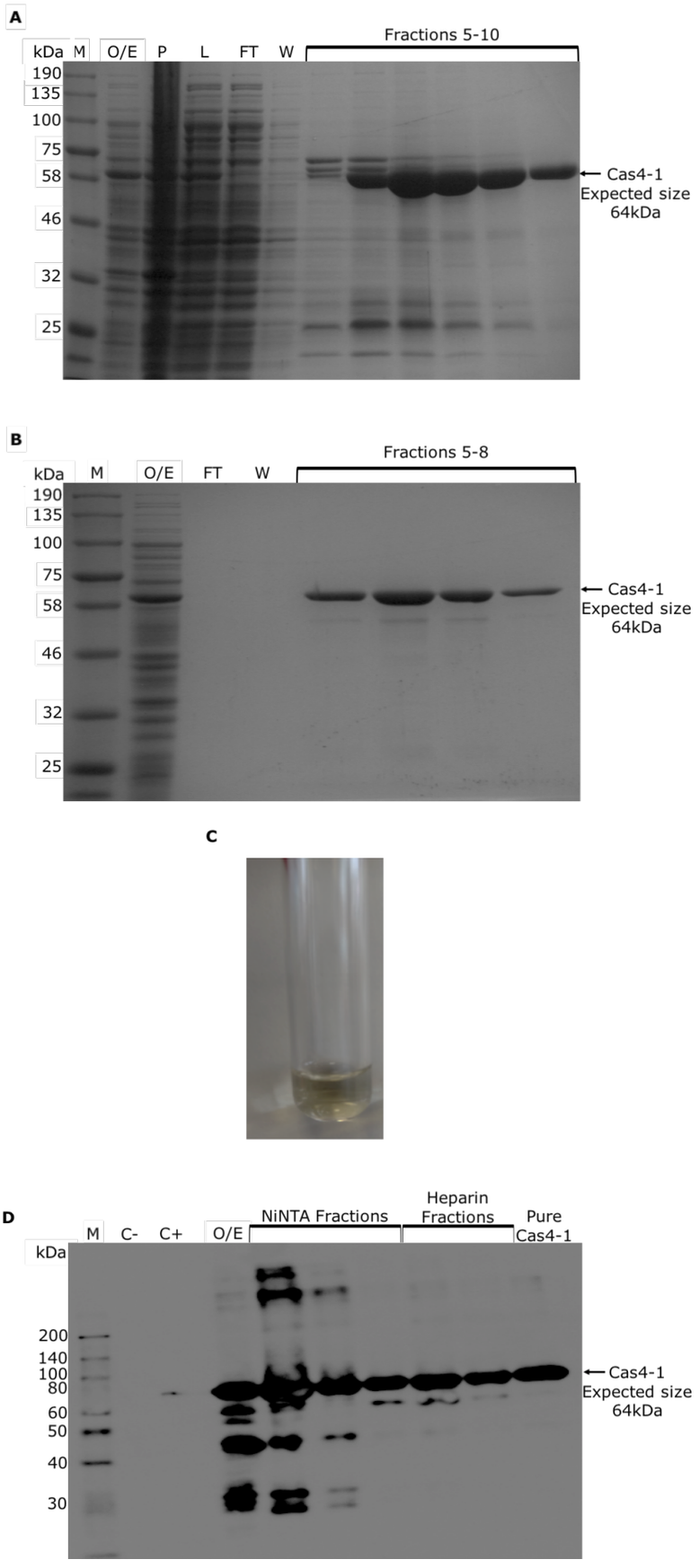

Figure 46: Cas4-1 Protein Purification. A) Cas4-1 was initially separated from clarified lysate by Ni^{2+}-NTA affinity chromatography. Fractions 8-10 were pooled for loading onto a heparin column. **B)** Heparin affinity chromatography purified Cas4-1 to near homogeneity. Fractions 6 and 7 were desalted before storage at -80°C **C)** Brown solution of purified Cas4-1 indicating the presence of an 4Fe:4S cluster. **D)** Confirmation of purification of a His-tagged protein, the approximate size of Cas4-1, via western blotting. Expression samples, pooled fractions and pure Cas4-1 were tested using Anti-His antibodies which detected a His-tagged protein at the size expected for Cas4-1. Smaller bands detected are degradation products of Cas4-1.

4.5.2 Protein Purification: Cas2

Purification of His-Cas2 required the use of a single column, Ni^{2+}-NTA as described in section 2.10.7. Analysis of purification by SDS-PAGE (Figure 47) was followed by detection of the hexahistidine tag by western blotting (Figure 48). After positive identification by western blotting, purified Cas2 was sent for mass spectrometry (University of Leicester) which confirmed purification of Cas2.

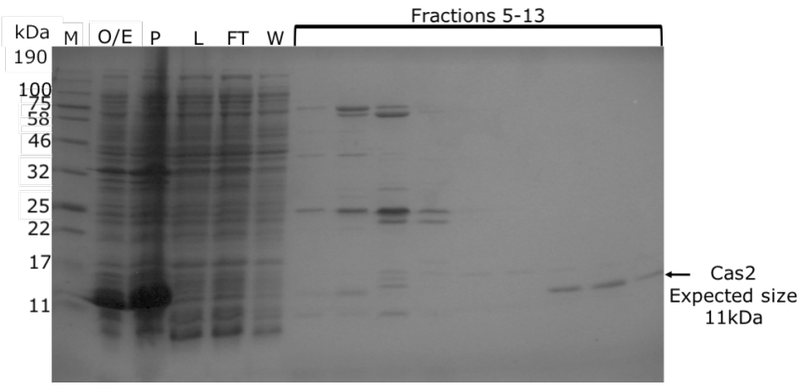

Figure 47: Purification of His-Cas2 Protein. His-Cas2, indicated by an arrow, was purified by Ni^{2+}-NTA affinity chromatography. Fraction 11-13 were pooled for dialysis before storage.

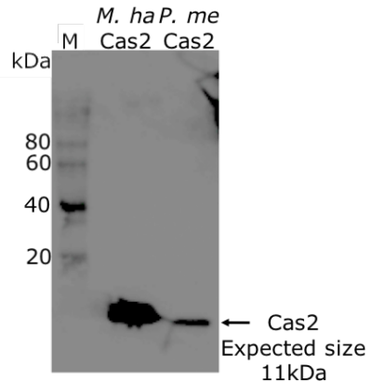

Figure 48: Western Blot Analysis of Purified Cas2. Purified His-Cas2 was analysed by western blot to confirm hexahistidine tagged protein at the size expected for Cas2. Purified Cas2 from *M. harundinacea* was used as a positive control and both proteins were detected at their expected sizes.

4.5.3 Protein purification: HPS

Purification of HPS to homogeneity also required a single column, Ni^{2+}-NTA described in section 2.10.8. The purification was analysed by SDS-PAGE, see Figure 49. Due to its small size, HPS was not verified further as it was highly unlikely to be a different protein. This was due to proteins that small being difficult to clarify on an acrylamide gel. Only the overabundance of HPS allows it to be visualised.

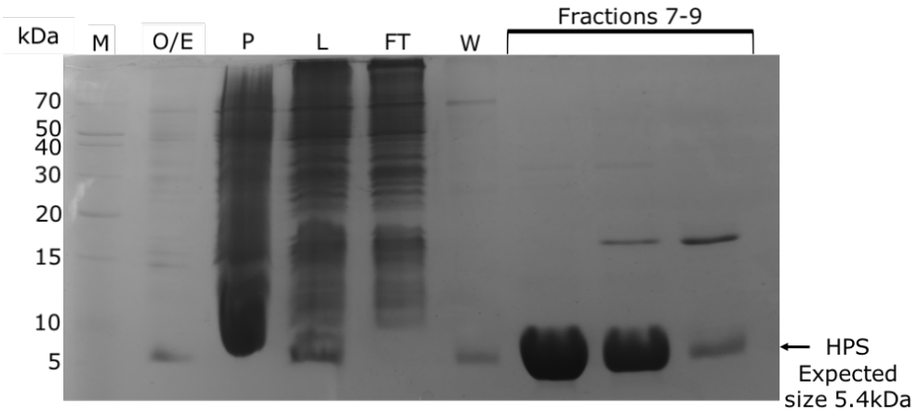

Figure 49: His-tagged HPS purification by Ni^{2+}-NTA Chromatography. HPS clarified lysate was purified through Ni^{2+}-NTA affinity chromatography. Fractions 7-8 were pooled for dialysis before storage at -80°C.

4.6 Analysis of Cas4-1 quaternary structure

Cas4-1 as described earlier contains structural and functional features of both Cas1 and Cas4. Having successfully purified Cas4-1 the oligomeric state of the protein was explored to see if it resembled oligomeric states reported for Cas1 or Cas4, or had its own unique oligomeric state. This was examined in three ways: Structural modelling using Galaxy Gemini, Blue Native PAGE (BN-PAGE) and Analytical Gel Filtration (AGF).

4.6.1 Modelling of Oligomeric State Analysis Using Galaxy Gemini

Cas4-1 predicted structure generated by Phyre2 (see section 4.3.2.1) was entered into Galaxy Gemini to generate a predicted oligomeric structure (Figure 50). The model is created by searching for profile similarity (stretches of up 20 homologous amino acids) and the best matches are searched for in the Protein Data Bank (PDB) for existing structures. The best PDB structures are then used as a basis to build the model. For Cas4-1 the PDB structure used to create the oligomeric structure was Cas1 from *A. fulgidus.* As can be seen from the model the oligomeric state was a dimer, consistent with a Cas1 oligomeric state as opposed to Cas4 oligomeric state. As the modelling was based upon a Cas1 structure, it is not surprising that the predicted oligomeric state was similar to Cas1. As discussed in section 4.3.2.1 a higher proportion of Cas4-1 protein contains structural and sequence similarities to Cas1, meaning Cas1 will have a higher profile hit than Cas4. There is also the issue that few oligomeric structures have been crystallised for Cas4 creating a smaller pool for modelling in comparison with Cas1. Despite this the model was a fair representation of a possible oligomeric state.

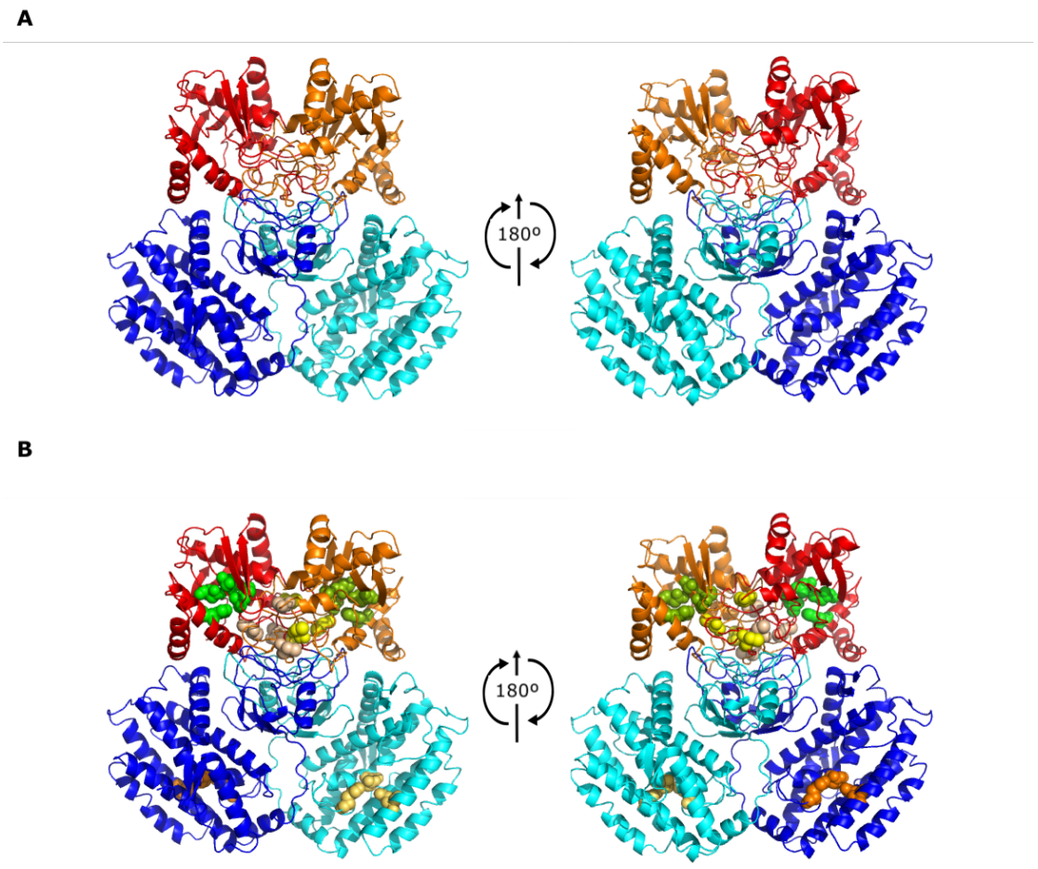

Figure 50: Oligomer Model for Cas4-1 Created by Galaxy Gemini. A) Cas4-1 predicted oligomeric state model is a dimer. Two monomer create the dimer, with Cas4 regions (red and orange) and Cas1 regions (dark blue and light blue) interact together. **B)** The predicted sites are shown within each monomer: Cas4 active site (Green and dark green), 4Fe:4S cluster (Yellow and pink) and Cas1 active site (Orange and yellow). This model, appears to show a region on interaction around the two 4Fe:4S cluster which could help form a strong dimeric structure.

4.6.2 Analysis of Oligomeric State by Blue Native PAGE

The predicted oligomeric state was a dimer. To test this experimentally blue native PAGE (BN-PAGE) was carried out. BN-PAGE is a technique used to detect protein oligomers and protein-protein interactions. SDS is omitted from the process so the negative charge for electrophoresis is instead provided by non-specific binding between the proteins and Coomassie blue G-250. As SDS is not present to denature proteins, the proteins run in their native states. The electrophoresis was carried out over a gradient gel to allow the separation and visualisation of complexes with a variety of molecular weights. BN-PAGE was carried out to experimentally verify the oligomeric state of Cas4-1. Purified Cas4-1 was incubated before loading in four different conditions: under normal buffer conditions, with TCEP (a reducing agent), with SDS and boiled at 95°C.

Purified Cas4-1 migrated in two bands around the 480kDa and 240kDa size markers (Figure 51). This indicated tetrameric and octameric states. This was different to the predicted oligomeric state, however as discussed above this model was based upon Cas1. The oligomeric state of Cas4-1 may be more dependent on Cas4, or have its own unique oligomeric state dissimilar to either Cas1 or Cas4.

Addition of the reducing agent TCEP induced no change in migration bands, although the bands become more defined. The presence of a 4Fe:4S cluster means Cas4-1 could be susceptible to oxidation and TCEP may protect the protein or reverse the effects of oxidation. Though there could be a slight stabilisation of the protein leading to more defined migration bands, there was no significant effect on Cas4-1 by the reducing agent. Smearing was seen in both lanes, indicative of aggregation.

SDS had a significant effect on the migration of Cas4-1 with a stronger band migrating between 146kDa and 66kDa size markers. A faint band was still present at 242kDa, and a further band slightly higher between the 480kDa and 242kDa size markers. It may be expected that the addition of SDS would denature the protein to its monomeric form, however the major form was a dimer. A dimer may be the most stable oligomeric state of Cas4-1, forming the building blocks to create larger complexes. Less smearing was seen in this sample, suggesting less protein was aggregated. This could account for the darker band; less protein was aggregated and therefore more was contained in bands.

Boiling of the sample showed no protein on the gel, this could be because the protein had been destroyed or as the 66kDa marker showed the limit of this gel, the monomeric form may have migrated off the gel.

Though not shown on this gel Cas2 was added to Cas4-1 and incubated for 30 minutes on ice before migration on a BN-PAGE gel. No difference was observed, so complex formation was not detected.

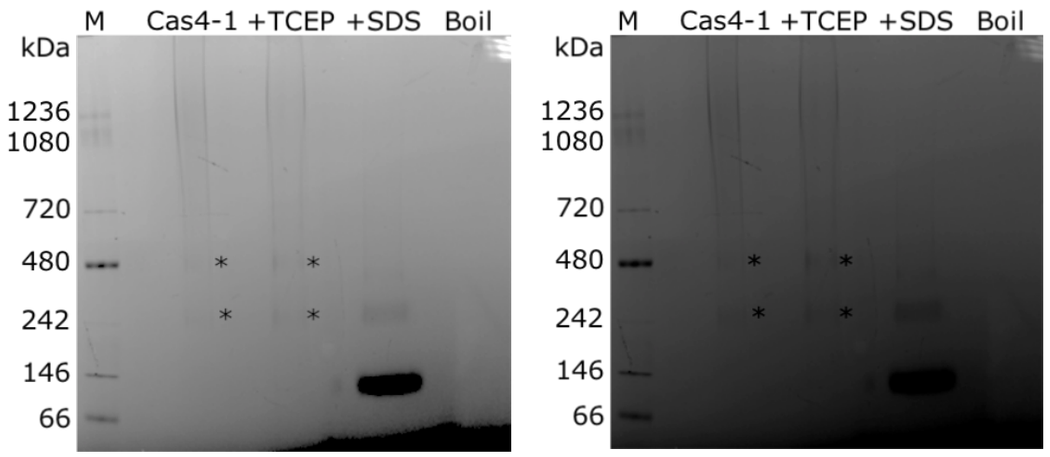

Figure 51: BN-PAGE analysis of Cas4-1 Oligomeric State. This image is shown in duplicate with two different contrasts to fully show all bands present. Native purified Cas4-1 was run on a 3-12% gradient acrylamide gel in four conditions. Under normal buffer conditions Cas4-1 migrated around 480kDa and 242kDa around 8 and 4 monomers. Addition of TCEP, a reducing agent, did not change the oligomeric state but the bands appeared more defined so may be stabilised. SDS addition to denature the sample did not produce a monomer as expected, instead the main product was a dimer, with some tetramer and hexamer present. The dimer may be the most stable complex from which higher oligomeric states are built. Cas4-1 was also boiled to produce a monomer, but no band appeared on the gel. But as the 66kDa band appears just before the limit of the gel it was likely that the Cas4-1 monomer may have run off the gel

4.6.3 Oligomeric State Analysis by Analytical Gel Filtration

The oligomeric state was further investigated using AGF. AGF is size exclusion chromatography where proteins are separated by size, the protocol is described in section 2.14. Initially two sets of standards were run on the column to determine the void volume of the column and create a standard line (Figure 52). Cas4-1 eluted in a single peak at around 13.5ml, at the same position over multiple runs. This value was converted into a Kav (a relationship between the void volume, column volume and elution volume, see section 2.14 for more details) and plotted on the standard line. The molecular weight of Cas4-1 was calculated using the line equation as 112kDa or 1.75 molecules. This is slightly below what is expected for a dimer, but there is some error associated with AGF. Therefore, it is more likely Cas4-1 is a dimer as opposed to a monomer.

Cas4-1 was also run through the column after incubation with Mg^{2+}, DNA and Cas2. However, no change was seen with any addition. This would suggest that Mg^{2+} and DNA have no effect on Cas4-1 oligomeric state. Cas4-1 from this data would appear to not complex with Cas2, but no peak was visible when Cas2 alone is run. It may be that Cas2 does complex with Cas4-1, but the amount was too small to detect.

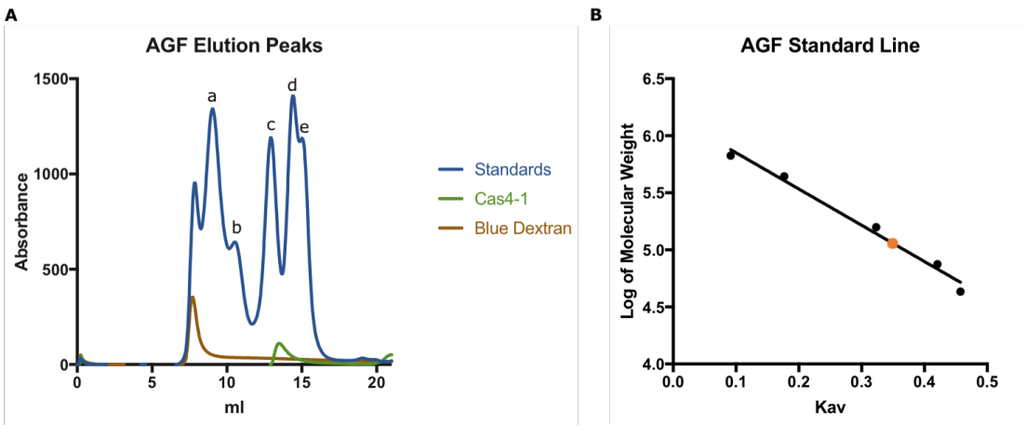

Figure 52: Analytical Gel Filtration Elution Values and Standard Line. A) Elution volumes and absorbance values for standards and Cas4-1 are shown in a graphical form. Different standards are as follows a- Thyroglobulin b- Ferritin c- Aldolase d- Conalbumin e- Ovalbumin. **B)** Standard line where the Kav is plotted against the log of molecular weight. Cas4-1 Kav places it on the line at the orange dot. This gives the molecular weight at around 112kDa.

AGF supports the predicted oligomeric state of a dimer, whereas the BN-PAGE shows a different view. However, in the presence of SDS a dimer is observed on BN-PAGE. Therefore, it may be that the dimer is the most stable oligomer, but that the process of BN-PAGE supports the formation of higher oligomeric structures compared with AGF. From this data, it would appear likely that Cas4-1 is a dimer.

4.7 Investigation of DNA binding by Cas4-1 via EMSAs

Cas1 has been shown to bind DNA and Cas4 is predicted to bind DNA, therefore DNA binding by Cas4-1 was investigated using EMSAs. Before analysis of the ability to bind different substrates, the pH for the binding was optimised. The initial substrate used for exploring the ideal pH was a flayed duplex as Cas1 has been shown to bind to a flayed duplex previously. pH6.0 to pH7.5 were tested as the pH range at which *P. methylaliphatogenes* can survive is between pH4.1-7.8, with an optimum at pH6.5. As seen in Figure 53 the flayed duplex was bound at different efficiencies across all four pH values. The highest binding was observed in pH6.0 and pH7.0 at the second highest concentration of 500nM. pH7.5 had the least aggregation over the concentration range, but (except for 1000nM) it had the lowest binding. Therefore, there is a possibility that this may lead to a decreased ability to detect lower binding capacities with unfavourable substrates. pH6.0 and pH7.0 were the best candidates for binding, and both are the same distance from the optimum pH of 6.5. pH7.0 was selected due to less aggregation overall. The preparation of Cas4-1 for pH optimisation and some initial EMSAs was depleted, so

for following EMSAs a new preparation of Cas4-1 was used with better binding capacities.

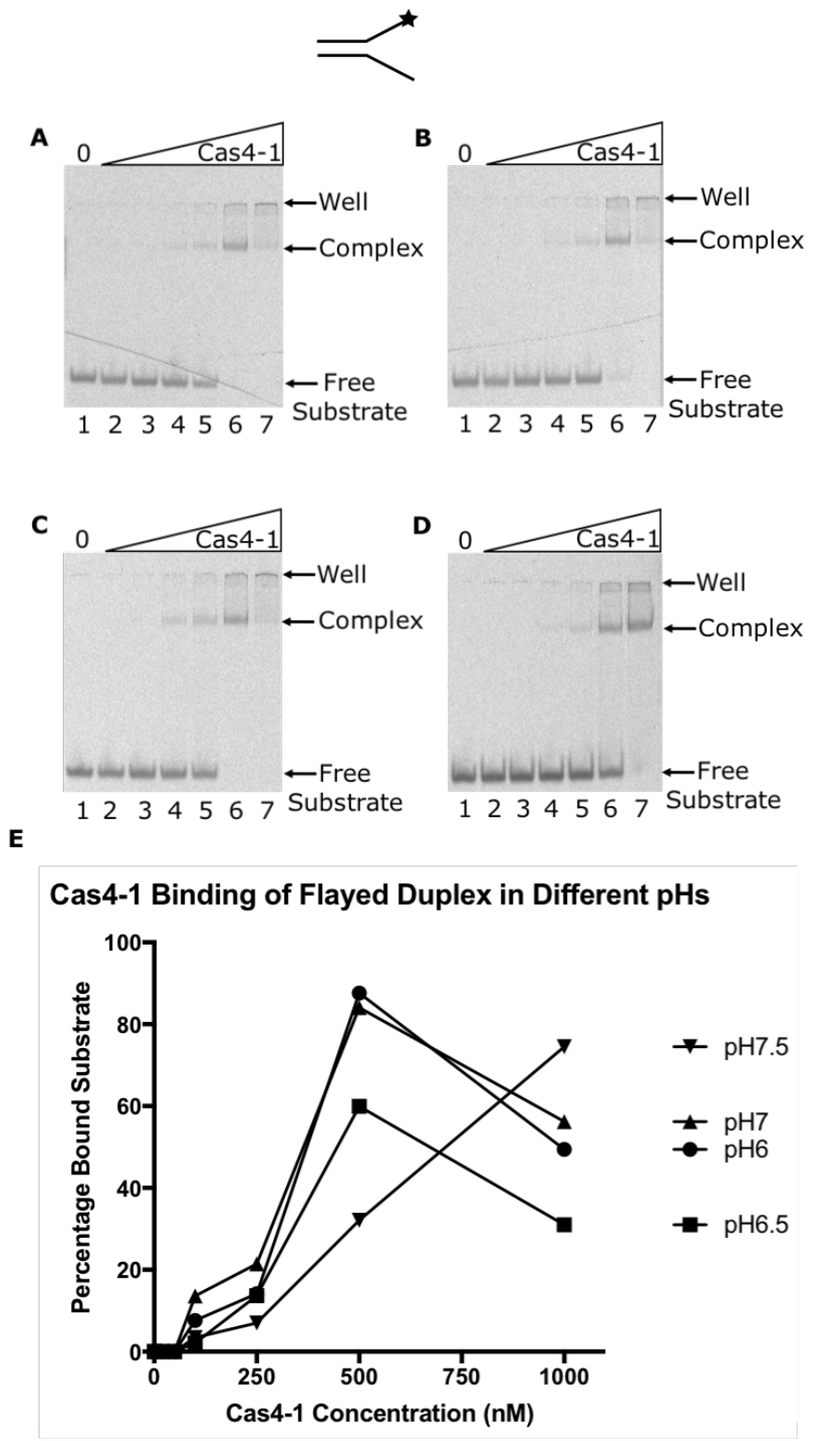

Figure 53: Cas4-1 Binding of Flayed Duplex Over Different pHs. Binding of a flayed duplex by Cas4-1 was tested over four different pHs: 6 **(A)**, 6.5 **(B)**, 7 **(C)** and 7.5 **(D)**. The EMSA gels were analysed using ImageJ and the percentage bound substrate was plotted against the protein concentration **(E)**. The best binding was observed at pH6 and pH7 at 500nM (Lane 6), both aggregated at the highest concentration, but pH7 to the lesser degree. Whilst pH7.5 had the best binding at the highest concentration (Lane 7), but there was still unbound substrate present and binding at lower concentrations was poor.

Initial binding experiments looked at the binding of minimal DNA substrates: ssDNA, dsDNA, DNA with a 5' overhang and DNA with a 3' overhang. Seen in Figure 54, Cas4-1 bound best to ssDNA and DNA with a 3' overhang, slightly to DNA with

a 5’ overhang and hardly to dsDNA. It was expected that Cas4-1 would bind to ssDNA and the overhangs, as Cas1 has been shown to bind ssDNA and 3’ overhangs and Cas4 is expected to bind 5’ overhangs due to its 5’-3’ exonuclease activity, so the results are unexpected as Cas4-1 struggled to bind the 5’ overhang. Therefore, Cas4-1 may have different activity to Cas4.

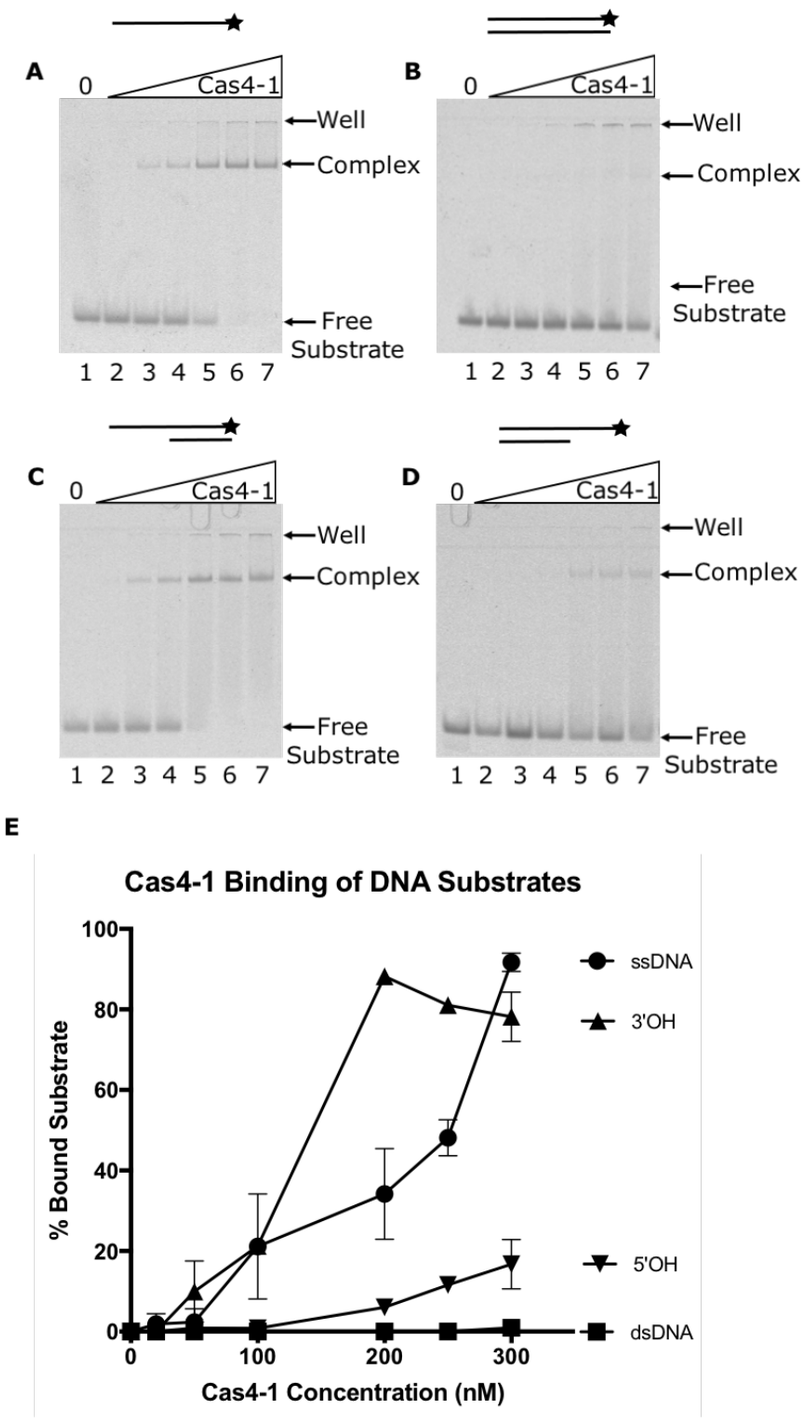

Figure 54: Cas4-1 Binding to Minimal DNA Substrates. Binding assays of Cas4-1 were carried out with ssDNA **(A)**, dsDNA **(B)**, DNA with 5’ overhang **(C)** and DNA with 3’ overhang **(D)**. These gels were analysed using ImageJ and presented in a graphical form **(E).** Cas4-1 bound best to ssDNA and DNA with a 3’ overhang. It slightly bound to the 5’ overhang, and slightly to dsDNA.

Substrates used above were labelled at the 5’ end, and as Cas4-1 was expected to bind to a 5’ overhang substrate, it was investigated whether the 5’ label interfered with Cas4-1. A 3’ cy5 labelled 5’ overhang substrate was made and EMSAs carried out. Figure 55 shows that moving the cy5 label dramatically increases the binding by Cas4-1 with an increase from 6% to 86% for 200nM Cas4-1 (Lane 5). Clearly cy5 interferes with binding of the 5’ overhang substrate. EMSAs were also carried out with a 3’ labelled 3’ overhang (not shown), however whilst there was a reduction in binding below 250nM the interference was overcome at 300nM resulting in the same binding. Interference was not consistent across substrates, but that is likely due to how Cas4-1 binds the substrates as opposed different effects by the same label. No research into interference by cy5 could be found in the literature. Cy5 labelling does interfere with binding here, but not enough is known about how Cas4-1 binds the DNA to make a conclusion to how cy5 is interfering.

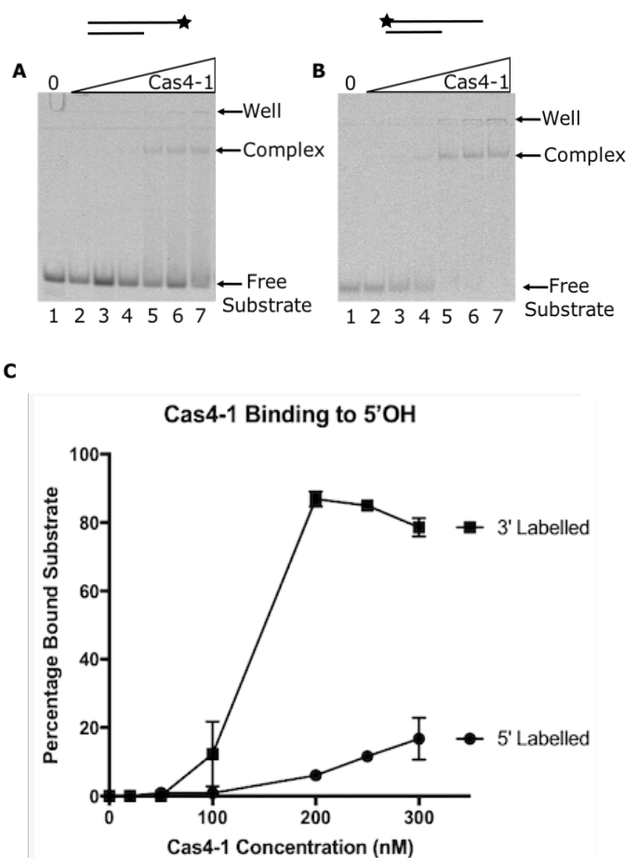

Figure 55: Effect of Labelling on Cas4-1 Binding. Cas4-1 binding was tested on both a 5’ labelled 5’ overhang **(A)** and a 3’ labelled 5’ overhang **(B).** As seen on the graph **(C)** Cas4-1 bound more readily to the 3’ labelled 5’ overhang than the 5’ labelled.

The next set of substrates to be tested were forked substrates. CRISPR implicates the involvement of replication forks in CRISPR-Cas, as spacers are often obtained from sites of stalled replication. Also during integration of a new spacer, a forked

substrate is created and bound by Cas1. EMSA gels and a summarising graph are seen in Figure 56. Cas4-1 bound readily to forks with ssDNA present (flayed duplex, leading and lagging strand forks), but struggled to bind to a full fork. This was consistent with the minimal DNA substrates, in that Cas4-1 bound substrates with a ssDNA region (the region was the same length in both substrates) but not to dsDNA. Unlike the 5' overhang substrate, the leading strand fork binding was not affected by the cy5 label. This could be due to a different binding mechanism for forks, or that each substrate was bound by a different binding site (Cas4 or Cas1).

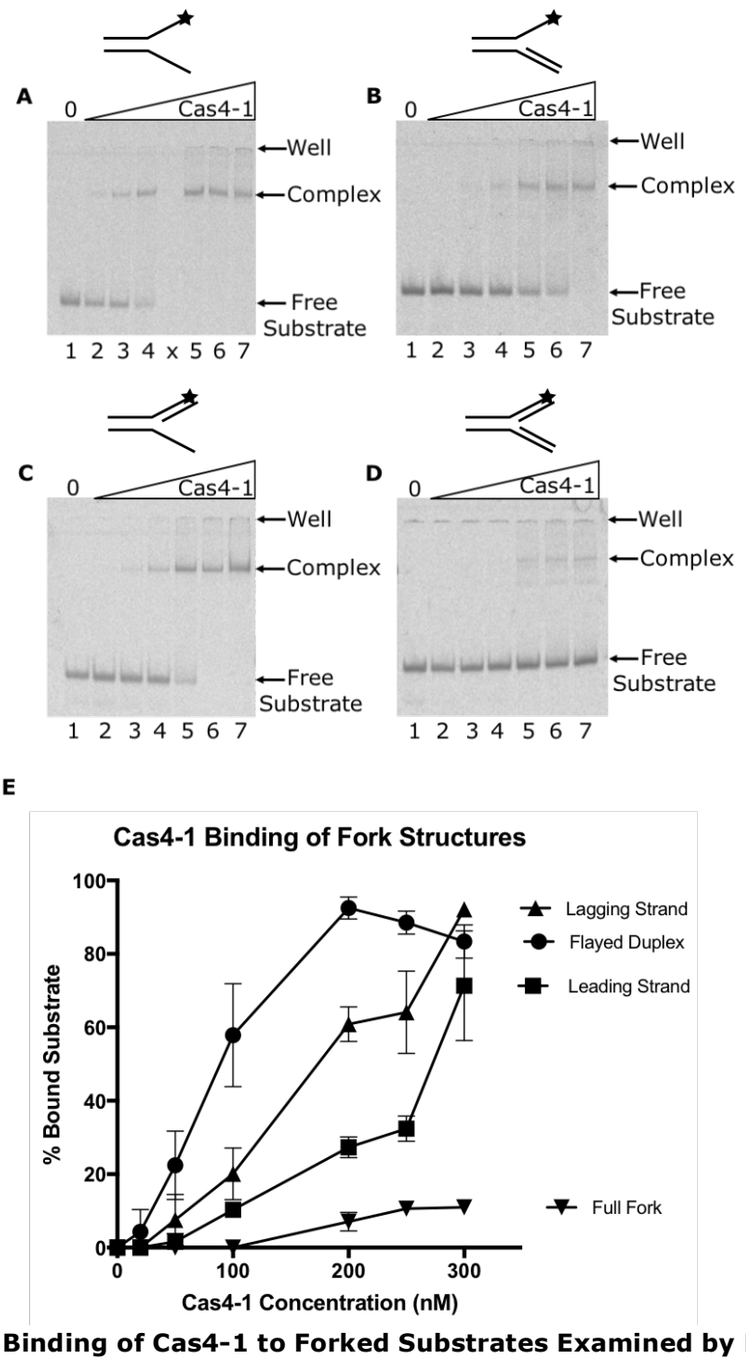

Figure 56: Binding of Cas4-1 to Forked Substrates Examined by EMSAs. A flayed duplex **(A)**, leading strand fork **(B)**, lagging strand fork **(C)** and full fork **(D)** were tested for Cas4-1 binding. Gels were analysed by ImageJ and presented in a graph **(E)**. Cas4-1 bound to all forks with ssDNA present (flayed duplex, leading and lagging strand forks), but struggled to bind the double stranded full fork.

4.8 Degradation of Cas4-1 and its Effect on DNA Binding

As mentioned in the previous section the pH optimisation was carried out with a different preparation of Cas4-1 than the bulk of the EMSAs. EMSAs for the forked substrates were initially begun using the first preparation, but as stocks grew short a new preparation was made and used. As detailed in Figure 57, the first preparation of Cas4-1, which was produced over a 3-day purification, had peak binding at 500nM. Whereas the second preparation, which was produced over a single day, had a peak binding at 200nM. It is likely that Cas4-1 was degrading during the purification and shortening the purification limited the amount of degradation. This is something to be considered later in this work, as until a true value of activity (i.e. Units/ml) can be defined for this protein small differences between preparations can be due to degradation as opposed to true differences.

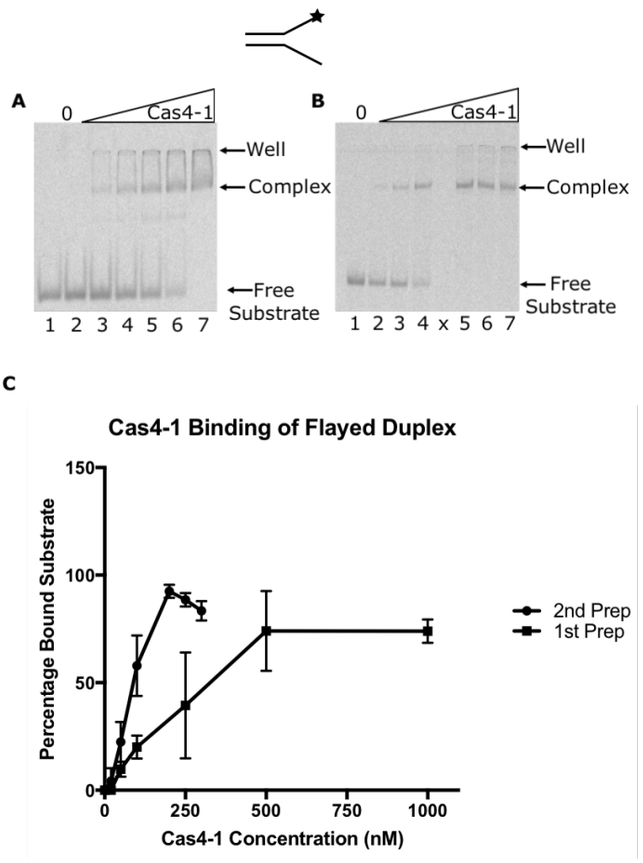

Figure 57: Comparison of Binding Ability by two preps of Cas4-1. A) Cas4-1 protein from the first purification, produced over 3 days, reached peak binding at 500nM. **B)** Cas4-1 protein from the second purification, produced during a single day, reached peak binding at 200nM. **C)** The graph shows the difference in preparations binding, with the second preparation binding at lower concentration and better.

4.9 Analysis of Cas2 and HPS DNA Binding by EMSAs

Cas2 has not been shown to bind DNA alone, only when complexed with Cas1. Cas2 binding assays were carried out to see if this Cas2 behaves as previously studied Cas2s. If Cas2 cannot bind DNA, then EMSAs can be used to investigate complex formation between Cas4-1 and Cas2 in the form of supershifting. This is where Cas4-1 binding to DNA is compared to Cas4-1 and Cas2 binding to DNA. If the shifted band dictating the complex shifts higher than Cas4-1 alone, then a complex between Cas4-1 and Cas2 has formed to bind the DNA. As can be seen in Figure 58 Cas2 binds to none of the substrates.

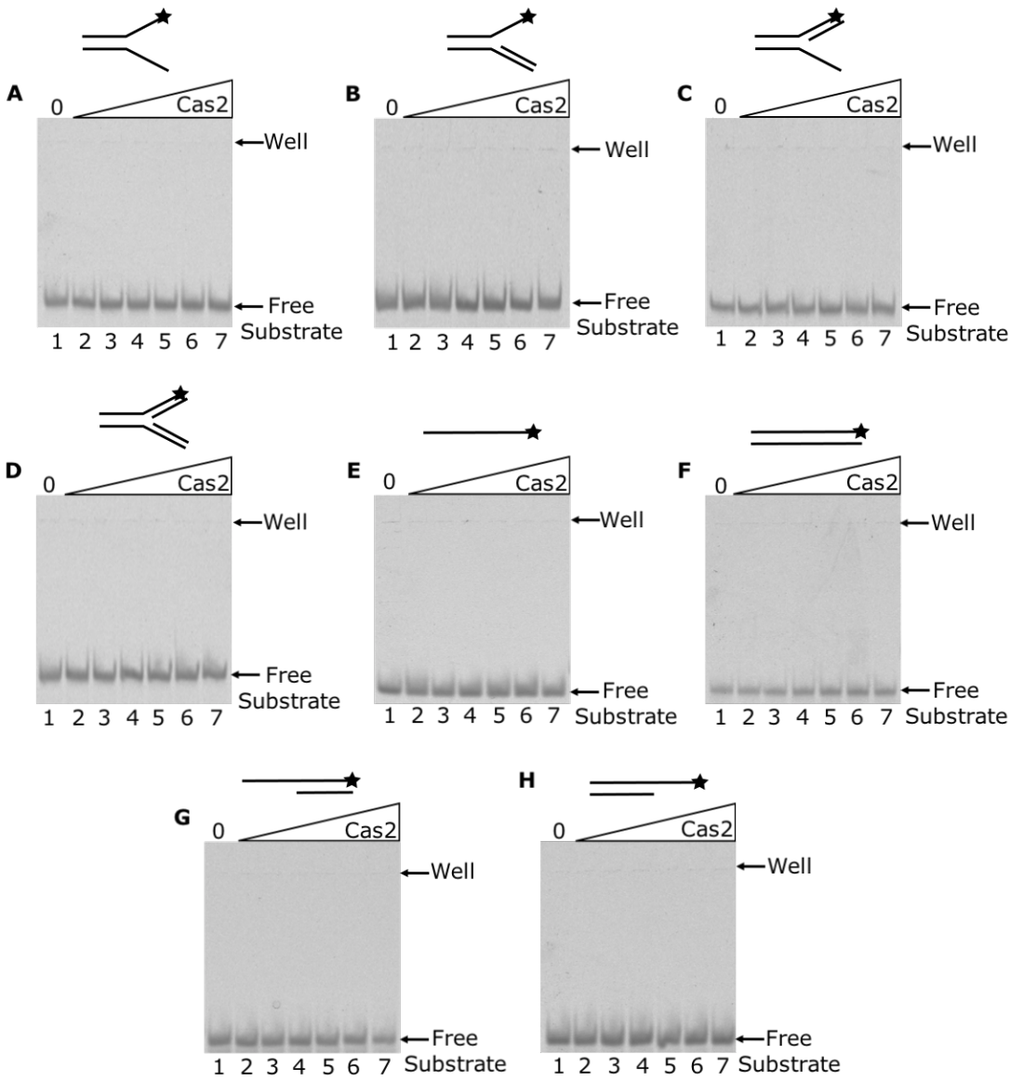

Figure 58: EMSA Analysis of Cas2 Binding. Cas2 bound to none of the substrates provided: flayed duplex **(A)**, leading strand fork **(B)**, lagging strand fork **(C)**, full fork **(D)**, ssDNA **(E)**, dsDNA **(F)**, 3' overhang **(G)** and 5' overhang **(H).**

Following this supershifting was attempted, but when Cas4-1 and Cas2 were incubated together there was no difference from Cas4-1 alone. Under the conditions of the EMSA no interaction was seen, this does not conclusively prove a

lack of interaction between Cas2 and Cas4-1. It could be Cas4-1 and Cas2 only interact transiently which cannot be observed in an EMSA, or that they only interact when a certain DNA molecule or conditions are encountered.

HPS was also tested for DNA binding. As discussed previously the role of HPS is unknown, therefore no hypothesis about the binding ability could be made. Although the predicted structure from Phyre2 was constructed with a ferredoxin-like fold from a transcription factor. A ferredoxin-like fold is often found in DNA binding proteins. However, HPS, like Cas2 did not bind to any of the substrates tested, Figure 59.

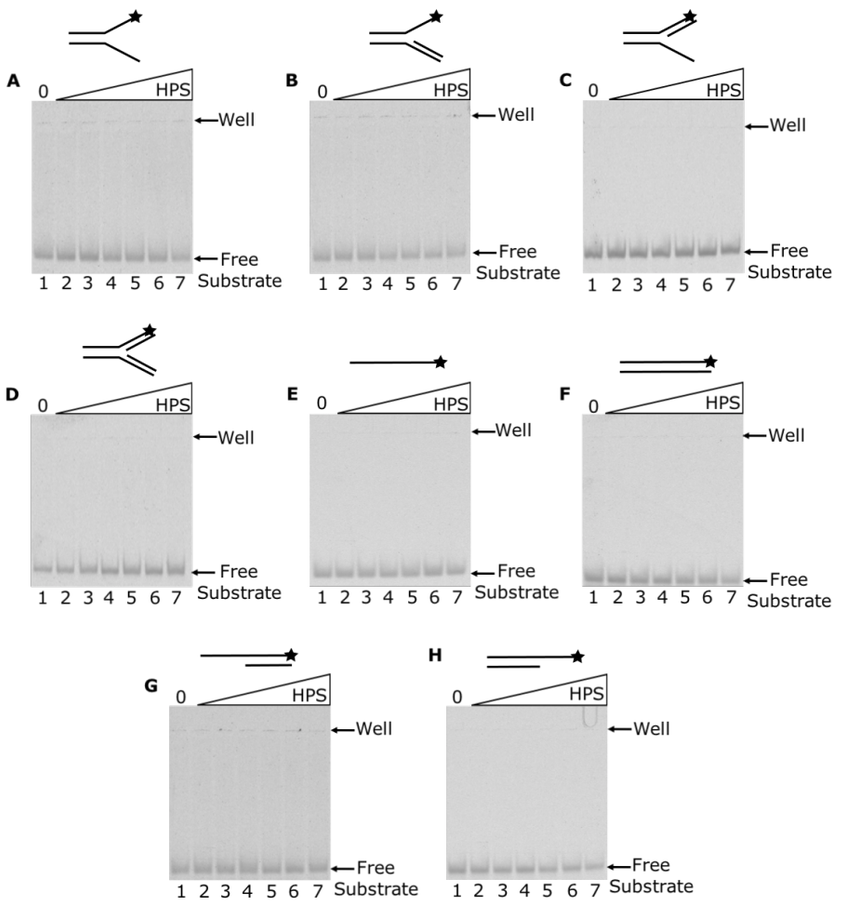

Figure 59: EMSA Analysis of HPS Binding. HPS bound to none of the substrates provided: flayed duplex **(A)**, leading strand fork **(B)**, lagging strand fork **(C)**, full fork **(D)**, ssDNA **(E)**, dsDNA **(F)**, 3' overhang **(G)** and 5' overhang **(H).**

HPS was also tested for the ability to supershift/form a complex with Cas4-1, but again as with Cas2, no difference was seen between Cas4-1 with HPS and Cas4-1 alone.

4.10 Exploring Nuclease Activity of Cas4-1 against M13 ssDNA and pUC18 dsDNA

Cas4 has been shown to cleave M13 in various publications (Zhang *et al*, 2012; Lemak *et al.*, 2013; Lemak *et al.*, 2014). Thus, the ability of Cas4-1 to cleave M13 was tested (Figure 60). Over a concentration range of 0-4μM Cas4-1, M13 was cleaved. 1.5μM Cas4-1 (Lane 6) was required to see a decrease in M13. No product was seen for the cleavage, but there was smearing showing degradation of M13. It is likely that the product produced is too small for visualisation on an agarose gel. There was a band present just below the 1kB marker in lanes containing high concentrations of Cas4-1 (Lanes 1, 9-11). As this was present in the EDTA control lane, where there was no degradation it is unlikely that this was a product. This was investigated, and it was shown that this DNA/RNA molecule was associated with Cas4-1 after purification. The investigation was inconclusive as to whether this was DNA or RNA.

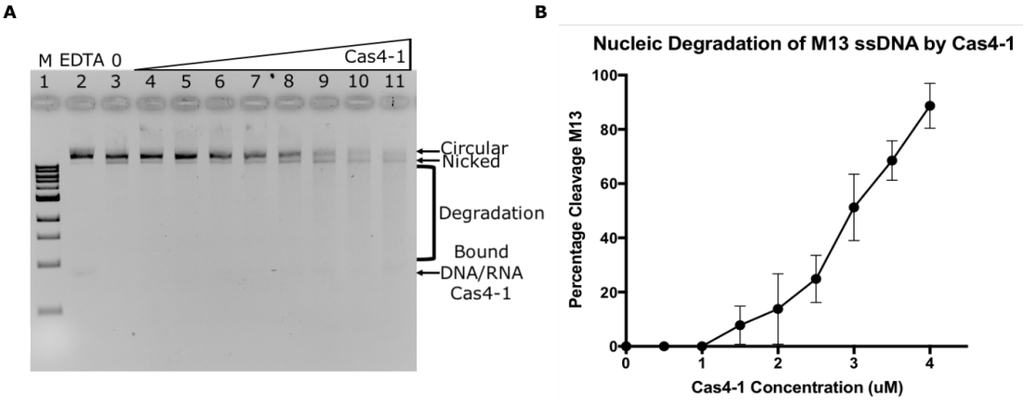

Figure 60: Nuclease Degradation of M13 by Cas4-1. A) Degradation of M13 by Cas4-1 was analysed on an agarose gel. Two bands were present in lanes which contained Mg^{2+} (3-11) which correspond to circular and nicked DNA. There is no singular product formed by the nuclease activity, just smearing on the gel. It is likely that the products are too small for analysis on an agarose gel. There was also a small band, present below the 1kB marker which was present in the EDTA control lane along with lanes 9-11, this was not a product but some DNA/RNA bound to Cas4-1. **B)** Graphical representation of cleavage activity, cleavage was first observed at 1.5μM Cas4-1 and steeply increased after this.

The concentration required for M13 cleavage was a high concentration of Cas4-1, even higher than the concentration required for *M. harundinacea* Cas4-1. It could be that Cas4-1 is not a particularly active enzyme, or that its activity has degraded over the purification and experiment. It could also be a background nuclease that has purified with Cas4-1 is also responsible for this degradation. Mutational studies were carried out later in this thesis to test the Cas4-1 active site for this activity.

Though Cas4 has not been shown to cleave dsDNA, a dsDNA substrate (pUC18) was tested to see if Cas4-1 can cleave it. Figure 61 shows that Cas4-1 can degrade pUC18 from 0.5μM of Cas4-1. No products are seen, but as with M13 it is likely they are too small to detect on an agarose gel. Cas4-1 has been shown to bind dsDNA poorly, so the mechanism of action is unknown. Cas1 has been shown to nick dsDNA, so this could potentially open up the dsDNA for binding and cleavage by Cas4.

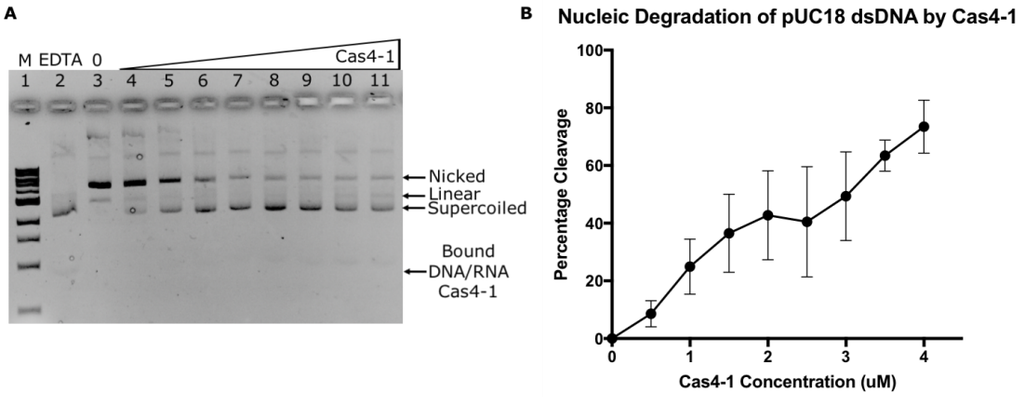

Figure 61: Nucleolytic Degradation of pUC18 by Cas4-1. A) As with M13 degradation, pUC18 degradation was analysed on an agarose gel. As with M13, nicked and linear products are present in the Mg^{2+} buffer. No degradation products are formed, though the substrate is decreased. **B)** The graph created after analysis of gels by ImageJ shows a decrease in pUC18 from 0.5μM Cas4-1, with increases with concentration.

Though no complex was seen between Cas4-1 and Cas2 in EMSAs, the effect of Cas2 on Cas4-1 activity was tested, using M13 as a substrate. Cas2 did not cleave M13 on its own, but when added to Cas4-1, increased the nucleic degradation of Cas4-1 (Figure 62). The mechanism of this increase or how this fits into the CRISPR-Cas mechanism is unknown. It is also possible that this activity is due to contamination of Cas4-1 or Cas2 preparation with a nuclease or topoisomerase

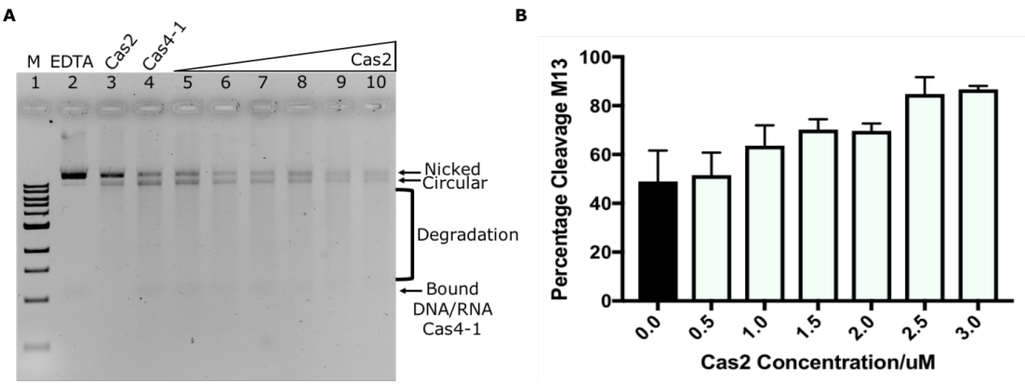

Figure 62: Effect of Cas2 on Cas4-1 degradation of M13. A) Degradation of M13 was carried out with 3μM Cas4-1 and increasing concentration of Cas2 from 0-3μM. Cas2 alone did not cleave DNA, but when added to Cas4-1 nucleic degradation increased. **B)** Analysis by ImageJ produced a graphical representation. The first bar shown in black shows the level of degradation 3μM Cas4-1, the remaining bars show the level of degradation upon the addition of Cas2. An upwards trend is seen across the samples.

4.11 Generation of Active Site Mutants by Site-Directed Mutants

Cas4-1 contains conserved residues for 3 active sites. Mutations were carried out for each residue of each active site to look at the role of each active site in Cas4-1. The residues that were mutated and their position in the predicted structure are shown in Figure 63. Primers were designed to mutate Cas1 and Cas4 active site residues to alanines and 4Fe:4S cysteines to serines. The Cas1 and Cas4 active site residues were altered to alanines as alanine is a non-reactive amino acid, so will remove the ability of that residue to catalyse reactions. The four cysteines that make up the 4Fe:4S cluster were mutated to serine, as serine is a very similar to cysteine but it lacks a sulfur. As only the sulfur is needed for the 4Fe:4S cluster serine removed this ability while keeping an amino acid of a similar size.

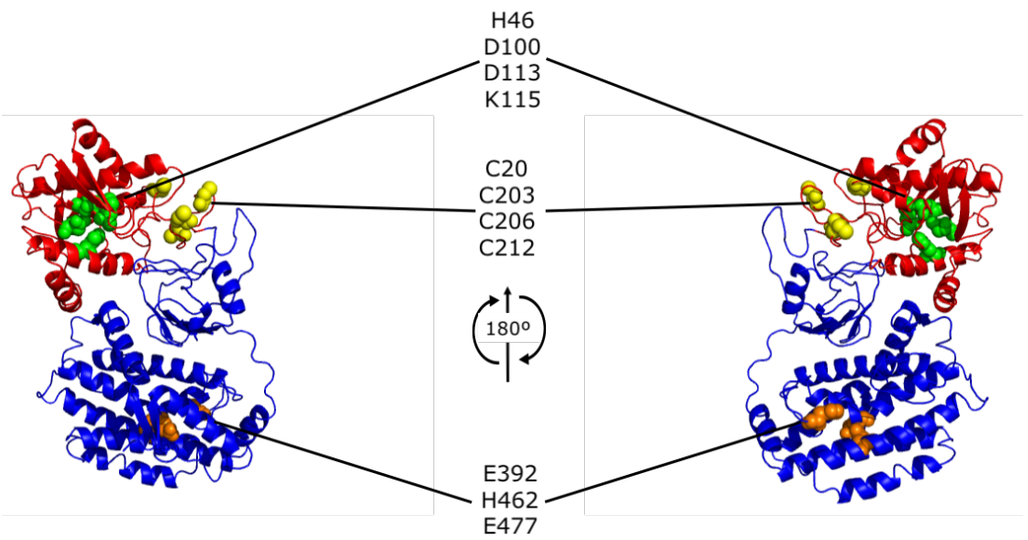

Figure 63: Positions of Active Site Residues. Three active sites are predicted in Cas4-1: Cas4 (Green), 4Fe:4S cluster (Yellow) and Cas1 (orange). The conserved residues for these active sites are shown in the centre.

Primers designed for site-directed mutagenesis were used in conjugation with NEB's Q5 mutagenesis kit (see section 2.8.4). Three of the PCRs were unsuccessful: H46A, D100A and C206S. The remaining mutations were created and verified by sanger sequencing (Source Bioscience).

Expression of proteins was possible for all mutations. Mutant proteins were expressed in the same way as the WT proteins. Overexpression of each mutant is shown in Figure 64.

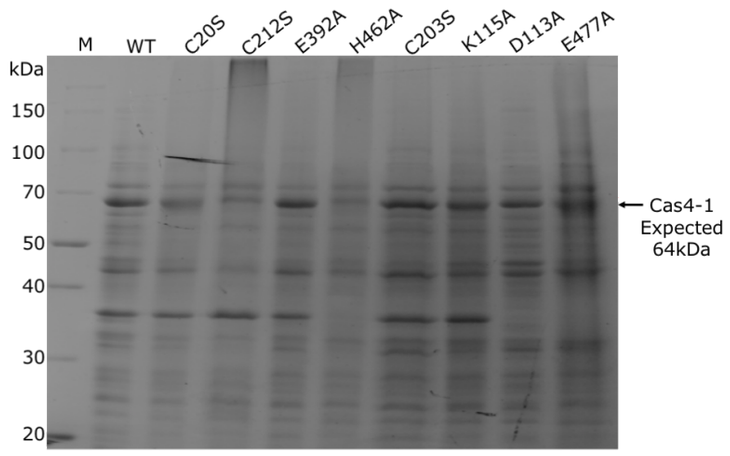

Figure 64: Summary of Mutant Overexpression. Overexpression samples from all 8 mutants were compared using SDS-PAGE. All mutants were the same size as the WT protein at around 64kDa.

Two mutant proteins were successfully purified using the exact method used for purifying the WT protein (section 2.10.6), C20S and K115A shown in Figure 65. 200ng of Cas4-1 and K115A and 50ng C20S were migrated on the gel. The reason for running less C20S was due to a lower concentration of protein available. An attempt was made to run all samples at 50ng, but the bands were weak and did not give a good image. Therefore Cas4-1 and K115A were run at a higher amount to allow better visualisation.

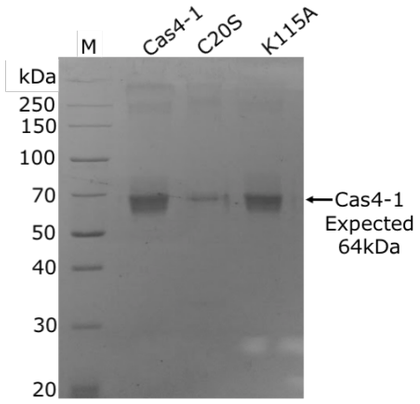

Figure 65: Purified Cas4-1 and mutant proteins. Cas4-1 and mutant proteins (C20S and K115A) were run together on a 10% acrylamide gel. 200ng of Cas4-1 and K115A and 50ng of C20S were run. Less C20S was used as the concentration of the sample was low. A gel was attempted with 50ng of all sample but the bands were not clear enough.

Three further purification were attempted for D113A, E392A and H462A but in all purifications the proteins aggregated during the desalting step. With time, it should be possible to alter the protocol to allow purification of these proteins.

4.12 Investigation of Effect of Active Site Mutants on DNA Binding Using EMSAs

DNA binding of C20S and K115A was tested using EMSAs for the same substrates tested for WT Cas4-1. The binding of each mutant will be explored first and a comparison of the binding of all three proteins will follow.

C20S, was a mutation in one of the conserved cysteine residues involved in the 4Fe:4S cluster. No brown solution was produced during C20S purification, suggesting a loss of the 4Fe:4S cluster. C20S binding to minimal substrates was tested using EMSAs, Figure 66 shows that C20S struggled to bind to any of the minimal substrates provided.

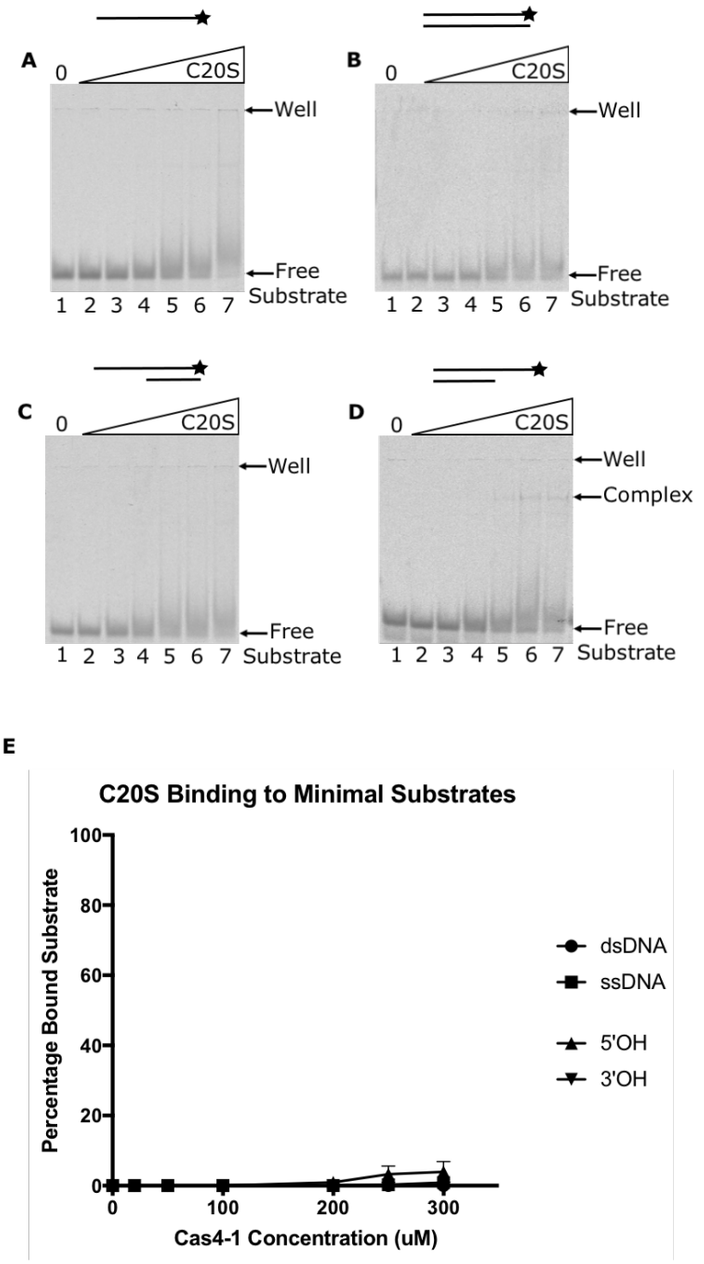

Figure 66: C20S Binding to Minimal DNA Substrates. Binding assays of C20S were carried out with ssDNA **(A)**, dsDNA **(B)**, DNA with 5' overhang **(C)** and DNA with 3' overhang **(D)**. These gels were analysed using ImageJ and presented in a graphical form **(E).** C20S struggled to bind to any of the substrates provided.

Next binding of forked substrates was tested. As seen in Figure 67, in comparison to the minimal substrates, C20S was capable of binding to forked substrates. As seen with the WT, C20S preferred to bind forks with ssDNA present.

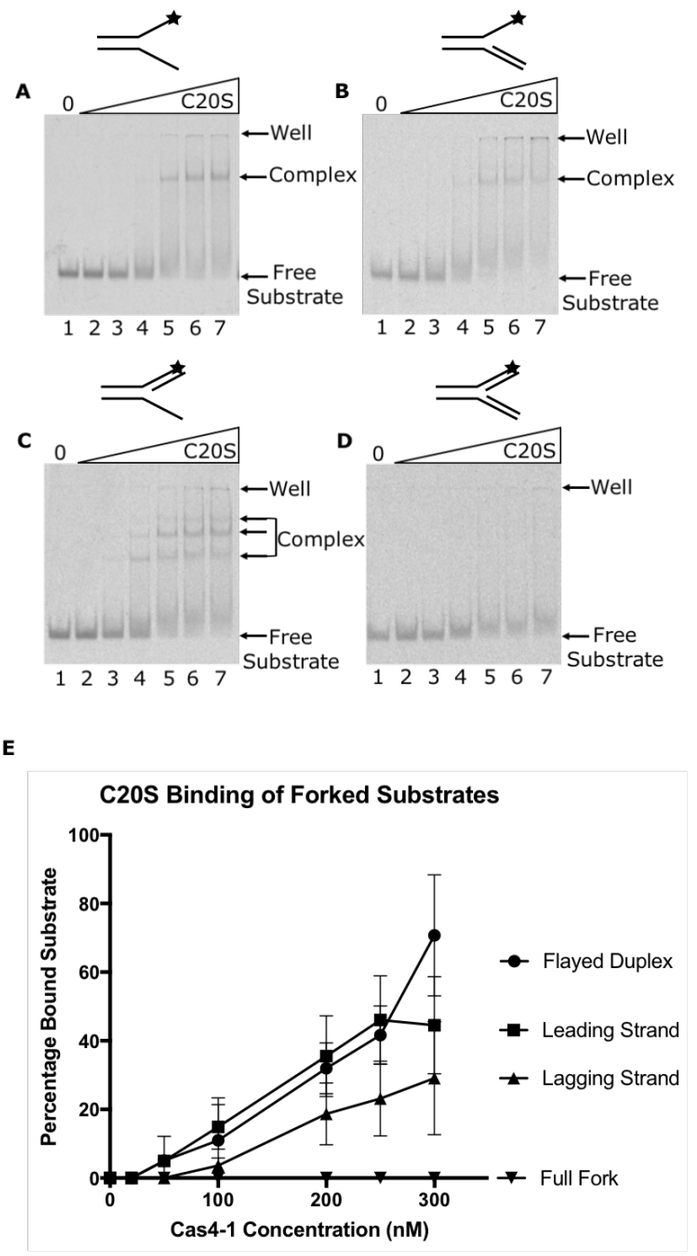

Figure 67: Binding of C20S to Forked Substrates Examined by EMSAs. A flayed duplex **(A)**, leading strand fork **(B)**, lagging strand fork **(C)** and full fork **(D)** were tested for Cas4-1 binding. Gels were analysed by ImageJ and presented in a graph **(E)**. C20S bound best to a flayed duplex, but bound all forks with ssDNA present. It had little to no binding of the full fork.

K115A, a mutation in a Cas4 active site residues, was not predicted to affect DNA binding. To test this K115A binding of minimal substrates was measured using EMSAs. Figure 68 shows the results of the binding, K115A bound well to ssDNA and 3’ overhangs, less so to a 5’ overhang and not at all to dsDNA. This was similar to WT binding.

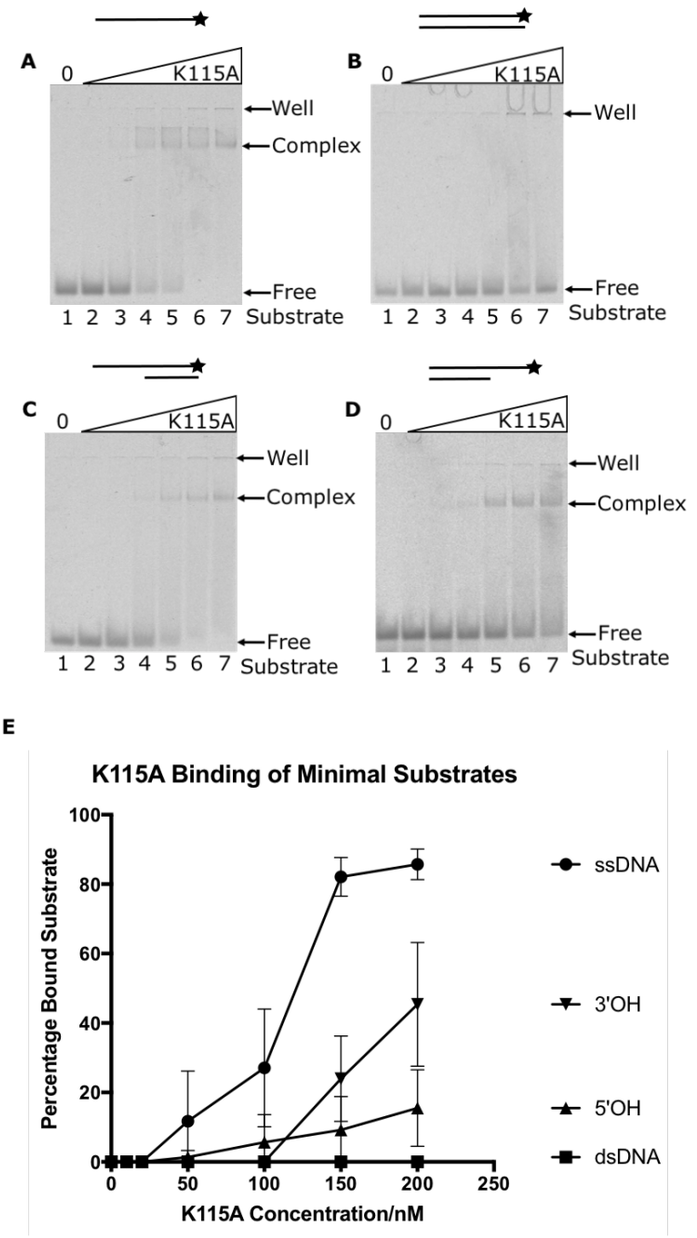

Figure 68: K115A Binding to Minimal DNA Substrates. Binding assays of K115A were carried out with ssDNA **(A)**, dsDNA **(B)**, DNA with 5' overhang **(C)** and DNA with 3' overhang **(D)**. These gels were analysed using ImageJ and presented in a graphical form **(E).** K115A bound best to ssDNA and 3' overhang best. It also bound slightly to 5' overhang, but not at all to dsDNA.

K115A DNA binding was also tested on forked substrates. K115A, as shown in Figure 69 bound to all forks that contain ssDNA present. This was consistent with WT Cas4-1.

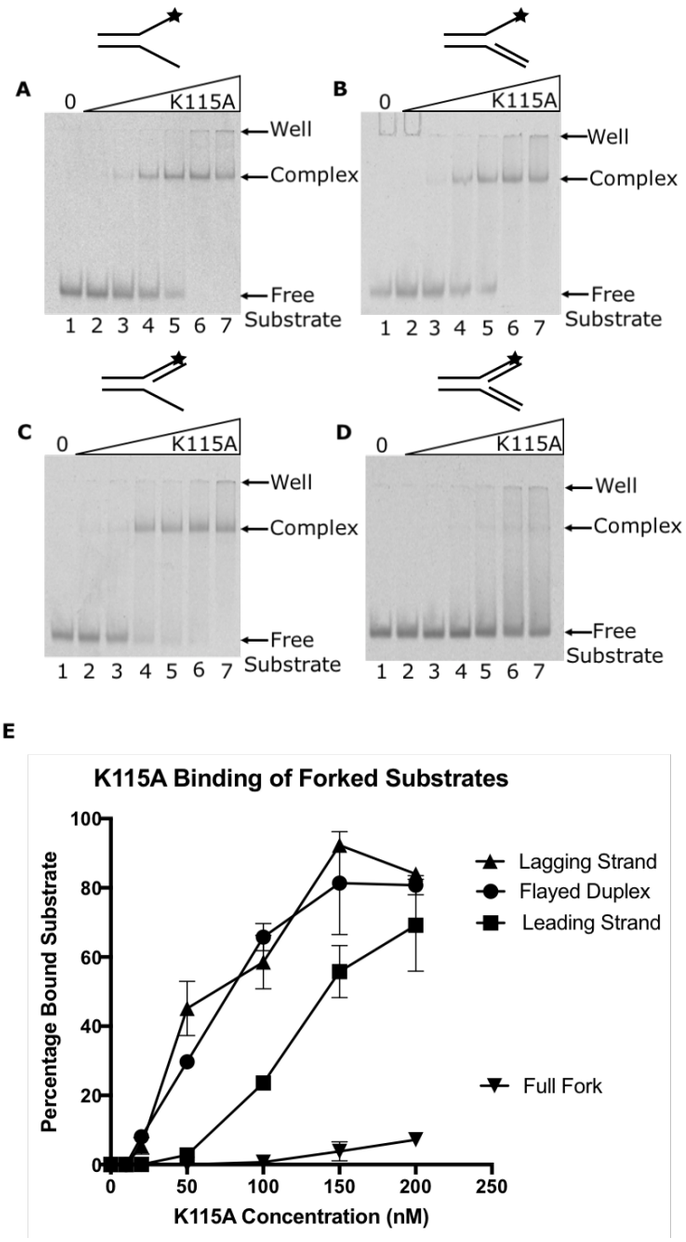

Figure 69: Binding of K115A to Forked Substrates Examined by EMSAs. A flayed duplex **(A)**, leading strand fork **(B)**, lagging strand fork **(C)** and full fork **(D)** were tested for Cas4-1 binding. Gels were analysed by ImageJ and presented in a graph **(E)**. K115A bound all forks with ssDNA present, it struggled to bind to the double stranded full fork.

EMSA and individual graphs for each mutant have been shown and described, the graphs for WT Cas4-1 and mutants are shown for comparison in Figure 70. K115A binding appears mostly unchanged in comparison to WT Cas4-1, which was expected as K115A is a mutation in the Cas4 active site, not the Cas4 binding site. There is some variability in the overall percentage binding of the substrates, but this could be due degradation occurring during purification as discussed earlier. The only result that differs substantially is the binding of the 3' overhang with a decrease from 88% to 44% compared with WT. The next greatest reduction was for binding to the flayed duplex which is reduced from 92% to 81%. The 3' overhang result would seem significant. As this is a mutation in the active site, it

may be that the active site engages with the 3′ end of ssDNA upon binding and this mutation alters its ability to do this. The binding to other substrates may be through a different mechanism that does not require the active site. Another hypothesis is that the mutation changes the fold, which alters its ability to interact with the 3′ end.

C20S has a much more dramatic change in binding ability. The greatest change was observed for the minimal substrates (Figure 70A compared with Figure 70B), where all substrate binding is reduced to under 4%. C20S is predicted to disrupt Cas4's DNA binding domain, as 4Fe:4S clusters are often involved in DNA binding. There was some reduction for the forked substrates, but the only significant reduction was for the leading strand fork from 71% to 46%. This fork contains a 3′ ssDNA end, and the Cas4 active site mutant has been shown to have reduction in the 3′ overhang substrates. It seems Cas4 preferably interacts with 3′ ssDNA ends. For the remaining forks the reduction was no more than 25%, which whilst still quite a reduction could possibly be due to degradation during purification. Though a degradation study was not completed, the loss of the 4Fe:4S cluster is likely to increase instability leading to more degradation. However, it can be concluded that Cas4 DNA binding site binds to both forked and non-forked substrates with a 3′ ssDNA end.

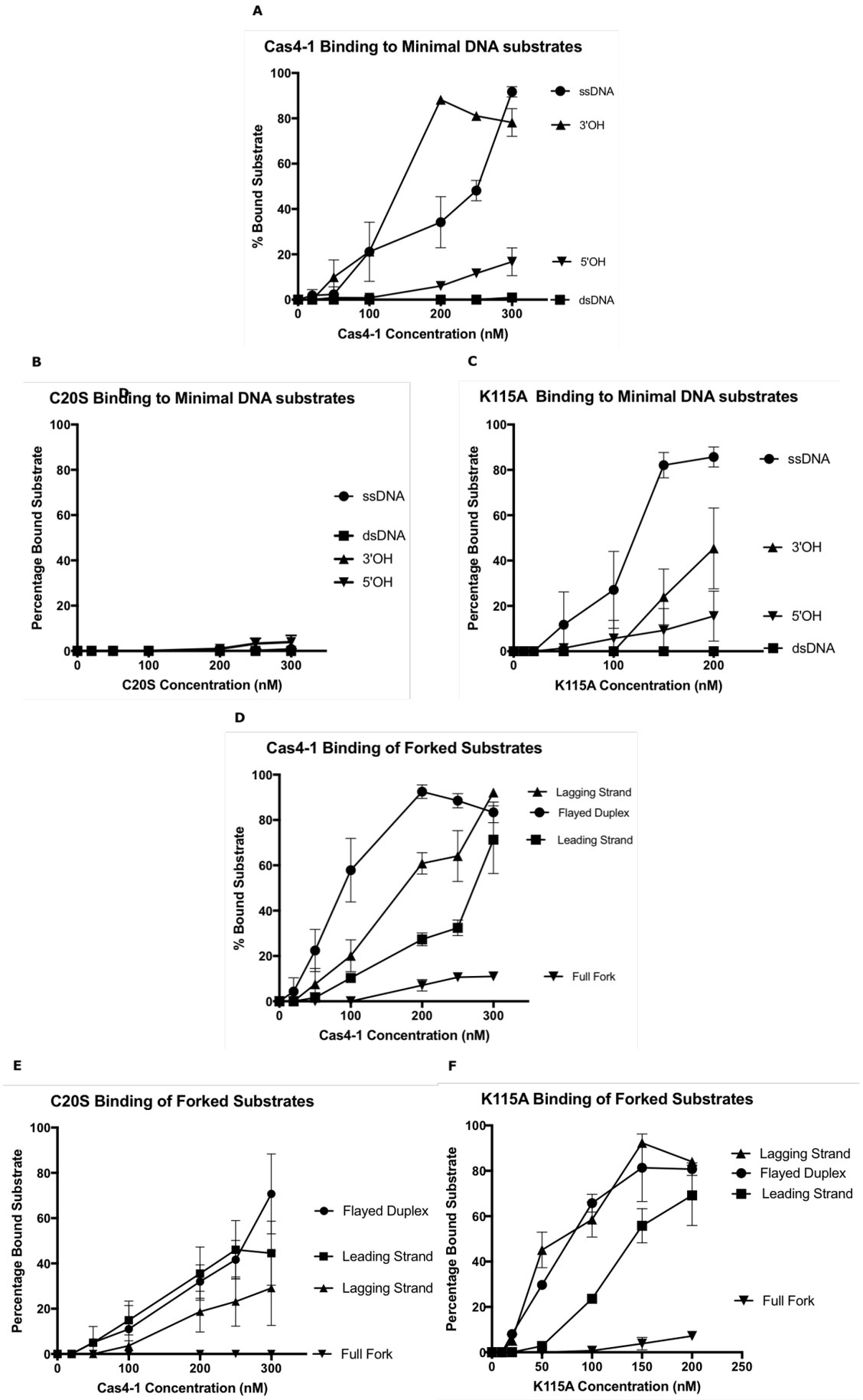

Figure 70: EMSA Summary Graphs for WT Cas4-1 and each mutant for each set of substrates. EMSA binding data is shown for WT-Cas4-1 (A & D), C20S (B & E) and K115A (C & F). C20S struggled to bind minimal substrates, but not the forked substrates. Whereas K115A bound all the substrates that WT Cas4-1 could.

It would be interesting to see whether the substrates unaffected by mutation of Cas4's DNA binding domain, are affected by loss of Cas1's binding site. However, this proves difficult as the DNA binding sites for Cas1 are only known for *E. coli* Cas1, and those residues are not conserved in *P. methylaliphatogenes* Cas4-1. Active site mutations in the Cas1 region of Cas4-1 may also provide some insight, but purification of these mutants has proved unsuccessful at this stage. For a better understanding any future research should look to complete a Cas1 active site mutant purification and analysis by EMSA.

4.13 Analysis of Cas4-1 Mutants Nuclease Activity against M13 and pUC18

The 4Fe:4S cluster mutant, C20S, was unable to be tested in nuclease assays due to the low purification concentration. C20S was purified at 1μM and the assays have been carried out with a range between 0.5-4μM, with only a small amount of activity observed at 1μM. Therefore, it would be impossible to tell if a lack of activity at 1μM was due to a loss of activity in the protein, or simply that the activity did not occur until a higher concentration. Purification was re-attempted to produce a higher concentration, but the concentration was not improved. C20S expressed poorly compared to WT Cas4-1, so even purification from an 8L overexpression culture did not improved the purification concentration.

K115A, Cas4 active site mutant, however was of a sufficient concentration (10μM) to be tested. As with WT, M13 circular ssDNA and pUC18 dsDNA were tested with K115A. Figure 71 shows K115A nuclease activity against M13. The mutation does not eliminate cleavage, as the substrate does decrease as K115A concentration increases. But as seen in Figure 71B, cleavage by K115A was decreased to 40% from 80% compared to WT at the highest concentration. Some Cas4 activity appears to be able to occur without K115A, so it is not an essential residue for the active site.

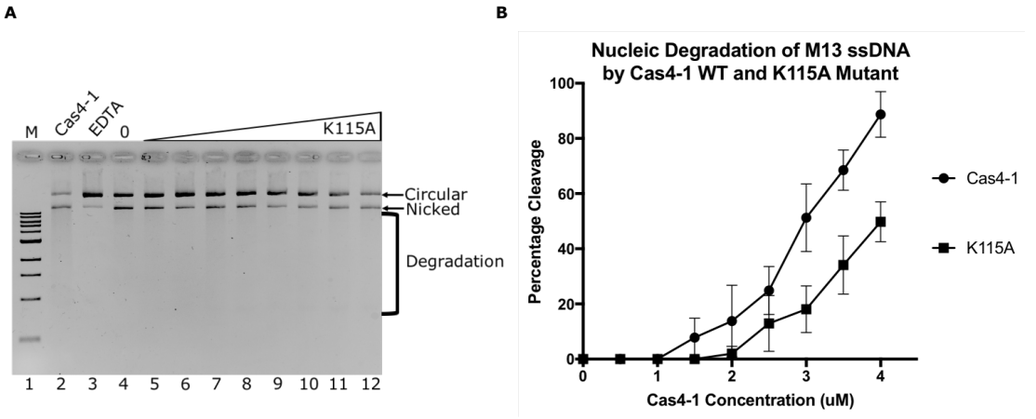

Figure 71: Nuclease Degradation of M13 by K115A A) Degradation of M13 by K115A was analysed on an agarose gel. Two bands were present in lanes (3-11), these correspond to circular and nicked DNA. There is no singular product formed by the nuclease activity, just smearing on the gel. It is likely that the products were too small for analysis on an agarose gel. **B)** Graphical representation of cleavage activity comparing K115A to WT Cas4-1. Cleavage by K115A was not detected until 2μM, and was decreased by 39% in comparison to WT at 4μM.

K115A nuclease activity against pUC18 was also tested. As shown in Figure 72 this result contrasts with M13 results. No smearing was seen indicating cleavage, instead as protein concentration was increased more supercoiled DNA was seen. As Mg^{2+} causes nicking in DNA, K115A may still be able to bind pUC18 without cleaving essentially protecting the DNA from the Mg^{2+}. This data would suggest that K115 residue is more involved in ssDNA cleavage than dsDNA cleavage.

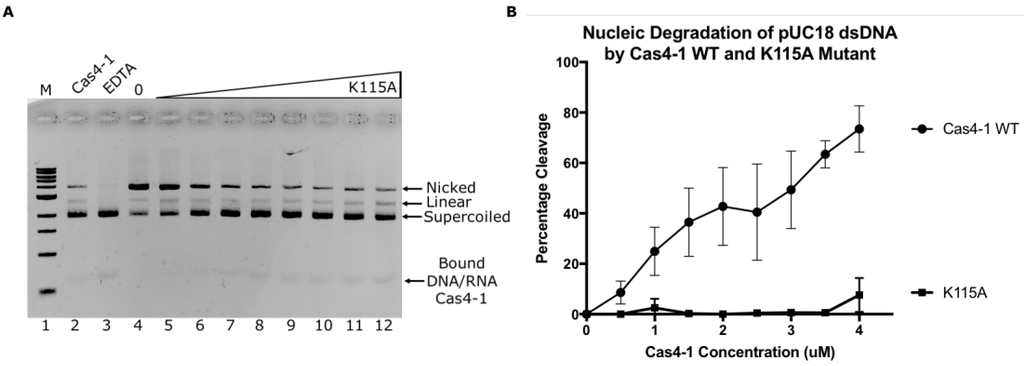

Figure 72: Nucleolytic Degradation of pUC18 by K115A. A) As with M13 degradation, pUC18 degradation was analysed on an agarose gel. As with M13, nicked and linear products are present in the Mg^{2+} buffer. No degradation products were formed and there is no decrease in DNA. Instead of degradation, there was an increase of supercoiled DNA as the protein concentration increased. **B)** The graph created after analysis of gels by ImageJ showed little to no cleavage with K115A.

4.14 Creating a single *in vitro* reaction for Adaptation

The ultimate aim of this project was to create a single *in vitro* reaction for adaptation. Having purified 3 out of 4 of the proteins from the *P. methylaliphatogenes* gene neighbourhood, a SPIN (spacer integration) assay was attempted (method is detailed in section 2.15). The premise of this assay is that Cas4-1 will degrade pUC18 creating spacer products for integration. In combination with Cas2 (possibly HPS), Cas4-1 will integrate this spacer into the *P. methylaliphatogenes* CRISPR locus which is located on a plasmid. Different combination of proteins and protein concentration were used along with different incubation orders. Figure 73 shows the results of the SPIN assay shown on an agarose gel. No real difference was seen between any of the lanes, there was potentially a drop in the amount of DNA in lane 11-13, though only slight. If a spacer has been integrated, it was unlikely to be fully integrated as that would require gap filling by a DNA polymerase. Therefore, integration would be expected to form a linear plasmid of ≈3kDa. However, if only a few integration events have occurred this may not be observable on the gel. Detection by PCR was attempted, however the concentration of pCRISPR within the sample when further diluted in the PCR mix was too low for detection. There was a further issue in that as full integration with gap filling cannot occur the PCR would not be able to proceed across the gaps. As not all pCRISPRs would be expanded to include a new spacer, the only product would be an amplification of pCRISPR with no expansion. A better detection method is required.

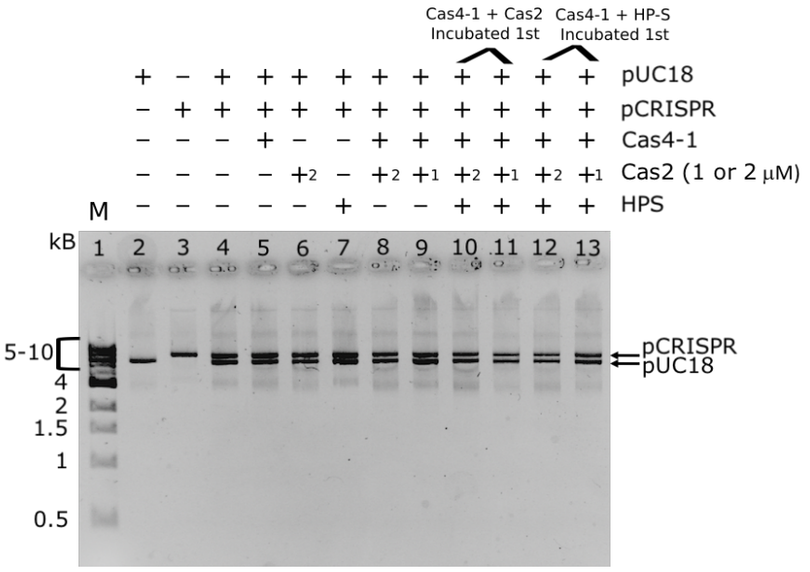

Figure 73: Spacer Integration Assay. An attempted SPIN assay where degradation of pUC18 was tested to provide spacers for integration into pCRISPR. The assay used 2μM Cas4-1, 2μM HPS and 1 or 2μM Cas2. No change was seen in any lane regardless of condition. Improvement and optimisation are required.

Chapter 5: Summary Discussion and Future Research

5.1 Discussion

After little success with the *M. harundinacea* system, the *P. methylaliphatogenes* system proved more fruitful. The organism was chosen because it contained a Cas4-1 fusion and was an aerobic bacterium. This meant the Cas4-1 was less likely to degrade upon exposure to oxygen, as it functioned in an oxygen containing environment. Also as a bacterium, the protein was likely to be more compatible with an *E. coli* system, producing a more stable protein. Expressing bacterial proteins within a bacterial system will limit the effects of codon bias. In addition, the expression machinery will less diverged and therefore will produce proteins with correct folding and post-translational modification. All proteins, bar HPL, were successfully cloned and purified, showing the improvements gained from changing to the *P. methylaliphatogenes* system.

5.1.1 Cas2 and HPS Activity and Function Remain Largely Undiscovered

Cas2 is not predicted to have any individual activity due to the lack of two conserved active site residues (Section 4.3.2.2) which provide the RNase activity for Cas2 in *S. solfataricus.* Cas2 has not been shown previously to bind DNA alone while not in complex to Cas1, and these results (Section 4.9) show that Cas2 cannot bind any of the substrates alone. Complexing was attempted between Cas2 and Cas4-1, but unfortunately no complexes were observed within an EMSA or AGF. This may be because certain conditions are required for complexing which have yet to be discovered, or in the case of AGF the concentration of Cas2 was not sufficient for detection. Despite the absence of Cas4-1-Cas2 complex, Cas2 was shown to promote Cas4-1 nuclease activity against M13 ssDNA. This suggests a link between Cas4-1 and Cas2, though not necessarily as a complex. The hypothesis for this promotion of Cas4-1 activity by Cas2 will be discussed later.

HPS unlike Cas2, has no significant similarities to other proteins. This makes any predictions about the properties and function of HPS difficult. The proteins with the most similarities in sequence (see section 4.3.2.2) are DNA binding proteins.

However, when tested against the same substrates as Cas4-1 and Cas2, no binding occurred (section 4.9). HPS may therefore not bind DNA, or only bind a specific DNA structure or sequence not tested here. It is also possible that HPS only binds DNA under certain conditions or in certain complexes. HPS was not tested on AGF for complex formation as the protein was purified later in the work than the other proteins, and after the conclusion of the AGF work. HPS was tested for any effects on Cas4-1 nuclease activity as Cas2 was, but the assays were inconclusive. The function of HPS (if a function exists) was not discovered by this work.

5.1.2 Cas4-1 likely exists as a dimer

As discussed in section 4.6 oligomeric state prediction and AGF showed Cas4-1 as a dimer, whereas BN-PAGE showed Cas4-1 as a tetramer and octamer. However, the addition of SDS to the Cas4-1 BN-PAGE sample gave a band consistent with a dimer. Therefore, it appeared likely that the most stable form of Cas4-1 is a dimer but that this dimer could form the building blocks to create higher oligomeric states. Recent research has examined the formation of a Cas1-Cas4 complex from *Bacillus halodurans* (Lee *et al.*, 2018). The complex which was stably observed in AGF, was characterised using single-particle electron microscopy (EM) (Figure 74A), giving a complex of $Cas4_2$-$Cas1_4$. Crystal structures from *P. calidifontis* and *E. coli* were docked into the model to show approximate fold structure within the complex (Figure 74B). This complex is similar to the predicted Cas4-1 model (Figure 74C), with Cas4 at the top of the complex and Cas1 at the bottom. The main difference is the absence of a Cas1 dimer, but is it possible that the tetramer observed in the BN-PAGE is the dimerization of the Cas1 regions.

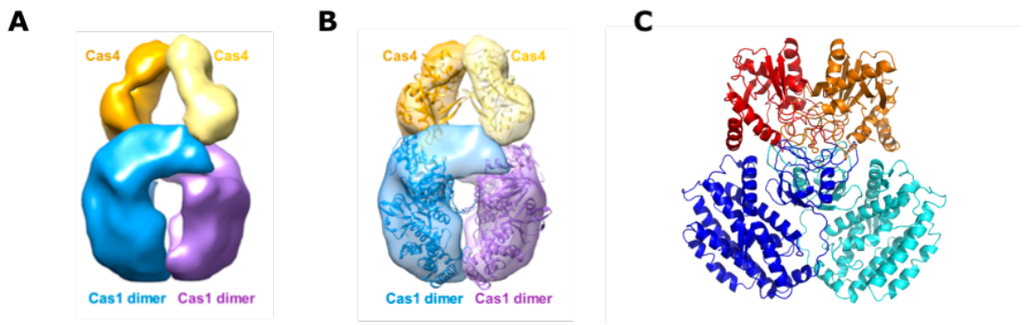

Figure 74: Comparison of *B. halodurans* Cas4-Cas1 complex with predicted Cas4-1 oligomeric state. A) Reconstruction of Cas4-Cas1 complex from *B. halodurans* as obtained from single-particle EM. A Cas4 dimer (yellows) are complexed with two Cas1 dimers (Blue and Purple) (Taken from Lee *et al.* 2018) **B)** Crystal structures of *P. calidifontis* Cas4 and *E. coli* Cas1 dimers docked into the reconstruction to show approximate fold structure of the complex (Taken from (Lee *et al.*, 2018) **C)** Predicted oligomeric state of Cas4-1 as generated by Galaxy Gemini, showing a dimer of Cas4-1 with Cas4 regions in red and orange and Cas1 regions in blues.

5.1.3 Cas4-1 binding of linear DNA with ssDNA ends: implications for protospacer processing

The data from section 4.7 and 4.12 show that Cas4-1 preferentially binds to DNA substrates with ssDNA present. Mutation of one of the conserved cysteines produces a protein lacking a brown solution, indicating the loss of the 4Fe:4S cluster. This mutant protein fails to bind linear DNA substrates which contain ssDNA (ssDNA, 3' overhang and 5' overhang). As 4Fe:4S clusters are involved DNA binding in other proteins (Yeeles *et al*, 2009), it is likely that this 4Fe:4S cluster is involved in DNA binding for the Cas4 region. This set of substrates, resembles protospacer and spacer substrates (bar ssDNA). Protospacers are often trimmed to create overhangs, or already contain overhangs that require trimming. Though at the start of this research we were hypothesising that the Cas4 region was involved in the original creation of the protospacer through degradation of DNA, recent research has been published showing that Cas4 is involved in trimming of protospacer to create processed spacers.

Cas4 is required for protospacer processing and correct integration (Kieper *et al.*, 2018; Lee *et al.*, 2018; Rollie *et al.*, 2018; Shiimori *et al.*, 2018). Recent research in *B. halodurans, Synechocystis, Pyrococcus furiosus* and *S. solfataricus* has revealed more information about the role of Cas4 in adaptation. In the absence of Cas4, integration can still occur, but the intergrated spacers are larger than the usual expected size for WT spacers (Kieper *et al.*, 2018; Lee *et al.*, 2018; Shiimori *et al.*, 2018). This is due to the requirement of Cas4 for processing of protospacers. Cas4 has been shown to trim the 3' overhangs of protospacers to generate the correctly sized spacer for integration (Kieper *et al.*, 2018; Lee *et al.*, 2018; Rollie *et al.*, 2018; Shiimori *et al.*, 2018). Two different methods have been suggested for this processing. The first requires Cas4 to complex with Cas1 and Cas2. Cas1-Cas2 bind the unprocessed protospacer and Cas4 complexes with Cas1-Cas2 to trims the 3' overhangs that are not bound and protected by Cas1-Cas2 (Lee *et al.*, 2018; Rollie *et al.*, 2018). The second involves two Cas4 molecules binding to either end of the pre-processed protospacer and to trim the 3' overhangs. However, this method could be unique to *P. furiosus* as the gene neighbourhood around the CRISPR locus encodes two separate Cas4 sequences which interact with each end of the protospacer (Shiimori *et al.*, 2018). While none of this research implicates

Cas4 in the original generation of the protospacer, Cas4 is required for SAM recognition. Without Cas4, no SAM sequence for the protospacers can be detected (Kieper *et al.*, 2018; Lee *et al.*, 2018; Shiimori *et al.*, 2018). This suggests that Cas4 is involved in generation of the protospacers and cleaves them at the SAM. Or that protospacers are generated by another mechanism and Cas4 is involved in the selection of DNA fragments containing a SAM.

This research project has been attempted to create a singe *in vitro* adaptation reaction based on the nuclease activity of Cas4 generating fragments for capture by Cas1-Cas2. This latest research however would seem to suggest a role in protospacer processing as opposed to protospacer generation. Though the requirement for a SAM only in the presence of Cas4 suggest a role in generation of protospacers. The nuclease activity examined in this research show that Cas4-1 does not appear generate DNA fragments large enough to be utilised as protospacers, but it is possible that these small fragments cannot be detected on an agarose gel. Considering this recent research, the role of Cas4-1 in protospacer processing should be explored.

A hypothesised mechanism extrapolating from recent publications and the results gather here would suggest that Cas4-1 (potentially in complex with Cas2) bind to an unprocessed protospacer with 3’ overhangs. Cas2 then stimulates the nuclease activity of the Cas4-1 to trim the 3’ overhangs creating a processed spacer. Cas4-1 could then integrate this spacer. This would explain the preference of binding of 3’ overhangs by Cas4 (via the 4Fe:4S cluster) and the relatively low activity of Cas4-1. If the only cleavage that Cas4-1 carries out is on protospacers, it would not be required to degrade large substrates as used in these experiments. The activity seen here may be sufficient when stimulated by Cas2 to cleave unprocessed protospacers.

To explore this mechanism, the integration of spacers both *in vivo* and *in vitro* needs to be examined in more detail. To test whether Cas4-1 can cleave unprocessed protospacer, the sequence that would be encountered by Cas4-1 as protospacer needs to be determined i.e. SAM sites and processing sites. Existing spacers in the *P. methylaliphatogenes* CRISPR locus were examined using BLAST to see if the spacers matched any existing sequences. Unfortunately no

homologous sequences were found, but this is likely due to the small amount of available viral sequences. To investigate the origin of spacers to determine a SAM sequence and *in vivo* assay needs to be developed. *In vivo* assays (spacer acquisition assays) have been used in *E. coli* (Ivancic-Bace *et al.*, 2015) to genetically explore spacer acquisition. An attempt was made to create a plasmid to express proteins in *E. coli* to test spacer acquisition and integration into a plasmid containing *P. methylaliphatogenes* CRISPR locus. However, the cloning of Cas4-1 and Cas2 into a single plasmid was unsuccessful.

If an *in vivo* assay could be generated then after spacer integration into the CRISPR locus, the locus could be sequenced and the spacers mapped back to their source. This mapping would provide information on the SAM. Having established a SAM sequence, two further experiments could be conducted. The first would be to supply an *in vitro* reaction with unprocessed spacers or fully processed spacers to test the ability of Cas4-1 to trim these spacers and integrate them. The second would be to provide Cas4-1 with a substrate for degradation which contains several SAM sequences to see if cleavage occurs at a SAM sequence, forming a protospacer. These experiments would provide enough information to allow a better understanding of the system and design an improved experiment to create a single *in vitro* adaptation reaction.

5.2 Future Research

The primary aim of this project was to produce a single *in vitro* naïve adaptation assay with a secondary aim of biochemically investigating Cas4-1 and the surrounding gene neighbourhood. Initial progress in these aims was slow due to issues arising from purification of the gene neighbourhood from *M. harundinacea.* But progress was made after selection of a new gene neighbourhood from *P. methylaliphatogenes.*

Unfortunately, the primary aim of this project was not reached, but this research has set up future research. Reliable purification protocols have been established for 3 out of 4 proteins, allowing research into the molecular mechanism of these proteins and the establishment of a single *in vitro* naïve adaptation reaction.

Experimental results have revealed DNA binding preferences of Cas4-1, along with nuclease activities and interaction between Cas4-1 and Cas2. This in combination with recent publications about Cas4 have brought about several research questions that may now be investigated, discussed below.

5.2.1 What effect does oxidation have on Cas4-1 activity?

As discussed throughout this chapter Cas4-1 is prone to degradation which causes an observerable change in activity. This is hypothesised to be due to oxidation of the 4Fe:4S cluster. 4Fe:4S clusters stabilise proteins and oxidation or disruption of them leads to conformational changes and instability. It was shown that an 4Fe:4S cluster mutant of Cas4-1 cannot be produced at the same concentration of WT or other mutant proteins, showing potential stability issues. To conclusively understand the effects of oxidation on Cas4-1 activity two approaches can be taken. The first, fully oxidise the 4Fe:4S cluster in purified Cas4-1 to assess the effects on activity and stability. The second, produce Cas4-1 in an anaerobic system and carry out experiments in an oxygen free/depleted environment to test whether this improves activity and stability. This would determine whether degradation issues encountered within this are due to 4Fe:4S cluster oxidation.

5.2.2 How do Cas1 active site mutants effect Cas4-1 activity?

Mutations in the 4Fe:4S cluster and Cas4 active site have been examined within this work. Though mutations were made in Cas1 active site residues, none were purified successfully as all purification resulted in aggregation during the desalting step. Attempts were made to change the final step to improve purification, but these were unsuccessful. With some alterations Cas4-1 Cas1 mutants could be purified. It is hypothesised that Cas1 mutations would have no effect on binding or nuclease activity. This is because the Cas1 DNA binding site is separate from the active site in other Cas1 proteins and as Cas1 is an integrase it would not be predicted to be involved in the nuclease activity of the Cas4 region. However, it would be interesting to see if the Cas1 mutant did effect the nuclease activity, and was potentially nicking dsDNA to give access for Cas4. The Cas1 mutant however, would be important upon the generation of the single *in vitro* naïve adaptation reaction to confirm integration is carried out by the Cas1 active site.

5.2.3 Development of the single *in vitro* naïve adaptation assay.

As has been previously discussed the latest research into the role of Cas4 in the CRISPR-Cas mechanism shows the involvement of Cas4 in protospacer processing as opposed to protospacer generation. Therefore, cleavage by Cas4-1 may not generate the protospacers for integration in the *in vitro* naïve adaptation assay. An *in vivo* assay conducted in *E. coli* if successful would allow the sequencing of integrated spacers and mapping back to the source. This would provide information about the SAM sequence (if one is required) and spacer selection.

Leading on from *in vivo* experiments both unprocessed and processed spacer could be provided to Cas4-1 for integration into the CRISPR locus in an *in vitro* assay. These spacers would be designed from the information gained from the *in vivo* assays. Cas4 has recently been shown to process spacers before integration by trimming 3’ overhangs. Therefore, by using unprocessed spacer within the *in vitro* assay this would test the ability of Cas4-1 to process spacers before integration. Providing a pre-established spacer sequence would improve detection, as primers could be designed to detect the spacer sequence. In this case both half-site and full integrations would be detected.

If *in vivo* assays and *in vitro* assays with unprocessed spacers proves successful, then the ability of Cas4-1 to generate protospacers can be tested. Protospacers are generated by cleavage at SAM sites, therefore DNA substrate can be used which contain the *P. methylaliphatogenes* SAM sequence. It may be that DNA already tested in the nuclease assays contains a SAM sequence. If this is the case, a better detection method may be needed to see generated protospacers.

Acknowledgments

I would firstly like to thank my supervisor Dr Edward Bolt for the opportunity to work on this project and for helping me find a path when the research proved particularly difficult. I'm grateful for all I have learned during this project and the experience has grown me as a person.

I would like to thank all members of the D53-D55 lab past and present in particular: Tom for your technical expertise and for the many afternoon tea and coffee breaks to discuss our pet peeves, Ryan for being utterly ridiculous but never failing to bring light and laughter to the lab and to Tabi who has taken on many of my responsibilities in her stride which was really helpful towards the end.

To my partner, Ed, thank you for supporting me through this whole process and never giving up on me. Always there to provide a cup of a tea and hug when the process became almost unbearable.

To my parents, who always believed that I could achieve this and any dream I had. Thank you for everything. You helped me both mentally and financially, I would not be where I was today without you.

To the cool guys, particularly to Linzy, Aaron, Meg, Ben and Jon who have been on this journey with me. We've all faced our demons along the way, but various game nights and terrible films have really helped. I wish you guys the best of luck with your own submissions and vivas.

And finally, to the dodgeballers, both at Balls of steel and Nottingham sheriffs. Dodgeball has really been my release during this process and even if most of you don't realise you helped you all did. My biggest thanks to Kathryn and Alex who have been there the whole time and were always to help.

Bibliography

Aeling, K. A., Opel, M. L., Steffen, N. R., Tretyachenko-Ladokhina, V., Hatfield, G. W., Lathrop, R. H. and Senear, D. F. (2006) 'Indirect recognition in sequence-specific DNA binding by Escherichia coli integration host factor: The role of DNA deformation energy', *Journal of Biological Chemistry*, 281(51), pp. 39236–39248. doi: 10.1074/jbc.M606363200.

Aklujkar, M. and Lovley, D. R. (2010) 'Interference with histidyl-tRNA synthetase by a CRISPR spacer sequence as a factor in the evolution of Pelobacter carbinolicus', *BMC Evolutionary Biology*, 10(1), pp. 16–19. doi: 10.1186/1471-2148-10-230.

Albà, M. M. (2001) 'Protein family review Replicative DNA polymerases', *Genome Biology*, 2(1), pp. 3–6.

Almendros, C., Guzmán, N. M., Díez-Villaseñor, C., García-Martínez, J. and Mojica, F. J. M. (2012) 'Target Motifs Affecting Natural Immunity by a Constitutive CRISPR-Cas System in Escherichia coli', *PLoS ONE*, 7(11). doi: 10.1371/journal.pone.0050797.

Amundsen, S. K., Taylor, a F., Chaudhury, a M. and Smith, G. R. (1986) 'recD: the gene for an essential third subunit of exonuclease V.', *Proceedings of the National Academy of Sciences of the United States of America*, 83(15), pp. 5558–5562. doi: 10.1073/pnas.83.15.5558.

Anders, C. and Jinek, M. (2014) *In vitro enzymology of cas9*. 1st edn, *Methods in Enzymology*. 1st edn. Elsevier Inc. doi: 10.1016/B978-0-12-801185-0.00001-5.

Aravind, L., Makarova, K. S. and Koonin, E. V (2000) 'Holliday junction resolvases and related nucleases: identification of new families, phyletic distribution and evolutionary trajectories', *Nucleic acids research*, 28(18), pp. 3417–3432. doi: 10.1093/nar/28.18.3417.

Arslan, Z., Hermanns, V., Wurm, R., Wagner, R. and Pul, Ü. (2014) 'Detection and characterization of spacer integration intermediates in type I-E CRISPR-Cas system', *Nucleic Acids Research*, 42(12), pp. 7884–7893. doi: 10.1093/nar/gku510.

Azeroglu, B. and Leach, D. R. F. (2017) 'RecG controls DNA amplification at double-strand breaks and arrested replication forks', *FEBS Letters*, 591(8), pp. 1101–1113. doi: 10.1002/1873-3468.12583.

Babu, M., Beloglazova, N., Flick, R., Graham, C., Skarina, T., Nocek, B.,

Gagarinova, A., Pogoutse, O., Brown, G., Binkowski, A., Phanse, S., Joachimiak, A., Koonin, E. V., Savchenko, A., Emili, A., Greenblatt, J., Edwards, A. M. and Yakunin, A. F. (2011) 'A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair', *Molecular Microbiology*, 79(2), pp. 484–502. doi: 10.1111/j.1365-2958.2010.07465.x.

Baldwin, R. L. (1986) 'Temperature dependence of the hydrophobic interaction in protein folding.', *Proceedings of the National Academy of Sciences*, 83(21), pp. 8069–8072. doi: 10.1073/pnas.83.21.8069.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A. and Horvath, P. (2007) 'CRISPR provides acquired resistance against viruses in prokaryotes', *Science*, 315(5819), pp. 1709–1712. doi: 10.1126/science.1138140.

Béguin, P., Charpin, N., Koonin, E. V., Forterre, P. and Krupovic, M. (2016) 'Casposon integration shows strong target site preference and recapitulates protospacer integration by CRISPR-Cas systems', *Nucleic Acids Research*, 44(21), p. gkw821. doi: 10.1093/nar/gkw821.

Beloglazova, N., Brown, G., Zimmerman, M. D., Proudfoot, M., Makarova, K. S., Kudritska, M., Kochinyan, S., Wang, S., Chruszcz, M., Minor, W., Koonin, E. V., Edwards, A. M., Savchenko, A. and Yakunin, A. F. (2008) 'A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats', *Journal of Biological Chemistry*, 283(29), pp. 20361–20371. doi: 10.1074/jbc.M803225200.

Benda, C., Ebert, J., Scheltema, R. A., Schiller, H. B., Baumgärtner, M., Bonneau, F., Mann, M. and Conti, E. (2014) 'Structural model of a CRISPR RNA-silencing complex reveals the RNA-target cleavage activity in Cmr4', *Molecular Cell*, 56(1), pp. 43–54. doi: 10.1016/j.molcel.2014.09.002.

Berg, T. M., Tymoczko, J. L. and Stryer, L. (2002) *Biochemistry*. 5th edn. New York: W. H. Freeman & company.

Bhattacharyya, B., George, N. P., Thurmes, T. M., Zhou, R., Jani, N., Wessel, S. R., Sandler, S. J., Ha, T. and Keck, J. L. (2014) 'Structural mechanisms of PriA-mediated DNA replication restart', *Proceedings of the National Academy of Sciences*, 111(4), pp. 1373–1378. doi: 10.1073/pnas.1318001111.

Bianco, P. R. and Kowalczykowski, S. C. (1997) 'The recombination hotspot Chi is recognized by the translocating RecBCD enzyme as the single strand of DNA containing the sequence 5′-GCTGGTGG-3′', *Proceedings of the National Academy*

of Sciences, 94(13), pp. 6706–6711. doi: 10.1073/pnas.94.13.6706.

Bianco, P. R. and Kowalczykowski, S. C. (2000) 'Translocation Step Size and Mechanism of the RecBC DNA Helicase', *Nature*, 405, pp. 368–372.

Bianco, P. R. and Lyubchenko, Y. L. (2017) 'SSB and the RecG DNA helicase: an intimate association to rescue a stalled replication fork', *Protein Science*, 26(4), pp. 638–649. doi: 10.1002/pro.3114.

Biek, D. P. and Cohen, S. N. (1986) 'Identification and characterization of recD, a gene affecting plasmid maintenance and recombination in Escherichia coli.', *J. Bacteriol.*, 167, pp. 594–603.

Bikard, D., Euler, C. W., Jiang, W., Nussenzweig, P. M., Goldberg, G. W., Duportet, X., Fischetti, V. a and Marraffini, L. a (2014) 'Exploiting CRISPR-Cas nucleases to produce sequence-specific antimicrobials', *Nat Biotechnol*. Nature Publishing Group, 32(11), pp. 1146–1150. doi: 10.1038/nbt.3043.

Bikard, D., Jiang, W., Samai, P., Hochschild, A., Zhang, F. and Marraffini, L. A. (2013) 'Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system', *Nucleic Acids Research*, 41(15), pp. 7429–7437. doi: 10.1093/nar/gkt520.

Blackwood, J. K., Rzechorzek, N. J., Bray, S. M., Maman, J. D., Pellegrini, L. and Robinson, N. P. (2013) 'End-resection at DNA double-strand breaks in the three domains of life: Figure 1', *Biochemical Society Transactions*, 41(1), pp. 314–320. doi: 10.1042/BST20120307.

Bolotin, A., Quinquis, B., Sorokin, A. and Dusko Ehrlich, S. (2005) 'Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin', *Microbiology*, 151(8), pp. 2551–2561. doi: 10.1099/mic.0.28048-0.

Bondy-Denomy, J., Pawluk, A., Maxwell, K. L. and Davidson, A. R. (2013) 'Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system', *Nature*. Nature Publishing Group, 493(7432), pp. 429–432. doi: 10.1038/nature11723.

Brouns, S. J. J., Jore, M. M., Lundgren, M., Westra, E. R., Slijkhuis, R. J. H., Snijders, A. P. L., Dickman, M. J., Makarova, K. S., Koonin, E. V. and van der Oost, J. (2008) 'Small CRISPR RNAs guide antiviral defense in prokaryotes.', *Science (New York, N.Y.)*, 321(5891), pp. 960–4. doi: 10.1126/science.1159689.

Brown, M. W., Dillard, K. E., Xiao, Y., Dolan, A. E., Hernandez, E. T., Dahlhauser, S., Kim, Y., Myler, L. R., Anslyn, E., Ke, A. and Finkelstein, I. (2017) 'Assembly

and translocation of a CRISPR-Cas primed acquisition complex', *bioRxiv*, 41(October), pp. 1–11. doi: https://doi.org/10.1101/208058.

Brutlag, D. and Kornberg, A. (1972) 'Enzymatic synthesis of deoxyribonucleic acid. XXXVI: A proofreading function for the 3'-5' exonuclease activity in deoxyribonucleic acid polymerases.', *The Journal of biological chemistry*, 247(1), pp. 241–248.

Burnette, W. N. (1981) '"Western Blotting": Electrophoretic transfer of proteins from sodium dodecyl sulfate-polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein A', *Analytical Biochemistry*, 112(2), pp. 195–203. doi: 10.1016/0003-2697(81)90281-5.

Buss, J. A., Kimura, Y. and Bianco, P. R. (2008) 'RecG interacts directly with SSB: implications for stalled replication fork regression', *Nucleic Acids Res.*, 36, p. 7029–7042.

Cady, K. C. and O'Toole, G. A. (2011) 'Non-identity-mediated CRISPR-bacteriophage interaction mediated via the Csy and Cas3 proteins', *Journal of Bacteriology*, 193(14), pp. 3433–3445. doi: 10.1128/JB.01411-10.

Cann, I. K. O. and Ishino, Y. (1999) 'Archaeal DNA Replication : Identifying the Pieces to Solve a Puzzle', *Genetics*, 152, pp. 1249–1267.

Cann, I., Komori, K., Toh, H., Kanai, S. and Ishino, Y. (1998) 'A heterodimeric DNA polymerase: evidence that members of Euryarchaeota possess a distinct DNA polymerase.', *Proc Natl Acad Sci USA*, 95(24), pp. 14250–5.

Chen, H., North, S. and Nakai, H. (2004) 'Properties of the PriA helicase domain and its role in binding PriA to specific DNA structures.', *J Biol Chem*, 279(37), pp. 38503–38512.

Cho, S. W., Kim, S., Kim, J. M. and Kim, J. S. (2013) 'Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease', *Nature Biotechnology*. Nature Publishing Group, 31(3), pp. 230–232. doi: 10.1038/nbt.2507.

Chylinski, K., Le Rhun, A. and Charpentier, E. (2013) 'The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems', *RNA Biology*, 10(5), pp. 726–737. doi: 10.4161/rna.24321.

Clark, a. J. and Margulies, a. D. (1965) 'Isolation and characterisation of recombination-deficient mutant of Escherichia coli K12', *Proceedings of the National Academy of Sciences of the United States of America*, 53(1), pp. 451–9. doi: 10.1073/pnas.53.2.451.

Cooper, G. M. (2000) 'Chapter 5: DNA replication', in *The Cell: A molecular*

approach. 2nd edn. Sunderland, USA.: Sinauer Associates.

Cooper, P. K. and Hanawalt, P. C. (1972) 'Role of DNA Polymerase I and the rec System in Excision-Repair in Escherichia coli', *Proc Natl Acad Sci USA*, 69(5), pp. 1156–1160.

Cox, M. M. and Lehman, I. R. (1981a) 'Directionality and polarity in recA protein-promoted branch migration.', *Proceedings of the National Academy of Sciences of the United States of America*, 78(10), pp. 6018–22. doi: 10.1073/pnas.78.10.6018.

Cox, M. M. and Lehman, I. R. (1981b) 'recA protein of Escherichia coli promotes branch migration, a kinetically distinct phase of DNA strand exchange.', *Proceedings of the National Academy of Sciences of the United States of America*, 78(6), pp. 3433–3437. doi: 10.1073/pnas.78.6.3433.

Cox, M. M. and Lehman, I. R. (1982) 'recA protein promoted DNA strand exchange.', *Journal of Biological Chemistry*, 257(14), pp. 8523–8532.

Crowe, M. A., Power, J. F., Morgan, X. C., Dunfield, P. F., Lagutin, K., Rijpstra, W. I. C., Rijpstra, I. C., Sinninghe Damste, J. S., Houghton, K. M., Ryan, J. L. J. and Stott, M. B. (2014) 'Pyrinomonas methylaliphatogenes gen. nov., sp. nov., a novel group 4 thermophilic member of the phylum Acidobacteria from geothermal soils', *International Journal of Systematic and Evolutionary Microbiology*, 64(PART 1), pp. 220–227. doi: 10.1099/ijs.0.055079-0.

Cui, Y., Xu, J., Cheng, M., Liao, X. and Peng, S. (2018) 'Review of CRISPR/Cas9 sgRNA Design Tools', *Interdisciplinary Sciences: Computational Life Sciences*. Springer Berlin Heidelberg, 10(2), pp. 455–465. doi: 10.1007/s12539-018-0298-z.

Datsenko, K. A., Pougach, K., Tikhonov, A., Wanner, B. L., Severinov, K. and Semenova, E. (2012) 'Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system', *Nature Communications*. Nature Publishing Group, 3(May), pp. 945–947. doi: 10.1038/ncomms1937.

Deltcheva, E., Chylinski, K., Sharma, C. M. and Gonzales, K. (2011) 'CRISPR RNA maturation by trans -encoded small RNA and host factor RNase III', *Nature*, 471(7340), pp. 602–607. doi: 10.1038/nature09886.CRISPR.

Deveau, H., Barrangou, R., Garneau, J. E., Labonté, J., Fremaux, C., Boyaval, P., Romero, D. A., Horvath, P. and Moineau, S. (2008) 'Phage response to CRISPR-encoded resistance in Streptococcus thermophilus', *Journal of Bacteriology*, 190(4), pp. 1390–1400. doi: 10.1128/JB.01412-07.

Díez-Villaseñor, C., Guzmán, N. M., Almendros, C., García-Martínez, J. and Mojica, F. J. M. (2013) 'CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of Escherichia coli', *RNA Biology*, 10(5), pp. 792–802. doi: 10.4161/rna.24023.

Dillingham, M. S., Spies, M. and Kowalczykowski, S. C. (2003) 'RecBCD enzyme is a bipolar DNA helicase', *Nature*, 423(6942), pp. 893–897. doi: 10.1038/nature01673.

Dixon, D. and Kowalczykowski, S. (1995) 'Role of the eshcerichia coli recombination hotspot, chi, in RecABCD-dependent homologous pairing.', *The Journal of Biological Chemistry*, 70(27), pp. 16360–16370.

Dohoney, K. M. and Gelles, J. (2001) 'Chi-sequence recognition and DNA translocation by single RecBCD helicase/nuclease molecules.', *Nature*, 409(6818), pp. 370–374. doi: 10.1038/35053124.

Efron, B. (1979) 'Bootstrap Methods: Another Look at the Jackknife', *The Annals of Statistics*, 7(1), pp. 1–26. doi: 10.1214/aos/1176344552.

Emmerson, P. T. (1968) 'Recombination Deficient Mutants of Escherichia Coli K12 That Map Between thyA and argA', *Genetics*, 60, pp. 19–30.

Fagerlund, R. D., Wilkinson, M. E., Klykov, O., Barendregt, A., Pearce, F. G., Kieper, S. N., Maxwell, H. W. R., Capolupo, A., Heck, A. J. R., Krause, K. L., Bostina, M., Scheltema, R. A., Staals, R. H. J. and Fineran, P. C. (2017) 'Spacer capture and integration by a type I-F Cas1-Cas2-3 CRISPR adaptation complex.', *Proceedings of the National Academy of Sciences of the United States of America*, p. 201618421. doi: 10.1073/pnas.1618421114.

Farah, J. a and Smith, G. R. (1997) 'The RecBCD enzyme initiation complex for DNA unwinding: enzyme positioning and DNA opening.', *Journal of molecular biology*, 272(5), pp. 699–715. doi: 10.1006/jmbi.1997.1259.

Faure, G., Makarova, K. S. and Koonin, E. V (2019) 'CRISPR – Cas : Complex Functional Networks and Multiple Roles beyond Adaptive Immunity', *Journal of Molecular Biology*. The Authors, 431, pp. 3–20. doi: 10.1016/j.jmb.2018.08.030.

Felsenstein, J. (1985) 'Confidence limits on phylogenies: an approach using the bootstrap', *Evolution*, 39(4), pp. 783–791. doi: 10.2307/2408678.

Finch, P. W., Storey, A., Brown, K., Hickson, I. D. and Emmerson, P. T. (1986) 'Complete nucleotide sequence of recD, the structural gene for the alpha subunit of exonuclease V of Escherichia coli.', *Nucleic Acids Research*, 14(21), pp. 8583–8594. doi: 10.1128/CMR.00125-13.

Forget, A. L. and Kowalczykowski, S. C. (2012) 'Single-molecule imaging of DNA pairing by RecA reveals a three-dimensional homology search', *Nature*. Nature Publishing Group, 482(7385), pp. 423–427. doi: 10.1038/nature10782.

Friedman, D. I. (1988) 'Integration host factor: a protein for all reasons.', *Cell*, 55(4), pp. 545–554. doi: 0092-8674(88)90213-9 [pii].

Fu, Y., Foden, J. A., Khayter, C., Maeder, M. L., Reyon, D., Joung, J. K. and Sander, J. D. (2013) 'High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells', *Nature Biotechnology*. Nature Publishing Group, 31(9), pp. 822–826. doi: 10.1038/nbt.2623.

Fukuoh, A., Iwasaki, H., Ishioka, K. and Shinagawa, H. (1997) 'ATP-dependent resolution of R-loops at the ColE1 replication origin by Escherichia coli RecG protein, a Holliday junction-specific helicase', *EMBO Journal*, 16(1), pp. 203–209. doi: 10.1093/emboj/16.1.203.

Garcia-Diaz, M. (2007) 'Multiple functions of DNA polymerases', *CRC Crit Rev Plant Sci*, 26(2), pp. 105–122.

Garneau, J. E., Dupuis, M. È., Villion, M., Romero, D. A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadán, A. H. and Moineau, S. (2010) 'The CRISPR/cas bacterial immune system cleaves bacteriophage and plasmid DNA', *Nature*, 468(7320), pp. 67–71. doi: 10.1038/nature09523.

Gasiunas, G., Barrangou, R., Horvath, P. and Siksnys, V. (2012) 'Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria', *Proceedings of the National Academy of Sciences*, 109(39), pp. E2579–E2586. doi: 10.1073/pnas.1208507109.

Gefter, M. L. (1975) 'DNA replication', *Annual Review of Biochemistry*, 44, pp. 45–78.

Gilson, E., Clément, J. M., Brutlag, D. and Hofnung, M. (1984) 'A family of dispersed repetitive extragenic palindromic DNA sequences in E. coli.', *The EMBO journal*, 3(6), pp. 1417–1421.

Gong, B., Shin, M., Sun, J., Jung, C.-H., Bolt, E. L., van der Oost, J. and Kim, J.-S. (2014) 'Molecular insights into DNA interference by CRISPR-associated nuclease-helicase Cas3', *Proceedings of the National Academy of Sciences*, 111(46), pp. 16359–16364. doi: 10.1073/pnas.1410806111.

Goodrich, J. A., Schwartz, M. L. and Mcclure, W. R. (1990) 'Searching for and Predicting the Activity of Sites for DNA-Binding Proteins - Compilation and Analysis of the Binding-Sites for Escherichia-Coli Integration Host Factor (Ihf)', *Nucleic Acids*

Research, 18(17), pp. 4993–5000. doi: DOI 10.1093/nar/18.17.4993.

Gorbalenya, A. E. and Koonin, E. V. (1993) 'Helicases: amino acid sequence comparisons and structure-function relationships.', *Curr. Opin. Struct. Biol.*, 3, pp. 419–429.

Goren, M. G., Yosef, I., Auster, O. and Qimron, U. (2012) 'Experimental definition of a clustered regularly interspaced short palindromic duplicon in escherichia coli', *Journal of Molecular Biology*. Elsevier Ltd, 423(1), pp. 14–16. doi: 10.1016/j.jmb.2012.06.037.

Gregg, A. V., McGlynn, P., Jaktaji, R. P. and Lloyd, R. G. (2002) 'Direct rescue of stalled DNA replication forks via the combined action of PriA and RecG helicase activities', *Molecular Cell*, 9(2), pp. 241–251. doi: 10.1016/S1097-2765(02)00455-0.

Grissa, I., Vergnaud, G. and Pourcel, C. (2007) 'The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats', *BMC Bioinformatics*, 8, pp. 1–10. doi: 10.1186/1471-2105-8-172.

Groenen, P. M., Bunschoten, a E., van Soolingen, D. and van Embden, J. D. (1993) 'Nature of DNA polymorphism in the direct repeat cluster of Mycobacterium tuberculosis; application for strain differentiation by a novel typing method.', *Molecular microbiology*, 10(5), pp. 1057–1065. doi: 10.1111/j.1365-2958.1993.tb00976.x.

Grossman, L., Braun, A., Feldberg, R. and Mahler, I. (1975) 'Enzymatic repair of DNA', *Annual Review of Biochemistry*, 44, pp. 19–43.

Gudbergsdottir, S., Deng, L., Chen, Z., Jensen, J. V. K., Jensen, L. R., She, Q. and Garrett, R. A. (2011) 'Dynamic properties of the Sulfolobus CRISPR/Cas and CRISPR/Cmr systems when challenged with vector-borne viral and plasmid genes and protospacers', *Molecular Microbiology*, 79(1), pp. 35–49. doi: 10.1111/j.1365-2958.2010.07452.x.

Haft, D. H., Selengut, J., Mongodin, E. F. and Nelson, K. E. (2005) 'A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/cas subtypes exist in prokaryotic genomes', *PLoS Computational Biology*, 1(6), pp. 0474–0483. doi: 10.1371 /journal.pcbi.0010060.

Handa, N., Yang, L., Dillingham, M. S., Kobayashi, I., Wigley, D. B. and Kowalczykowski, S. C. (2012) 'Molecular determinants responsible for recognition of the single-stranded DNA regulatory sequence, , by RecBCD enzyme', *Proceedings of the National Academy of Sciences*, 109(23), pp. 8901–8906. doi:

10.1073/pnas.1206076109.

Hastings, P. J., Hersh, M. N., Thornton, P. C., Fonville, N. C., Slack, A., Ryan, L., Ray, M. P., Harris, R. S., Leal, S. M. and Rosenberg, S. M. (2010) 'Competition of Escherichia coli DNA Polymerases I , II and III with DNA Pol IV in Stressed Cells', *PLoS ONE*, 5(5), p. e10862. doi: 10.1371/journal.pone.0010862.

Hayes, R. P., Xiao, Y., Ding, F., Van Erp, P. B. G., Rajashankar, K., Bailey, S., Wiedenheft, B. and Ke, A. (2016) 'Structural basis for promiscuous PAM recognition in type I-E Cascade from E. coli', *Nature*. Nature Publishing Group, 530(7591), pp. 499–503. doi: 10.1038/nature16995.

Heller, R. C. and Marians, K. J. (2006) 'Replisome assembly and the direct restart of stalled replication forks', *Nat. Rev. Mol. Cell Biol.*, 7, pp. 932–943.

Hermans, P. W. M., Van Soolingen, D., Bik, E. M., De Haas, P. E. W., Dale, J. W. and Van Embden, J. D. A. (1991) 'Insertion element IS987 from Mycobacterium bovis BCG is located in a hot-spot integration region for insertion elements in Mycobacterium tuberculosis complex strains', *Infection and Immunity*, 59(8), pp. 2695–2705. doi: PMID:1649798.

Heussler, G. E., Miller, J. L., Price, C. E., Collins, A. J. and O'Toole, G. A. (2016) 'Requirements for Pseudomonas aeruginosa type I-F CRISPR-Cas adaptation determined using a biofilm enrichment assay', *Journal of Bacteriology*, 198(22), pp. 3080–3090. doi: 10.1128/JB.00458-16.

Hickman, A. B. and Dyda, F. (2015) 'The casposon-encoded Cas1 protein from Aciduliprofundum boonei is a DNA integrase that generates target site duplications', *Nucleic Acids Research*, 43(22), pp. 10576–10587. doi: 10.1093/nar/gkv1180.

Hickman, A. B. and Dyda, F. (2016) 'DNA Transposition at Work', *Chemical Reviews*, 116(20), pp. 12758–12784. doi: 10.1021/acs.chemrev.6b00003.

Higgins, C. F., Ames, G. F. L., Barnes, W. M., Clement, J. M. and Hofnung, M. (1982) 'A novel intercistronic regulatory element of prokaryotic operons', *Nature*, 298(5876), pp. 760–762. doi: 10.1038/298760a0.

Hoe, N., Nakashima, K., Grigsby, D., Pan, X., Dou, S. J., Naidich, S., Garcia, M., Kahn, E., Bergmire-Sweat, D. and Musser, J. M. (1999) 'Rapid molecular genetic subtyping of serotype M1 group A Streptococcus strains', *Emerging Infectious Diseases*, 5(2), pp. 254–263. doi: 10.3201/eid0502.990210.

Holmes, A. M. and Haber, J. E. (1999) 'Double-Strand Break Repair in Yeast Requires Both Leading and Lagging Strand DNA Polymerases', *Cell*, 96, pp. 415–424.

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

Howard-Flanders, P. and Theriot, L. (1966) 'Mutants of Escherichia coli K-12 defective in DNA repair and in genetic recombination.', *Genetics*, 53, pp. 1137–1150.

Howard, J. A. L., Delmas, S., Ivančić-Baće, I. and Bolt, E. L. (2011) 'Helicase dissociation and annealing of RNA-DNA hybrids by *Escherichia coli* Cas3 protein', *Biochemical Journal*, 439(1), pp. 85–95. doi: 10.1042/BJ20110901.

Hsieh, P., Camerini-Otero, C. S. and Camerini-Otero, R. D. (1992) 'The synapsis event in the homologous pairing of DNAs: RecA recognizes and pairs less than one helical repeat of DNA.', *Proceedings of the National Academy of Sciences of the United States of America*, 89(14), pp. 6492–6. doi: 10.1073/pnas.89.14.6492.

Hudaiberdiev, S., Shmakov, S., Wolf, Y. I., Terns, M. P., Makarova, K. S. and Koonin, E. V. (2017) 'Phylogenomics of Cas4 family nucleases', *BMC Evolutionary Biology*. BMC Evolutionary Biology, 17(1), pp. 1–14. doi: 10.1186/s12862-017-1081-1.

Hulton, C. S. J., Higgins, C. F. and Sharp, P. M. (1991) 'ERIC sequences: a novel family of repetitive elements in the genomes of Escherichia coli, Salmonella typhimurium and other enterobacteria', *Molecular Microbiology*, 5(4), pp. 825–834. doi: 10.1111/j.1365-2958.1991.tb00755.x.

Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. and Nakata, A. (1987) 'Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product.', *Journal of Bacteriology*, 169(12), pp. 5429–5433. doi: 10.1128/jb.169.12.5429-5433.1987.

Ishioka, K., Iwasaki, H. and Shinagawa, H. (1997) '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, pp. 91–99.

Ito, J. and Braithwaite, D. K. (1991) 'Compilation and alignment of DNA polymerase sequences', *Nucleic Acids Research*, 19(15), pp. 4045–4057.

Ivancic-Bace, I., Cass, S. D., Wearne, S. J. and Bolt, E. L. (2015) 'Different genome stability proteins underpin primed and nai ve adaptation in E. coli CRISPR-Cas immunity', *Nucleic Acids Research*, pp. 1–10. doi: 10.1093/nar/gkv1213.

Ivančić-Bace, I., Cass, S. D., Wearne, S. J. and Bolt, E. L. (2015) 'Different genome stability proteins underpin primed and Naïve adaptation in E. Coli CRISPR-Cas immunity', *Nucleic Acids Research*, 43(22), pp. 10821–10830. doi:

10.1093/nar/gkv1213.

Ivančić-Baće, I., Al Howard, J. and Bolt, E. L. (2012) 'Tuning in to interference: R-loops and cascade complexes in CRISPR immunity', *Journal of Molecular Biology*, 422(5), pp. 607–616. doi: 10.1016/j.jmb.2012.06.024.

Jaktaji, R. P. and Lloyd, R. G. (2003) 'PriA supports two distinct pathways for replication restart in UV-irradiated Escherichia coli cells', *Molecular Microbiology*, 47(4), pp. 1091–1100. doi: 10.1046/j.1365-2958.2003.03357.x.

Jansen, R., Embden, J. D. A. van, Gaastra, W. and Schouls, L. M. (2002) 'Identification of genes that are associated with DNA repeats in prokaryotes', *Molecular Microbiology*, 43(6), pp. 1565–1575. doi: 10.1046/j.1365-2958.2002.02839.x.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A. and Charpentier, E. (2012) 'A Programmable Dual-RNA – Guided DNA Endonuclease in Adaptive Bacterial Immunity', *Science*, 337(August), pp. 816–822.

Johnson, M. K. (1998) 'Iron-sulfur proteins : new roles for old clusters', *Current Opinion in Chemical Biology*, 2, pp. 173–181.

Jore, M. M., Lundgren, M., Van Duijn, E., Bultema, J. B., Westra, E. R., Waghmare, S. P., Wiedenheft, B., Pul, Ü., Wurm, R., Wagner, R., Beijer, M. R., Barendregt, A., Zhou, K., Snijders, A. P. L., Dickman, M. J., Doudna, J. A., Boekema, E. J., Heck, A. J. R., Van Der Oost, J. and Brouns, S. J. J. (2011) 'Structural basis for CRISPR RNA-guided DNA recognition by Cascade', *Nature Structural and Molecular Biology*. Nature Publishing Group, 18(5), pp. 529–536. doi: 10.1038/nsmb.2019.

Jovin, T. M., Englund, P. T. and Bertsch, L. L. (1969) 'Enzymatic synthesis of deoxyribonucleic acid: XXVI: Phyiscal and chemical studies of a homogenous deoxyribonucleic acid polymerase', *The Journal of biological chemistry*, 244(11), pp. 2996–3008.

Jovin, T. M., Englund, P. T. and Kornberg, A. (1969) 'Enzymatic synthesis of deoxyribonucleic acid. XXVII: Chemical modifications of deoxyribonucleic acid polymerase.', *The Journal of biological chemistry*, 244(11), pp. 3009–3018.

Kelly, R. G., Atkinson, M. R., Huberman, J. A. and Kornberg, A. (1969) 'Excision of thymine dimers and other mismatched sequences by DNA polymerase of Escherichia coli', *Nature*, 224, pp. 495–501.

Kieper, S. N., Behler, J., Vink, J. N. A., Hess, W. R., Brouns, S. J. J., Behler, J., Mckenzie, R. E. and Nobrega, F. L. (2018) 'Cas4 Facilitates PAM-Compatible Spacer Selection during CRISPR Adaptation Report Cas4 Facilitates PAM-Compatible

Spacer Selection during CRISPR Adaptation', *Cell Reports*, 22, pp. 3377–3384. doi: 10.1016/j.celrep.2018.02.103.

Killelea, T. and Bolt, E. L. (2017) 'CRISPR-cas Adaptive Immunity and the Three Rs', *Bioscience Reports*, 0(July), p. BSR20170297. doi: 10.1042/BSR20160297.

Killelea, T., Hawkins, M., Howard, J. L., McGlynn, P. and Bolt, E. L. (2018) 'DNA replication roadblocks caused by Cascade interference complexes are alleviated by RecG DNA repair helicase', *RNA Biology*. Taylor & Francis, 00(00), pp. 1–6. doi: 10.1080/15476286.2018.1496773.

Kim, T. Y., Shin, M., Huynh Thi Yen, L. and Kim, J. S. (2013) 'Crystal structure of Cas1 from Archaeoglobus fulgidus and characterization of its nucleolytic activity', *Biochemical and Biophysical Research Communications*. Elsevier Inc., 441(4), pp. 720–725. doi: 10.1016/j.bbrc.2013.10.122.

Klett, B. Y. R. P., Cerami, A. and Reich, E. (1968) 'Exonuclease VI, a new nuclease activity associated with E. coli DNA polymerase', *Biochemistry*, 60(1968), pp. 943–950.

Konrad, E. B. and Lehman, I. R. (1974) 'A Conditional Lethal Mutant of Escherichia coli K12 Defective in the 5'-3' exonuclease associated with DNA polymerase I.', *Proc Natl Acad Sci USA*, 71(5), pp. 2048–2051.

Koonin, E., Makarova, K. and Zhang, F. (2017) 'Diversity, classification and evolution of CRISPR-Cas systems.', *Curr Opin Microbiol.*, 7, p. 67–78.

Koonin, E. V. and Krupovic, M. (2015) 'Evolution of adaptive immunity from transposable elements combined with innate immune systems', *Nature Reviews Genetics*. Nature Publishing Group, 16(3), pp. 184–192. doi: 10.1038/nrg3859.

Korangy, F. and Julin, D. A. (1994) 'Efficiency of ATP hydrolysis and DNA unwinding by the RecBC enzyme from Escherichia coli.', *Biochemistry*, 33, pp. 9552–9560.

Kornberg, A. (1969) 'Active Center of DNA Polymerase', *Science*, 163(3874), pp. 1410–1418.

Kowalczykowski, S. C. (1991) 'Biochemistry of genetic recombination: energetics and mechanism of DNA strand exchange.', *Annual review of biophysics and biophysical chemistry*, 20, pp. 539–575. doi: 10.1146/annurev.bb.20.060191.002543.

Kowalczykowski, S. C., Dixon, D. A., Eggleston, A. K., Lauder, S. D. and Rehrauer, W. M. (1994) 'Biochemistry of homologous recombination in Escherichia coli.', *Microbiological reviews*, 58(3), pp. 401–65. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7968921%5Cnhttp://www.pubmedcentral.

nih.gov/articlerender.fcgi?artid=PMC372975.

Krajewski, W. W., Fu, X., Wilkinson, M., Cronin, N. B., Dillingham, M. S. and Wigley, D. B. (2014) 'Structural basis for translocation by AddAB helicase-nuclease and its arrest at χ sites', *Nature*. Nature Publishing Group, 508(7496), pp. 416–419. doi: 10.1038/nature13037.

Krupovic, M., Béguin, P. and Koonin, E. V. (2017) 'Casposons: mobile genetic elements that gave rise to the CRISPR-Cas adaptation machinery', *Current Opinion in Microbiology*, 38, pp. 36–43. doi: 10.1016/j.mib.2017.04.004.

Krupovic, M., Makarova, K. S., Forterre, P., Prangishvili, D. and Koonin, E. V (2014) 'Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity.', *BMC biology*, 12(1), p. 36. doi: 10.1186/1741-7007-12-36.

Krupovic, M., Shmakov, S., Makarova, K. S., Forterre, P. and Koonin, E. V. (2016) 'Recent mobility of casposons, self-synthesizing transposons at the origin of the CRISPR-Cas immunity', *Genome Biology and Evolution*, 33(0), p. evw006. doi: 10.1093/gbe/evw006.

Lecointe, F., Serena, C., Velten, M., Costes, A., McGovern, S., Meile, J. C., Errington, J., Ehrlich, S. D., Noirot, P. and Polard, P. (2007) 'Anticipating chromosomal replication fork arrest: SSB targets repair DNA helicases to active forks', *EMBO J.*, 26, p. 4239–4251.

Lee, H., Park, H., Ko, J. and Seok, C. (2013) 'GalaxyGemini: A web server for protein homo-oligomer structure prediction based on similarity', *Bioinformatics*, 29(8), pp. 1078–1080. doi: 10.1093/bioinformatics/btt079.

Lee, H., Zhou, Y., Taylor, D. W. and Sashital, D. G. (2018) 'Cas4-Dependent Prespacer Processing Ensures High-Fidelity Programming of CRISPR Arrays', *Molecular Cell*. Elsevier Inc., 70(1), p. 48–59.e5. doi: 10.1016/j.molcel.2018.03.003.

Lee, M. S. and Marians, K. J. (1989) 'The Escherichia coli primosome can translocate actively in either direction along as DNA strand', *Journal of Biological Chemistry*, 264(24), pp. 14531–14542.

Lehman, I. R., Bessman, M. J., Simms, E. S. and Kornberg, A. (1958) 'Enzymatic synthesis of Deoxyribonucleic Acid. I. Preparation of substrates and partial purification of an enzyme from Escherichia coli', *Journal of Biological Chemistry*, 233(1), pp. 163–170.

Lehman, I. R. and Richardson, C. C. (1964) 'The deoxyribonucleases of escherichia

coli. IV: An exonuclease activity present in purified preparations of deoxyribonucleic acid polymerase.', *The Journal of biological chemistry*, 239(1), pp. 233–241.

Lemak, S., Beloglazova, N., Nocek, B., Skarina, T., Flick, R., Brown, G., Popovic, A., Joachimiak, A., Savchenko, A. and Yakunin, A. F. (2013) 'Toroidal structure and DNA cleavage by the CRISPR-associated [4Fe-4S] cluster containing Cas4 nuclease SSO0001 from sulfolobus solfataricus', *Journal of the American Chemical Society*, 135(46), pp. 17476–17487. doi: 10.1021/ja408729b.

Lemak, S., Nocek, B., Beloglazova, N., Skarina, T., Flick, R., Brown, G., Joachimiak, A., Savchenko, A. and Yakunin, A. F. (2014) 'The CRISPR-associated Cas4 protein Pcal_0546 from Pyrobaculum calidifontis contains a [2Fe-2S] cluster: crystal structure and nuclease activity', *Nucleic Acids Res*, 42(17), pp. 11144–11155. doi: 10.1093/nar/gku797.

Lemak, S., Nocek, B., Beloglazova, N., Skarina, T., Flick, R., Brown, G., Joachimiak, A., Savchenko, A. and Yakunin, A. F. (2014) 'The CRISPR-associated Cas4 protein Pcal 0546 from Pyrobaculum calidifontis contains a [2Fe-2S] cluster: Crystal structure and nuclease activity', *Nucleic Acids Research*, 42(17), pp. 11144–11155. doi: 10.1093/nar/gku797.

Levy, A., Goren, M. G., Yosef, I., Auster, O., Manor, M., Amitai, G., Edgar, R., Qimron, U. and Sorek, R. (2015) 'CRISPR adaptation biases explain preference for acquisition of foreign DNA', *Nature*, 520(7548), pp. 505–510. doi: 10.1038/nature14302.

Li, M., Wang, R., Zhao, D. and Xiang, H. (2014) 'Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process', *Nucleic Acids Research*, 42(4), pp. 2483–2492. doi: 10.1093/nar/gkt1154.

Lillestøl, R. K., Redder, P., Garrett, R. A. and Brügger, K. (2006) 'A putative viral defence mechanism in archaeal cells', *Archaea*, 2(1), pp. 59–72. doi: 10.1155/2006/542818.

Lin, Y., Cradick, T. J., Brown, M. T., Deshmukh, H., Ranjan, P., Sarode, N., Wile, B. M., Vertino, P. M., Stewart, F. J. and Bao, G. (2014) 'CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences', *Nucleic Acids Research*, 42(11), pp. 7473–7485. doi: 10.1093/nar/gku402.

Liu, J. and Marians, K. J. (1999) 'PriA-directed assembly of a primosome on D loop DNA', *Journal of Biological Chemistry*, 274(35), pp. 25033–25041. doi:

10.1074/jbc.274.35.25033.

Liu, J., Xu, L., Sandler, S. J. and Marians, K. J. (1999) 'Replication fork assembly at recombination intermediates is required for bacterial growth.', *Proceedings of the National Academy of Sciences of the United States of America*, 96(7), pp. 3552–5. doi: 10.1073/pnas.96.7.3552.

Liu, T., Liu, Z., Ye, Q., Pan, S., Wang, X., Li, Y., Peng, W., Liang, Y., She, Q. and Peng, N. (2017) 'Coupling transcriptional activation of CRISPR-Cas system and DNA repair genes by Csa3a in Sulfolobus islandicus', *Nucleic Acids Research*. Oxford University Press, 45(15), pp. 8978–8992. doi: 10.1093/nar/gkx612.

Lloyd, R. G. (1991) 'Conjugational recombination in resolvase-deficient ruvC mutants of Escherichia coli K-12 depends on recG', *Journal of Bacteriology*, 173(17), pp. 5414–5418. doi: 10.1128/jb.173.17.5414-5418.1991.

Lloyd, R. G. and Buckman, C. (1991) 'Genetic analysis of the recG locus of Escherichia coli K-12 and of its role in recombination and DNA repair.', *Journal of bacteriology*, 173(3), pp. 1004–1011.

Lopper, M., Boonsombat, R., Sandler, S. and Keck, J. (2007) 'A hand-off mechanism for pri- mosome assembly in replication restart.', *Mol Cell*, 26(6), pp. 781–793.

Lovett, S. T. (2006) 'Replication arrest-stimulated recombination: Dependence on the RecA paralog, RadA/Sms and translesion polymerase, DinB', *DNA Repair (Amst.)*, 5, pp. 1421–1427.

Lovett, S. T., Drapkin, P. T., Sutera Jr, V. A. and Gluckman-Peskind, T. J. (1993) 'A sister- strand exchange mechanism for recA-independent deletion of repeated DNA sequences in Escherichia coli.', *Genetics*, 135, pp. 631–642.

Lupski, J. R. and Weinstock, G. M. (1992) 'Short , Interspersed Repetitive DNA Sequences in Prokaryotic Genomes', *Journal of Bacteriology*, 174(14), pp. 4525–4529.

Ma, K., Liu, X. and Dong, X. (2006) 'Methanosaeta harundinacea sp. nov., a novel acetate-scavenging methanogen isolated from a UASB reactor', *International Journal of Systematic and Evolutionary Microbiology*, 56(1), pp. 127–131. doi: 10.1099/ijs.0.63887-0.

Mahdi, A. A., McGlynn, P., Levett, S. D. and Lloyd, R. G. (1997) 'DNA binding and helicase domain of the Escherichia coli recombination protein RecG.', *Nucleic Acids Res*, 25, pp. 3875–3880.

Mahdi, A. A., Sharples, G. J., Mandal, T. N. and Lloyd, R. G. (1996) 'Holliday

junction resolvases encoded by homologous rusA genes in Escherichia coli K-12 and phage', *J. Mol. Biol.*, 82(257), pp. 561–573.

Makarova, K. S., Grishin, N. V, Shabalina, S. a, Wolf, Y. I. and Koonin, E. V (2006) 'A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action.', *Biology direct*, 1, p. 7. doi: 10.1186/1745-6150-1-7.

Makarova, K. S., Haft, D. H., Barrangou, R., Brouns, S. J. J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F. J. M., Wolf, Y. I., Yakunin, A. F., van der Oost, J. and Koonin, E. V (2011) 'Evolution and classification of the CRISPR-Cas systems.', *Nature reviews. Microbiology*. Nature Publishing Group, 9(6), pp. 467–477. doi: 10.1038/nrmicro2577.

Makarova, K. S., Wolf, Y. I., Alkhnbashi, O. S., Costa, F., Shah, S. A., Saunders, S. J., Barrangou, R., Brouns, S. J. J., Charpentier, E., Haft, D. H., Horvath, P., Moineau, S., Mojica, F. J. M., Terns, R. M., Terns, M. P., White, M. F., Yakunin, A. F., Garrett, R. A., van der Oost, J., Backofen, R. and Koonin, E. V. (2015) 'An updated evolutionary classification of CRISPR–Cas systems', *Nature Reviews Microbiology*, (September), pp. 1–15. doi: 10.1038/nrmicro3569.

Makarova, K. S., Wolf, Y. I., Alkhnbashi, O. S., Costa, F., Shah, S. A., Saunders, S. J., Barrangou, R., Brouns, S. J. J., Charpentier, E., Haft, D. H., Horvath, P., Moineau, S., Mojica, F. J. M., Terns, R. M., Terns, M. P., White, M. F., Yakunin, A. F., Garrett, R. A., Van Der Oost, J., Backofen, R. and Koonin, E. V. (2015) 'An updated evolutionary classification of CRISPR-Cas systems', *Nature Reviews Microbiology*, 13(11), pp. 722–736. doi: 10.1038/nrmicro3569.

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., Norville, J. E. and Church, G. M. (2013) 'RNA-Guided Human Genome Engineering via Cas9 Prashant', *Science*, 339(6121), pp. 823–826. doi: 10.1126/science.1232033.RNA-Guided.

Marraffini, L. A. and Sontheimer, E. J. (2010) 'Self versus non-self discrimination during CRISPR RNA-directed immunity', *Nature*. Nature Publishing Group, 463(7280), pp. 568–571. doi: 10.1038/nature08703.

Masepohl, B., Gorlitz, K. G. and Bohme, H. (1996) 'Long tandemly repeated repetitive (LTRR) sequences in the filamentous', *Biochimiea et Biophysica Acta*, 1307, pp. 26–30.

McCool, J. D. and Sandler, S. J. (2001) 'Effects of mutations involving cell division,

recombination, and chromosome dimer resolution on a priA2::kan mutant.', *Proceedings of the National Academy of Sciences of the United States of America*, 98(15), pp. 8203–8210. doi: 10.1073/pnas.121007698.

McEntee, K., Weinstock, G. M. and Lehman, I. R. (1979) 'Initiation of general recombination catalyzed in vitro by the recA protein of Escherichia coli.', *Proceedings of the National Academy of Sciences of the United States of America*, 76(6), pp. 2615–9. doi: 10.1073/pnas.76.6.2615.

McGlynn, P., Al-Deib, a a, Liu, J., Marians, K. J. and Lloyd, R. G. (1997) 'The DNA replication protein PriA and the recombination protein RecG bind D-loops.', *Journal of molecular biology*, 270(2), pp. 212–221. doi: 10.1006/jmbi.1997.1120.

McGlynn, P. and Lloyd, R. G. (2000) 'Modulation of RNA polymerase by (p)ppGpp reveals a RecG-dependent mechanism for replication fork progression.', *Cell*, 101(1), pp. 35–45. doi: 10.1016/S0092-8674(00)80621-2.

McGlynn, P. and Lloyd, R. G. (2001) '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.', *Proceedings of the National Academy of Sciences of the United States of America*, 98(15), pp. 8227–8234. doi: 10.1073/pnas.111008698.

McGlynn, P., Lloyd, R. G. and Marians, K. J. (2000) 'Characterisation of the catalytically active form of RecG helicase.', *Nucleic Acids Res*, 28, pp. 2324–2332.

McGlynn, P., Lloyd, R. G. and Marians, K. J. (2001) 'Formation of Holliday junctions by regression of nascent DNA in intermediates containing stalled replication forks: RecG stimulates regression even when the DNA is negatively supercoiled.', *Proc. Natl. Acad. Sci. U.S.A.*, 98, pp. 8235–8240.

Mojica, F. J., Juez, G. and Rodríguez-Valera, F. (1993) 'Transcription at different salinities of Haloferax mediterranei sequences adjacent to partially modified PstI sites.', *Molecular microbiology*, 9(3), pp. 613–621. doi: 10.1111/j.1365-2958.1993.tb01721.x.

Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. and Almendros, C. (2009) 'Short motif sequences determine the targets of the prokaryotic CRISPR defence system', *Microbiology*, 155(3), pp. 733–740. doi: 10.1099/mic.0.023960-0.

Mojica, F. J. M., Ferrer, C., Juez, G. and Rodríguez-Valera, F. (1995) 'Long stretches of short tandem repeats are present in the largest replicons of the Archaea Haloferax mediterranei and Haloferax volcanii and could be involved in replicon partitioning', *Molecular Microbiology*, 17(1), pp. 85–93. doi: 10.1111/j.1365-

2958.1995.mmi_17010085.x.

Nuñez, J. K., Bai, L., Harrington, L. B., Hinder, T. L. and Doudna, J. A. (2016) 'CRISPR Immunological Memory Requires a Host Factor for Specificity', *Molecular Cell*, 62(6), pp. 824–833. doi: 10.1016/j.molcel.2016.04.027.

Nuñez, J. K., Kranzusch, P. J., Noeske, J., Wright, A. V., Davies, C. W. and Doudna, J. A. (2014) 'Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity', *Nature Structural and Molecular Biology*, 21(6), pp. 528–534. doi: 10.1038/nsmb.2820.

Nurse, P., Liu, J. and Marians, K. J. (1999) 'Two modes of PriA binding to DNA', *Journal of Biological Chemistry*, 274(35), pp. 25026–25032. doi: 10.1074/jbc.274.35.25026.

Okazaki, R. (1968) 'In vivo mechanism of DNA chain growth', *Cold Spring Harb Symp Quart Biol*, 33, pp. 129–43.

van der Oost, J., Jore, M. M., Westra, E. R., Lundgren, M. and Brouns, S. J. J. (2009) 'CRISPR-based adaptive and heritable immunity in prokaryotes', *Trends in Biochemical Sciences*, 34(8), pp. 401–407. doi: 10.1016/j.tibs.2009.05.002.

Painter, R. B. and Scaefer, A. (1969) 'State of newly synthesised HeLa DNA', *Nature*, 221, pp. 1215–1217.

Plagens, A., Tjaden, B., Hagemann, A., Randau, L. and Hensel, R. (2012) 'Characterization of the CRISPR/Cas subtype I-A system of the hyperthermophilic crenarchaeon Thermoproteus tenax', *Journal of Bacteriology*, 194(10), pp. 2491–2500. doi: 10.1128/JB.00206-12.

Pul, Ü., Wurm, R., Arslan, Z., Geißen, R., Hofmann, N. and Wagner, R. (2010) 'Identification and characterization of E. coli CRISPR-cas promoters and their silencing by H-NS', *Molecular Microbiology*, 75(6), pp. 1495–1512. doi: 10.1111/j.1365-2958.2010.07073.x.

Resnick, M. A. (1976) 'The repair of double-strand breaks in DNA: A model involving recombination', *Journal of Theoretical Biology*, 59(1), pp. 97–106. doi: 10.1016/S0022-5193(76)80025-2.

Rice, P. A., Yang, S. W., Mizuuchi, K. and Nash, H. A. (1996) 'Crystal structure of an IHF-DNA complex: A protein-induced DNA U-turn', *Cell*. Cell Press, 87(7), pp. 1295–1306. doi: 10.1016/S0092-8674(00)81824-3.

Richardson, C. C., Schildkraut, C. L., Aposhian, H. V and Kornberg, A. (1964) 'Enzymatic synthesis of deoxyribonucleic acid. XIV: Further purification and properties of deoxyribonucleic acid polymerase of Escherichia coli.', *The Journal of*

biological chemistry, 239(1), pp. 222–233.

Richter, C., Dy, R. L., McKenzie, R. E., Watson, B. N. J., Taylor, C., Chang, J. T., McNeil, M. B., Staals, R. H. J. and Fineran, P. C. (2014) 'Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer', *Nucleic Acids Research*, 42(13), pp. 8516–8526. doi: 10.1093/nar/gku527.

Rocha, E. P. C., Cornet, E. and Michel, B. (2005) 'Comparative and Evolutionary Analysis of the Bacterial Homologous Recombination Systems', *PLoS Genetics*, 1(2), p. e15. doi: 10.1371/journal.pgen.0010015.

Rollie, C., Graham, S., Rouillon, C. and White, M. F. (2018) 'NAR breakthrough article: Prespacer processing and specific integration in a type I-A CRISPR system', *Nucleic Acids Research*. Oxford University Press, 46(3), pp. 1007–1020. doi: 10.1093/nar/gkx1232.

Rollie, C., Schneider, S., Brinkmann, A. S., Bolt, E. L. and White, M. F. (2015) 'Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition', *eLife*, 4(AUGUST2015), pp. 1–19. doi: 10.7554/eLife.08716.

Roman, L. J. and Kowalczykowski, S. C. (1989a) 'Characterization of the adenosinetriphosphatase activity of the Escherichia coli RecBCD enzyme: relationship of ATP hydrolysis to the unwinding of duplex DNA.', *Biochemistry*, 28, pp. 2873–2881.

Roman, L. J. and Kowalczykowski, S. C. (1989b) 'Characterization of the helicase activity of the Escherichia coli RecBCD enzyme using a novel helicase assay.', *Biochemistry*, 28, pp. 2863–2873.

Rosamond, J., Telander, K. M. and Linn, S. (1979) 'Modulation of the action of the recBC enzyme of Escherichia coli K-12 by Ca2+', *Journal of Biological Chemistry*, 254(17), pp. 8646–8652.

Van Rosmalen, M., Krom, M. and Merkx, M. (2017) 'Tuning the Flexibility of Glycine-Serine Linkers to Allow Rational Design of Multidomain Proteins', *Biochemistry*, 56(50), pp. 6565–6574. doi: 10.1021/acs.biochem.7b00902.

Rudolph, C. J., Upton, A. L., Briggs, G. S. and Lloyd, R. G. (2010) 'Is RecG a general guardian of the bacterial genome?', *DNA Repair*, 9(3), pp. 210–223. doi: 10.1016/j.dnarep.2009.12.014.

Rudolph, C. J., Upton, A. L., Harris, L. and Lloyd, R. G. (2009) 'Pathological replication in cells lacking RecG DNA translocase', *Molecular Microbiology*, 73(3), pp. 352–366. doi: 10.1111/j.1365-2958.2009.06773.x.

Rutkauskas, M., Sinkunas, T., Songailiene, I., Tikhomirova, M. S., Siksnys, V. and Seidel, R. (2015) 'Directional R-loop formation by the CRISPR-cas surveillance complex cascade provides efficient off-target site rejection', *Cell Reports*. The Authors, 10(9), pp. 1534–1543. doi: 10.1016/j.celrep.2015.01.067.

Saitou, N. and Nei, M. (1987) 'The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees'', *Science*, 4(4), pp. 406–425.

Samai, P., Smith, P. and Shuman, S. (2010) 'Structure of a CRISPR-associated protein Cas2 from Desulfovibrio vulgaris', *Acta Crystallographica Section F: Structural Biology and Crystallization Communications*. International Union of Crystallography, 66(12), pp. 1552–1556. doi: 10.1107/S1744309110039801.

Sapranauskas, R., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P. and Siksnys, V. (2011) 'The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli', *Nucleic Acids Research*, 39(21), pp. 9275–9282. doi: 10.1093/nar/gkr606.

Sashital, D. G., Wiedenheft, B. and Doudna, J. A. (2012) 'Mechanism of Foreign DNA Selection in a Bacterial Adaptive Immune System', *Molecular Cell*. Elsevier Inc., 46(5), pp. 606–615. doi: 10.1016/j.molcel.2012.03.020.

Savic, D. J., Jankovic, M. and Kostic, T. (1990) 'Cellular role of DNA polymerase I', *J. Basic. Microbiol.*, 30(10), pp. 769–784.

Savitskaya, E., Semenova, E., Dedkov, V., Metlitskaya, A. and Severinov, K. (2013) 'High-throughput analysis of type I-E CRISPR / Cas spacer acquisition in E . coli', *RNA biology*, 10(5), pp. 716–725. doi: 10.4161/rna.24325.

Seed, K. D., Lazinski, D. W., Calderwood, S. B. and Camilli, A. (2013) 'A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity', *Nature*. Nature Publishing Group, 494(7438), pp. 489–491. doi: 10.1038/nature11927.

Semenova, E., Savitskaya, E., Musharova, O., Strotskaya, A. and Vorontsova, D. (2016) 'Highly efficient primed spacer acquisition from targets destroyed by the Escherichia coli type I-E CRISPR-Cas interfering complex', *PNAS*, 113(27). doi: 10.1073/pnas.1602639113.

Shah, S. A., Erdmann, S., Mojica, F. J. M. and Garrett, R. A. (2013) 'Protospacer recognition motifs: Mixed identities and functional diversity', *RNA Biology*, 10(5), pp. 891–899. doi: 10.4161/rna.23764.

Sharples, G. J., Ingleston, S. M. and Lloyd, R. G. (1999) 'Holliday junction processing in bacteria: Insights from the evolutionary conservation of RuvABC,

RecG, and RusA', *Journal of Bacteriology*, 181(18), pp. 5543–5550.

Sharples, G. J. and Lloyd, R. G. (1990) 'A novel repeated DNA sequence located in the intergenic regions of bacterial chromosomes', *Nucleic Acids Research*, 18(22), pp. 6503–6508. doi: 10.1093/nar/18.22.6503.

Shibata, T., Dasgupta, C. and Cunningham, R. P. (1979) 'Purified Escherichia coli recA protein catalyses homologous pairing of superhelical DNA and single-stranded fragments', *Proc Natl Acad Sci USA*, 76(4), pp. 1638–1642.

Shiimori, M., Garrett, S. C., Chambers, D. P., Glover, C. V. C., Graveley, B. R. and Terns, M. P. (2017) 'Role of free DNA ends and protospacer adjacent motifs for CRISPR DNA uptake in pyrococcus furiosus', *Nucleic Acids Research*, 45(19), pp. 11281–11294. doi: 10.1093/nar/gkx839.

Shiimori, M., Garrett, S. C., Graveley, B. R. and Terns, M. P. (2018) 'Cas4 Nucleases Define the PAM, Length, and Orientation of DNA Fragments Integrated at CRISPR Loci', *Molecular Cell*, 70(5), p. 814–824.e6. doi: 10.1016/j.molcel.2018.05.002.

Shlomai, J. and Kornberg, A. (1980) 'A prepriming DNA replication enzyme of Escherichia coli. II. Actions of protein n': a sequence-specific, DNA-dependent ATPase', *Journal of Biological Chemistry*, 255(14), pp. 6794–6798.

Singleton, M. R., Dillingham, M. S., Gaudier, M., Kowalczykowski, S. C. and Wigley, D. B. (2004) 'Crystal structure of RecBCD enzyme reveals a machine for processing DNA breaks', *Nature*, 432(7014), pp. 187–193. doi: 10.1038/nature02988.

Singleton, M. R., Scaife, S. and Wigley, D. B. (2001) 'Structural analysis of DNA replication fork reversal by RecG', *Cell*, 107(1), pp. 79–89. doi: 10.1016/S0092-8674(01)00501-3.

Soltis, D. and Lehman, I. (1983) 'recA protein promoted DNA strand exchange.', *Journal of Biological Chemistry*, 257(14), pp. 8523–8532. Available at: http://www.jbc.org/content/258/10/6073.short.

Spies, M., Bianco, P. R., Dillingham, M. S., Handa, N., Baskin, R. J. and Kowalczykowski, S. C. (2003) 'A molecular throttle: The recombination hotspot Chi controls DNA translocation by the RecBCD helicase', *Cell*, 114(5), pp. 647–654. doi: 10.1016/S0092-8674(03)00681-0.

Spies, M. and Kowalczykowski, S. C. (2006) 'The RecA binding locus of RecBCD is a general domain for recruitment of DNA strand exchange proteins', *Molecular Cell*, 21(4), pp. 573–580. doi: 10.1016/j.molcel.2006.01.007.

Steczkiewicz, K., Muszewska, A., Knizewski, L., Rychlewski, L. and Ginalski, K.

(2012) 'SURVEY AND SUMMARY: Sequence, structure and functional diversity of PD-(D/E)XK phosphodiesterase superfamily', *Nucleic Acids Research*, 40(15), pp. 7016–7045. doi: 10.1093/nar/gks382.

Stern, M. J., Ames, G. F., Smith, N. H., Robinson, E. C. and Higgins, C. F. (1984) 'Repetitive Extragenic Palindromic Sequences : A Major Component of the Bacterial Genome', *Cell*, 37, pp. 1015–1026.

Sticht, H. and Rösch, P. (1998) 'The structure of iron-sulfur proteins', *Progress in Biophysics and Molecular Biology*, 70(2), pp. 95–136. doi: 10.1016/S0079-6107(98)00027-3.

Storm, P. K., Hoekstra, W. P. M., De Haan, P. G. and Verhoef, C. (1971) 'Genetic recombination in Escherichia coli. IV. Isolation and characterization of recombination- deficient mutants of Escherichia coli K12.', *Mutat. Res.*, 13, pp. 9–17.

Sugimoto, K., Okazaki, T. and Okazaki, R. (1968) 'Mechanism of DNA chain growth. II: Accumulation of newly synthesised short chains in E. coli infected with ligase-defective T4 phages.', *Biochemistry*, 60, pp. 1356–1362.

Sugino, A., Hirose, S. and Okazaki, R. (1972) 'RNA-Linked Nascent DNA Fragments in Escherichia coli', *Proc Natl Acad Sci USA*, 69(7), pp. 1863–1867.

Sun, J. Z., Julin, D. A. and Hu, J. S. (2006) 'The nuclease domain of the Escherichia coli RecBCD enzyme catalyzes degradation of linear and cir- cular single-stranded and double-stranded DNA.', *Biochemistry*, 45, pp. 131–140.

Swarts, D. C., Mosterd, C., van Passel, M. W. J. and Brouns, S. J. J. (2012) 'CRISPR interference directs strand specific spacer acquisition', *PLoS ONE*, 7(4), pp. 1–7. doi: 10.1371/journal.pone.0035888.

Szostak, J. W., Orr-Weaver, T. L., Rothstein, R. J. and Stahl, F. W. (1983) 'The double-strand-break repair model for recombination', *Cell*, 33(1), pp. 25–35. doi: 10.1016/0092-8674(83)90331-8.

Takeuchi, N., Wolf, Y. I., Makarova, K. S. and Koonin, E. V. (2012) 'Nature and intensity of selection pressure on CRISPR-associated genes.', *J. Bacteriol.*, 194, pp. 1216–1225.

Talib, J., Cook, N., Pattison, D. and Davies, M. (2014) 'Disruption of the iron-sulfur cluster of aconitase by myeloperoxidase-derived oxidants', *Free Radical Biology and Medicine*, 75, pp. S27–S28. doi: 10.1016/j.freeradbiomed.2014.10.752.

Tanaka, T., Mizukoshi, T., Sasaki, K., Kohda, D. and Masai, H. (2007) 'Escherichia coli PriA protein, two modes of DNA binding and activation of ATP hydrolysis',

Journal of Biological Chemistry, 282(27), pp. 19917–19927. doi: 10.1074/jbc.M701848200.

Tanaka, T., Mizukoshi, T., Taniyama, C., Kohda, D., Arai, K. I. and Masai, H. (2002) 'DNA binding of PriA protein requires cooperation of the N-terminal D-loop/arrested-fork binding and C-terminal helicase domains', *Journal of Biological Chemistry*, 277(41), pp. 38062–38071. doi: 10.1074/jbc.M204397200.

Tanaka, T., Taniyama, C., Arai, K. I. and Masai, H. (2003) 'ATPase/helicase motif mutants of Escherichia coli PriA protein essential for recombination-dependent DNA replication', *Genes to Cells*, 8(3), pp. 251–261. doi: 10.1046/j.1365-2443.2003.00630.x.

Tang, T.-H., Bachellerie, J.-P., Rozhdestvensky, T., Bortolin, M.-L., Huber, H., Drungowski, M., Elge, T., Brosius, J. and Hüttenhofer, A. (2002) 'Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus.', *Proceedings of the National Academy of Sciences of the United States of America*, 99(11), pp. 7536–7541. doi: 10.1073/pnas.112047299.

Tang, T. H., Polacek, N., Zywicki, M., Huber, H., Brugger, K., Garrett, R., Bachellerie, J. P. and Hüttenhofer, A. (2005) 'Identification of novel non-coding RNAs as potential antisense regulators in the archaeon Sulfolobus solfataricus', *Molecular Microbiology*, 55(2), pp. 469–481. doi: 10.1111/j.1365-2958.2004.04428.x.

Taylor, A. F. and Smith, G. R. (2003) 'RecBCD enzyme is a DNA helicase with fast and slow motors of opposite polarity.', *Nature*, 423, pp. 889–893. doi: 10.1038/nature01747.1.

Touchon, M., Charpentier, S., Clermont, O., Rocha, E. P. C., Denamur, E. and Branger, C. (2011) 'CRISPR Distribution within the Escherichia coli Species Is Not Suggestive of Immunity-Associated Diversifying Selection', *Journal of Bacteriology*, 193(10), pp. 2460–2467. doi: 10.1128/JB.01307-10.

Touchon, M. and Rocha, E. P. C. (2010) 'The small, slow and specialized CRISPR and anti-CRISPR of Escherichia and Salmonella', *PLoS ONE*, 5(6). doi: 10.1371/journal.pone.0011126.

Upton, A. L., Grove, J. I., Mahdi, A. A., Briggs, G. S., Milner, D. S., Rudolph, C. J. and Lloyd, R. G. (2014) 'Cellular location and activity of Escherichia coli RecG proteins shed light on the function of its structurally unresolved C-terminus', *Nucleic Acids Research*, 42(9), pp. 5702–5714. doi: 10.1093/nar/gku228.

Vincent, S. D., Mahdi, A. A. and Lloyd, R. G. (1996) 'The RecG branch migration

protein of Escherichia coli dissociates R-loops', *Journal of Molecular Biology*, 264(4), pp. 713–721. doi: 10.1006/jmbi.1996.0671.

Wagar, M. A. and Huberman, J. A. (1973) 'Evidence for the attachment of RNA to pulse-labeled DNA in the slime mold, Physarum polycephalum.', *Biochem. Biophys. Res. Commun.*, 51, pp. 174–80.

Wang, J., Chen, R. and Julin, D. A. (2000) 'A single nuclease active site of the Escherichia coli RecBCD enzyme catalyzes single-stranded DNA degrada- tion in both directions.', *J. Biol. Chem.*, 275, pp. 507–513.

Wang, J., Chen, R. and Julin, D. A. (2000) 'A single nuclease active site of the Escherichia coli RecBCD enzyme catalyzes single-stranded DNA degradation in both directions', *Journal of Biological Chemistry*, 275(1), pp. 507–513. doi: 10.1074/jbc.275.1.507.

Wang, J., Li, J., Zhao, H., Sheng, G., Wang, M., Yin, M. and Wang, Y. (2015) 'Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems', *Cell*. Elsevier Inc., 163(4), pp. 840–853. doi: 10.1016/j.cell.2015.10.008.

Weinstock, G. M., McEntee, K. and Lehman, I. R. (1979) 'ATP-dependent renaturation of DNA catalyzed by the recA protein of Escherichia coli.', *Proceedings of the National Academy of Sciences of the United States of America*, 76(1), pp. 126–130. doi: 10.1073/pnas.76.1.126.

Westergaard, O., Brutlag, D. and Kornberg, A. (1973) 'Initiation of Deoxyribonucleic Acid Synthesis. IV: Incorporation of the ribonucleic acid primer into the phage replicative form.', *The Journal of biological chemistry*, 248(4), pp. 1361–1364.

Westra, E. R., van Erp, P. B. G., Künne, T., Wong, S. P., Staals, R. H. J., Seegers, C. L. C., Bollen, S., Jore, M. M., Semenova, E., Severinov, K., de Vos, W. M., Dame, R. T., de Vries, R., Brouns, S. J. J. and van der Oost, J. (2012) 'CRISPR Immunity Relies on the Consecutive Binding and Degradation of Negatively Supercoiled Invader DNA by Cascade and Cas3', *Molecular Cell*, 46(5), pp. 595–605. doi: 10.1016/j.molcel.2012.03.018.

Westra, E. R., Pul, Ü., Heidrich, N., Jore, M. M., Lundgren, M., Stratmann, T., Wurm, R., Raine, A., Mescher, M., Van Heereveld, L., Mastop, M., Wagner, E. G. H., Schnetz, K., Van Der Oost, J., Wagner, R. and Brouns, S. J. J. (2010) 'H-NS-mediated repression of CRISPR-based immunity in Escherichia coli K12 can be relieved by the transcription activator LeuO', *Molecular Microbiology*, 77(6), pp.

1380–1393. doi: 10.1111/j.1365-2958.2010.07315.x.

Whitby, M. C. and Lloyd, R. G. (1998) 'Targeting Holliday junctions by the RecG branch migration protein of Escherichia coli', *Journal of Biological Chemistry*, 273(31), pp. 19729–19739. doi: 10.1074/jbc.273.31.19729.

Whitby, M. C., Vincent, S. and Lloyd, R. G. (1994) 'Branch migration of holliday junctions: identification of RecG protein as a junction specific DNA helicase.', *Embo J*, 13, pp. 5220–5228.

Willetts, N. S. and Clark, A. J. (1969) 'Characteristics of some multiply recombination-deficient strains of Escherichia coli.', *Journal of Bacteriology*, 100(1), pp. 231–239.

Wong, C. J., Rice, R. L., Baker, N. A., Ju, T. and Lohman, T. M. (2006) 'Probing 3'-ssDNA Loop Formation in E. coli RecBCD/RecBC-DNA Complexes Using Non-natural DNA: A Model for "Chi" Recognition Complexes', *Journal of Molecular Biology*, 362(1), pp. 26–43. doi: 10.1016/j.jmb.2006.07.016.

Wright, M., Buttin, G. and Hurwitz, J. (1971) 'The isolation and characterization from Escherichia coli of an adenosine triphosphate-dependent deoxyribonuclease directed by RecB,C genes.', *J. Biol. Chem.*, 246, pp. 6543–6555.

Xiao, Y., Luo, M., Hayes, R. ., Kim, J., Ng, S., Ding, F., Liao, M. and Ke, A. (2017) 'Structure Basis for Directional R-loop Formation and Substrate Handover Mechanisms in Type I CRISPR-Cas System.', *Cell.*, 170(1), pp. 48–60.

Xue, C., Zhu, Y., Zhang, X., Shin, Y. . and Sashital, D. G. (2017) 'Real-Time Observation of Target Search by the CRISPR Surveillance Complex Cascade', *Cell. Rep.*, 21(13), pp. 3717–3727.

Yan, B. and Sun, Y. (1996) 'Glycine Residies Provide Flexibility for Enzyme Active Sites', *Journal of Biological Chemistry*, 272(6), pp. 3190–3194.

Yang, C. C. and Nash, H. a (1989) 'The interaction of E. coli IHF protein with its specific binding sites.', *Cell*, 57(5), pp. 869–880. doi: 10.1016/0092-8674(89)90801-5.

Yeeles, J. T. P., Cammack, R. and Dillingham, M. S. (2009) 'An iron-sulfur cluster is essential for the binding of broken DNA by AddAB-type helicase-nucleases', *Journal of Biological Chemistry*, 284(12), pp. 7746–7755. doi: 10.1074/jbc.M808526200.

Yin, H., Xue, W., Chen, S., Bogorad, R. L., Benedetti, E., Grompe, M., Koteliansky, V., Sharp, P. A., Jacks, T. and Anderson, D. G. (2014) 'Genome editing with Cas9 in adult mice corrects a disease mutation and phenotype', *Nature Biotechnology*,

32(6), pp. 551–553. doi: 10.1038/nbt.2884.

Yoganand, K. N. R., Sivathanu, R., Nimkar, S. and Anand, B. (2017) 'Asymmetric positioning of Cas1-2 complex and Integration Host Factor induced DNA bending guide the unidirectional homing of protospacer in CRISPR-Cas type I-E system', *Nucleic Acids Research*, 45(1), pp. 367–381. doi: 10.1093/nar/gkw1151.

Yosef, I., Goren, M. G. and Qimron, U. (2012) 'Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli', *Nucleic Acids Research*, 40(12), pp. 5569–5576. doi: 10.1093/nar/gks216.

Yu, M., Souaya, J. and Julin, D. A. (1998) 'Identification of the nuclease active site in the multifunctional RecBCD enzyme by creation of a chimeric enzyme', *Journal of Molecular Biology*, 283(4), pp. 797–808. doi: 10.1006/jmbi.1998.2127.

Yu, M., Souaya, J. and Julin, D. A. (1998) 'The 30-kDa C-terminal domain of the RecB protein is critical for the nuclease activity, but not the helicase activity, of the RecBCD enzyme from Escherichia coli.', *Proc. Natl. Acad. Sci. USA*, 95, pp. 981–986.

Zhang, J., Kasciukovic, T. and White, M. F. (2012) 'The CRISPR Associated Protein Cas4 Is a 5′ to 3′ DNA Exonuclease with an Iron-Sulfur Cluster', *PLoS ONE*, 7(10). doi: 10.1371/journal.pone.0047232.

Zhu, J., Zheng, H., Ai, G., Zhang, G., Liu, X. and Dong, X. (2012) 'The genome characteristics and predicted function of methyl-group oxidation pathway in the obligate aceticlastic methanogens, Methanosaeta spp.', *PLoS ONE*, 7(5), p. e36756.