Molecular investigations of the segregation proteins of plasmid pNOB8, harboured by a novel *Sulfolobus islandicus* strain

John Armstrong

PhD

University of York
Biology

December 2021

Abstract

Genome segregation is a fundamental process, occurring across all domains of life. Bacterial chromosomal and plasmid partitioning systems are well understood, however, little is known about genome segregation in archaea. This work utilises the low copy number plasmid pNOB8, from the thermophilic *Sulfolobus* NOB8-H2, to investigate archaeal DNA segregation mechanisms. The plasmid pNOB8 harbours a partition cassette, analogous to bacterial segregation systems, encoding proteins AspA, ParB and ParA. AspA is a site-specific DNA-binding protein, which binds with high affinity to a palindromic DNA sequence upstream of the *aspA* gene, and previous studies established that AspA can spread along the DNA from this motif and form an extended nucleoprotein complex.

Another identical palindrome is located elsewhere on pNOB8, and this work describes the interactions of AspA with the DNA at this second site. Band-shift and DNase I footprinting assays demonstrated that AspA binds avidly to the second palindrome *in vitro*, and forms two discrete binding regions on the DNA, located in the putative promoter region of an uncharacterised operon. We speculate that AspA may act as a transcriptional regulator at the second palindrome, and propose a model describing the functional roles of AspA at each binding site. Furthermore, an array of AspA mutants were generated, and the specific amino acids that contribute to crucial properties of the protein, such as DNA-binding, dimer-dimer interactions, and dimerisation are reported.

Additionally, to assess the dynamic interactions between pNOB8 and the chromosome, the host genome was sequenced, and phylogenetic analyses revealed NOB8-H2 to be a novel strain of *Sulfolobus islandicus*. Strain NOB8-H2 possesses two CRISPR-Cas systems, a dissection of which provides evidence of ongoing evolutionary competition between plasmid and host, and the possibility that pNOB8 may encode an anti-CRISPR protein provides an interesting avenue for future studies.

292 words

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to sincerely thank my main project supervisor, Professor Daniela Barillà, for her unending support, patience, intellectual and practical advice, and for her encouragement to persevere when things were tough. A PhD is a very long and difficult journey, and I could not have completed it without her help. I would also like to extend my gratitude to co-supervisor Prof James Chong, and TAP members Dr Paul Pryor and Dr Michael Plevin, for their advice and support throughout the last four (and a half) years.

I also would not have been able to complete this work without the considerable and selfless assistance of current and past Barillà lab members; particularly Dr Azhar Kabli, Dr Cecilia Pennica, Dr Iman Alnaqshabandy and Dr Nicholas Read. I am also grateful to other staff members in the Department of Biology who assisted on various aspects of the project, including, but not limited to; Dr Andrew Leech, Dr John Davey, Dr James Robson, and Dr Sally James. Other notable *Homo sapiens* who have provided me with scientific assistance, advice and most importantly friendship are too numerous to mention; but include the people of L1 corridor and A Block, fellow PhD cohort members (one of whom is sadly no longer with us), teaching, infrastructure and administrative staff who I have come to know, and of course the lovely Jenny in Cookies, who kept me fed and watered.

Special thanks must go to fellow PhD candidate Theresa Leslie, who became my musical collaborator and valuable friend during the last four years. Playing guitar and singing with Tess has provided me with much joy, and helped me to preserve some sanity during difficult times. A particular mention also goes to Dr Aritha Dornau, another fellow musical and scientific confidant, with whom I was able to share some adventures along the way.

Lastly, I also thank my friends and family back in Liverpool (and other parts of the globe) who I see much less of now than I once did, but for whose company I am always grateful. To close, I must mention my wonderful niece Maya, who was diagnosed with cancer during the third year of this PhD. Her remarkable equanimity and continued good nature (along with her Periodic Table bedroom wallpaper!) in the face of such adversity remain a constant source of inspiration and pride to me.

Author's Declaration

I, John Armstrong, declare that this thesis is a presentation of original work, and that I am the sole author. Any work conducted by other persons is noted in the text. This work has not been previously been submitted for an award at this, or any other University. All sources are acknowledged as references.

John Armstrong

Chapter 1

Introduction

Chapter 1

Introduction

1.1 Plasmids

Genome segregation, the partitioning of newly-replicated DNA molecules and their subsequent delivery to daughter cells, is a fundamental and vital cellular process that occurs across the three biological domains. Segregation of genomic material, comprising both chromosomes and plasmids, must occur accurately in order for correct ploidy to be maintained in subsequent cellular generations.

Plasmids are extrachromosomal genetic elements; predominantly circular double-stranded DNA molecules, capable of autonomous replication independent of the chromosome (Sherratt 1974). Plasmids are prevalent in bacteria, but are also found to inhabit archaeal cells (Zillig *et al*. 1996), and some eukaryotes, including yeasts and plant mitochondria (Handa 2008, Utatsu *et al*. 1987). Plasmids are small molecules compared to the host chromosome, but vary in size, from a few kilobases to greater than one megabase. They are usually present in multiple copies; ranging from larger size, low copy-number plasmids, to artificially derived cloning plasmids, which may be present in hundreds of copies per cell. Copy number is usually stable within a given host under defined growth conditions (del Solar & Espinosa 2000).

Typically, the presence of a plasmid induces a metabolic burden on the host (Million-Weaver & Camps 2014), however in return, plasmids confer benefits to their hosts, encoding non-essential proteins that aid pathogenicity and virulence (Wang 2017), or that allow the host to survive in different environments and compete with other microorganisms in the same specific niche (Heuer & Smalla 2012). Bacterial plasmids frequently encode antibiotic resistance genes, allowing the host to survive and persist when challenged with antibiotics (Bennett 2008, Jacob & Hobbs 1974). Furthermore, these resistance genes can be disseminated on plasmids throughout a population (Li *et al.* 2019), providing resistance *en masse*. Plasmids are also important drivers of evolution via horizontal and vertical gene transfer, due to their ability to transfer between cells,

modulate gene expression levels with changes in copy number, and integrate into the host chromosome (Hülter *et al.* 2017, She *et* al. 2004, Shintani *et al*. 2015). Plasmids also typically harbour genes whose products self-regulate such processes as plasmid copy number control (del Solar *et al.* 2002), conjugation mechanisms (Firth & Skurray 1992) and maintenance and partitioning systems (Pilla & Tang 2018).

1.2 Mechanisms of plasmid maintenance

1.2.1 Random diffusion model

Given the usefulness of plasmids to their hosts, it is clearly important that plasmids are maintained in a population, and after DNA replication, are successfully inherited by future cellular generations. In the case of high-copy number plasmids, which, unlike their low copy-number counterparts, do not harbour genes encoding active partitioning systems (Baxter & Funnell 2014), random diffusion alone has been advanced to explain their delivery to daughter cells (Durkacz & Sherratt 1973). The random diffusion model assumes that if plasmids are present in sufficient number, they will stochastically be segregated and each daughter cell should inherit at least one plasmid copy. In mathematical terms, the probability of a plasmid-free daughter cell after binary fission can be calculated using the equation: $\mathrm{p_o} = 2^{1-n}$, where n represents the number of plasmids in the cell. This means that for a cell containing ten plasmid copies, the probability of a plasmid-free daughter cell is approximately 1 in 500, and clearly for cells harbouring much greater copy numbers, this probability becomes incredibly small. The stochastic dissemination of plasmids would therefore be sufficient for their stable maintenance in a population, and removes the requirement for an active segregation system **(Figure 1.1)**.

However, data from microscopy experiments to visualise plasmid localisation has led to the random diffusion model being questioned. The bacterial plasmid ColE1 was found to localise mainly at cell poles, and was excluded from the nucleoid, the volume of the chromosome (Reyes-Lamothe *et al*. 2014). Another study using fluorescently-labelled high copy-number plasmids again found them to cluster in large numbers at both the cell

poles and mid-cell, with plasmid replication also occurring in the nucleoid-free spaces at the cell poles. Here, occlusion from the nucleoid and subsequent positioning at cell poles ensures that daughter cells receive newly-replicated plasmids (Hsu & Chang 2019). These observations and many others have generated alternative hypotheses for the segregation of high-copy number plasmids. One suggestion is that high-copy number plasmid segregation is a regulated process, relying on interactions of the plasmid origin of replication with chromosomally-encoded proteins rather than those present on the plasmid itself, although these factors await identification (Million-Weaver & Camps 2014). Alternatively, it has been demonstrated that high-copy number plasmids are present both in clusters, and also single copies that are randomly distributed throughout the cell volume, including within the nucleoid volume. This 'hybrid distribution' model posits that a combination of random distribution plus clustering is sufficient to maintain a stable population of plasmids without requiring any active partitioning mechanism (Wang *et al*. 2016, Wang 2017).

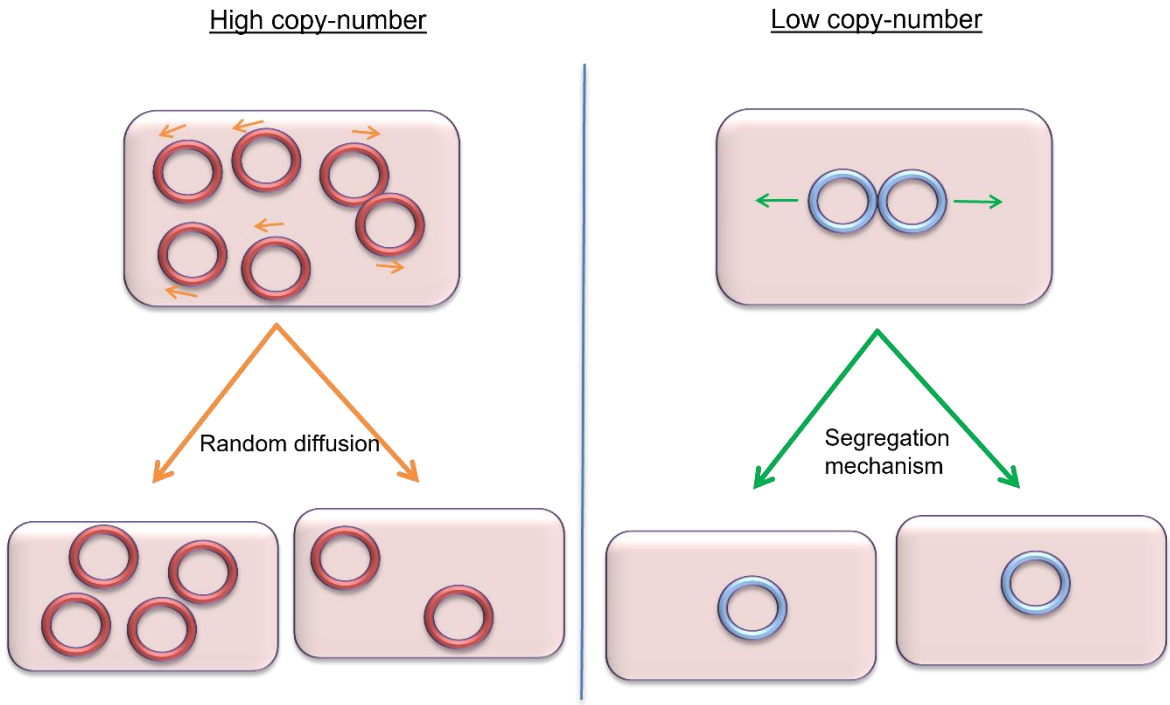


Figure 1.1. Segregation of high and low copy-number plasmids. (Left) Plasmids occurring at sufficiently high numbers within the cell could be stochastically transmitted to daughter cells by random diffusion alone. (Right) Low copy-number plasmids cannot rely on random diffusion, instead encoding an active segregation system.

1.2.2 Post-segregational killing

Post-segregational killing, also known as toxin/antitoxin (TA) or addiction systems, is another strategy by which plasmid loss within a population may be controlled. The basis for this is that daughter cells which do not inherit a copy of the plasmid because of replication and/or segregation defects will be killed, thus removing plasmid-free cells from the population **(Figure 1.2a)** (Hayes 2003, Tsang 2017). The TA system was first elucidated several decades ago, and was found encoded on the low copy number *Escherichia coli* plasmids F and R1 (Ogura & Hiraga 1983, Gerdes *et al*. 1986). TA systems generally comprise two genes, encoding a toxin and an antitoxin, of which the toxin is a protein, and the antitoxin may be a protein or non-coding RNA. The toxin protein kills or disarms cells from within, targeting a range of different structures or processes necessary for the cell's growth or survival; including disrupting protein synthesis by RNA cleavage, interfering with chromosome topology and DNA replication, and damaging the cell membrane (Kędzierska & Hayes 2016).

The antitoxin protein may either bind the toxin and neutralise its activity, or, if the antitoxin is a small RNA, it may inhibit translation of the toxin mRNA. The TA complex is inherited by daughter cells even in the absence of the plasmid, and after degradation of the antitoxin by host enzymes, the toxin is free to damage the host cell as the antitoxin cannot be regenerated (Hayes 2003). TA systems are also found to be chromosomally-encoded, and multiple different TA modules may be found on the same chromosome (Kędzierska & Hayes 2016), although the function of chromosomal TA loci is to ensure persistence and adaptation under different environmental stresses (Gerdes *et al.* 2005, Ramage *et al*. 2009).

TA systems are characterised according to the mechanistic basic of neutralisation by the antitoxin, and currently six classes or types have been identified, with the Type I and III antitoxin gene product being a non-coding RNA, and Types II, IV, V and VI antitoxin genes encoding a small protein (Page & Peti 2016). The first Type I TA system to be identified was the *hok/sok* gene pair in plasmid R1, where *hok* (**h**ost **k**illing) encodes the toxin protein, a short transmembrane peptide that disrupts the host cell membrane (Gerdes & Wagner 2007). The *sok* gene (**s**uppressor of **k**illing) encodes a rapidly-decaying antisense

RNA that base-pairs with *hok* mRNA, which originally was thought to directly inhibit Hok translation. However, a third gene *mok* (**m**odulation of **k**illing) was discovered that overlaps with and regulates *hok* expression, and it is the translation of *mok* that is blocked by Sok RNA (Thisted & Gerdes 1992). Thus, Sok RNA inhibits *hok* translation indirectly, whilst the RNA duplex formed by Sok RNA and *hok* mRNA is targeted for cleavage by RNase III **(Figure 1.2b)** (Gerdes & Wagner 2007).

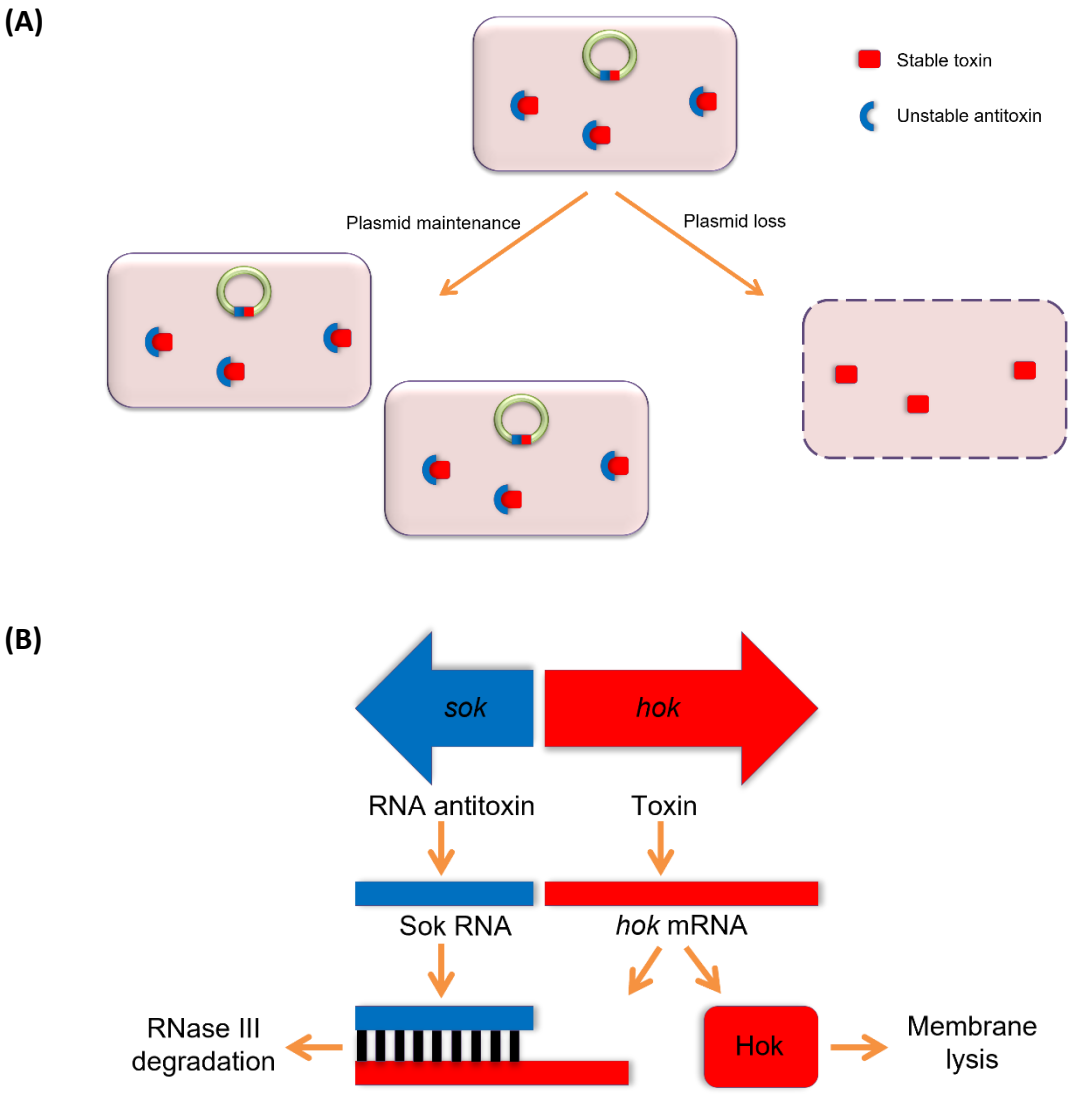


Figure 1.2. Post-segregational killing, or Toxin/Antitoxin (TA) plasmid maintenance. **(A)** The differing fates of cells which either inherit plasmids, or suffer plasmid loss after segregation. If plasmids are lost, anti-toxins will be degraded and not replenished, causing cell death. Adapted from Tsang 2017. **(B)** Cartoon representation of the *hok/sok* Type I TA system, showing the toxin protein in red, and the RNA antitoxin in blue. The system has been simplified to not include the *mok* gene, which overlaps with *hok*. Adapted from Page & Peti 2016.

1.2.3 Active plasmid partitioning mechanisms

Given that plasmids occurring in low copy numbers (e.g. less than ten copies per cell) cannot rely on stochastic diffusion mechanisms alone for their maintenance within a population, an active mechanism by which plasmids can be segregated and delivered to daughter cells is necessary (Bouet & Funnell 2019). This mechanism, mediated by DNA partition systems, is widespread across microbial taxa, and is found to be both chromosomally and plasmid-encoded (Baxter & Funnell 2014). Given their relative simplicity, bacterial plasmids have been utilised as model systems to understand genome segregation systems through study of plasmid partition systems (Hayes & Barillà 2006b, Schumacher 2008).

The active partition systems (*par* for short) have a relatively simple genetic organisation, typically being comprised of three elements; two proteins and a centromere-like DNA sequence that acts as a site of recruitment of the proteins and assembly of a nucleoprotein complex dubbed the segrosome (Hayes & Barillà 2006b). The DNA partition sites, which may be thought of as functional analogues of eukaryotic centromeres, can be located either upstream or downstream of the partition operon and differ in their organisation, but comprise repeat DNA sequences that may be direct or inverted, or contain palindromic nucleotide arrangements (Bouet & Funnell 2019, Hayes & Barillà 2006a, Schumacher 2008). The DNA partition site is generally denoted as *parS*, however the nomenclature of the centromere differs across systems **(Figure 1.3)**.

The two proteins of the partition system are a DNA-binding factor, often termed ParB, which exhibits specificity for the *parS* site(s), and an ATPase or GTPase motor protein denoted ParA, which is recruited into the nucleoprotein complex formed by ParB-*parS* interactions. The *parA* and *parB* genes are typically encoded in the same operon (Schumacher 2008). Thus, the typical genetic organisation of a plasmid partition module is *parABS*, and this nomenclature will be used henceforth when describing these components in general, although the names of individual proteins and centromeres may change dependent on the particular system. The *parABS* archetypal partition system was first described for low copy number P1 plasmid of *E. coli* (Austin & Abeles 1983).

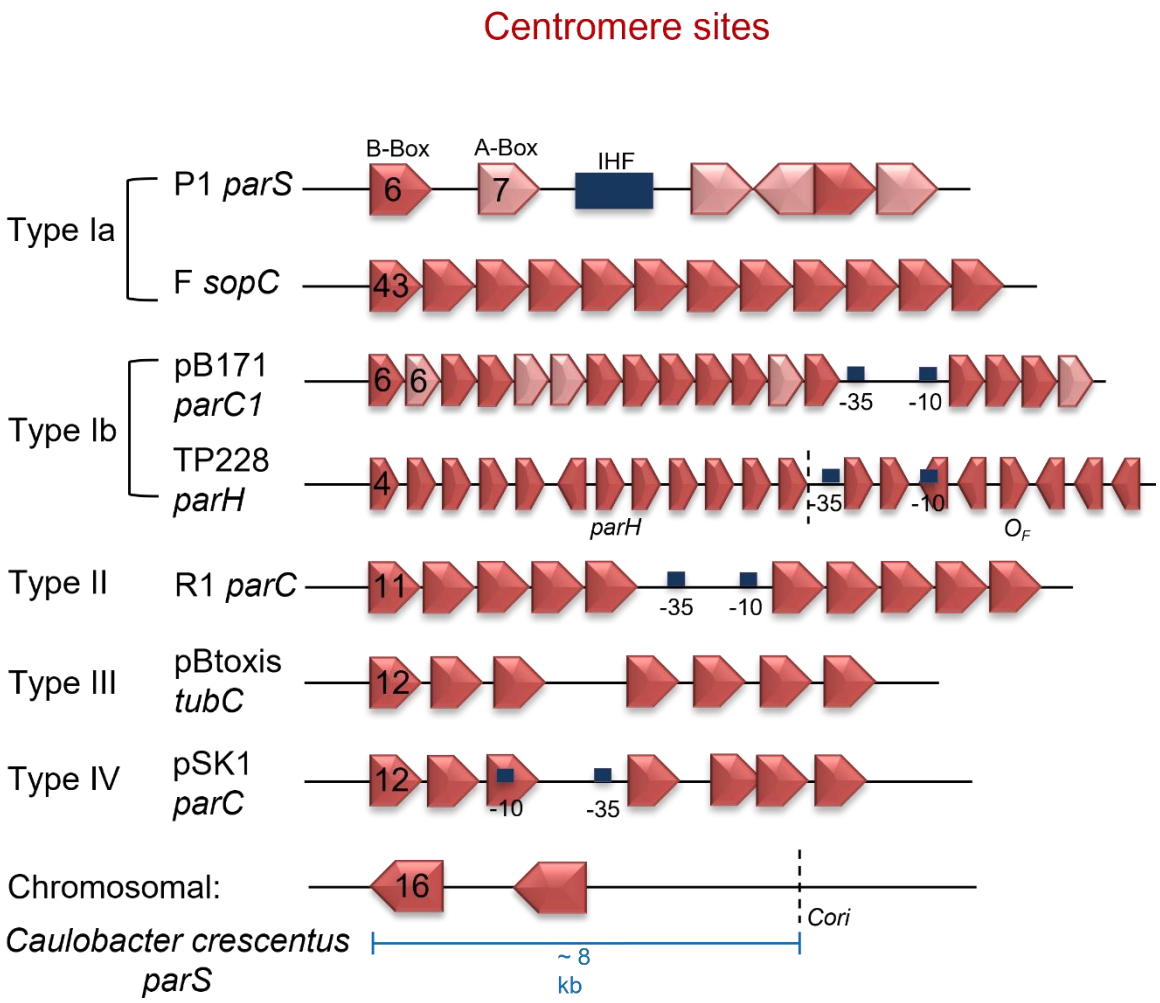


Figure 1.3. The diversity of *par* centromere-like sequences. A schematic representation of the different organisations of centromeric sites found in each *par* system. The arrows represent the orientation of the direct or inverted repeats that comprise the centromere. The numbers inside the first arrows indicate the length of the repeat in base-pairs. Different shadings represent different repeat motifs. The -35 and -10 promoter boxes are displayed where relevant. All centromeres depicted are from plasmids, except the bottom one which depicts the chromosomal *parS* sites of *C. crescentus*, which is located ~ 8 kb from the origin of replication (*Cori*). The *par* arrows are not to scale. IHF = integration host factor. Adapted from Hayes & Barillà 2006a, Toro *et al.* 2008, Bouet & Funnell 2019.

Partition systems are subdivided into several classes, based on the characteristics of the ParA motor protein. Currently, three main types of partition system are known, with a fourth segregation system type more recently proposed. Type I systems, the first to be elucidated and including the *parABS* module from plasmid P1, encode a Walker-type ATPase, named for its Walker-A motif that mediates binding to ATP. Type I systems are the most widespread of the partition systems, and are subdivided into Types Ia and Ib, based on the relative size of the Par proteins and the positioning of the centromere relative to the *par* genes. In Type Ia systems, which include those from *E. coli* plasmids P1 and F, the Par proteins are larger, and the centromere is located downstream of the *par* operon. In Type Ib systems, their Par protein counterparts are smaller, and the centromeric site is found upstream of the *par* operon **(Figure 1.4)**.

Type II partition systems include an ATPase protein which belongs to the actin/heat-shock protein superfamily and is therefore an evolutionary homologue of eukaryotic actin, which performs a structural role in the formation of cytoskeletal filaments in eukaryotic cells. The Type II ATPase is therefore labelled actin-like. Type III systems encode an GTPase motor protein which again has ancestral homology to another eukaryotic cytoskeletal protein tubulin, and is labelled tubulin-like. More recently, non-canonical partition systems have been uncovered on the broad host-range plasmid R388 and plasmid pSK1 from *Staphylococcus aureus* (Guynet & de la Cruz 2011, Dmowski & Jagura-Burdzy 2013). In both cases, it appears that only a single DNA-binding protein is required for plasmid stability without the requirement for an ATPase motor protein. Although the mechanisms underpinning segregation are unknown, these modules are proposed as Type IV partition systems. The genetic organisation and componentry of the four types of partition system are shown in **Figure 1.4**, and have been reviewed extensively (Hayes & Barillà 2006a, Schumacher 2008, Barillà 2010, Million-Weaver & Camps 2014, Baxter & Funnell 2014, Misra *et al*. 2018, Bouet & Funnell 2019). Type I partition systems are not only encoded on plasmids, but are also found on many bacterial chromosomes (Wang *et al*. 2013), with bacterial chromosomes often harbouring several *parS* sites close to the replication origin (Badrinarayanan *et al*. 2015, Funnell 2016).

Although knowledge of prokaryotic DNA segregation systems has primarily come from studies of bacteria, similar partition cassettes have more recently been described on archaeal chromosomes and plasmids, one of which is the focus of this study. The chromosomal *segAB* locus of *S. solfataricus* was found to encode SegA, a ParA orthologue, and SegB, an archaea-specific factor that nevertheless displayed sites-specific DNA-binding activity similar to bacterial ParB proteins (Kalliomaa-Sanford *et al*. 2012). The *segAB* cassette is not specific to *S. solfataricus,* as it was found to be present in a variety of species within the Crenarchaea and Euryarchaea phyla (Barillà 2016).

Of particular relevance to this study is the partition cassette encoded on the conjugative plasmid pNOB8, harboured by the *Sulfolobus* strain NOB8-H2. This cassette is unusual as it comprises three genes rather the more common bicistronic arrangement found in bacteria. Two of the gene products share homology with bacterial ParB and ParA family proteins, whilst the third, AspA, did not (Schumacher *et al*. 2015). Similar to *segAB*, the *aspA-parB-parA* locus was not confined to pNOB8, but is harboured by a range of crenarchaeal species, on chromosomes and plasmids (Schumacher *et al*. 2015).

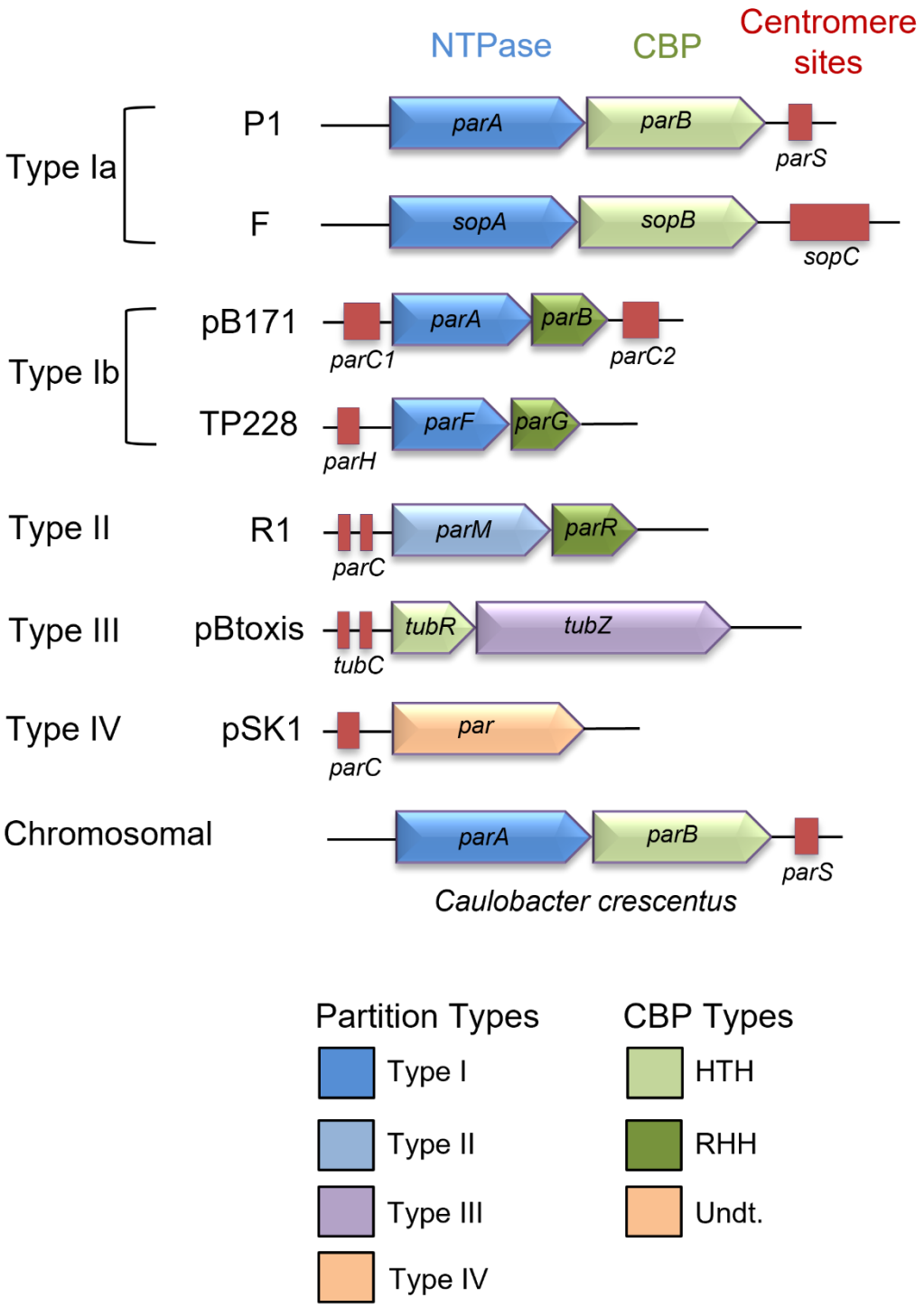


Figure 1.4. Genetic organisation of different partition system classes. The partition cassettes of different *par* systems types are shown. Partition types are named based on their NTPase proteins, whose genes are coloured dark and light blue for Types I and II, and purple for Type III. The genes encoding centromere-binding proteins (CBPs) are coloured light and dark green for helix-turn-helix and ribbon-helix-helix protein motifs respectively. The Type IV system is labelled orange, and the mode of binding of its Par protein is undetermined. The chromosomally-encoded *par* system from *C. crescentus* is also shown. Genes are drawn approximately to scale. Adapted from Bouet & Funnell 2019.

1.3 Segregation system types

1.3.1 Type I segregation systems

Type I partition systems, defined by their Walker-type ATPase motor protein, were the first to be described for plasmids P1 and F in *E. coli*, and are the most prevailing type found on plasmids, having also been characterised for plasmids RK2 and pB171 in *E. coli*, and plasmid TP228 in *Salmonella enterica* (Bouet & Funnell 2019). The CBP for type I systems is generally known as ParB, although homologues encoded on distinct plasmids may have different names (e.g. SopB), and non-homologous but functionally analogous proteins such as ParG of TP228 also belong to Type I systems **(Figure 1.4)**. Type I systems can be further subdivided into Type Ia and Type Ib; in each case the *parA* gene is upstream of *parB*, however the centromere location varies; in Type Ia systems it is downstream of *parB*, but upstream of the *par* operon in Type Ib systems **(Figure 1.4)** (Schumacher 2008). The size of the partition proteins is also a distinguishing factor between Type Ia and Type Ib systems; in Type Ia systems both proteins are on average larger than their Type Ib counterparts. Typical sizes for ParA of both systems are 251-420 amino acids (Ia) *cf.* 208-227 (Ib), and for ParB they range from 182-336 (Ia) *cf*. 46-113 (Ib) (Schumacher 2008). The larger size of Type Ia ParA proteins has allowed additional functionality, as here, ParA is involved in both segregation and the transcriptional regulation of the *par* operon. This regulatory activity is mediated by an additional N-terminal region of around 100 amino acids, containing a helix-turn-helix (HTH) DNA-binding domain. In the case of P1 ParA, binding to adenosine diphosphate (ADP) induces a conformational change that enables it to bind to the *par* operator and suppress transcription (Dunham *et al*. 2009). Type Ib ParA proteins do not act as transcriptional regulators, but in both Type I subgroups, the ParA motor proteins are capable of binding non-specifically to nucleoid DNA (Baxter & Funnell 2014). The autoregulation of the *par* operon is of vital importance to the correct maintenance of the plasmid; overproduction of either ParA or ParB has been shown to disrupt the partitioning of plasmid P1 (Abeles *et al*. 1985, Funnell 1988a). In Type Ib systems, it is the CBP which acts to autoregulate transcription, as demonstrated by ParG of TP228 (Carmelo *et al*. 2005).

1.3.1.1 Type Ia systems: *parABS* and *sopABC*

Two of the most widely studied examples of Type Ia partition systems are those harboured by *E. coli* plasmids P1 and F. The archetypal *parABS* partition system from P1 was described by Abeles and colleagues in the 1980s, who defined the *par* region as being ~2.7 kb in size, incorporating the genes encoding ParA and ParB proteins, along with the *parS* site downstream of *parAB*. The *par* region was shown to be essential for plasmid maintenance, and so described as functionally analogous to eukaryotic centromeres (Austin & Abeles 1983, Abeles *et al*. 1985). The *parS* centromere is roughly 80 bp, and displays a high level of complexity, as it comprises two flanking portions that contain a non-symmetrical arrangement of heptameric and hexameric nucleotide motifs, named the A-box and B-box respectively **(Figure 1.5)** (Hayes & Austin 1994). An additional central region between the two flanking arms had previously been shown to act as a binding site for the endogenous integration host factor (IHF) protein (Funnell 1988b). Binding of IHF to this central sequence promotes bending of the DNA in such a way to mediate the recognition of both A- and B-boxes by the CBP ParB, which binds as a dimer and bridges the flanking *parS* regions, bringing them together **(Figure 1.5)** (Hayes & Barillà 2006a, Schumacher 2007). This nucleoprotein complex, formed at *parS* by the interactions of both IHF and ParB with specific DNA sequences, forms an assembly to which the motor protein ParA is recruited to drive plasmid movement throughout the segregation process (Erdmann *et al*. 1999).

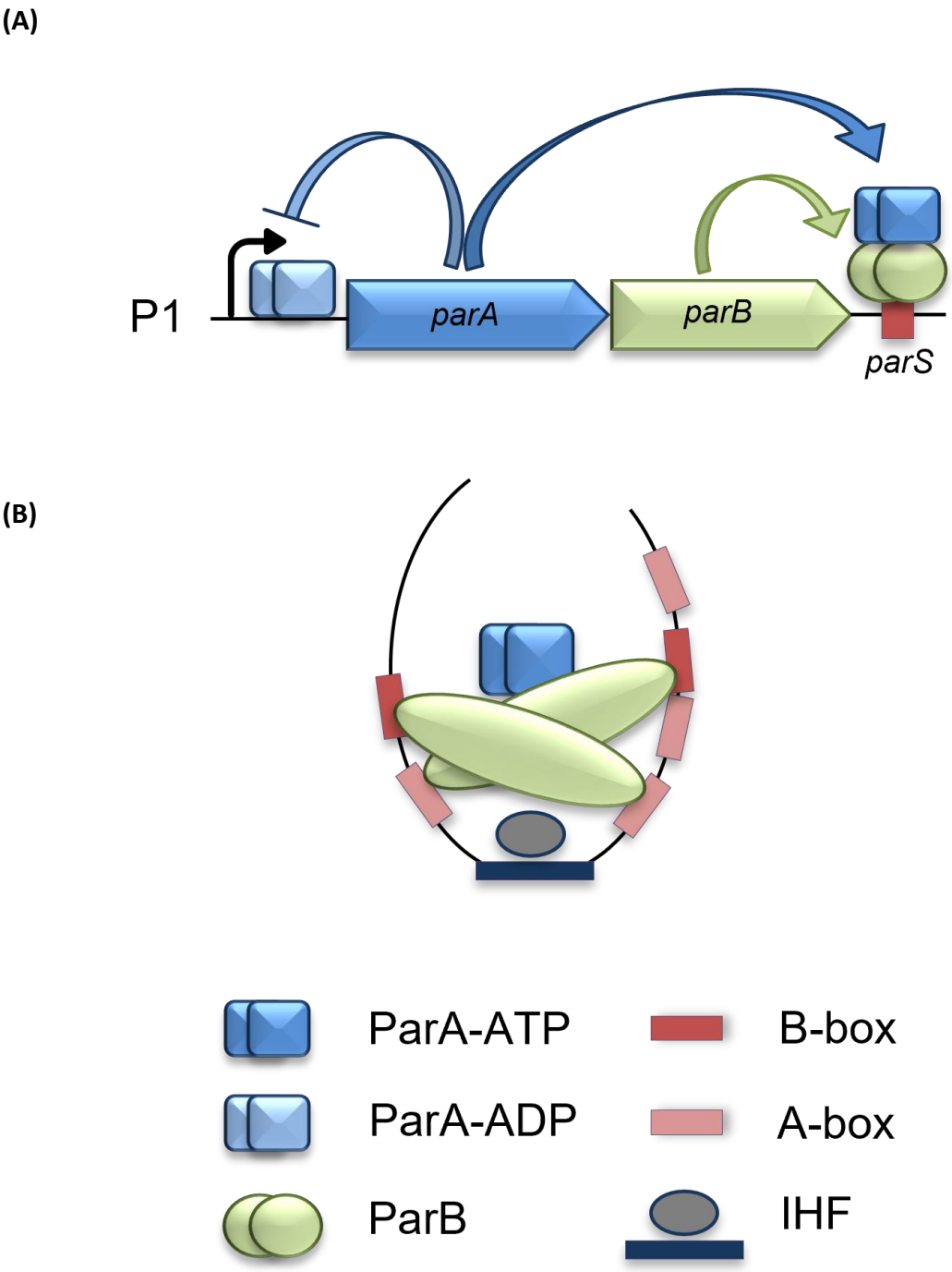


Figure 1.5. ParAB*S* of plasmid P1. **(A)** The partition cassette of plasmid P1 showing the *parB* and *parA* genes, and the downstream *parS* centromere. The ParB protein binds as a dimer to the *parS* repeat motifs. The ADP-bound form of ParA binds in the promoter region to auto-regulate transcription, whereas the ATP-bound form of ParA is recruited into the partition complex to function in plasmid partition. **(B)** Cartoon representation of the partition complex; integration host factor (IHF) binds to the DNA and induces bending, allowing the ParB dimer to contact both DNA arms via interactions with both A- and B-box motifs within *parS*. Colour scheme follows Figs 1.3, 1.4. Adapted from Hayes & Barillà 2006a.

Another extensively studied Type Ia segregation system is that of the *E. coli* F plasmid. Here, the partition system was identified within a ~3 kb segment outside of, but adjacent to the region containing the mini-F plasmid origin of replication, *ori*. (Ogura & Higara 1983). Similar to that of P1, the F plasmid segregation module contains two genes; *sopA* and *sopB*, and a downstream centromeric site, *sopC*, where *sop* is stability of plasmid. SopA is the ATPase motor protein, and SopB the centromere-binding protein. All three components were shown to be necessary for correct plasmid partitioning (Ogura & Higara 1983). The P1 and F operons are of similar size, however one difference is that whilst *sopC* is larger than *parS*, its genetic arrangement is simpler; *sopC* consists of 12 consecutive 43 bp direct repeats with a central 14 bp inverted repeat **(Figure 1.3)** (Mori *et al*. 1986). SopB was demonstrated to bind *in vitro* to each of the inverted repeats within the 12 direct repeats, whilst SopA was shown to recognise and bind to specific sequences within the *sopAB* promoter region, with binding enhanced at some sequences by the addition of SopB (Mori *et al*. 1989). SopA was therefore suggested to perform an autoregulatory role, similar to that of ParA, with its relatively weak intrinsic transcriptional repression activity enhanced by SopB acting as a corepressor (Hirano *et al*. 1998). This is consistent with the observation that ParB, in conjunction with ParA, down-regulated *parAB* expression to a greater extent than its attenuation by ParA alone (Friedman & Austin 1988).

One difference in the formation of the partition complex by SopB-DNA interactions compared with that of ParB, is that no additional factor is required for CBP binding. The SopB protein binds as a dimer to *sopC* without distorting the DNA, facilitated by its HTH DNA-binding domain. The SopB C-terminal dimerisation domain was found to display no DNA-binding activity, unlike that of P1 ParB, however, the central DNA binding domain was also shown to have 'secondary' dimerisation properties, which could act to bring together SopB dimers located on separate DNA duplexes (Schumacher *et al*. 2010). The specific DNA-binding activity of SopB at the *sopC* site, and formation of the partition complex, then acts as a site of recruitment for SopA, with SopB interacting with SopA via the first 45 amino acids of its N-terminal domain (Ravin *et al*. 2003). The specific mechanisms by which the CBPs and motor proteins act in concert to promote plasmid segregation will be discussed in later sections.

1.3.1.2 Type Ib system: *parFGH* of TP228

Plasmid TP228 is a multidrug resistant conjugative plasmid that was first identified in *Salmonella enterica* serovar Newport. After transformation of the plasmid into *E. coli*, TP228 was shown to replicate at low copy number, but with stable maintenance after 25 generations, indicating that this may be due to an active segregation system (Hayes 2000). The stable maintenance of TP228 was shown to be due to a partition system similar to those previously described, comprising two genes and a centromeric site, named *parFGH*. Here, the ATPase motor protein is ParF, which contains a Walker-type motif and thus is a homologous member of the ParA superfamily. The CBP in this system is ParG, which is unrelated to ParB proteins, but which performs an analogous function by binding to a specific region upstream of the *parFG* genes, denoted *parH* **(Figure 1.6)** (Barillà & Hayes 2003).

This upstream centromeric site is in contrast to the downstream location of *parS* and *sopC* relative to the partition genes, in part defining the TP228 partition system as belonging to Type Ib. The *parH* site consists of 12 tetrameric repeats separated by 4 bp AT-rich spacer sequences. In addition, downstream of the putative promoter region, and upstream of the start of the *parFG* genes, another eight tetrameric motifs are present, which were designated as the operator site, O_F **(Figure 1.3, Figure 1.6)** (Zampini *et al*. 2009). The CBP ParG was demonstrated to bind as a dimer to each repeat motif in both the *parH* and O_F sites. ParG acts as a repressor to autoregulate transcription of the *parFG* operon when bound to the operator sites (Carmelo *et al*. 2005).

The role of ParG in plasmid stability was measured by progressive deletion of *parH* motif pairs, resulting in reduced levels of plasmid retention levels. Interestingly, the O_F region alone also displayed the ability to act as a centromere, though in reduced capacity compared to *parH* (Wu *et al*. 2011). ParG binds as a dimer to the specific DNA sequence by way of a ribbon-helix-helix (RHH) C-terminal folded domain. This folded domain, comprising intertwined C-terminal helices from each ParB monomer, also serves as the dimerisation interface of the protein (Golovanov *et al*. 2003). ParG also has an unstructured N-terminal domain, which performs several roles: modulation of DNA

binding at both the *parH* and O_F sites (Carmelo *et al*. 2005, Wu *et al*. 2011), along with interactions with the partner protein ParF to promote plasmid segregation. These ParG-ParF interactions, and a model for the partitioning of TP228, will be discussed in later sections.

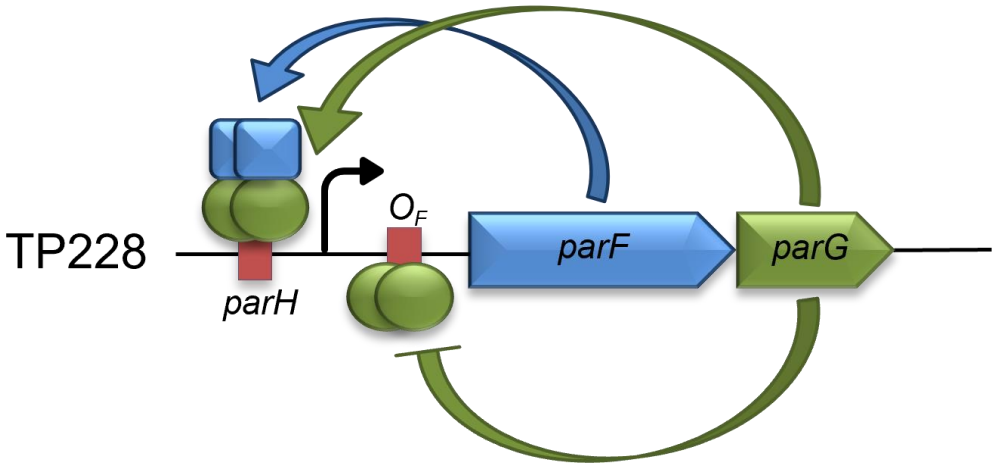


Figure 1.6. ParFG*H* partition system of plasmid TP228. The genetic organisation of the *parFGH* cassette, with genes encoding the CBP ParG, and the ATPase ParF shown in dark green and blue respectively. ParG binds to O_F, the operator region of the promoter to autoregulate transcription, and to *parH* to form the partition complex, where it recruits ParF. Colour scheme follows Figs 1.3, 1.4.

1.3.2 Type II system: *parMRC* of R1 plasmid

Bacterial plasmid segregation systems are categorised according to the type of motor protein they encode. Type I systems utilise a Walker-type ATPase, whilst in Type II systems, the ATPase proteins contain an actin-like fold (Schumacher 2008). The Type II system found on the antibiotic resistant plasmid R1 from *E. coli* is a well-studied example, in which the *parMRC* locus comprises the actin homologue ParM, the DNA-binding protein ParR, and the upstream centromere-like region *parC* (Sengupta & Austin 2011). The genetic organisation of the locus, the relative size of the two proteins, and their respective functions, is similar to that of Type Ib *par* systems **(Figure 1.7A)**. The upstream centromere-like *parC* site was found to be composed of ten 11 bp direct repeats, organised into two equal clusters of five **(Figure 1.3)**. The two clusters are discontinuous, being separated by a region of 39 bp, in which the *parMR* promoter lies. The property of transcriptional repression of the locus was mapped to ParR, whilst the actin-like ParM was not implicated in repression (Jensen *et al*. 1994). The contribution of both the promoter element and the direct repeats was assessed by progressive deletion of the repeats, with plasmid stability decreasing concomitantly, whilst replacement of the promoter did not negatively affect stability (Breüner *et al*. 1996).

ParR was also found to exhibit specific DNA-binding activity i*n vitro* towards the *parC* site via electron microscopy studies, here functioning to pair together two separate DNA molecules, thus demonstrating a role in plasmid segregation (Jensen *et al*. 1998). The physical interaction of ParR at *parC* was shown to be mediated by the formation of a dimeric RHH structure at the N-termini of the protein, based on structural analysis of a ParR homologue from *E. coli* plasmid pB171 (Møller-Jensen *et al*. 2007). ParR dimers bind to the *parC* repeat sequences in a cooperative fashion, causing the DNA to bend into a U-shaped structure that wraps around ParR molecules to form a ring **(Figure 1.7B)** (Møller-Jensen *et al*. 2007, Hioschen *et al*. 2008, Saljie & Löwe 2008). Interactions with the partner ATPase ParM are mediated by the C-terminus of ParR, which binds to ParM filaments (Saljie & Löwe 2008), whose polymerisation dynamics act to drive plasmid movement (Møller-Jensen *et al*. 2003).

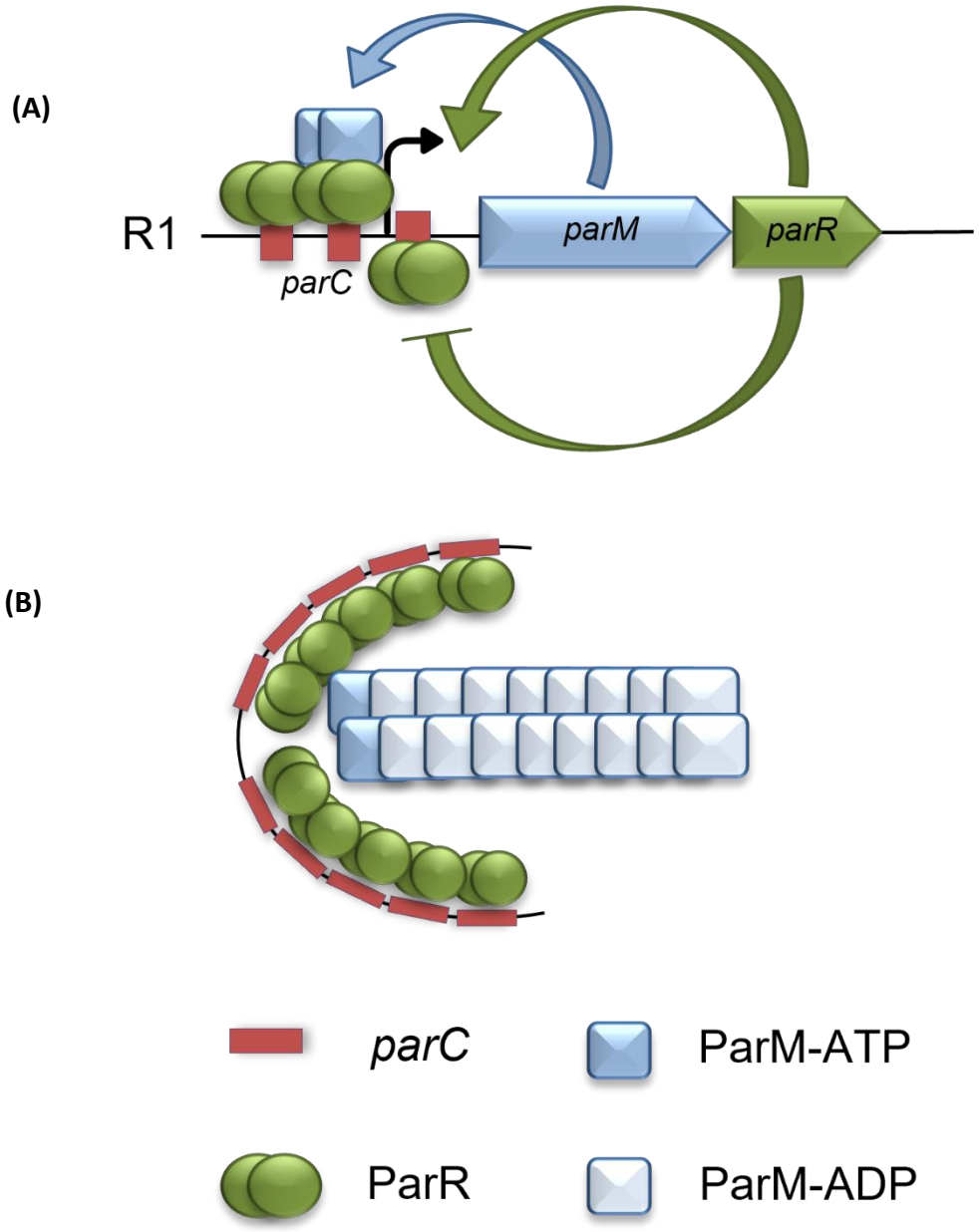


Figure 1.7. Par MRC of plasmid R1. **(A)** The genetic organisation of the *parMRC* cassette. The CBP ParR and the ATPase ParM are coloured dark green and light blue respectively. ParR binds to the operator region of the promoter to autoregulate transcription, and to *parC* for plasmid segregation. ParR recruits ParM into the partition complex. Not all *parC* repeats are shown. **(B)** The model for *parMRC* segrosome assembly. ParR binding to the *parC* sites bends the DNA into a ring-like structure. Filaments are formed from ADP-bound ParM, whereas ATP-bound ParM is inserted at the ParR interface. Colour scheme follows Figs 1.3, 1.4.

1.3.3 Type III and Type IV systems

Type III and Type IV partition systems are those most recently described. Type III systems are distinct from those of Types I and II as their motor protein is a homologue of tubulin/FtsZ, containing neither Walker-type motifs or actin-like folds (Schumacher 2008). Type III systems are similar to those previously described in that they comprise two genes and a centromere-like site, however the genetic organisation differs, as the CBP-encoding gene comes before that of the motor protein in the operon **(Figure 1.4)**. A Type III system was identified on the pBtoxis virulence plasmid, harboured by *Bacillus thuringiensis*, comprising the GTPase motor TubZ, the DNA-binding protein TubR, and the centromere *tubC*, located upstream of the *tubZR* genes (Tang *et al*. 2006, Ni *et al.* 2010). The *tubC* centromere was originally found to contain 4 sets of 12 bp direct repeats (Tang *et al*. 2006). However later analysis extended the centromeric region to comprise two clusters of three and four repeats respectively, separated by 54 bp **(Figure 1.3)** (Aylett & Löwe 2012). TubR was found to bind to repeats as a dimer, and although the protein contains a N-terminal winged HTH motif, the DNA-interaction mechanism was found to be distinct from other HTH proteins, with recognition helices and wings able to insert into adjacent major and minor grooves respectively (Ni *et al*. 2010). TubR, similar to other partitioning proteins, was demonstrated to negatively regulate expression of *tubZ* (Larsen *et al*. 2007). The GTPase TubZ binds to TubR-DNA via its flexible C-terminus, and has the ability to form filamentous polymers in a GTP-dependant manner, that act dynamically to transport plasmid-bound cargo within the cell in a process dubbed treadmilling (Larsen *et al*. 2007, Ni *et al.* 2010, Barillà 2010). This mechanism will be described in a later section.

Lastly, there are a few examples of potentially new partition systems distinct from Types I, II and III. The first, found on the *Staphylococcus aureus* plasmid pSK1, is unusual as it comprises a single gene, rather than two. The gene, called *par*, was shown to mediate pSK1 stability, implying that it plays a role in active partitioning (Simpson *et al*. 2003). The candidate centromere-like site is located upstream of *par*, and comprises seven direct repeats plus one inverted repeat that could represent binding sites for the Par protein **(Figure 1.3)**. The N-terminal domain was shown to harbour a HTH DNA-binding motif, whilst the region towards the C-terminus was predicted to form a

coiled-coil, suggesting that the protein may be able to form oligomers (Firth *et al*. 2000, Simpson *et al*. 2003). The Par protein was found to lack an ATP-binding motif, therefore it is unknown whether Par has dual functionality as both CBP and motor protein in this system (Dmowski & Jagura-Burdzy 2013).

A second plasmid partition system that only requires one protein for stable maintenance was found encoded on the broad host-range plasmid R388. Here, an operon containing three genes is present, but only the first, *stbA*, is required for plasmid stability, whilst the second, *stbB* was implicated in plasmid conjugation (Guynet *et al*. 2011). The centromere-like site is located upstream of the operon, and comprises two sets of five 9 bp repeats, to which StbA bound site-specifically *in vitro*. Interestingly, StbB harbours Walker-type motifs and is thus a putative motor protein, however StbB was not required for plasmid stability. It is conjectured by the authors that the segregation of R388 either requires only one protein, as described above for pSK1, or that it requires an endogenous motor protein supplied by the host (Guynet *et al*. 2011, Guynet & De la Cruz 2011). Both the pSK1 and pR388 segregation apparatus have been described in the literature as potential Type IV partition systems. Selected plasmid- and chromosomally-encoded partition systems found in bacteria and archaea are summarised below in Table 1.1.

Table 1.1 Summary of bacterial and archaeal partition systems

Partition Type - location	Name	Organism	Reference
Type Ia – plasmid P1	*parABS*	*E. coli*	Austin & Abeles 1983
Type Ia – plasmid F	*sopABC*	*E. coli*	Ogura & Higara 1983
Type Ib – plasmid TP228	*parFGH*	*Salmonella* Newport	Barillà & Hayes 2003
Type II – plasmid R1	*parMRC*	*E. coli*	Jensen *et al*. 1994
Type III – plasmid pBtoxis	*tubZRC*	*B. thuringiensis*	Tang *et al.* 2006
Type IV – plasmid pSK1	*par*	*S. aureus*	Simpson *et al.* 2003
Chromosome	*parAB*	*C. crescentus*	Mohl & Gober 1997
Chromosome	*soj-spo0J*	*B. subtilis*	Ireton *et al.* 1994
Chromosome	*segAB*	*S. solfataricus*	Kalliomaa-Sanford *et al.* 2012
Plasmid pNOB8	*aspA-parB-parA*	*Sulfolobus* NOB8-H2	Schumacher *et al*. 2015

1.3.4 Chromosomal Par systems

So far, a variety of bacterial plasmid partition systems have been introduced. However similar active DNA segregation systems are also found encoded on chromosomes. Interestingly, systems based on *parABS* or plasmid Par homologues are not found in *E. coli*, although they are widespread in over 60% of bacterial taxa (Livny *et al*. 2007, Badrinarayanan *et al*. 2015). Most studies involving chromosomal *par* loci systems have focussed on both Gram-positive and Gram-negative bacteria; *Bacillus subtilis*, *Caulobacter crescentus*, *Pseudomonas* spp., and *Vibrio cholerae* (Schumacher 2008). Chromosomal *par* systems appear to be an amalgam of plasmid Type Ia and Ib systems, as the ParA motor proteins are the smaller Walker-types of Type Ib, and the ParB proteins harbour HTH DNA-binding motifs similar to Type Ia CBPs (Schumacher 2008). Chromosomal *parS* sites are usually found proximal to the origin of replication (*oriC*), and similar to those found on plasmids, may exist as multiple iterations of a repeat sequence. The *parS* site of the *B. subtilis* chromosome occurs ten times across a region 20% proximal to *oriC*, whilst in *C. crescentus*, seven *parS* sites were found clustered within 8 kb of the origin (Lin & Grossman 1998, Tran *et al*. 2018).

The identification of Par homologues encoded on bacterial chromosomes came initially from studies involving *B. subtilis* proteins involved in sporulation. One of these, Spo0J, was found to be a ParB homologue, whilst Soj, the product of the gene upstream of *spo0J*, was similar to ParA family members (Ireton *et al*. 1994). Spo0J was not only involved in regulation of *B. subtilis* sporulation, but was also required for correct chromosome partitioning in the vegetative growth phase, and was demonstrated to bind site-specifically to the DNA at its cognate *parS*-like sites (Ireton *et al*. 1994, Lin & Grossman 1998). Soj was originally thought unnecessary for segregation, as its inactivation did not result in obvious segregation defects (Lin & Grossman 1998). However, later work uncovered distinct roles for Soj, both in DNA replication, and also when working in conjunction with Spo0J and another host factor, SMC (Structural Maintenance of Chromosomes) to contribute to origin segregation (Lee & Grossman 2006, Murray & Errington 2008, Wang *et al*. 2014). The structure of *Thermus thermophilus* Spo0J revealed a HTH DNA-binding domain, and both a C-terminal primary

dimerisation domain and secondary N-terminal dimerisation domain, which positions the HTH motifs from each monomer to allow binding to *parS* (Leornard *et al*. 2004).

The ParAB system of *C. crescentus* has also been extensively explored, with the *parAB* locus originally identified within 80 kb of the origin of replication. Unlike *spo0J* and *soj* in *B. subtilis*, *parA* and *parB* of *C. crescentus* were found to be essential for cell viability (Mohl & Gober 1997). ParB was found to exhibit high specificity when interacting with the cognate binding site *parS* immediately downstream of *parAB*. ParB was implicated in chromosome segregation as it was shown to localise initially to a single cell pole, then to both cell poles, in a manner approximating the movement of replicated chromosomes during the cell cycle. In addition, overexpression of both *parA* and *parB* caused defects in chromosome partitioning (Mohl & Gober 1997). *C. crescentus* ParB was later found to bind with differing degrees of affinity to several *parS* sites close to the origin, with the protein capable of spreading up to 10 kb from the initial *parS* nucleation site onto the adjacent DNA to form extended nucleoprotein complexes (Tran *et al*. 2018). This spreading phenomenon of CBPs, often from multiple *parS* sites, has been observed in many other species, including *B. subtilis*, *Pseudomonas aeruginosa* and *Vibrio cholerae*, and is deemed to be a general feature of ParB proteins (Bartosik *et* al. 2004, Breir & Grossman 2007, Kusiak *et al*. 2011, Baek 2014, Graham *et al*, 2014). This finding was not unexpected, as plasmid-encoded ParB proteins from P1 and F plasmids had also displayed the ability to spread from centromeric sites to flanking DNA (Lynch & Wang 1995, Rodionov *et al*. 1999). The further implications of the spreading of ParB along the DNA to form an extended partition complex, and the various mechanisms by which the CBP and ParA motor proteins interact to drive genome segregation, will be discussed in later sections.

1.4 DNA Partition proteins

1.4.1 Centromere-Binding Proteins (CBPs)

Active partitioning systems are responsible for the segregation of newly replicated DNA molecules, and these systems are encoded on low copy number plasmids as well as chromosomes (Badrinarayanan 2015, Bouet & Funnell 2019). In the majority of cases, two proteins, along with the centromere-like DNA sequence, comprise active *par* systems. The two proteins are the centromere-binding protein (CBP), which is often called ParB (or ParB-like), and the NTPase motor protein, ParA (or ParA-like). The initial step in the partitioning process involves the CBP recognising, and subsequently binding to its cognate DNA sequence, the *parS* site (Schumacher 2008). The formation of this nucleoprotein complex serves as a scaffold for further recruitment of the ParA motor protein, followed by the dynamic transport of the replicated DNA within the cell volume. The CBP therefore performs multiple roles during segregation: binding of multiple CBPs to DNA to form higher-order partition complexes, the pairing of plasmids or distal chromosomal DNA, binding to the motor protein via protein-protein interactions and stimulation of the NTPase activity of the motor protein to drive segregation, and in some cases transcriptional regulation (Funnell 2005, Ringgaard *et al.* 2007, Schumacher 2008, Oliva 2016, Bouet & Funnell 2019).

CBPs have been classified according to their structure and DNA-binding motif, and comprise two main groups: helix-turn-helix (HTH) and ribbon-helix-helix (RHH) containing proteins (Bouet & Funnell 2019). HTH (also denoted as HTH_2 for the dimer form) CBPs are found in Type Ia plasmid partition systems, and all chromosomal *par* system studied to date. The level of sequence conservation amongst the HTH CBPs is low, however they all share similar domain arrangements: a central HTH DNA-binding domain, a C-terminal dimerisation domain, and a flexible N-terminal region which both interacts with the cognate NTPase and mediates oligomerisation of the CBP (Baxter & Funnell 2014, Bouet & Funnell 2019). The existence of the central HTH domain common to these CBPs results in high structural conservation at this region (Oliva 2016). The three domains are separated by flexible linker regions, and in the case of ParB from *E. coli* plasmid P1, this flexibility permits the rotational movement of individual domains, allowing the protein to

bind to a range of A- and B-box motifs within the *parS* centromere via both HTH and dimerisation domains **(Figure 1.8A)** (Schumacher & Funnell 2005). The structures of the HTH domains of several plasmid-encoded and chromosomal ParBs in complex with their cognate DNA binding sites have been solved, and demonstrate that there are large similarities, but also subtle differences in how the proteins interact with DNA. For P1 ParB, the *parS* A-box is bound by the recognition helix within the HTH motif only, whereas SopB of F plasmid utilises the recognition helix plus additional residues that lie outside of it (Schumacher *et al*. 2010, Bouet & Funnell 2019). Chromosomal ParBs display similar binding patterns to their plasmid counterparts. Spo0J (ParB) of *Helicobacter pylori* binds in a similar manner to that of SopB, with additional residue-base interactions that may reflect species-specific recognition of *parS* motifs (Chen *et al*. 2015). This was also seen in ParB of *C. crescentus,* where additional helices separate from the main recognition helix of the HTH domain contributed to interactions at the protein-DNA interface (Jalal *et al*. 2020a). Interestingly, *B. subtilis* Spo0J (ParB) was found to not only bind *parS* DNA via its central HTH domain, but in addition, to bind non-specific DNA via a lysine-rich area located in the C-terminal dimerisation domain (Fisher *et al*. 2017). This additional binding interface was also shown to be required for condensing the DNA and thus facilitating genomic segregation. This property of ParB will be discussed in a later section.

The N-terminal domains of ParB proteins are generally flexible, and also fulfil several functional roles, both in forming higher order oligomeric complexes with other ParBs, and in interactions with the partner NTPase (Oliva 2016, Bouet & Funnell 2019). The crystal structure of *Helicobacter pylori* Spo0J (ParB) demonstrated that protein-protein interactions were facilitated by two conserved motifs (one harbouring an arginine-rich patch) within the N-terminus. Additionally, binding to the DNA induces a conformational change that favours N-terminal interactions both with neighbouring Spo0J (ParB) molecules adjacent on the same stretch of DNA, and to Spo0J (ParB) on more distal sections of DNA **(Figure 1.8B)** (Chen *et al*. 2015, Oliva 2016). The arginine patch motif is highly conserved amongst plasmid and chromosomal ParBs, and mutations in this region have previously demonstrated the motif to be required for the higher-order complex formation of ParB from F Plasmid, and in *P. aeruginosa* and *B. subtilis* (Yamaichi & Niki 2000, Kusiak *et al*. 2011, Graham *et al*. 2014, Debaugny *et al*. 2018).

In addition to mediating CBP oligomerisation, the ParA interaction domain has also been mapped to the extreme N-terminus of ParB proteins. This region not only contacts the NTPase, but also stimulates ParA nucleotide hydrolysis, with a crucial arginine residue implicated in some cases (Baxter & Funnell 2014). The N-terminal 45 residues of SopB, in particular Arg-36 providing the arginine-finger motif, and 20 amino acids in the N-terminus of Spo0J (ParB), with mutation of Arg-10 abrogating ATPase stimulation, were shown to be important for ParB-ParA interactions (Ravin *et al*. 2003, Leonard *et al*. 2005, Ah-Seng *et al*. 2009). However, the NTPase interaction domain is not always found in the N-terminus of the CBP: ParB of *P. aeruginosa* interacts with its cognate ParA via its C-terminal dimerisation domain (Bartosik *et al*. 2004).

RHH (also denoted as RHH_2 for the dimer form) CBPs are found in Type Ib and Type II segregation systems. The RHH motif was first identified in the Arc and MetJ transcriptional repressors, whose solved crystal structures in complex with DNA revealed that specific DNA base contacts were mediated by antiparallel β-sheets, rather than α-helices (Schreiter & Drennan 2007). The RHH CBPs, which tend to be smaller in size than their HTH counterparts, also act as transcriptional repressors. The difference between Type Ib and Type II RHH CBPs lies in their domain organisation: Type Ib CBPs interact with DNA and ATPases via their C-terminal and N-terminal domains respectively, whilst in Type II CBPS this is the other way around (Baxter & Funnell 2014).

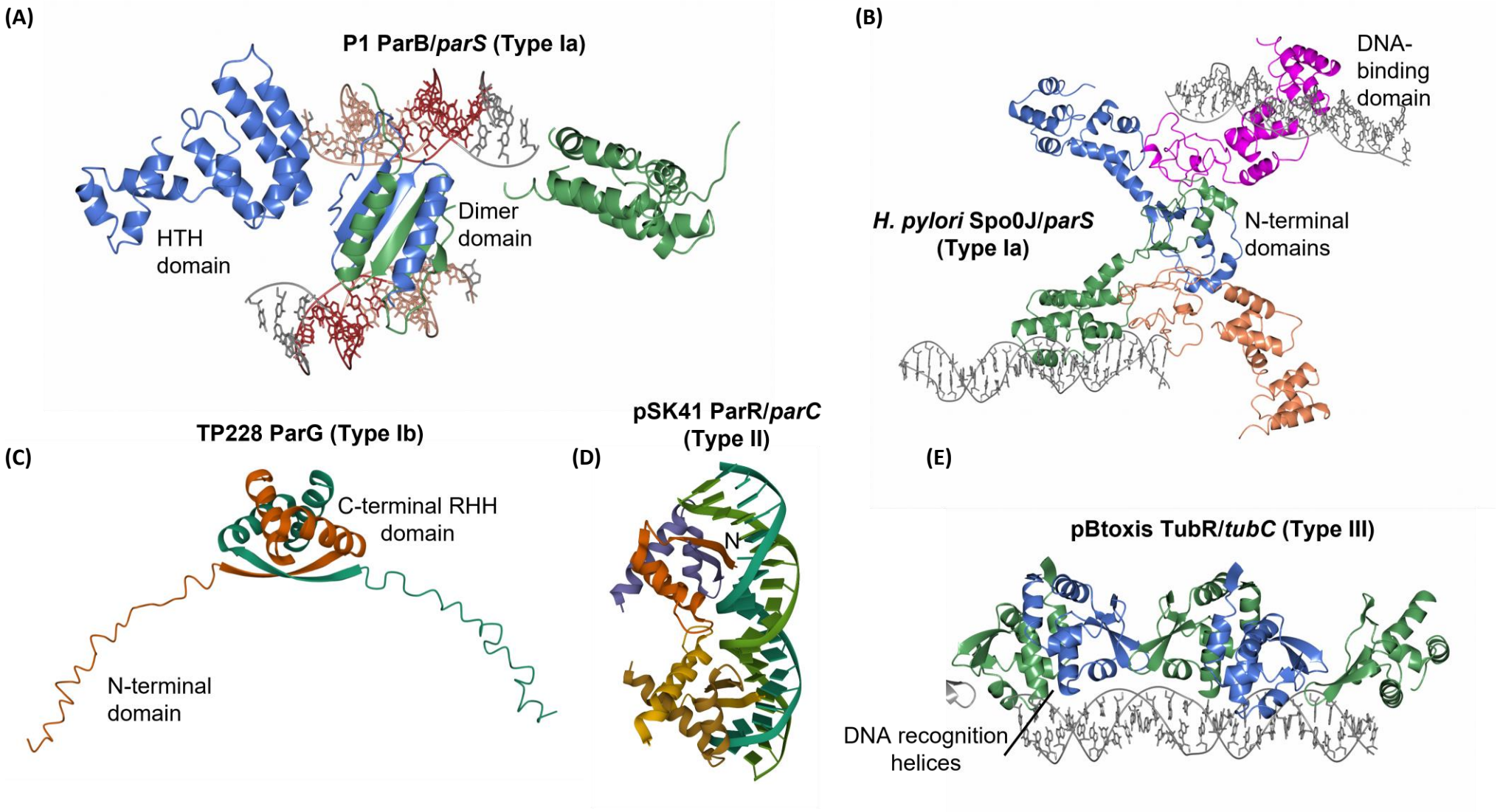
An example of a Type Ib RHH CBP is that of ParG, encoded on plasmid TP228 harboured by *Salmonella* Newport. ParG consists of a highly flexible N-terminal tail, and a C-terminal DNA-binding domain, and in the ParG dimer, the C-termini intertwine to form the RHH fold which contacts the DNA **(Figure 1.8C)** (Golovanov *et al*. 2003). The flexible N-termini extensions perform multiple roles: modulation of transcriptional repression by establishing higher-order complexes on the DNA via contacts with the folded RHH domain, promoting the oligomerisation of the partner ParF, and enhancing the nucleotide hydrolysis of ParF via an arginine-finger motif (Carmelo *et al*. 2005, Barillà *et al*. 2007, Wu *et al*. 2011).

Structural data for a Type II RHH CBP is available for ParR from *E. coli* pB171. Here, the two ParR monomers, each consisting of an N-terminal β-strand followed by five α-helices,

form a tight antiparallel homodimer, with the RHH_2 fold formed by the dimer N-termini (Møller-Jensen *et al*. 2007). The ParR dimers assemble cooperatively into a helical structure, stabilised by inter-dimer contacts, with the N-termini RHH_2 DNA-binding domains facing outwards to form basic patches, whilst the C-termini point inwards to the helix centre (Møller-Jensen *et al*. 2007). Another ParR, from pSK41, was solved in complex with 20-mer centromeric DNA, and again showed a super-helical array formed from assembled dimer-of-dimers **(Figure 1.8D)**. The N-terminal RHH_2 folds of ParR dimers form an electropositive surface that interacts with the DNA, whilst the C-termini act to promote higher-order assembly on the DNA, along with mediating contact with the cognate ATPase, ParM (Schumacher *et al*. 2007, Baxter & Funnell 2014).

Type III CBPs harbour a HTH fold as their DNA-binding interface, although the method of binding differs from canonical HTH proteins, as the recognition helix is buried in the dimer core. The N-termini of the recognition helices of the dimer were proposed to insert into a single major groove of the DNA **(Figure 1.8E)** (Ni *et al*. 2010, Aylett & Löwe 2012). The formation of the TubR-DNA superstructure is reminiscent of that observed with ParR, as some TubR proteins spread and form ring-like protein-DNA filaments, interacting in turn to polymerise the cognate motor protein TubZ (Aylett & Löwe 2012, Martin-Garcia *et al*. 2018). Lastly, the Par protein of putative Type IV segregation from plasmid pSK1 contains an N-terminal HTH motif, half of which is conserved and confers non-specific DNA-binding, whilst the second half of the HTH motif is not conserved, and the exact mode of how Par interacts with the DNA remains unclear (Simpson *et al*. 2003, Dmowski & Jagura-Burdzy 2013).

Figure 1.8. Centromere-binding proteins (following page). Crystal structures of various CBPs. **(A)** ParB of plasmid P1. Two ParB monomers are shown in blue and green. The dimerisation domain is shown here bridging between two DNA B-boxes. The HTH domains also contact DNA A-boxes (not shown). A- and B-boxes are coloured as in Fig. 1.3. (PDB code 2NTZ). **(B)** The chromosomal ParB homologue Spo0J from *Helicobacter pylori*. Four monomers are shown bound to DNA. *ParS* sites are bound by the HTH DNA-binding domain, however N-terminal domain adjacent interactions allow spreading, and transverse interactions mediate higher-order complex formation (PDB code 4UMK). **(C)** ParG of plasmid TP228. The ParG dimer is shown, with each monomer in green and orange. The C-terminal RHH dimerisation and DNA-binding domain, and the flexible N-terminal domain are labelled (PDB code 1P94). **(D)** ParR of pSK41. Two ParR dimers are shown, with dimer subunits coloured purple/orange and yellow/brown. The N-terminal RHH DNA-binding domain is labelled N. The DNA is coloured green (PDB code 2Q2K). **(E)** TubR of pBtoxis bound to *tubC*. The TubR dimers are shown with each subunit coloured blue and green. The N-termini of paired recognition helices from one dimer which insert into a DNA major groove are labelled (PDB code 4ASS). All structures were generated using CCP4MG (McNicholas *et al*. 2011), except **(C)** and **(D)**, which used the Protein Data Bank website interface.


(A)
P1 ParB/*parS* (Type Ia)
HTH domain
Dimer domain
(B)
DNA-binding domain
H. pylori Spo0J/*parS* (Type Ia)
N-terminal domains
(C)
TP228 ParG (Type Ib)
C-terminal RHH domain
N-terminal domain
(D)
pSK41 ParR/*parC* (Type II)
N
(E)
pBtoxis TubR/*tubC* (Type III)
DNA recognition helices

1.4.2 NTPase motor proteins

The second functional component in DNA partition systems is the motor protein, often called ParA. The recruitment of ParA into the CBP-DNA complex mediates the next step in segregation, driving the separation of replicated DNA molecules (Schumacher 2008). The different segregation systems are distinguished by the signatures of their motor proteins, and these ATPases or GTPases are the proteins that provide the required energy for DNA movement (Bouet & Funnell 2019). The most common type of NTPases are the Walker-type ParA proteins found in Type I systems, and these proteins have several properties: including adenine nucleotide-influenced dimerisation, ATP binding and hydrolysis, ATP-dependent binding to non-specific DNA, and the ability to form higher-order structures (Baxter & Funnell 2014). ParA of plasmid P1 was found to exist as a dimer in physiologically relevant conditions, and interestingly, binding to different adenine nucleotides induced conformational changes that affected function. In the ATP-bound state, ParA mediates plasmid partition, whereas the ADP-bound protein acts as a transcriptional repressor at the *par* operator site **(Figure 1.9A)** (Bouet & Funnell 1999, Dunham *et al*. 2009). The ATP-dependent binding of ParA to the bacterial nucleoid, in complex with plasmid-bound ParB, has led to proposed models of plasmid segregation, which will be discussed in a later sections (Vecchiarelli *et al*. 2013b).

ParA proteins are known to have low inherent ATPase activity, that is enhanced to varying degrees by combinations of specific and non-specific DNA, and the cognate CBP. Experiments with SopA of F Plasmid demonstrated that its ATPase activity was greatly enhanced in the presence of non-specific DNA and SopB, however the activity was increased three-fold when specific DNA (the *sopC* centromere) was used (Ah-Seng *et al*. 2009). SopA was also demonstrated to polymerise into filaments *in vitro* in the presence of ATP, but not ADP (Bouet *et al.* 2007). Chromosomal ParA homologues exhibit similar properties: Soj of *T. thermophilus* was shown to dimerise and bind to non-specific DNA in an ATP-dependent fashion, form polymers in the presence of ATP and DNA, and its maximal levels of ATP hydrolysis occurred in the presence of its CBP Spo0J (ParB) and *parS* DNA (Leonard *et al*. 2005), and *C. crescentus* ParA showed similar behaviours (Ptacin *et al*. 2010, Lim *et al*. 2014).

The properties of the Type Ib motor protein ParF, encoded on plasmid TP228, have also been elucidated. ParF is a member of the wider superfamily of ParA ATPases, and similar to those described above, ParF was shown to dimerise when bound to ATP, and to form filaments *in vitro* in a manner stimulated by ATP, but inhibited by ADP **(Figure 1.9B)** (Barillà *et al.* 2005, Schumacher *et al.* 2012). Its cognate CBP, ParG, demonstrated the ability to stimulate the ATPase activity of ParF via a single Arg-19 amino acid acting as an arginine finger-like residue (Barillà *et al.* 2007), as is seen in CBP-ATPase interactions in Type Ia systems (Ravin *et al*. 2003). Thus, although Type Ib motor proteins do not regulate transcription like their Type Ia counterparts, there are many similarities in their roles in segregation.

A well-studied example of a Type II (actin-like) motor protein is that of ParM from plasmid R1, which also does not function in transcription repression. Interestingly, although ParM was initially demonstrated to bind ATP, it was subsequently shown to also bind to GTP, and form polymers in the presence of either nucleotide, although with greater affinity to ATP than GTP **(Figure 1.9C)** (Galkin *et al*. 2009, Rivera *et al*. 2011). The CBP ParR slightly augmented ParM ATPase activity, and the addition of *parC* greatly increased nucleotide hydrolysis, whilst ParM filamentation *in vivo* required the ParR-*parC* complex to act as a nucleation point for polymerisation (Jensen & Gerdes 1997, Møller-Jensen *et al*. 2002).

The tubulin like motor proteins, such as pBtoxis TubZ, have GTPase activity, and also assemble into polymers *in vivo*, though unlike ParM, TubZ does not require its cognate CBP TubR to form filaments **(Figure 1.9D)** (Larsen *et al*. 2007). The assembly and disassembly dynamics of the TubZ filament also differ from that of ParM: ParM filaments can polymerise and depolymerise from both ends, whereas TubZ filaments are polar as they grow from one end and disassemble at the other, in a process called treadmilling (Aylett *et al*. 2010).

Various models have been put forth to describe how the CBP and motor proteins interact to drive DNA segregation in Type I, II, III, and chromosomally-encoded *par* systems, and these will be enumerated in the following sections.

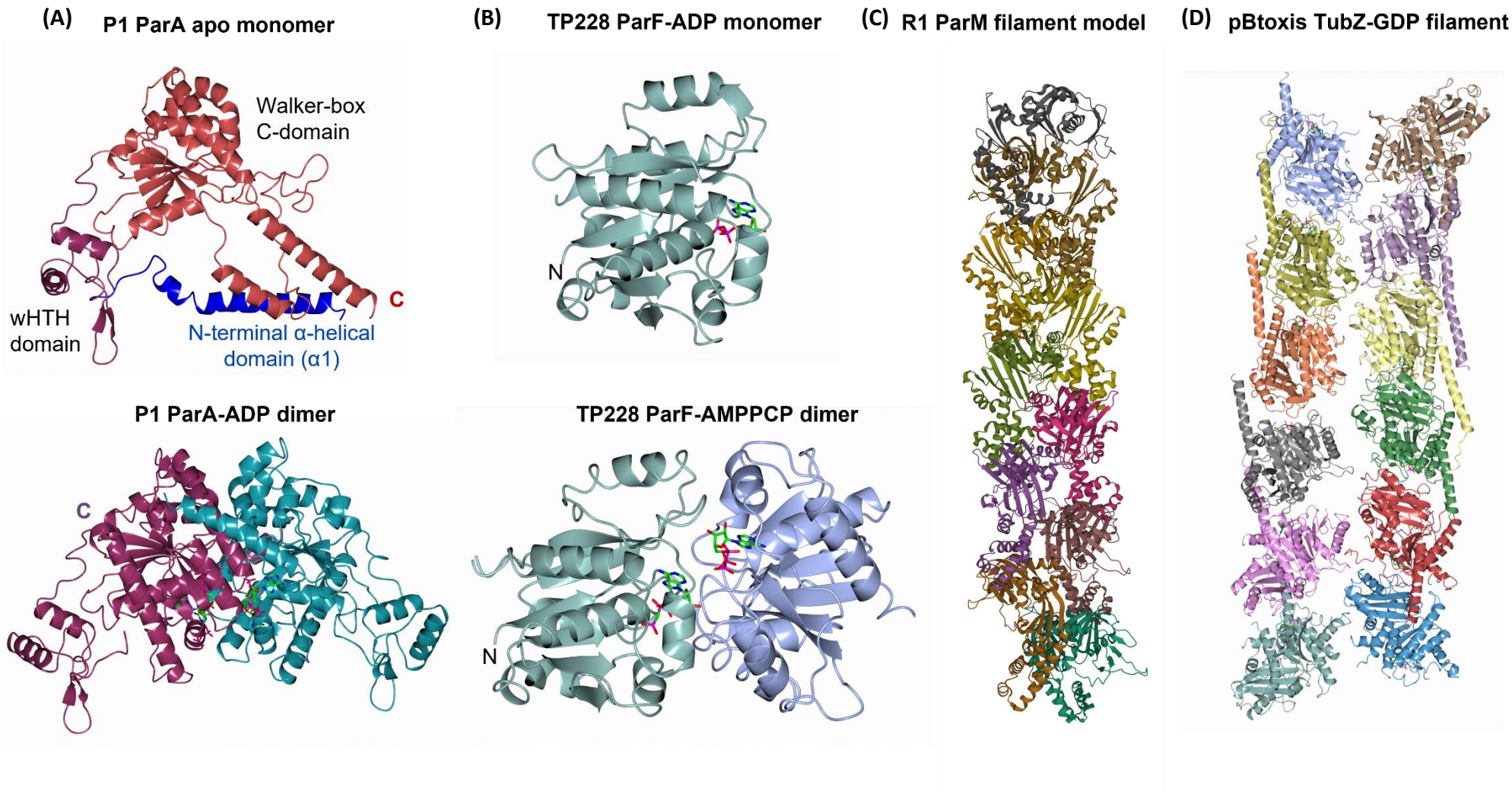

(A)
P1 ParA apo monomer
Walker-box
C-domain
C
wHTH
domain
N-terminal α-helical
domain (α1)
(B)
TP228 ParF-ADP monomer
N
(C)
R1 ParM filament model
(D)
pBtoxis TubZ-GDP filament
P1 ParA-ADP dimer
C
TP228 ParF-AMPPCP dimer
N

Figure 1.9. NTPase motor proteins (previous page). Crystal structures of various partition motor proteins. **(A)** ParA from P1 plasmid. (Top) The monomeric apo-form. The three domains; N-terminal domain, wHTH domain, and Walker-box C-terminal domain are shown in blue, magenta and red respectively. (Bottom) The ParA dimer in the ADP-bound form, which acts to regulate transcription. The two monomers are shown in dark purple and dark cyan. The ADP molecules are shown at the monomer-monomer interface as cylinders. The dark purple monomer is in the same orientation as the apo-form to demonstrate the change of conformation when binding ADP (note the position of the C-terminal helix). The ATP-bound form is not shown (PDB codes 3EZ7, 3EZ2). **(B)** ParF from plasmid TP228. The monomeric ADP-bound form (Top), and the ParF dimer bound to the non-hydrolysable ATP analogue, AMPPCP, are shown (Bottom). Nucleotides are shown as cylinders (PDB codes 4E03, 4E07). **(C)** The R1 plasmid ParM filament model. Ten ParM subunits are shown in different colours, and are bound together to form a linear polymer (PDB code 2QU4). **(D)** The TubZ filament, from the pBtoxis plasmid partition system. The twelve TubZ subunits are shown in different colours, and here are in the GDP-bound form, with nucleotides shown as cylinders (PDB code 2XKB). All structures were generated using CCP4MG except **(C),** which used the Protein Data Bank website interface.

1.5 Mechanisms of plasmid segregation

1.5.1 Type I push/pull mechanism

One of the first models to suggest a mechanism to drive segregation of newly replicated plasmid molecules came from observations of filament formation by the ATPase motor proteins of Type I systems (Baxter & Funnell 2014). Polymerisation has been observed for motor proteins ParF, Soj, SopA and ParA of plasmid pB171, in an ATP-dependant manner (Barillà *et al*. 2005, Leonard *et al*. 2005, Bouet *et al*. Ringgaard *et al*. 2009). These polymerisation and depolymerisation dynamics were also shown to be correlated with the transit of DNA within the cell volume (Marston & Errington 1999, Ebersbach & Gerdes 2004). The initial model suggested that ParA filaments alone were enough to move plasmids to opposite cell poles, via their association with their cognate CBPs bound to the centromere, and that replicated plasmids could potentially be either pulled or pushed within the cell (Barillà *et al*. 2005, Hayes & Barillà 2006).

The involvement of the nucleoid, and ParA associations with it, led to a modification of the model, suggesting that the ParA filaments in some cases polymerised on non-specific DNA (Leonard *et al.* 2005, Baxter & Funnell 2014). However, in some systems, e.g. F plasmid, SopA was found to polymerise in the presence of ATP without the requirement of DNA, and in fact that both specific and non-specific DNA inhibited SopA polymerisation (Bouet *et al.* 2007). Nevertheless, Ringgaard and colleagues proposed a model based on the assembly and disassembly of ParA filaments, in conjunction with chromosomal non-specific DNA, to transit plasmid cargo within the cell (Ringgaard *et al*. 2009). Once the plasmid is replicated, the CBP binds to the centromere. ParA dimers in the ATP bound form bind to non-specific DNA in a cooperative fashion, forming a polymeric filament whose tip associates with the partition complex. The stimulation of the ATPase activity of the ParA terminal subunit by the CBP causes hydrolysis to its ADP-bound form, which subsequently dissociates from the filament, resulting in polymer shortening. The depolymerisation of ParA filaments will ultimately release the ParB-bound plasmid, where it can interact with newly-formed ParA polymers which could move the plasmid in the opposite direction **(Figure 1.10)** (Ringgaard *et al*. 2009).

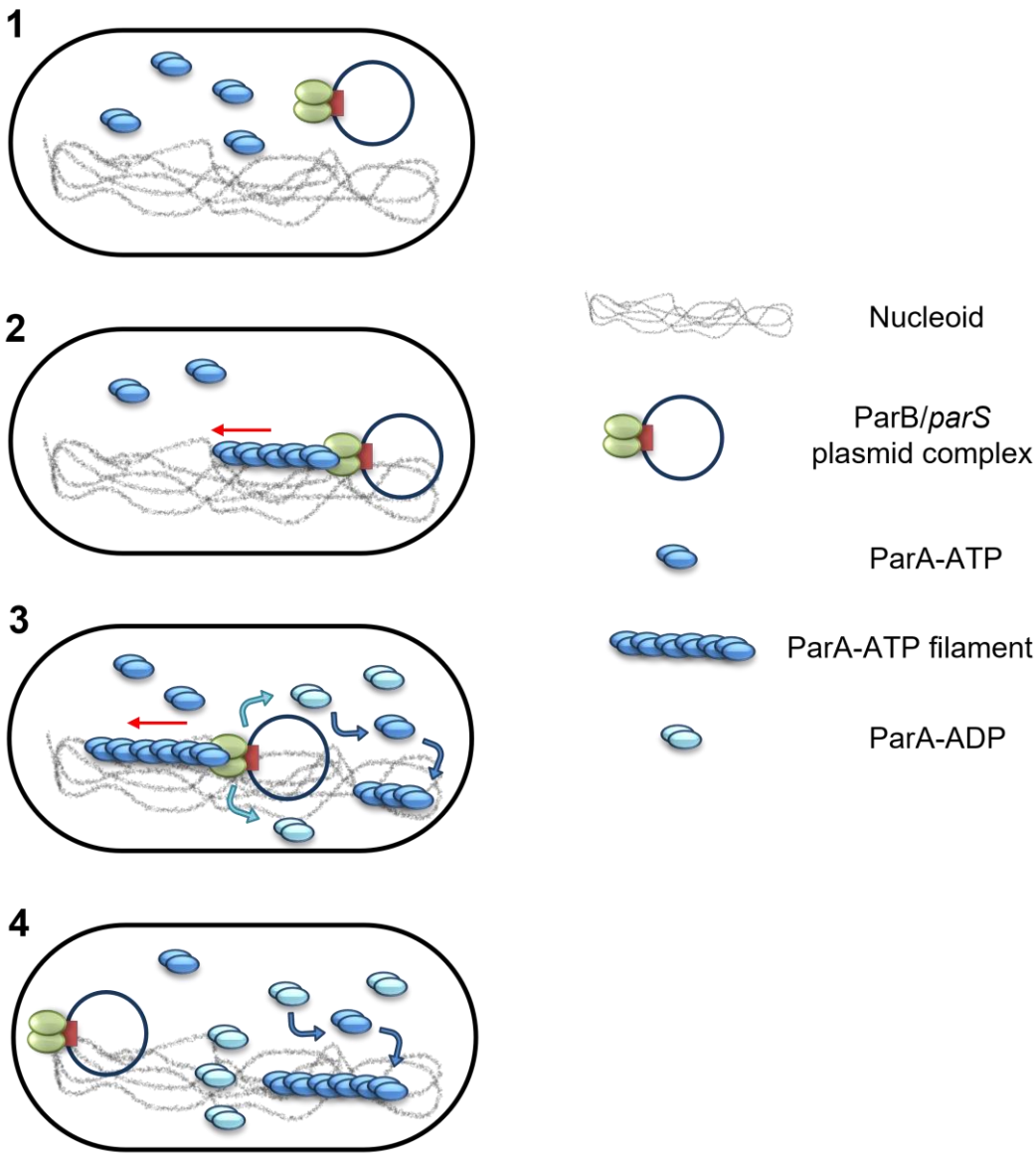


Figure 1.10. Type I pulling segregation mechanism. **(1)** The CBP ParB binds to the *parS* centromere on the plasmid. **(2)** ParA motor proteins form a filament when bound to ATP in the dimer-state, and bind the non-specific DNA of the nucleoid in a cooperative fashion. Interactions with ParB attach the plasmid to the ParA filament. **(3)** Stimulation of the ATPase activity of ParA by ParB leads to hydrolysis to ADP, release from the nucleoid, and depolymerisation of the filament. The ParB-plasmid cargo may contact the next ParA-ATP at the tip of the filament. Here, pulling of the plasmid is shown (red arrows), but a pushing mechanism would also be viable. New filament growth could occur elsewhere in the cell, behind the plasmid. **(4)** The plasmid has been translocated to one cell pole, where filament disassembly results in it being dropped. Adapted from Ringgaard *et al*. 2009.

1.5.2 Type I Brownian diffusion ratchet mechanism

This model proposes a different mechanism, as various *in vitro* and *in vivo* data disagreed with the formation of discrete ParA filaments (Baxter & Funnell 2014). Although some ParA family proteins had been observed to form filament-like structures *in vivo*, ParA of plasmid P1 was not observed to do so in fluorescent microscopy studies (Hatano & Niki 2010). Instead, the distribution of ParA was observed to correspond to that of the nucleoid, indicating non-specific binding to chromosomal DNA. Vecchiarelli and colleagues proposed a different model called the diffusion-ratchet (or Brownian-ratchet) mechanism, based on observations that, upon binding ATP, ParA undergoes a slow conformational change, enabling it to bind non-specific DNA. This more gradual cycling between DNA-binding and non-binding forms of ParA creates an uneven distribution across the nucleoid, which acts as a matrix for plasmid movement (Vecchiarelli *et al*. 2010). The stimulation of ParA ATPase activity by ParB removes ParA from the nucleoid close to the partition complex, granting a directionality to the plasmid cargo due to the low concentration gradient of ParA in its wake. Meanwhile, the time-delay in the ParA conformational change allows it to randomly diffuse and bind nucleoid DNA within the cell **(Figure 1.11)** (Vecchiarelli *et al*. 2010).

Although ParA of pB171 was shown to form filaments, leading the authors to propose the pulling mechanism previously described, ParA was also shown to form dynamic helical clouds within the nucleoid, meaning components of both models may not be mutually exclusive (Ringgaard *et al*. 2009). The diffusion-ratchet model was also proposed as a result of cell-free studies of ParA and F plasmid SopA, where a DNA-carpeted flow cell was used to mimic the bacterial nucleoid, and here the motor proteins again formed dynamic patterns on the DNA (Hwang *et al.* 2013, Vecchiarelli *et al*. 2013a). For the chromosomal ParAB system of *C. crescentus*, a variation of this model has been proposed. Here, although ParA binds to chromosomal DNA, diffusion was insufficient to lead to directional motion of the ParB-*parS* complex. An alternative 'DNA-relay' mechanism was proposed, in which the elastic force of chromosomal DNA propels the plasmid cargo from one cell pole to the other (Lim *et al*. 2014).

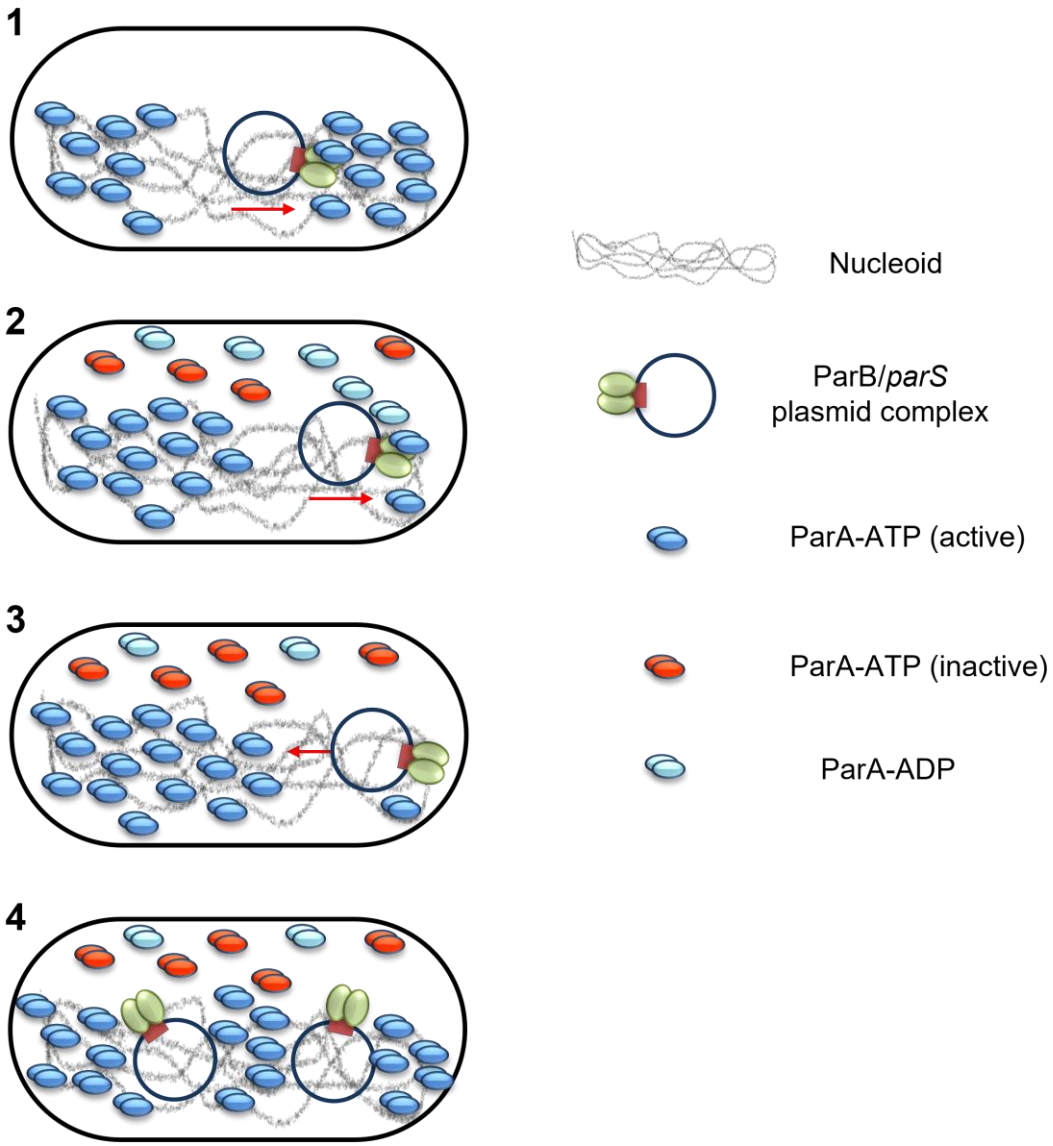


Figure 1.11. Type I diffusion ratchet partition mechanism. **(1)** ParB binds to the plasmid *parS* centromere. The ParA-ATP dimer (DNA-binding, active form), binds non-specifically to the nucleoid. **(2)** ParB recruits ParA into the partition complex, and stimulates its ATPase activity. This hydrolyses ATP to ADP, releasing ParA-ADP from the nucleoid and allowing it to diffuse throughout the cell. The plasmid continues in one direction as there is a lower ParA concentration in its wake. ParA exchanges ADP for ATP, but is in an inactive form due to a time-delay, and cannot yet bind the nucleoid. After diffusing and undergoing a conformational change, ParA-ATP can again bind the nucleoid in a random fashion. **(3)** The plasmid changes direction after reaching one cell pole, and moves due to uneven distribution of ParA on the nucleoid. **(4)** After plasmid replication, plasmids are moved apart due to ATP hydrolysis and removal of ParA between them, and are positioned at approximately ¼ ad ¾ positions within the cell length. Adapted from Vecchiarelli *et al*. 2010.

1.5.3 Type I Venus flytrap model

The Type Ib segregation system of plasmid TP228, incorporating the CBP ParG, and the Walker-type ATPase ParF, has been extensively studied. ParF had previously been found to form polymeric filament-like structures *in vitro*, in an ATP-dependent manner, and that filament formation was required for correct plasmid segregation (Barillà *et al*. 2005). This observation, along with evidence of other ParA proteins forming filaments, led to the proposal of the push/pull model described earlier (Hayes & Barillà 2006, Ringgaard *et al*. 2009). The structural basis of ParF polymerisation was determined, as crystals structures of the protein in its ADP- and ATP-bound (where a non-hydrolysable analogue was used) states revealed ParF-ADP to be monomeric, and ParF-ATP to be dimeric, capable of forming dimer-of-dimer polymeric filaments (Schumacher *et al*. 2012).

However, three-dimensional super-resolution microscopy investigations showed that ParF did not form linear polymers that spanned the length of the nucleoid, but instead, it formed higher-order structure patches that oscillated dynamically across the nucleoid from pole to pole, in a manner dependent on the stimulation of its ATPase activity by ParG (McLeod *et al*. 2017). Microscopy revealed that ParF formed a meshwork that permeates throughout the nucleoid volume, entrapping and transporting the ParG-plasmid complex throughout the cell. ParF was found to bind non-specific DNA *in vitro* in an ATP-dependent manner (McLeod *et al*. 2017).

These data led to the proposal of the Venus flytrap model of segregation, where transient and dynamic remodelling of the ParF meshwork on the nucleoid effects plasmid segregation. The meshwork consists of a denser leading edge, and a looser lagging edge, containing fewer oligomers. The duplicated ParG-bound plasmids are engulfed into the meshwork via interactions with ParF, and transported towards one cell pole. The ParF polymer network grows between the plasmids, separating them, and one plasmid becomes detached from the lagging edge after ATP hydrolysis and polymer disassembly, causing deposition of the plasmid at one pole. The other plasmid, attached to the leading edge of the ParF meshwork, is translocated to the other cell pole and eventually released, again due to ATP hydrolysis by ParF **(Figure 1.12)** (McLeod *et al*. 2017).

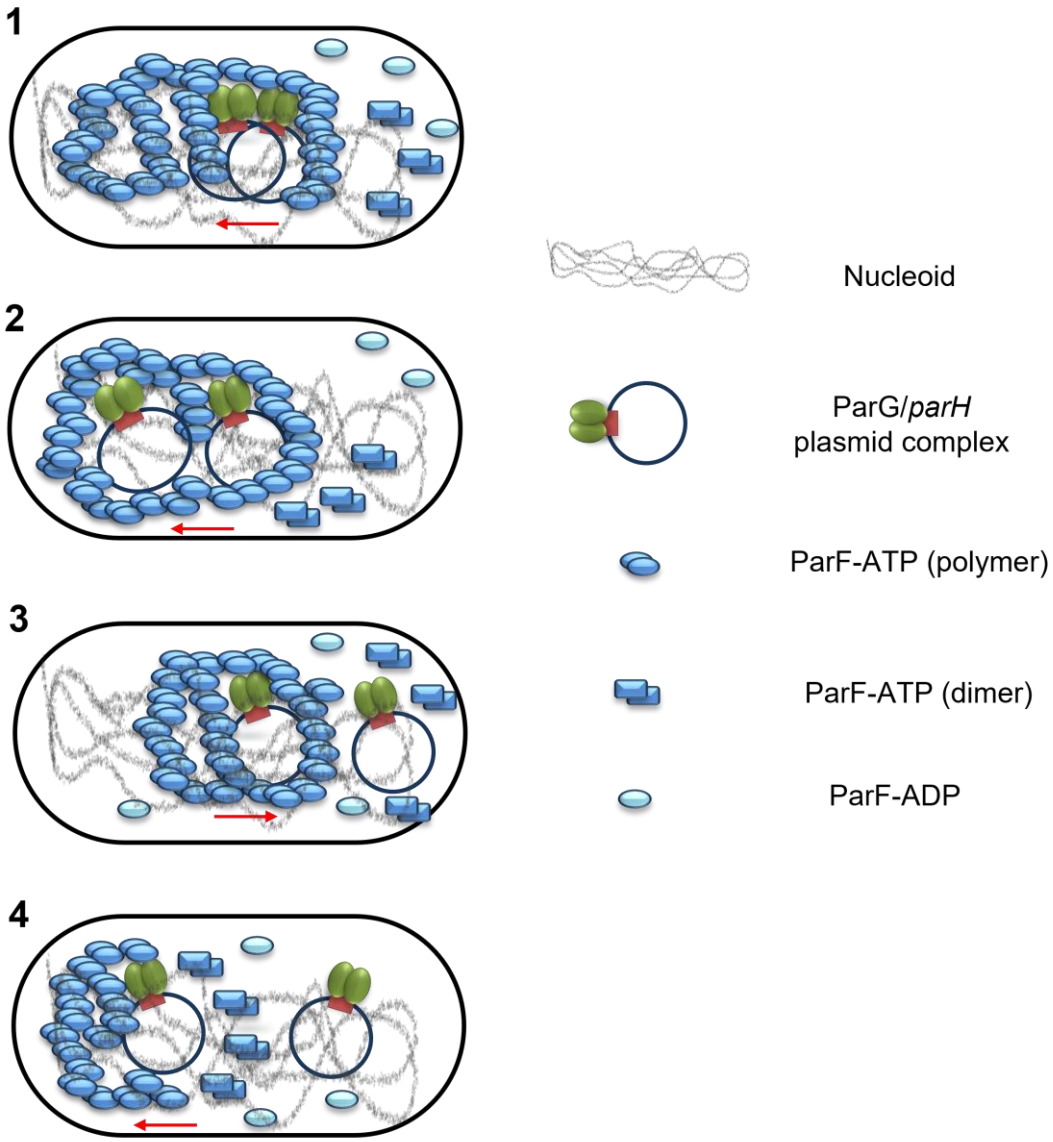


Figure 1.12. Type Ib Venus flytrap partition mechanism. **(1)** The newly-replicated plasmids are bound at *parH* by ParG dimers. ParF dimerises upon binding ATP, and binds to the nucleoid to form a meshwork of higher-order oligomers, which form a three-dimensional matrix throughout the nucleoid volume. The ParF meshwork comprises a denser, leading edge and a less dense tail. **(2)** The plasmids are engulfed within the meshwork and transported to one cell pole, during which time ParF polymers grow between them. **(3)** After reaching a cell pole, one plasmid is released due to stimulation of ParF ATPase activity and meshwork disassembly. **(4)** The other plasmid will remain attached to the ParF meshwork and is moved to the opposite pole. The ParF meshwork oscillates repeatedly across the nucleoid, repositioning the plasmids at ¼ cell and ¾ cell locations prior to cell division. Adapted from McLeod *et al*. 2017.

1.5.4 Pushing mechanism of Type II systems

The segregation mechanism of Type II systems such as ParMR*C* encoded on plasmid R1 has been described as pushing, filament driven, or insertional polymerisation (Møller-Jensen *et al*. 2003, Schumacher 2008, Bouet & Funnell 2019). The actin-like ParM ATPase was demonstrated to form polymeric filaments both *in vitro* and *in vivo*, in an ATP-dependent manner, and required both the CBP ParR and the centromeric site *parC* (Møller-Jensen *et al*. 2002). The ParR dimers bind to single repeats within the *parC* sites, resulting in the formation of a U-shaped/helical open-ring protein-DNA superstructure, that can act as the nucleation site for ParM (Møller-Jensen *et al*. 2002, 2007, Hoischen *et al*. 2008). The ParM filaments were observed via fluorescence microscopy to extend from pole to pole within the cell, and had plasmids attached at each end (Møller-Jensen *et al*. 2003, Campbell & Mullins 2007).

The ParM monomers polymerise to form an antiparallel bipolar spindle, comprised of two ParM filaments that form left-handed helices in a head-to-tail arrangement, with a pointed and barbed end on each filament. The ParR-*parC* complex caps the filament at one end, however new ATP-bound ParM monomers can be added to the filament at the ParR-*parC* interface, elongating the spindle. As the spindle is antiparallel and comprises two ParM filaments, replicated plasmids situated at each end can therefore be propelled to opposite cell poles, in a longitudinal filament manner similar to that initially proposed for Type I systems **(Figure 1.13)** (Salje & Löwe 2008, Gayathri *et al*. 2012, Bharat *et al*. 2015).

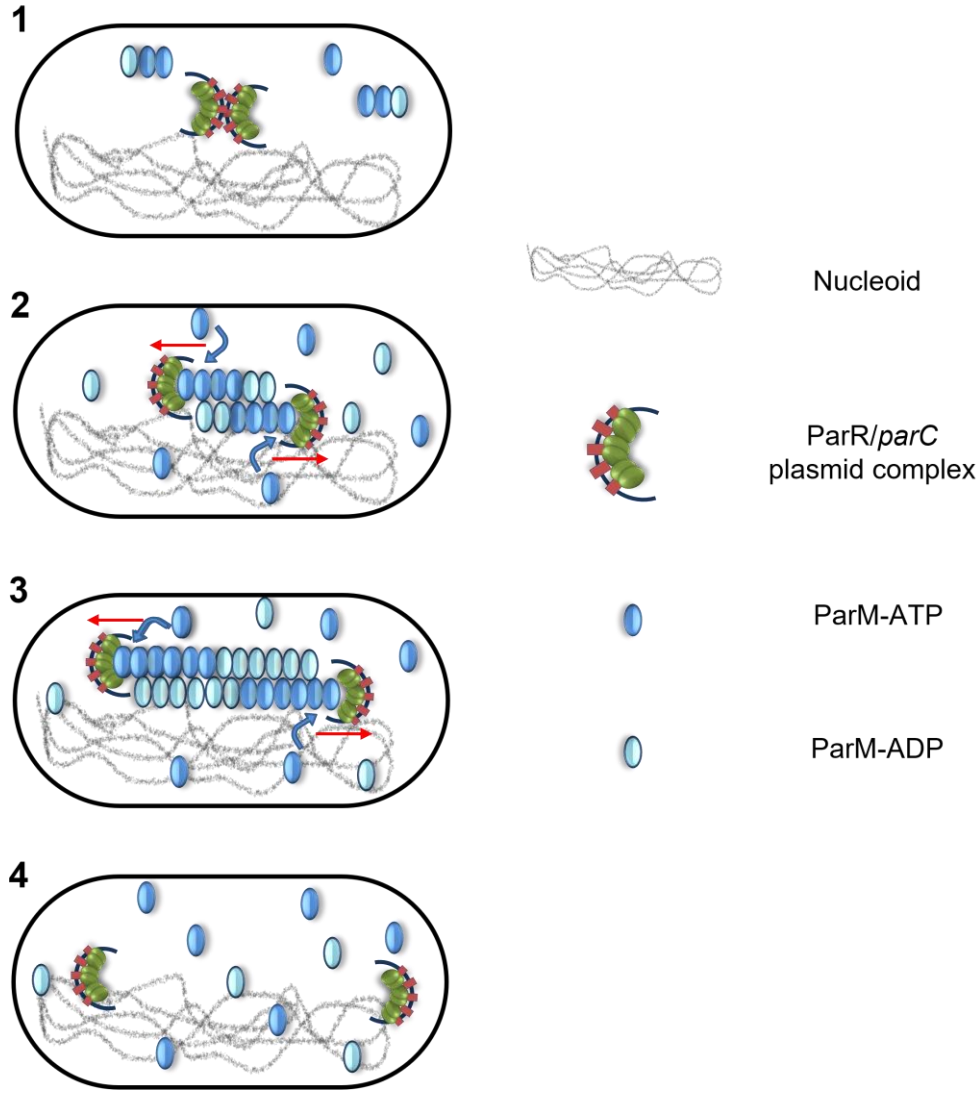


Figure 1.13. Type II insertional polymerisation segregation mechanism. (1) The newly-replicated plasmids are shown at mid-cell. ParR dimers bind to each repeat of the *parC* site on the plasmid centromere and cause the DNA to bend into a circular U-shape. **(2)** The ParR-*parC* partition complex is the nucleation site for ParM. ParM-ATP monomers are added at the nucleation site. **(3)** The ParM filament attaches to both plasmids, forming an antiparallel spindle, which elongates and pushes the plasmids apart. ATP hydrolysis slightly releases the partition complex and allows addition of another ParM-ATP subunit. The ADP-bound ParM filament is less stable, but is temporarily protected by the ATP cap and binding to the ParR-*parC* partition complex. **(4)** The ParM filaments eventually disassemble, releasing the plasmids at opposite cell poles.

1.5.5 Treadmilling mechanism of Type III systems

The TubZR*C* segregation system encoded on the *Bacillus thuringiensis* plasmid pBtoxis has also been proposed to involve a filamentation mechanism. Type III systems are known as tubulin-like as the motor protein, TubZ, is a member of the FtsZ/tubulin superfamily of GTPases (Larsen *et al*. 2007). TubZ was found to assemble into linear polymers *in vivo*, which moved dynamically through the cell, changing direction once at the cell pole to move to the pole opposite (Larsen *et al*. 2007). TubZ was found to exhibit polarity, with a plus end and a minus end, however unlike ParM, TubZ polymerises from the leading plus end by addition of TubZ-GTP, and depolymerises by dissociation from the minus end, in a process dubbed treadmilling (Aylett *et al*. 2010, Bouet & Funnell 2019).

As in other segregation systems, the CBP TubR binds to repeat sequences in the *tubC* centromeric site (Ni *et al*. 2010). The C-terminal flexible tail of TubZ was implicated both in interactions with TubR-*tubC*, and polymerisation into filaments in a GTP-dependent manner (Ni *et al*. 2010, Fuentes-Pérez *et al.* 2017). Fink and colleagues used fluorescently-labelled TubR and TubZ to demonstrate that the TubR-*tubC* complexes preferentially bound to the shrinking, minus end of the TubZ filament, where TubZ-GDP depolymerises, and never bound to the plus, or growing ends (Fink & Löwe 2015). Thus, in this partition system, the TubR-bound plasmids are not pushed via insertional polymerisation in the manner described for ParMR*C*, but rather translocated from midcell to the cell pole by the treadmilling assembly/disassembly kinetics of the TubZ filament (Fink & Löwe 2015). The mechanism by which plasmids are deposited at the cell poles is unclear, but it has been proposed that reaching the pole induces bending in the TubZ filament, causing strain at the binding interface with TubR-*tubC*, and subsequently detaching the plasmid cargo **(Figure 1.14)** (Ni *et al*. 2010).

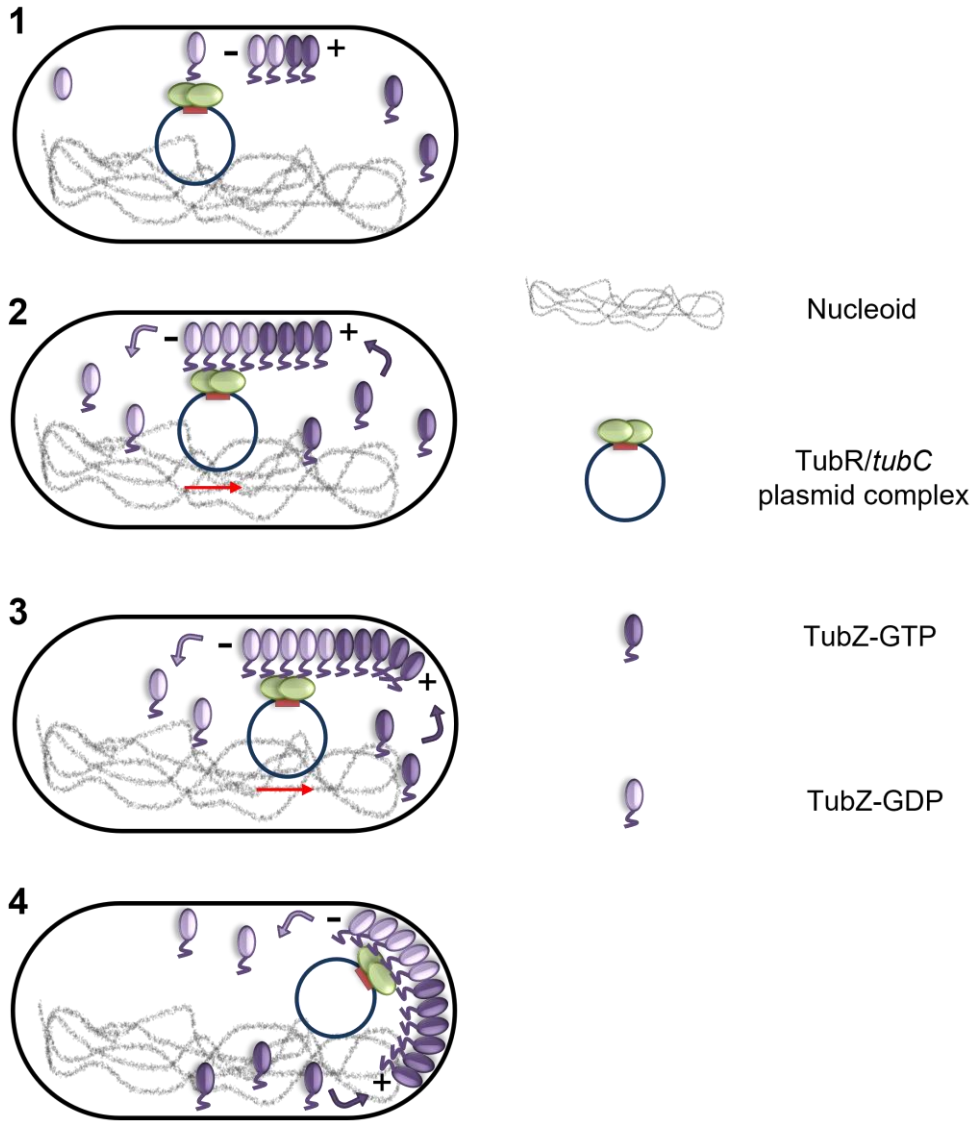


Figure 1.14. Type III segregation system treadmilling mechanism. **(1)** The CBP TubR binds to the *tubC* centromere on the replicated plasmid. TubZ forms filaments in both GDP- and GTP-bound states, however the filament exhibits a polarity, in which TubZ-GTP is added to the plus end, whilst TubZ-GDP disassociates from the minus end. **(2)** The TubR-*tubC* complex attached to GDP-bound forms of TubZ, via the flexible C-terminal tail of TubZ. **(3)** The successive polymerisation and depolymerisation from the plus and minus ends of TubZ respectively, acts to move the plasmid cargo across the cell in a process called treadmilling. **(4)** The TubZ filament curves upon reaching the cell interior at one pole, and the torsional stress placed on the binding interface with TubR is thought to cause the plasmid to detach from the filament.

1.6 Par AB*S* involvement in chromosome segregation

ParAB*S* segregation systems also mediate bacterial chromosome partitioning, as over 60% of species have been found to harbour *parAB* homologues (Livny *et al*. 2007). The *parS* sites are often located near to the origin of replication, and the ParB protein binds site-specifically to these sequences. ParB has been observed to bind DNA non-specifically and spread for several kilobases from the *parS* site, potentially functioning to bridge together and condense more distant DNA regions (Lin & Grossman 1998, Bartosik *et al*. 2004, Graham *et al*. 2014, Tran *et al.* 2018). Chromosome-encoded ParA has weak ATPase activity that is stimulated by ParB. A ParA filament push/pull model of segregation was proposed, similar to that in plasmid models (Ptacin *et al*. 2010). A more recent model in *C. crescentus*, which extended the plasmid diffusion-rachet mechanism to include the elastic forces imparted by the nucleoid DNA, has also been proposed to drive chromosome segregation (Lim *et al*. 2014).

Other proteins alongside ParAB have been implicated in chromosome segregation, such as SMC (Structural Maintenance of Chromosomes), which forms a homodimer, and along with subunit proteins ScpA and ScpB, forms a ring-like structure known as condensin. In *B. subtilis*, SMC was shown to be recruited to chromosomal regions proximal to the replication origin by Spo0J (ParB) bound to *parS* sites, and that SMC recruitment was required to promote chromosome segregation (Gruber & Errington 2009, Sullivan *et al*. 2009). Cells that lacked SMC or in which SMC was inactivated were unable to correctly segregate replicated chromosomes (Wang *et al*. 2014). The capacity of Spo0J to form dimer-dimer interactions and bring together distal DNA regions was suggested as a means by which SMC could be loaded onto the resultant DNA loops (Graham *et al*. 2014). This 'loop extrusion' model for DNA condensation was previously proposed for eukaryotic chromosome organisation, and bacterial SMC shares homology with its eukaryotic counterpart (Kamada & Barillà 2018, Srinivasan *et al*. 2018). The recruitment and loading of SMC at *parS* sites by ParB has also been observed in *C. crescentus*, where SMC complexes can progressively align and tether the two chromosome arms together (Tran *et al*. 2017).

The SMC complex is known to have ATPase activity which aids DNA entrapment, and studies of *B. subtilis* condensin demonstrated that ATP-bound dimers could recognise and bind to ParB-*parS* nucleoprotein complexes, and upon ATP hydrolysis were released from *parS* sites and able to relocate large distances along the chromosome, again causing the extrusion of DNA loops (Minnen *et al*. 2016, Wang *et al*. 2018).

Intriguingly, recent evidence has demonstrated that ParB also has nucleotide hydrolysis capacity, as *B. subtilis* ParB displayed CTPase activity. The ability to bind and hydrolyse CTP to CDP was promoted by *parS* DNA, and the formation of extended partition complexes through ParB spreading from *parS* required the CTP-bound form, with ParB acting as a sliding clamp on the DNA (Soh *et al*. 2019, Balaguer *et al*. 2021). These findings were recapitulated elsewhere; ParB of *Myxococcus xanthus* was shown to have CTPase activity which was crucial for partition complex formation *in vivo*, whilst in *C. crescentus*, CTP was required for ParB spreading *in vitro* (Osorio-Valeriano *et al*. 2019, Jalal *et al*. 2020b). The CTP-binding region of *C. crescentus* ParB was mapped to the N-terminal domain, and CTP-binding was shown to alter the conformational state of ParB. This could facilitate its relocation away from *parS* and thus allow spreading on the DNA, whereas hydrolysis to CDP was proposed to release ParB from the DNA (Jalal *et al*. 2021). Therefore, the rate of CTP hydrolysis by ParB determines the length of time ParB can spend spreading on the DNA, and concomitantly controls the resultant size of the ParB-DNA partition complex (Osorio-Valeriano *et al*. 2021).

These recent findings have therefore contributed to the overall picture of the partition process in bacteria, uncovering some of the mechanisms by which Par proteins interact with each other, the nucleoid DNA, and act in combination with additional necessary proteins such as SMC-condensin, to faithfully potentiate the segregation of replicated chromosomes **(Figure 1.15)**.

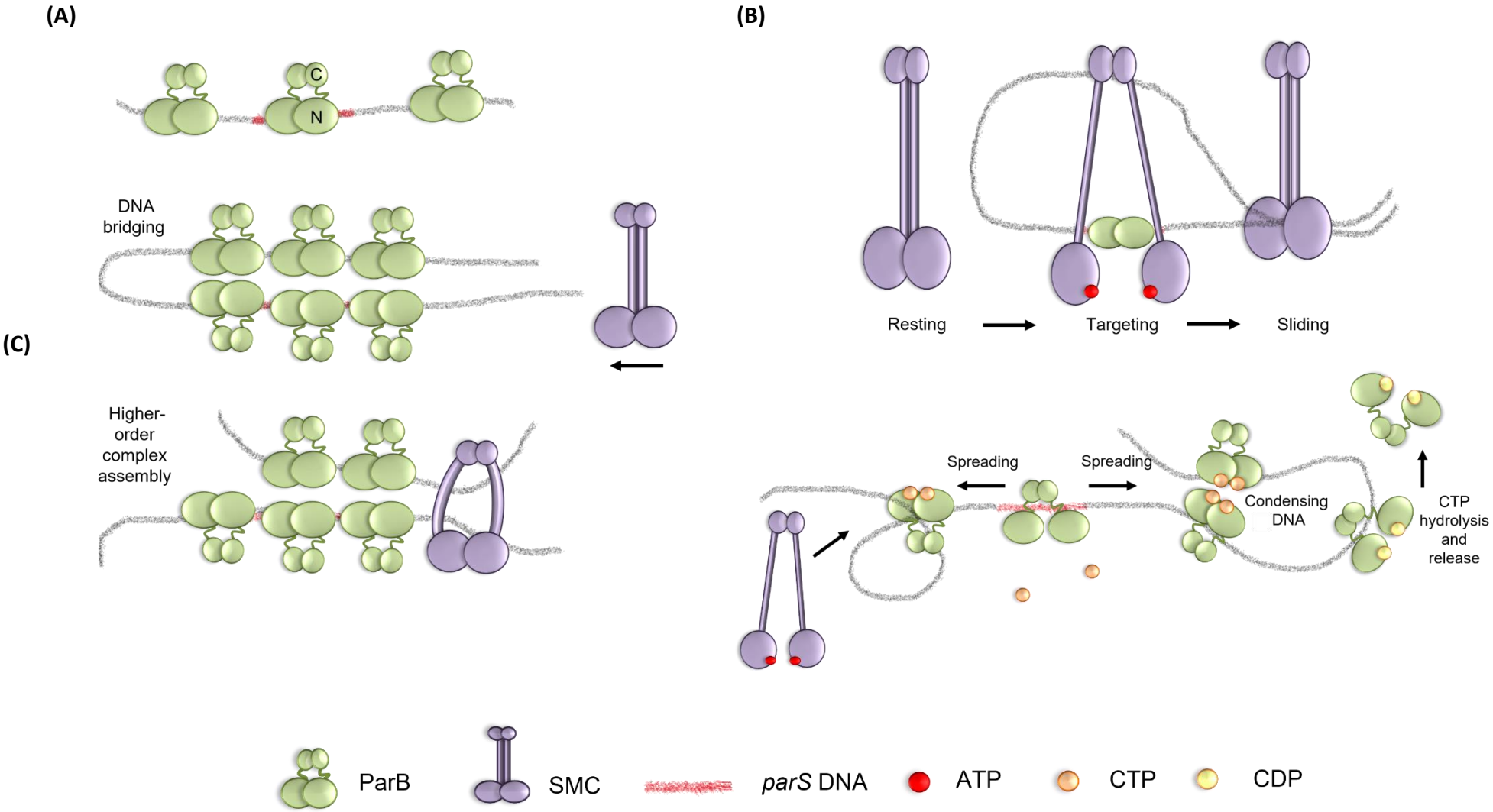

(A)
C
N
DNA bridging
Higher-order complex assembly
(B)
Resting
Targeting
Sliding
Spreading
Spreading
Condensing DNA
CTP hydrolysis and release
(C)
ParB
SMC
parS DNA
ATP
CTP
CDP

Figure 1.15. Chromosome segregation mechanisms (previous page). **(A)** Model of *B. subtilis* ParB (Spo0J) spreading and bridging DNA. ParB monomers form dimers via their C-termini, and bind site-specifically to *parS* via the N-terminal DNA-binding domain. Spreading laterally via nearest-neighbour interactions, or DNA bridging via N-terminal interactions with dimers on more distal DNA is possible, compacting the chromosome. Higher-order complex assembly can be mediated by the recruitment of SMC/condensin by ParB, which can then entrap DNA loops. **(B)** SMC/condensin is targeted to chromosomal *parS*/ParB via ATP-binding and opening of the SMC arms to a ring-like configuration. ATP hydrolysis then allows SMC/condensin to relocate to more distant parts of the chromosome away from *parS*, driving DNA loop-extrusion. **(C)** Recent data show that some ParBs have CTP hydrolase activity, leading to a model of ParB engagement and movement on the chromosome. ParB dimers nucleate at *parS* in a non-CTP-bound state, and CTP binding induces a conformational change that locks the dimer on the DNA, and allows spreading from *parS*. Spreading and N-terminal interactions in the CTP-bound state could engender DNA condensation and compaction. CTP hydrolysis to CDP leads to a reversion to a more open conformational state of the ParB dimer and so release from the DNA. SMC/condensin could be recruited by spreading/sliding ParB dimers. Note that this figure depicts general models, specifics e.g. whether CTP binding is required for *parS* loading may differ between ParBs. Adapted from Graham *et al*. 2014, Minnen *et al*. 2016, Osorio-Valeriano *et al*. 2019, Jalal *et al.* 2020b, Osorio-Valeriano *et al*. 2021.

1.7 A brief introduction to the Archaea domain

1.7.1 Archaea, their discovery and phylogeny

Thus far, this introduction has described genome segregation in bacteria, as this process has been extensively studied and is relatively well understood in this domain of life. However, this study relates to the partitioning system encoded by an archaeal plasmid, and as such, a brief detour into the discovery of archaea, archaeal biology, and their evolutionary relationship to the other primary domains of life, is useful.

Archaea are unicellular microorganisms, which are abundant on our planet and inhabit a diverse array of habitats and environmental niches (Dombrowski *et al*. 2019). However, given the ubiquity of archaea, it is surprising to find that this entire domain of life was only recognised a little over forty years ago, when they were taxonomically reclassified, previously being recognised as bacteria. This pioneering work was conducted in the 1970s by Carl Woese and George Fox, who set out to construct a universal phylogeny based on 16S and 18S ribosomal RNA sequences, due to the universality of the molecule and its slow mutation rate over time (Woese & Fox 1977). This resulted in two main phylogenetic groups, dubbed 'eubacteria' and 'urkaryotes' (prokaryotes and ancestral eukaryotes respectively). However, the anaerobic methanogens, prokaryotes that were previously classified as bacteria, formed a third distinct grouping in the ribosomal RNA data, and it led the authors to state that these organisms appeared 'no more related to typical bacteria than they are to eukaryotic cytoplasms'. Not only was this true when comparing the 16S rRNA genes of several methanogenic species with bacteria (Balch *et al.* 1977, Fox *et al*. 1977), but also when comparing phenotypes between methanogens and bacteria. Though morphologically similar in cell size and shape, there were fundamental differences in structure. Methanogens lacked peptidoglycan in their cell walls (Kandler & Hippe 1977), possessed distinguishing enzymes involved in methane formation (Woese & Fox 1977), and had characteristic tRNAs (Woese *et al*. 1978). Other features of archaea, such as their RNA polymerases and transcription factors, were more similar to those found in eukaryotes compared to bacteria (Walsh & Doolittle 2005).

The fundamental molecular and phenotypic differences between the methanogens and bacteria led Woese and Fox to pronounce that the methanogenic 'bacteria' were in fact organisms representing a third 'urkingdom' named 'archaebacteria', and that, therefore, the tree of life contained three main branches, rather than two (Woese & Fox 1977). An early universal phylogenetic tree constructed by comparison of 16S rRNA sequences, is shown below in **Figure 1.16A**, and subdivides the archaebacteria into three main subkingdoms: extreme halophiles, methanogens and extreme thermophiles (Woese 1987). In the following decade, evolutionary relationships between organisms were routinely being constructed using molecular data, and Woese proposed a revision of the current taxonomical hierarchy, with Domain as the highest rank. The Eubacteria were renamed Bacteria, the Archaebacteria as Archaea, and the classic three-domains phylogenetic tree, in which Archaea and Eukarya are positioned as monophyletic sister groups, with Bacteria more distantly related, was published **(Figure 1.16B)** (Woese *et al.* 1990). The tree shows the two main archaeal lineages known at the time (named kingdoms but now more commonly described as phyla): the Euryarchaeota, which includes methanogens, halophiles and some thermophiles, and the Crenarchaeota, predominantly extremely thermophilic organisms including the genus *Sulfolobus*, a strain of which is the focus of this study.

As more archaeal species were discovered, the phylogeny of the main archaeal lineages has necessarily expanded, to include additional superphyla. The Crenarchaeota were proposed to belong to the 'TACK' superphylum, along with the Thaumarchaeota, Korarchaeota and Aigarchaeota, forming a distinct clade separate to the Euryarchaeota (Guy & Ettema 2011). Another archaeal superphylum, the DPANN (named after the initial five groups incorporated: Diapherotrites, Parvarchaeota, Aenigmararchaeota, Nanoarchaeota and Nanohaloarchaeota) are characterised as possessing small cell and genome sizes (Rinke *et al.* 2013). Whilst there is some controversy over the phylogenetic positioning of DPANN and lineages therein, most studies have supported the DPANN superphylum being a deeply-branching sister group to the Euryarchaeota/TACK lineage (Williams *et al.* 2017, Dombrowski *et al.* 2019). Research in the last decade has uncovered another novel archaeal superphylum that has shed further light on the evolutionary relationships between archaea and eukaryotes, and the topology of the tree of life has undergone a proposed revision. The 'Lokiarchaeota' phylum was proposed after using

metagenomic sequencing on sediments collected from a deep-sea underwater vent called Loki's Castle (Spang *et al*. 2015). The Lokiarchaeota were found to form a monophyletic group with eukaryotes, and interestingly, further genetic analysis revealed the presence of genes encoding eukaryotic signature proteins (ESPs), including actin homologues and small GTPases. These findings resulted in the paradigm-shifting observation that all of eukaryotes may have arisen from within an archaeal lineage, implying that there are only two primary domains of life rather than three **(Figure 1.16C)** (Spang *et al*. 2015, Williams *et al*. 2020).

More recently, additional closely-related phyla (also named after Norse gods) have been discovered, including the Odinarchaeota, Thorarchaeota and Heimdallarchaeota, with these groups incorporated into the Asgard superphylum. Further phylogenomic studies reinforced those conducted with Lokiarchaeota alone, showing eukaryotes arising from within the Asgard group and most closely-related to the Heimdallarchaeota (Zaremba-Niedzwiedzka *et al.* 2017). Although there still remains some debate as to the tree of life comprising two or three domains, a continual expansion of the Asgard superphylum with the inclusion of a greater number of metagenome-assembled genomes, provides increasingly robust support for the two-domain topology (Liu *et al*. 2021).

Despite these recent discoveries, the exact archaea to eukaryote evolutionary transition remains unexplained, partly due to the dearth of cultured Asgard specimens. One group recently reported their success in culturing an archaeon related to Lokiarchaeota (Imachi *et al*. 2021). Isolation and cultivation of the archaeon (*Candidatus* Prometheoarchaeum syntrophicum) over a period of ten years allowed Imachi and colleagues to characterise the morphology of this strain. The cells exhibit long branching protusions, leading to a proposed model of eukaryogenesis in which these protusions capture and engulf the bacterial future mitochondrion (Imachi *et al*. 2021). A similar model, supported by data showing the Heimdallarchaeota as the closest relative to eukaryotes, has recently been prosed in which a Heimdall-like archaeon subsumes an endosymbiotic bacterium, thus establishing the first eukaryotic common ancestor (Wu *et al*. 2022).

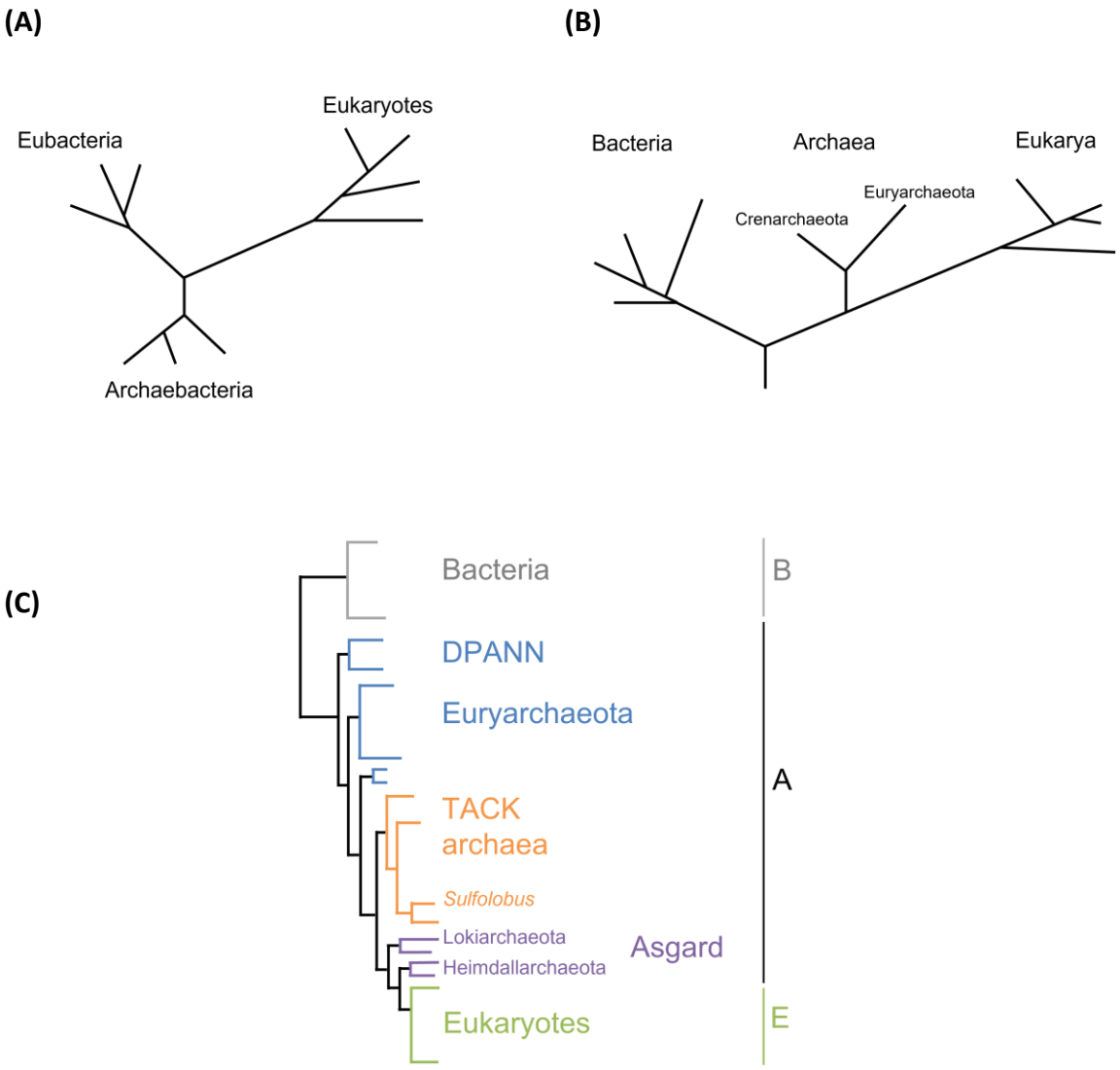


Figure 1.16. Early to modern phylogenetic trees of the biological domains. **(A)** An early universal unrooted tree showing the three 'urkingdoms' (Figure adapted from Woese 1987). **(B)** The rooted universal tree showing the proposed three domains of life. The two main lineages of archaea, Crenarchaeota and Euryarchaeota, are indicated (adapted from Woese *et al.* 1990). **(C)** Example of a modern phylogenetic tree, after the discovery of the Asgard archaea, which demonstrates the evolution of eukaryotes from within archaea in a two-domains tree of life. The TACK and DPANN superphyla are shown.. The *Sulfolobus* genus is indicated within the TACK superphylum. B – Bacteria, A – Archaea, E – Eukaryotes. Adapted from Williams *et al*. 2020. All trees are simplified versions of those originally published.

1.7.2 The crenarchaea genus *Sulfolobus*

The archaeal strain which harbours a plasmid-encoded segregation system under investigation in this study belongs to the genus *Sulfolobus* (family Sulfolobaceae), and is a member of the crenarchaeal phylum, one of the two originally proposed archaeal lineages **(Figure 1.16B)**. The Crenarchaea were originally only thought to be extreme sulphur-metabolizing thermophiles, with all cultured species exhibiting this characteristic. Since then, however, crenarchaea have been discovered in great abundance in low-temperature terrestrial and aquatic environs, including species that oxidize ammonia (DeLong 1992, Fuhrman *et al*. 1992, Könneke *et al*. 2005). The genus *Sulfolobus* (*sulfo* - sulphur metabolism, *lobus* – spherical but with irregular lobes) is thermophilic and acidophilic, and was first isolated from a variety of high temperature thermal habitats such as acidic soils and hot springs, with low pH (<3), and temperatures between 65 and 95°C (Brock *et al*. 1972). The type species *Sulfolobus acidocaldarius* was isolated from a hot spring in Yellowstone National Park, with subsequent described species often named after their habitat of discovery, e.g. *Sulfolobus solfataricus*, isolated from the Solfatara crater near Naples, Italy (Zillig *et al*. 1980). It has recently been suggested that *S. solfataricus* and *S. shibatae* be reclassified as belonging to the newly-proposed genus *Saccharolobus*. This is based on 16S and 23S rRNA phylogenies indicating *S. solfataricus* belongs to a monophyletic clade distinct from the grouping containing the *Sulfolobus* type species, *S. acidocaldarius.* The 16S rRNA sequence similarity score between *S. solfataricus* and *S. acidocaldarius* supports this, being below the 95% boundary used to distinguish between genera (Sakai & Kurosawa 2018). The *Sulfolobus* strain used in this study has the designation NOB8-H2, after its isolation from acidic thermal springs in Noboribetsu, Japan (Schleper *et al*. 1995).

Archaea, like bacteria, are distinguished from eukaryotes in lacking an internal compartment in which genetic material is confined. In contrast, the chromosome (and any extra-chromosomal genetic elements) occupy a region within the cell volume called the nucleoid. Archaeal genomes comprise a circular chromosome, and often extrachromosomal elements such as plasmids. Chromosome copy-number differs between archaeal phyla, with the Euryarchaea generally being polyploid, whilst

crenarchaeal genera, including *Sulfolobus*, possess a single chromosome (Barillà 2016). *Sulfolobus* species typically harbour a genome between 2 and 3 Mb in size, with *S. solfataricus* genomes tending to be slightly larger than that of *S. acidocaldarius* (Dai *et al.* 2016).

1.7.3 Archaeal DNA organisation and segregation

As in both eukaryotes and bacteria, archaea also employ various mechanisms to aid the organisation of genetic material, both in terms of large-scale three-dimensional genome architecture, and in the accurate segregation of replicated chromosomes and plasmids to daughter cells. Compared with other domains of life, in which these processes are better understood, our knowledge of chromosome organisation and segregation in archaea is more primitive (Barillà 2016). Chromatin proteins, such as eukaryotic histones and SMCs, and bacterial nucleoid-associated DNA-binding proteins, including the previously-mentioned IHF, impart order by wrapping, bending, folding and condensing DNA into more compact structures (Zhang *et al*. 2019). Different archaeal phyla utilise distinct sets of proteins to organise their genetic material. For example, the Euryarchaeota, which have multiple chromosome copies, encode both histone and SMC homologues, whereas crenarchaeal species do not harbour histones with the exception of a few species, and do not encode SMC-condensin proteins (Kamada and Barillà 2018, Zhang *et al*. 2019, Maruyama *et al*. 2020). Crenarchaeal species have instead been shown to utilise small basic proteins such as Cren7, CC1 and Sul7 to effectively package DNA, and both Crenarchaea and Euryarchaea employ the DNA-binding protein Alba to shape the genome by bridging and compacting the DNA (Luo *et al.* 2007, Guo *et al.* 2008, Laurens *et* al. 2012, Zhang *et al*. 2012, Maruyama *et al*. 2020).

Interestingly, although *Sulfolobus* spp. lack canonical condensins, Bell and colleagues recently described the higher-order structuring of chromosomes in *S. solfataricus* and *S. islandicus*. Their findings indicated that *Sulfolobus* chromosomes are organised into two distinct compartments: the A-compartment harbours the three *Sulfolobus* replication origins and transcriptionally active genes, whereas the B-compartment contains genes

with lower levels of transcription. Moreover, the B-compartment was also enriched with a novel SMC-superfamily protein dubbed coalescin, whose gene was found only in the Sulfolobales order, which could function in large-scale chromosomal organisation in the absence of condensin (Takemata *et al.* 2019, Takemata & Bell 2020).

DNA partition cassettes analogous to the bacterial *parABS* systems described earlier have also been found in archaea. A chromosome segregation system comprising two proteins and a centromere-like region was discovered in *S. solfataricus*. One protein, SegB is an archaea-specific factor, however it functions analogously to bacterial ParBs by binding as a dimer to specific consensus sequences near to the partition cassette, sites which could be centromeres or regulatory regions. The other protein, SegA, is an orthologue of bacterial, Walker-box ParA motor proteins, and was shown to assemble into polymers *in vitro* when bound to ATP (Kalliomaa-Sanford *et al*. 2012). This *segAB* cassette was also found to be widespread across archaeal species from both crenarchaeal and euryarchaeal phyla, indicating possible conservation of this mechanism of chromosome segregation (Barillà 2016). Intriguingly, during the course of writing this thesis, the crystal structures of both SegA and SegB were published (Yen *et al*. 2021). SegA was found to adopt a different dimer conformation compared to a typical ParA 'sandwich dimer' configuration. Moreover, SegA did not require nucleotides in order to bind DNA, unlike other ParA superfamily proteins (Yen *et al*. 2021). The structure of SegB also provided novel findings, as it was shown to bind DNA via a RHH motif, whereas previously described chromosomal ParB proteins utilise a HTH fold for DNA interactions. SegB was also found to stimulate SegA ATPase activity via its unstructured N-terminal tail, similar to other RHH CBPs (Yen *et al*. 2021). A summary of DNA organisation strategies employed by selected archaeal phyla is shown overleaf in Table 1.2.

Table 1.2 Summary of DNA organisation methods in selected archaeal phyla

Phylum	Histone-like	SMC	Other	Reference
Crenarchaea	No[a]	No		Kamada and Barillà 2018
			DNA organisation proteins:	
			Cren7	Guo *et al*. 2007
			CC1	Luo *et al.* 2007
			Sul7/Sac7	McAfee *et al*. 1996
			Alba	Laurens *et al*. 2012
			Coalescin	Takemata *et al.* 2019
			Partition systems:	
			SegA, SegB	Kalliomaa-Sanford *et al*. 2012
			AspA, ParB, ParA	Schumacher *et al*. 2015
Euryarchaea	Yes			Nishida & Oshima 2017
		Yes		Soppa 2001
			DNA organisation proteins:	
			Alba	Maruyama *et al*. 2020
			HTa, MC1	Zhang *et al*. 2012
Lokiarchaeota	Yes			Nishida & Oshima 2017

a Histones are absent from Crenarchaea except for species in genera *Caldivirga*, *Thermofilum* and *Vulcanisaeta* (Nishida & Oshima 2017).

1.7.4 *Sulfolobus* NOB8-H2 and plasmid pNOB8

This study concerns the crenarchaeal *Sulfolobus* strain NOB8-H2, isolated by Wolfram Zillig from Japanese thermal springs (Schleper *et al*. 1995). The strain is of interest as it accommodates a conjugative plasmid, pNOB8, which harbours a partition cassette exhibiting similarity to bacterial *par* systems. The pNOB8 sequence was determined, and the products of two open reading frames (ORFs) were found to share homology with

bacterial ParA and ParB family proteins, giving clues as to their potential function. Furthermore, there is circumstantial evidence suggesting that the *parBA* cassette of this plasmid encodes a *bona fide* partition system: when pNOB8 is transferred by conjugation or transformation into a different *Sulfolobus* strain, the plasmid undergoes a genetic rearrangement due to a single recombination event, which produces the deletion variant pNOB8-33 (She *et al.* 1998). This plasmid presents a deletion of an ~8 kb region, that results in the loss of the *parBA* cassette, and is not stably maintained (She *et al.* 1998). Furthermore, a third ORF was found to overlap the putative *parB* gene, and the three ORFs together (pNOB8 *orf44, orf45/parB* and *orf46/parA*) appeared to form a single transcriptional entity. The product of *orf44* did not display homology to hitherto characterised segregation proteins.

The genetic arrangement of the putative pNOB8 partition cassette is therefore unusual as it is tricistronic, rather than the bicistronic *parAB* cassettes described earlier **(Figure 1.17)**. Interestingly, bioinformatic searches revealed that this tripartite cassette was conserved across both chromosomes and plasmids in different genera within the crenarchaeal phylum (Schumacher *et al*. 2015). The protein product of *orf44* was named AspA (archaeal segregation protein A), as biochemical and structural studies demonstrated that AspA bound as a dimer to an upstream centromere-like palindrome, and that it displays a winged helix-turn-helix motif, similar to Type Ia ParB proteins, indicating that it may function as a centromere-binding protein (Schumacher *et al*. 2015).

The crystal structure was solved (Schumacher *et al*. 2015), and showed that AspA comprises four alpha-helices and two beta-strands, with the wHTH module followed by a C-terminal dimerisation helix (α4). The majority of DNA base and phosphate contacts are mediated by the N-terminal helices, which demonstrated the ability to insert into both major and minor grooves, and glutamine residues from α3, the recognition helix (Schumacher *et al*. 2015). AspA was also observed to extend its binding to regions upstream of the putative partition cassette from the initial nucleation site at higher concentrations, and form an extended superhelical complex on the DNA. The ability of AspA to spread on the DNA, a property observed with bacterial ParBs, is also mediated by the recognition helix, as adjacent dimers are able to insert their α3 helices

into the same DNA major groove, thus generating a superhelical DNA-protein complex (Schumacher *et al*. 2015).

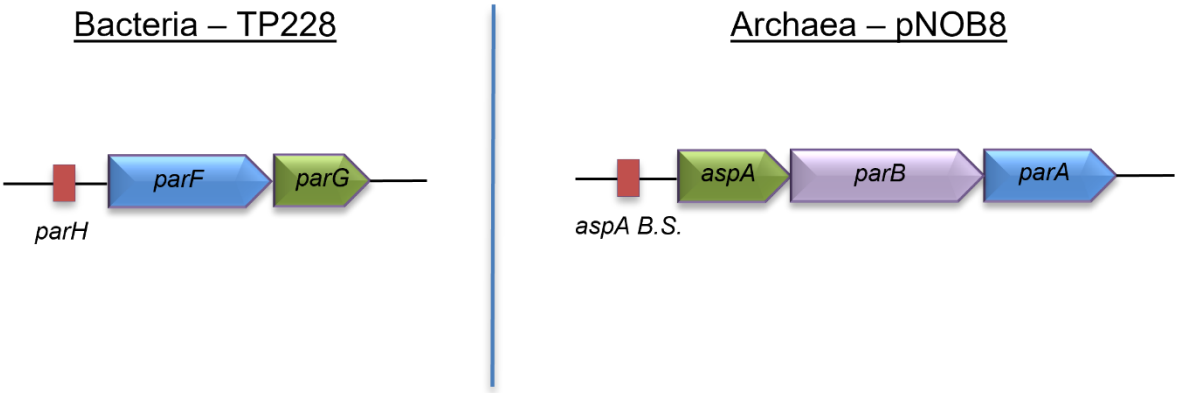


Figure 1.17. Partition cassettes harboured by bacterial plasmids and pNOB8 segregation cassette. Typical genetic organisation of a bacterial plasmid, here *parFGH* from Salmonella Newport TP228 (left) and pNOB8 of *Sulfolobus* NOB8-H2 (right). Here, genes encoding functionally analogous proteins are represented in the same colour. *AspA B.S.* is the palindromic binding site for the AspA protein upstream of the cassette.

The protein encoded by *orf45/parB* was found to comprise two distinct domains separated by a flexible linker: an N-terminal domain which shares homology with bacterial ParB proteins, and a C-terminus which is structurally similar to a eukaryotic centromere-specific histone variant, CenpA (Schumacher *et al*. 2015). The N-terminal domain of a chromosomal ParB homologue from *S. solfataricus* 98:2 was used for crystallographic analysis. The ParB-N structure was found to be similar to that of the bacterial *Thermus thermophilus* ParB, Spo0J. However, crucially, whereas Spo0J harbours a HTH motif consistent with its role as a CBP, the *S. solfataricus* 98:2 ParB-N does not, indicating that it performs a role unrelated to DNA-binding. Instead, ParB-N was demonstrated to interact with AspA, and ParB-C was found to bind to non-specific DNA, whereas the flexible linker was implicated in ParA binding. These findings indicated that ParB may therefore function as an adaptor protein within the partition complex (Schumacher *et al*. 2015).

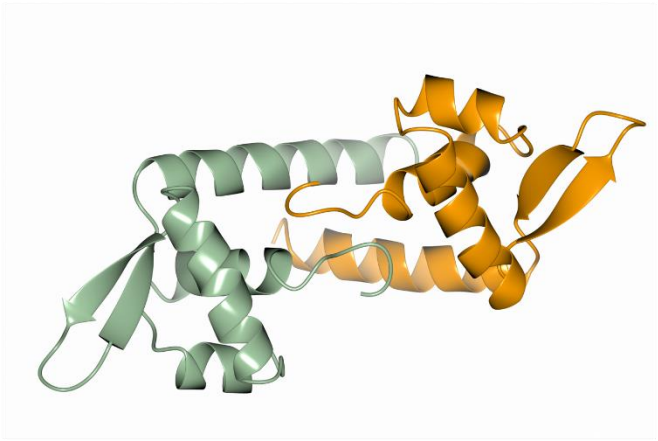
Interestingly, the recently-discovered CTPase activity of bacterial ParB proteins described earlier may be potentially shared by pNOB8 ParB. The structure of the N-terminal domain of pNOB8 ParB superimposes closely with ParB structures from *B. subtilis* and *Myxococcus xanthus* (Soh *et al*. 2019, Osorio-Valeriano *et al.* 2019). It appears that the CTP-binding pocket/motif is highly conserved and present in pNOB8 ParB, therefore it is possible that this archaeal protein also possesses CTPase activity like bacterial ParBs.

The third protein in the partition system, encoded by *orf46*, was found to possess a very similar structure to that of bacterial Walker-box ATPases such as ParF, and was thus named ParA. ParA demonstrated nucleotide binding properties, and though the protein formed dimers in both apo- and nucleotide-bound states, binding to ATP induced a conformational change optimal for ATP hydrolysis. ParA also demonstrated the capability to bind non-specific DNA in an ATP-dependent manner, a property of bacterial ParA proteins (Schumacher *et al*. 2015).

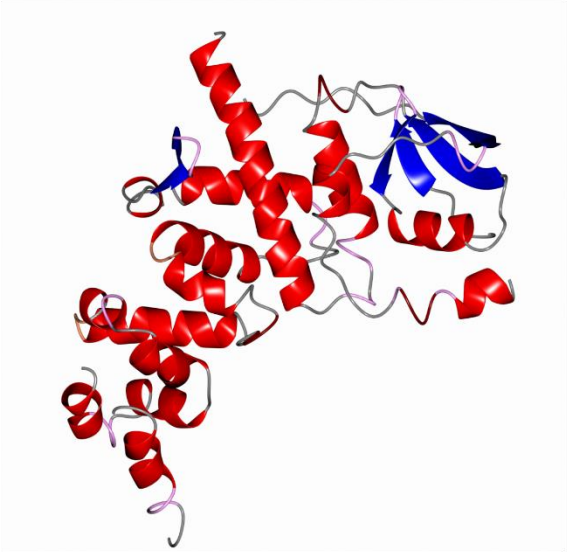
Thus, the pNOB8 segregation system, comprising AspA, ParB and ParA, fuses genome partitioning elements from both bacterial and eukaryotic lineages: ParA contains a bacterial Walker-box fold, whilst ParB comprises two separate domains, one homologous to bacterial proteins (ParB-N), and the other structurally similar to proteins involved in eukaryotic DNA segregation (ParB-C, Schumacher *et al.* 2015). The crystal structures of the three pNOB8 partition proteins, AspA, ParB and ParA are shown below in **Figure 1.18**.

Figure 1.18. Structures of pNOB8 partition proteins (following page). (A) (Top) The apo-AspA dimer. One monomer is in green, the other in orange (PDB code 4RS8). (Bottom) The AspA-DNA structure, showing three AspA dimers (dark green/blue, red/yellow, sea green/pink) in complex with a 32-mer DNA (PDB code 5K1Y). **(B)** (Top) ParB-N structure, coloured according to secondary structure elements: alpha-helices are red, beta sheets are blue. The structure is derived from a chromosomal ParB homologue from *S. solfataricus* 98:2 (PDB code 5K5A). (Bottom) The ParB-C dimer, with one monomer shown in sea green, the other in dark purple (PDB code 4RS7). **(C)** (Top) A superposition of pNOB8 ParA and ParF monomers from TP228, demonstrating the structural similarity of the proteins. ParA is coloured light blue, ParF is maroon. (Bottom) The ParA-dimer bound to the ATP analog adenylyl-imidodiphosphate (AMP-PNP). One monomer is coloured light blue, the other coral, and AMP-PNP molecules are shown as cylinders (PDB code 5K5Z). All figures were produced using CCP4MG.

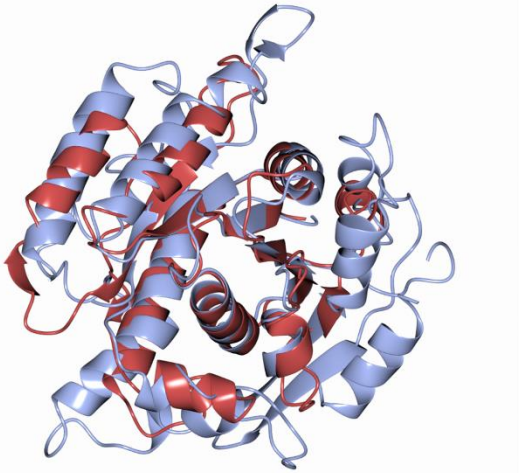
(A) AspA dimer



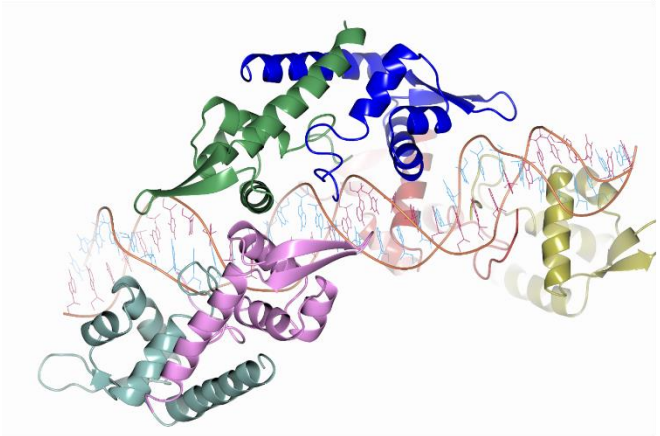
(B) ParB-N



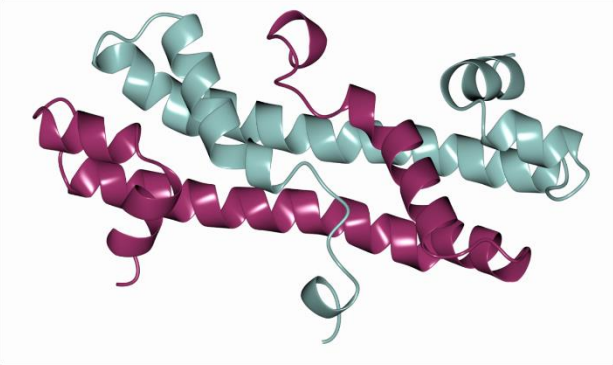
(C) ParA/ParF superposition



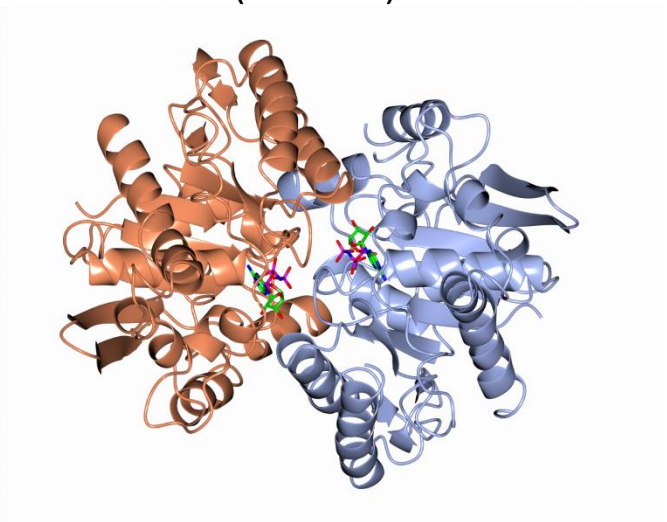
AspA-DNA complex



ParB-C dimer



ParA-(AMP-PNP) dimer



1.7.5 Plasmid-host interactions and archaeal CRISPR systems

An additional property of the conjugative plasmid pNOB8 is its ability to integrate into, and excise from the NOB8-H2 host chromosome, via an integrase-dependent mechanism. Episodes of plasmid insertions and excisions could therefore provide a means for horizontal gene transfer and evolution of the genome (She *et al*. 2004). Plasmids can be thought of as invasive genetic elements, as they often impart a burden on the host cell, therefore flux between plasmid and host chromosome and any resulting genetic changes may be one mechanism employed by plasmids such as pNOB8 to avoid host immunity (Wang *et al*. 2015). Alternatively, the stable maintenance of the pNOB8 in its extrachromosomal form could indicate some intrinsic property of the plasmid that aids evasion of host defence mechanisms.

Archaea, like bacteria, harbour CRISPR-Cas (Clustered Regularly Interspaced Short Palindromic Repeats and CRISPR-associated proteins) defence systems against invading genetic elements such as viruses and plasmids (Gudbersdottir *et al.* 2011, Lillestøl *et al*. 2006). Indeed, it is estimated that CRISPR loci are found on 80% of archaeal genomes (Makarova *et al*. 2015b). CRISPR-Cas systems have been extensively reported in Crenarchaea including *Sulfolobus* spp., with almost all species harbouring (alongside a module for spacer acquisition) both a Type I and a Type III system, which in some cases grants interference against both DNA and RNA (Makarova *et al*. 2011, Deng *et al*. 2013, Peng *et al*. 2015). To counter the host CRISPR defence mechanisms, invasive elements have responded by evolving anti-CRISPR proteins, which were first identified in bacteriophages, but have more recently been found encoded on bacterial plasmids and archaeal viral genomes (Bondy-Denomy *et al*. 2013, Mahendra *et al*. 2020, Peng *et al*. 2020). Currently, no examples of archaeal plasmid-encoded anti-CRISPRs have been reported. Therefore, sequencing the *Sulfolobus* NOB8-H2 genome will permit a more in-depth characterisation of the interactions between plasmid and chromosome, and possibly shed light on the mechanisms pNOB8 employs to permit its continued existence within the host cell.

1.8. Project Aims

The accurate segregation of newly replicated genetic information is a requirement for all biological organisms. Segregation systems, encoded on bacterial low-copy number plasmids and chromosomes, act to ensure the precise compartmentalisation of the replicated DNA prior to cell division. This process is less well-understood in Domain Archaea, although partition systems analogous to those found in bacteria have been discovered on archaeal chromosomes and plasmids. The low-copy number plasmid pNOB8, which is stably maintained in the thermophilic crenarchaeal strain *Sulfolobus* NOB8-H2, harbours one such segregation system: comprising three proteins, AspA, ParB and ParA, and two centromere-like palindromic DNA sequences. Structural information is available for the three proteins, and how they interact both with each other and with specific and non-specific DNA has been characterised (Schumacher *et al.* 2015). However, the precise molecular mechanism of how this segregation system works to transport pNOB8 within the cell is currently unknown, therefore, this project aims to further elucidate these processes, and increase our understanding of genome segregation in archaea.

Specifically, three main aims will be addressed, each comprising one data chapter of this thesis:

1. AspA interactions at the first palindrome upstream of the partition cassette have previously been characterised. However, there is an identical palindrome elsewhere on pNOB8, and unpublished data established that AspA also binds here *in vitro*. The affinity of AspA to the DNA, and its patterns of binding at the second site will be evaluated, in order to formulate a hypothesis as to the functional role(s) the protein plays at each palindrome. In addition, mutagenesis will be performed to characterise the relationship between specific AspA residues and function.

2. AspA is known to interact with the N-terminus of ParB, which is thought to perform the role of adaptor protein. The crystal structure of ParB-N is based on a

closely-related chromosomal homologue. Therefore, the N-terminal domain of pNOB8 ParB will be delineated using bioinformatic and structure prediction approaches, and a ParB-N construct will be used to define the binding interface with AspA.

3. The plasmid pNOB8 is known to interact dynamically with the NOB8-H2 chromosome. The NOB8-H2 genome will be sequenced in order to further understand the interplay between plasmid and host, with the aim of understanding how pNOB8 is stably maintained within the cell. Additionally, sequencing of the NOB8-H2 chromosome will provide insight into the position this strain occupies in the phylogenetic tree of *Sulfolobus*.

Chapter 2

Materials and Methods

2.1 Bacterial strains and plasmids used

2.1.1 Bacterial strains

Strains of *E. coli* used in this study were obtained from glycerol stocks already present in the laboratory, stored at –80°C. A sterile loop was used to streak out the glycerol stock onto Luria-Bertani (LB) agar, supplemented with antibiotics when required, and grown overnight at 37°C. A single colony was taken from the plate and inoculated into 10 ml LB (+/- antibioitics), and the culture grown overnight at 37°C. Bacterial strains are listed in Table 2.1.

Table 2.1 List of *E. coli* strains used in this study

***E. coli* strain**	**Genotype**	**Application**	**Antibiotic selection**
DH5α	F^- ϕ80*lac*ZΔM15 Δ(*lac*ZYA-*arg*F)U169 r*ec*A1 *end*A1 *hsd*R17(r_K^-, m_K^+) *pho*A *sup*E44 λ^-*thi*-1 *gyr*A96 *rel*A1	Cloning, plasmid mini-prep, glycerol stock storage.	-
BL21(DE3) Codon Plus	F^- *ompT hsdS*(r_B^- m_B^-) *dcm*$^+$ Tet^r *gal* λ(DE3) end*A* Hte [*argU ileY leuW* Cam^r]	Protein overproduction	Chloramphenicol

2.1.2 Plasmids

Plasmids used during this work are described below in Table 2.2. Plasmids were either available in the laboratory, or constructed during this study.

Table 2.2 List of plasmids used in this study

Plasmid	**Description**	**Selection**	**Source**
pET-22b(+)	Overexpression plasmid containing the T7 promoter, a MCS, and a C-terminal 6xHis tag.	Ampicillin	Novagen
pET-Orf44 (AspA)	pET-22b(+) containing the wild-type *aspA* gene cloned between *Xho*I and *Nde*I restriction sites. C-terminal 6xHis tag.	Ampicillin	Barillà lab
pET-AspA-R49A	pET-22b(+) containing the *aspA-R49A* mutant gene cloned between *Xho*I and *Nde*I restriction sites. C-terminal 6xHis.	Ampicillin	Barillà lab

pET-AspA-A53K	pET-22b(+) containing the *aspA-A53K* mutant gene cloned between *Xho*I and *Nde*I restriction sites. C-terminal 6xHis.	Ampicillin	Barillà lab
pET-AspA-Y41A	pET-22b(+) containing the *aspA-Y41A* mutant allele. C-terminal 6xHis tag.	Ampicillin	This study
pET-AspA-Q42A	pET-22b(+) containing the *aspA-Q42A* mutant allele. C-terminal 6xHis tag.	Ampicillin	This study
pET-AspA-L52K	pET-22b(+) containing the *aspA-L52K* mutant allele. C-terminal 6xHis tag.	Ampicillin	This study
pET-AspA-E54A	pET-22b(+) containing the *aspA-E54A* mutant allele. C-terminal 6xHis tag.	Ampicillin	This study
pET-AspA-L12G	pET-22b(+) containing the *aspA-L12G* mutant allele. C-terminal 6xHis tag.	Ampicillin	This study
pET-AspA-I85G	pET-22b(+) containing the *aspA-I85G* mutant allele. C-terminal 6xHis tag.	Ampicillin	This study
pET-AspA-V89G	pET-22b(+) containing the *aspA-V89G* mutant allele. C-terminal 6xHis tag.	Ampicillin	This study
pET-AspA-I89GV89G	pET-22b(+) containing the *aspA-I89GV89G* double mutant allele. C-terminal 6xHis tag.	Ampicillin	This study
pET-ParB	pET-22b(+) containing *parB* cloned between *Xho*I and *Nde*I restriction sites. C-terminal 6xHis tag.	Ampicillin	Barillà lab
pET-ParB-N	pET-22b(+) containing the region that encodes for the ParB N-terminus cloned between the *Xho*I and *Nde*I restriction sites. C-terminal 6xHis tag.	Ampicillin	This study
pET-ParB-N & linker	pET-22b(+) containing the region that encodes for N-terminus plus flexible linker region of ParB cloned between the *Xho*I and *Nde*I restriction sites. C-terminal 6xHis tag.	Ampicillin	Barillà lab

pUC18	Cloning plasmid containing the *lac* promoter, and a MCS located within the *lacZ* gene.	Ampicillin	Barillà lab
pJA1-200	pUC18 containing a 200 bp fragment from plasmid pNOB8 harbouring the second *aspA* binding site, cloned between the *Pst*I and *Eco*RI restriction sites.	Ampicillin	This study
pJA2-1.7	pUC18 containing a 1.68 kbp fragment from plasmid pNOB8 harbouring both *aspA* binding sites, cloned between the *Pst*I and *Eco*RI restriction sites.	Ampicillin	This study

2.2 Media and antibiotics used

2.2.1 Luria-Bertani (LB)

Strains of *E. coli* used in this study were grown in liquid LB broth (Fisher Scientific) or on solid LB Agar (Fisher Scientific), supplemented with antibiotics when required. Media were prepared according to the manufacturer's instructions by adding a specific amount of powder to a particular volume of distilled water, e.g. 7.5 g LB broth per 300 ml water. Media was sterilised by autoclave at 121°C. Antibiotics and/or induction compounds were added after the media had sufficiently cooled. LB agar plates were poured in a laminar flow hood, pre-sterilised using 70% ethanol. The composition of LB media is listed in Table 2.3.

Table 2.3 Luria-Bertani composition

Compound	Concentration (g/L)
Tryptone	10
Yeast extract	5
Sodium Chloride	10
Agar (solid medium)	12

2.2.2 Brock's medium

Strains of *Sulfolobus* NOB8-H2 used in this study were grown in the following medium, after Brock (Brock *et al.* 1972). Individual components were sterile filtered using a 0.22 µm filter, and Milli-Q water was autoclaved prior to addition. The pH of the final medium was adjusted by adding H_2SO_4.

Table 2.4 Brock's medium composition

Stock component and concentration	**Volume added per 400 ml (final concentration)**
Brock I (1000 X)	0.4 ml (1X)
Brock II (100 X)	4 ml (1X)
Brock III (200 X)	2 ml (1X)
20% Sucrose or glucose	4 ml (0.2%)
20% Tryptone	4 ml (0.2%)
2% Iron (II) chloride solution	0.8 ml (0.004%)
Brock components	**(per 100 ml Milli-Q water)**
Brock I	7 g $CaCl_2.2H_2O$
Brock II	13 g $(NH_4)_2SO_4$ 2.5 g $MgSO_4.7H_2O$ 150 µl 1:2 H_2SO_4
Brock III	5.6 g KH_2PO_4 10 ml trace elements 150 µl 1:2 H_2SO_4
Trace elements (1000 X)	Per 1 L Milli-Q water: 1.8 mg $MnCl_2$ 0.22 mg $ZnSO_4$ 0.05 mg $CuCl_2$ 0.03 mg $VOSO_4$ 0.01 mg $CoSO_4$ 4.5 mg $Na_2B_4O_7$ 0.03 mg $NaMoO_4$

2.2.3 Antibiotics

Antibiotic stock solutions were prepared by dissolving the required amount of powder in suitable solvent, filter sterilising with 0.22 µm filters (Millipore), then storing aliquots at –20°C until use. Antibiotics used in this study are shown in Table 2.5.

Table 2.5 Antibiotics used and relevant concentrations

Antibiotic	Solvent	Stock conc. (mg/ml)	Working conc. (µg/ml)
Ampicillin	Milli-Q water	100	100
Chloramphenicol	100% ethanol	34	34

2.3 Recombinant DNA techniques

2.3.1 Preparation of competent cells

E. coli strains DH5α and BL21(DE3)CodonPlus are required to be chemically competent for DNA uptake prior to transformation. Non-competent cells were streaked from glycerol stocks onto LB agar plates, and a single colony inoculated into 10 ml LB (plus antibiotics if necessary) and incubated overnight at 37°C with shaking at 200 rpm. 0.3 ml of the overnight culture was inoculated into 60 ml LB broth, and grown at 37°C with shaking until the optical density at a wavelength of 550 nm (OD_{550}) was between 0.4 and 0.6. After reaching the required OD_{550} value, the culture was incubated on ice for 10 minutes. The culture was split into two 50 ml Falcon tubes, and centrifuged at 7,000 rpm for 5 minutes at 4°C to harvest the cells. The supernatant was discarded, and cells resuspended in 18 ml of buffer RF1 (Table 2.6), equivalent to 1/3 of the culture volume. The resuspended cells were incubated on ice for 1 hour. The cells were again pelleted by centrifugation at 7,000 rpm for 5 minutes at 4°C. The supernatant was discarded, and cells were resuspended in 4.2 ml of buffer RF2 (Table 2.6), equivalent to 1/12.5 of the culture volume. After incubation on ice for 15 minutes, the competent cells were aliquoted and stored at –80°C. Cells were tested for competence by DNA transformation.

Table 2.6 Competent cell preparation buffers

Compound	RF1 (pH 5.8)	RF2 (pH 6.8)
Glycerol	15%	15%
RbCl	100 mM	10 mM
$MnCl_2$	50 mM	-
CH_3CO_2K	30 mM	-
$CaCl_2$	10 mM	75 mM
MOPS	-	10 mM

2.3.2 Bacterial transformation

Competent *E. coli* strains DH5α or BL21(DE3)CodonPlus were thawed on ice. An aliquot of 100 µl of competent cells was used for each transformation. 1 µl of plasmid DNA was added to the aliquot of competent cells, and an additional 100 µl aliquot without DNA was used as a negative control. The cells plus plasmid mixture was incubated on ice for 40 minutes, then subjected to heat shock by incubation at 42°C for 90 seconds in a heat-block. The cells were then placed back on ice for 2 minutes, 400 µl of LB added, and the tube incubated at 37°C for 1 hour with shaking, to allow expression of the plasmid-encoded antibiotic resistance gene and subsequent translation into protein. Typically, 100 µl of cells were then spread on LB agar plates, plus relevant antibiotics, and incubated overnight at 37°C.

2.3.3 Plasmid DNA extraction

E. coli DH5α cells were transformed with plasmid DNA as described above. One transformant colony was inoculated into a suitable volume (usually 5 or 10 ml) of sterile LB, supplemented with the relevant antibiotic. The inoculum was grown overnight at 37°C with shaking at 200 rpm. The cells were harvested the next day by centrifugation at 12,000 rpm for 5 minutes. The plasmid DNA was extracted using the Macherey-Nagel NucleoSpin® Plasmid miniprep kit, according to the manufacturer's instructions. Briefly, the cell pellet is resuspended, cells are subjected to alkaline lysis, and the lysate neutralised to precipitate proteins, chromosomal DNA and cell debris. After centrifugation, the lower molecular weight plasmid DNA remains in the supernatant, and this is extracted by passing over a silica membrane. Isolated plasmid DNA is then eluted into elution buffer or Milli-Q water via centrifugation, and plasmid DNA is stored at -20°C.

2.3.4 Primer design

All forward and reverse primers used in this study are listed in Table 2.7. Primers were used to amplify a required section of DNA via PCR, using a DNA template (either genomic DNA extract or plasmid miniprep). Each oligonucleotide primer was between ~20 and ~30 nucleotides, depending on downstream application. Primers used for cloning purposes included the relevant restriction endonuclease site at either 5' or 3' end, plus an additional 6 bp 'tail', 5' of the restriction sequence. Primers for amplification of DNA for use in EMSA assays were biotinylated at the 5' end. Primer sequences were input into the Sigma-Aldrich OligoEvaluator™ site, to assess parameters such as: melting temperatures, GC%, run length of any repeated bases, and secondary structure and primer dimer formation. Primers were manufactured by Sigma-Aldrich/Merck, and lyophilised primers were resuspended in sterile Milli-Q water to a stock concentration of 100 µM.

Table 2.7 List of primers used in this study

Primer name	Sequence (5' to 3')
AspA BS2 247 bp F	[Btn]CTCTGGACTTACTTTGAATA
AspA 200 bp R	CTTAATGTTCTCCGACATA
200 bp cloning F	AAGGAGCTGCAGTAATACGTAAAAAACTGA
200 bp cloning R	ATATATGAATTCCTTAATGTTCTCCGACATA
1.68 kb cloning F	AAGGAACTGCAGCCTTCAGATAATACGTAA
1.68 kb cloning R	ATATATGAATTCGCTTTAGCCTTACCTACC
AspA-Y41A F	CACACAGATCCCAGCTCAAACCGTAATACAG
AspA-Y41A R	CTGTATTACGGTTTGAGCTGGGATCTGTGTG
AspA-Q42A F	CACAGATCCCATATGCAACCGTAATACAGAAT
AspA-Q42A R	ATTCTGTATTACGGTTGCATATGGGATCTGTG
AspA-L52K F	GAATATTAGGTGGTTAAAAGCTGAAGGATATGTAG
AspA-L52K R	CTACATATCCTTCAGCTTTTAACCACCTAATATTC
AspA-E54A F	GGTGGTTACTAGCTGCAGGATATGTAGTAAAAGAGC
AspA-E54A R	GCTCTTTTACTACATATCCTGCAGCTAGTAACCACC
AspA-L12G F	ACAAATACATCTTCGGAACTCCTAGAGCATA
AspA-L12G R	TATGCTCTAGGAGTTCCGAAGATGTATTTGT
AspA-I85G F	CGGAACTAGAAAAAGGTAGAAAATTAGTAGA
AspA-I85G R	TCTACTAATTTTCTACCTTTTTCTAGTTCCG

AspA-V89G F	AAATTAGAAAATTAGGAGAGGTGGTTCAATG
AspA-V89G R	CATTGAACCACCTCTCCTAATTTTCTAATTT
ParB-FL F / ParB-N cloning F	AAGGAACATATGAGTAAGCTGAAAGAGTAT
ParB-FL R	AAGGAACTCGAG TAACTTCCCCTCCAAGAC
ParB-N cloning R	AAGGAACTCGAGCCTCTGCAGTTTCTCTAG
126 bp BS2 F	AGGTTCTCTTTACGTAAC
126 bp BS2 R	CCCTCATATTATGCTCTA
1.6 middle F#2	TTACGAGATCCACTCATCTT
1.6 middle R#2	GGTTAACAATCTAATTGAGGC
C45 pNOB8 1 F	CTCTTTGTTGCGCTGCTCTTC
C45 3 R	TCCTGGAGGCTTCCTGCTTC
P1 Ref 2 F	AGATCGAAGTAACTGGCGGAC
C45 pyrE F	AGTAGGAATAGCCACTGGAG
C45 pyrE R	CCTCCACCGTTAAGAATCTC
pKEF9 Orf153 F	ACGTGCTTATCGTCTCCCAA
pKEF9 Orf153 R	TGCCAGAGAAAGTAAGGTCT
T7 Promoter F	TAATACGACTCACTATAGGG
T7 Terminator R	GCTAGTTATTGCTCAGCGG
M13 pUC18 F	CCCAGTCACGACGTTGTAAAACG
M13 pUC18 R	AGCGGATAACAATTTCACACAGG

2.3.5 Polymerase chain reaction (PCR)

DNA sequences were amplified by polymerase chain restriction (PCR) using an Eppendorf Mastercycler thermal cycling machine. A typical PCR reaction mixture, using GoTaq G2 polymerase, is outlined below in Table 2.8. Alternative DNA polymerases were used depending on the application. The total reaction volume was 60 µl, and was prepared on ice. A working solution of each deoxynucleotide triphosphate (dNTP) at 5 mM was prepared by diluting each dNTP from 100 mM stocks (Roche) in sterile Milli-Q water. Template DNA was typically added to a final mass of 60 ng.

Table 2.8 Components of a typical PCR reaction

Reaction Component	Volume	Final concentration/amount
5X GoTaq buffer	12 µl	1X
DNA template	X µl	60 ng
dNTPs (5 mM)	2.4 µl (each dNTP)	200 µM
Forward primer (5 µM)	3 µl	15 pmol
Reverse primer (5 µM)	3 µl	15 pmol
Sterile Milli-Q water	to 60 µl	-
GoTaq G2 polymerase (5 U/ µl)	0.5 µl	2.5 U

PCR program parameters were altered depending on the calculated melting temperatures of primers, the size of the DNA amplicon, and the DNA polymerase used. The number of repeated cycles of denaturation, annealing and extension was generally 30 cycles. A typical PCR program is outlined in Table 2.9.

Table 2.9 Typical PCR thermocycler program settings

Program step	Temperature (°C)	Time (minutes)
Initial Denaturation	95	3
Denaturation	95	1
Annealing	42	1
Extension	72	1
(Repeat x 29)		
Final Extension	72	6
Hold	10	-

2.3.6 Restriction endonuclease digest

Restriction enzyme (RE) digests were used primarily when cloning a DNA fragment insert into the multiple cloning site (MCS) of a recipient vector. RE digests produced linearized vector, inserts with complementary digest sites, and were also used in diagnostic digests to confirm the inserts were cloned into the vector correctly. Typical RE digest reactions using two different enzymes were performed in a total volume of 30 µl, as shown in Table 2.10. RE buffers were supplied by the manufacturer, and in the case of a double

digest, a buffer suitable for both enzymes was used. Reactions were typically incubated at 37°C for 3 hours, and heat inactivated for 20 minutes if necessary.

Table 2.10 Typical restriction enzyme reaction components

Reaction component	Volume (µl) and amount
Restriction enzyme 1	1 (10 U)
Restriction enzyme 2	1 (10 U)
Buffer (e.g. 10X)	3
DNA	20 (500 ng)
Milli-Q water	5
Total	30

2.3.7 Ethanol precipitation

DNA purification using ethanol (EtOH) precipitation was used either to remove potential contaminants from the DNA, or to concentrate the DNA following another process e.g. restriction enzyme digestion, before use in downstream applications. The volume of DNA (resuspended in e.g. Milli-Q water) was supplemented with 0.1 volumes of 3 M sodium acetate pH 5.2 and 2.5 volumes of ice-cold 100% ethanol. 0.05 volumes of glycogen were added to visualise the pellet unless downstream applications included sequencing. The components were gently mixed, and incubated at either -20°C overnight, or -80°C for 2 – 4 hours. The precipitated DNA was collected by centrifugation at 13,000 rpm at 4°C for 20 minutes. The supernatant was carefully withdrawn, and the pellet resuspended in 500 µl of ice-cold 70% ethanol. The pellet was again centrifuged at 13,000 rpm at 4°C for 20 minutes. The resuspension in 70% ethanol and centrifugation were repeated a second time. The supernatant was carefully removed, and the pellet left to air dry, usually for 5 – 10 minutes at 37°C to remove residual traces of ethanol. The pellet was then resuspended in an appropriate volume of Milli-Q water and stored at -20°C until required.

2.3.8 Cloning protocol overview

Cloning of a gene, or partial gene, into a suitable expression plasmid was a multi-step process, involving PCR amplification of the gene of interest, restriction enzyme digestion of the recipient plasmid and ligation of the amplified fragment into the plasmid. This was followed by a diagnostic restriction digest, colony PCR to screen for positive transformants and sequencing of selected clones to verify the correct insertion (and orientation) of the gene of interest into the expression plasmid. The following is an outline of the process used to clone *parB-N* into pET-22b(+), and is representative of other cloning experiments conducted during this study.

2.3.8.1 Restriction digest of pET-22b(+) and PCR amplification of *parB-N*

The plasmid pET-22b(+) was digested with *Xho*I and *Nde*I restriction enzymes as detailed in Section 2.3.6. This produced a sufficient quantity of linearised plasmid (~2 µg) to use in the following stages. Concurrently, the section of the gene encoding the N-terminus of ParB was amplified by PCR as outlined in Section 2.3.5. Here, however, Phusion DNA polymerase (ThermoFisher Scientific) was used due to its higher fidelity compared to other polymerases (~50 X *Taq* polymerase). 1 U of Phusion was used in the reaction. A range of different DNA templates were used, typically either a genomic DNA extract, or a large-scale plasmid prep, both of which contain the gene of interest. The annealing temperature and extension time were optimised to give the greatest amount of PCR product. The *parB-N* PCR product was run on an agarose gel and the correct band was excised and purified as detailed in Sections 2.3.11 and 2.3.12. The purified *parB-N* fragment was then also digested with *Xho*I and *Nde*I restriction enzymes as detailed in Section 2.3.6., and purified again by ethanol precipitation as outlined in Section 2.3.7.

2.3.8.2 Alkaline phosphatase treatment of DNA

The digested, linearised pET-22b(+) was subjected to alkaline phosphatase treatment to de-phosphorylate the plasmid by removal of the 5’ phosphate, to prevent its re-circularisation. A typical reaction involved the digested plasmid (20 µl), 10X alkaline phosphatase buffer (10 µl), with the reaction adjusted to a total volume of sterile Milli-Q water. 0.5 µl of Antarctic phosphatase (NEB, 5 U/µl) was added to the reaction, and the mixture incubated at 37°C for 30 minutes. Another 0.5 µl of phosphatase was added, and

the mixture incubated again at 37°C for 30 minutes. Then, 10 µl of 200 mM ethylene glycol-bis(β-aminoethyl ether)-N,N,N'N'-tetraacetic acid (EGTA) was added to the reaction, and the mixture incubated at 75°C for 10 minutes. The linearised, de-phosphorylated plasmid was then purified by ethanol precipitation as outlined in Section 2.3.7.

2.3.8.3 DNA ligation

The digested, linearised pET-22b(+) plasmid backbone and the digested *parB-N* fragment were then combined in a ligation reaction to insert the fragment between the *Xho*I and *Nde*I restriction sites. A typical reaction used between 1:1 and 1:5 molar ratios of vector to insert (e.g. 100 ng vector and 50 ng insert gave a 1:3 molar ratio). T4 DNA ligase (NEB, 5 U/µl) was used to ligate the insert into the linearised plasmid backbone. Two control reactions were set up in parallel; one without the insert, and the second without either the insert or T4 ligase, to test both the phosphatase efficiency and restriction digest efficiency respectively. A typical set of ligation reactions are shown below in Table 2.11. The reactions were incubated at room temperature for 3 hours, then heat-inactivated by placing the reactions at 65°C for 20 minutes. The ligation reactions, including controls, were then transformed into *E. coli* strain DH5α as detailed in Section 2.3.2.

Table 2.11 Components of a typical DNA ligation reaction

Reaction Component	Real ligation (3:1)	No insert	No insert/no ligase
Insert	X µl (50 ng)	-	-
Vector	Y µl (100 ng)	Y µl (100 ng)	Y µl (100 ng)
10X T4 DNA ligase buffer	3 µl	3 µl	3 µl
Sterile Milli-Q water	to 30 µl	to 30 µl	to 30 µl
T4 DNA ligase (5 U/ µl)	1 µl	1 µl	-

2.3.8.4 Colony PCR and diagnostic restriction digest

Colony PCR is a technique used to screen for positive clones, i.e. those that harbour the cloned insert. Here, a PCR master mix was prepared in a single tube, containing enough reagents for 20 reactions, then divided into 20 separate PCR tubes. The primers used were the same as initially used to amplify the insert. The transformant colonies from the previous step were used to provide the DNA template. A sterile pipette tip was used to

take a small amount from a single colony, touched first to a separate LB agar plate (plus antibiotics), then swirled in the PCR reaction mix to transfer the remaining bacterial transformants to the tube. A control reaction was set up in parallel by taking a colony from the 'No insert' agar plates. The PCR program was the same as initially used to amplify the insert. Colony PCR products were loaded onto an agarose gel as detailed in Section 2.3.9., and those colonies that contained the correct size insert were used to grow overnight cultures before extracting the plasmid as outlined in Section 2.3.3. As an additional verification, the plasmids were then digested with the same restriction enzymes originally used for the cloning, and the digested products run on an agarose gel. Those plasmids positive for the insert after digestion were sent for sequencing to confirm that the cloning was successful, and that the insert contained no mutations.

2.3.9 Agarose gel electrophoresis

Gel electrophoresis was used as a diagnostic tool to verify the success of e.g. PCR reactions, restriction enzyme digests, plasmid minipreps etc, or used to separate the correct size DNA fragment before gel extraction and purification. The percentage of agarose used (w/v) varied according to the size(s) of the DNA fragments being loaded, but typically was between 1 and 2%. The agarose gel was prepared by dissolving the appropriate mass of powder in 50 ml (or 100 ml for larger gel casting trays) of 1X TAE buffer (40 mM Tris-HCl, 20 mM acetic acid, 1mM ethylenediaminetetraacetic acid (EDTA)). Once dissolved, SYBER Safe DNA gel stain (Invitrogen) was added at a 1 in 10,000 fold dilution, the agarose was poured into a casting tray, an agarose gel comb was inserted, and the gel was left to set for ~30 minutes. The DNA sample to be loaded was mixed with 1X final concentration of DNA loading dye (NEB), and loaded into the well. An appropriate molecular weight DNA marker ladder was also loaded into another well to enable the DNA fragment size to be accurately estimated. Typically, a GeneRuler 1 kb (Thermo Fisher Scientific) or PCR marker (NEB) was used. The gels were loaded whilst submerged in running buffer in the tank, which was typically 1X TAE buffer (EMSA agarose gels were made and run using 1X TBE buffer (90 mM Tris-HCl, 90 mM Boric Acid, 2 mM EDTA)), and run at 100 V for an appropriate time until the DNA fragment(s) had run through ¾ of the gel. DNA agarose gels were visualised using a Bio-Rad Gel Doc EZ imager, or on an ultraviolet transilluminator (UV) if band excision was required.

2.3.10 DNA extraction from agarose gels and purification

DNA fragments such as PCR products that were to be used in downstream experiments such as cloning required extraction from an agarose gel. After running the DNA on an agarose gel as outlined in Section 2.3.9, the DNA fragments of interest were visualised on a UV transilluminator. A sterile scalpel was used to excise the band, taking care to remove only the band and as little of the surrounding agarose as possible. The gel slice was placed into a sterile, pre-weighed Eppendorf tube. The DNA was purified from the agarose using a Macherey-Nagel PCR clean-up and gel extraction kit. The gel slice was first weighed, then 2 volumes of NTI buffer were added to the tube. The tube was incubated at 50°C for ~10 minutes, with periodic vortexing, until the gel slice was completely dissolved. The solution was transferred to a NucleoSpin column and centrifuged at 12,000 x g for 1 minute. The flow-through was discarded, and the column washed twice with 700 µl of buffer NT2 (containing ethanol) with centrifugation at 12,000 x g for 1 minute. The flow-through was discarded and any residual ethanol removed by centrifugation at 12,000 x g for 2 minutes. The NucleoSpin column was placed into a new, sterile Eppendorf tube, and the DNA was eluted by adding an appropriate volume (typically 20 – 50 µl) of Milli-Q water or TE buffer, incubating at room temperature for 2 minutes, then centrifuging at 12,000 x g for 1 minute. The extracted, purified DNA was assessed if required by visualising a small aliquot on an agarose gel.

2.3.11 Site-directed mutagenesis using the QuikChange system

Site-directed mutagenesis was performed to generate the *aspA* mutants described in this study, by substitution of specific DNA base(s) that encode the target amino acid. The QuikChange Lightening Site-Directed Mutagenesis Kit (Agilent Technologies) was used to construct mutants. Briefly, the principle of this technique is to use a wild-type parental plasmid as the DNA template. The template used was pET-22b(+) containing the wild-type *aspA* gene cloned between the *Xho*I and *Nde*I restriction sites. This plasmid is denatured, and primers containing the mutation anneal and are extended via DNA synthesis as in a normal PCR reaction. Forward and reverse primers for this protocol are typically designed to be ~30 bp in length, with the mutation located centrally in each primer. The PCR reaction then results in a new mutant plasmid, which is present alongside the parental

wild-type plasmid. Then the restriction enzyme *Dpn*I is used to digest the methylated or hemimethylated parental plasmid DNA, leaving behind the mutated plasmid. The mutant plasmid is then transformed into the provided ultracompetent cells (XL10-Gold), purified, and sent for sequencing to verify the mutation is present. A typical PCR reaction for the mutagenesis protocol, and the thermocycler settings, are detailed below in Tables 2.12 and 2.13.

Table 2.12 Components of a typical QuickChange mutagenesis PCR reaction

Reaction Component	**Volume**	**Final concentration/amount**
10X reaction buffer	5 µl	1X
DNA template	X µl	50 ng
dNTP mix (proprietary)	1 µl	-
Forward primer (5 µM)	3 µl	15 pmol
Reverse primer (5 µM)	3 µl	15 pmol
Quick solution	1.5	
Sterile Milli-Q water	to 50 µl	
QuikChange Lightning polymerase (proprietary)	1 µl	-

Table 2.13 Typical QuikChange mutagenesis PCR thermocycler program settings

Program step	**Temperature (°C)**	**Time**
Initial Denaturation	95	2 mins
Denaturation	95	20 secs
Annealing	60	10 secs
Extension	68	3 mins (30 secs/kb)
(Repeat x 17)		
Final Extension	68	5 mins
Hold	4	-

After the PCR reaction is complete, 2 µl of proprietary *Dpn*I was added, and the reaction incubated for 10 minutes at 37°C to digest the wild-type parental plasmid. At the same time, ultracompetent cells (*E. coli* XL10-Gold strain) were thawed on ice, with 45 µl of cells used for each reaction added to a prechilled 15 ml Falcon tube. 2 µl of the supplied

β-mercaptoethanol was added to the cells, to aid transformation efficiency. The cells were incubated for 2 minutes on ice, before the addition of 2 µl of the PCR reaction containing the mutant plasmid, followed by a further incubation on ice for 30 minutes. The suggested medium for the transformation is NZY^+ broth, which was previously prepared (per Litre; 10 g casein hydrolysate, 5 g yeast extract, 5 g NaCl, 12 mM $MgCl_2$, 12 mM $MgSO_4$, 0.4% (w/v) glucose, pH 7.5). The NZY^+ broth was preheated to 42°C using a water bath, then the ultracompetent cells were transformed by heat-shock at 42°C for exactly 30 seconds. The cells were then incubated on ice for 2 minutes, 0.5 ml of preheated NZY^+ broth was added, and the tubes incubated at 37°C for 1 hour with shaking at 225 rpm. 100 µl of the cells were spread onto LB agar plates containing ampicillin and incubated overnight at 37°C. A number of colonies were then selected for overnight growth in selective media and plasmid isolation, followed by sequencing to confirm the presence of the mutation and absence of additional changes.

2.3.12 DNA sequencing

DNA sequencing was used to verify mutagenesis and cloning. Sanger sequencing was performed by Eurofins Genomics, Germany, after following the suggested samples preparation guidelines. Aliquots of 10 µl of sample were sent in a barcoded 1.5 ml Eppendorf tube. The sample consisted of 5 µl of plasmid DNA (at 50 – 100 ng/µl), plus 5 µl of the appropriate primer (at 5 µM). DNA sequence traces were analysed using SnapGene software.

2.4 Protein production and related techniques

2.4.1 Gene overexpression and protein overproduction

The 6xHis-tagged recombinant AspA and ParB proteins used in this study were all overproduced using the same methodology. The expression plasmid pET-22b(+) containing the gene of interest was first transformed into the *E. coli* overexpression strain BL21 (DE3)CodonPlus, using the standard transformation protocol, and spread onto LB agar plates plus antibiotics. A number of colonies (5 – 10) were taken from the agar plate using a sterile loop, and used to inoculate 15 ml LB broth plus antibiotics in a small conical flask. The cells were grown for two hours at 37°C with shaking, or in some cases overnight if the overexpression was to be carried out the next day. After two hours, the inoculum was decanted into a 2 L conical flask containing 300 ml autoclaved LB broth, plus antibiotics. The culture was grown for 3 – 4 hours until the OD_{550} reached 0.8 - 0.9. A 100 µl aliquot was taken and kept as an uninduced control. Overexpression of the gene of interest from the pET-22b(+) plasmids was then induced by adding 1 mM of Isopropyl β-d-1-thiogalactopyranoside (IPTG), an analogue of allolactose, to the culture. The culture was grown for an additional 3 hours at 37°C with shaking. At hourly intervals, a 100 µl aliquot was taken to test for overproduction levels of the protein. These aliquots, along with the aliquot taken before induction, were centrifuged at 11,000 rpm, 4°C, for 1 minute, the supernatant removed, and the pellets resuspended in 20 µl of binding buffer (see table 2.14). To assess overproduction levels, the resuspended pellet was mixed with 20 µl of 2X SDS loading buffer (Table 2.17) and analysed by SDS-PAGE. The remaining 300 ml culture was removed from 37°C after 3 hours, and split into two centrifuge bottles. The culture was centrifuged at 12,000 x g, 4°C, for 25 minutes in a Sorval high-speed centrifuge. The supernatant was discarded and the pellets stored at -20°C until required.

2.4.2 Protein solubility assay

After overproduction trials of the mutant protein, the remaining cell culture (~14 ml) was pelleted, and resuspended in 1 ml binding buffer (20 mM Tris pH7.5, 500 mM NaCl, 10 mM Imidazole). 14 µl of lysozyme (10 mg/ml was added) and the cells incubated for 15 minutes at 30°C. The suspension was sonicated (6 x 15 secs on, 30 secs off), then centrifuged for 30 minutes at 14,000 RPM, 4°C. 500 µl of the supernatant was removed, the rest discarded, then the pellet was resuspended in 500 µl of binding buffer. 20 µl of the supernatant and pelleted fractions were mixed with 20 µl 2X SDS loading buffer (Table 2.17), denatured, and analysed by SDS-PAGE.

2.4.3 Protein purification by Ni^{2+} affinity chromatography

The 6xHis-tagged recombinant Asp and ParB proteins used in this study were all purified using the same methodology. The two flasks of cell pellets collected in Section 2.4.1 were thawed at room temperature, then each resuspended in 11.25 ml binding buffer (total 22.5 ml binding buffer). A list of all buffers used in this protocol is shown in Table 2.14. 150 µl of lysozyme (10 mg/ml) and 1 tablet of protease inhibitor cocktail (Roche) were added to each 11.25 ml suspension. Tablets were crushed inside a sterile Eppendorf tube using a sterile pipette tip, then dissolved in the cell suspension. Both suspensions were thoroughly mixed, then transferred to a 50 ml Falcon tube. The suspension (~23 ml) was incubated at 30°C for 15 minutes, another 150 µl of lysozyme (10 mg/ml) was added, and the suspension incubated at 30°C for a further 15 minutes. The cells were then lysed by sonication; 7 times for 30 seconds at 40% power, separated by 1-minute pause intervals between each sonication step. The cell lysate was then centrifuged at 10,000 rpm, 4°C, for 40 minutes. Meanwhile, the column used for purification (C10 chromatography column, GE Healthcare) was assembled. 5 ml of 50% His-Bind resin slurries (Merck) were pipetted into the column, and washed with 6 column volumes (CVs) of filtered Milli-Q water (all solutions used in the purification, including the cell lysate, were pre-filtered using a 0.22 µm syringe filter). Solutions were passed over the column using a peristaltic pump, typically at flow rates of 3.5 ml/min. The resin was then charged with 5 CVs of 50 mM $NiSO_4$, then equilibrated with 6 CVs of binding buffer. Prior to passing the cell lysate over the column, a 100 µl aliquot was taken for later analysis.

The cell lysate was passed over the column and re-circulated for 2 – 3 hours to allow an optimal quantity of His-tagged protein to bind to the resin. A 100 µl aliquot of lysate flow-through was taken at the end of the circulation. Then, the column was washed with 6 CVs of binding buffer, followed by 10 CVs of wash buffer. 100 µl aliquots were taken at the end of each stage and kept for later analysis. The protein was then eluted from the column using 3 CVs of elution volume, and 12 x 1 ml fractions were collected. The concentrations of these fractions were estimated via Bradford Assay, using 10 µl aliquots of protein. The six most concentrated fractions were kept for buffer exchange (Section 2.4.4) and the rest discarded. The 100 µl aliquots taken from each stage of the purification were analysed by SDS-PAGE.

Table 2.14 Protein purification buffers used in this study

Buffer	**AspA proteins**	**ParB proteins**
Binding buffer (1X)	20 mM Tris-HCl, pH 7.5 500 mM NaCl 10 mM Imidazole	20 mM Tris-HCl, pH 8.0 500 mM NaCl 10 mM Imidazole
Wash buffer (1X)	20 mM Tris-HCl, pH 7.5 1 M NaCl 60 mM Imidazole	20 mM Tris-HCl, pH 8.0 500 mM NaCl 20 mM Imidazole
Elution buffer (1X)	20 mM Tris-HCl, pH 7.5 500 mM NaCl 400 mM Imidazole	20 mM Tris-HCl, pH 8.0 500 mM NaCl 400 mM Imidazole
Storage buffer	50 mM HEPES, pH 7.5 50 mM KCl	20 mM HEPES , pH 7.0 150 mM NaCl

2.4.4 Buffer exchange

After purification, the AspA and ParB proteins were immediately buffer-exchanged into storage buffer using a 5 ml HiTrap Desalting column (GE Healthcare). The column was initially washed with 5 CVs of sterile, filtered Milli-Q water, followed by equilibration with 5 CVs of storage buffer. Then, 1.5 ml of protein fraction was passed over the column,

followed by 2 ml of storage buffer, collecting the flow-through in volumes of 1 ml. This process was repeated until all protein fractions had been passed over the column. The column was then stripped with 4 CVs of Strip buffer (20 mM Tris-HCl pH 8.0, 100 mM EDTA, 50 mM NaCl), washed with 4 CVs of filtered Milli-Q water, and stored in 20% ethanol at 4°C. The column was reused for the same protein if further purification and buffer exchange was required. The concentration of the buffer-exchanged protein fractions was again measured by Bradford assay, and the six most concentrated were aliquoted into sterile Eppendorf tubes in 100 µl volumes, flash-frozen in liquid nitrogen, and stored at – 80°C until required.

2.4.5.1 Protein concentration measurement by Bradford assay

After purification of a protein, it is necessary to measure its concentration before use in downstream applications. The Bradford protein colorimetric assay was used for this purpose. Here, a standard curve was first constructed using 2 mg/ml Bovine Gamma Globulin (BGG, Thermo Scientific). The BGG was diluted to a stock solution of 0.2 mg/ml in 0.9% NaCl solution. Seven reactions were then prepared in triplicate in a 1ml cuvette, as shown in Table 2.15, with 200 µl of Bradford reagent dye (BioRad) used each time, such that the total reaction volume was 1 ml. All reactions were vortexed and left to stand for 5 minutes. The mean optical densities at 595 nm (OD_{595}) for each BGG concentration were then plotted as a function of BGG concentration. This produced a linear plot, with the equation of the straight line in the form y = mx + c; a typical plot is shown in Figure 2.1. After acquiring the equation of the standard curve, the concentration of the protein of interest was measured by adding 10 µl of protein to 790 µl of Milli-Q water, plus 200 µl of Bradford dye, in triplicate, and again the mean OD_{595} was calculated. This value was then input into the equation above (where y is the OD value), and solving the equation for x gave the concentration of the protein in mg/ml.

Table 2.15 Bradford assay reaction components

BGG 0.2 mg/ml (µl)	Milli-Q water (µl)	Bradford dye (µl)	BGG mass (µg)
0	800	200	0
2	798	200	0.4
5	795	200	1
10	790	200	2
20	780	200	4
30	770	200	6
50	750	200	10

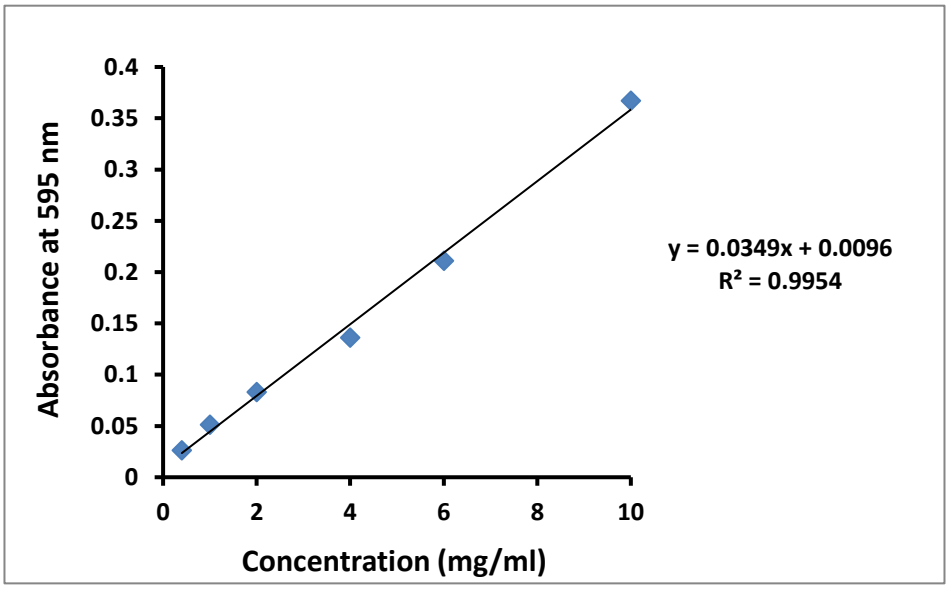


Figure 2.1 Typical Bradford assay standard curve. The equation of the straight line, along with the R^2 value, is shown.

2.4.5.2 Protein concentration measurement by UV spectrophotometry

For some Circular Dichroism experiments, protein concentrations were measured with a Jasco V560 spectrophotometer using 1 cm path length cuvettes at absorbance of 280 nm. The program SEDNTERP3 was used to calculate protein extinction coefficients based on the amino acid sequence, and these data used to calculate the UV absorption-based concentration. These measurements were performed by Dr Andrew Leech.

2.4.6 Sodium Dodecyl Sulphate-Polyacrylamide Gel Electrophoresis (SDS-PAGE)

2.4.6.1 Gel preparation and electrophoresis

SDS-polyacrylamide gels were prepared by mixing the components listed in Table 2.16. Resolving gels were typically 15%, but other percentage gels were used when appropriate. Stock solutions of Tris-HCl and SDS were used, however the ammonium persulphate (APS) was always made fresh on the day of gel casting. Components were added to a 50 ml Falcon tube in the order shown in the table, with APS and tetramethylethylenediamine (TEMED) added immediately prior to casting the gel, as polymerisation starts to occur on their addition. The gel solution was pipetted between the two glass plates, and ~500 µl of isopropanol was pipetted on top of the gel to seal it from the air. The resolving gel typically took ~30 minutes to set, after which the isopropanol was removed by wicking using Whatman filter paper. The stacking gel solution was prepared, pipetted on top of the resolving gel, and a comb was placed into the stacking gel. The gel was left for ~30 minutes to polymerise, after which time the comb was carefully removed. The gel was either used immediately, or kept at 4°C inside paper towels soaked with 1X SDS running buffer, and wrapped in cling film. If using the gel straight away, it was transferred to the gel tank (Mini-PROTEAN Tetra, Bio-Rad) and submerged in 1X SDS running buffer. Running buffer was pipetted into the gel wells to remove excess acrylamide. Before loading, protein samples were mixed with an equal volume of 2X SDS-loading dye (Table 2.17) and denatured by incubation at 95°C for 10 minutes. The volume of sample loaded depended on the size of the glass plates used; typically, this was either 20 or 40 µl. Samples were loaded alongside 5 – 10 µl of PageRuler Plus prestained molecular weight marker (Thermo Scientific). Typically, gel electrophoresis was performed at 150 V for 25 minutes, then 190 V for ~30 minutes, until the molecular weight marker had reached the bottom of the gel.

Table 2.16 Components used to prepare a typical SDS-PAGE gel

Component	**15% resolving gel (10 ml)**	**5% stacking gel (4 ml)**
Milli-Q water	2.3 ml	2.7
30% acrylamide:bisacrylamide (29:1)	5.0	0.67
1.5 M Tris-HCl pH 8.8	2.5	-
1 M Tris-HCl pH 6.8	-	0.5
10% SDS	0.1	0.05
10% APS	0.1	0.05
TEMED	0.004	0.004

Table 2.17 Buffers used in SDS-PAGE

Buffer	**Components**
10X running buffer (1X working concentration)	144 g Glycine 30.3 g Tris base 10 g SDS
2X SDS-PAGE loading dye (1X working concentration)	100 mM Tris-HCl, pH 6.8 4% (w/v) SDS 20% (v) Glycerol 0.2% (w/v) Bromophenol Blue 200 mM β-mercaptoethanol

2.4.6.2 SDS-PAGE gel staining

After gel electrophoresis had completed, the gel was carefully removed from between the glass plates. The gel was placed in a plastic box, and covered with Coomassie blue stain (Table 2.18). The gel was stained for ~1 hour with gentle shaking. The Coomassie blue was siphoned off, and could be reused with another gel. The gel was then covered with de-stain solution (Table 2.18), for several hours, or overnight. The de-stain solution was changed periodically. The gel was removed from de-stain solution when it appeared mostly transparent, and the protein bands were clearly visible. Gels were photographed using the Gel Doc EZ Imager (Bio-Rad) and associated Image Lab 4.0.1 software.

Table 2.18 Solutions used to stain and de-stain SDS gels

Solution	Components	Amount
Coomassie blue stain	Coomassie blue	0.1% (w/v)
	Methanol	5% (v/v)
	Glacial acetic acid	10% (v/v)
	Distilled water	To required volume
De-stain	Methanol	5% (v/v)
	Glacial acetic acid	10% (v/v)
	Distilled water	To required volume

2.4.7 Dialysis of proteins

Proteins were dialysed using SnakeSkin dialysis tubing with a 7 kDa MWCO (molecular weight cut-off, Thermo Scientific). The dialysis tubing was cut to an appropriate length using sterilised scissors, then soaked in the filtered buffer that would be used to dialyse the protein against. One end of the dialysis tubing was sealed with clips, and the protein aliquot(s) were pipetted into the tubing. The other end of the tubing was sealed with a clip, and the tubing placed in a beaker filled with 500 ml filtered buffer. The beaker was placed on a magnetic stirrer and the protein was dialysed at 4°C for 2 hours at medium stirring speed. After two hours, the buffer was replaced with 500 ml fresh buffer and dialysed overnight at 4°C using medium stirring speed. The following day, the protein solution was withdrawn from the dialysis tubing using a pipette and transferred to a sterile Eppendorf tube. The protein concentration was measured by Bradford assay and compared with the pre-dialysis concentration to assess the percentage of recovery.

2.4.8 Size-Exclusion Chromatography-Multi Angle Laser Light Scattering (SEC-MALLS)

SEC-MALLS provides a robust method for determining the molecular weight (Mw) of proteins and protein complexes in solution, which may be vital in molecular biology research, e.g. to ensure the protein of interest has been produced correctly. Size-exclusion chromatography alone may be used to estimate the Mw of a protein, however

this is an imprecise technique as it relies on elution volume, and Mw values may be false if e.g. a protein sample displays non-favourable column interactions (Some *et al*. 2019). Combining SEC with multi-angle laser light scattering and differential refractive index (dRI) and/or UV detectors results in an absolute measurement of Mw, due to the known relationship between molar mass, scattered light amount, and sample concentration. The resultant Mw value is independent of the elution volume, and not affected by column interactions (Folta-Stogniew & Williams 1999). SEC-MALS can also measure molecular size (given as root means square radius or radius of gyration, R_g), and determine if eluting peaks are homogenous or heterogeneous, giving information about the behaviour of molecules in solution (Some *et al*. 2019).

Here, protein samples that had previously been purified were used for SEC-MALLS. Samples were typically provided at concentrations of 2-3 mg/ml, as measured by Bradford Assay. 120 µl of sample was provided, of which 100 µl was injected into a Superdex S200 10/300 GL column (GE Healthcare). The column was pre-equilibrated with running buffer; for AspA samples this was 50 mM HEPES, 200 mM KCl, pH 8.0, for ParB samples it was 20 mM HEPES, 150 mM NaCl, pH 7.0. Buffers were filtered using a 0.2 µm syringe filter before use. The sample flow rate over the column was 0.5 ml/min, with the run lasting 60 minutes. The SEC-MALLS system comprised a Wyatt HELEOS-II multi-angle light scattering detector and a Wyatt rEX refractive index detector linked to a Shimadzu High-performance liquid chromatography (HPLC) system. The refractive index increment (*dn/dc*) was normalised using BSA as a calibration standard. The UV absorbance detection was at 280 nm, and data were analysed using Astra V software, using the Zimm fitting model with a fit degree of 1. All SEC-MALLS experiments were conducted by Dr Andrew Leech.

2.4.9 Circular Dichroism (CD)

Circular Dichroism (CD) is a spectroscopic technique that is commonly used to assess protein structure, as it gives information about secondary structure elements. This is due to particular secondary structure elements such as alpha-helices and beta-sheets differentially absorbing left and right-handed circularly polarised light, resulting in characteristic CD spectra for each structural element (Micsonai *et al.* 2015).

This differential results in light being elliptically polarised after absorption, where the ellipticity (measured in millidegrees) at a certain wavelength indicates the amount of absorbance difference, or dichroism. Different secondary structure elements have characteristic CD spectra, e.g. alpha helices have negative peaks at 208 and 222 nm (Greenfield 2006). Purified protein aliquots were used for CD. The samples were diluted from storage buffer to 0.3 and 0.1 mg/ml samples, as measured by Bradford Assay, using CD buffer (10 mM HEPES, 50 mM KCl, pH 7.5, 0.2 µm filtered). A 400 µl sample volume was used, and sample collection was performed in a 1mm path-length quartz cuvette using a Jasco J-1500 spectrometer, with a temperature control setting of 20°C. The data were collected over a wavelength of 190 to 260 nm, at 0.5 nm intervals with a bandwidth of 1.00 nm. The scanning speed was 50 nm/min, and each spectrum was the result of 5 repeat accumulations. The CD buffer spectrum was also collected and subtracted from sample spectra. The data were analysed using Jasco Spectra Manager v2 software, and additional analysis of secondary structure proportions was performed using the BeStSel secondary structure prediction server (Micsonai *et al*. 2018), by uploading CD data (wavelength and measured ellipticity (mdeg)) to https://bestsel.elte.hu/index.php.

2.4.10 DMP chemical cross-linking

The cross-linking reagent Dimethyl pimelimidate (DMP) was used to assess the oligomerisation properties of AspA, and the interactions between AspA and ParB. DMP was freshly prepared before use by diluting in cross-linking buffer (50 mM HEPES, 50 mM KCl, 5 mM M_gCl_2, pH 8.5) to a working concentration of 20 mM. Proteins were diluted in either cross-linking buffer or storage buffer, or buffer-exchanged into cross-linking buffer, depending on the experiment. DMP was serially diluted such that final concentrations typically ranged from 0.1 to 10 mM. 10 µl of DMP at the required concentration was added to 10 µl of protein. The reaction was incubated at 80°C for one hour, then quenched by the addition of 1 µl 0.5 M Tris-HCl, pH 6.8. Samples were mixed with 20 µl 2X SDS loading dye, heated at 95°C for 5 minutes then analysed by SDS-PAGE, typically using a 15% acrylamide gel. Parameters such as cross-linking temperature, mass or molar ratio of protein used, percentage of gel, presence/absence of DNA were sometimes altered, and this is mentioned in the text.

2.4.11 BS3 chemical cross-linking

The cross-linking reagent bis[sulfosuccinimidyl] suberate (BS3) was also used in some experiments. Here, a typical reaction was similar to when using DMP. The 2 mg aliquot of BS3 was dissolved in 70 µl of cross-linking buffer (50 mM HEPES, 50 mM KCl, 5 mM M_gCl_2, pH 8.5) to provide a 50 mM stock solution, as per the manufacturer's instructions. BS3 is recommended to be used at between 20 and 50-fold molar excesses for protein concentrations of <5 mg/ml, and at final concentrations of 0.25 – 5 mM. The specific concentrations and -fold molar excess used in each experiment is mentioned in the text. The mass and molar concentrations of proteins and DNA used for individual experiments are noted in the text. The reactions were typically incubated at 80°C for 30 minutes, then quenched by adding 50 mM Tris-HCl, pH 7.5. Samples were mixed with 20 µl 2X SDS loading dye, heated at 95°C for 5 minutes then analysed by SDS-PAGE.

2.5 DNA-Protein interaction assays

2.5.1 Electrophoretic Mobility Shift Assay (EMSA)

2.5.1.1 Sample preparation and gel electrophoresis

EMSA assays were used to assess the interaction between AspA and the second palindromic site. In EMSA, the labelling of DNA fragments is required to visualise DNA, therefore a 247 bp fragment harbouring the second palindrome centrally was amplified by PCR, using primers of which the forward primer had a biotin label at the 5' end. The reverse primer was not modified. After PCR, the products were run on a 2% (w/v) agarose gel, extracted, and purified as in Section 2.3.10. DNA concentration was measured using the Qubit 3 Fluorometer (Invitrogen), and diluted to the required stock concentration in Milli-Q water.

EMSA reactions were set up as follows: Biotinylated DNA fragments at a final concentration of 0.12 nM were mixed with 10X binding buffer (100 mM Tris-HCl pH 7.5, 500 mM KCl, 10 mM DTT), glycerol, NP-40 and $MgCl_2$. The synthetic polymer Poly(dI-dC) (Poly(deoxyinosinic-deoxycytidylic) acid, Thermo Fisher) was added as a competitor DNA to reduce non-specific interactions. One reaction contained DNA only, and the others

contained the AspA protein at increasing concentrations, e.g. from 10 to 500 nM. The final reaction volume was 20 µl, and a typical set of reactions is shown in Table 2.19.

Reactions were incubated at 50°C for 30 minutes, during which time an agarose gel [typically 1.2% or 2% agarose dissolved in 0.5X TBE (40 mM Tris pH 8.3, 45 mM Boric Acid, 1 mM EDTA)] was pre-run in 0.5X TBE at 100 V, 4°C. After incubation, the reactions were loaded onto the agarose gel, with the addition of 5 µl of 5X loading dye to the 'DNA only' only reaction. The gel was run at 100 V, 4°C typically for ~2 hours at which point the blue dye had migrated to near the bottom of the gel.

Table 2.19 Typical components of an EMSA reaction

Component (final concentration)	**DNA only (µl)**	**10 nM (µl)**	**20 nM (µl)**	**40 nM (µl)**	**50 nM (µl)**	**100 nM (µl)**	**200 nM (µl)**	**500 nM (µl)**
Milli-Q water	8.74	2.90	2.90	2.90	1.44	1.44	1.44	1.44
Binding Buffer (1X)	2	2	2	2	2	2	2	2
Glycerol (2.5%)	1	1	1	1	1	1	1	1
$MgCl_2$ (5 mM)	1	1	1	1	1	1	1	1
NP-40 (0.05%)	1	1	1	1	1	1	1	1
Poly(dI-dC) (1 µg)	1	1	1	1	1	1	1	1
DNA (0.12 nM)	5.26	5.26	5.26	5.26	5.26	5.26	5.26	5.26
Protein	-	5.84	5.84	5.84	7.30	7.30	7.30	7.30
Final volume	20	20	20	20	20	20	20	20

2.5.1.2 DNA transfer onto positively charged membrane

Whilst running the gel, a piece of positively-charged nylon membrane (Roche) and four pieces of Whatman 3 mm filter paper were cut to the size of the agarose gel. The membrane and filter paper were soaked for at least 10 minutes in 0.5X TBE. When electrophoresis had finished, a 'sandwich' was constructed to transfer the DNA from within the gel to the positively-charged nylon membrane. The sandwich comprised of some folded blue paper towel, then two pieces of pre-soaked filter paper, then the nylon

membrane. The gel was placed carefully on top of the membrane, and pressed down to ensure equal contact and remove air. Then, two more pieces of filter paper, followed by more blue paper towel were placed on top, completing the transfer sandwich. The transfer was aided by placing heavy weights (e.g. lab books/autoradiographic cassettes) on top of the sandwich to ensure effective transfer. Transfer took place overnight at room temperature. The next day, the sandwich was disassembled and the membrane carefully wrapped in cling film. It was then exposed with DNA face-down to 302 nm UV light for 5 minutes on a UVP Transilluminator (Jena Analytic), covalently cross-linking the DNA to the positively-charged nylon membrane.

2.5.1.3 Detection of DNA on X-ray film

A LightShift Chemiluminescent kit (Thermo Scientific) was used to detect the membrane-bound biotinylated DNA. The buffers listed below in Table 2.20 were provided in the kit, or prepared separately when necessary, ensuring that they were syringe-filtered using a 0.22 µm filter. Buffers were kept at 4°C, therefore the blocking and wash buffers were placed in a water bath at ~40°C to dissolve precipitated SDS before use. The membrane was transferred to a clean tray, using sterilised forceps at all times to avoid contamination. All following steps were performed at room temperature. The membrane was incubated for 15 minutes in 20 ml blocking buffer with gentle shaking. The blocking buffer was discarded, and the membrane incubated with 10 ml blocking buffer plus 35 µl stabilised streptavidin-horseradish peroxidase conjugate for 15 minutes with shaking. The blocking buffer was discarded, and the membrane washed four times with 20 ml of 1X wash buffer with shaking. The membrane was transferred to a new clean box, and incubated with 30 ml equilibration buffer without shaking for 5 minutes. Meanwhile, the detection solution was prepared by mixing together 2 ml of luminol/enhancer and 2 ml stable peroxide solutions. The membrane was drained using Whatman filter paper, and placed in a new clean box. The detection solution was carefully poured on top of the membrane, ensuring complete coverage, and incubated for 5 minutes. The membrane was again drained on filter paper, and when dry, placed inside an autoradiographic cassette, and left for 5 – 10 minutes to increase the signal intensity. The membrane was then exposed onto X-ray film for a time dependant on the intensity of the signal.

Table 2.20. Buffers used in EMSA and DNase I footprinting detection

Solution	Components
Blocking buffer	100 mM Tris-HCl pH 7.5, 5% SDS, 1% (w/v) BSA
1X wash buffer	100 mM Tris-HCl pH 7.5, 5% SDS
Equilibration buffer	100 mM Tris-HCl pH 7.5

2.5.1.4 Data quantification and analysis

To derive ligand-binding curves, the relative band intensities of unbound DNA bands were measured compared to the 'DNA only' lane using the Gel-Doc and Image Lab 4.0.1 software (Bio-Rad). The fraction of DNA bound was plotted against protein concentration, and the apparent dissociation constant calculated using the one-site binding equation:

$$\mathrm{y} = B_{max}\,\frac{[\mathrm{AspA}]}{K_d + [\mathrm{AspA}]}$$

where Y is the fraction of DNA bound, B_{max} is the maximal binding, K_d is the equilibrium dissociation constant and [AspA] is the protein concentraion. The Microsoft Excel plugin Solver was used to fit the data using a non-linear regression by maximising the R^2 value.

2.5.2 DNase I Footprinting

A Maxam-Gilbert sequencing ladder, in which purines are chemically modified, was made by adding 50 µl of formic acid to 12 µl of 30 nM DNA (the same 247 bp biotinylated fragment used in EMSA), and incubated for 2.5 mins at 22°C. The reaction was stopped by adding 200 µl of 300 mM sodium acetate, pH 7.0, followed by ethanol precipitation. The dried pellet was resuspended in 100 µl of 1 M piperidine, and incubated for 30 mins at 90°C. 10 µl of 3 M sodium acetate, pH 7.0 was added, followed by ethanol precipitation. The dried pellet was resuspended in 20 µl of loading buffer (95% formamide, 20 mM EDTA, 0.05% bromophenol blue, 0.05% xylene cyanol), and denatured for 10 mins at 99°C before loading 5 µl on the sequencing gel. Footprinting reactions were set up using the same conditions as EMSA, except using a final DNA concentration of 5 nM. The reactions were incubated for 20 mins at 50°C, then DNase I (1.2 U) was added, and incubated for 70 secs at 25°C. The digestion reaction was stopped by adding 200 µl of stop solution

(10 mM EDTA, 300 mM sodium acetate), then the reactions were subjected to phenol:chloroform extraction followed by ethanol precipitation, with the addition of 1 µl of glycogen to visualise the pellet. Pellets were dried and resuspended in 12 µl loading buffer, denatured, and 5 µl loaded on a pre-warmed 6% acrylamide sequencing gel. The gel was run in 1X TBE at 60 W for ~2.5 hrs. DNA fragments were transferred to a positively charged nylon membrane, covalently crosslinked using UV, and detected using the Lightshift chemiluminescence detection substrate followed by exposure onto film. The same protocol as outlined in Sections 2.5.1.2 and 2.5.1.3 was followed, except in the use of different volumes of solutions due to the larger size of membrane.

2.6 Atomic Force Microscopy

Atomic force microscopy (AFM) is a microscopy technique that, due to its high resolution of less than a nanometre, can be employed to acquire information about a wide variety of materials and sample types, both non-biological and biological, e.g. molecular interactions between DNA and protein. The AFM apparatus consists of a cantilever, on the end of which is a tip or probe which contacts the sample surface, and which typically has a radius on the order of nanometres. On contact with the surface, the cantilever is deformed or bent, and the amount of deformation is measured by reflecting a laser off the cantilever into a detector, thus providing information about the sample (e.g. height of a molecule above the sample surface). Various scanning modes exist, such as contact mode, in which the tip is dragged across the sample surface. However for biological or sensitive samples, 'tapping' modes are frequently used in which the tip makes transient contacts with the surface, thus reducing force and lessen damage or deformation to the sample. The proprietary PeakForce Tapping mode (Bruker) has provided significant advances applicable to the life sciences, for example enabling resolution of individual major and minor grooves in plasmid DNA (Pyne *et al*. 2014).

2.6.1 Sample preparation

A typical AFM sample reaction contained either the 200 bp or 1.68 kbp linear fragments, or the 4.4 kbp circular plasmid. The DNA concentration was measured using the Qubit 3 Flourometer, and was diluted in filtered Milli-Q water. The final concentration of the DNA

was typically 0.5 ng/µl. Filtered AspA (between 50 and 300 nM final concentration) and M_gCl_2 (10 mM final concentration) were added in a total reaction volume of 50 µl. Reactions were incubated at 37°C for 20 minutes. Meanwhile, 9.9 mm mica discs (Agar Scientific) mounted on metal support discs were stripped of ~5 top surface layers using tape, ensuring that the surface appeared flat. 20 µl of reaction volume was pipetted onto the disc using filter tips, and incubated at room temperature for 5 minutes. The discs were washed with 1 ml filtered Milli-Q water and gently dried using air passed through a 0.22 µm filter before microscopy.

2.6.2 Microscopy and image analysis

A Bruker BioScope Resolve microscope fitted with a ScanAsyst-Air-HR cantilever probe (Bruker) was used to scan the samples using the QNM (quantitative nanomechanical mapping in air scanning mode). The probe tip has a radius of 2 nm and a spring constant of 0.4 N/m. Typical scanning parameters included; peak force amplitude 20 nm, peak force setpoint 50 pN, scan rate 1.21 Hz, and samples per line 1024. These and other parameters were adjusted to achieve the best image. The scanning area size depended on the size of the DNA being imaged, but was typically between 0.5 and 2 µm square. Gywddion software was used to process AFM images. Data processing including image flattening, mean plane subtraction and correction of horizontal scarring was performed prior to analysis. The height of the DNA and AspA-DNA complexes above the mica surface was measured by using the 'extract profile' tool on a cross-section of a molecule; this is visualised using the scale bar which changes colour as a function of height above the mica surface.

2.7 Bioinformatics methods and tools

2.7.1 Sequencing and assembly of the NOB8-H2 chromosome

Sulfolobus NOB8-H2 was grown in Brock's medium from glycerol stocks prior to DNA isolation. 30 ml of culture was grown at 75°C with agitation until the OD_{600} reached 0.3-0.5. Cells were pelleted at 6,000 xg for 10 minutes at 4°C. The pellet was resuspended in 1 ml Milli-Q water, transferred to a fresh 1.5 ml Eppendorf tube, and pelleted again at 12,000 xg for 1 minute at room temperature. Genomic DNA was isolated using the Sigma GenElute Genomic DNA Kit, following the manufacturer's instructions. DNA was sequenced using the MinION Flongle sequencer, and the resulting sequence was assembled using Canu. The genomic DNA was also sequenced via Illumina, which produced ~60x of coverage, and both sequencing data sets were combined to produce a polished assembly of the chromosome of ~2.81 Mbp. The genome was compared to published *Sulfolobus* genomes using the MASH program (Ondov 2016). Prokka (Seemann 2014) was used to annotate the genome, and Artemis (Rutherford 2000) used to visualise the annotated sequence. Basic Local Alignment Search Tool (BLAST) was used to identifythe multiple insertions of the pNOB8 plasmid within the NOB8-H2 chromosome, and DNA Plotter (Carver 2009) used to depict the circular genome.

2.7.2 Phylogenetic analysis

2.7.2.1 Single gene tree

MEGA-X (Hall 2013) was used to construct the phylogeny based on the 16S rRNA gene. Homologous sequences were retrieved from NCBI using the Megablast parameter against the nucleotide collection. Sequences from 26 strains were used to build the phylogeny, including that of *Sulfolobus* NOB8-H2. Sequences from complete genomes were used in all cases, except for one sequence. Once imported into MEGA, sequences were first aligned using the MUSCLE alignment algorithm, aligning by codons. Once aligned, phylogenetic trees were constructed using Maximum-Likelihood (ML), Maximum Parsimony and Neighbour-Joining methods, each giving similar topologies. For ML trees, the 'find best model' feature was used and this model (TN93+G+I) used to generate the tree. The 'Gaps/Missing Data Treatment' option was set to Partial Deletion to retain more sequence information (Hall 2013). For 'Test of Phylogeny', bootstrapping was used to test

the reliability of the tree, with the number of bootstrap replicates set to 1000. To root the tree, *Metallosphaera sedula* was used as the outgroup, as it belongs to the same family, the Sulfolobaceae, as *Sulfolobus*.

2.7.2.2 Concatenated phylogenetic tree

The Comparative Genomics function at http://www.genoscope.cns.fr/agc/microscope/home/index.php was used to perform an analysis of the pan/core genome of *Sulfolobus,* using the genomes of five species from the order Sulfolobales that were available in the database. This produced a core genome of 96 genes, of which ten were chosen to construct the phylogenetic tree (Table 2.21). The ten core genes were found in the *Sulfolobus* NOB8-H2 strain with BLAST, using the closely related strain *S. solfataricus* 98/2 for the query gene sequences. The NOB8-H2 sequences were then BLASTed against all of the other strains that were used to construct the 16S rRNA phylogeny, with the exception of three more distantly-related strains which did not produce any BLAST hits. The BLASTn (somewhat similar sequences) search algorithm was used. In total, 23 strains were used to construct the concatenated phylogeny. The ten gene sequences were concatenated into a single fasta file using http://www.bioinformatics.org/sms2/combine_fasta.html, ensuring synteny was preserved. These were then imported into MEGA, and aligned using MUSCLE as before. Here, the best Maximum Likelihood model was GTR+G+I, and this was used to construct an ML tree with 1000 bootstrap replicates, again using Partial Deletion. *Metallosphaera sedula* was used as the outgroup to root the tree.

Table 2.21. Genes used in concatenated phylogenetic tree

Gene	**Accession number (*S. solfataricus* 98:2)**
30S ribosomal subunit protein S11	161207|25034276|ACUK_v1_260022
HTH-type transcriptional regulator LysM	229438|25035296|ACUK_v1_1110021
Superoxide dismutase [Fe]	232659|25035455|ACUK_v1_1110180
Putative transcriptional regulators, CopG/Arc/MetJ family	235122|25034193|ACUK_v1_230047
ORC1-type DNA replication protein 3	277507|25036425|ACUK_v1_2500059
TATA-box-binding protein	306259|25036602|ACUK_v1_2690081
Proteasome subunit alpha	306371|25035920|ACUK_v1_1800201
Elongation factor 2	306395|25035910|ACUK_v1_1800191
DNA repair and recombination protein RadA	311342|25035388|ACUK_v1_1110113
DNA-directed RNA polymerase subunit B	1377733|25035366|ACUK_v1_1110091

2.7.3 *Sulfolobus* whole genome comparisons

DNA-DNA hybridisation was originally a wet-lab technique in which DNA from two different organisms is mixed, and the amount of hybridisation between the two measured. Though used less frequently, now, it was once a gold standard technique for defining distinct prokaryotic species, based on a DDH score of 70% being the boundary for species delineation (Meier-Kolthoff *et al*. 2013). The wet-lab approach has become less popular with the advances in whole-genome sequencing, and now *in silico* DDH methods are available, such as the online Genome-to Genome Distance Calculator (GGDC), in which pairwise alignments between two genomes are transformed in into a genome-to genome distance value (Meier-Kolthoff *et al*. 2013). DNA-DNA hybridisation (DDH) values were computed by uploading selected Sulfolobaceae genome FASTA DNA files to the Genome-to-Genome Distance Calculator 2.1 located at http://ggdc.dsmz.de/, using the default settings.Blast Ring Image Generator (BRIG, Alikhan *et al*. 2011) was used to visualise the NOB8-H2 chromosome and compare it to other sequenced genomes of the family Sulfolobaceae. The NOB8-H2 chromosome was used as the reference sequence, and all other genomes used were downloaded from NCBI and set as the query sequences. The most recent version of BLAST (BLAST+ version 2.10.0) was installed to enable local sequence alignments. Genomes were compared using BLASTn, using nucleotide identity values of 100%, 90% and 70%.

2.7.4 COG Analysis

The COG database was designed to classify proteins, which may have an undefined functional role, based on their orthology, as orthologous proteins typically possess the same function and domain structure (Tatusov *et al*. 2000). The 26 COG functional classes contain previously assigned orthologues, therefore a match is strongly indicative of orthology (Galperin *et al*. 2019). COG originally stood for 'Clusters of Orthologous Groups of Proteins', but has since been rebranded as 'Clusters of Orthologous Genes' to reflect the greater complexities inherent in the evolutionary relationships between genes (Galperin *et al*. 2019). The COG database has expanded since its inception to cover over 5,000 COG groups, and archaea-specific searches are now possible using the arCOG database (Makarova *et al*. 2015a). Here, the annotated *Sulfolobus* NOB8-H2 Genbank file was converted to a protein sequence FASTA file using the tools at

https://rocaplab.ocean.washington.edu, then uploaded to the web server WebMGA (http://weizhong-lab.ucsd.edu/webMGA/. WebMGA annotates input peptide sequences against the NCBI COG database using RPS-BLAST (Wu *et al*. 2011). The COG and Pfam orthologous protein group databases were used to assign protein functions to the annotated genome. For COGs, proteins were assigned to one of 25 categories, excluding the recently added Mobilome category X (Galperin *et al*. 2019).

2.7.5 NOB8-H2 CRISPR spacer analysis

The first stage of the CRISPR-Cas adaptive immunity response is the adaptation phase, in which short DNA sequences from the invading element(s) (e.g. virus or plasmid) are incorporated into the host CRISPR array as spacers (Makarova *et al*. 2015b). The spacers are inserted in a linear fashion, and provide the host with a record of past encounters that can be accessed to prevent future invasion. Obtaining information about the invasive elements to which the spacers were derived is useful in an ecological context, as it details past encounters with mobile genetic elements, but in the context of this study, may provide clues as to how the conjugative plasmid pNOB8 is stably maintained within *Sulfolobus* NOB8-H2. To analyse the NOB8-H2 spacers, a list of those viruses and plasmids which are known to interact with crenararchaea was derived from a recent thesis on *Sulfolobus* extrachromosomal genetic elements (Liu thesis 2015). The accession numbers of these 53 elements were used in a BLAST search against spacers of the two CRISPR arrays, with the top three hits for each spacer returned. Matches with a bit score of >30 were deemed as significant, giving e-values of <0.01 and % coverage of >85. Code for this BLAST search was written by Dr James Robson.

2.7.6 pNOB8 analysis

The pNOB8 plasmid has previously been sequenced (accession number NC_006493), with potential functions assigned to ~20% of the 52 gene products (She *et al*. 1998). Therefore, the Genbank file for the plasmid was used to conduct BLASTp searches for each ORF manually, using the BLASTp algorithm. COG analysis was performed in the same way as for the NOB8-H2 chromosome.

2.8 Structural analysis software and tools

Molecular structures were visualised using CCP4MG (McNicholas *et al.* 2011) and PyMOL (DeLano 2002). For protein superpositions, the SSM (secondary structure matching) method of CCP4MG was used.

To delineate the pNOB8 ParB domains, the PSIPRED Analysis Workbench was used to predict secondary structure. Secondary structure is based on position-specific scoring matrices, and has a prediction accuracy of 84% (Buchan & Jones 2019). The server is available at http://bioinf.cs.ucl.ac.uk/psipred/. The IUPred3 web interface was also used to identify putative disordered regions, based on the assumption that amino acids comprising disordered regions cannot form favourable interactions (Erdös & Dosztányi 2020). The server is available at https://iupred.elte.hu/. The amino acid sequence of pNOB8 ParB was uploaded to both PSIPRED and IUPred3 servers, using default settings.

The Phyre2 server can be used to predict protein structures based on acquisition of homologous sequences and comparison with known structures (Kelley *et al*. 2015). Protein models are generated using a four-stage method: (i) gathering homologous sequences and using PSIPRED to predict secondary structure, (ii) Conversion to a hidden Markov model (HMM) and alignment against HMMs of known structures to create a crude backbone model, (iii), loop modelling to resolve any insertions or deletions in the backbone model, (iv), side chain fitting, placing side chains in the most probable conformation that avoid steric clashes to create a final model (**Figure 2.2,** Kelley *et al*. 2015). Phyre2 can also predict the effect of mutations of a particular amino acid, and the likelihood of any phenotypic effect due to the mutation. Phyre2 was used in 'normal' mode to model the structure of pNOB8 ParB-N.

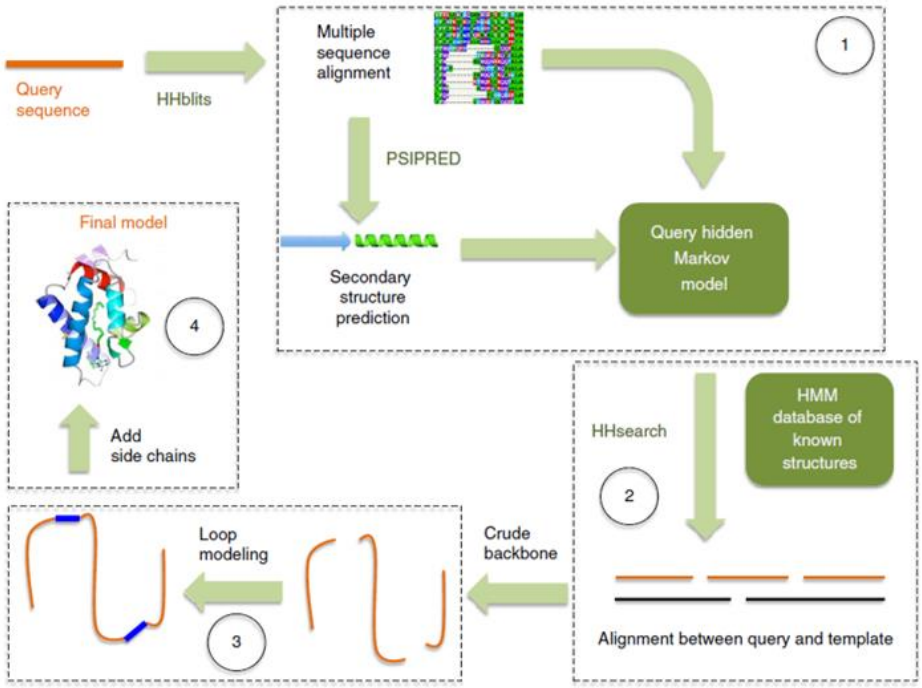


Figure 2.2 Phyre2 algorithmic workflow. The four stages of the Phyre2 pipeline in normal mode. Adapted from Kelley *et al*. 2015.

The latest iteration of the neural network-based structure prediction model Alphafold was released during the writing of this thesis. Alphafold generates structures with atomic accuracy even in the absence of homologous structures, with accuracy across all atoms of 1.5 Å (Jumper *et al*. 2021). The network processes input sequences in a novel way by embedding multiple sequence alignments (MSAs) and pairwise features using a neural network block named Evoformer. The output structure is refined in an iterative manner by recursively feeding outputs back into the network, making incremental enhancements until the structure cannot be improved (**Figure 2.3,** Jumper *et al*. 2021). The Alphafold v2.1 source code is available, but here, the web-based AlphaFold Colab server was used to model pNOB8 ParB from its amino acid sequence using default parameters. The server uses a simplified version of AlphaFold v2.1, however accuracy is near-identical for most targets.

The AlphaFold server is available at: https://colab.research.google.com/github/deepmind/alphafold/blob/main/notebooks/AlphaFold.ipynb.

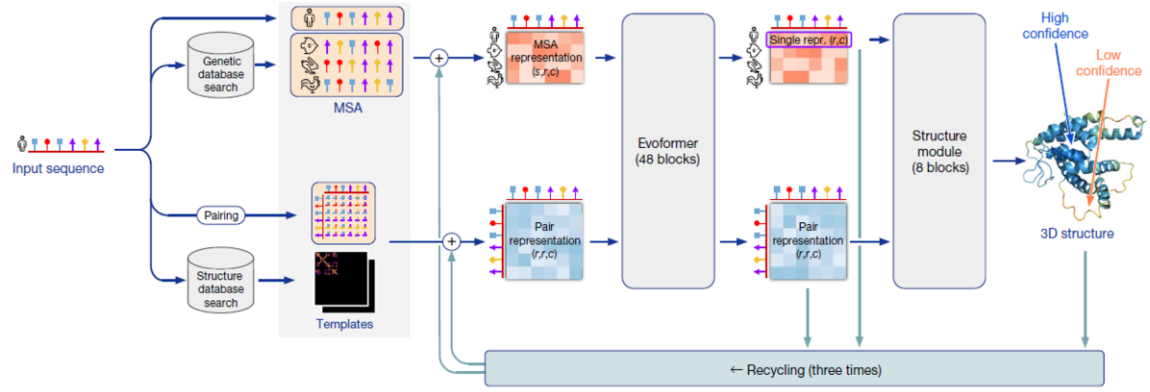


Figure 2.3 AlphaFold model architecture. The AlphaFold neural network structure prediction pipeline. Adapted from Jumper *et al*. 2021.

The ClusPro server was used for protein-protein docking and to model the interface between AspA and ParB-N. ClusPro is a direct docking method, which seeks the most energetically favourable structure: first by sampling billions of structure conformation, then clustering of the lowest-energy structures based on root-mean-square deviation, followed by energy minimisation refinement (Kozakov *et al*. 2017). ParB-N was specified as the receptor molecule, and AspA as the ligand, and a number of docking iterations were run with parameters such as attractive residues were altered to approximate the derived SAXS model for the interaction (Schumacher *et al*. 2015). The ClusPro server is available at https://cluspro.bu.edu/home.php.

Chapter 3

AspA-DNA interactions

Chapter 3

Probing AspA-DNA interactions at the second palindrome

3.1. Introduction

The *Sulfolobus* strain NOB8-H2, isolated from hot springs at Noboribetsu, Japan, harbours a 41 kb conjugative plasmid, pNOB8, which contains about 50 genes (She *et al*. 1998). The plasmid contains a partition or segregation cassette, similar in genetic organisation to those found on bacterial plasmids and chromosomes (Hayes & Barillà 2006a, Schumacher 2008, Schumacher 2012, Wang *et al.* 2013). The partition cassette contains three genes arranged in a tricistronic operon (pNOB8 *orfs 44, 45* and *46*), and although the *aspA-parB-parA* cassette is found on other crenarchaeal plasmids and chromosomes, a bicistronic partitioning system is more commonplace in bacteria, and this arrangement also comprises the chromosomal *segAB* cassette of *S. solfataricus* (Kalliomaa-Sanford *et al*. 2012). The amino acid sequences of two of the proteins were found to be similar to the bacterial partition proteins ParA and ParB (pNOB8 ORFs 46 and 45 respectively), which perform an important role in the active partitioning of plasmids, ensuring their correct dissemination to future cellular generations (Baxter & Funnell 2014). This similarity to ParA and ParB provided a clue to the possible function of ORFs 45 and 46 as segregation proteins, potentially involved in the correct partitioning of pNOB8. The third protein in the partition cassette, ORF 44, did not display any similarity to previously characterised segregation proteins (Schumacher *et al*. 2015). The protein encoded by *orf 44* was demonstrated, via band-shift and DNase I footprinting assays, to bind a palindromic sequence of 23 bp, immediately upstream of the partition cassette, and was dubbed AspA (Archaeal segregation protein A, Schumacher *et al*. 2015). AspA binds to the palindrome as a dimer, and has high affinity for this sequence *in vitro*. Moreover, AspA demonstrated the ability to form an extended protein-DNA superhelical structure, capable of spreading along the DNA **(Figure 3.1)** (Schumacher *et al*. 2015). Here then, AspA performs the role of site-specific DNA-binding protein (also known as centromere-binding protein, or CBP) that is more usually the function of ParB in bacterial segregation systems.

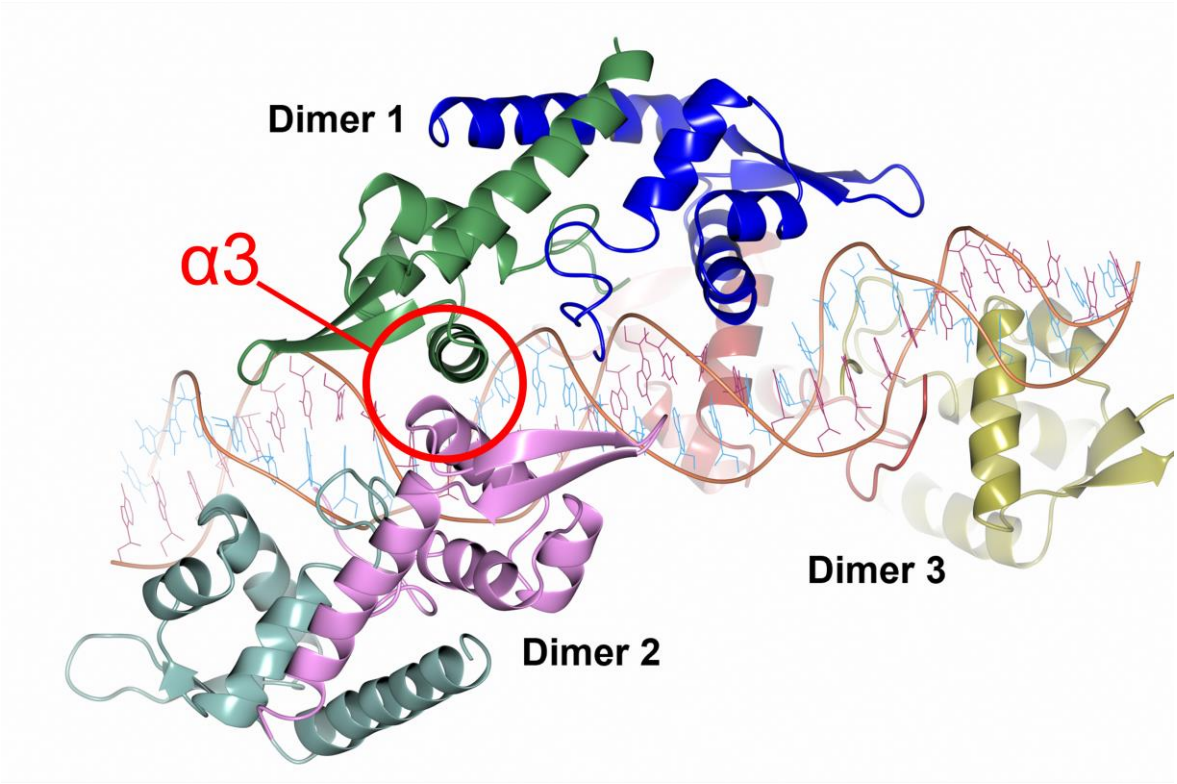


Figure 3.1. AspA binding and spreading on DNA. Structural and biochemical data show AspA to be dimeric. Shown are three AspA dimers: coloured blue/green, cyan/purple and red/yellow. Dimer 1 (blue/green) interacts with the palindromic sequence upstream of the cassette, but at increasing concentrations, AspA dimers can spread along the DNA in both 5' and 3' directions, generating a protein-DNA superhelical structure. Adjacent AspA dimers can each insert alpha helix 3 into the same major groove, facilitating spreading. The third alpha-helices from adjacent dimers both occupying the same major groove is shown. Adapted from Schumacher 2015 using CCP4MG (PDB entry 5k5q).

The crystal structure of AspA has previously been solved, in both apo- and DNA-bound states, and show the protein has a winged helix-turn-helix (wHTH) component, known to be a DNA-binding motif across the three domains of life (Aravind *et al.* 2005), plus a C-terminal helix involved in dimerisation. Both analytical ultracentrifugation (AUC) and size-exclusion chromatography – multi-angle laser light scattering (SEC-MALLS) showed AspA to be predominantly dimeric. This dimerisation of bacterial ParB CBPs has previously been demonstrated to facilitate binding to their cognate DNA sequences (Delbrück *et al*. 2002, Schumacher & Funnell 2005, Schumacher *et al*. 2010).

Spreading of AspA from the palindromic sequence upstream of the partition cassette is facilitated by its third alpha helix, as adjacent dimers can insert one α3 helix each into the same major groove of the DNA, allowing multiple dimers to bind DNA non-specifically in a contiguous fashion and enable the generation of a protein-DNA superhelix **(Figure 3.1)**. This non-specific binding and resultant spreading pattern from the centromere has previously been observed with multiple bacterial ParB CBPs (Lynch & Wang 1995, Rodionov *et al*. 1999, Breier & Grossman 2007, Tran *et al*. 2017), with the protein able to spread tens of kilobases in some cases (Jalal *et al*. 2021). This spreading is thought to have a structural basis, where the formation of extended partition complexes increases interactions with both the ParA motor protein and other more distal ParB-DNA assemblies (Schumacher 2012, Graham *et al.* 2014, Sanchez *et al.* 2015, Funnell 2016, Song *et al*. 2017). However, in addition, ParB spreading along the DNA may act to repress the transcriptional activity of neighbouring genes (Bartosik *et al*. 2004, Kusiak *et al*. 2011). CBP proteins are also known to autoregulate expression of their own operon, in this way controlling the cellular concentration of both the CBP and motor proteins (Carmelo *et al*. 2005, Schumacher 2008).

Aside from the palindrome upstream of the *aspA-parB-parA* cassette, there is another identical sequence found on pNOB8, located approximately 1.5 kb away **(Figure 3.2)**, and unpublished footprinting experiments in the Barillà group have established that AspA binds to this second site *in vitro*. The existence of this second identical palindrome on pNOB8 is intriguing, raising questions about its role in plasmid segregation, and how AspA interacts at this location. The AspA structure bears similarity to that of PadR superfamily transcription factors found in bacteria, which possess both wHTH and C-terminal dimerisation domains, and interact with palindromic DNA sequences to control gene expression (Fibriansah *et al*. 2012, Park *et al*. 2017). Therefore, it is hypothesised that in *Sulfolobus* NOB8-H2, AspA could perform a dual function similar to bacterial ParBs. The existence of two identical palindromes on pNOB8 may support this, as perhaps AspA performs different functions at each palindrome. Understanding the similarities of binding properties and spreading behaviours of AspA at each palindromic sequence may provide clues as to the functionality of the protein at each site.

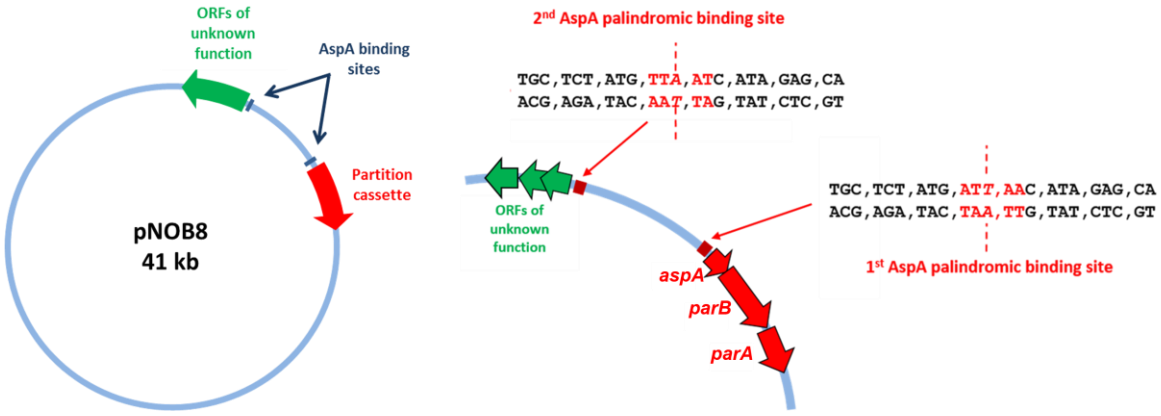


Figure 3.2. Map of pNOB8 and position of partition cassette and AspA binding site(s). (Left) Cartoon of pNOB8. **(Right)** Expanded view of partition cassette and palindromes. Positions of *aspA*, *parB* and *parA* genes on pNOB8 are shown as red arrows. The position of the first 23 bp palindromic binding site of AspA is indicated just prior to the start of the *aspA* gene; red dashed lines indicate the palindrome centre of symmetry. The additional palindromic binding site of AspA is located approximately 1.5 kb upstream of the partition cassette. The palindrome at site 2 is an inverted form of that at site 1 (see 5 bp highlighted in red), and is upstream of three genes of unknown function (green arrows). Adapted from (She 1998), and (Schumacher 2015).

3.1.1 Aims

This chapter will aim to detail the interactions of AspA at the second palindromic site on pNOB8, using a variety of *in vitro* experimental techniques.

These main aims will be addressed:

1) Whether AspA binds to the second palindrome, and if so, does it bind with an affinity similar to that observed for the first palindrome?

2) Which amino acids are important for DNA binding and spreading activity, and does mutating theses residues result in a pattern distinct from the wild-type?

3) What is this pattern of spreading from the second palindrome compared to the first, and does this give any indication as to the function of the protein at each site?

3.2 Results

3.2.1 AspA binds to the second palindromic site

3.2.1.1 WT AspA overproduction and purification

In order to assess the ability of AspA to bind the second palindromic site on pNOB8, the protein first was first overproduced and purified to homogeneity. AspA was available in the laboratory collection, with the *aspA* gene previously cloned into the multiple cloning site of the pET22b(+) vector used for overexpression. The pET22-AspA construct contains *aspA* cloned in-frame with a C-terminal hexa-His tag, and is located downstream of a T7 promoter and *lac* operator. The vector was transformed into the host expression strain *E. coli* BL21-Codon-Plus (DE3) competent cells, and expression of *aspA* was induced by the addition of IPTG, an analogue of allolactose (see Materials and Methods 2.4.1). Sufficient overproduction of the protein was measured by taking aliquots from the culture medium immediately before the addition of ITPG, then at hourly intervals for three consecutive hours afterwards. The AspA protein began to be produced after one hour, and after three hours, was produced at sufficient levels to allow the protein to be purified using Ni^{2+} affinity chromatography. The protein can be purified in this way due to the affinity of imidazole side-chains on each histidine of the hexa-His tag to the nickel column. After purification and collection of six 1.5 ml fractions of AspA, the concentration was measured by Bradford assay, and aliquots of the four most concentrated fractions were run on a 15% SDS-polyacrylamide gel, along with samples taken from the column flow-through at each purification stage. The results of the AspA overproduction and purification are seen in **Figure 3.3**. The protein fractions ranged from 1.08 to 2.83 mg/ml, giving ~12 mg of AspA from 300 ml of culture, and the protein appeared homogeneous and of sufficient purity on the resulting SDS gel. The protein appeared to be the correct size, as AspA-$(His)_6$ is approximately 11.7 kDa, as calculated from the amino acid sequence using the Expasy online tool.

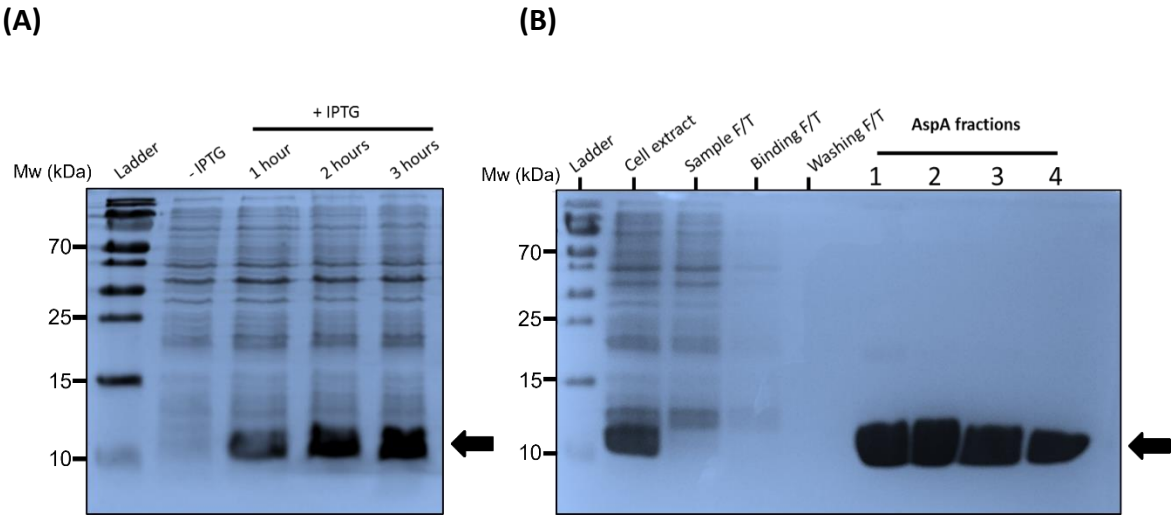


Figure 3.3. Overproduction and purification of AspA. (A) Presence of the overproduced AspA protein (Mw 11.7 kDa including 6xHis tag) was monitored in uninduced culture, then 1, 2 and 3 hours after IPTG induction, with the results seen via SDS-PAGE. **(B)** SDS gel showing the results of purification of AspA using Ni^{2+} affinity chromatography. Lanes marked Cell Extract, Sample F/T, Binding F/T and Washing F/T contain aliquots taken from various stages of the purification; Cell Extract is the crude lysate prior to chromatography, Sample F/T is after circulation over the column for 2-3 hours, and Binding F/T and Washing F/T after two washing steps with binding buffer and wash buffer respectively. The final four lanes are aliquots taken from eluted fractions of AspA. Bands representing AspA are indicated with black arrows. The Mw ladder used in both cases is PageRuler Plus prestained protein ladder (Thermo Scientific).

3.2.1.2 Generation of a biotinylated DNA fragment for EMSA studies

In order to assess AspA-DNA binding interactions *in vitro* via electrophoretic mobility shift assay (EMSA), a fragment of DNA was required that contains the second AspA palindromic binding site. The fragment should be large enough to allow potential spreading of several AspA dimers from the palindrome, a phenomenon seen at the first palindrome upstream of the *aspA-parB-parA* cassette (Schumacher 2015). Thus, primers were designed to amplify a 247 bp region of the plasmid pNOB8 that contained the second palindrome central on the fragment, in addition to the start of *orf41* and its promoter region either side **(Figure 3.4A)**. Primers were ordered from Sigma-Aldrich/Merck, with the additional 5' biotinylation modification that is required for EMSA assays. A list of primers used in this study is given in Table 2.7, Materials and Methods. The 247 bp biotinylated fragment was amplified by PCR, typically using 60 ng of either a NOB8-H2 genomic preparation, or mini/maxi-prep plasmid isolation as template DNA. The amplified fragment was verified as being the correct size by agarose gel electrophoresis **(Figure 3.4B)**, before gel extraction, purification and quantification of DNA concentration using the Qubit Fluorometer (Invitrogen).

(A) **(B)**

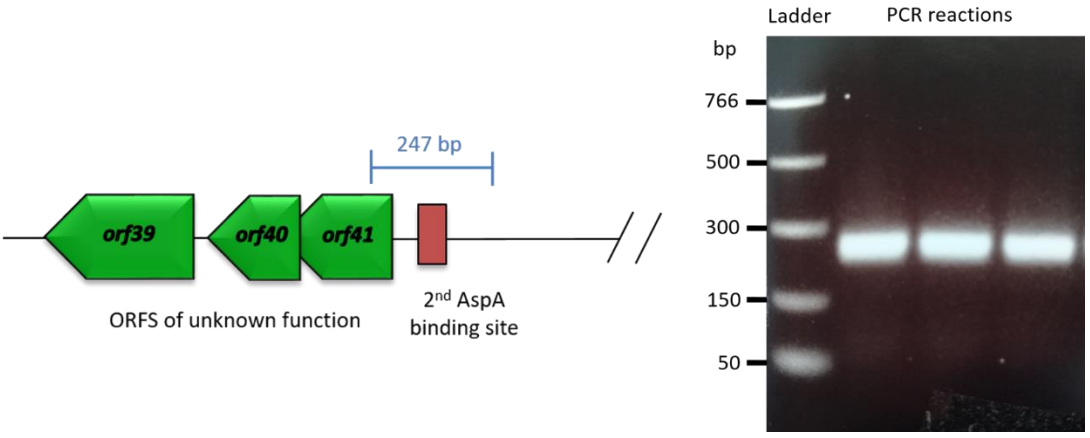


Figure 3.4. Amplification of biotinylated DNA fragment for EMSA assays. **(A)** Schematic showing the location of the 247 bp fragment to be amplified on plasmid pNOB8. The second AspA palindromic binding site is upstream of three genes of unknown function. The binding site is central in the amplified fragment. **(B)** Typical 2% agarose gel of the 247 bp biotinylated PCR products, showing distinct bands of the correct size. Bands were excised from the gel and purified before use in subsequent assays.

3.2.1.3 AspA-R49A and AspA-A53K overproduction and purification

In addition to the WT AspA protein, the laboratory also possessed two AspA mutants that could be used in EMSA experiments as controls and/or to assess their properties when binding and spreading from the second palindromic binding site. The first of these mutants, AspA-R49A, was previously shown to be incapable of binding to DNA. From the crystal structure of AspA bound to DNA, it was shown that arginine 49 makes contact purely with the phosphate backbone of the DNA (Schumacher 2015). The side chain of arginine is positively-charged, whilst the phosphate groups of the DNA are negatively-charged, resulting in a strong charge interaction between the DNA and AspA. Interestingly, mutation of this single amino acid to alanine completely abrogated DNA-binding activity, indicating that R49 is a crucial residue for correct protein function, and allowing AspA-R49A to be used as a negative control in EMSA assays. A second AspA mutant, AspA-A53K, has DNA binding activity but its ability to spread is abrogated, due to steric restriction of adjacent dimers occupying the same DNA major groove, caused by the replacement of the small alanine side chain by the more substantial lysine. This was demonstrated by DNAse I footprinting assays showing a lack of spreading from the palindrome for the A53K mutant when compared to WT AspA (Schumacher *et al*. 2015) Both mutants were already available as pET22 constructs, and were overproduced and purified by Ni^{2+} affinity chromatography, resulting in a similar yield of protein as seen for WT AspA (data not shown).

3.2.1.4 AspA binds to the second palindromic site with high affinity

Prior to conducting EMSA experiments with DNA and protein, the assay requires some optimisation to define the optimal DNA concentration that produces a clear and distinct band after transfer to the positively charged nylon membrane, and detection and exposure to film (Materials and Methods 2.5.1.2). An example of this is shown in **Figure 3.5A**, where a range of final DNA concentrations between 0.03 and 0.06 nM all

produced distinct bands on X-ray film. Initially, EMSA experiments were carried out using a native (non-denaturing) acrylamide gel. The 247 bp biotinylated fragment was incubated with increasing concentrations of WT AspA (50 to 500 nM). A shift to a higher molecular weight complex, indicating protein binding to the DNA, was observed at 50 nM concentration **(Figure 3.5B)**. Due to the high affinity of AspA to the DNA and the resultant immediate band-shift at 50 nM, a further EMSA was carried out using a range of concentrations from 10 nM upwards, in order to obtain a more precise measurement of the AspA concentration at which binding starts to occur. It was decided to use agarose gels rather than acrylamide gels at this point, as agarose appeared to give better quality data. Subsequent EMSA images in this chapter all depict agarose gels. The final DNA concentration was again optimised for use with agarose gels (see Appendix 1), with a concentration of 0.02 ng/µl, or 0.12 nM used in all subsequent EMSA assays.

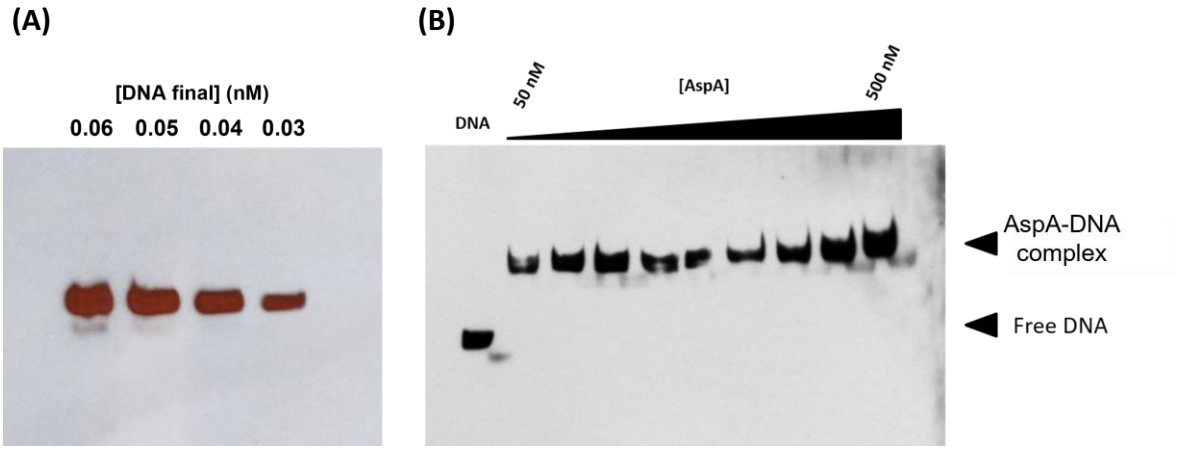


Figure 3.5. Optimisation of EMSA conditions. **(A)** Initial determination of the amount of biotinylated 247 bp DNA fragments to use in subsequent EMSA assays. The DNA was loaded onto a 6% native acrylamide gel using a total volume of 20 µl at the final concentrations shown. **(B)** Initial EMSA with wild-type AspA incubated with the 247 bp DNA fragment. This image shows reactions loaded onto a 4% native acrylamide gel. The gel shift indicative of AspA binding to the DNA occurs immediately at 50 nM final protein concentrations. AspA concentrations used were 0, 50, 75, 100, 125, 150, 175, 200, 250 and 500 nM.

Using a range of protein concentrations from 10 to 500 nM allowed the point of binding to be observed more clearly. At 20 nM, two distinct bands appear, indicative of the formation of low molecular weight AspA-DNA complexes **(Figure 3.6A)**. Complex 1 is presumably the initial nucleation event at the palindrome by a single AspA dimer. At increasing concentrations of protein above 200 nM, the unbound DNA disappears, and at the highest concentration of 500 nM a super-shift of a higher-order complex is visible, where it is likely that AspA has completely spread along the DNA fragment. It should be noted that previous data revealed an AspA:DNA stoichiometry of 3 dimers per 32-mer (Schumacher 2015), therefore the 247 bp fragment used in these assays could potentially accommodate ~23 AspA dimers. The AspA-A53K mutant, capable of binding but not spreading along the DNA, displayed a distinct shift pattern to that of the WT protein. The formation of protein:DNA complexes occurred slightly later at 40 nM concentration, and none of the DNA fragments became fully bound, indicating that spreading was indeed abolished **(Figure 3.6B)**. The AspA-R49A mutant did not bind to the DNA, as evidenced by no observed shift on the gel, indicating that protein:DNA complexes are not formed **(Figure 3.6C)**.

EMSA is a semi-quantitative method of analysing the affinity of a protein to the DNA, allowing an estimation of binding affinity by calculating the apparent dissociation constant ($K_d$app). This was done by measuring the intensity of the unbound DNA bands using a Gel-Doc and associated Image Lab 4.0.1 software (Bio-Rad). Relative band intensities were quantified compared to the 'DNA only' lane, and the mean of three experimental replicates used. The ligand-binding curve was derived by plotting the fraction of bound DNA against AspA concentration, and $K_d$app calculated using the one-site binding equation:

$$\mathrm{y} = B_{max}\,\frac{[\mathrm{AspA}]}{K_d + [\mathrm{AspA}]}$$

where Y is the fraction of DNA bound, B_{max} is the maximal binding, K_d is the equilibrium dissociation constant and [AspA] is the protein concentration. The one-site specific binding model was used due to the initial nucleation event presumably comprising one AspA dimer binding to the palindrome, before additional dimers bind and occupy the DNA

at higher concentrations **(Figure 3.6A)**. A limitation of this model is that it does not incorporate non-specific binding, which may decrease the accuracy of the derived K_d. Additionally, if the protein binds cooperatively, incorporating the Hill coefficient (h) would indicate this, with h > 1.0 indicating positive cooperativity. The $K_d$app of WT AspA was calculated as 21 ± 0.77 nM, indicating high affinity to the DNA, and in close agreement to the previously derived affinity at the first palindrome (~50 nM). AspA-A53K showed reduced affinity with a $K_d$app of 98 ± 4.5 nM **(Figure 3.6D)**.

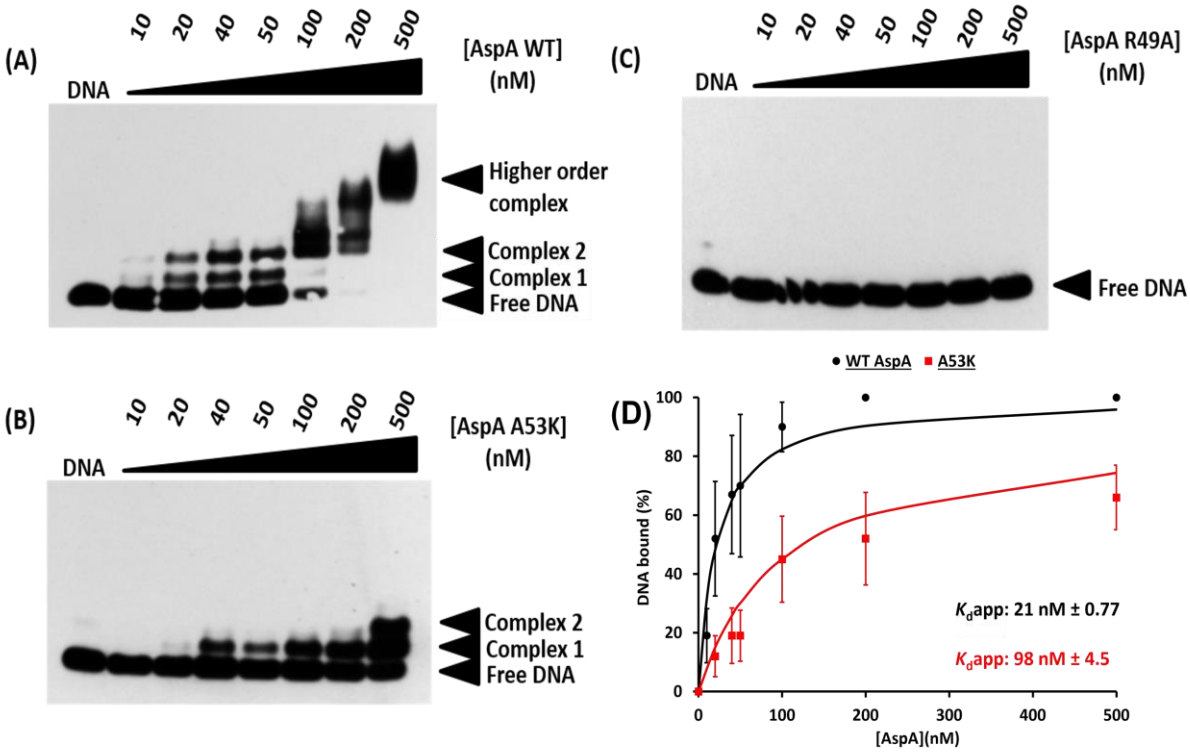


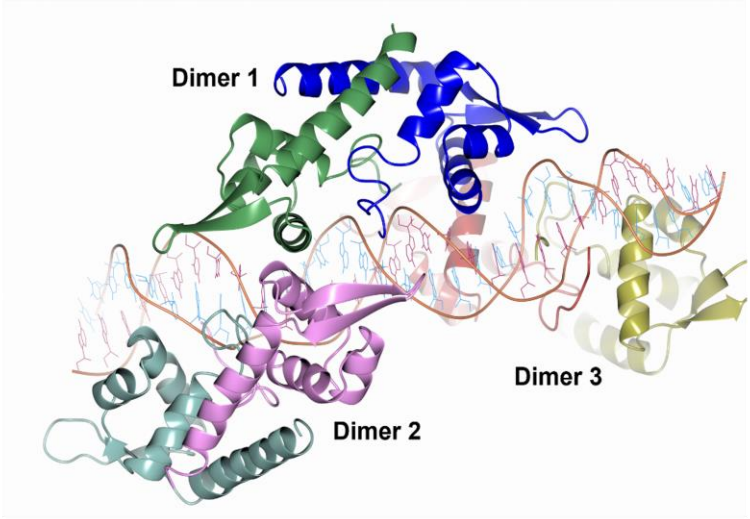
Figure 3.6. EMSA of AspA at the second binding site. **(A)** Representative EMSA in which WT AspA was incubated with the biotinylated 247 bp DNA fragment containing the second binding site. The DNA was used at a final concentration of 0.12 nM, with AspA used at the final concentrations indicated. Free DNA indicates unbound fragments, Complex 1 indicates the formation of a first higher molecular weight AspA:DNA complex, and the Higher order complex indicates a supershift in which all DNA is bound. **(B)**, **(C)** The AspA-A53K and AspA-R49A mutants were also assayed under the same conditions. AspA-A53K is able to bind DNA but is deficient in spreading, whereas AspA-R49 is unable to bind DNA and was used as a negative control. All gel images depict reactions loaded onto a 1.2% agarose gel. **(D)** Ligand binding curve for the AspA WT and AspA-A53K proteins. The percentage of DNA bound was calculated by measuring pixel intensities of unbound DNA using the BioRad Gel-Doc and Image Lab 4.0.1 software, and a mean taken from three experimental replicates. The apparent K_d was calculated with Microsoft Excel (with the additional Solver plugin) using the one-site specific binding model equation. Error bars represent the standard error of the mean.

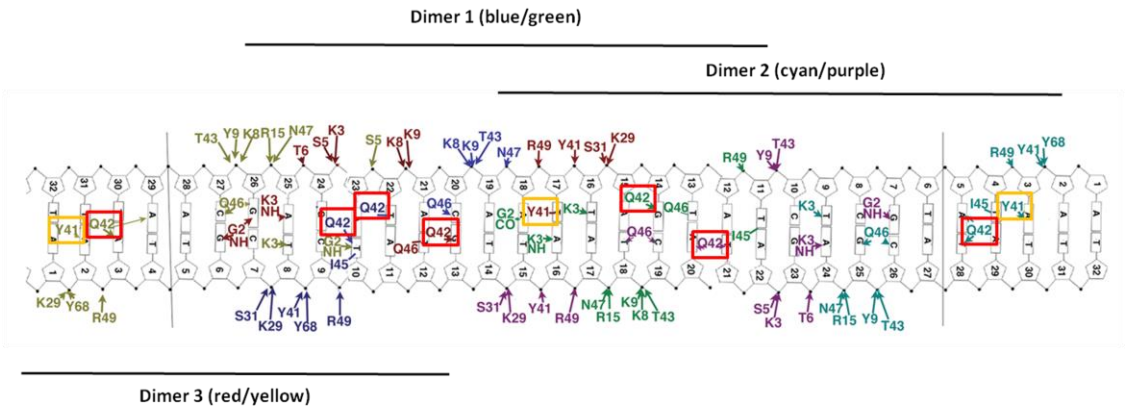
3.2.2 Identification of further AspA residues important for function

3.2.2.1 Rationale for mutant creation

To further probe the interactions of AspA with the DNA at the second palindromic site, four novel mutant proteins were created. The co-crystal structure of AspA-DNA has already been solved (Schumacher 2015), and was used to determine which residues make contacts with the DNA **(Figure 3.7, top)**. Glutamines at positions 42 and 46 from each monomer contact single bases, and tyrosine 41 contacts DNA bases in half of the monomers in the structure (three monomers: yellow, green and cyan **Figure 3.7, middle**), along with several contacts with the phosphate backbone. An alignment of the AspA protein against other homologues in the genus *Sulfolobus* was produced using Clustal Omega **(Figure 3.7, bottom)**. Tyrosine 41 was conserved in all strains of *S. islandicus* and *S. solfataricus* used in the alignment, and glutamine 42 was conserved in half the strains. The 'mutational sensitivity' feature of Phyre2 was used to determine which amino acid substitutions would be predicted to leave the mutant protein structure unaffected (Kelley 2015). Alanine was chosen as the replacement residue; it is non-bulky due to its side chain not extending beyond the beta-carbon and therefore is not expected to cause steric interference or considerable conformational change (Ziolkowska 2006). Alanine substitutions (dubbed ‘scanning mutangenesis’) has previously been utilised in the Barillà lab when investigating the function of the site-specific DNA-binding protein ParG (Barge, thesis 2015), and has been used as a strategy to characterise the functional residues of a *Bacillus thuringiensis* toxin (Howlader 2010).

To further assess AspA-DNA interactions, two AspA residues were chosen for mutation that are hypothesised to be involved in dimer-dimer interactions, and thus facilitate spreading of the protein on the DNA (Schumacher 2015). The negatively-charged glutamic acid 54 of one dimer interacts with positively-charged lysine 59 of the adjacent dimer, whilst hydrophobic residues leucine 52 and alanine 53 from adjacent dimers are subject to van der Waals interactions. The Clustal Omega alignment shows that E54 is well conserved at this position, and L52 is conserved in about half the strains **(Figure 3.7)**.






```
                                                                  41 42          52 54
S. tokodaii              MINMLKQKRYLLFSPEGE-ILIFLYKNPGFHNIDEISESTRIPRSTISQKIPELQIYELI      59
M. sedula                MINMLKQKRYLLFSPEGE-ILIFLYKNLGFHNIDEISEITRIPRSTISQKIPQLQIYQLI      59
S. islandicus LD85       MVGER----DPKGIPM-YRILAFLKKKKEAKTITEISEEADIPYNTVRDNIYKLLALGAI      55
S. solfataricus 98/2     --------MDRLLEPEIQ-ILLNL--SKGKKKRIEIAKECNIPYTTLIYLIPKLEAEGLV      49
S. islandicus M1627      --------MDRLLEPEIQ-ILLNL--SKGKKKRIEIAKECNIPYTTLIYLIPKLEAEGLV      49
S. islandicus LS215      --------MDRLLEPEIQ-ILLNL--SKGKKKRVEIARECNIPYTTLIYLIPKLEAEGLV      49
NOB8H2 pNOB8 AspA        -MGKISTDKYIFLTPRAYIIVHLL--KVGKAKASEISENTQIPYQTVIQNIRWLLAEGYV      57
S. islandicus YG57       -MGKISTDKYIFLTPRAYIVIYLL--KVGKAKASEIAESTQIPYQTVIQNIRWLLAEGYV      57
S. islandicus YN15       -MGKISTDKYIFLTPRAYIVIYLL--KVGKAKASEIAESTQIPYQTVIQNIRWLLAEGYV      57
S. islandicus M164       -MGKISTDKYIFLTPRAYIVIYLL--KVGRAKASEISESTQIPYQTVIQNIRWLLAEGYV      57
```

Figure 3.7. AspA mutational analysis. (Top) The AspA-DNA structure showing three dimers bound to 32-mer DNA. PDB 5K5Q. (Middle) Schematic of AspA-DNA interactions at the palindromic sequence. Shown is the DNA 32-mer that accommodates three AspA dimers, as depicted in **Fig 3.1**. Glutamine (Q) 42 and tyrosine (Y) 41 residues contacting DNA bases are highlighted in red and orange, respectively. Adapted from Schumacher 2015. **(Bottom)** Clustal Omega protein sequence alignment of AspA. *S* = *Sulfolobus*, *M* = *Metallosphaera*. Residues in positions 41, 42, 52 and 54 are highlighted in bold, conserved glutamines and tyrosines are in red, leucines and glutamic acids in blue.

E54 was mutated to alanine, both to negate any charge interactions and due to its non-bulky side-chain. L52 was mutated to lysine, due to its much lower hydrophobicity, and the Phyre2 mutational sensitivity feature was again used to predict a more likely observable phenotypic effect. The hypothesis here was that these mutations would not reduce AspA binding affinity for the DNA, but may reduce the ability of the protein to spread, demonstrating the importance of these residues for dimer-dimer interactions. Here, forward and reverse primers were designed to insert the desired base changes, and AspA mutants AspA-Y41A, AspA-Q42A, AspA-L52K and AspA-E54A were constructed using the QuikChange Lightning Site-Directed Mutagenesis Kit (Agilent Technologies). The locations and various interactions of these amino acids within the AspA:DNA structure are shown in **Figure 3.8**.

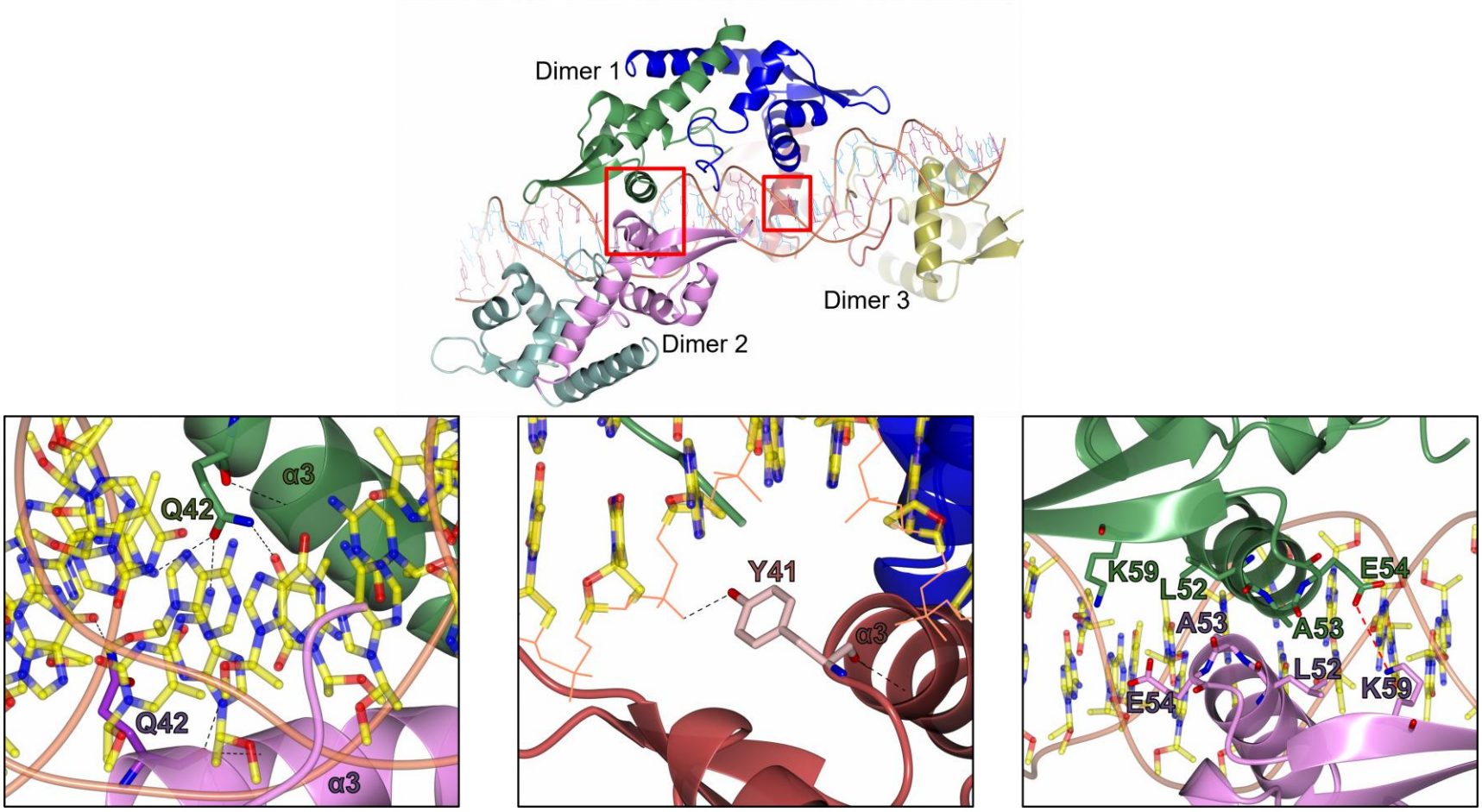


Figure 3.8. Location of mutated residues within the AspA-DNA crystal structure. (Top) AspA:DNA structure showing three AspA adjacent dimers covering the DNA. Red boxes indicate the third alpha helices in three of the monomers. **(Bottom)** As per the schematic in **Fig 3.7**, residues Q42 and Y41 make DNA base and phosphate backbone contacts via hydrogen bonding (black dotted lines, left, middle). Interactions between adjacent dimers are strengthened via electrostatic interactions and van der Waals forces (right). This is shown by a red dotted line between E54 of dimer 1 and K59 of dimer 2; also note the proximity of L52 and A53 between adjacent dimers. In all images, dimers are coloured as in the above structure, DNA bases are yellow, and the DNA backbone is orange. Relevant resides and the alpha helices are labelled. Figure generated using CCP4MG and PDB file 5K5Q.

3.2.2.2 Overproduction and purification of AspA mutants

The correct mutations in the *aspA* gene were confirmed by sequencing **(Figure 3.9A)**. The *aspA* mutant alleles were overexpressed **(Figure 3.9B)**, and the overproduced protein purified using nickel affinity chromatography as previously described, yielding fractions of between 10 and 18 mg for each protein from 300 ml of culture **(Figure 3.9C)**. Before undertaking EMSA experiments with the mutants, various assays were employed to ensure that the mutants behaved similarly to the wild-type AspA protein. Firstly, protein solubility assays were conducted, as mutagenesis can increase or decrease the solubility of a protein, even when there is no underlying structural change (Maxwell *et al*. 1999). Therefore, the solubility of the mutants was assessed by SDS-PAGE before use in downstream assays. This was done by pelleting the culture from a small-scale overexpression (~14 ml of culture) 3 hours after induction with 1 mM IPTG. After cell lysis, the cell suspension was centrifuged, and an equal volume of supernatant and resuspended pellet compared by loading onto a 15% SDS gel (Materials and Methods 2.4.2), alongside aliquots from the overexpression. The majority of the recombinant protein was present in the soluble fraction, compared to the resuspended pellet, as seen by comparing the lower band in the two lanes marked S and P **(Figure 3.9B)**. The higher observed band represents another endogenous protein. All other endogenous proteins are removed after affinity chromatography **(Figure 3.9C)**.

Table 3.1. Summary of AspA mutants produced in this section

Mutant	Description	Source
AspA R49A	DNA binding is abrogated	Barillà lab
AspA A53K	Can bind palindrome but unable to spread	Barillà lab
AspA Y41A AspA Q42A	Y41 contacts bases and phosphate backbone therefore assessing DNA-binding capacity. Q42 contacts bases therefore assessing DNA-binding	This study
AspA L52K AspA E54A	L52 and E54 are hypothesised to be involved in dimer-dimer interactions and aid spreading of AspA on the DNA. Mutants assess for spreading capabilities.	This study

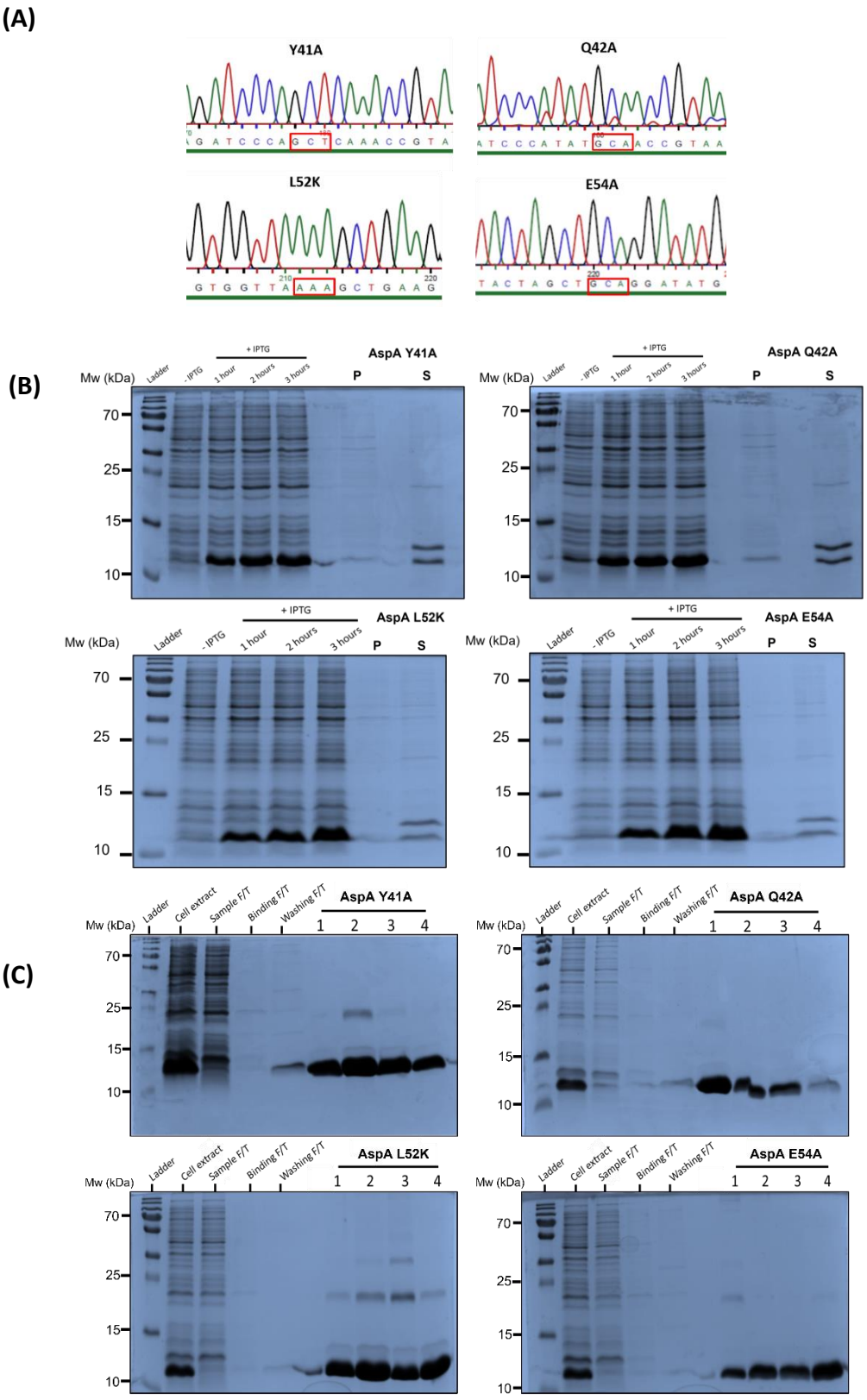


Figure 3.9. Overproduction and purification of AspA mutants. (A) The correct mutations were confirmed by sequencing. **(B)** The mutant AspA proteins were overproduced via IPTG induction as previously described. The mutants were assessed for their solubility; lanes marked P and S indicate the amount of protein present in the pelleted and soluble fractions, respectively, after cell lysis. **(C)** The mutant proteins were purified using Ni^{2+} affinity chromatography, with binding to the column assessed via SDS-PAGE as described for the WT AspA protein. The four most concentrated protein elution fractions were also loaded onto the same 15% polyacrylamide gel.

3.2.3 Amino acid changes do not affect protein structure or behaviour

3.2.3.1 DMP chemical cross-linking

The cross-linking reagent dimethyl pimelimidate (DMP) was then used to demonstrate that the dimerisation, and higher-order oligomerisation properties of the AspA mutants were unaffected by the amino acid substitutions *in vitro*. Chemical cross-linking has been utilised to study protein-protein interactions, as a means of determining sites of interaction and potential binding interfaces (Arora *et al*. 2017, Mintseris & Gygi 2020). A combination of chemical cross-linking and mass spectrometry has been used to define the interaction between an archaeal virus protein and the host *Sulfolobus* RNA polymerase (Sheppard *et al*. 2016). The cross-linking reagent DMP has amine-reactive imidoester groups separated by a 9.2 Å spacer, and covalently links amine groups on lysine side-chains along with the N-termini of peptides. AspA contains eleven lysine residues, of which four are located in the C-terminal dimerisation helix, therefore the protein should be amenable to cross-linking.

Initially, WT AspA was cross-linked alone at 37°C, as bacterial segregation proteins had previously demonstrated cross-linking capacity at this temperature. AspA dimerisation was observed at 37°C, however it was decided to repeat the cross linking at 80°C, as this is a more physiologically relevant temperature for *Sulfolobus* proteins. All subsequent assays were performed at 80°C. Cross-linking experiments were performed as described in Materials and Methods 2.4.10, using a final DMP concentration of 0.1-10 mM. In these experiments, the mass of protein used was not constant across all treatments, as a quantitative measurement of the relative number of dimers, trimers etc. was not planned. Generally, between 20 and 30 µg of protein was used in each reaction, and a qualitative assessment via SDS-PAGE showed that in each case, mutations did not negatively affect the behaviour of the proteins *in vitro*, evidenced by the ability of the proteins to form dimers, trimers, tetramers, and higher order oligomers, similar to wild-type AspA **(Figure 3.10)**.

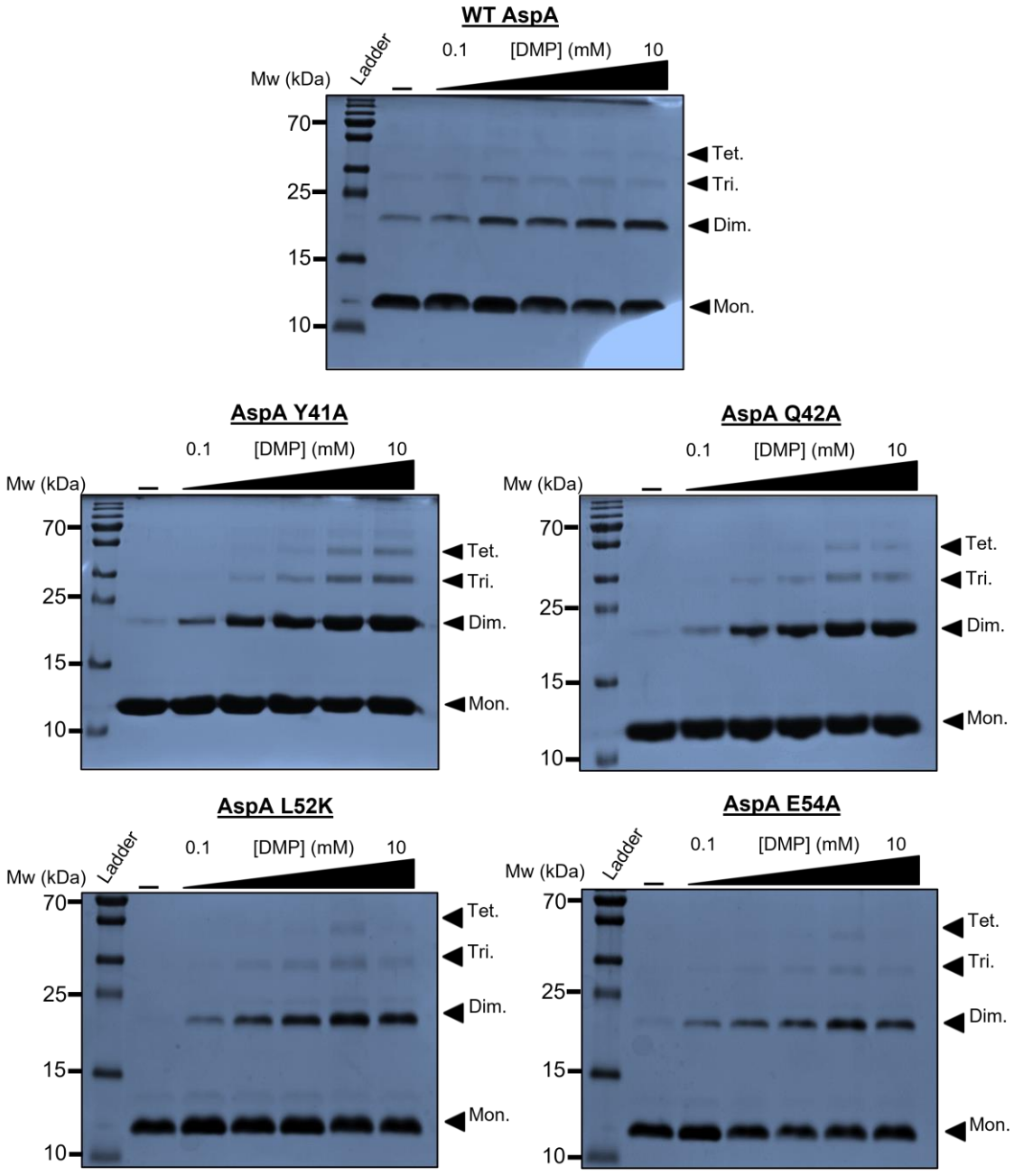


Figure 3.10. DMP cross-linking of AspA mutants. The wild-type AspA and AspA-Y41A, AspA-Q42A, AspA-L52K and AspA-E54A mutants were chemically cross-linked with DMP and analysed via SDS-PAGE. Between 20 and 30 µg of protein was used in each reaction. DMP was added at final concentrations of 0, 0.1, 0.5, 1, 5 and 10 mM. Reactions were incubated at 80°C for one hour, and loaded onto a 15% acrylamide gel. The AspA Mw (monomer) is 11.7 kDa, and bands equating to monomers, dimers, trimers and tetramers are indicated with black arrows. The Mw ladder used in each case is the PageRuler Plus prestained marker (Thermo Scientific).

3.2.3.2 SEC-MALLS and Circular Dichroism

Size Exclusion Chromatography coupled with Multiple Laser Light Scattering (SEC-MALLS) was employed to determine the molecular weights of the mutant proteins in solution, and to assess the oligomeric states of the mutant proteins compared to the wild-type. It is a reliable method for characterising different protein species, and the levels of homogeneity or heterogeneity of the eluted peaks (Some *et al.* 2019). Previous SEC-MALLS data for the WT AspA showed one main peak at ~25 kDa, and a smaller peak from ~60 kDa to ~110 kDa, indicating that the protein formed predominately dimers, but also higher-order oligomeric species in lesser proportions (Schumacher *et al.* 2015). Here, the mutant proteins showed a similar behaviour to the wild-type, with a single peak at approximately the molecular weight of the AspA dimer (theoretical Mw of 23.4 kDa). The SEC-MALLS experiments were conducted in two separate tranches, due to the mutants being constructed at different times. All SEC-MALLS experiments were conducted by Dr Andrew Leech, following provision of the purified proteins (Materials and Methods 2.4.8). The AspA-Y41A and AspA-Q42A proteins had experimental molecular weight estimates of 24.9 and 25.3 kDa respectively, with the WT control being 23.8 kDa. **(Figure 3.11A, left)**. The Mw values for the mutants are within 5% of that of the wild-type, which is within the range attributable to experimental error, and the samples appear homogenous as seen by the Mw lines being close to horizontal across the peaks. Interestingly, no smaller peak at an earlier elution time, equating to higher order oligomers, was seen for either the wild-type or mutant proteins (data not shown). This could be possibly be explained by the increased salt concentrations used in the running buffer compared to the storage buffer (200 mM KCl *cf*. 50 mM), which may prevent association of the protein into greater oligomeric states.

For the AspA-L52K and AspA-E54A mutants, again, there was a similar profile to the WT protein, with one main peak at the approximate molecular weight corresponding to the dimeric protein **(Figure 3.11A, right - red, magenta and blue lines)**. The molecular weight estimates given for the WT, AspA-L52K and AspA-E42A proteins are 21.1, 24.3 and 20.8 kDa respectively, two which are slightly lower than the expected value of 23.4 kDa.

The AspA-E54A mutant Mw estimation curve is consistently horizontal across the peak, indicating homogeneity of the sample and no dissociation (as also seen with the WT). However, the AspA-L52K shows a slight shoulder at a molecular weight of ~31 kDa, indicating a heterogeneous mix of dimeric and trimeric species **(magenta line)**. This appears to be a concentration dependent effect, as at the lower concentrations of 0.6 mg/ml, AspA-L52K displays a horizontal Mw estimate across the entirety of the peak **(Figure 3.11A, right - brown line)**. The diluted AspA-E54A sample **(Figure 3.11A, right - green line)** clearly elutes later than the other samples, and has a lower estimated Mw of 16 kDa, closer to the monomeric form. This could mean that AspA-E54A is reaching a dissociation equilibrium at the lower concentrations, although it should be noted that the trace is quite noisy and this reduces the reliability of the Mw estimate.

In addition, Circular Dichroism (CD) was used to confirm that the mutant proteins retained the same secondary structure elements, and that the overall tertiary structure had not been perturbed due to the mutations of particular residues. CD is a spectroscopic method that relies upon different secondary structure elements within proteins differentially absorbing circular polarised light, resulting in spectra patterns determined by secondary structure composition (Micsonai 2015). All four mutants were included in the CD analysis, along with the WT protein as a control (Materials and Methods 2.4.9). Initially, the spectra for the WT and AspA-Y41A mutant were incomplete (not shown), probably caused by a contaminant (e.g. imidazole remnants from the initial purification). To remedy this, the WT and AspA-Y41A samples were dialysed overnight against the protein storage buffer (50 mM HEPES pH 7.5, 50 mM KCl), which successfully removed the contaminant. The proteins were provided at nominal dilutions of 0.3 mg/ml, as measured by Bradford assay, however, because of the previous contamination it was suggested that sample concentration may be unreliable. To address this, UV absorption-based sample concentrations were determined, and used to scale the CD spectra to that expected for 0.2 mg/ml samples (this stage was performed by Dr Andrew Leech). The resultant CD spectra are consistent across the WT and mutant proteins **(Figure 3.11B)**, and display a predominantly alpha-helical structure (65% of amino acids comprise the four α-helices of AspA). The minima at ~209 and 222 nm are known characteristics of α-helical proteins (Greenfield 2007) along with a peak at 193 nm (not shown).

The CD spectra show that the residue changes introduced in the AspA mutants do not appear to have affected the native secondary structure of the proteins when compared to wild-type.

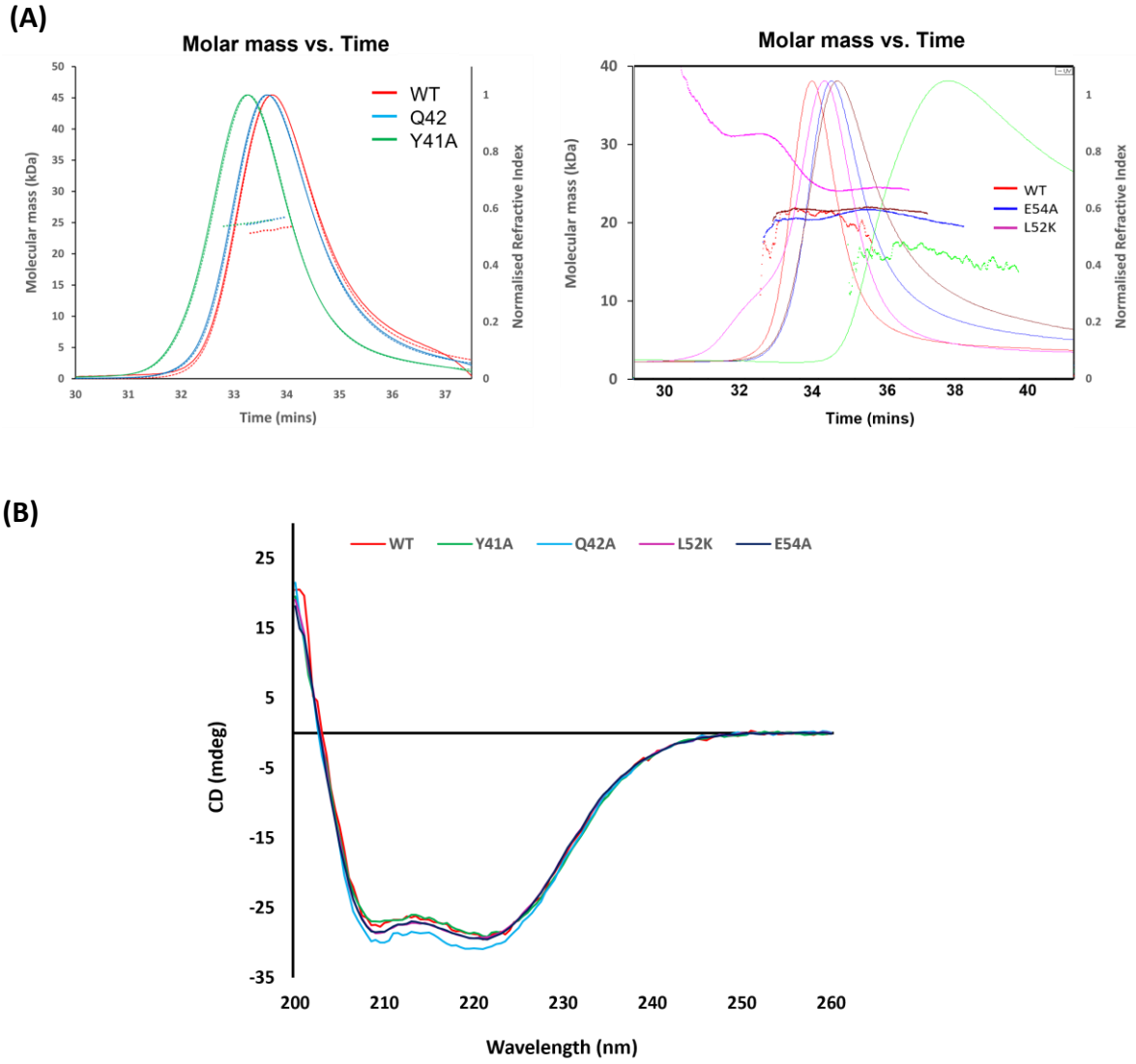


Figure 3.11. SEC-MALLS and Circular Dichroism of AspA mutants. (A) SEC-MALLS. Samples were injected onto a Superdex 200 10/300 GL size-exclusion column. The solid lines represent the refractive index (RI) chromatograms for each protein, dashed lines are the UV traces, with dotted lines across each peak representing the RI molecular weight estimates. (Left) The red, light blue and green curves correspond to the WT, Q42A and Y41A samples at concentrations of 2-3 mg/ml. (Right) The red, magenta and dark blue curves correspond to the UV traces of WT, L52K and E54A samples at concentrations of (2-3 mg/ml). The brown and green curves correspond to diluted L52K and E54A samples at concentrations of ~0.6 mg/ml. The lines across each peak are UV-determined molecular weight estimates. **(B)** CD. Samples were run on a Jasco J-1500 Spectrometer. The CD spectra here was scaled to 0.2 mg/ml for each sample. Only the wavelength over the valid range between 200 and 260 nm is shown. The colour scheme follows that of the SEC-MALLS figure. The figure in **(A**, right**)** was produced by Andrew Leech.

The secondary structure determination sever BeStSel (Micsonai *et al.* 2018) was also used to assess any potential structural changes resulting from mutation. The CD data (wavelength and measured ellipticity) were uploaded to the server, which analyses these data and returns the relative proportions of secondary structure elements. CD data for WT AspA and four mutants, as depicted in **(Figure 3.11B)** were used, and the BeStSel output shown below in Table 3.2. There appear some differences in structure proportions between the WT and mutant proteins, predominantly an increase in the percentage of helices and decrease in beta strands for Q42A, L52K and E54A (Table 3.2). The Y41A mutant shows an increase in both parallel and antiparallel beta strand elements, with a concomitant decrease in 'other', which represents irregular or disordered structures (Micsonai *et al.* 2018).

Therefore, despite the similarity of the CD spectra, some alterations in the folded state for the mutant proteins cannot be ruled out. It should be noted however that the helix percentage of 45.7% for WT AspA is considerably less than the 65% figure mentioned previously, which was derived by counting the number of amino acids in the four helices of the published structure.

Table 3.2. BestSel analysis of AspA WT and mutant secondary structure elements

Protein	Helix (%)	Antiparallel (%)	Parallel (%)	Turn (%)	Other (%)
AspA WT	45.7	7.2	1.1	12.8	33.2
AspA Y41A	44.3	14.3	4.2	13.0	24.2
AspA Q42A	53.3	0.2	1.3	11.6	33.6
AspA L52K	55.3	2.7	0.0	15.0	27.1
AspA E54A	57.3	3.8	0.0	15.9	22.9

3.2.4 EMSA of mutant AspA proteins

3.2.4.1 EMSA of AspA-Y41A and AspA-Q42A exhibit decreased DNA binding

EMSA experiments were again used to investigate the binding affinity of the mutant AspA proteins to the second palindromic site, with initial assays involving AspA-Y41A and AspA-Q42A. As these residues are known to make DNA base and phosphate backbone contacts, it was hypothesised that these mutations may decrease or completely abrogate DNA binding activity, as seen with the AspA-R49A mutant. Previously, a high binding affinity had been observed for wild-type AspA at the second site, with an apparent K_d of ~20 nM **(Figure 3.6)**. Here, both the AspA-Y41A and AspA-Q42A mutants showed decreased DNA-binding capacity, with a band shift only starting to occur at higher concentrations towards 500 nM, as evidenced by the smearing pattern on the gel, however, no discrete protein:DNA complex is apparent **(Figure 3.12, left)**.

Interestingly, these amino acid substitutions do not appear to completely abolish DNA-binding activity as with the previously assayed AspA-R49A mutant, indicating that they may not be essential for this purpose. Another previously assayed mutant, AspA-A53K, is able to bind the DNA at lower affinity compared to the WT, but is unable to spread, due to the bulky lysine side-chain preventing adjacent dimers occupying the same DNA major groove. Here, even at maximum concentrations of 500 nM, the AspA-Y41A and AspA-Q42A EMSA mutants do not show a disappearance of unbound DNA, nor a complete super-shift that may reflect the protein completely coating the DNA. Given the small size of the alanine side chain, lack of spreading is presumably not due to steric hindrance as with the AspA-A53K mutant, indicating that reduced affinity alone is the most likely explanation. It is possible that a spreading/coating pattern would be observed at higher protein concentrations.

Here, and in subsequent band-shift experiments, assays were conducted as triplicate biological replicates, with a wild-type control run simultaneously using the same experimental conditions. EMSA figures such as **Figure 3.12** show one replicate, plus the WT control, and other replicates are included in **Appendix 1.1** to demonstrate the reproducibility of the assay. One limitation of this experimental design is that reactions

were loaded onto four separate agarose gels, therefore this does not control for differences between gels, e.g. gel thickness, and it is possible that any such differences could contribute to the observed shift-pattern.

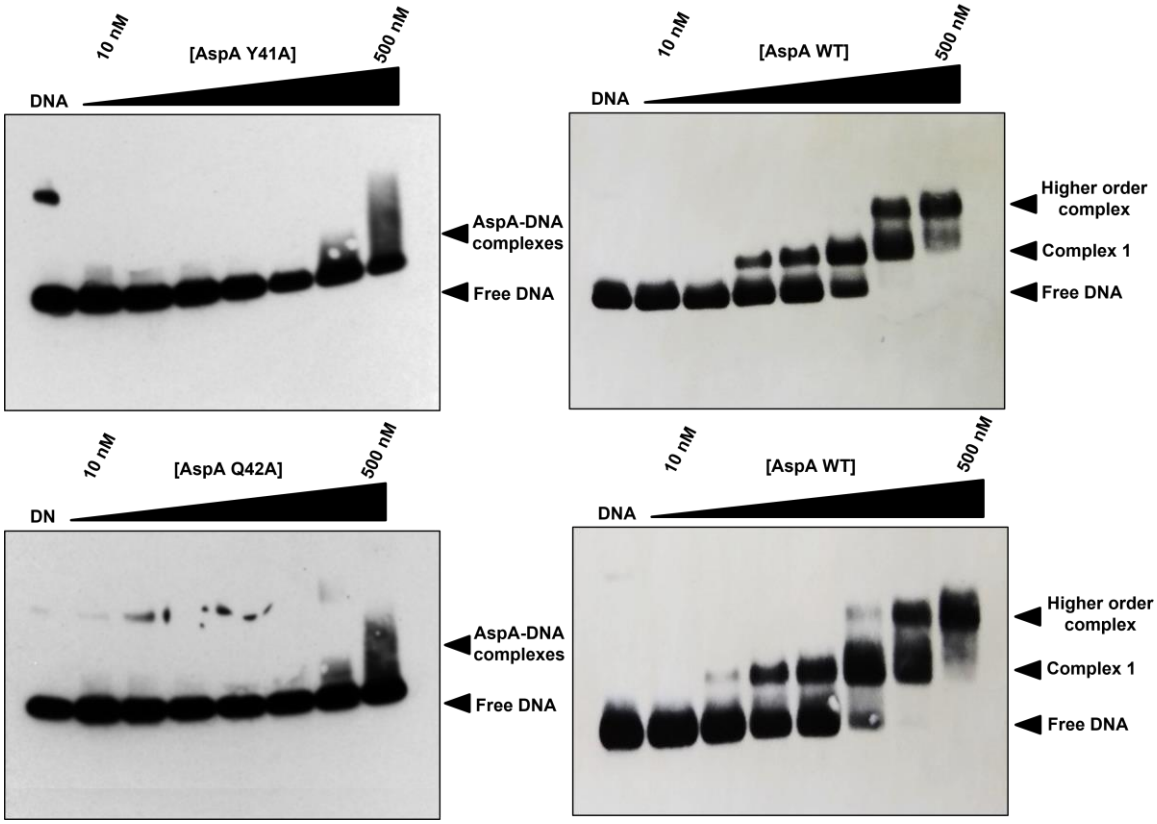


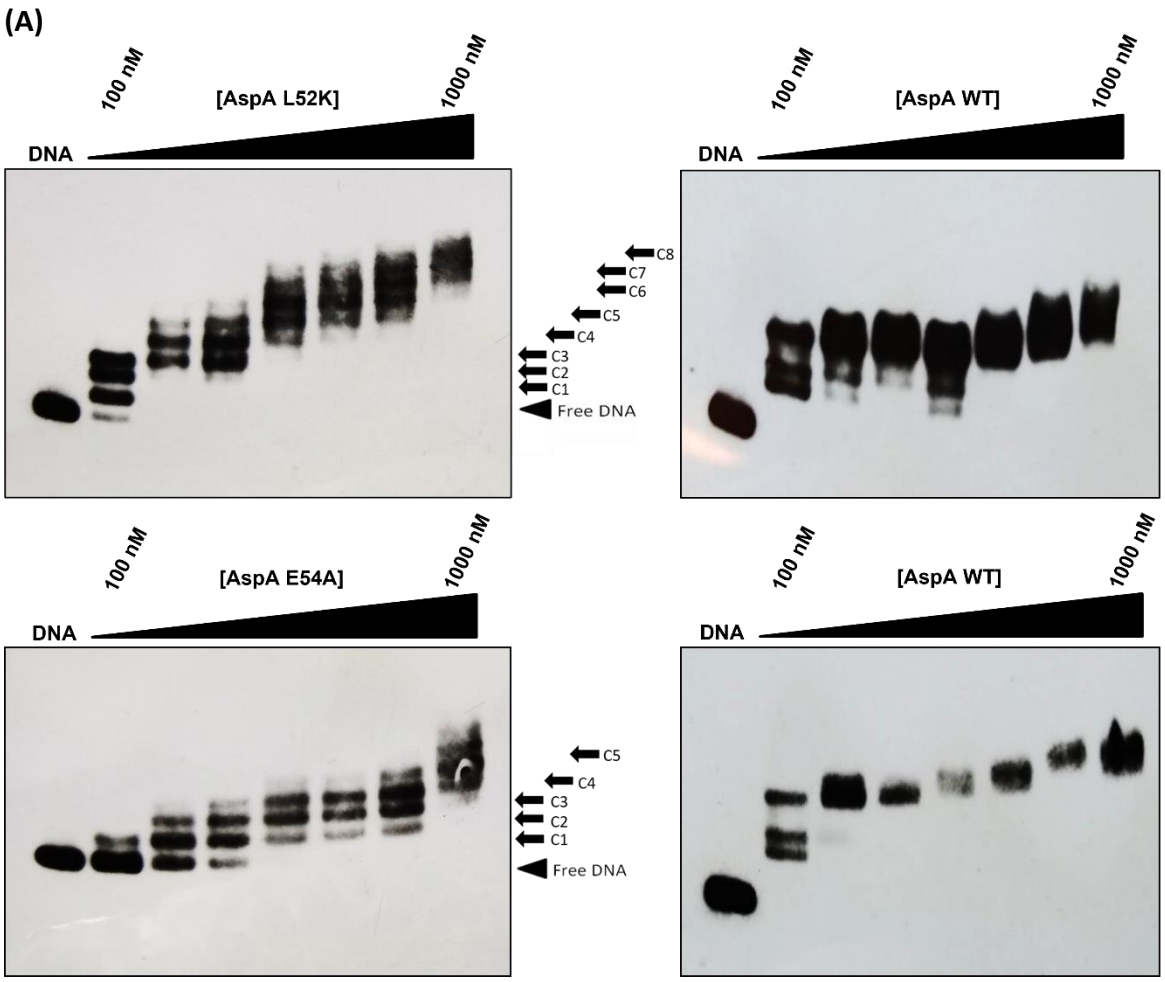
Figure 3.12. EMSA of AspA-Y41A and AspA-Q42A mutants. Proteins were incubated at increasing concentration with a biotinylated 247 bp DNA fragment containing the second palindromic binding sequence, and analysed by agarose gel electrophoresis. Shown are representative images from three experimental repeats for the AspA-Y41A and AspA-Q42A mutants (left). A control EMSA using the WT AspA protein was run concurrently (right). The final concentration of DNA in each reaction was 0.12 nM, and the final protein concentrations were 0, 10, 20, 40, 50, 100, 200 and 500 nM. Free DNA indicates unbound fragments, Complex 1 indicates the formation of a first higher molecular weight AspA:DNA complex, and the Higher order complex indicates a supershift in which all DNA is bound. Reactions were loaded onto a 2% agarose gel.

3.2.4.2 EMSA with AspA-L52K and AspA-E54A dimer-dimer interaction mutants

EMSA experiments were then conducted with the AspA-L52K and AspA-E54A mutants to assess any effects of altering dimer-dimer interactions on both affinity to the DNA, and capability of spreading. Initially, EMSA assays were performed using a higher concentration of protein, 100 – 1000 nM. Both the AspA-L52K and AspA-E54A mutants showed similar band-shift patterns, particularly at lower protein concentrations, where multiple individual, distinct bands are observed, as opposed to one or two bands followed by a smearing pattern as seen with WT AspA **(Figure 3.13A, left)**. These bands could equate to individual AspA dimers, separated by 'gaps' on the DNA due to the mutations lessening dimer-dimer interactions and thus preventing the protein from completely coating the DNA. A control EMSA using the WT AspA protein was run simultaneously under the same experimental conditions **(Figure 3.13A, right)**.

The stoichiometry of the wild-type AspA has previously been determined to be 3 dimers per 32-mer of DNA (Schumacher 2015), therefore for this size of DNA fragment, this equates to ~23 AspA dimers, if the DNA was completely coated **(Figure 3.13B, top)**. If dimer-dimer interactions are abrogated, one such model for mutant AspA binding would be to remove 'every other' dimer, allowing a maximum of ~11 dimers for this size fragment. This is unlikely to be the case however, and it is more probable that dimers bind the DNA in a stochastic fashion after the palindrome is bound, rather than being equally spaced, leading to a mixed population of molecules bound by differing arrangements of mutant dimers **(Figure 3.13B, bottom)**. It appears that the AspA-L52K mutant is able to form more complexes on the DNA compared to the AspA-E54 mutant, with ~8 complexes compared to 5 at the highest concentration of 1000 nM **(Figure 3.13A, left)**. The AspA-L52K mutant also appears to have higher affinity to the DNA, as a complete disappearance of free DNA was seen almost immediately (*cf*. third lane for both mutants).

Complexes were not measured quantitatively, but represent a qualitative interpretation of the band shift data that is consistent with a decrease in dimer-dimer interactions, and therefore adjacent binding and subsequent coating of the DNA, being inhibited. Each discrete complex could hypothetically represent an additional mutant dimer binding non-adjacently to the DNA, and even when no free DNA is present (e.g. E54A lane 5, 500 nM), distinct bands still remain, in contrast to WT. WT AspA previously displayed a band-shift pattern of smearing at 200 – 500 nM, indicating a full range of different molecular weight species up to full covered DNA **(Figure 3.6)**, whereas this is not observed with AspA L53K and AspA E54A at these concentrations. As the distinct complexes are quite difficult to determine, this experiment could be repeated using a larger (taller) agarose gel and longer electrophoresis run to better resolve the individual bands.



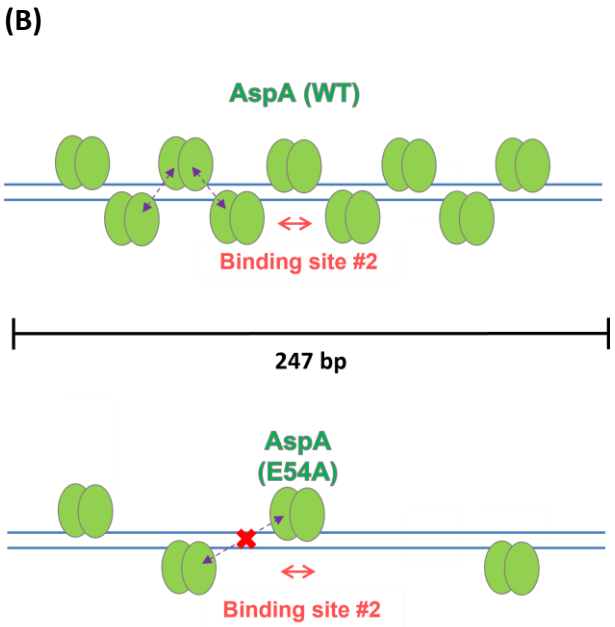


Figure 3.13. EMSA of AspA-L52K and AspA-E54A mutants at higher concentrations. (A) Proteins were incubated with a 247 bp biotinylated DNA fragment containing the second binding site, and analysed by agarose gel electrophoresis. Shown are representative images from three experimental repeats for the AspA-L52K and AspA-E54A mutants (left). A control EMSA using the WT AspA protein was run concurrently (right). The final concentration of DNA in each reaction was 0.12 nM, and the final protein concentrations were 0, 100, 250, 350, 500, 650, 800 and 1000 nM. Free DNA represents unbound molecules. Distinct AspA-DNA complexes are marked by black arrows (C1-C8). Reactions were loaded onto a 2% agarose gel.

(B) Model for how E54A mutations affect spreading. (Top) WT AspA coats the DNA in a helical fashion, with dimer-dimer interactions shown by black arrows. (Bottom) After the initial palindrome binding event, dimers bind but cannot form an extended complex covering the DNA.

To perform semi-quantitative analysis and determine the apparent dissociation constant ($K_d$app), the experiment was repeated at lower protein concentrations of 10 - 500 nM **(Figure 3.14, left)**. The band intensity of the unbound DNA was quantified as previously described, and plotting the percentage of bound DNA against AspA concentration enabled the binding curves and derivation of $K_d$app for each mutant. **(Figure 3.14, right)**. The $K_d$app values for AspA-L52K and AspA-E54A were 2.5 ± 2.1 nM and 78 ± 4.6 nM respectively. It appears that these two mutations do not dramatically reduce binding affinity, as the $K_d$app of WT AspA was 23 nM. Although the binding affinity of AspA-L52K is greater than that of the WT, this may be due to the decreased sensitivity of the assay when using a starting concentration of 10 nM, and repeating these experiments at lower starting concentration may result in a more precise saturation binding curve.

The patterns of the bands on the EMSA figures could also be a result of decreased cooperativity in binding for these two mutant proteins. Amino acid changes at the dimer-dimer interface that act to weaken dimer-dimer interactions may reduce the degree of cooperative binding to the DNA. The initial palindrome binding event could be unaffected due to the high affinity to the DNA as demonstrated by the $K_d$app values, however the lessening of dimer-dimer interactions might result in dissociation of some dimers from the DNA, due to decreased cooperative binding by AspA-L52K and AspA-E54A. This could generate a mixture of molecules comprising AspA-DNA complexes that contain differing numbers of dimers bound to the DNA. Modelling these data has inherent complexities due to potential combinations of specific and non-specific binding, plus the additional factor of cooperative binding following the initial nucleation event at the palindrome. Employing a range of quantitative assays such as Microscale thermophoresis (MST), Isothermal titration calorimetry (ITC) or Surface plasmon resonance (SPR) may allow the level of cooperative binding to be determined and would help address this question.

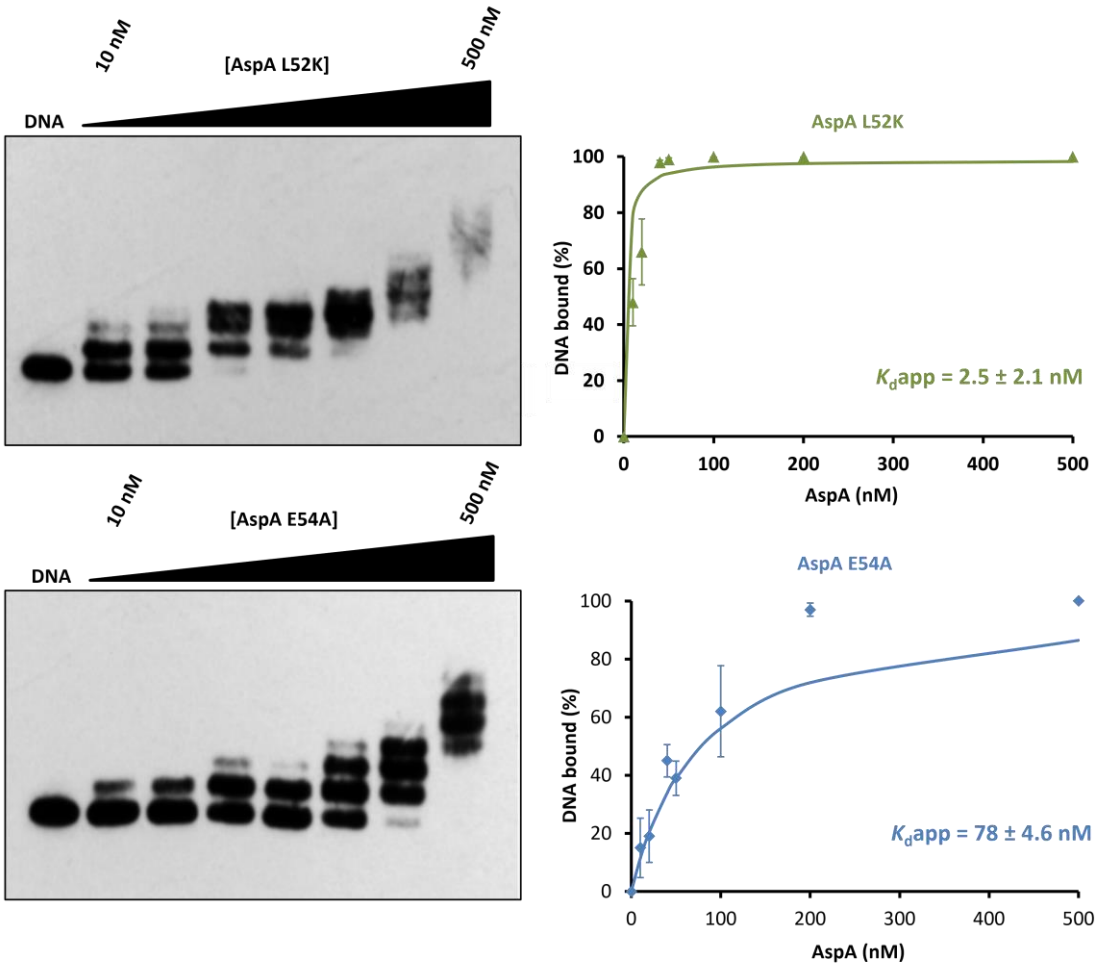


Figure 3.14. EMSA of AspA-L52K and AspA-E54A mutants at lower concentrations. (A) (Left) Representative images from three experimental repeats for the AspA-L52K and AspA-E54A mutants. The final DNA concentration was 0.12 nM, and final protein concentrations were 0, 10, 20, 40, 50, 100, 200 and 500 nM. (Right) Ligand-binding curves were generated using the Excel Solver plugin, using the one-site binding equation. Error bars represent the standard error of the mean.

3.2.5 DNase I footprinting identifies two discrete regions of protection at the second site

3.2.5.1 Optimisation of assay and pilot DNase I footprinting

DNase I footprinting experiments were used to give detailed information about the patterns of spreading of AspA around the second palindromic site. The assay indicates where on the DNA the protein binds, as the bound protein protects the DNA from enzymatic cleavage by deoxyribonuclease (DNase I), producing a distinct 'area of protection' when run on an acrylamide sequencing gel. An A+G sequencing ladder is run alongside the reactions, enabling the protein binding site to be mapped at single nucleotide resolution. This method was used to initially characterise the 23 bp palindrome upstream of the partition cassette to which AspA binds **(Figure 3.15, left)**, and also demonstrated the spreading of the protein upstream from the *aspA* gene at higher concentrations, extending the area of protection (Schumacher 2015). Intriguingly, unpublished data (DB group) shows a different footprinting pattern at the second palindrome, as several discrete zones of protection are apparent, rather than the observed larger region of spreading from the first site. Footprinting assays using the wild-type AspA at the second palindrome were conducted to corroborate and expand on these data, to assess if the different footprinting pattern at the second palindrome was experimentally reproducible.

Initially, as with EMSA experiments, some optimisation of DNA and protein concentrations is required. The DNA used is the same 247 bp biotinylated fragment used in previous EMSA assays, however the final DNA concentration is not necessarily the same, and so a range of concentrations were tested, with a final concentration of 5 nM proving optimal (data not shown). The DNA was incubated with wild-type AspA at increasing concentrations from 100 to 1000 nM, then treated with DNase. At concentrations greater than 500 nM, a discrete region of protection is seen, indicating the site of AspA binding **(Figure 3.15, right)**. However, no information about the DNA sequence that is bound by the protein can be derived, as in this initial experiment, the A+G sequencing ladder that is required to map areas of protection to the actual sequence was not included.

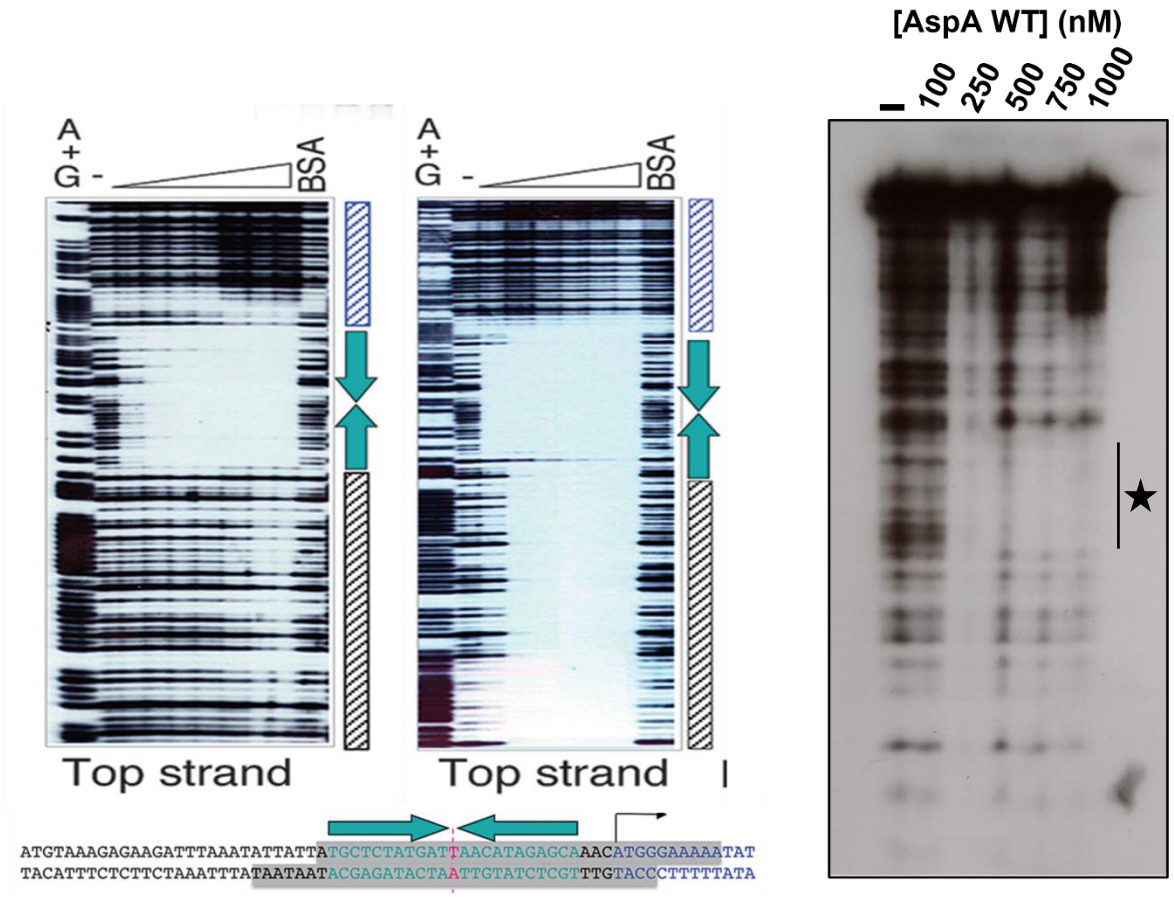


Figure 3.15. Pilot DNase I footprinting at the second palindromic site. (Left) DNase footprinting showing the area bound by AspA at the first binding site, and region of spreading at higher concentrations. The protein binds the 23 bp palindrome (inverted arrows), but also spreads upstream of the *aspA* start codon at higher concentrations (black hatched area). The blue hatched area is the *aspA* gene. Left: 0 – 750 nM, right: 0 – 3000 nM. The gray shaded sequence below the footprints indicates the extent of AspA binding at low to medium protein concentrations. Figure adapted from Schumacher *et al*. 2015. **(Right)** Pilot DNase footprint using the 247 bp DNA fragment which contains the second AspA site, incubated with the wild-type protein at concentrations of 100 - 1000 nM. The final DNA concentration is 5 nM. The first lane indicates no protein, and the putative region of protection where AspA has bound the DNA is indicated with a star.

3.2.5.2 DNase I footprinting with WT, AspA-Y41A and AspA-E54A proteins

After optimisation of the experimental conditions, footprinting was performed at the second AspA binding site. In addition to WT AspA, two mutants were also used to assess their spreading patterns on the DNA and compare with EMSA data. Since the AspA-Y41 and AspA-Q42 residues are involved in protein-DNA interactions, and AspA-L52 and AspA-E52 mediate protein-protein interactions, one mutant of each type was selected for footprinting analysis. Therefore, WT, AspA-Y41A and AspA-E54A proteins were used in this experiment. The same biotinylated 247 bp DNA fragment used in previous EMSA experiments was incubated with wild-type or mutant AspA at two different sets of concentrations (0 - 1000 nM) and (1000-3000 nM), then treated with DNase I, and the purified DNA run on the sequencing gel (Materials and Methods 2.5.2). As a negative control, bovine serum albumin (BSA) was used, as it is not expected to bind to the palindromic site. At concentrations of 750 nM and upwards of WT AspA, a clear window of protection is visible, which corresponds to the 23 bp palindrome **(Figure 3.16, left; Figure 3.17)**. This region of protein binding becomes more apparent at higher concentrations of 1500 - 3000 nM. There is also a second, distinct region of protection higher up on the gel, further upstream of *orf41* and the palindrome **(Figure 3.17)**. This second region has previously been observed in unpublished experiments in the Barillà group, and differs from the region of protection at the first site, where the spreading of AspA for several hundred bases is apparent at concentrations of ~1500 nM upwards (Schumacher 2015). The WT protein may be able to spread from the second site at the highest concentration of 3000 nM, however this effect is not seen across all replicates **(*cf.* last lane Figure 3.16, left; Figure 3.17)**.

The footprints of the two mutant AspA proteins appear to support the EMSA data. The AspA-Y41A mutant, which has previously demonstrated reduced DNA-binding affinity and lack of spreading pattern in EMSA assays **(Figure 3.12)**, here does not result in any region of protection, even at the highest concentrations of 3000 nM, where the pattern of the DNA is the same as at 1000 nM **(Figure 3.16, middle)**. The AspA-E54A mutant, which binds DNA but spreads in a reduced-occupancy fashion, appears to conform to this shift-pattern here. At higher concentrations, the DNA bands become fainter, indicating some binding events are occurring, however a full region of protection does not appear at either the palindrome or the second region of protection **(Figure 3.16, right)**. The palindrome can accommodate 2 - 3 dimers, so in the case of the AspA-E54A mutant, where dimer-dimer interactions are lessened, this region is presumably only being occupied by one or two (separated, non-adjacent) dimers, leaving a region of unoccupied DNA. Protection of the DNA by AspA-E54A is definitely observable at both the palindrome and the upstream second region of protection, however the level of protection is not as clear as with WT AspA. As with EMSA, DNase footprinting is an ensemble technique, comprising many different DNA-protein complexes. Therefore, some DNA molecules may be completely covered at the palindrome and second region of protection by AspA-E54A, whereas other molecules may be only partially occupied, with the resultant image depicting the aggregate of all DNA-protein complexes. Replicate footprinting gel images are included in **Appendix 2.**

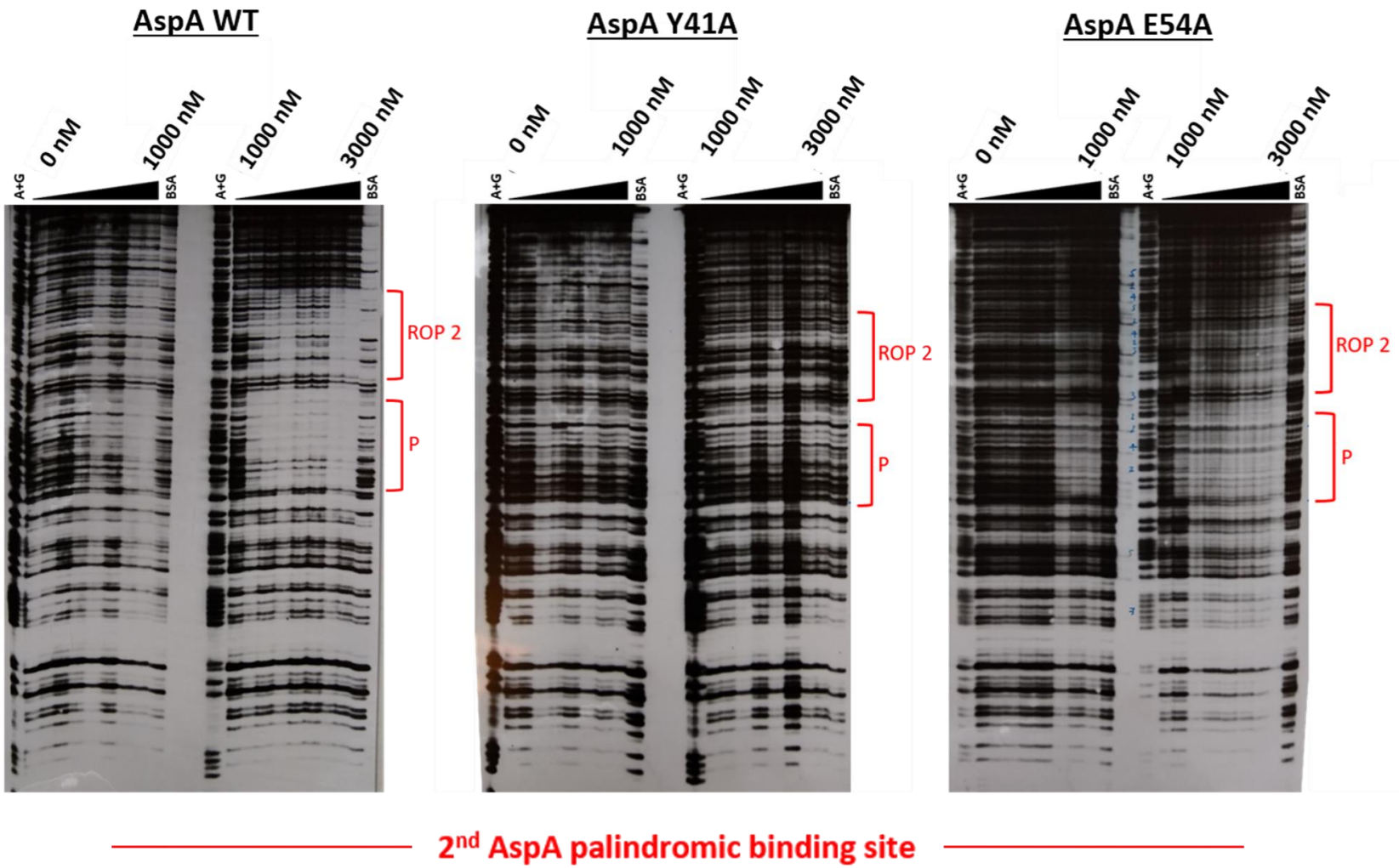


Figure 3.16. DNase I footprinting at the second AspA binding site. A 247 bp biotinylated DNA fragment (final concentration 5 nM) was incubated with WT, AspA-Y41A or AspA-E54A AspA at the following concentrations: lower - 0, 25, 50, 100, 200, 500, 750, 1000 nM; higher - 0, 1000, 1250, 1500, 1750, 2000, 2500, 3000 nM. The BSA control was used at the highest concentration in each set of reactions, i.e. 1000 and 3000 nM. The 23 bp palindrome is marked P, and the second region of protection marked ROP 2. A+G indicates the sequencing ladder, which is the same DNA fragment cleaved at every purine position. These footprints are representative images of at least two experimental replicates.

3.2.5.3 Mapping the second region of protection

The second region of protection bound by the WT protein was mapped back to the sequence using the A+G ladder as reference. This second region is larger than the palindrome, at 35 bp, and is located upstream of the palindrome and pNOB8 *orf41* **(Figure 3.17)**. A putative TATA box, just upstream of the palindrome, has been labelled, as these are often centred ~26-28 nucleotides from the transcription start site in archaeal promoters, and are usually 7 nucleotides in length (Soppa 1999, Peng 2011). The location of the TATA box in archaeal promoters is similar to that found in many eukaryotic promoters, where the TATA box lies ~30 bp upstream of the transcriptional start (Soppa 1999, Xu *et al.* 2016).

Another *cis* acting gene regulatory element found in archaeal promoters is the BRE (Transcription Factor IIB (TFIIB) recognition element), which is also putatively labelled. This sequence, of 7 bp in length and situated immediately upstream of the TATA box appears less well conserved in *Sulfolobus* (Ao 2013). TFIIB (TFB in archaea) is part of the pre-initiation complex that recruits and aids RNA polymerase binding to the promoter, and therefore plays an important role in transcriptional activity. Archaeal TFB is a homologue of eukaryotic TFIIB, and archaea also possess the additional transcription factor TATA box-binding protein (aTBP), which again is homologous to its eukaryotic counterpart (Micorescu 2008). The importance of these two transcription factors in archaea was demonstrated in DNAse I footprinting experiments, where *S. acidocaldarius* TBP and TFB functioned together to bind the promoter and recruit RNA polymerase (Bell & Jackson 2000). Here, the putative TATA box and BRE lie slightly further upstream of the transcription start site than usual, with the TATA box centred at – 37 bp, therefore this region of the sequence requires further characterisation. This could be due to the presence of the palindrome at this location, as several other pNOB8 genes have a putative TATA box at ~ 28 bp upstream of the transcriptional start. However, it does appear that the second region of protection bound by AspA may cover part of the BRE site, leading to the speculation that *in vivo*, AspA could act as a transcriptional regulator at the second site by binding the DNA and preventing the assembly of the pre-initiation complex.

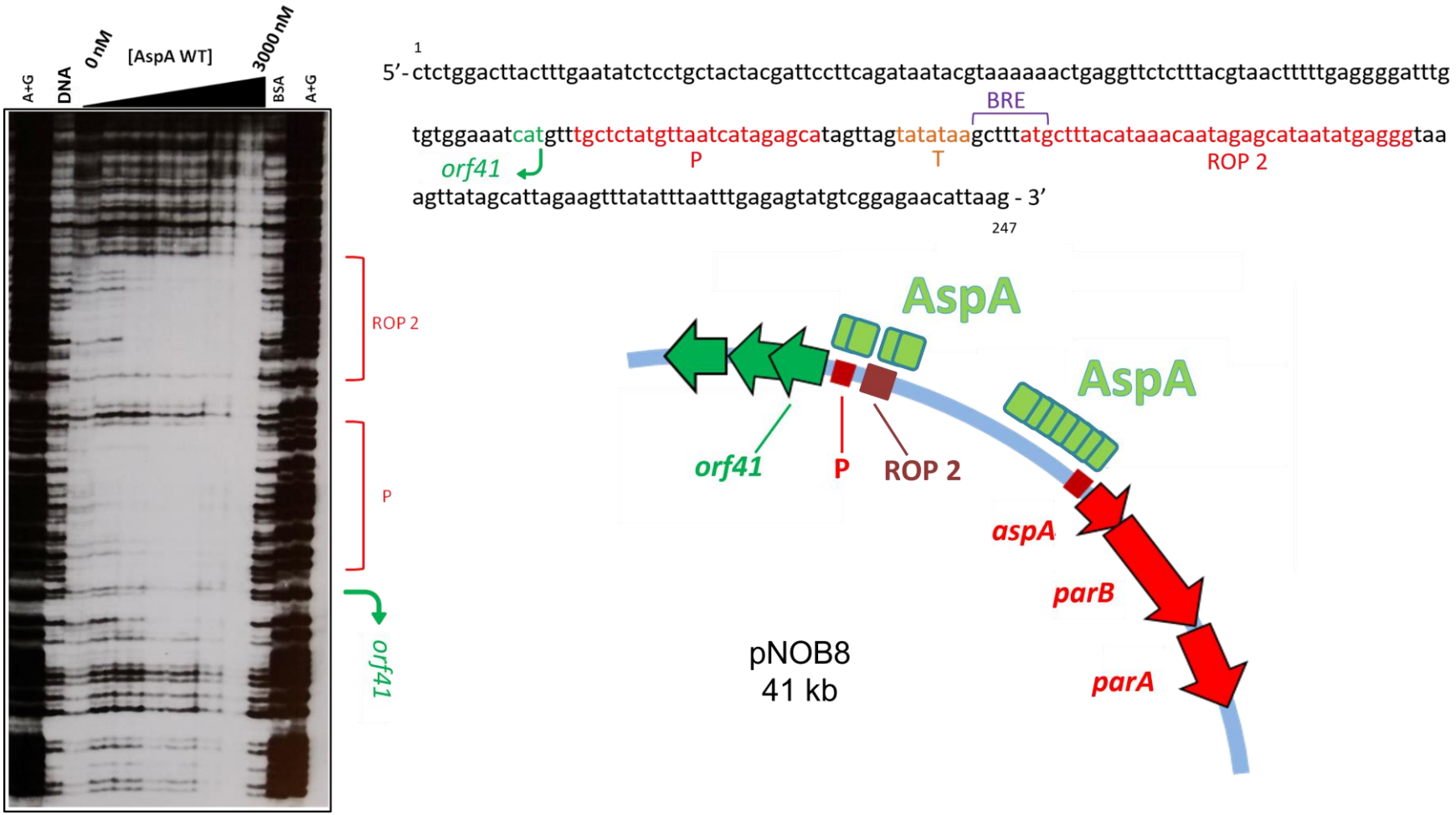


Figure 3.17. The second region of protection and model of transcriptional repression by AspA. (Left) Replicate of WT AspA footprint as in the previous figure. Only the higher concentration of protein (0, 1000, 1250, 1500, 1750, 2000, 2500, 3000 nM) is shown. The palindrome and second region of protection is labelled as before. The green arrow marks the start codon of *orf41* on pNOB8. **(Right)** (Top) Sequence of the 247 bp DNA fragment used in this and EMSA assays. The 5' end of the fragment corresponds to the bottom of the sequencing gel. The palindrome and second region of protection are marked in red. The start codon of pNOB8 *orf41* is marked in green. The putative TATA box within the *orf41* promoter is marked in orange. The approximate region corresponding to the putative B recognition element (BRE) is shown in purple. (Bottom) A region of pNOB8 showing regions of AspA binding; upstream of the *aspA-parB-parA* cassette at the first palindrome, and upstream of *orf41* at the second palindrome as depicted in the footprint.

3.2.6 Assessing residues involved in AspA dimerisation

3.2.6.1 Rationale for mutant creation

In the previous sections, AspA residues were identified which were hypothesised to contribute to the correct functionality of the protein via both DNA-protein interactions, and dimer-dimer interactions, and the effect on both DNA-binding affinity and spreading capability along the DNA for the relevant AspA mutants was measured. Another property of AspA that may be considered important for its correct function is the ability to dimerise, and to bind to DNA in this dimeric form. This property is true for many other site-specific DNA-binding proteins, including those involved in DNA segregation in both bacteria and archaea (Delbrück *et al*. 2002, Schumacher & Funnell 2005, Schumacher *et al*. 2010, Kalliomaa-Sanford *et al*. 2012).

The crystal structure of the apo AspA dimer has previously been solved (Schumacher 2015), and was used to identify residues at the monomer-monomer interface which may be important for dimerisation of the protein and subsequent binding to the DNA. An initial visual assessment of the AspA dimer using the CCP4MG software shows that the monomer-monomer interface is closest at two points: at the N-terminal loops and fourth alpha-helix of each monomer (dimerisation domain). The portion of the N-terminal loops from each monomer in close proximity comprises amino acids 9-12, with resides 82-92 of α4 also being separated by only a few Ångstroms, making them putative candidates for mutagenesis. The 'display close contacts/hydrogen bonds' function of CCP4MG was used to measure both the distance and type of interactions between individual atoms for each of the residues listed above (data not shown). At the N-terminus, both phenylalanine 11 and leucine 12 were promising candidates, displaying both hydrogen bonding and van der Waals interactions, however leucine 12 was more conserved and so was chosen for mutagenesis. At the C-terminus, the process was repeated for the residues listed above in the fourth alpha helix. All hydrogen bonding was intramolecular and involved in forming the alpha-helical structure, whereas all residues displayed intermolecular van der Waals interactions. Isoleucine 85 and valine 89 were chosen for mutagenesis as

these two amino acids contacted two and three amino acids from the other monomer respectively. The structure of the apo AspA dimer, showing the location and intermolecular interactions between these amino acids is shown in **Figure 3.18**. Phyre2 was again used to assess the mutational sensitivity at each position to predict whether these mutations were likely to engender a phenotypic and therefore potential functional effect. All three residues chosen for mutagenesis (L12, I85 and V89) had the highest degree of mutational sensitivity when mutating to glycine, presumably due to glycine possessing the smallest side chain, thereby reducing intermolecular contacts between monomers.

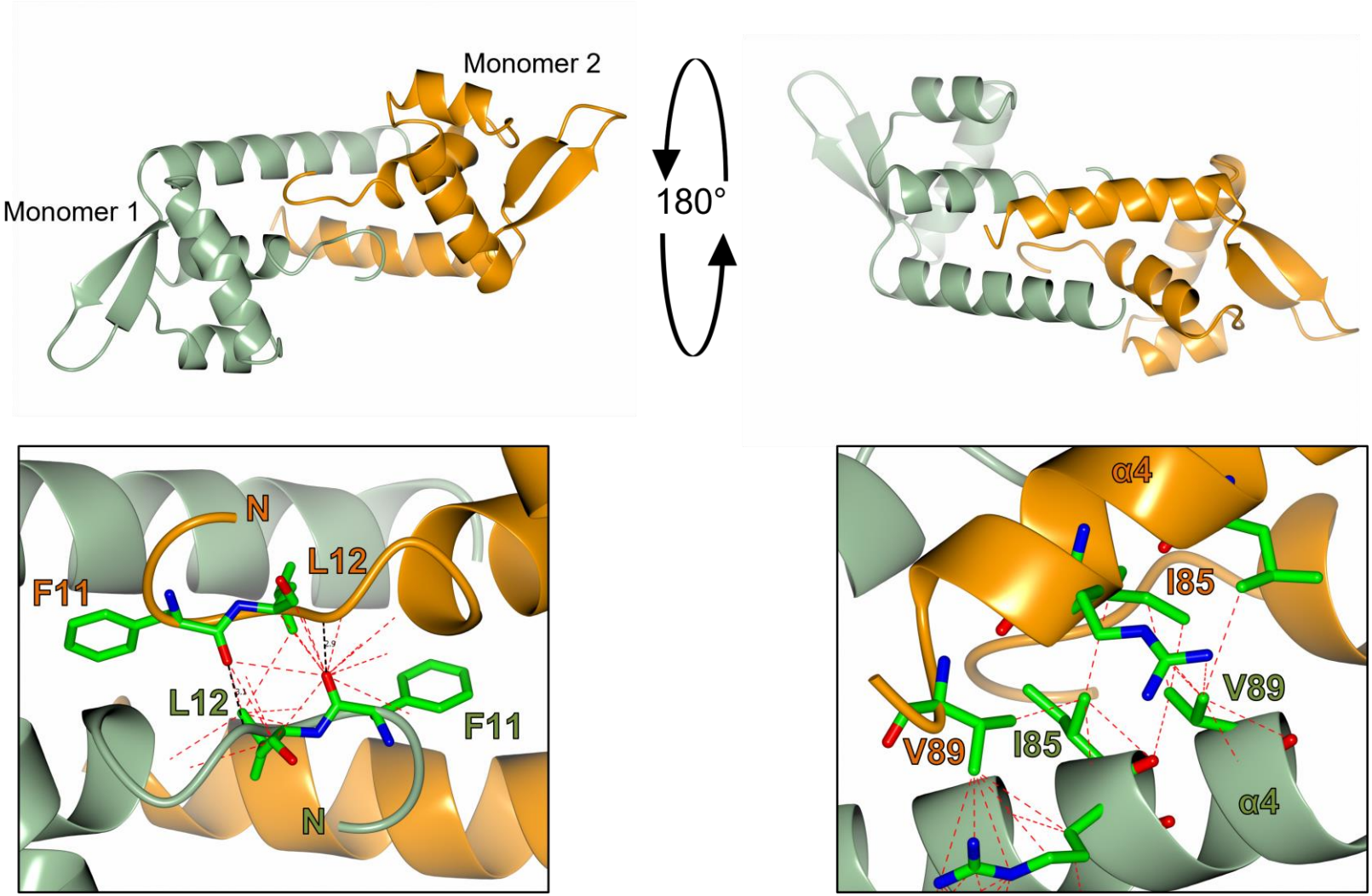


Figure 3.18. AspA dimer structure and location of mutated residues. (Top) The AspA dimer, with one monomer coloured sea green and the other orange. The dimer is rotated 180° to show both the N-terminal loop and the fourth alpha helix. **(Bottom)** Close-up view of the N-terminal loop (left), and α4 (right), with mutated amino acids depicted as cylinders. Hydrogen bonds are depicted as black dotted lines, with the distance in Ångstroms. Red dotted lines indicate close inter- and intra-molecular contacts. Figure generated with CCP4MG (PDB 4RS8).

3.2.6.2 Purification of mutant proteins, SEC-MALLS and CD

Forward and reverse primers were designed to insert the desired base changes, using the pET-22:*aspA* WT construct as a template, and AspA-L12G, AspA-I85G, and AspA-V89G were constructed using the QuikChange Lightning Site-Directed Mutagenesis Kit (Agilent Technologies). The mutagenesis was confirmed by sequencing (not shown), and the AspA-L12G, AspA-I85G and AspA-V89G proteins were overproduced and purified by nickel affinity purification as previously described. Typical yield for each of the proteins was 2 – 3 mg/ml from 300 ml of culture, with fractions totalling 5 ml collected for each protein. Aliquots of each fraction were assessed for purity by SDS-PAGE as previously described, and all showed distinct bands equating to the molecular weight of the monomeric form, with faint dimer bands appearing in some of the AspA-I85G and AspA-V89G fractions **(Figure 3.19A)**.

The mutant proteins were again subjected to analysis by SEC-MALLS and CD to assess any effect of residue change on protein structure and behaviour. SEC-MALLS analysis proved problematic due to a contamination issue in the supplied buffer, which resulted in an increase in light scattering, rendering the data unusable (not shown). The experiment was repeated after some troubleshooting, ensuring that for example all buffer reagents (KCl and HEPES) were freshly prepared and filtered, and all bottles used for buffer preparation were thoroughly cleaned. This alleviated the contamination problems somewhat, but did not completely remove them as there was still a higher degree of light scattering present in the running buffer compared to the sample storage buffer, reducing the precision of the protein molecular weight estimates. Nevertheless, the WT control appeared to have approximately the correct Mw for a dimer, at 20.5 kDa. The measured molecular weights of AspA-L12G and AspA-I85G are slightly lower than to be expected for a monomer, at 8.3 and 10.1 kDa, whereas AspA-V89G is slightly greater at 14.8 kDa, indicating there could be a small degree of dimerisation taking place for this mutant **(Figure 3.19B, left)**. The molar mass lines are also not horizontal for the three mutants, indicating there may be some heterogeneity of species for these samples. These values were for proteins at

concentrations of ~3 mg/ml, reduced concentration samples (at ~0.6 mg/ml) were also used, and gave similar Mw values. These data should be interpreted cautiously given the reduced quality, but are suggestive that the residue changes have affected the dimerisation of the protein in solution.

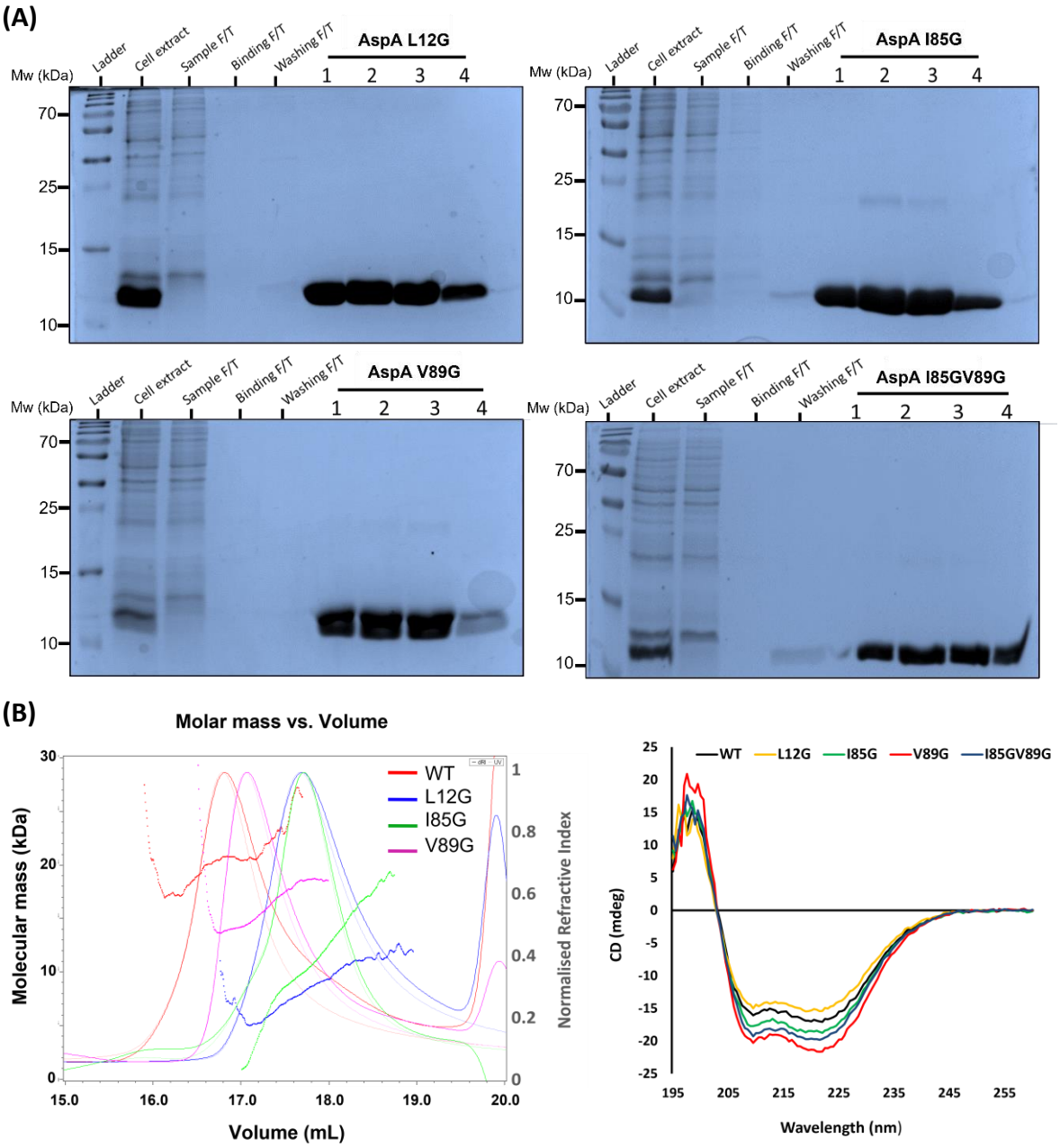


Figure 3.19. Purification and structural assessment of AspA dimerisation mutants. (A) The four AspA mutant proteins were purified using Ni^{2+} affinity chromatography, and the results assessed via SDS-PAGE as described for the WT AspA protein. The four most concentrated protein elution fractions were also loaded onto the same 15% polyacrylamide gel. **(B)** (Left) SEC-MALLS for the monomer mutants. Solid-line peaks represent the refractive index, dotted-line peaks are UV traces. Molar mass estimates are derived from the light-scattering lines across the peaks. (Right) CD. Samples were run on a Jasco J-1500 Spectrometer. The CD spectra were obtained at a nominal concentration of 0.3 mg/ml for each sample. Only the wavelength over the valid range between 195 and 260 nm is shown.

Circular dichroism experiments were untaken with the proteins, to again assess any potential structural change brought about by mutagenesis. Prior to this, a double mutant was constructed, using the single *aspA*^{I85G} gene as a template, and using the QuikChange site-directed mutagenesis kit as previously described to create *aspA*^{I85GV89G}. The rationale for doing this was to see if any single-mutant property was amplified if a double mutation further decreased the dimerisation of the protein. The mutation was confirmed by sequencing, the protein overproduced and purified and analysed by SDS-PAGE **(Figure 3.19A, bottom right)**. The AspA-I85GV89G mutant was produced after the SEC-MALLS experiment, and so is not included in those data, however, CD experiments were conducted using all four mutants plus the WT control.

The proteins were all dialysed against AspA storage buffer overnight to avoid any potential contamination issues that occurred previously. It was thought that a residual contaminant from the purification process negatively affected CD experiments for AspA WT and AspA-Y41A proteins (Section 3.2.3.2), therefore dialysis was performed prior to CD for this set of proteins. Protein concentration was measured by Bradford assay, and proteins were diluted to a nominal concentration of 0.3 mg/ml. All mutants displayed similar CD spectra to WT AspA **(Figure 3.19B, right)**, with minima at wavelengths characteristic of alpha-helical proteins (208 and 222 nm, Greenfield 2006). However there was some difference in amplitude, i.e. the amount of circular dichroism (as measured on the y-axis). The differences in amplitude are probably due to concentration differences as previously, UV-based concentration values were derived and were found to differ from the nominal value. This was not done for these samples, and so some structural changes due to mutation cannot be ruled out.

The BestSel web server was again used to assess the proportions of secondary structure elements of the WT AspA and mutant proteins (Micsonai *et al*. 2018). The CD data (wavelength and measured ellipticity) was uploaded to the server, and the secondary structure proportions for the five proteins are shown below in Table 3.3. Strikingly, the proportion of helices is far below what is expected for AspA based on the structure (65%), with all proteins showing helices accounting for less than 30% of structural elements. However, the proportions are consistent between the proteins, particularly WT AspA, AspA-L12G and AspA-I85G, indicating that secondary structure is preserved (Table 3.3).

The BestSel server requires the user to input the molar concentration of the protein, and changes in concentration dramatically alter the structural proprtions. Reducing the molar concentration results in a much greater percentage of helices (not shown), therefore it is likely that the nominal concentration of 0.3 mg/ml is incorrect and too high. The experiment should be repeated with exactly equal starting concentrations in order to give more reliable spectra.

Table 3.3. BestSel analysis of AspA WT and mutant secondary structure elements

Protein	Helix (%)	Antiparallel (%)	Parallel (%)	Turn (%)	Other (%)
AspA WT	16	31.3	0.0	15.1	37.7
AspA L12G	16.5	28.7	0.5	14.1	40.2
AspA I85G	18.7	27.1	0.0	15.4	38.8
AspA V89G	23.5	23.8	0.0	14.2	38.5
AspA I85GV89G	22.2	26.7	0.8	13.1	37.3

3.2.6.3 DMP cross-linking shows AspA dimer mutants form fewer complexes

Dimethyl pimelimidate (DMP) was again used in chemical cross-linking experiments, as WT AspA and the four mutant proteins previously created were shown to dimerise at the physiologically relevant temperature of 80°C (Section 3.2.3.1). Here, DMP cross-linking was conducted in a quantitative fashion to measure any reduction in the propensity of the mutant proteins to dimerise. None of the mutations targeted lysine residues, therefore the proteins should still be amenable to DMP cross-linking, however the hypothesis was that a reduction in intramolecular contacts between residues of different monomers may result in more transient interactions and therefore less dimerisation. These experiments were conducted using a fixed mass of 20 µg of protein in each reaction, and were conducted in triplicate. The reactions were incubated at 80°C for 1 hour with increasing concentrations of DMP from 0.1 to 10 mM. The amount of cross-linking was assessed by SDS-PAGE. For each of the mutant proteins, along with the wild type, dimerisation and formation of higher-order oligomers was apparent, though to differing extents **(Figure 3.20)**. Replicate gel images are included in **Appendix 3.**

To quantify the amount of dimerisation, the intensity of each dimer band was measured using a Gel-Doc (Bio-Rad) and associated Image Lab 4.0.1 software. The dimer bands were measured for each DMP concentration, using an equal-sized portion of the gel image background as a relative intensity of 1 for comparison (this was done separately for each gel image as the background intensity was unequal across images). The mean relative band intensities from three experimental replicates (including WT) were plotted as a function of DMP concentration. Interestingly, it appeared that the AspA-L12G mutant formed a greater proportion of dimers compared to the wild-type, along with more trimers and tetramers **(Figure 3.20)**. One explanation could be that this mutation induced a slight change in structure, or the flexible N-termini loops could be brought in closer proximity due to the smaller glycine side-chain, however this result does not correlate with the SEC-MALLS data that showed the AspA-L12G mutant having the smallest estimated molecular weight **(Figure 3.19B)**. It is possible that chemical cross-linking captures the short-range but transient interactions of AspA-L12G monomers.

The other two single mutants however, AspA-I85G and AspA-V89G, formed a lesser proportion of dimers compared to wild-type, as predicted, and interestingly, the double mutant AspA-I85GV89G formed the lowest proportion of dimers out of all the proteins **(Figure 3.20)**. This suggests that perhaps alpha helix 4 plays a more prominent role in monomer-monomer interactions, given that mutations here negatively affect dimerisation more so than the mutation in the N-terminal loop. These results also correlate more positively with the SEC-MALLS data that was suggestive of the mutant proteins being predominantly monomeric in solution. The proteins are spatially close enough to be cross-linked as evidenced by the gel images, however perhaps in their native state (without the addition of a cross-linker), the intramolecular forces are sufficiently weakened to lessen, or prevent dimerisation. The band intensities of the dimers at 10 mM were plotted separately, and the relative band intensities were found to be significantly different between both the WT and AspA-V89G mutant ($p < 0.05$), and the WT and double mutant AspA-I85GV89G ($p < 0.01$). A two-tailed, unpaired T-Test was used to test for significance. These differences are based on a limited sample size of three replicates however, and thus, should be interpreted with a degree of caution.

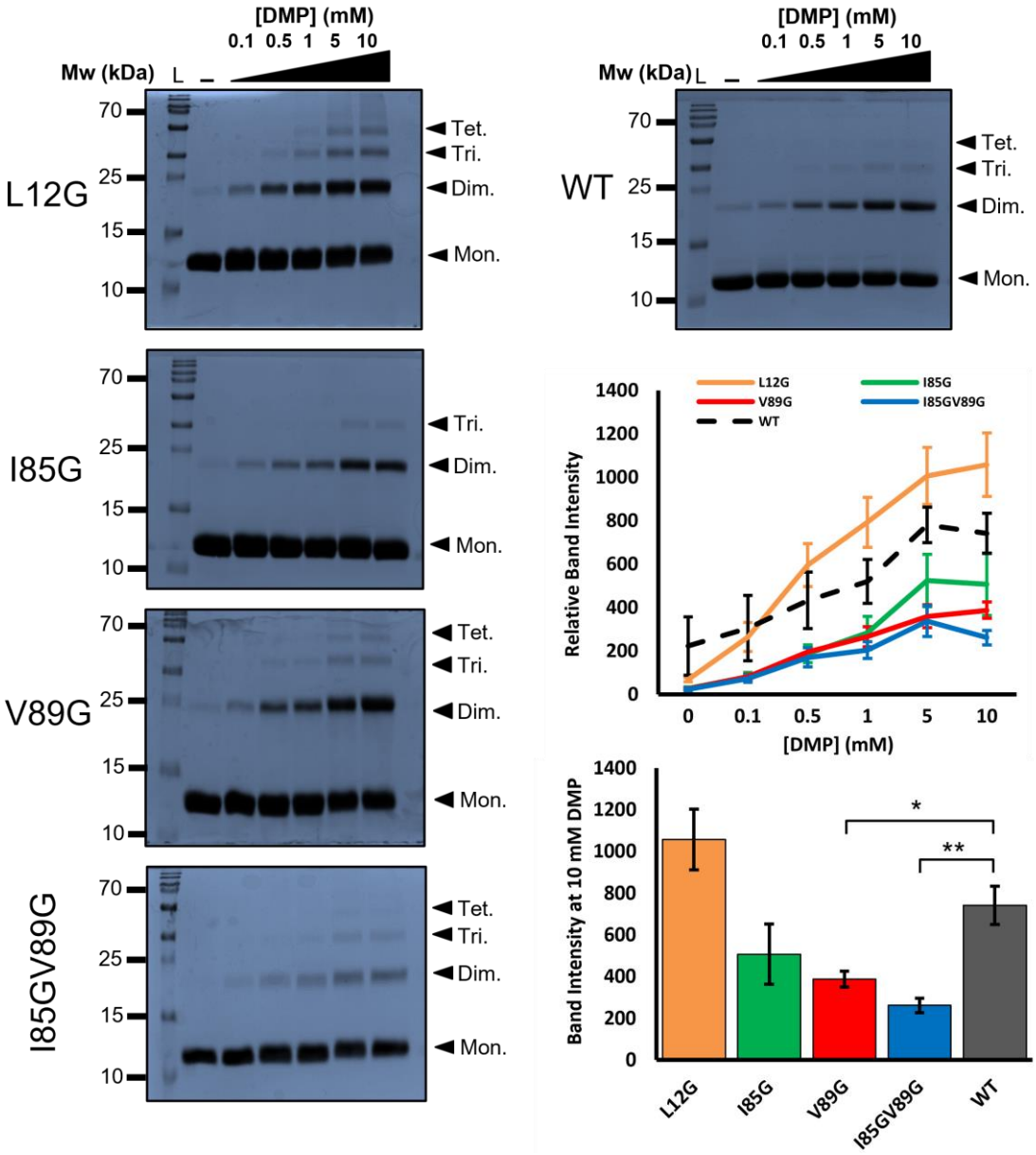


Figure 3.20. DMP cross-linking of AspA dimerisation mutants. (Left) The AspA-L12G, AspA-I85G, AspA-V89G and AspA-I85GV89G mutants were chemically cross-linked with DMP and analysed via SDS-PAGE. A fixed mass of 20 µg protein was used in each reaction. DMP was added at final concentrations of 0, 0.1, 0.5, 1, 5 and 10 mM. Reactions were incubated at 80°C for one hour, and loaded onto a 15% acrylamide gel. The AspA Mw (monomer) is 11.7 kDa, and bands equating to monomers, dimers, trimers and tetramers are indicated with black arrows. The Mw ladder used in each case is the PageRuler Plus prestained marker (Thermo Scientific). Images are representative of three experimental replicates, and a WT control was performed simultaneously alongside each mutant (a representative image is shown top right). **(Right)** The relative band intensity was calculated by comparison of dimer bands at each DMP concentration relative to the gel background, using the BioRad Gel-Doc and Image Lab 4.0.1 software. Means of three experimental replicates were plotted (middle), with error bars representing the standard error of the mean. The relative band intensities of the 10 mM dimer bands of each mutant plus WT were plotted separately (bottom), and a two-tailed, unpaired T-Test was used to test significance between each protein (* = $p < 0.05$, ** = $p < 0.01$). Error bars represent standard error of the mean.

3.2.6.4 EMSA shows AspA dimerisation mutants bind less avidly to the DNA

The AspA mutants were used in EMSA assays, to assess their capacity to dimerise and thus bind DNA, and to correlate any resultant shift-pattern with previous experiments. EMSA assays were conducted according to the standard protocols used previously (Materials and Methods 2.5.1). The proteins were used at concentrations from 100 – 1000 nM rather than 10 to 500 nM, as the intention was to qualitatively assess binding ability, rather than quantify the apparent dissociation constant as with the WT, AspA-L52K and AspA-E54A proteins **(Figures 3.6, 3.14)**. EMSA assays were conducted in triplicate for each mutant, and the WT protein, was used concurrently as a control to ensure the assay was working correctly. Replicate band-shift data is shown in **Appendix 1**.

The AspA-L12G mutant showed no binding to the DNA, as there was no apparent band shift on the agarose gel even at 1000 nM protein concentration **(Figure 3.21)**. This would indicate that DNA binding activity is abrogated due to the inability of AspA-L12G to dimerise, which is in agreement with the SEC-MALLS data that demonstrates that AspA-L12G appears mostly monomeric **(Figure 3.19B)**. However, this contradicts the cross-linking data which showed greater dimerisation than the WT **(Figure 3.20)**. It is possible that AspA-L12G monomers are close enough spatially to be chemically cross-linked, but under native conditions, interactions are less stable and too transient to favour the formation of permanent dimers and thus permit DNA binding. It is interesting that this single mutation in the AspA N-terminal loop abrogates DNA binding altogether, suggesting that L12 is required for the protein to dimerise.

The AspA-I85G and AspA-V89G mutants did bind to the DNA, albeit with slightly lower affinities than the WT, with a complete disappearance of free DNA occurring at 350 – 500 nM, compared with 100 – 250 nM for WT AspA, suggesting that these mutations may slightly reduce the spreading capability of the proteins **(Figure 3.21)**. The EMSA data correlates with the DMP cross-linking results, where both these single mutants formed a lower proportion of dimers compared to the WT protein, which would result in decreased binding capacity. Interestingly, the double mutant AspA-I85GV89G was unable to bind the DNA at all, and demonstrated the same band-shift pattern as for

AspA-L12G. Again, this finding is in agreement with the cross-linking data, as AspA-I85GV89G formed the lowest proportion of dimers **(Figure 3.20)**. These EMSA data suggest that intramolecular interaction between residues in the dimerisation helix of AspA (α4) may have a degree of redundancy, and that single mutations here are not sufficient to abolish dimerisation and subsequent DNA-binding, whereas a single mutation in the N-terminal loop abrogates DNA binding completely. Both the N and C-terminal domains are therefore important (though perhaps to differing extents) in stabilising the AspA dimer and thus its subsequent interactions with the DNA.

It should be noted that there is some variability in the four WT control band-shifts, with a full shift occurring mainly at 250 nM, but at 100 nM one instance **(Figure 3.21)**. Theses band-shift experiments, although performed in triplicate plus the WT control, were done using four different agarose gels, and so it is possible that gel-specific effects are responsible for the observed differing shift patterns.

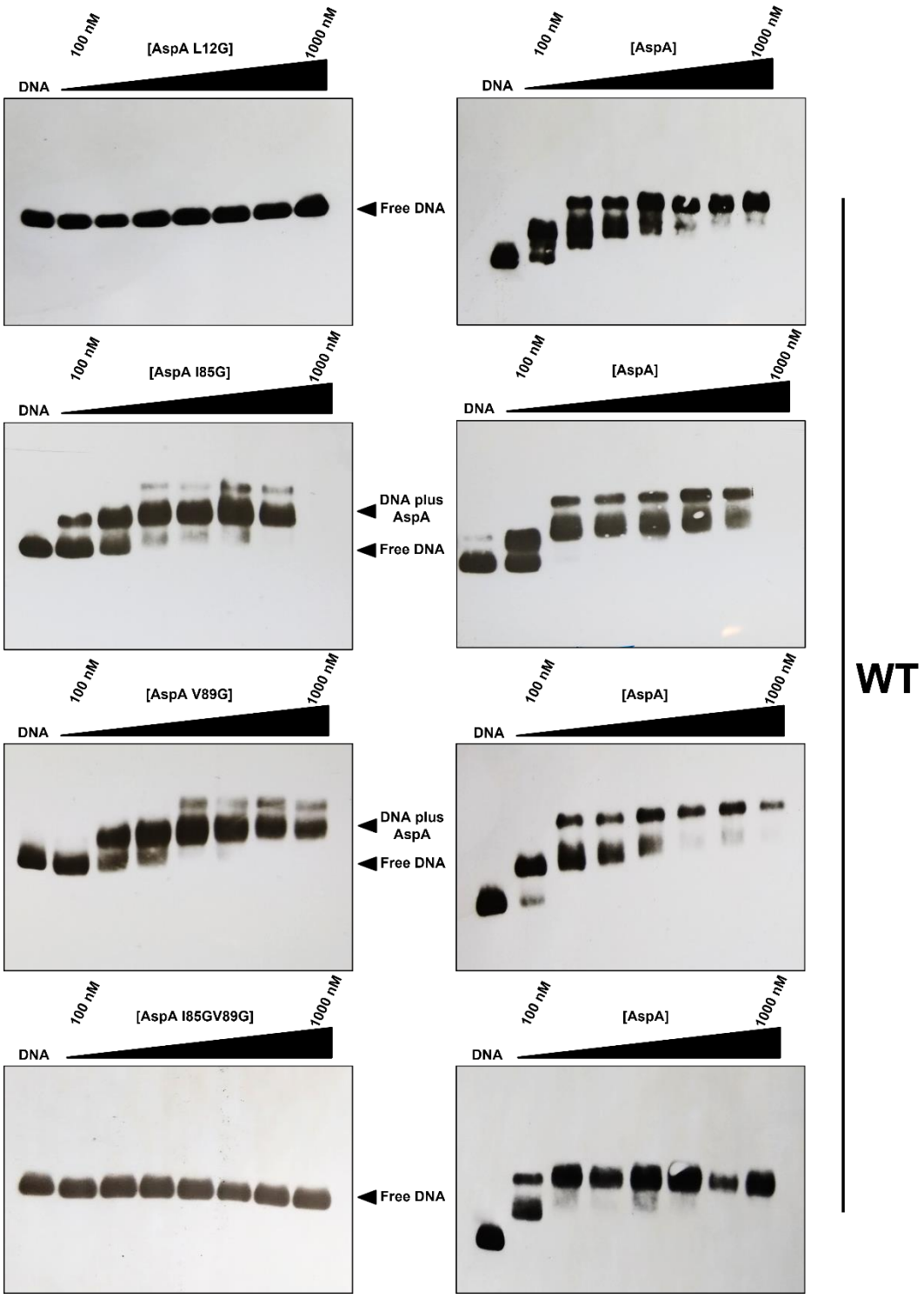


Figure 3.21. EMSA of AspA dimerisation mutants. Proteins were incubated with a 247 bp biotinylated DNA fragment containing the second binding site, and analysed by agarose gel electrophoresis. Shown are representative images from three experimental repeats for the AspA-L12G, AspA-I85G, AspA-V89G and AspA-I85GV89G mutants (left). A control EMSA using the WT AspA protein was run concurrently alongside each mutant (right). The final concentration of DNA in each reaction was 0.12 nM, and the final protein concentrations were 0, 100, 250, 350, 500, 650, 800 and 1000 nM. Free DNA represents unbound molecules, and AspA-DNA complexes are also marked. Reactions were loaded onto a 2% agarose gel.

3.2.7 Atomic Force Microscopy analysis of AspA-DNA interactions

Atomic Force Microscopy (AFM) is a microscopy technique that has sufficient resolution to enable interactions between DNA and protein to be characterised at the single molecule level *in vitro*. It has previously been shown to be effective in understanding the mechanisms of binding of bacterial plasmid-encoded proteins to their cognate binding sites (Pratto *et al*. 2009), and archaeal chromatin organisation via protein-DNA interactions (Laurens *et al*. 2012). The topological state of DNA in the presence and absence of protein has also been investigated, with several studies demonstrating the ability of the protein to bring together distinct sections of the DNA molecule (in both intra- and inter-molecule fashion), bridging the DNA and thus acting to condense the molecule in a way that may be beneficial for its segregation (Laurens *et al*. 2012, Murugesapillai *et al*. 2014, Andres *et al.* 2019).

Here, the 41 kb size of the plasmid pNOB8 is a limiting factor when using AFM, with plasmids used for AFM analysis typically being no larger than 9 kb. Therefore, artificial plasmid constructs were designed by cloning regions of pNOB8, and inserting them by restriction digest into the vector pUC18 (2.7 kb). Two regions of pNOB8 were chosen for amplification: a shorter 200 bp fragment containing the second AspA binding site centrally, and a larger ~1.7 kb fragment that contains both palindromes. This produced two new constructs, dubbed pJA1-200 and pJA2-1.7, which are ~2.9 kb and 4.4 kb in size respectively. A schematic of the plasmids and overview of the cloning protocol is shown in **Figure 3.22A and 3.22B**. Both constructs were sequenced to ensure they were correct. Fragments were then amplified by PCR from the new constructs, and purified PCR products used for AFM experiments. The reason that these fragments were cloned into another plasmid and not just amplified from pNOB8 was that it was planned that the linearised plasmids would also be used, along with plasmids in their circular form to observe the effect of DNA supercoiling.

AFM studies of this nature often use mica as a substrate for the DNA. As mica, along with DNA, is negatively charged, a source of divalent cations from for example $MgCl_2$ is required to allow adsorption of the DNA onto the substrate (Pang 2015). The DNA also is diluted to a concentration such that enough molecules can be observed on the mica surface; here, 0.5 ng/µl was deemed optimal. Once the DNA, or DNA-protein complexes are deposited on the mica and imaged, it is necessary to distinguish DNA alone from DNA-protein complexes. As the mica should be atomically flat, this can be done by measuring the height of a cross-section of DNA (+/- protein) above the surface in nm. Example AFM images and a description of the methodology used to distinguish between DNA and DNA-protein complexes are shown in **Figure 3.22C**.

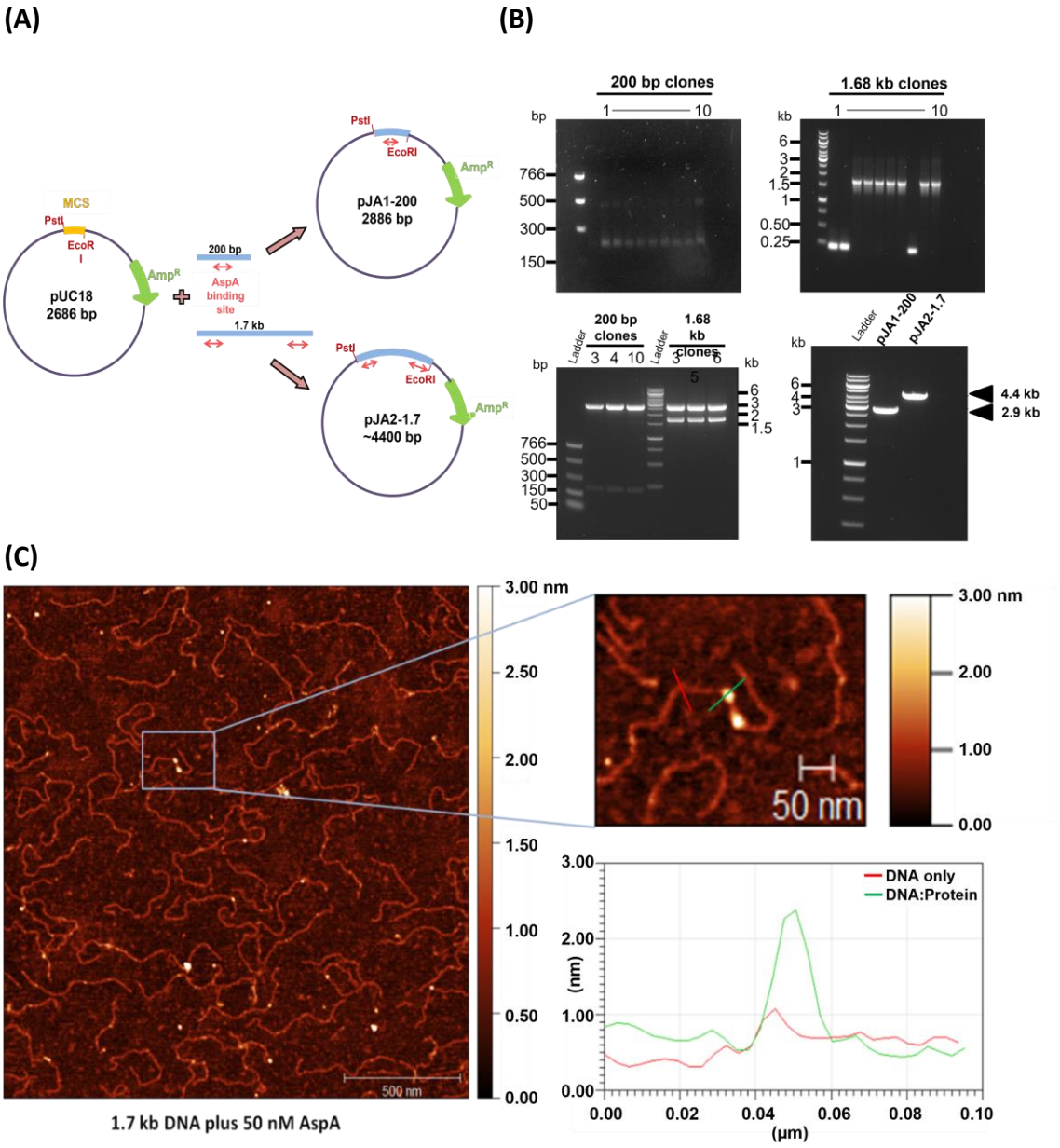


Figure 3.22. AFM fragment cloning and example analysis. (A) Two DNA fragments containing either one or both AspA binding sites were cloned into the cloning vector pUC18 using restriction enzymes PstI and EcoRI. The newly created vectors pJA1-200 and pJA2-1.7 were transformed into *E. coli* DH5-α cells. **(B)** (Top) Ten transformed colonies for each vector were chosen for colony PCR and three clones harbouring the correct size fragment were chosen. (Bottom left) A diagnostic restriction digest was performed to verify the inserts. (Bottom right) The vectors were digested with PstI alone to create linearised plasmid fragments containing the AspA binding site(s) for use in AFM experiments. **(C)** Representative AFM image of DNA incubated with 50 nM AspA and immobilised on the mica surface. The DNA is a 1.7 kb fragment containing both AspA binding sites. The magnified section of the image (right, top) shows a DNA molecule bound by AspA, and the red and green lines are 100 nm cross-sections through DNA and DNA-AspA respectively. The height of the DNA and AspA-DNA cross-sections above the surface (right, bottom) was measured using the 'extract profile' tool in Gywddion software. The scale bar on the right shows colour change as a function of height above the mica surface in nm. This method is used to determine protein-DNA binding events in the following figures.

Initial AFM experiments were conducted with the 200 bp fragment (containing the second AspA binding site only), either alone or with the addition of AspA at two different concentrations of 50 and 300 nM. Here, it was hypothesised that at 50 nM, binding events may be seen, as EMSA assays demonstrated protein binding at this concentration. Protein complexes can be seen at 50 nM, along with the appearance of several different DNA molecules seemingly recruited into the complex **(Figure 3.23)**. This effect is amplified at greater AspA concentrations of 300 nM, where a larger cluster of AspA is apparent, and a concomitant increase in the number of DNA molecules bound to the complex is seen. It should be noted that these effects were not seen over a large number of molecules, and therefore no quantitative analysis has been conducted for these conditions.

Next, the 1.7 kb fragment was used for similar experiments. Here, the fragment, which contains both AspA binding sites, was incubated with 50, 150 and 300 nM protein, and the number of complexes per molecule measured **(Figure 3.24)**. The hypothesis here was that more complexes would be seen at higher protein concentrations, not only due to both sites potentially being occupied, but also due to the ability of the protein to spread non-specifically on the DNA. 80 DNA molecules were analysed for each condition. Surprisingly, the number of complexes per molecule did not increase with protein concentration, with the mean number of complexes of 0.34 at 50 nM being slightly greater compared to 0.14 and 0.16 at 150 nM and 300 nM respectively, the opposite of that which was hypothesised. The number of complexes was significantly different between the 50 nM and 150 nM conditions (two-tailed, unpaired t-test: $p < 0.05$). Two or more complexes per molecule, indicating occupation at both sites, was observed only 5 times, which is surprisingly low, especially at 300 nM concentration. These experiments were not completed in full due to reproducibility problems with the assay, and so it is possible that these results are not representative of the normal *in vitro* AspA-DNA interaction, and therefore require repetition. It should also be noted that these sets of experiments lack controls due to time constraints. Additional controls such as a random-sequence DNA fragment without the palindrome are required, and the use of an asymmetrical DNA molecule that would allow the location of the palindrome to be measured, would be beneficial in future studies.

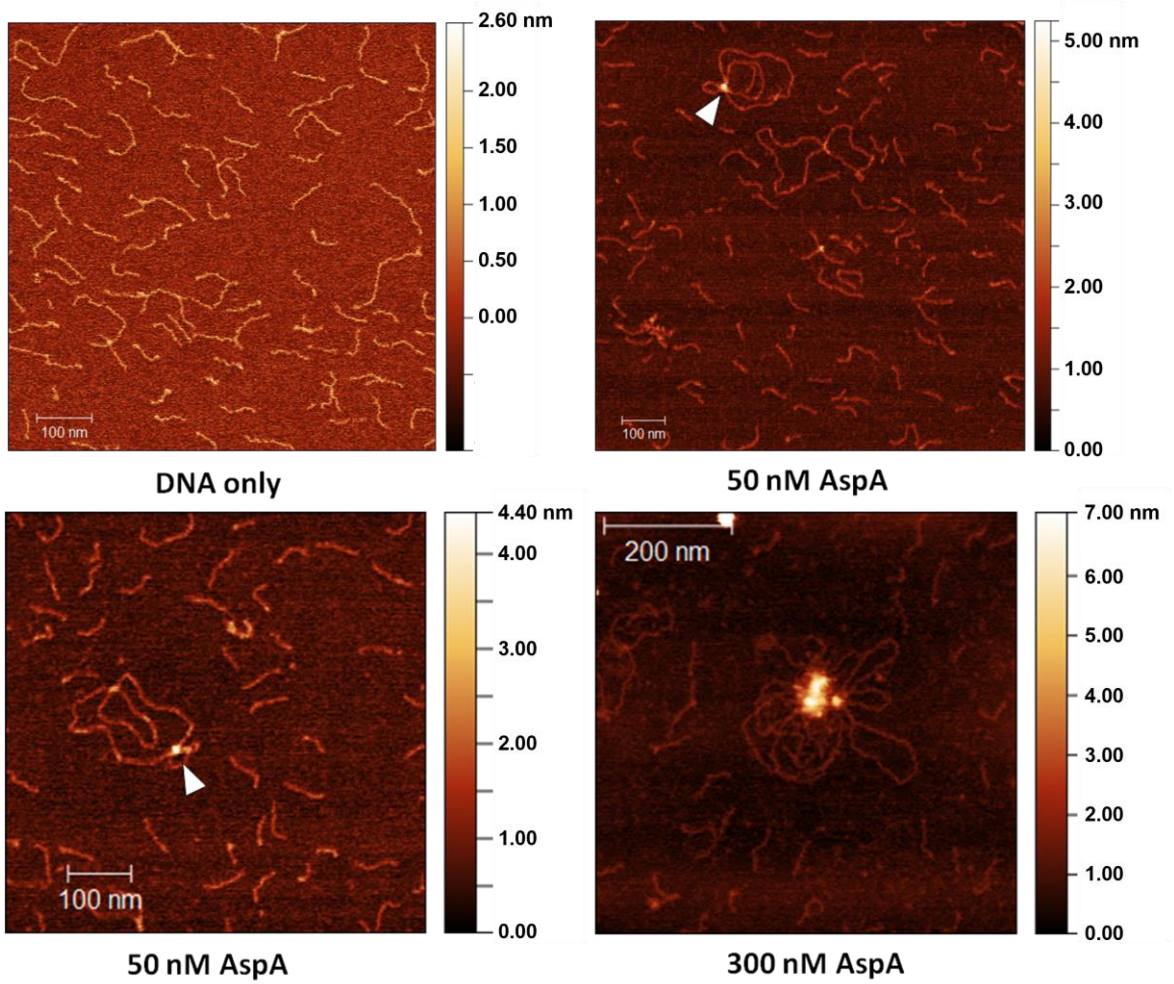


Figure 3.23. Qualitative analysis of increased protein concentration on DNA binding. All panels show the 200 bp DNA fragment containing the second AspA binding site, at a concentration of 0.5 ng/µl (3.85 nM). The top left panel shows DNA only, with the remaining panels showing DNA plus the wild-type AspA protein, at concentrations of 50 nM and 300 nM. The white arrows show a protein-DNA binding event, as measured using the method described in the previous figure.

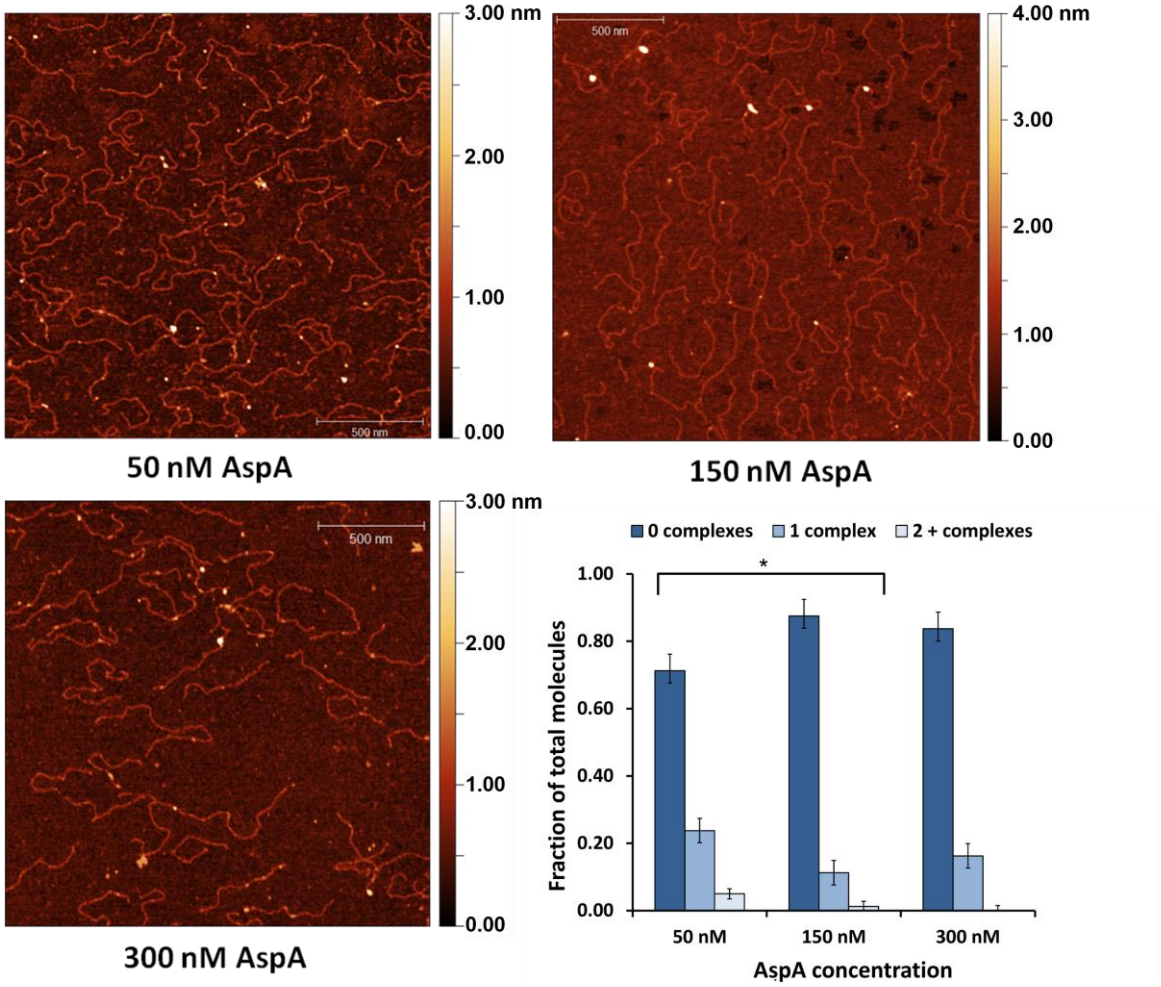


Figure 3.24. Quantitative analysis of number of protein complexes as a function of concentration. All AFM image panels show the 1.7 kb DNA fragment, containing both AspA binding sites, incubated with the stated concentration of AspA. The DNA is at a concentration of 0.5 ng/µl (0.5 nM). Protein complexes were defined as stated previously in **Figure 3.23**. (Bottom right) The fraction of total molecules with 0, 1 or 2+ complexes, for the three AspA concentrations. A two-tailed, unpaired Student's T-Test was used to test significance between each protein concentration (* = $p < 0.05$). n = 80 molecules were measured for each protein concentration. Error bars represent standard error of the mean.

The circular plasmid containing both binding sites (4.4 kb in size) was also used, to assess any effect of increased protein concentration on the topology of the DNA. Here, in addition to the number of protein complexes per molecule, the number of 'bridging events' was also measured. A bridging event was defined as two separate regions of the plasmids (or two separate plasmids) being brought together in the presence of a protein complex. Natural plasmid topologies where the molecule crossed over itself, but this was not accompanied by a protein complex, were not counted as bridging events.

The total number of protein complexes observed per molecule (n = 33 molecules) increased from 26 at 50 nM, to 89 at 300 nM, as might be expected given a six-fold increase in protein. Six molecules had two or more complexes at 50 nM protein concentration, compared with 27 at 300 nM, indicating that at higher concentrations, the protein binds to the plasmid non-specifically. Indeed, the maximum number of complexes observed bound to a single molecule at 50 nM was two, compared with one instance of six complexes at 300 nM. It should be noted that 'complex' does not necessarily equate to one AspA dimer, and could represent multiple dimers, as the size of complexes was not measured. The mean number of complexes per molecule was 0.79 at 50 nM, compared with 3.00 at 300 nM, indicating a ~four-fold increase in binding at six-fold increase in concentration.

The increased number of complexes observed at 300 nM brought a concomitant increase in the number of bridging events. There were 22 total bridging events at 50 nM, compared with 73 at 300 nM. At 50 nM, most molecules displayed either none, or one bridging event, compared to two or three bridging events being commonplace at 300 nM **(Figure 3.25)**. The mean number of bridging events per molecule was significantly greater at 300 nM AspA concentration compared to 50 nM, at 2.21 and 0.67 respectively (two-tailed, unpaired t-test: $p < 0.001$).

As mentioned previously, these experiments were not completed, and also lack controls such as a plasmid backbone that does not incorporate the cloned section harbouring the palindrome.

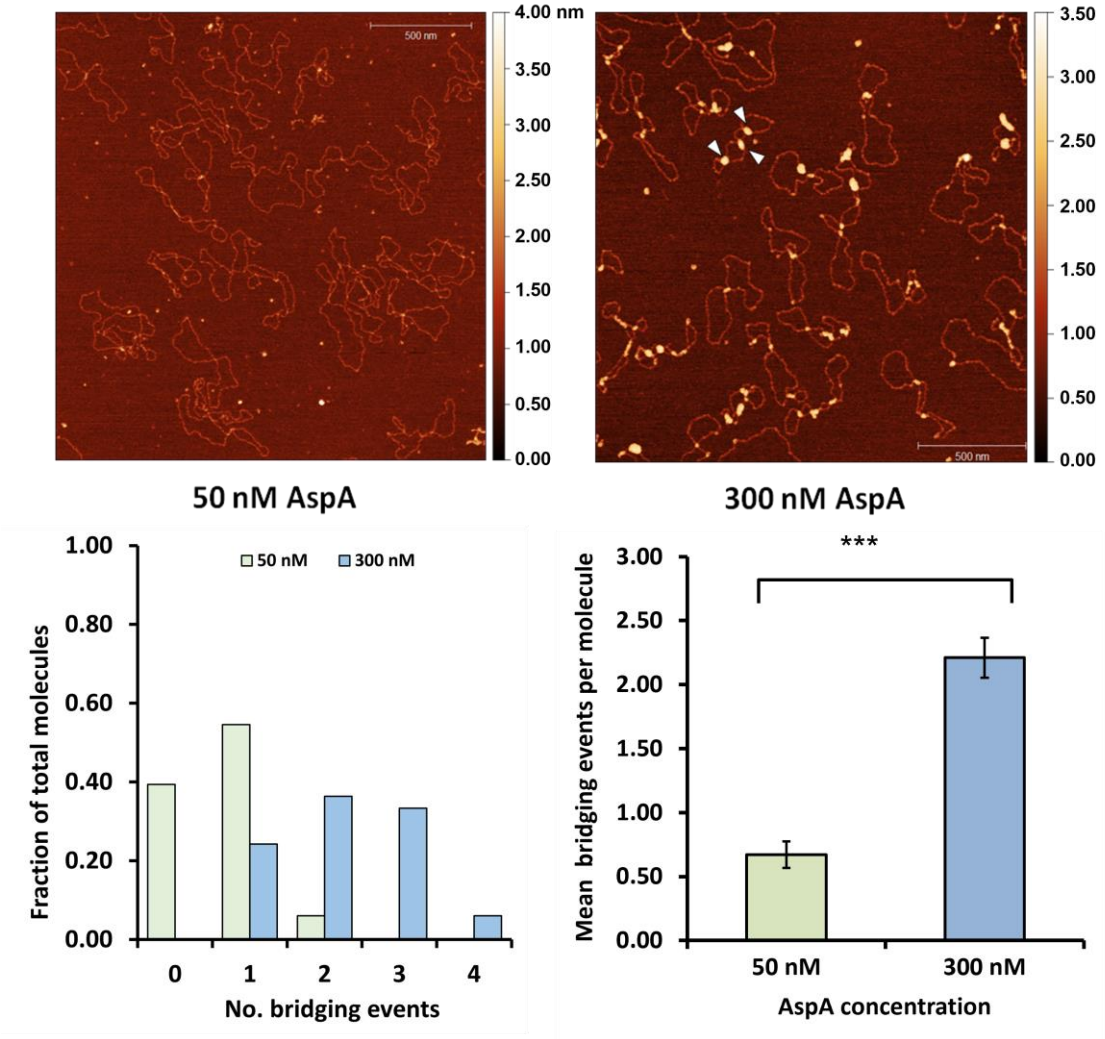


Figure 3.25. Quantitative analysis of number of bridging events as a function of protein concentration. The top two panels show representative images of the circular plasmid, 4.4 kb in length, at a concentration of 0.5 ng/µl (0.17 nM), incubated with the stated concentrations of AspA. Bridging events are defined as a region of DNA coming together in the presence of a protein-DNA complex (e.g. white arrows in the top-right panel). (Bottom-left) The number of bridging events seen per molecule as a fraction of total molecules, with 50 nM or 300 nM AspA (Bottom-right). The mean number of bridging events per molecule at 50 nM and 300 nM protein. A two-tailed, unpaired Student's T-Test was used to test significance between each protein concentration (*** = p <0.001). n = 33 molecules were measured in each condition. Error bars represent the standard error of the mean.

3.3 Conclusions and discussion

DNA segregation is an active mechanistic process that ensures the faithful dissemination of replicated genetic material to future generations across all domains of life (Nasmyth 2002). In prokaryotes, the archetypal partitioning system *parABS* was first identified on *E. coli* P and F plasmids (Austin & Ables 1983, Abeles *et al.* 1985), and comprises the centromeric DNA sequence *parS*, the centromere-binding protein (CBP) ParB, and the motor protein ParA (Schumacher 2008). These partitioning systems, or similar variants thereof, have been found across a diverse array of bacterial plasmids and chromosomes (Hayes & Barillà 2006a, Broedersz *et al*. 2014, Badrinarayanan *et al*. 2015).

The DNA segregation system of the archaeal plasmid pNOB8, harboured by the strain *Sulfolobus* NOB8-H2, differs from the predominant two-gene system found in bacteria, as it has a tricistronic arrangement (Schumacher *et al*. 2015). The partition cassette is comprised of three genes; *aspA*, *parB* and *parA*, and there are two identical palindromic centromere-like sequences on the plasmid. In this segregation system, the role of site-specific DNA-binding protein is fulfilled by AspA, and not ParB. AspA has been demonstrated to bind with high affinity to the palindrome immediately upstream of the *aspA-parB-parA* cassette, and can spread upstream along the DNA at higher concentrations to form an extended DNA-protein complex. Exactly how these three proteins work in conjunction to effectively partition pNOB8 is currently unknown. This chapter has focussed on the interactions of the AspA protein at the second DNA palindrome, ~1.5 kb away on pNOB8, along with the biochemical characterisation of a number of AspA residues hypothesised to be important to function: those involved in DNA binding activity, and both dimer-dimer and monomer-monomer interactions.

AspA is known to bind with high affinity to the palindrome upstream of the partition cassette, and band-shift assays demonstrated an equally strong degree of interaction at the second palindrome *in vitro*. This could imply that neither site is favoured over the other by AspA *in vivo*, although this would require experimental investigation. The importance of several AspA residues for correct function was assessed by mutagenesis and further band shift assays. It was previously demonstrated that a single amino acid

substitution, arginine to alanine at position 49, completed abolished the DNA binding activity of ApsA (Schumacher *et al*. 2015), and further mutations in residues tyrosine 41 and glutamine 42, which contact DNA bases and the backbone, also vastly reduced binding. Further mutagenesis of AspA residues that were hypothesised to be important for dimer-dimer interactions, and thus spreading of the protein on the DNA was conducted. Mutagenesis of residues important for dimerisation of the protein and subsequent DNA binding, highlighted the importance of these single residues (or a combination thereof) for correct functioning of AspA. This specificity is perhaps unsurprising given the enormous timescales natural selection has acted over; indeed it is thought that bacterial chromosomal *par* loci, and their partner CBPs, must have arisen very early on evolutionary time (Livny *et al.* 2007), and it is probable that this was also the case with archaeal chromosomes and plasmids. Jalal and colleagues recently studied the specificity of chromosomal ParB and the closely related DNA-binding protein Noc, whose cognate binding sites differ by only a few base pairs. They demonstrated that this high degree of specificity of the two proteins was dependent on only four amino acids at the protein-DNA interface (Jalal *et al*. 2020a).

The spreading pattern of AspA at the second palindrome was also investigated using DNAse I footprinting assays. It is known that CBPs can spread proximally along the DNA from the initial nucleation site, sometimes for many kilobases (Rodionov *et al*. 1999, Tran *et al*. 2017). The footprinting data demonstrate that a different pattern of protection is formed by AspA at the second site, upstream of *orf41*, compared with the pattern observed at the first palindrome adjacent to the start of the *aspA* gene. At the first site, the region of protection extends ~300 bp upstream of the *aspA-parB-parA* cassette at higher concentrations, perhaps indicating that the protein is performing dual roles; both structural (in aiding partition complex formation), and regulatory, by modulating transcription of the partition genes. At the second site, the second region of protection is much smaller, leading to the hypothesis that the protein only performs the role of transcriptional regulation at this site. However it is unknown if these patterns would be replicated *in vivo*, and it is possible that the AspA-DNA complex could theoretically extend from one palindrome to another if enough endogenous protein was present.

The function of the protein encoded by *orf41* is currently unknown; recent database searches using pNOB8 ORFs failed to give any new insight into this particular gene. Future experiments could look at transcriptional regulation by the WT and mutant AspA proteins using a plasmid-based, *in vitro* cell lysate assay, as has recently been described in *S. solfataricus* (Lo Gullo 2019).

Although incomplete, preliminary AFM experiments provided useful data for future studies, particularly when assessing the ability of AspA to bring together distal sections of DNA, via a combination of protein-DNA and protein-protein interactions. These experiments appeared to show AspA 'bridging' two DNA strands together. This behaviour has previously been characterised for the DNA-repair protein Ctp1, where both intra- and inter-molecular bridging events are seen (Andres *et al*. 2019), and paired newly-replicated plasmids bound to the partition protein Omega (ParB) in *Streptococcus pyogenes* has been observed under AFM (Pratto *et al*. 2009). Here, it is unknown whether AspA is forming bridges between DNA strands or just binding to natural plasmid topological links. Bridging events have previously been described as two or more strands connected by or more proteins, or protein interacting with two molecules of dsDNA (Murugesapillai *et al*. 2014, Andres *et al*. 2019). A volumetric measurement of AspA-DNA complexes would help answer this question, as it may demonstrate that e.g. two or more AspA dimers are bound together between DNA strands. Non-specific AspA binding at higher concentrations could also act to further condense the plasmid, in the case of more than two bridging events, which were observed at higher protein concentrations.

One limitation with the experiments in this chapter relates to the purification of the AspA WT and mutant proteins. Use of a cleavable hexa-histidine tag rather than the non-cleavable C-terminal tag used would have allowed the use of the native protein without additional amino acids comprising the tag. Additionally, further purification steps such as size-exclusion chromatography following the affinity purification would be useful to remove higher-order oligomers observed on some AspA mutant SDS gels. Finally, some assays suffered from a lack in accuracy of protein concentration measurement when using the Bradford assay, therefore in future, more accurate concentrations could be obtained using UV spectrophotometric measurements.

Chapter 4

AspA-ParB interactions

Chapter 4

Investigating AspA-ParB interactions

4.1. Introduction

In a typical bacterial DNA segregation system, which comprises two genes, *parA* and *parB* and a DNA centromeric site, the protein encoded by *parB* acts as the centromeric binding protein (CBP). This arrangement is observed both on plasmid-encoded and chromosomal partition cassettes (Baxter & Funnell 2014, Funnell 2016, Bouet & Funnell 2019, Jindal & Emberley 2019). A bicistronic operon is also found on the chromosome of the crenarchaeaon *S. solfataricus*, where the protein SegB, although not homologous to bacterial ParBs, performs an analogous function as a site-specific DNA-binding protein (Kalliomaa-Sanford *et al*. 2012). The segregation cassette of plasmid pNOB8, harboured by the *Sulfolobus* strain NOB8-H2, is atypical in its genetic organisation, comprising three genes *aspA*-*parB*-*parA* **(See Figure 1.17, Introduction)**, although this arrangement is not confined to pNOB8, as it is also harboured on both chromosomes and plasmids of other chrenarchaeal species (Schumacher *et al*. 2015). It has been established that in this system, AspA functions as the CBP, and interactions of AspA at its second cognate binding site were discussed in the previous chapter. ParB is therefore atypical in that it does not perform the role of CBP as this is effected by AspA, and instead was hypothesised to possess a different functionality. A range of biochemical assays were previously used to describe the behaviour of ParB; EMSA demonstrated that ParB binds DNA non-specifically, whilst surface plasmon resonance (SPR) established that ParB interacts with both AspA and the ATPase ParA, meaning that ParB may function as an adaptor protein in this system, potentially forming a bridge between the AspA-DNA palindrome and non-specific DNA elsewhere on pNOB8 or the nucleoid (Schumacher *et al*. 2015). ParB is the largest protein in the pNOB8 segregation system at 470 aa, compared to both AspA and ParA (93 and 315 aa respectively), and is composed of two domains; an N-terminal domain and a C-terminal domain, separated by a flexible linker region. The N-terminal

domain is the larger of the two, at 320 aa, and it was the N-terminus that was demonstrated via isothermal titration calorimetry (ITC) to interact with AspA in a 1:1 stoichiometric ratio; whilst fluorescence polarisation (FP) assays indicated that ParB-N alpha helices 11-13 were required for correct binding to AspA (Schumacher *et al*. 2015). The C-terminal domain of ParB (aa 370 – 470) was demonstrated using FP to be responsible for the non-specific DNA-binding activity of the protein, which could help compact pNOB8 by binding elsewhere on the plasmid, or allow the plasmid to be anchored to the nucleoid DNA (Schumacher *et al*. 2015). Microscale thermophoresis (MST) experiments suggested that the inter-domain flexible linker of ParB (aa 321 – 369) was the region of ParA binding (Schumacher *et al*. 2015). A model for inter-protein interactions, and with both plasmid and chromosomal DNA, and how these functionalities may mediate effective segregation of the pNOB8 plasmid, is shown below in **Figure 4.1**.

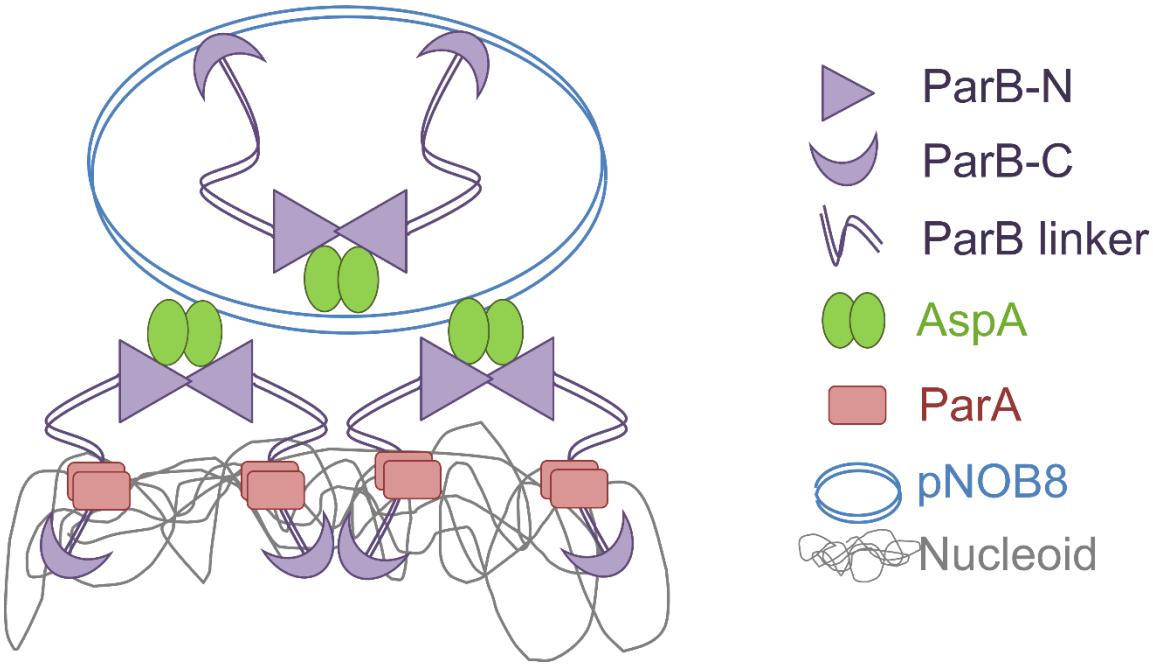


Figure 4.1. Model of pNOB8 plasmid segregation. Cartoon representing the interior of the *Sulfolobus* NOB8-H2 cell. One of two replicated pNOB8 plasmids is shown (blue lines). AspA dimers bind to pNOB8 and spread along the plasmid DNA (here not shown to scale). The N-terminal domain of ParB (purple) binds to AspA in a 1:1 stoichiometric ratio. The C-terminus of ParB binds non-specific DNA, and here is shown binding to both pNOB8 and the NOB8-H2 nucleoid (chromosomal DNA, grey). The flexible linker is bound to ParA dimers (red), which also binds to nucleoid DNA, therefore potentially tethering the plasmid to the chromosome via its association with ParA. The mechanism by which the replicated plasmids are transported to the cell poles is unknown. Adapted from Schumacher *et al*. 2015.

The crystal structures of both domains of ParB have previously been solved, in order to further characterise their functions. The C-terminal structure was determined using a pNOB8 ParB-C construct, however the N-terminus of pNOB8 ParB was less amenable to crystallographic diffraction, and so structural studies were performed using a homologous protein. Here, the N-terminal domain of chromosomal ParB from *S. solfataricus* 98:2, which has 40% amino acid identity to pNOB8 ParB and comprises amino acids 1 – 350, was used **(Figure 4.2, left)**. Small-angle X-ray scattering (SAXS) experiments using the 98:2 ParB-N structure were used to describe a model in which ParB-N alpha helices 11-13 are implicated in both ParB dimerisation, and in binding interactions with the AspA C-terminal helix **(Figure 4.2, right)**. The SAXS model was consistent with the previous ITC-derived stoichiometric ratio of ParB-N:AspA binding of 1:1, and the importance of ParB-N alpha helices 11-13 was confirmed when a ParB-N truncation mutant lacking these helices displayed no binding to AspA (Schumacher *et al*. 2015). Thus, the DNA:AspA:ParB-N multi-protein complex **(Figure 4.1**., **Figure 4.2, right)** is supported by a range of biochemical and structural data.

Originally, an additional aim of the project was to investigate the interaction of ParB not only with AspA, but also with ParA, in order to describe further the region of ParB which mediates this interaction. This would involve the creation of several constructs based on the pNOB8 ParB domains, and to assess their interactions with the ATPase ParA, to confirm the hypothesis that it is the flexible linker region of ParB that binds to ParA **(Figure 4.1)**. A range of constructs were to be created alongside full-length ParB, (e.g. ParB-N, ParB-C, ParB-N plus linker) overproduced, purified, and employed with ParA in protein-protein interaction assays such as MST. However, due to time constraints, these experimental plans were altered, and as such the main focus would be on the interaction of pNOB8 ParB-N with AspA, with the aim of characterising the residues involved at the interface between the two proteins. To do this, the ParB-N construct previously planned would still be required, therefore the initial aim was to confirm the boundaries of the N and C-terminal domains, and the flexible linker of pNOB8 ParB, using the structure determined for the ParB homologue.

4.1.1 Aims

In this chapter, the interaction between pNOB8 ParB and AspA will be investigated, with the overall aim of identifying residues involved at the binding interface between the two proteins.

The main aims of this chapter are:

1) To define the domain boundaries of pNOB8 ParB, based on bioinformatic and structural analysis, and compare with the previously designated domains.
2) The cloning of the DNA region that encodes the N-terminus of ParB into a suitable expression plasmid, the overproduction of the ParB-N protein, and its purification.
3) Assessing interactions between ParB-N and AspA using chemical cross-linking, followed by liquid chromatography coupled to mass spectrometry to identify residues at the interaction interface.

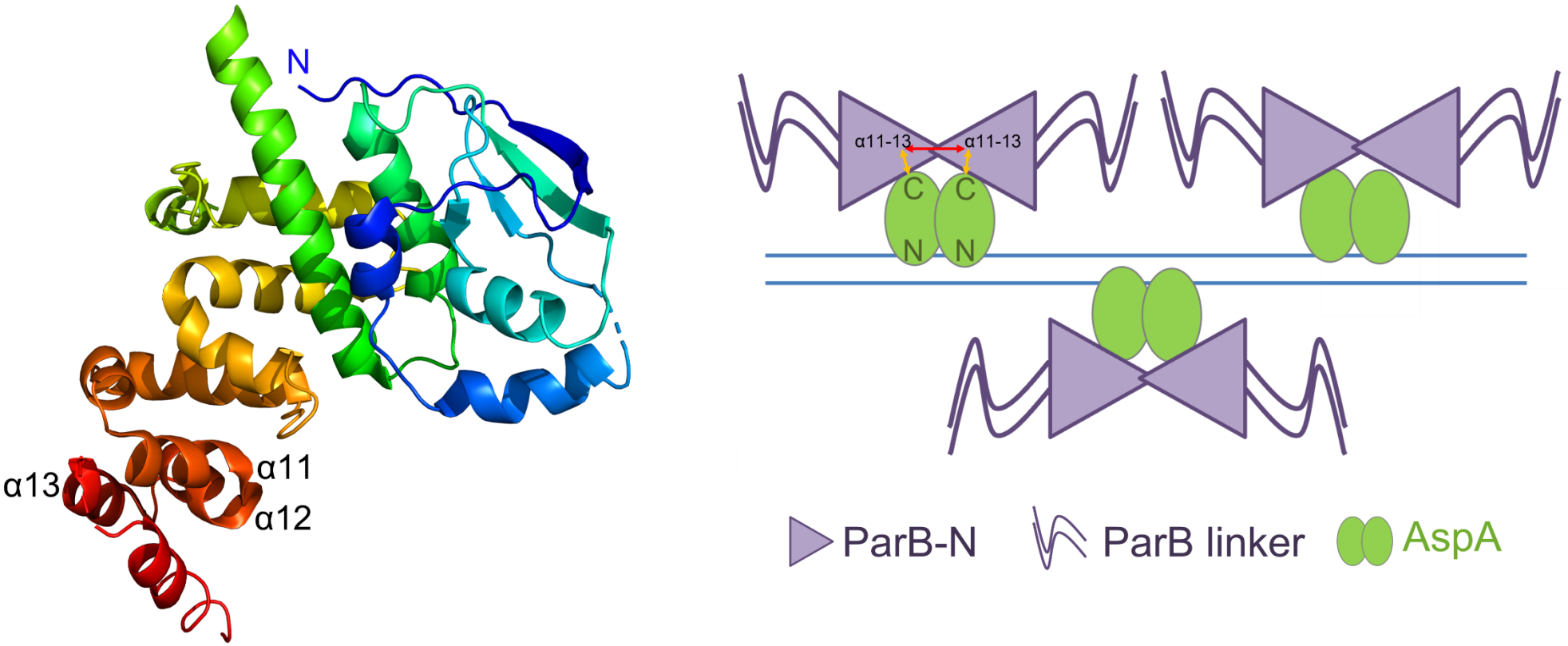


Figure 4.2. ParB-N crystal structure and model of DNA:AspA:ParB-N complex. (Left) Structure of the *S. solfataricus* 98:2 chromosomal ParB homologue N-terminal domain. The N-terminus, and alpha helices 11-13 are labelled. The figure was generated using PyMOL, with the PDB file 5K5A. **(Right)** Model of ParB-N binding to AspA, based on biochemical and modelling data. The ParB-N alpha helices 11-13 are involved in both dimerisation (red arrow), and interactions with the C-terminal helices of AspA (orange arrows), to form the DNA:AspA:ParB structure. The ParB C-terminal domain is not shown. Blue horizontal lines represent pNOB8 DNA. Figure is not to scale. Adapted from Schumacher *et al*. 2015.

4.2 Results

4.2.1 Using AlphaFold to predict the pNOB8 ParB structure

During the writing of this chapter, the latest iteration of the machine-learning based protein structure prediction algorithm AlphaFold was made available. This approach to predicting protein structures with atomic accuracy compared to experimental data was demonstrated to be considerably more accurate than competing algorithms, representing a revolution in computational approaches to structure determination (Jumper *et al*. 2021). The crystal structure of the C-terminus of pNOB8 ParB is already know, but since the structure of the N-terminal domain was derived using a homologous protein, it was decided to use AlphaFold to model the pNOB8 ParB structure. The model could also be used to compare to the domain boundaries assigned later in this chapter (Section 4.2.2). The AlphaFold v2.0 source code is available, but an alternative web-based approach is provided using AlphaFold Colab, (Methods 2.8) which uses a simplified version of the full software, but nevertheless produces highly accurate structures.

The protein sequence of pNOB8 ParB was uploaded to the AlphaFold Colab server. The predicted structure of ParB is shown in **Figure 4.3**. The AlphaFold structure output is coloured according to a per-residue confidence score on a scale of 1-100, pLDDT (predicted Local Distance Difference Test). A pLDDT score of >90 is classed as a highly accurate prediction, with progressively lower scores giving decreased confidence in the model. A region with a pLDDT score of <50 is a strong predicter of disorderly, unstructured regions (Jumper *et al*. 2021). The ParB structure is predominantly scored between 70 and 90, or greater than 90 (light blue, dark blue respectively), indicating confident or very high pLDDT confidence scores **(Figure 4.3, left)**. Areas coloured orange, indicating unstructured regions, include the inter-domain flexible linker, along with the alpha-helices 5 and 6 of ParB-N which are either side of another flexible region which separates the two ParB-N sub-domains. Using the pLDDT scores/colour scheme, it appears that the flexible linker appears to be longer than previously thought (AlpaFold prediction; aa 303-374, previous prediction; aa 321-369, Schumacher *et al*. 2015). The unstructured regions are also clearly outlined on the pLDDT plot **(Figure 4.3, right, top)**.

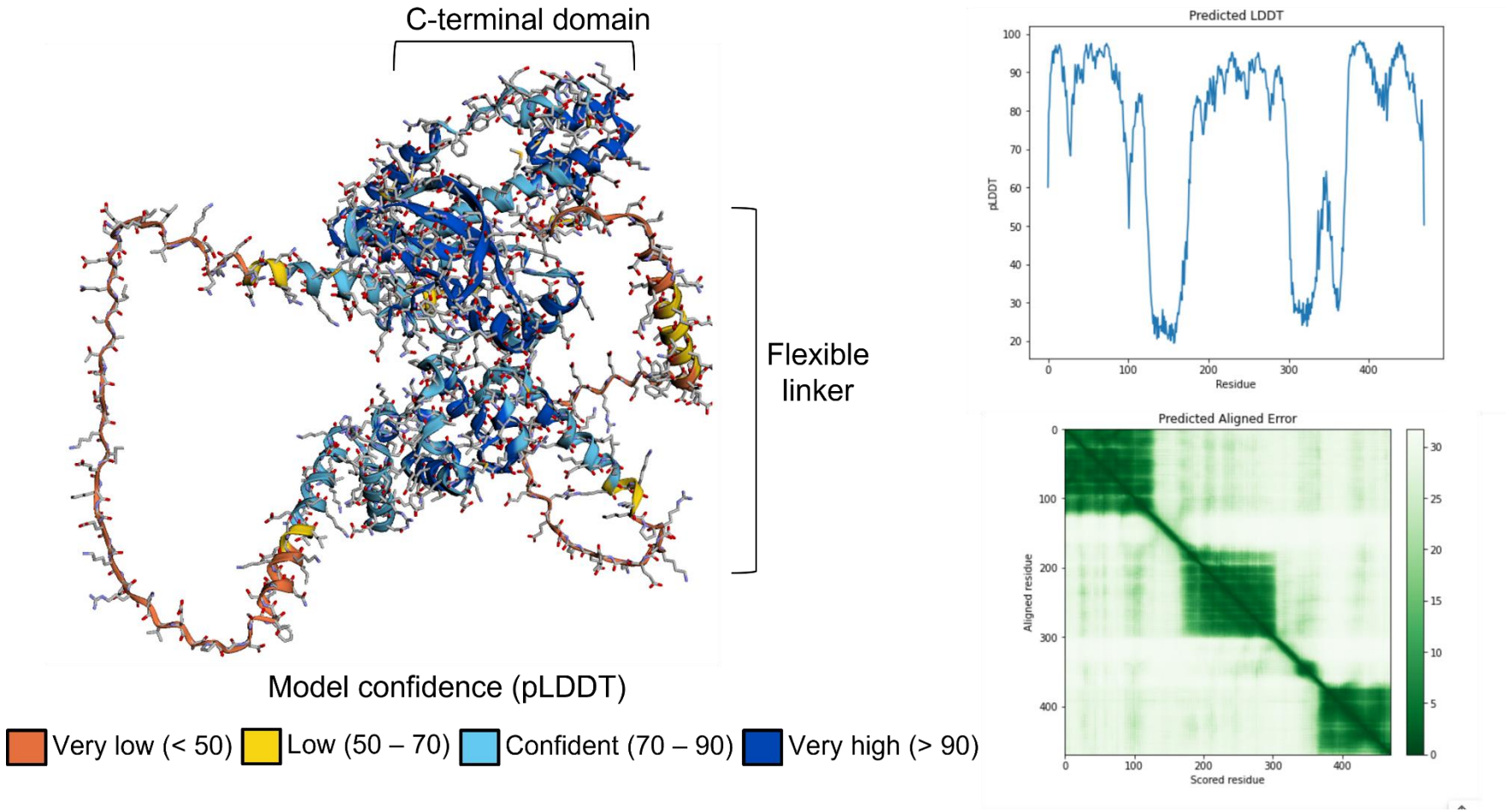


Figure 4.3. Predicted structure of pNOB8 ParB using AlphaFold. (Left) The structure was derived using the online AlphaFold Colab sever, using the pNOB8 ParB amino acid sequence. The colour scheme indicates the model confidence based on the per-residue pLDDT score. **(Right)** (Top) A 2D plot of the pLDDT score (y-axis) by residue number (x-axis), showing clearly the regions of low model confidence equating to unstructured regions. (Bottom) The Predicted Aligned Error (PDE) plot, which gives the predicted position error in Ångstroms of the scored residue (x-axis), when aligned on a residue on the y-axis. The dark green boxes indicate the two N-terminal sub-domains plus the C-terminal domain..

Another useful output is the Predicted Aligned Error, which gives the expected distance error in Ångstroms of one residue (x), when the predicted structure is aligned on a different residue (y). The three stably folded domains (C-terminal domain and two N-terminal sub-domains) are clearly identifiable as dark squares on the plot **(Figure 4.3, right, bottom)**. The dark colour indicates a low expected distance error, meaning that there is high confidence of *intra*-domain positioning (i.e. positioning of residues of one sub-domain with respect to each other), however the lighter portions indicate less confidence in the relative, *inter*-domain predicted positions. The AlphaFold model also shows the presence of a large (35 aa) unstructured region in the N-terminus between the two sub-domains **(Figure 4.3, left)**. The crystal structure of the ParB-N 98:2 homologue also comprises two sub-domains, separated by a central helix (Schumacher *et al*. 2015), and it appears that the intrinsically disordered region absent from the structure is smaller, at 17 residues. Thus, pNOB8 ParB appears to contain larger regions of intrinsic disorder than previously thought, and compared to the 98:2 homologue structure.

4.2.2 Mapping the domains of pNOB ParB-N

In order to assess any interactions between the ParB and AspA proteins, which was previously demonstrated to involve the N-terminal domain of ParB, a new construct of this domain was required. The pNOB8 ParB domain structure has previously been outlined: N-terminus; residues 1-320, C-terminus; residues 370-470, flexible linker; residues 321-369 (Schumacher *et al*. 2015). However, given that the ParB-N construct was to be cloned anew, and that the 98:2 ParB-N homologue utilised for crystallography studies (residues 1-350) was used to generate the SAXS AspA:ParB-N model, it was decided to employ a variety of secondary structure prediction software to define the pNOB8 ParB domains, and compare these with those previously outlined and with the AlphaFold model shown above.

Initially, CLUSTAL OMEGA was used to generate an amino acid sequence alignment between the full-length pNOB8 ParB (470 aa), and the 98:2 ParB (first 350 aa), as the PDB

file of the 98:2 ParB-N crystal structure (PDB reference 5K5A) incorporates residues 1 -298 only, therefore this may indicate the flexible linker region starts at this point. The alignment showed a good level of conservation up to aa 352 of pNOB8 ParB (not shown), although this may include some of the flexible linker.

Next, a range of online secondary structure prediction tools were used to aid delineation of the pNOB8 ParB domain boundaries. The PSIPRED 4.0 server, which predicts protein secondary structure with 84% accuracy (Buchan & Jones 2019), shows disordered/flexible regions from amino acids 293-336 and 344-361 **(Figure 4.4, top)**. The secondary structure prediction tool Quick2D, available at the Max Plank Institute Bioinformatics Toolkit server (Gabler *et al*. 2020) was also used; here, the consensus disordered region from four prediction tools runs from ~300-370 aa. The output from one such disorder prediction program, IUPred3 (Erdös & Dosztányi 2020), in which disordered regions are predicted to comprise amino acids that cannot interact with each other, is shown in **Figure 4.4, middle**, where disorder >0.5. There are also a number of amino acids in this region (~300-370) that are frequently found in flexible/disordered regions of proteins, e.g. arginine, proline, lysine, glutamic acid, serine, aspartic acid and glutamine (Hansen *et* al. 2006, Dyson 2016). 73% of the residues in this region are one of these seven amino acids. Based on these analyses, it appears that the flexible linker region of pNOB8 ParB could be longer than previously thought, comprising residues 300 – 374, with the ParB N-terminal domain being concomitantly smaller (aa 1-299). These predicted domain boundaries align well with the AlphaFold model, which shows the N-terminal domain comprising residues 1-302. A schematic of these domain boundaries compared to those previously defined is shown in **Figure 4.4, bottom**.

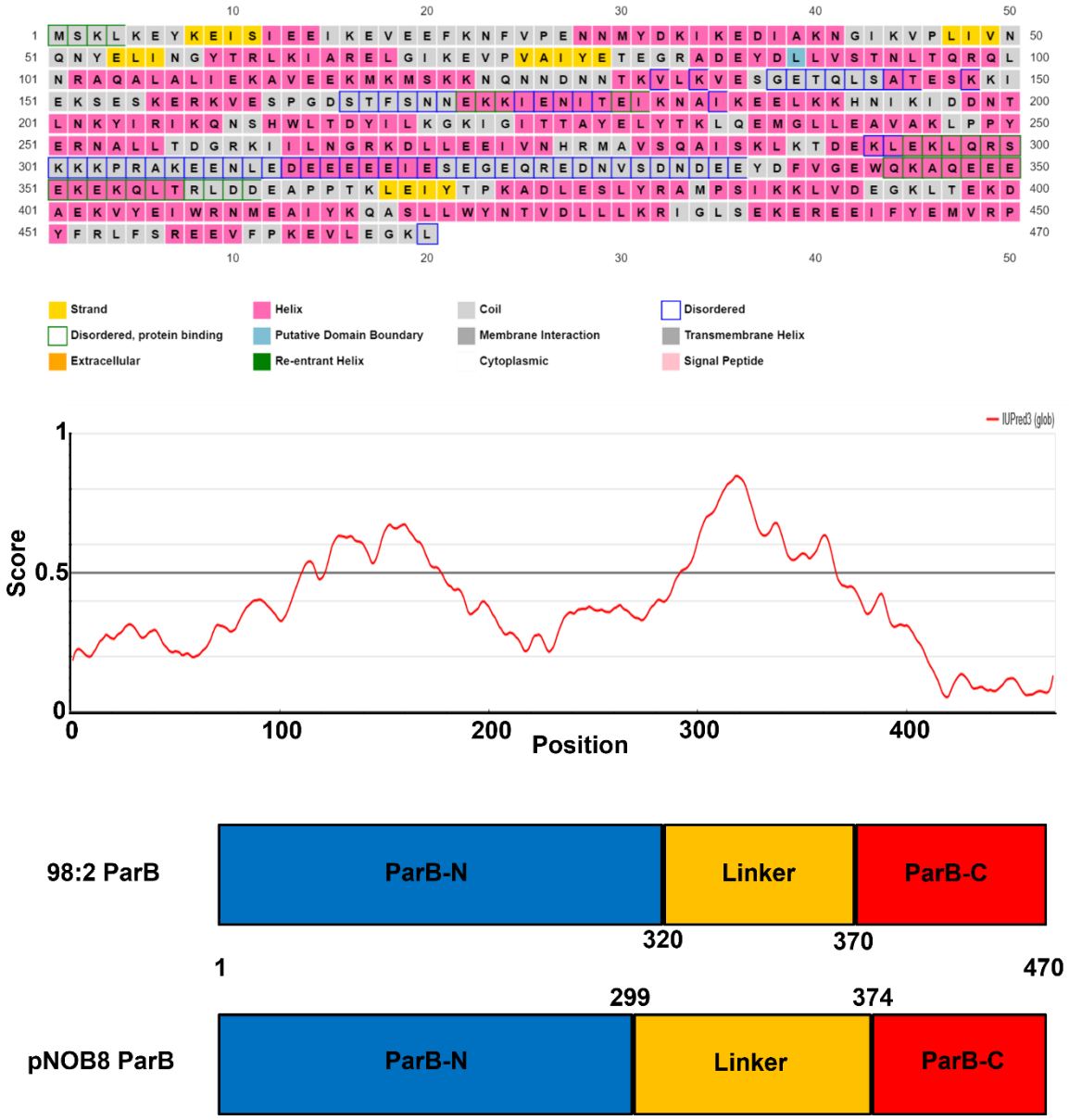


Figure 4.4. Determination of pNOB8 ParB domain boundaries. A variety of bioinformatics tools were used to predict the domains and regions of disorder of pNOB8 ParB. **(Top)** The PSIPRED secondary structure prediction, showing predicted helices, strand, disordered regions etc. **(Middle)** The IUPRED3 web interface showing predicted regions of disorder, with these regions defined as having a score >0.5. The score is shown on the y-axis, and the amino acid number (total 470 aa) on the x-axis. **(Bottom)** Domain map of pNOB8 ParB compared to the previously defined domain boundaries for 98:2 ParB. Numbers indicate amino acid position. Schematic not drawn to scale.

4.2.3 Modelling pNOB8 ParB-N using Phyre2

Given that the N-terminal domain of pNOB8 ParB-N was now defined as comprising amino acids 1-299, and therefore slightly smaller than previously thought (aa 1 – 320), it was decided to model its structure and compare it to that of the 98:2 ParB homologue. The Phyre2 protein modelling server was used to build the model based on closely related homologues with known structures (Kelley *et al*. 2015). This methodology of structure prediction is very different to that performed by AlphaFold: here, homologous structures are used to predict structures, firstly by alignment against known models to create a backbone, before fitting side chains in conformations that avoid steric clashes. AlphaFold uses a neural network to refine the predicted structure in an iterative fashions by sending outputs back through the network until the structure cannot be improved (see Methods 2.8).

Unsurprisingly, the closest matching homologue was that of 98:2 ParB, and the resultant pNOB8 ParB structure was modelled with a high degree of certainty (100% confidence across 84% of amino acids). The Phyre2 predicted structure of pNOB8 ParB-N is shown below **(Figure 4.5, left)**. The model and the structure were superposed using the secondary structure matching (SSM) algorithm in CCP4MG, which iteratively superposes backbone Cα atoms of equivalent secondary structure elements (Krissinel & Henrick 2004). This gave a high degree of overall structural similarity, as measured by a root mean square deviation (RMSD) of 0.56 Å, where the RMSD is the average distance given between Cα atoms, with a lower value indicating greater similarity between structures. However, the two structures differed in their relative positions of alpha helices 11-13, those suggested to be involved in dimerisation and binding to AspA **(Figure 4.5, right)**. This could mean that the interface between the 98:2 ParB-N homologue and AspA in the SAXS model is slightly different to that of the endogenous interaction between pNOB8 ParB and AspA, and that a distinct set of residues are involved in binding.

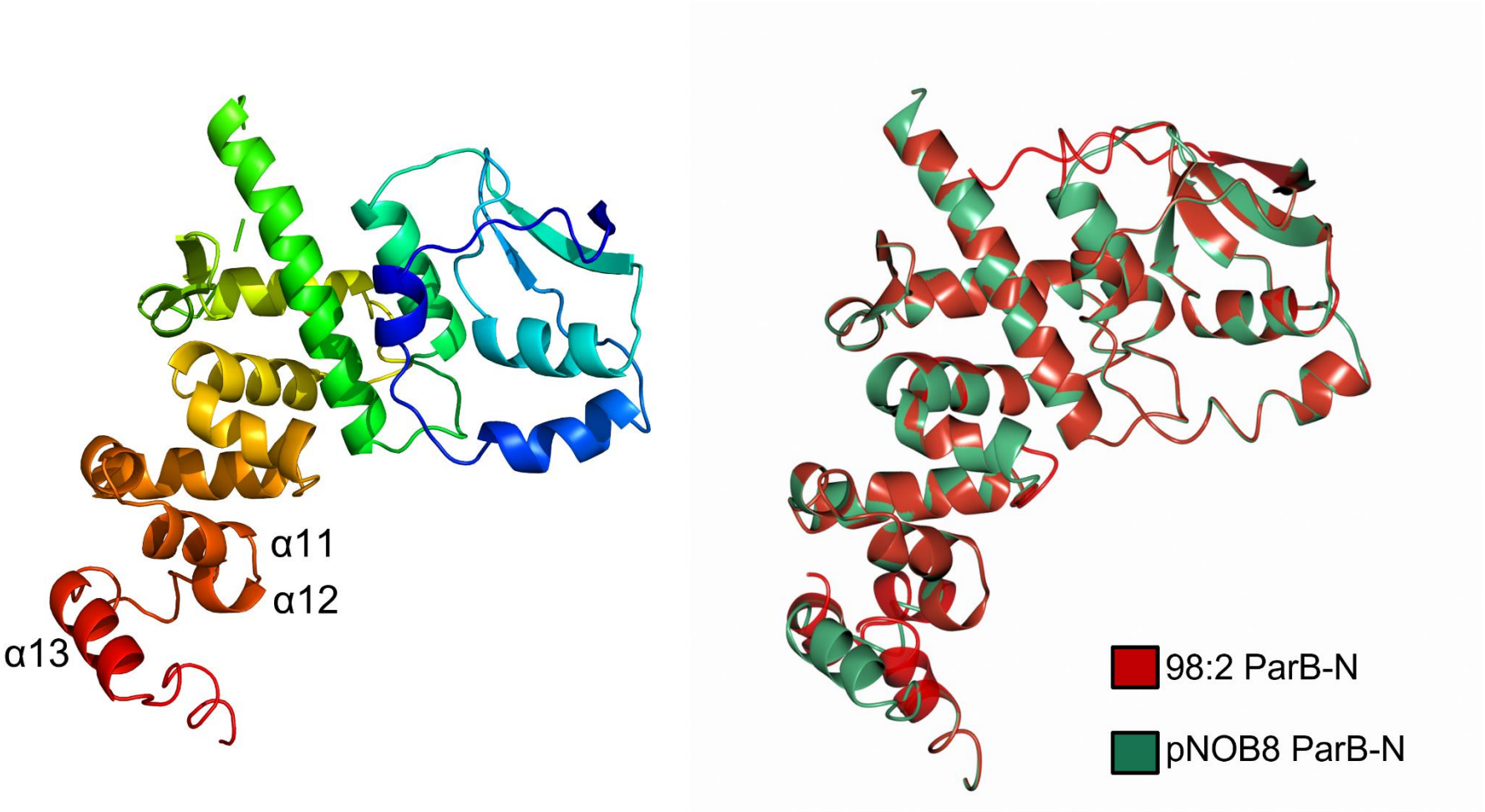


Figure 4.5. Predicted model of pNOB8 ParB-N and superposition with 98:2 ParB-N. (Left) The predicted model of ParB-N, based on the domain boundaries previously assigned, was generated using the PHYRE2 server, and visualised in PyMOL. Alpha helices 11-13 are indicated. **(Right)** Superposition of the N-terminal domain of the 98:2 Par-B homologue (red), and the predicted PHYRE2 ParB-N (green). The superposition was generated using CCP4MG.

4.2.4 Modelling the ParB-N: AspA interaction using ClusPro

Prior to beginning chemical cross-linking experiments with the AspA and ParB proteins, the molecular docking program ClusPro 2.0 was used to model the interface between the proteins *in silico*, potentially to compare against future experimental data, and also as a rationale in targeting specific amino acids for subsequent mutagenesis. The ClusPro server models protein-protein docking using two PDB input files, and constructs likely models based on algorithmic refinement of lowest-energy docked structures (Kozokov *et al*. 2017). Various parameters can be altered to increase the likelihood of the output being close to the native structure of the complex; such as the removal of unstructured regions, and labelling specific residues as attractive or repulsive if they are known to be involved in binding based on prior experimental data. ClusPro was chosen to model the interaction due to its superior predictive performance compared to other molecular docking software (Kozokov *et al*. 2017). Here, both the SAXS and structural models for ParB dimerisation and binding to AspA (using the 98:2 ParB-N homologue structure) indicate that the alpha helices 11-13 of ParB-N appear to mediate dimer-dimer interactions. ParB-N α12 appears proximal to the C-terminal alpha-helix 4 of AspA and therefore may aid formation of the ParB:AspA interface **(Figure 4.2)** (Schumacher *et al.* 2015).

Initially, the 98:2 ParB-N homologue (PDB file 5K5A) was used to dock with AspA, in an iterative manner, to see if the output model was close to the previously generated SAXS data. ParB-N was defined as the receptor and AspA as the ligand, as this is deemed computationally favourable (Kozokov *et al*. 2017). The first docking runs were without any alterations to the structure modification parameters, nor specified any attractive residues on either protein. The first iteration used the AspA dimer in the docking run, however this did not provide a good approximation of the SAXS-generated model **(Figure 4.6)**, therefore successive iterations specified a monomer of AspA (Chain A) docking with the ParB-N monomer. This is a limitation of this approach, as, although the binding stoichiometry of the interaction is 1:1 (one ParB-N molecule to one AspA monomer), AspA binds to DNA as a dimer, and the ParB-N helices 11-13 also dimerise. Subsequent docking iterations were performed, adjusting parameters such that unstructured

terminal residues were removed, along with labelling specific residues as attractive, based on the SAXS data (aa 242-272 for ParB-N α11-13 and aa 71-92 for AspA α4). This iterative refinement resulted in a closer approximation of the SAXS model depicted in **Figure 4.6**.

The process was repeated, this time using the Phyre2 generated model of pNOB8 ParB-N **(Figure 4.5)**. The pNOB8 ParB-N residues found specifically in α11-13 (aa 249-257, 259-262 and 274-283 respectively) were input as the attractive residues for this protein, along with aa 71-92 for AspA α4. The ClusPro docking model is shown in **(Figure 4.7)**, with AspA α4 appearing in close enough proximity to pNOB8 ParB-N α11-13 to generate four pairs of intramolecular contacts, two of which (ParB-N:AspA; Asn-253:Val-92 and Arg-260:Glu-83, respectively) are shown below (**Figure 4.7, right)**. Two sets of these amino acids have oppositely charged side chains, which may help to reinforce interactions between the two proteins. This complex is only representative of how the actual ParB-N:AspA interface could be formed, however these residues represent potential targets for mutagenesis to test for any effect on complex formation.

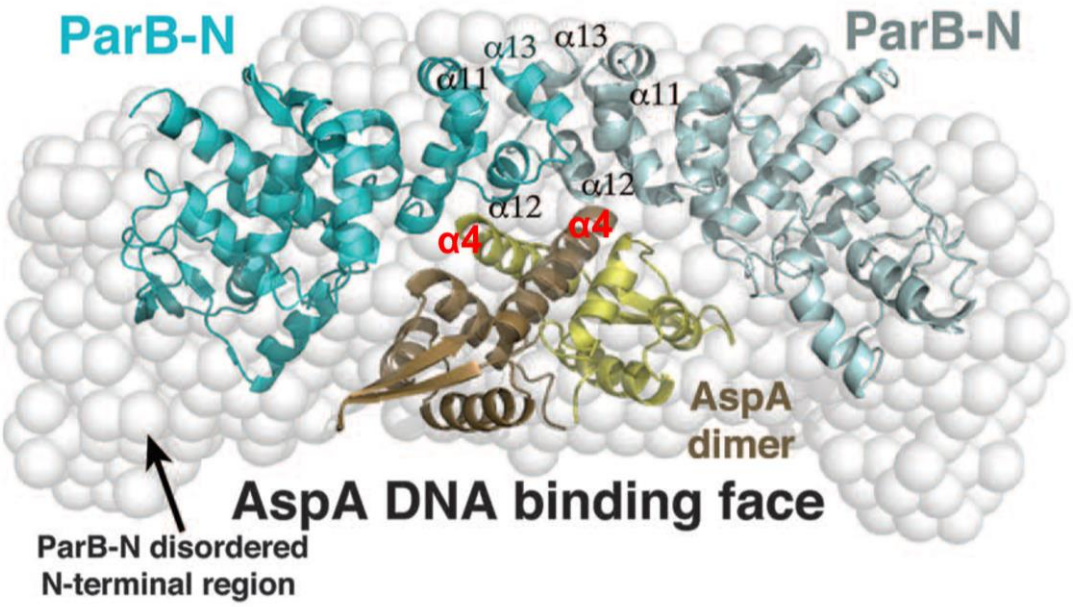


Figure 4.6. AspA-ParB-N SAXS model. The small angle x-ray scattering (SAXS) generated model for the AspA-ParB-N interaction places the AspA dimer centrally, flanked by ParB-N molecules. The ParB-N α11-α12-α13, which dimerise in the model, grasp the AspA C-teminal alpha helix (α4, red). The model is consistent with biochemical data showing a 1:1 AspA-ParB-N stoichiometry, and leaves the AspA N-terminal face free to contact the DNA. Figure adapted from Schumacher *et al*. 2015.

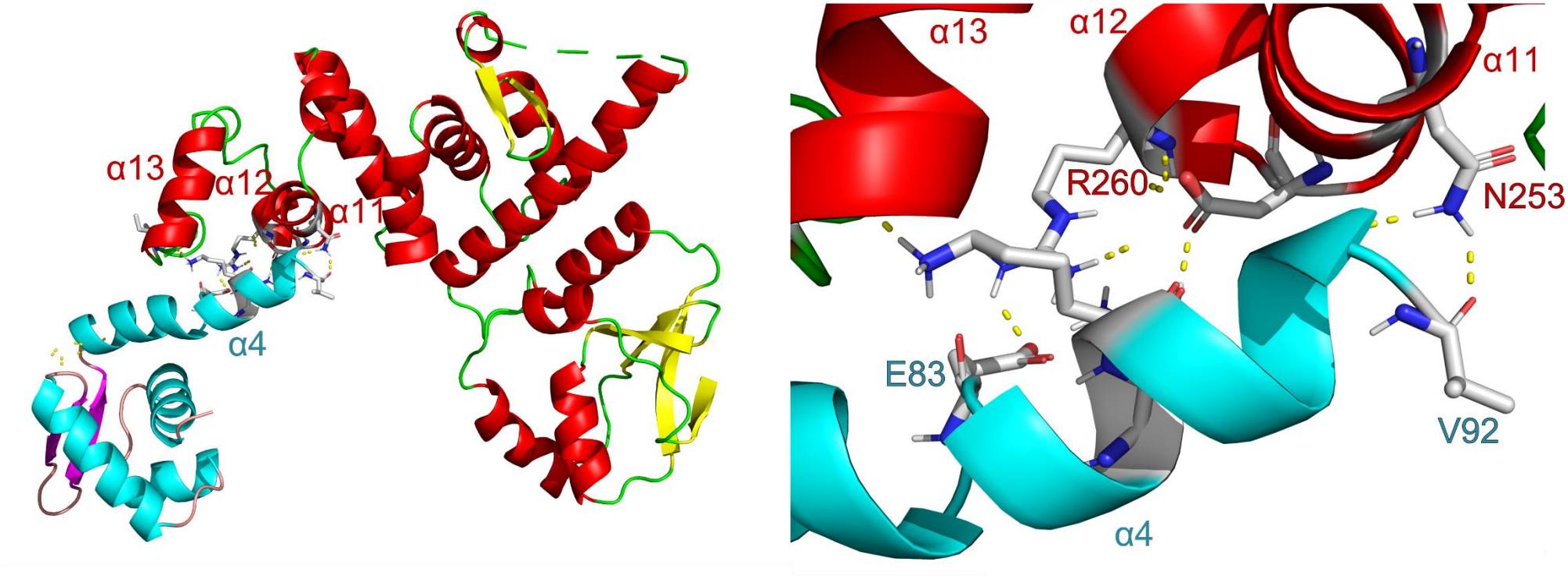


Figure 4.7. ClusPro molecular docking. The ClusPro molecular docking server was used to assess the interface between ParB-N and AspA. **(Left)** The pNOB8 ParB-N PHYRE2 model was docked with AspA. ParB-N is shown in red/yellow, AspA in cyan/purple. The alpha helices thought to be involved in interface formation are labelled, with residues forming intramolecular interactions shown as white sticks. **(Right)** Close-up of interface region. The interactions between the two sets of amino acids for ParB-N and AspA are shown, with hydrogen bonds depicted as yellow dashed lines.

4.2.5 ClusPro docking of AlphaFold pNOB8 ParB and AspA

The ParB model was coloured according to the AlphaFold predicted domain boundaries using Pymol, allowing the domains and unstructured regions to be clearly defined **(Figure 4.8, left)**. The AlphaFold server was also used to predict the structure of ParB-N terminus alone, with the aim of constructing another ParB-N:AspA docking model using ClusPro. However, comparison of the full-length ParB and ParB-N predicted structures showed significant difference in the spatial arrangement of alpha helices 11-13 (those potentially involved in AspA binding), due to α13 being the last helix before the start of the flexible linker, when superposed together (not shown). Therefore, the full-length ParB predicted structure was used in ClusPro docking simulations with AspA, using ParB as the receptor molecule, and the AspA:DNA structure as the ligand. ClusPro outputs four sets of structures based on scoring schemes derived from the type of interaction energy coefficients between the two proteins; balanced, electrostatic-favoured, hydrophobic-favoured, and van der Waals plus electrostatic effects. The ClusPro authors state that the size of a cluster is proportional to the model probability, and that lower-energy conformations are not necessarily closest to native structures (Kozakov *et al*. 2017). For the ParB-AspA structure, hydrophobic-favoured docking produced clusters with the most members **(Table 4.1)**, however on closer inspection these models were incorrect, showing AspA as binding to ParB-C rather than the N-terminus at helices 11-13.

Table 4.1 Summary of ClusPro output for pNOB8 ParB-AspA docked structure

Scoring coefficient	Cluster members	Lowest energy score
Balanced	104	-1807.70
Electrostatic-favoured	91	-1837.60
Hydrophobic-favoured	176	-2161.10
VdW plus electrostatic	90	-374.60

In actuality, the electrostatic-favoured coefficient models produced docked structures that appeared more probable based on previous modelling and biochemical data. The electrostatic-favoured docked structure is shown below **(Figure 4.8, right)**, and in this model, pNOB8 ParB α13 is in close proximity to AspA α4. This arrangement also leaves

AspA α3 free to be inserted into the DNA major groove. It should be repeated that this model is based on the interactions of the monomeric proteins only, and this is one of the limitations when using the software. PyMOL was then used to map the interface between the two proteins by looking at polar contacts between the residues of ParB α13 and AspA α4. Hydrogen bonding between lysine-287 of ParB, and valine-89, valine-92 and glutamine-93 of AspA, and between lysine-293 of ParB and glutamic acid-81 of AspA is depicted below **(Figure 4.8, right, close-up)**. Due to time constraints, and the fact that no apparent chemical cross-linking was observed, it was not possible to experimentally validate the interface shown in this model using mass-spectrometry, an approach recently used in the laboratory to assess the interface between the ParA and ParB proteins of the *E. coli* plasmid pB171 (Alnaqshabandy thesis 2020).

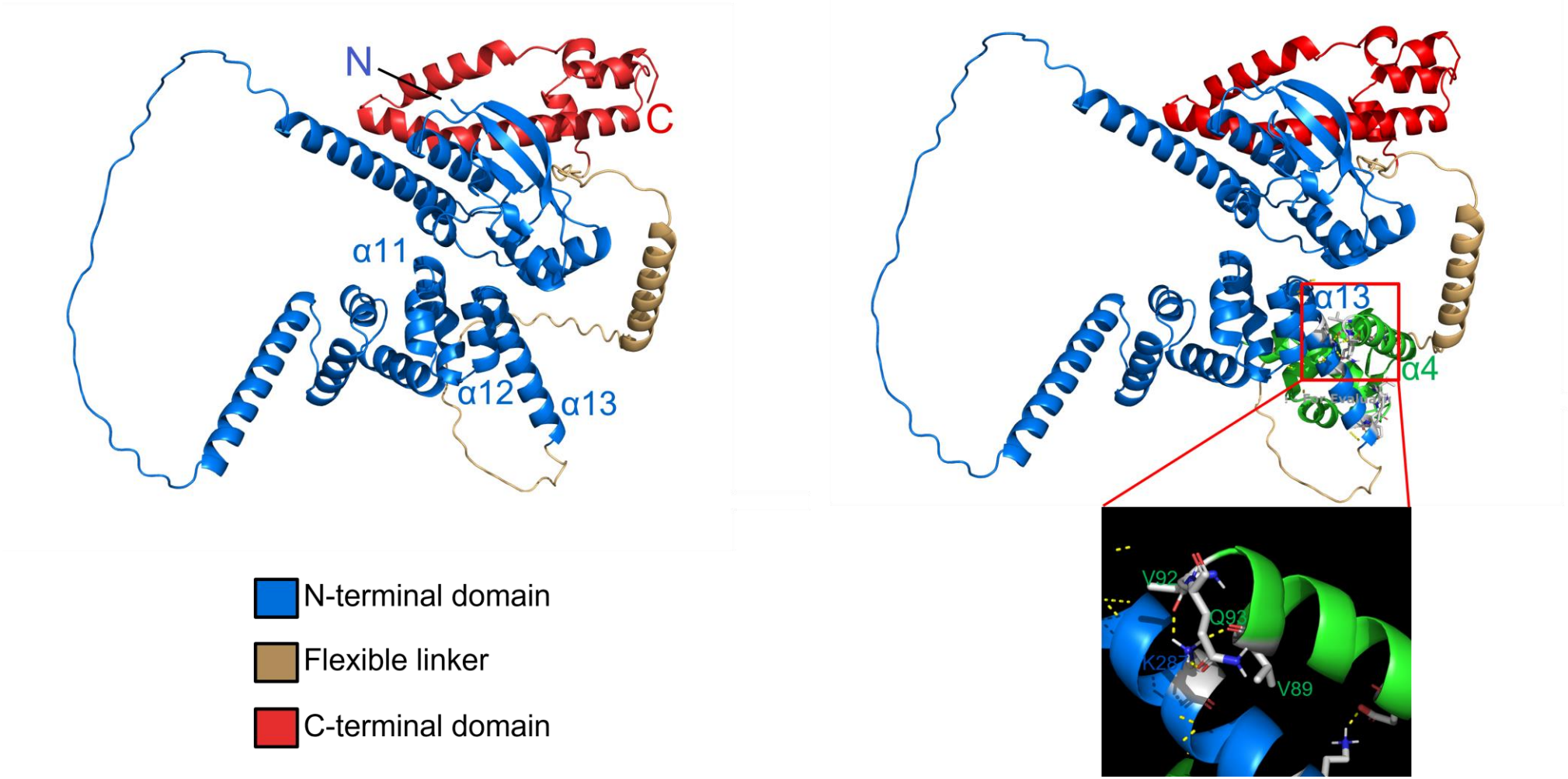


Figure 4.8. Pymol visualisation of AlphaFold predicted ParB structure. (Left) The structure is in the same orientation as in **Figure 4.3** for comparison. The N-terminal domain is coloured blue, and is comprised of two-subdomains separated by an unstructured region. The already structurally defined C-terminus is shown in red. The flexible linker separating the N- and C-terminal domains is coloured gold. The N and C-termini, along with alpha helices 11-13 of the N-terminal domain, are labelled. **(Right)** ClusPro model of AspA (green) docked with pNOB8 ParB. The AspA α4 helix is labelled. A close up of the interface is shown below, with residues making polar contacts coloured white, and hydrogen bonds as yellow dashed lines.

4.2.6 Generation of the pNOB8 ParB-N construct

The designation of the domain boundaries of pNOB8 ParB now allowed the creation of various ParB constructs, originally for assessing interactions with ParA, as outlined above in Section 4.1.1. The constructs relevant to this chapter were ParB-N (residues 1 to 299) and ParB-N plus the flexible linker (residues 1 to 374, henceforth denoted ParB-N & L). The expression vector chosen for cloning was pET-22b(+), previously used to express the *aspA* gene, which harbours a sequence encoding a non-cleavable 6xHis tag at the -3' end of the multiple cloning site. Using a different cloning vector, such as pET-15b, which incorporates a cleavable N-terminal tag, would perhaps have been better, as the C-terminal tag could potentially interfere with cross-linking (see Conclusions and discussion). The restriction sites chosen for insertion of the *parB-N* sequence were *Xho*I and *Nde*I. Therefore, primers were designed incorporating the restriction site sequences for *Xho*I and *Nde*I (CTCGAG and CATATG respectively, 5' to 3'), plus a 6 bp tail (AAGGAA) to aid the restriction digest. The primer sequences for the *parB-N* and *parB-N & L* constructs are reported in Table 2.7, Materials and Methods.

7 µg of pET-22b(+) was used in the restriction digest, and the successful digestion was verified by agarose gel electrophoresis, with the linearised plasmid (~5.5 kb) appearing at the correct location on the gel **(Figure 4.9A)**. The *parB-N* encoding fragment was amplified by PCR using Phusion DNA polymerase due to its increased fidelity, and run on an agarose gel to confirm its correct size (~0.9 kb). The fragments were excised from the agarose gel **(Figure 4.9B)**, purified, and digested using the same restriction enzymes *Xho*I and *Nde*I. The digested fragments were ligated into the linearised pET-22b(+) backbone, and ligation reactions transformed into *E. coli* DH5α competent cells. After growing transformed cells on selective media, 18 colonies were selected (plus a negative control without the *parB-N* insert) for analysis by colony PCR. The colony was used as the DNA template in the PCR reaction, as outlined in Methods 2.3.8.4. All PCR reactions were then run on an agarose gel to check for presence of the *parB-N* insert, indicating that cloning had potentially been successful **(Figure 4.9C)**. Purified plasmids from those clones deemed to harbour the correct insert were then subjected to restriction digest with *Xho*I and *Nde*I, and the size of the products verified by gel electrophoresis **(Figure 4.9D)**.

The clones which incorporated the correctly sized insert were then verified by sending each plasmid for sequencing, confirming the correct insertion of *parB-N*. Cloning of the *parB-N & L* construct was conducted by Dr Nicholas Read.

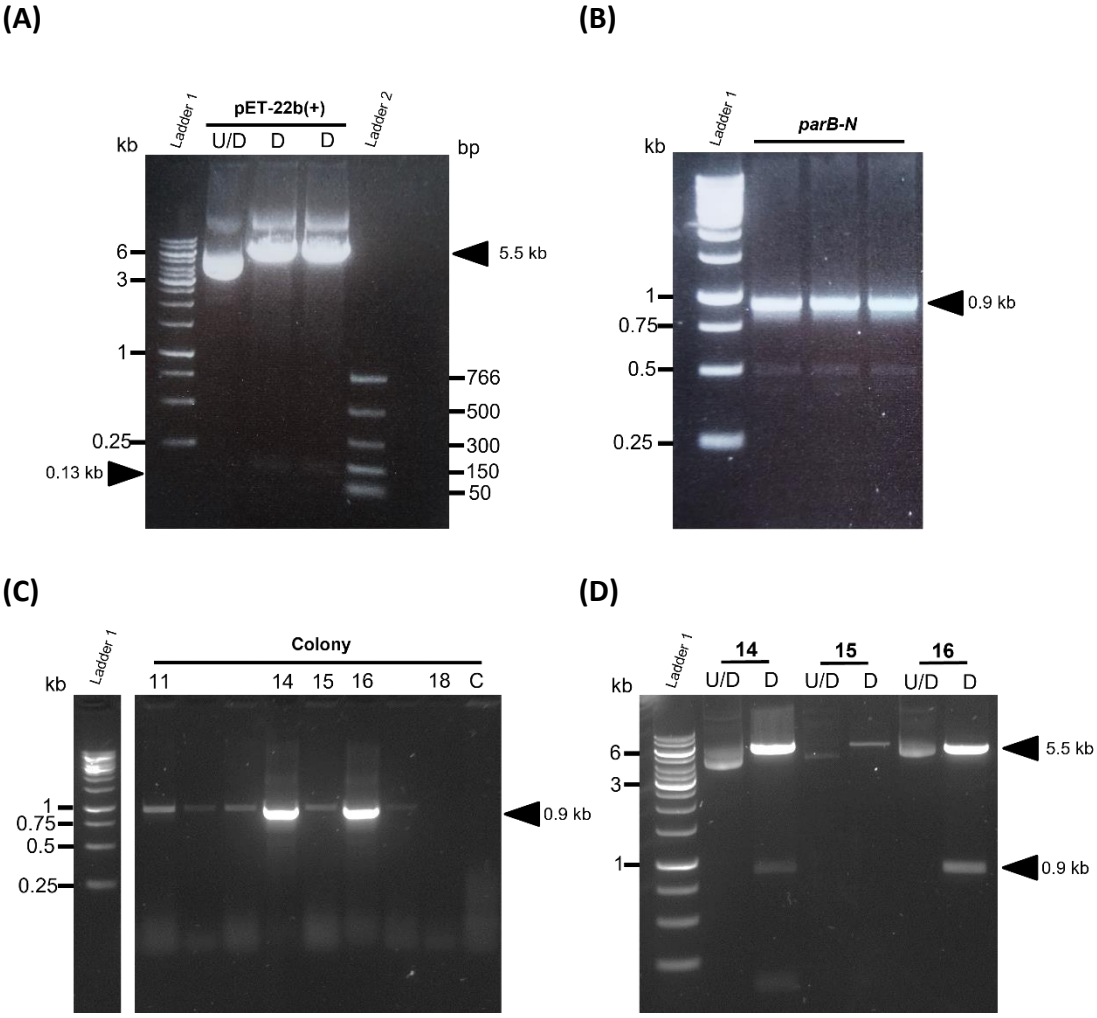


Figure 4.9. Overview of cloning procedure for *parB-N*. (A) Agarose gel showing the plasmid pET-22b(+) digested with restriction enzymes *Xho*I and *Nde*I, producing the linearised form of ~5.5 kb. The 130 bp region excised is just visible at the bottom of the gel. U/D = undigested, D = digested. **(B)** Agarose gel showing a typical PCR amplification of the *parB-N* fragment. The fragment encoding the N-terminus is 897 bp. **(C)** Agarose gel showing the colony PCR results for ligation transformants. 18 colonies were selected, only colonies 11-18 are shown. Lane C represents a negative control colony where no DNA insert was used in the reaction. The size of the insert (897 bp) is indicated. **(D)** Diagnostic digest of colonies positive for the insert. Plasmids from colonies 14-16 were digested with *Xho*I and *Nde*I, with the band at 0.9 kb indicating that the *parB-N* fragment was successfully incorporated. U/D = undigested, D = digested. All agarose gels shown are 1% w/v. In all cases Ladder 1 indicates GeneRuler 1 kb marker (Thermo Scientific), and Ladder 2 indicates PCR marker (NEB).

4.2.7 Overproduction and purification of ParB proteins

The pET-22b(+) plasmid containing the *parB-N* insert was transformed into the *E. coli* overexpression strain BL21 (DE3) CodonPlus competent cells. This overexpression system was previously used to overproduce the AspA proteins utilised in the previous chapter. A standard overexpression protocol was used, in which a liquid culture of 300 ml of BL21 (DE3) cells containing the plasmid was grown until the OD_{550} had reached 0.8 – 0.9, at which point expression of *parB-N* was induced by the addition of 1 mM IPTG (see Materials and Methods 2.4.1). Aliquots of the culture were taken prior to induction, then every hour for 3 hours. Overproduction of the ParB-N protein was analysed by SDS-PAGE, which indicated sufficient levels of production after 3 hours of incubation at 37°C. **(Figure 4.10A)**. The ParB-N domain is estimated to be ~35.5 kDa (monomer), including the 6xHis tag, as calculated by inputting the primary sequence into the online tool Expasy. The ParB-N protein appeared to migrate slightly lower than the 35 kDa molecular weight marker band on SDS-gels, and there also appeared to be a secondary band at ~20 kDa **(Figure 4.10A)**. The full length ParB protein has previously been shown to be susceptible to proteolysis, due to the accessibility of its unstructured flexible domain to proteases (not shown). Here, it appears a similar proteolysis may be occurring, as the IUPRED3 disorder plot predicts another region of disorder between residues ~120-180 **(Figure 4.4, middle)**. Proteolysis occurring near either boundary of this disordered region, or centrally, would produce a protein fragment of a similar size to that observed on the gel, at around 18-20 kDa.

The overproduced ParB-N construct was purified using Ni^{2+} affinity chromatography, as outlined in Materials and Methods Section 2.4.3. Initial purification resulted in large quantities of the proteolytic fragment of ~20 kDa (not shown), therefore the purification was repeated, this time doubling the number of protease inhibitor cocktail tablets that were used, adding additional tablets in both the initial cell suspension, and also to the filtered cell lysate. This reduced the amount of the proteolytic fragment in the purified fractions, but did not completely eliminate it **(Figure 4.10B)**. After buffer exchange into storage buffer (20 mM HEPES, 150 mM NaCl, pH 7.0), the concentration of the purified ParB-N fractions was measured by Bradford assay, averaging 1.4 mg/ml per fraction.

The full-length pET-ParB construct was available in the laboratory collection, therefore this protein was also overproduced and purified using the same methodology as described for the ParB-N construct. The predicted molecular weight of ParB, including the 6xHis tag, is ~56 kDa (monomer), and overproduction after induction with IPTG showed that ParB was present after 3 hours **(Figure 4.10C)**. Again, a secondary band was visible on the SDS gel, of approximately 45 kDa. This would equate to the size of the N-terminus plus the flexible linker, as ParB-N is 35.5 kDa, and the estimated molecular weight of the linker is 9 kDa. Therefore, it is likely that proteolysis of ParB may be predominantly removing the C-terminus of the protein, although there are other bands present further down the gel (including one at ~20 kDa) indicating that proteolysis is occurring at other regions within the protein **(Figure 4.10C)**. The protein was purified in the same manner as for ParB-N, again increasing the number of protease inhibitors used. This removed some of the proteolytic fragments, but not all, though the majority of the protein appears as the full length ParB, when assessed via SDS-PAGE **(Figure 4.10D)**. After the eluted fractions were buffer exchanged into storage buffer, the concentration of the ParB fractions was measured by Bradford assay, averaging over 3 mg/ml per fraction. In both the purification of ParB and ParB-N, a second affinity chromatography, passing elution fractions from the first purification over a second nickel column could have been performed to obtain purer protein samples. In addition, SEC could have been employed to separate the proteins from the proteolytic fragments to further increase sample purity.

The ParB-N+L construct was overproduced and purified by Dr Nicholas Read, in the same manner as for the ParB-N and full-length ParB proteins. The monomeric form of ParB-N+L has a predicted molecular weight of 44.5 kDa.

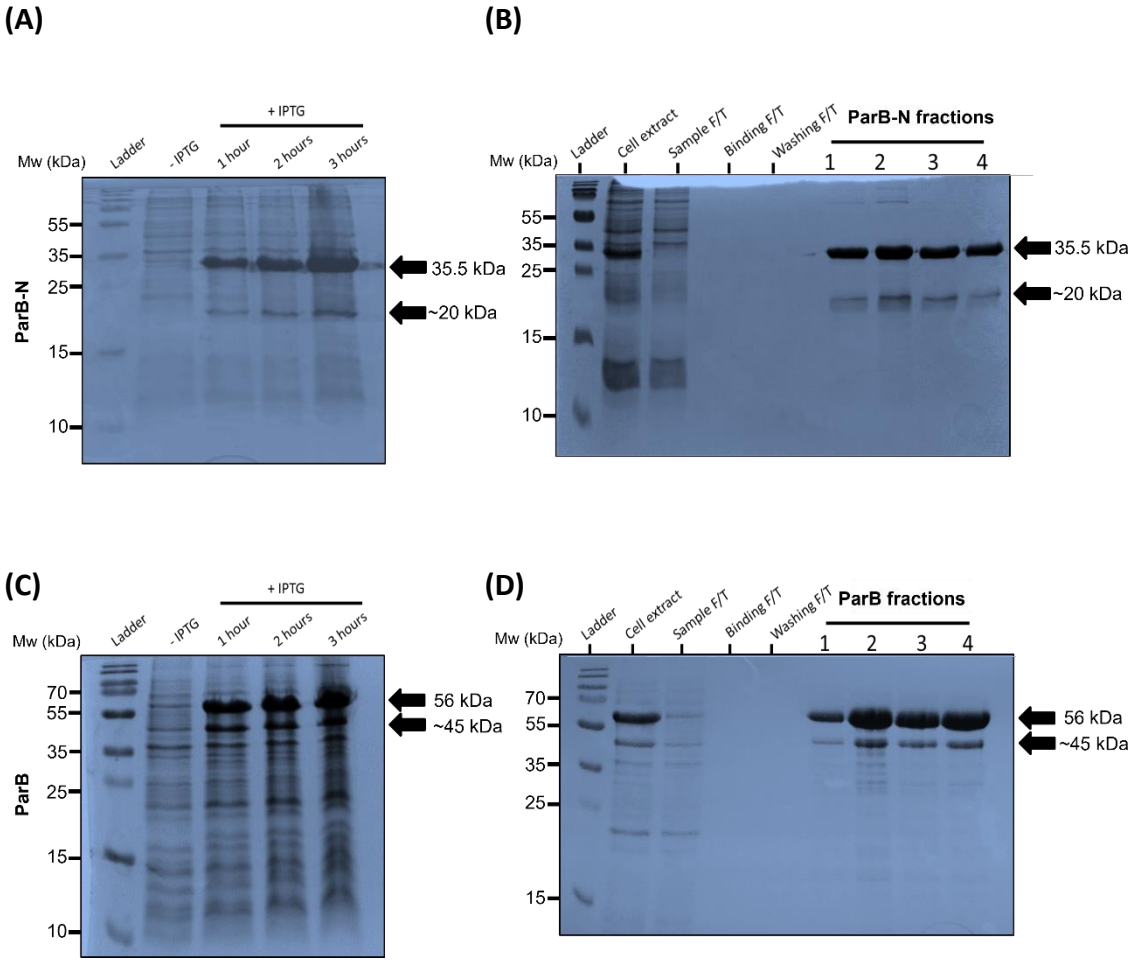


Figure 4.10. Overproduction and purification of full-length ParB and ParB-N. (A) The presence of overproduced ParB-N was measured in uninduced culture, then 1, 2 and 3 hours after induction with IPTG, with the results assessed by SDS-PAGE. ParB-N (~35.5 kDa) and the proteolytic fragment (~20 kDa) are indicated with black arrows. **(B)** SDS gel showing the results of purification of ParB-N. Lanes marked Cell Extract, Sample F/T, Binding F/T and Washing F/T contain aliquots taken from various stages of the purification; Cell Extract is the crude lysate prior to chromatography, Sample F/T is the cell extract after circulation over the column for 2-3 hours, and Binding F/T and Washing F/T after two washing steps with binding buffer and wash buffer respectively. The final four lanes are aliquots taken from eluted fractions of ParB-N. Bands representing ParB-N and the proteolytic fragment are indicated with black arrows. **(C)** The same protocol was used to induce the overproduction of the full-length ParB protein. The full-length ParB (~56 kDa) and proteolytic fragment (~45kDa) are indicated with black arrows. **(D)** SDS gel showing the results of purification of ParB-N. Lanes are the same as in **(B)**. The Mw ladder used in both cases is PageRuler Plus prestained protein ladder (Thermo Scientific).

4.2.8 DMP chemical cross-linking of the AspA and ParB proteins

Chemical cross-linking experiments were performed to assess interactions between AspA and pNOB8 ParB-N. Dimethyl pimelimidate (DMP) was initially used as the chemical cross-linker, as it had been shown to cross-link AspA to itself and promote the formation of dimers, trimers and tetramers at concentrations of 10 mM (see section 3.2.3.1). The hypothesis was that DMP would cross link AspA and ParB-N, and that this would result in a complex of ~47.5 kDa (based on molecular weights for each monomer) that could be resolved on an SDS gel. An additional aim was to repeat the experiment using a different cross-linking reagent, followed by mass-spectrometry to identify the individual lysine residues that are in close enough proximity for cross-linking to occur, and thus define the interaction interface between the two proteins.

The two proteins were buffer-exchanged into cross-linking buffer (50 mM HEPES, 50 mM KCl, 5 mM $MgCl_2$, pH 8.5), and their concentrations measured by Bradford assay. The two proteins were incubated individually, in the presence or absence of 10 mM DMP as controls, with the expectation that dimers and/or higher order oligomers would be formed. Then, AspA and ParB were incubated together at an equimolar ratio, in the presence or absence of 10 mM DMP. All reactions were incubated for 80°C for 1 hour, a temperature at which AspA had successfully been cross-linked before, and the reactions were loaded onto a 12% SDS-polyacrylamide gel. AspA formed a dimer as expected, but surprisingly, no dimeric form of ParB-N was evident when cross-linked alone upon the addition of 10 mM DMP **(Figure 4.11A)**. The proteolytic fragment of ~20 kDa was evident, which appears to migrate to the same location on the gel as the AspA dimer. When AspA and ParB-N were incubated together in the presence of DMP, no complex was visible at the expected molecular weight of ~47.5 kDa for two interacting monomers **(Figure 4.11A, red arrow)**, nor any higher molecular weight species equating to a dimer-dimer complex. There is a weakening of the AspA monomer band, and the appearance of a AspA dimer band, however this dimer band is quite faint, and the ParB-N band also appears slightly weaker in the presence of DMP **(Figure 4.11A, last two lanes)**. This could suggest an AspA:ParB-N interaction, as it appears that the reduction in AspA monomers does not produce a concomitant increase in dimers. It is possible that a larger

AspA:ParB-N complex is being formed that is not resolved on the gel. The experiment was conducted twice, using both cross-linking buffer and storage buffer as the diluent, however this did not alter the result.

One potential explanation for the lack of cross-linking of ParB-N could be that the 6xHis-tag, located at the C-terminal end of the construct may be hindering dimerisation of the protein via alpha helices 11-13. To test this, the ParB-N&L construct was used alongside ParB-N, as here the 6xHis-tag is positioned at the end of the flexible linker. The same mass of protein (13 µg) was used for each reaction, and the proteins were incubated with 0, 5 or 10 mM DMP. The reactions were incubated at 80°C for 1 hour, then analysed by SDS-PAGE. No cross-linking was evident for either ParB-N or ParB-N&L **(Figure 4.11B)**. The experiment was repeated at 37°C, to test whether ParB requires a different incubation temperature, however this did not alter the result (not shown).

The full-length ParB protein was then used, as previous data had showed it to be amenable to DMP cross-linking. A larger mass of protein was used in each reaction (30 µg), along with a greater maximum DMP concentration of 20 mM. The reactions were incubated at 80°C for 1 hour, with the resulting SDS gel showing the gradual increase in formation of ParB dimers (Mw ~112 kDa) at increasing DMP concentrations, along with the presence of higher order oligomers **(Figure 4.11C)**. The experiment was repeated once more, using all proteins cross-linked alone in the presence or absence of 10 mM DMP. An equal mass of protein was used in each reaction (16 µg), and reactions were incubated at 80°C for 1 hour. Analysis by SDS-PAGE showed that only AspA and full-length ParB formed distinct dimers **(Figure 4.11D)** although there is a faint smearing visible for ParB-N + DMP at the approximate molecular weight equal to a dimer (~70 kDa), indicating a small amount of cross-linking may be occurring. It is possible that the cross-linking of AspA hinders the ability of ParB-N to bind to the AspA dimer, perhaps due to the cross-linked AspA dimer adopting a different conformation to the endogenous form. Alternative approaches such as fluorescence polarisation and isothermal titration calorimetry have previously demonstrated this interaction (Schumacher *et al.* 2015), though in both cases ParB-N bound to the AspA-DNA complex, therefore a lack of DNA in this reaction may also prevent binding between the two proteins (see below).

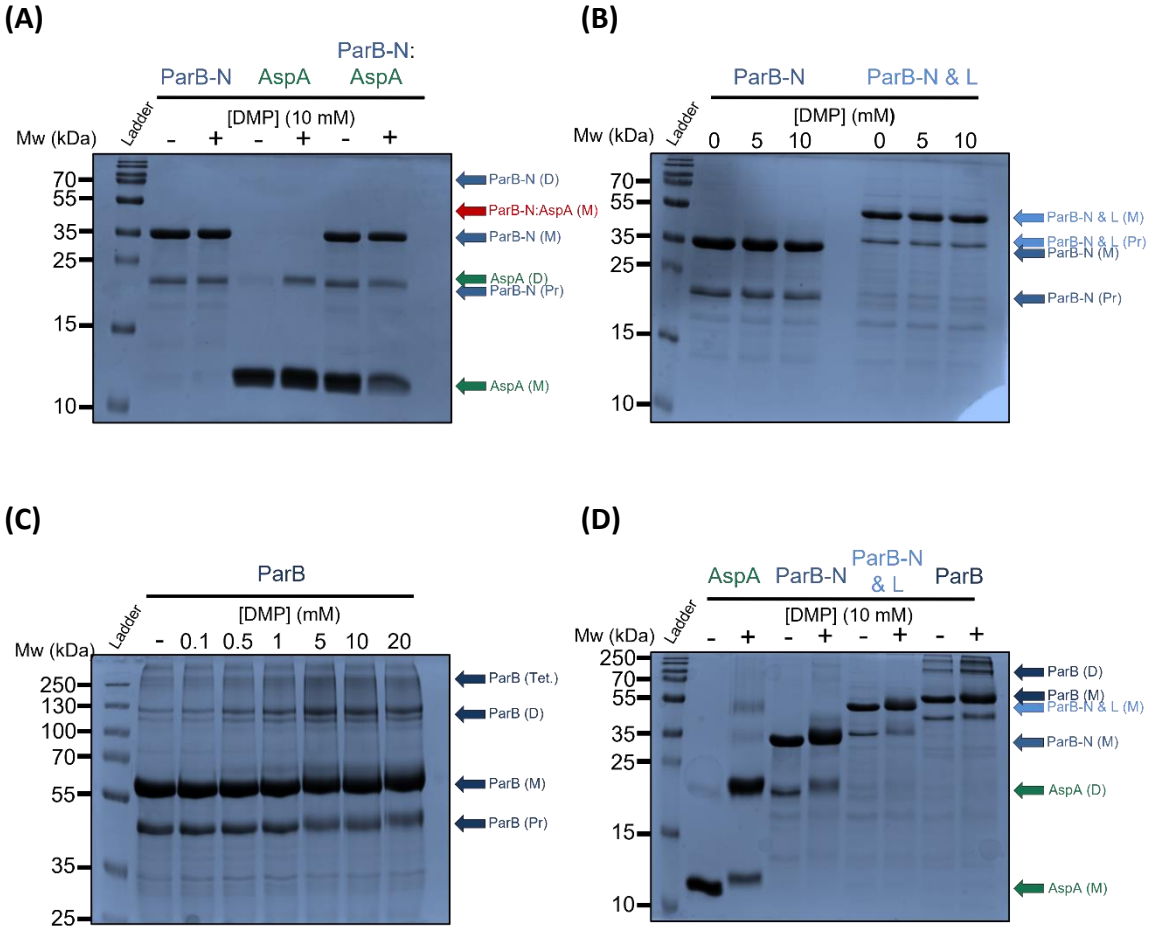


Figure 4.11. DMP cross-linking of ParB and AspA proteins. (A) SDS-polyacrylamide gel showing DMP cross-linking reactions containing ParB-N alone, AspA alone, and ParB-N and AspA. Each reaction was either without DMP or with 10 mM DMP. Arrows to the right indicate molecular weights equating to a particular oligomeric state; (M) = monomer, (D) = dimer, (Pr) = proteolytic fragment. The red arrow equated to the location on the gel where a predicted crosslinked ParB-N:AspA monomeric complex should appear. **(B)** Cross-linking of ParB-N and ParB-N & linker, at DMP concentrations of 0, 5 and 10 mM. Arrows depict the monomeric form and proteolytic fragments. **(C)** Cross-linking of the full-length ParB protein, at increasing concentrations of DMP from 0 to 20 mM. (Tet.) = tetramer. **(D)** Cross-linking of AspA alone, ParB-N alone, ParB-N & linker alone, and ParB full-length alone, either without DMP or with 10 mM DMP. In all gel images the Mw ladder used is the PageRuler Plus prestained protein ladder (Thermo Scientific).

4.2.9 SEC-MALLS of ParB proteins

Due to the lack of dimerisation of the ParB-N and ParB-N & L proteins observed in the DMP cross-linking experiments, SEC-MALLS was carried out to assess the oligomeric state of the proteins in solution. The full-length ParB protein was also included in the analysis, as this was shown to dimerise upon the addition of DMP, and unpublished SEC-MALLS experiments performed in the Barillà group show the protein to be dimeric in solution. The purified proteins were run on a Superdex S200 column before MALLS analysis, using 20 mM HEPES, 150 mM NaCl, pH 7.0 as running buffer, and all runs were conducted by Dr Andrew Leech. Each protein was provided as an aliquot in storage buffer (the same composition as running buffer), but also after dialysis overnight against the running buffer, in order to minimise any light-scattering problems that occurred previously with AspA. The dialysed proteins were typically 60 – 70% of the concentration of the pre-dialysis aliquots, as measured by Bradford assay after dialysis. Here, Bradford assay may not be the most appropriate method for measuring the concentration of ParB and ParB constructs, due to the presence of disordered regions within the proteins. An assumption in using colorimetric assays such as Bradford is that the target proteins exhibit uniform behaviour approximate to the globular proteins (here, bovine gamma globulin) used as standards in the assay (Contreras-Martos *et al.* 2018). The SEC-MALLS data shown below in **Table 4.2** and **Figure 4.12** are taken from the non-dialysed samples; the dialysed samples showed very similar plots, but molecular weight estimates were less accurate due to the decreased concentrations, and were therefore omitted.

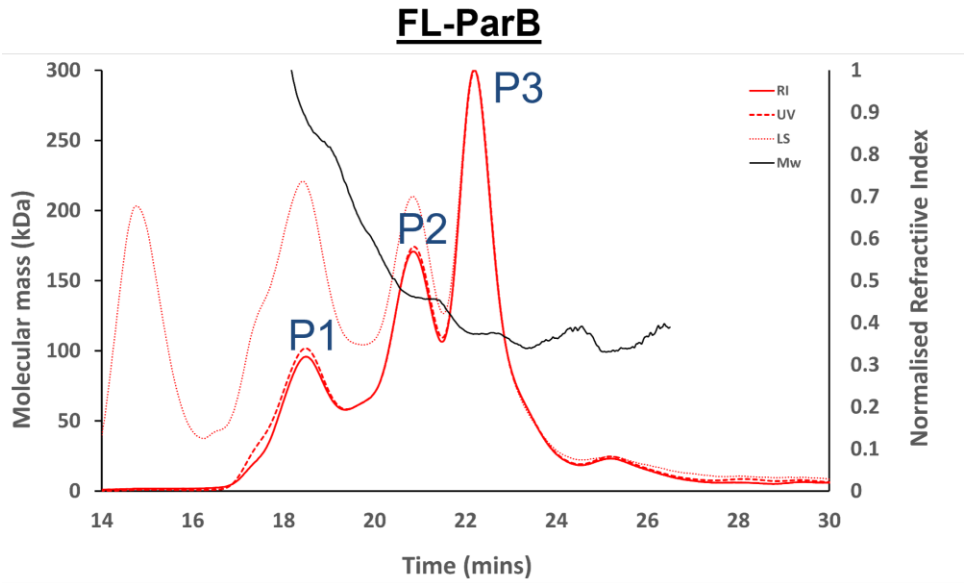
The full length ParB protein eluted as three main peaks as shown by the refractive index lines, with the majority of the protein eluting in peak 3 (P3), with a molecular weight of 115 kDa, which is in close agreement to the theoretical molecular weight for the dimer of 112 kDa **(Figure 4.12, top). Table 4.2** shows the theoretical molecular weights for different oligomeric states of each protein, the estimated amount of protein under each peak, and its corresponding molecular weight. The estimated mass of protein, in µg, is derived by integrating the area under each peak. There are two other main peaks; P2 has an experimental Mw of 149 kDa, less than a trimer, and so may represent a heterogenous mix of dimers and trimers, particularly as the molar mass line is not horizontal across the

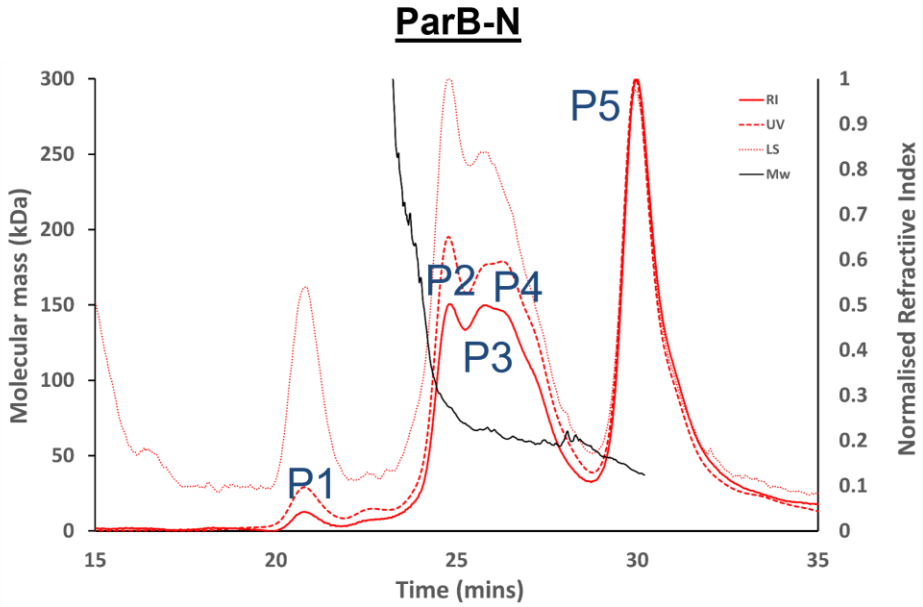
peak. P1 has a Mw of 272 kDa and so represents higher-order oligomeric structures; these are also apparent in the light-scattering traces which show small amounts of aggregated material at 15 minutes elution time **(Figure 4.12, top)**. The ParB-N plot showed a mixture of peaks, however the peak equivalent to the greatest mass (P5) has a Mw of 38 kDa, very close to the theoretical molecular weight of 35.5 for a ParB-N monomer **(Figure 4.12, middle)**. Peaks 2-4 overlap somewhat, but their molecular weight ranges suggest a heterogeneous mix of dimers and trimeric species. For ParB-N & L, again the peak equivalent to the largest mass (P3) has an experimental Mw closest to that of a monomer; however here the difference is larger than the 5% variation normally observed with clearly resolved peaks (53.8 kDa *cf.* theoretical monomer Mw of 44.5 kDa). The second main peak, P2, equates to a similar mass of protein, and is in close agreement with the theoretical molecular weight of a dimer (91 *cf.* 89 kDa), therefore it appears that ParB-N & L is a roughly even mix of monomeric and dimeric species in solution **(Figure 4.12, bottom)**. These SEC-MALLS data in part provide an explanation for the lack of cross-linking seen with ParB-N and ParB-N & L. ParB-N elutes predominantly as a monomer, whereas ParB-N & L appears an amalgam of monomers and dimers. This could be evidence of the hexa-His tag preventing dimerisation of the ParB-N protein, as hypothesised, however this explanation is not plausible for ParB-N & L. Perhaps the unstructured linker, when unattached to the C-terminus of ParB and not in its native state, adopts a range of conformations, some of which act to inhibit binding at the monomer-monomer interface of the protein.

Table 4.2 Molecular weight estimates of ParB proteins used in SEC-MALLS

Protein	Theoretical Mws (kDa)				Amount under peak (µg)					Mw for peak (kDa)				
	M.	Di.	Tri.	Tet.	P1	P2	P3	P4	P5	P1	P2	P3	P4	P5
FL-ParB	56	**112**	168	224	16.9	27.2	**36.2**			272	149	**115**		
ParB-N	**35.5**	71	106.5	142	0.4	4.7	73	4.6	**14.6**	763	89	65	59	**38**
ParB-N & L	**44.5**	89	133.5	178	2.5	18.2	**19.5**			209	91	**53.8**		

M = monomer, Di. = dimer, Tri. = Trimer, Tet. = Tetramer. The peaks with the greatest mass, their corresponding molecular weights, and the closest theoretical oligomeric Mw, are in bold-face.


FL-ParB
P1
P2
P3
RI
UV
LS
Mw
Molecular mass (kDa)
Normalised Refractiive Index
Time (mins)


ParB-N
P1
P2
P3
P4
P5
RI
UV
LS
Mw
Molecular mass (kDa)
Normalised Refractiive Index
Time (mins)

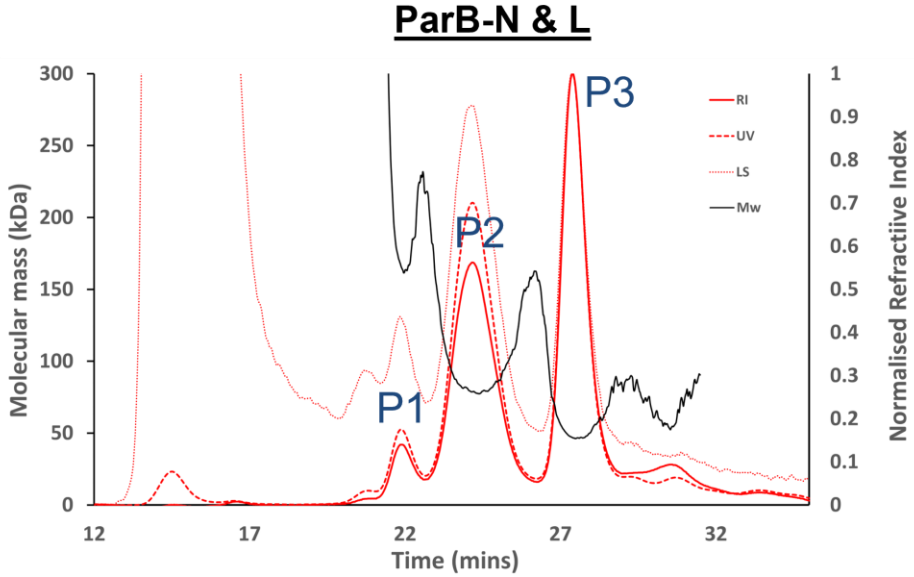

ParB-N & L
P1
P2
P3
RI
UV
LS
Mw
Molecular mass (kDa)
Normalised Refractiive Index
Time (mins)

Figure 4.12. (Previous page) SEC-MALLS of ParB proteins. Molar mass vs. time plots for WT ParB (FL-ParB), ParB-N, and ParB-N plus linker (ParB-N & L) proteins. In each case, the solid red lines represent the refractive index, the dashed line the UV trace, and the small dotted line the light scattering. The black line under each peak indicates an estimate of the molecular weight for the given peak at that elution time. The individual elution peaks for each protein are labelled P1, P2 etc, and correspond to the data in Table 4.2.

Given that pNOB8 ParB-N and ParB-N&L did not appear amenable to cross-linking, it was decided to use the full-length ParB protein to assess interactions with AspA. The cross-linking experiments were repeated, again incubating AspA and ParB alone, or together in the presence or absence of 10 mM DMP. For the AspA-ParB reaction, the proteins were used at an equimolar ratio. Due to the fact that the molecular weights of AspA and ParB are quite different, and therefore they appear on different regions of the gel, it was apparent that resolving all dimers for both proteins, plus a potential AspA:ParB complex on a single gel may be difficult. To overcome this, reactions were made in twice the volume, incubated at 80°C for 1 hour, then split equally across two SDS gels. One gel was run until the 10 kDa molecular weight marker band was near the bottom **(Figure 4.13A)**, whilst the other was run until the 25 kDa marker protein was the bottommost band **(Figure 4.13B)**. In this way, AspA monomers and dimers could be seen on the first gel, whilst ParB dimers and tetramers were resolved on the second gel. However, no band at the molecular weight equivalent to a AspA:ParB monomer-monomer complex was observed on either gel **(Figure 4.13A&B, red arrows)**.

This raised the possibility that the presence of DNA may be required in the reaction to aid formation of a complex between the two proteins. The model derived from SAXS data showed that the ParB-N molecules conform to the already present AspA:DNA complex, suggesting that ParB-N binding may only occur once the AspA:DNA structure is formed, with AspA-DNA forming a template for ParB-N binding (Schumacher *et al.* 2015). This could be brought about by AspA undergoing a conformational change, as the N-terminal arms are known to adopt a range of different configurations upon DNA binding (Schumacher *et al.* 2015). To test this possibility, a region of DNA 126 bp in length, incorporating the second AspA binding site was amplified, extracted from an agarose gel, and purified. Additionally, the C-terminus of ParB has been demonstrated to bind DNA non-specifically, therefore it was also possible that the

addition of non-specific DNA to the cross-linking reaction this may be required for and/or enhance ParB interactions with AspA.

The next set of cross-linking experiments therefore included combinations of AspA and ParB crosslinked alone, AspA and ParB crosslinked in the presence and absence of the 126 bp fragment (specific DNA) and an aliquot of NOB8-H2 genomic DNA extract (non-specific DNA). AspA and ParB were cross-linked alone at equal mass (21 µg per 20 µl reactions), and for AspA:ParB cross-linking, they were used at equimolar ratios (11.25 µM final concentration). The specific DNA was used at a final concentration of 0.3 µM, and 60 ng of non-specific DNA was added to the final reaction. Reactions were made in twice the volume as before, incubated at 80°C for 1 hour, then split equally across two SDS gels to better resolve the bands as previously described. There was no distinct band equating to an AspA:ParB complex upon analysis by SDS-PAGE **(Figure 4.13C&D)**, nor did the addition of specific DNA, nor specific plus non-specific DNA appear to promote the formation of any complex. However, as mentioned previously, changes in band intensities are apparent, which may indicate complex formation. The AspA monomer band is lessened with the addition of DMP, as is expected, however the dimer band is also quite faint **(Figure 4.13A, *cf.* lanes 5&6)**. There is also some smearing close to the expected Mw of a AspA-ParB monomer complex which may indicate some interaction **(Figure 4.13B, *cf.* lane 6)**, and additionally, larger complexes (e.g. dimer-dimers) may be forming which are not resolved on the gel, or at a similar molecular weight to another species (e.g. at 140 kDa).

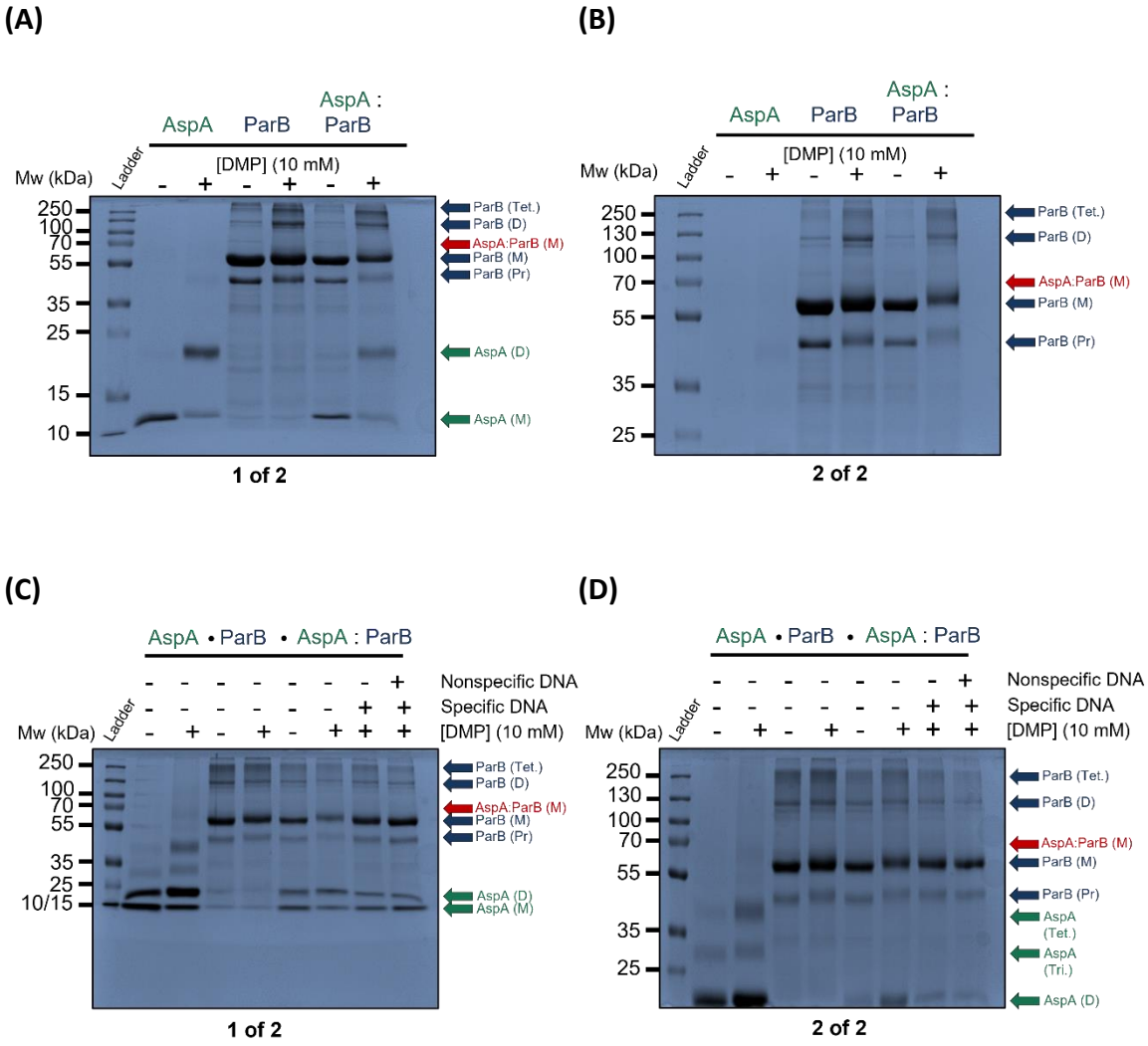


Figure 4.13. DMP cross-linking of ParB and AspA in the presence of DNA. (A) SDS-polyacrylamide gel showing DMP cross-linking reactions containing ParB alone, AspA alone, and ParB and AspA together. Each reaction was either without DMP or with 10 mM DMP. Arrows to the right indicate molecular weights equating to a particular oligomeric state; (M) = monomer, (D) = dimer, (Tet.) = tetramer, (Pr) = proteolytic fragment. The red arrow signifies the location on the gel where a predicted crosslinked ParB:AspA monomeric complex should appear. 1 of 2 signifies that the reaction was split across two gels, each run for a different length of time. **(B)** 2 of 2; the reactions are the same as in **(A)**, but run for longer to better resolve the higher molecular weight complexes. **(C)** Cross linking as in **(A)**, but with the addition of specific DNA and specific plus non-specific DNA to the final two lanes. 1 of 2. **(D)** 2 of 2; reactions as in **(C)**, but run for longer until the 25 kDa marker band was near to the bottom of the gel. In all gel images the Mw ladder used is the PageRuler Plus prestained protein ladder (Thermo Scientific).

4.2.10 BS3 cross-linking of AspA and ParB

Given that DMP did not promote cross-linking between ParB and AspA, it was decided to repeat the experiment using a different cross-linker, bis[sulfosuccinimidyl] suberate (BS3). Similar to DMP, BS3 forms stable amide bonds with primary amines in lysine side chains or N-termini of polypeptides, although BS3 has a slightly longer spacer arm of 11.4 Å. Initially, BS3 cross-linking was conducted with ParB proteins (full-length ParB, ParB-N and ParB-N & L) alone, to test if the different cross-linker promoted complex formation with ParB-N and ParB-N & L. A 50 mM stock solution of BS3 was used (Materials and Methods 2.4.11). BS3 is recommended to be used at between 20 and 50-fold molar excesses for protein concentrations of <5 mg/ml, and at final concentrations of 0.25 – 5 mM. BS3 was used at a final concentration of 2 mM, which gave molar excesses of between 24 and 35-fold for the three proteins. 40 µg of each protein were used, and the reaction incubated at 80°C for 30 minutes. The reactions were analysed by SDS-PAGE, which showed, as when DMP was used, cross-linking is apparent for the full-length ParB, but no complexes of ParB-N nor ParB-N & L were formed **(Figure 4.14A)**. However, the ParB-N & L monomer and proteolytic fragment bands become much less intense upon addition of BS3, without the concomitant appearance of a dimer band **(Figure 4.14A, *cf* lanes 5&6)**. It is possible that some material was stuck in the well of the SDS gel and so did not electrophorese correctly; alternatively, larger complexes could be forming which were not resolved on the gel. The experiment was also performed using AspA cross-linked alone, alongside the ParB proteins, at both 37°C and 80°C, with the same results (not shown).

The experiment was repeated again with the addition of specific and non-specific DNA, to test if this resulted in any complex formation. 40 µg of protein were used for the AspA only and ParB only reactions, whereas for AspA plus ParB reactions, proteins were used at equimolar ratios, at final concentrations of 10.6 µM. BS3 was used at a final concentration of 0.5 mM, a 50-fold molar excess. The 126 bp specific DNA was used at final concentrations of 0.3 µM, whilst 250 ng of non-specific DNA was added to the final reactions. The reactions were incubated for 30 minutes at 80°C. Instead of splitting the reactions across two SDS gels, a gradient polyacrylamide gel was used to achieve optimal resolution on a single gel. Analysis by PAGE showed that there was no complex equating

to ParB:AspA visible on the gel **(Figure 4.14B)**. At equimolar ratios, the mass of AspA that was loaded was quite small, resulting in faint bands (*cf.* first two lanes with last four lanes) and so decreasing the likelihood that an AspA:ParB complex would be observed. Therefore the experiment was repeated using 15 µg each of AspA and ParB (for the AspA:ParB reactions), however this did not alter the result (not shown).

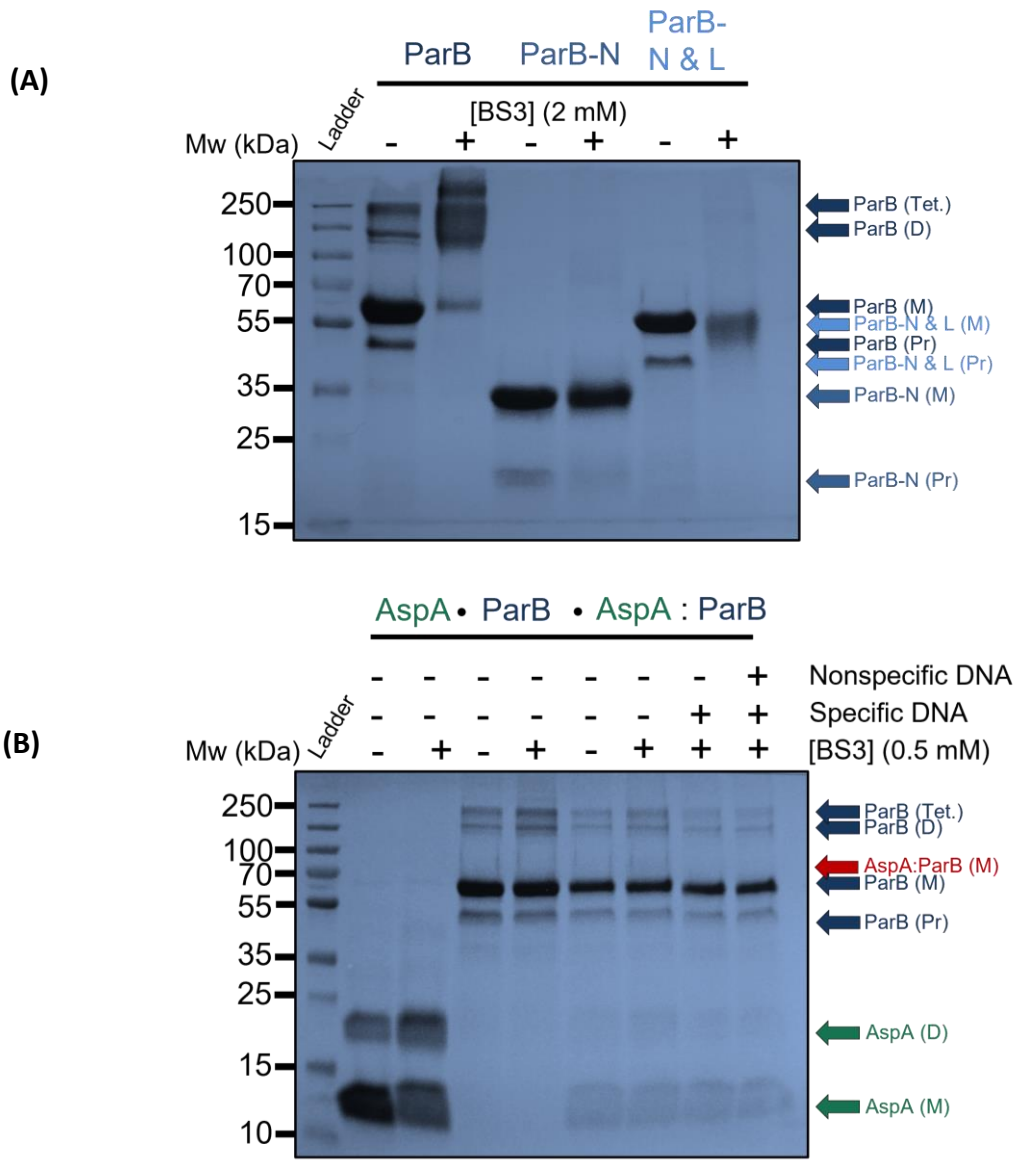


Figure 4.14. BS3 cross-linking of AspA and ParB proteins. (A) SDS-polyacrylamide gel showing BS3 cross-linking reactions containing ParB alone, ParB-N alone, and ParB-N&L alone. Each reaction was either without DMP or with 10 mM DMP. Arrows to the right indicate molecular weights equating to a particular oligomeric state; (M) = monomer, (D) = dimer, (Tet.) = tetramer, (Pr) = proteolytic fragment. **(B)** Gradient polyacrylamide gel showing BS3 cross-linking reactions containing AspA, ParB alone, and ParB and AspA together. The red arrow signifies the location on the gel where a predicted crosslinked ParB:AspA monomeric complex should appear. In all gel images the Mw ladder used is the PageRuler Plus prestained protein ladder (Thermo Scientific).

4.3 Conclusions and discussion

The effective and accurate segregation of genetic material via active partitioning mechanisms relies not only on protein-DNA transactions, but also the interactions between the partition proteins. In bacterial systems, once the centromere-binding protein (ParB type) has bound to the DNA to form the partition complex, it then interacts with its partner protein, the ATPase motor protein (ParA type) (Barillà *et al*. 2007, Schumacher 2008, Baxter & Funnell 2014, Bouet & Funnell 2019). Here, the segregation cassette of the archaeal plasmid pNOB8 encodes three proteins, and the previous chapter assessed the binding of AspA, the CBP in this system, to its cognate binding site. The ParB protein of pNOB8 therefore does not act as the CBP, but has been demonstrated to play an adaptor-like role, as it can interact with the other two proteins, AspA and ParA, whilst also exhibiting non-specific DNA-binding activity *in vitro* (Schumacher *et al*. 2015). Of particular interest in this chapter is the interaction between ParB, specifically its N-terminal domain, and AspA, although originally, broader aims involving ParA binding experiments were also planned.

When first solving the crystal structure of ParB-N, a homologue from the closely related *S. solfataricus* strain 98:2 was used in place of pNOB8 ParB-N. After solving the structure, small angle X-ray scattering (SAXS) was used to generate an interaction model, in which the alpha helices at the C-terminal end of 98:2 ParB-N, α11-13, were involved in both the dimerisation of ParB, and the interaction with AspA via its C-terminal alpha helix (α4). The SAXS data was corroborated by *in vitro* experiments with pNOB8 ParB, which demonstrated the importance of helices 11-13 in AspA interactions.

In this chapter, the aim was to characterise further the interaction interface between pNOB8 ParB and AspA, using chemical cross-linking to form a ParB-N:AspA complex, followed by mass-spectrometry to identify the specific residues that had been brought into close proximity by the interaction. Firstly, the domain boundaries of pNOB8 ParB were re-evaluated, and it appeared from bioinformatic analysis that the flexible linker between the N- and C-terminal domains was longer than previously thought, at 75 amino acids rather than ~50. The linker is thought to play an important role in the overall

mechanism of pNOB8 segregation, as unpublished data suggests that this region of ParB interacts with the ParA motor protein, and also enhances the ATPase activity of ParA (Rodriguez-Castañeda, unpublished). The interplay between the CBP and motor protein is observed in bacterial systems, where ParB (or functional analogues) stimulate ATP hydrolysis of the motor protein (Barillà *et al*. 2007, Ringgaard *et al.* 2009, Vecchiarelli *et al*. 2010, Lim *et al*. 2014). In archaea, the *S. solfataricus* chromosomal CBP SegB elicits increased polymerisation of the motor protein SegA (Kalliomaa-Sanford *et al*. 2012), and whilst pNOB8 ParB is not the CBP in this system, nevertheless its interactions with ParA are presumably vital for effective plasmid segregation. Unfortunately, it was not possible to test the interactions of the ParB linker with ParA during this project due to time constraints, however the creation of new protein constructs based on newly-defined domain boundaries of ParB will be of use in future experiments.

In this chapter, experiments focussed on the interactions between the pNOB8 ParB N-terminus and the CBP AspA. DMP chemical cross-linking was employed, using the ParB-N protein construct and AspA. It was apparent that ParB-N was not amenable to cross-linking, as no ParB-N dimer, nor ParB-N:AspA complex was observed, and the same result was obtained when the ParB-N & L construct was used. It is possible that the location of the hexa-histidine tag at the 3' end of *parB-N* was problematic when translated, and prevented dimerisation of the protein, though this less likely to be the case with the ParB-N & L construct. In hindsight, using a different cloning vector which for example incorporates a cleavable hexa-histidine tag at the N-terminus, may have alleviated this problem. SEC-MALLS data appeared to corroborate the lack of dimerisation of the constructs compared to the wild-type ParB, therefore full-length ParB was used in subsequent cross-linking experiments. However, no cross-linking and formation of a ParB:AspA complex was observed, using either DMP or BS3 cross-linkers. This was also true upon the addition of DNA containing the AspA binding site, as AspA:DNA may function as a scaffold for ParB binding. It is possible that in these experiments, an insufficient amount of specific DNA was used, as the molar ratio used (0.3 µM) was approximately half the amount required for AspA to completely cover this size of DNA fragment. Previous *in vitro* assays that demonstrated ParB-N binding to AspA either used a pre-formed AspA:DNAcomplex (at binding saturation concentrations), or a molar excess

of ParB-N to AspA:DNA (Schumacher *et al*. 2015), therefore these conditions could be replicated in future experiments with full-length ParB.

Chapter 5

Sulfolobus NOB8-H2 genome analysis

Chapter 5

Analysis of the *S. islandicus* NOB8-H2 genome and the conjugative plasmid pNOB8

5.1 Introduction

The *Sulfolobus* strain NOB8-H2, harbouring the conjugative plasmid pNOB8, was first isolated from thermal hot springs at Noboribetsu, on the Japanese island of Hokkaido (Schleper *et al*. 1995). Members of the hyperthermophilic genus *Sulfolobus* have been used as model organisms to study fundamental biological processes occurring in the crenarchaeal phylum, such as the cell cycle, DNA replication and cell division (Bernander 2000, 2007), the DNA-damage response (Feng *et al*. 2018), cell-envelope structure and function (Zhang, *et al*. 2018), and carbon metabolism (Quehenberger, *et al*. 2017). There are now over 20 complete *Sulfolobus* genome sequences available in the NCBI database at the time of writing, including newly proposed species (Dai *et al*. 2016).

The first conjugative plasmid isolated from an archaeon, pNOB8, was found in *Sulfolobus* NOB8-H2, and later sequencing of the plasmid gave insights into mechanisms involved in plasmid stability, conjugation and copy number control, and how these processes may differ in archaea compared with those of bacterial plasmids (She, *et al*. 1998). The plasmid pNOB8 was also shown to be able to integrate into the host chromosome at a particular tRNA integration site via a 'pNOB8-type' integrase, and this integration could be a means by which horizontal gene transfer and evolutionary change can occur (She *et al*. 2004). Subsequently, many *Sulfolobus* extrachromosomal elements, including plasmids and viruses, have been identified and characterised (Li 2015).

The segregation mechanism of pNOB8 is also under investigation. Bacterial plasmid partitioning systems are well-described, but little is known about how genomic segregation mechanisms operate in archaea. The bacterial partition system generally comprises three elements: two proteins and a centromeric DNA sequence that acts as a site of recruitment of the proteins and assembly of a nucleoprotein complex dubbed the segrosome (Hayes & Barillà 2006b). The two proteins are a DNA binding factor, called

ParB, which exhibits specificity for the centromeric site *parS*, and an ATPase denoted ParA, which is recruited into the nucleoprotein complex formed by ParB-DNA interactions.

Two pNOB8 open reading frames (ORFs) shared sequence similarity with bacterial *parA* and *parB* genes (She 1998), providing clues to their potential role in segregation of the plasmid. Further study of the partition cassette showed it to have a tricistronic organisation, with a third ORF that did not display similarity to previously characterised bacterial genes, however the protein encoded by this third gene (dubbed AspA) was shown to be a site-specific DNA-binding protein (Schumacher 2015).

5.1.1 Aims

In this chapter, the sequencing and subsequent analysis of the *Sulfolobus* NOB8-H2 genome using a variety of bioinformatic tools and approaches will be described. Interactions between the chromosome and pNOB8 are also discussed.

These main aims will be addressed:

1) Where does *Sulfolobus* NOB8-H2 sit within the phylogenetic tree of *Sulfolobus*, and does NOB8-H2 represent a novel strain, or perhaps a novel species?

2) What are the main features of the *Sulfolobus* NOB8-H2 genome, and how do these features compare with other sequenced species/strains?

3) What are the interactions and commonalities between *Sulfolobus* NOB8-H2 and the plasmid pNOB8, and are there any novel features of pNOB8 that may be of interest to future studies?

5.2 Results

5.2.1 Sequencing of the *Sulfolobus* NOB8-H2 genome

The total genomic DNA of the *Sulfolobus* strain NOB8-H2 was isolated from log-phase culture, and sequenced using the MinION (Nanopore). Assembly of the genome produced 17 contigs, most of which were between 3 and 60 kbp in size, and contained many pNOB8 fragments. One contig however, contig 45, was 3.07 Mbp in size and had a GC content of 36%, similar to other sequenced *Sulfolobus* strains (Dai *et al*. 2016), and presumably represented the NOB8-H2 chromosome The genome distance estimation program MASH was used to assign a similarity score for the NOB8-H2 genome in comparison with other sequenced *Sulfolobus* genomes in the NCBI database. MASH estimates genome similarity by comparing the pairwise distances of small sections, or 'hashes' of the genome (of e.g. 1 kb), rather than across the entire genome (Ondov *et al.* 2016). Here, 1000 of these selected hashes were used, and the similarity scores for NOB8-H2 against other *Sulfolobus* strains are depicted in **Table 5.1**. The MASH data show that *Sulfolobus* strain NOB8-H2 is 100% identical to *S. solfataricus* P1 across these hashes. An NCBI BLAST alignment of the entirety of contig 45 against the *S. solfataricus* P1 reference (accession NZ_LT549890) showed a >99% sequence similarity, confirming that the laboratory strain NOB8-H2 was in fact *S. solfataricus* P1. The genomes of other closely-related *Sulfolobus* strains are also present in the MASH data, for example, strain P2 has shared hashes of 941/1000. This does not imply that there are multiple separate genomes present, but reflects the similarity of closely-related strains, e.g. P2 to P1. Other, more distantly-related species of *Sulfolobus* that had a lower number of shared hashes were also returned by MASH, but have been omitted from **Table 5.1** for brevity.

Table 5.1. MASH results for *Sulfolobus* NOB8-H2[a]

Shared Hashes	Coverage	Organism name	RefSeq assembly/accession
1000/1000	58	*Sulfolobus solfataricus* strain P1	GCF_900079115.1
1000/1000	155	*Sulfolobus* sp. NOB8-H2 plasmid pNOB8	NC_006493.1|
1000/1000	118	*Sulfolobus islandicus* plasmid pKEF9	NC_006422.1|
994/1000	72	*Sulfolobus* virus 2	GCF_000842405.1
941/1000	58	*Sulfolobus solfataricus* P2	GCF_000007005.1

[a] Only the first five MASH results are shown, in order of shared hashes. Coverage indicates an estimate of the average number of sequencing reads in the raw read set.

The conjugative plasmid pNOB8 was also present with 100% of shared sequences **(Table 5.1)**. Given that the method of DNA extraction used should have purified genomic DNA, and not plasmid DNA, this raised the following possibilities: pNOB8 was detected as it is integrated into the host chromosome, or that free-floating, extrachromosomal pNOB8 was detected due to the extraction process not eliminating all plasmid DNA, or a combination of the two. To assess this, BLAST searches were conducted using the pNOB8 sequence against the various NOB8-H2 contigs produced by the assembly, revealing that there are multiple insertions of pNOB8 into the host chromosome. Fragments of pNOB8 inserted into contig 45 are shown in **Figure 5.1**, and range from 28 bp to 9118 bp in size.

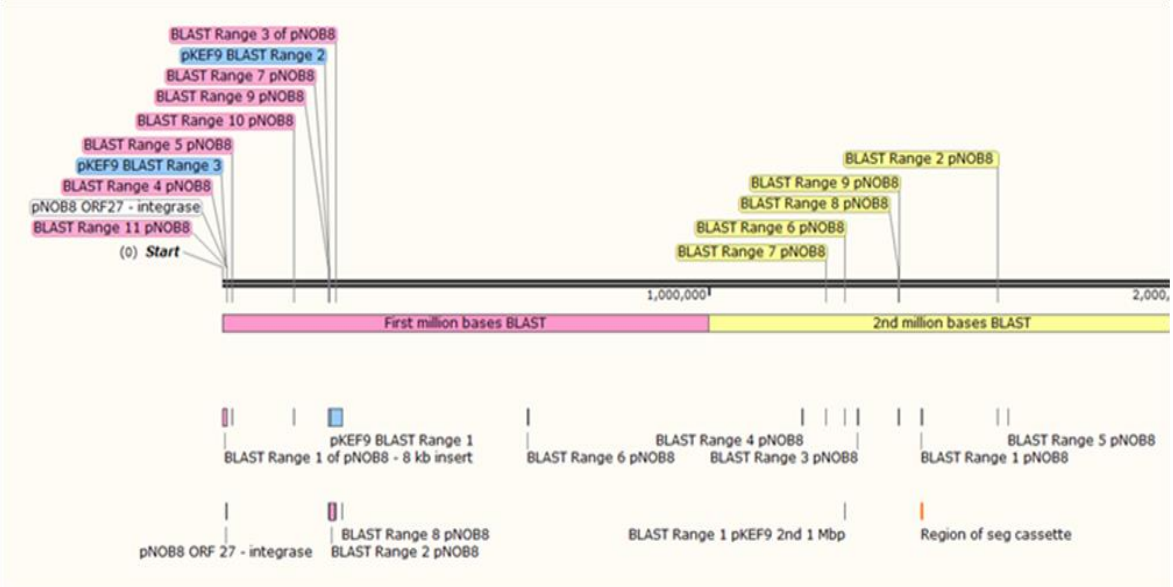
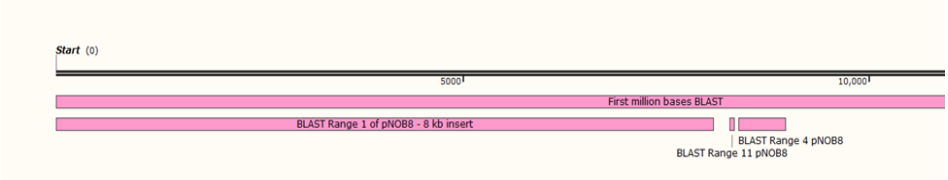


Figure 5.1. Schematic of pNOB8 insertions in contig 45. BLAST searches of pNOB8 against NOB8-H2 contig 45 were visualised using SnapGene Viewer. The pink and yellow bars underneath the linear contig represent 1 Mbp sections used for BLAST searches. The various integrated fragments of pNOB8 are shown above and below in the same colours, and are labelled 'BLAST Range X pNOB8'. Fragments of the plasmid pKEF9 are shown in blue. Only the first 2 Mbp of the contig is displayed - pNOB8 insertions in the 3rd Mb are omitted.

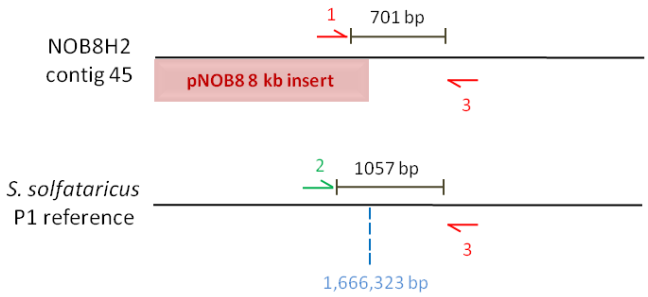
The NOB8-H2 contig 45, shown to be equivalent to *S. solfataricus* P1 by sequence identity, could not be completely circularised to form a closed chromosome, due to the presence of an 8 kb region of pNOB8 found at the start of the contig and spanning the point of circularisation **(Figure 5.2A)**. It was hypothesised that this 8 kb section might not be present in all reads, i.e. that there may be a mixed population of NOB8-H2 chromosomes, some containing this 8 kb insertion, and some without it. To test this, PCR was performed using primers spanning the plasmid fragment insertion point on contig 45, and the equivalent section on the P1 reference genome where the 8 kb fragment would

occur **(Figure 5.2B)**. The PCR products would be of different sizes if the pNOB8 fragment was inserted (701 bp), or not inserted (1057 bp), meaning a mixed population of chromosomes either harbouring the insertion, or not, could be distinguished on an agarose gel **(Figure 5.2C)**. A relative ratio 8:1 of chromosomes harbouring the insertion against those without the insertion was determined via densitometry of the PCR bands **(Figure 5.2C)**, although it is possible that PCR amplification bias due to differences in amplification efficiencies between the reactions affected this ratio (Silvia *et al*. 2005).

(A)



(B)



(C)

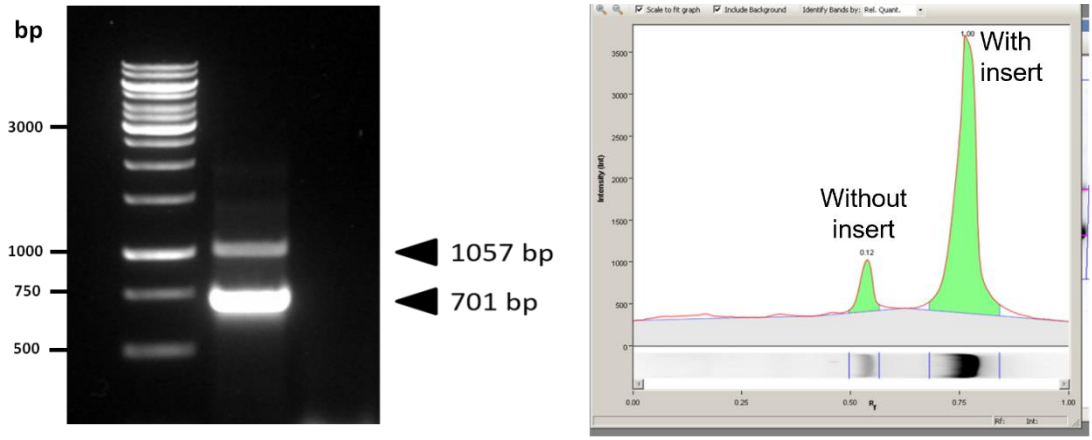


Figure 5.2. Strain NOB8-H2 has a mixed population of chromosomes. **(A)** BLAST result of pNOB8 against contig 45 revealed a fragment (denoted BLAST Range 1) of 8 kb located at the immediate start of the contig. **(B)** Schematic of PCR used to assess whether there was a mixed population of chromosomes. Primers 1 and 3, spanning the pNOB8 8 kb insert in contig 45, produce a 701 bp product (top). (Bottom) The location of the insert at the equivalent position of the P1 reference was calculated to be at position ~1.6 Mbp (blue dashed lines). Primers 2 and 3, annealing to chromosomes without the insert, produce a 1057 bp product. **(C)** (Left) Agarose gel of PCR products as in **(B)**. Both sets of primers were included in the same reaction. (Right) Band intensity was quantified using Image Lab software, giving a ratio of ~8:1 of chromosomes with the 8 kb insert to no insert.

Determining that the NOB8-H2 strain was comprised of a mixed population of cells, each containing different arrangements of pNOB8 insertions into the chromosome, demonstrated the complexity of the interactions between plasmid and host. It was likely that pNOB8 did not exist within the cell as an independent plasmid, but instead, the entire plasmid was incorporated into the chromosome in multiple different-sized fragments.

Surprisingly, another conjugative plasmid, pKEF9, was also found to be integrated into the *Sulfolobus* NOB8-H2 chromosome **(Table 5.1)**. pKEF9 shares approximately 10 kb sequence identity with pNOB8, therefore a region was chosen to amplify via PCR which was particular to pKEF9, thus demonstrating the presence of another conjugative plasmid. Plasmid pKEF9 was originally isolated from *Sulfolobus* cultures obtained in Iceland by Wolfram Zillig (Greve *et al*. 2004), but along with other low-copy number *Sulfolobus* plasmids, has been propagated in non-native hosts such as *S. solfataricus* P2 in order to increase their respective yields (Prangishvili *et al*. 1998). pKEF9 was previously used in a study to assess its conjugative and replicative properties inside different recipient *Sulfolobus* hosts (Liu, She & Garrett 2016); here, the recipient was *S. islandicus* REY15A, however the donor strain was a stable strain of *S. solfataricus* P1 that had previously been established by Zillig (Zillig *et al*. 1998).

Therefore, the discovery that the 'laboratory strain' of *Sulfolobus* NOB8-H2:

(i) Was in actuality *S. solfataricus* P1

(ii) Contained another plasmid, pKEF9 that had previously been stably propagated inside P1 to create a donor strain

led to the conclusion that the strain of *Sulfolobus* NOB8-H2 used thus far was not the actual original NOB8-H2 strain**.** It appears that this *Sulfolobus* NOB8-H2 strain (actually *S. solfataricus* P1) had also been used for conjugation experiments involving plasmid pNOB8, and that perhaps some unfortunate mix-up had occurred prior to acquisition of the strain by our laboratory. This meant that the 'original' NOB8-H2 strain should be obtained before conducting any further experiments relating to the interactions between pNOB8 and its natural parental host. In the following text, the 'laboratory strain' of

NOB8-H2 will be named as *S. solfataricus* P1, with the subsequently acquired (see below) 'original' strain named *Sulfolobus* NOB8-H2.

5.2.2 Sequencing of the original *Sulfolobus* NOB8-H2 strain

The original *Sulfolobus* NOB8-H2 isolate was first sampled from hot springs in Hokkaido, Japan (Schelper *et al*. 1995). The strain was obtained from the laboratory of Sonja Albers, University of Freiburg. The strain was grown in medium similar to that of *S. solfataricus* P1, with the exception of using sucrose as the carbon source rather than glucose (see Table 2.4, Materials and Methods). As before, cells were grown to log-phase before harvesting and extracting the genomic DNA. Initially, the extracted genomic DNA appeared to be contaminated with sugars (Dr Sally James, personal communication), therefore a modified version of the CTAB (cetyl trimethylammonium bromide) extraction protocol was used to remove polysaccharides prior to sequencing (performed by Dr Sally James). Initial sequencing was performed using the MinION Flongle, producing an assembly of 18 contigs. This was supplemented with an Illumina sequencing run, giving an additional 200 Mb of sequence, equating to genome coverage of ~60X. A polished, circularised chromosome was produced, and again was compared to other *Sulfolobus* genomes using MASH. Interestingly, *Sulfolobus* NOB8-H2 was found to be different to the previously sequenced strain *S. solfataricus* P1, and MASH data revealed NOB8-H2 to be less than 50% related (shared hashes 484/1000) to any other *Sulfolobus* strains in the NCBI database **(Table 5.2).** The MASH analysis, along with the assembly, polishing and annotation (see below) of the NOB8-H2 genome was performed by Dr John Davey.

Table 5.2. MASH results for the 'original' strain of *Sulfolobus* NOB8-H2 [a]

Shared Hashes	Coverage	Organism name	RefSeq assembly/accession
970/1000	89	*Sulfolobus* sp. NOB8-H2 plasmid pNOB8	NC_006493.1
484/1000	8	*Sulfolobus islandicus* M.14.25,	GCF_000022405.1
480/1000	8	*Sulfolobus islandicus* M.16.47	GCF_000245275.1
479/1000	8	*Sulfolobus islandicus* M.16.4,	GCF_000022445.1
478/1000	8	*Sulfolobus islandicus* M.16.13	GCF_000245135.1
		—	
140/1000	7	*Sulfolobus solfataricus* P2	GCF_000007005.1

[a] The first five MASH results are shown for *S. islandicus*, in order of shared hashes, then the first hit for *S. solfataricus*

Sulfolobus NOB8-H2 was between 43 and 48% similar to 20 other strains of *Sulfolobus islandicus*, but only 12 - 14% similar to strains of *S. solfataricus*. Therefore, NOB8-H2 was found to be more closely related to *S. islandicus* than *S. solfataricus*, but appears sufficiently different, having a maximum similarity of 48%, to be described as a novel strain of *S. islandicus*. Plasmid pNOB8 was also present in the sequencing data, with a shared hash value of 97%, whereas pKEF9 was not, demonstrating that this was likely to be the original isolate NOB8-H2, rather than a strain used for plasmid conjugation experiments as detailed above. Prior to the complete assembly of the genome, as a diagnostic tool, a region of pKEF9 that is not present in pNOB8 was chosen for PCR amplification. The previously sequenced *S. solfataricus* P1 contained a complete insertion of pKEF9, and pNOB8 fragments found at certain chromosomal locations. Using extracted genomic DNA as a template, PCR was conducted against the pKEF9 region, and a region spanning the pNOB8 insertion, for both the *S. solfataricus* P1 and *Sulfolobus* NOB8-H2 strains, with the resulting agarose gel demonstrating the difference between the two strains **(Figure 5.3)**.

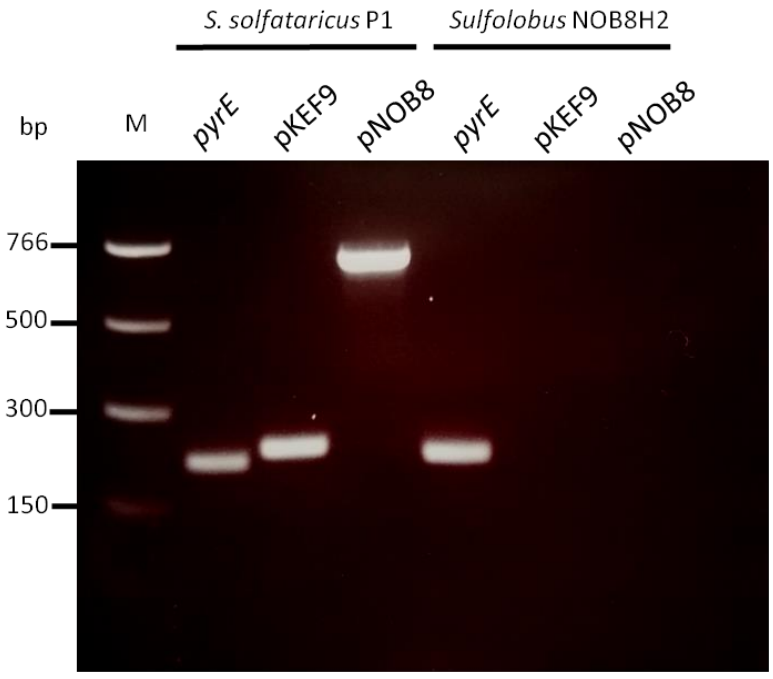


Figure 5.3. *Sulfolobus* NOB8-H2 does not contain plasmid pKEF9. A region of pKEF9 that is not present in pNOB8 was chosen as a diagnostic marker between the two strains. Amplification of this region should produce a fragment of 238 bp. The pNOB8 band of 701 bp is the same region that was amplified in **Figure 5.2**. A region of the *pyrE* gene was used as a positive control; this was previously known to be present in a NOB-H2 contig, along with the *S. solfataricus* P1 reference genome, and produces a fragment 216 bp in size. M = PCR marker ladder (NEB).

5.2.3 General properties of the *Sulfolobus* NOB8-H2 genome

The assembled, polished genome consists of a circular chromosome, 2,808,935 bp in length, with a G+C content of 35.8%. An annotation of the genome was performed using the prokaryotic genome annotation tool Prokka, a command line tool which uses external prediction tools to identify genomic features such as coding sequences, transfer RNA genes, non-coding RNA etc (Seeman 2014). 3,159 coding sequences were identified, and an initial annotation using the *S. islandicus* L.S.2.15 strain as a reference resulted in Prokka annotating 722 (23%) of the coding sequences. Increasing the number of references used for the annotation by including all 21 *S. islandicus* genomes raised the percentage of NOB8-H2 coding sequences annotated. Currently, 2,339 NOB8-H2 coding sequences are annotated (74%), with 820 remaining unannotated.

The NOB8-H2 chromosome contains 46 tRNA genes, three putative origins of replication, a CRISPR-Cas locus containing two distinct CRISPR systems, and 55 matches to plasmid pNOB8 when BLASTing against the chromosome **(Figure 5.4**., **green lines)**. Further details about these properties are given in subsequent sections. These general features of the NOB8-H2 genome are compared with other sequenced complete genomes from Sulfolobaceae family members in **Table 5.3**, and a map of the circularised NOB8-H2 chromosome is shown in **Figure 5.4**.

Table 5.3. Summary of properties of *S. islandicus* NOB8-H2 and other Sulfolobaceae complete genomes.

Strain	NCBI RefSeq	Genome size (Mb)	GC%	No. of Genes	No. of tRNAs	Habitat/Location isolated
***S. islandicus* NOB8-H2**	**-**	**2.81**	**35.8**	**3159**	**46**	**Hokkaido, Japan**
***S. islandicus* LAL 14/1**	NC_021058	2.47	35.1	2745	45	Iceland
***S. islandicus* REY15A**	NC_017276	2.52	35.30	2780	46	Iceland
***S. islandicus* HVE 10/4**	NC_017275	2.66	35.10	2914	44	Iceland
***S. islandicus* Y.G.57.14**	NC_012622	2.70	35.40	3018	48	Yellowstone, USA
***S. islandicus* M.16.27**	NC_012632.	2.69	35.00	2945	45	Kamchatka, Russia
***S. islandicus* L.S.2.15**	NC_012589	2.74	35.10	3045	45	Lassen, USA
***S. islandicus* Y.N.15.51**	NC_012623	2.85	35.31	3221	46	Yellowstone, USA
***S. islandicus* M.16.4**	NC_012726	2.59	35.00	2841	45	Kamchatka, Russia
***S. solfataricus* P1**	NZ_LT549890	3.03	35.80	3279	45	Naples, Italy
***S. solfataricus* P2**	NC_002754	2.99	35.80	3213	45	Naples, Italy
***S. solfataricus* 98/2**	NC_017274	2.67	35.80	2949	45	Yellowstone, USA
***S. acidocaldarius* SUSAZ**	**-**	2.06	36.3	2228	46	Los Azufres, Mexico
***S. acidocaldarius* DSM 639**	NC_007181	2.23	36.70	2347	48	Yellowstone, USA
***S. acidocaldarius* Ron12/1**	NC_020247	2.22	36.7	2341	30	Ronneburg, Germany
***S. acidocaldarius* N8**	NC_020246	2.18	36.7	2299	48	Hokkaido, Japan
***Sulfolobus sp.* A20**	NZ_CP017006	2.69	34.8	2726	45	Las Pallas, Costa Rica
***Sulfurisphaera. tokodaii str.* 7**	NC_003106	2.69	32.8	2951	46	Kyushu Island, Japan
***Acidianus brierleyi* DSM 1651**	NZ_CP029289	2.95	31.9	3165	46	Yellowstone, USA

a The species *S. solfataricus* has recently been designated as belonging to the genus *Saccharolobus*, not *Sulfolobus* (Sakai & Kurosawa 2018), therefore all *S. solfataricus* strains listed are actually *Saccharolobus solfataricus*.

The *Sulfolobus* NOB8-H2 chromosome is 2.81 Mb in size, which is intermediate in size between the *S. islandicus* and the *S. solfataricus* genomes. The G-C% of 35.8% is slightly greater than that of the *S. islandicus* strains, but exactly the same as the three *S. solfataricus* strains in **Table 5.3**. G-C content is a common genomic metric used when describing different prokaryotic species (Rossellò-Mora & Amann 2001); here this supports the data in **Table 5.2** that NOB8-H2 is most closely related to *S. islandicus* and *S. solfataricus*.

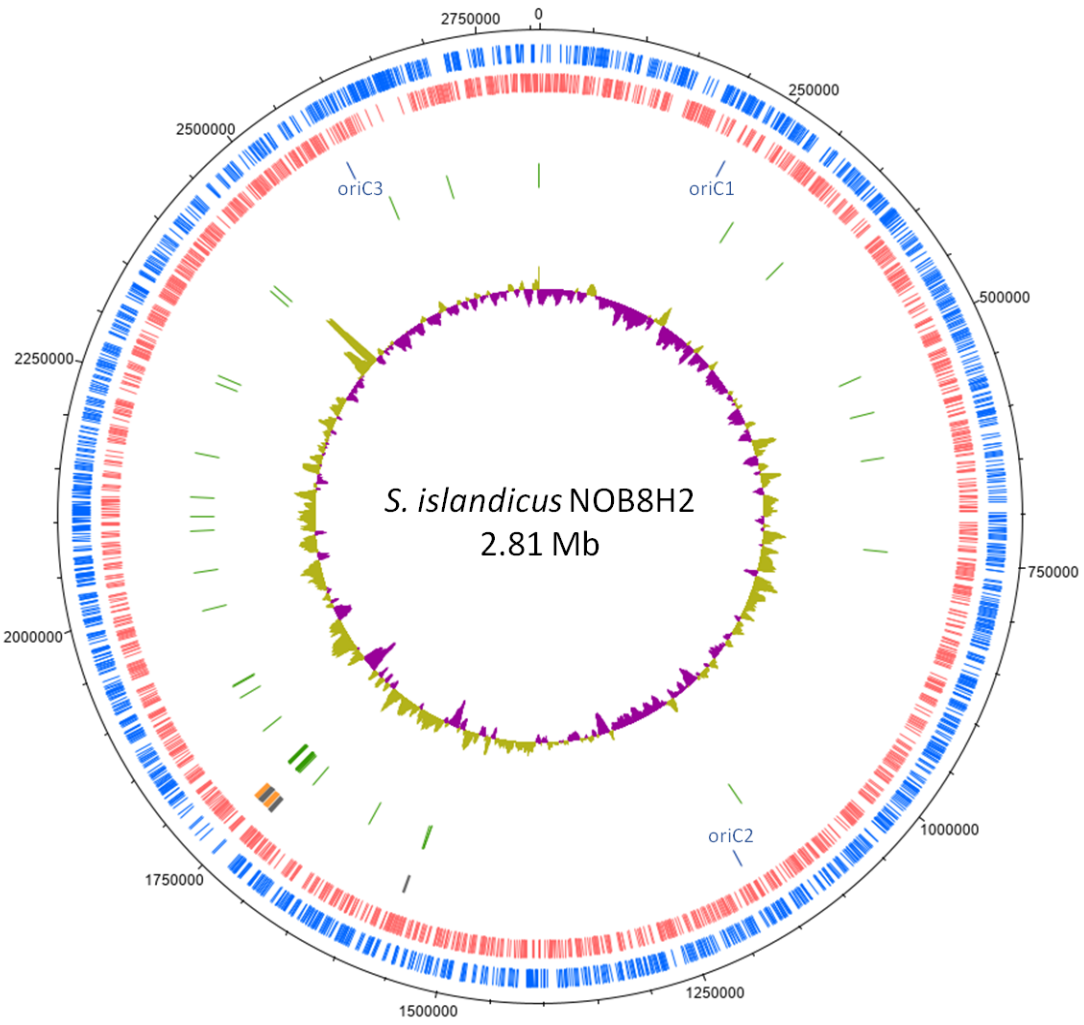


Figure 5.4. *Sulfolobus islandicus* NOB8-H2 chromosome. Circular representation of the chromosome of the newly-characterised strain *S. islandicus* NOB8-H2. From outer to inner: numbers indicate genome coordinates in base pairs, red and blue lines are genes on the forward and reverse strands, orange bars at 1,750,000 bp are two CRISPR repeat sequence arrays with grey bars representing *cas* genes, dark blue lines indicate the three origins of replication (*oriC1-C3*), and green lines are BLAST-derived matches to plasmid pNOB8 sequences. The innermost circle plots G-C content; gold is above average, purple is below average. The figure was generated using DNAPlotter, and BLAST was used to map the pNOB8 insertions in the chromosome.

5.2.4 *Sulfolobus* NOB8-H2 origins of replication

Sulfolobus chromosomes are known to have three origins of replication (Lundgren *et al.* 2004, Ausiannikava & Allers 2017), therefore a search was conducted using the DoriC database (Luo & Gao 2018) to find homologous sequences to known origins, using origins from the closely related strain *S. islandicus* REY15A for comparison. Origins *oriC1* and *oriC2* were found to be 96 and 97% similar to those found in REY15A, and were found to be upstream of the *cdc6-1* and *cdc6-3* genes respectively. However, *oriC3* is known to be found upstream of the gene encoding the replication initiator protein WhiP (Robinson & Bell 2007, Samson 2013), therefore BLAST searches were used to identify a *whiP* homologue in NOB8-H2. Replication origins in archaea are typically AT-rich (Wu 2014), and the 84 bp intergenic region upstream of the *whiP* homologue was found to have an AT composition of 76%, and is therefore considered as a putative *oriC3*. Origin regions have been found to contain repeat sequence elements to which the Cdc6 proteins bind: *oriC1* of *S. solfataricus* P2 contains three sets of a 36 bp element dubbed ORBs (Origin Recognition Boxes), with Cdc6-1 binding demonstrated by DNase I footprinting (Robinson *et al.* 2004). Here, the *oriC1* of NOB8-H2 was assessed for any putative ORBs. Interestingly, two elements of 19 and 23 bp aligned with the *S. solfataricus* P2 ORB1, with only two mismatches, and one elements is inverted with respect to the other, as also seen in the *S. solfataricus* ORBs. Therefore these sequence elements represent putative Cdc6-1 binding sites within the NOB8-H2 *oriC1.*

5.2.5 *Sulfolobus* NOB8-H2 phylogenetic analysis

The 16S rRNA gene is still often used as an initial starting point when building phylogenetic trees of bacteria and archaea (Rinke *et al*. 2013, Dai *et al.* 2016), its usefulness as an evolutionary genetic marker being established by Woese and Fox in the 1970s (Woese & Fox 1977). An initial BLAST search of the NOB8-H2 16s rRNA gene against the nucleotide database showed that it shared the greatest sequence identity with *S. solfataricus* at 99.67%, slightly greater than that with *S. islandicus* strains (99.60%).

The program MEGA-X (Molecular Evolutionary Genetics Analysis, Hall 2013) was used to construct a phylogenetic tree based on the 16s rRNA gene, using sequences from most of the Sulfolobaceae members listed in **Table 5.3,** along with *S. tengchongensis*, *S. shibitae*, and *S. metallicus* strains, and *Metallosphaera sedula*, a slightly more distantly related species within the same family that could act as the outgroup to root the tree.

Homologous sequences were aligned using the MUSCLE alignment algorithm, as generally this gives better alignments (Hall 2013). A Maximum Likelihood (ML) tree was constructed, as in general, ML is a more robust method compared with other commonly used approaches like Maximum Parsimony (MP) and Neighbour Joining (NJ) (Ogden 2006). However, MP and NJ trees were also derived, and gave similar topologies (data not shown). The 16s rRNA ML tree shows that NOB8-H2 is most closely-related to *S. solfataricus* 98/2, and that *S. solfataricus* and *S. islandicus* are more closely related to each other than to *S. acidocaldarius*. **(Figure 5.5).**

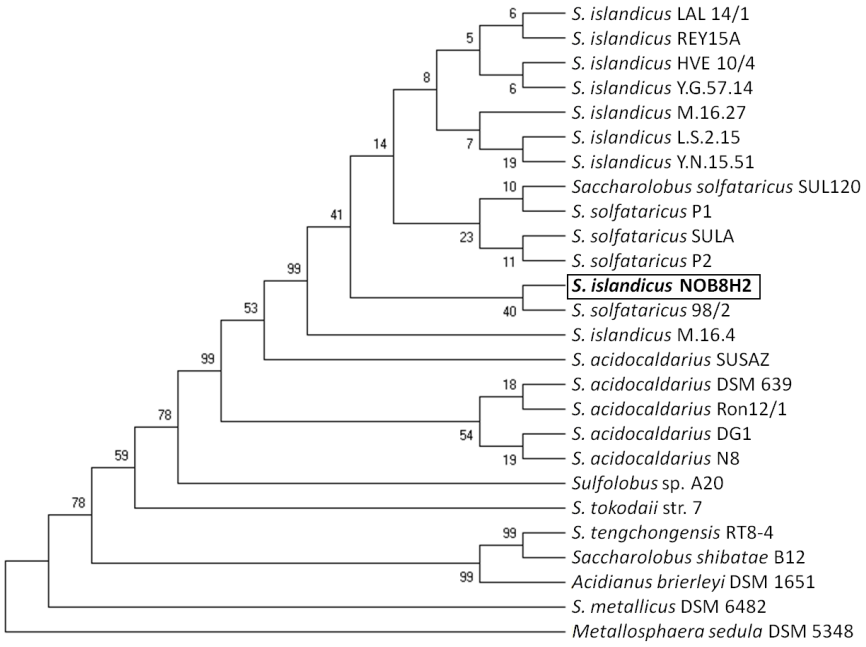


Figure 5.5. Maximum Likelihood phylogenetic tree of *Sulfolobus* 16S rRNA genes. The tree was rooted using *M. sedula* as the outgroup. Numbers at branch nodes indicate bootstrap percentages using 1000 bootstrap replicates.

Given that *Sulfolobus* NOB8-H2 appears more similar to *S. islandicus* when comparing a larger gene set **(Table 5.2)**, a more robust phylogeny would require larger number of genes. To do this, firstly a core genome set for the genus *Sulfolobus* was derived, using the pan/core genome tool available from the online Genoscope server (see Methods 2.7.2.2). The core genome of four representative *Sulfolobus* species: *Sulfolobus solfataricus* 98/2, *Sulfolobus islandicus* M.16.27, *Sulfolobus tokodaii* 7, *Sulfolobus acidocaldarius* DSM 639 plus *M. sedula* DSM 5348, comprises 96 genes **(Figure 5.6)**.

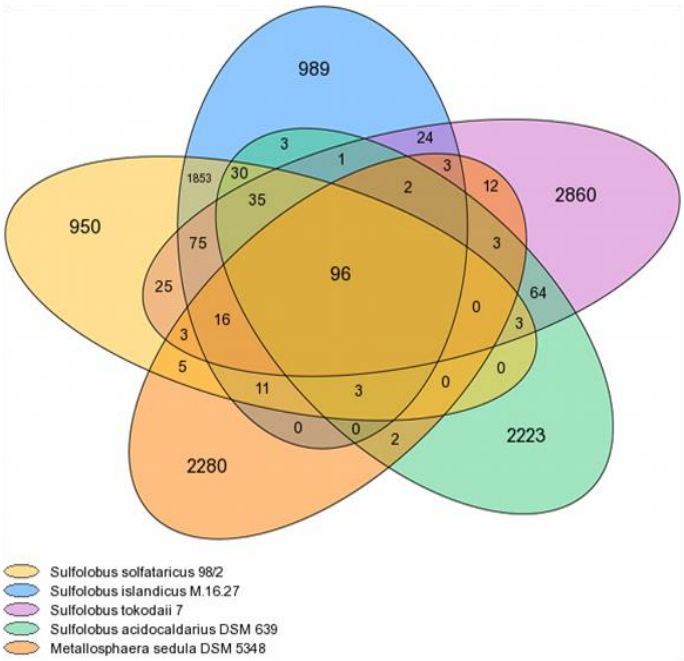


Figure 5.6. Venn diagram of core genes of the genus *Sulfolobus*. The pan and core genomes of five species of the genus *Sulfolobus* were derived using the comparative genomics serve at Genoscope.cns.fr.

From this core genome, a subset of ten genes were chosen at random to construct the multi-gene tree (Methods 2.7.2.2). Here, a trade-off between number of genes and processing time was reached; increasing the number of genes used should produce a more accurate phylogeny, but comes with a concomitant increase in resource requirement. Using ten concatenated gene sequences has been demonstrated to give a

robust degree of accuracy of >95% (Gadagkar 2005), therefore the sequences of the ten genes were found using BLAST, for the 23 strains used to construct the tree. The sequences were concatenated into a single file in the same order to preserve synteny, and a ML tree constructed using MEGA-X as before. The multi-gene tree is shown below in **(Figure 5.7)**. In this tree, *Sulfolobus* NOB8-H2 appears more closely related to the *S. islandicus* strains, forming a distinct group separate to *S. solfataricus*.

Thus, the multi-gene phylogeny supports the MASH data that NOB8-H2 is likely to be a novel strain of *S. islandicus*.

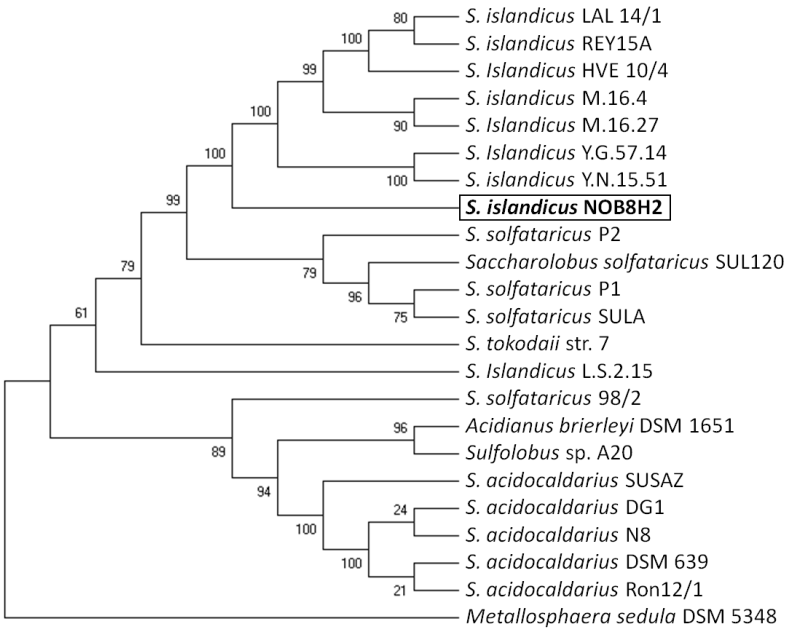


Figure 5.7. Maximum Likelihood phylogenetic tree of ten concatenated *Sulfolobus* genes. The genes used for the concatenated tree are listed in Table 2.21, Materials and Methods. The tree was rooted using *M. sedula* as the outgroup. Numbers at branch nodes indicate bootstrap percentages using 1000 bootstrap replicates.

5.2.6 Whole-genome comparison of *Sulfolobus* NOB8-H2 to other *Sulfolobus* spp.

So far, genome comparisons between *Sulfolobus* NOB8-H2 and other closely related strains have been conducted by either comparing sections of the genome (shared hashes), or by constructing single-gene and multi-gene phylogenetic trees. To gain more insight into the similarities of the NOB8-H2 genome to these other strains at a whole-genome level, *in silico* DNA-DNA hybridisation (DDH) was employed. DDH has previously been utilised to delineate prokaryotic species, i.e. whether a newly-identified bacterial or archaeal strain could in fact be classified as a species (Auch *et al.* 2010b). It has been suggested that the 16s rRNA gene sequence alone can be used to define a species, with a threshold value of 97% proposed (Tindall *et al*. 2010). Here, NOB8-H2 has greater than 99% identity at the level of the 16s rRNA sequence, however it is recommended in this case that other methods, e.g. DDH are also used, and that a DDH value of 70% represents a threshold for species definition (Auch *et al.* 2010a, Tindall *et al*. 2010). DDH was originally a wet-lab technique, but now, with whole-genome sequencing being commonplace, deriving DDH values *in silico* is possible using the online Genome-to-Genome Distance Calculator (GGDC, Meier-Kolthoff *et al*. 2013).

Initially, the genome sequences of *Sulfolobus* NOB8-H2 and selected other Sulfolobaceae strains were submitted to the GGDC, and the resulting DDH matrix is shown in **Table 5.4**. The DDH values for NOB8-H2 appear to support the multi-gene phylogeny in **Figure 5.7.** The highest values are seen in comparison to *S. islandicus*, then *S. solfataricus*, then *S. acidocaldarius* and other species, reflecting the increasing evolutionary distance between these organisms. Interestingly, the DDH values were below 70% in all cases, though only slightly below in comparison to the two *S. islandicus* strains REY15A and YG.57.14 (65.40% and 64.20% respectively).

Table 5.4. DNA-DNA hybridisation (DDH) matrix of selected Sulfolobaceae members[a]

Strain	NOB8-H2	*S. islandicus*		*S. solfataricus*		*S. acidocaldarius*		*S. toko.*	*S.* A20	*A. bri.*
		REY15A	YG.57.14	P1	P2	DSM 639	N8	Str. 7	A20	DSM 1651
NOB8-H2	—	65.40	64.20	36.10	37.10	19.50	23.10	21.10	16.60	20.20
REY15A		—	86.60	37.50	39.30	18.10	18.10	21.10	16.70	26.40
YG.57.14			—	38.80	40.60	18.10	18.10	21.70	16.60	31.40
P1				—	94.80	18.80	18.80	24.0	16.80	22.30
P2					—	18.20	18.20	23.0	16.80	22.20
DSM 639						—	100.00	15.70	23.10	21.30
N8							—	15.70	20.20	21.30
Str. 7								—	19.60	18.10
Sp. A20									—	16.00
DSM 1651										—

DDH values (%) were obtained by uploading genomes to the Genome-to-Genome distance calculator at http://ggdc.dsmz.de/ggdc.php

[a] *S. toko., S. tokodaii; S.* A20, *Sulfolobus* sp. A20; *A. bri., Acidianus brierleyi*.

Because of these DDH values being less than, but close to 70%, the analysis was repeated for NOB8-H2 against all *S. islandicus* strains, to assess the DDH values in relation to the threshold figure **(Table 5.5)**. All DDH values were <70%, but again were quite close to this figure, ranging from 62.90% to 65.40%. The recently characterised *Sulfolobus* sp. A20 was posited to be a novel species rather than strain, in part based on DDH values of <30% (Dai *et al.* 2016). Here, the DDH value being so close to the threshold of 70% means that we cannot confidently assert that *Sulfolobus* NOB82 represents a novel species of *Sulfolobus* rather than a new strain of *S. islandicus*; to do this would require more thorough taxonomic investigations that are beyond the scope of this thesis.

Table 5.5. DNA-DNA hybridisation percentage values of NOB8-H2 against *S. islandicus* strains

Strain	NOB8-H2	*S. islandicus*								
		REY15A	Y.G.57.14	HVE 10/4	L.S.2.15	LAL 14/1	M.14.25	M.16.4	M.16.27	Y.N.15.51
NOB8-H2	—	65.40	64.20	65.00	63.80	65.40	65.20	65.00	64.70	62.90

The previous comparative genomics techniques give an indication of the amount of similarity the *Sulfolobus* NOB8-H2 genome shares with other strains, but this does not inform us of any particular areas of the genome which may differ significantly. To do this, a more visual approach is required, therefore BRIG (Blast Ring Image Generator) was used to depict whole-genome similarities and differences. BRIG identifies the similarities between a reference sequence, here *Sulfolobus* NOB8-H2, and other chosen sequences,

using defined BLAST percentage identities, with each genome in the output displayed as a series of concentric rings (Alikhan *et al*. 2011). Here, nine other Sulfolobaceae genomes were compared with the NOB8-H2 genome, using BLAST identity values of 100, 90 and 70% for each genome (see Materials and Methods 2.7.3), and the result is seen in **Figure 5.8**. The NOB8-H2 genome appears very similar overall compared to the *S. islandicus* and *S. solfataricus* strains, with many regions appearing at least 90% similar, supporting the analyses already conducted. There are some areas of difference, and this can be seen at around the 1.9 Mb point, where there is a gap in the *S. islandicus* rings, indicating a region of non-homology. This could be explained by the fact that the NOB8-H2 genome is ~0.3 Mb larger than that of *S. islandicus* REY15A.

The MAUVE genome alignment viewer was used to compare these two genomes and study this region further (Darling *et al*. 2007). MAUVE allows genome rearrangements such as inversions, duplications, and large-scale reordering to be observed. The NOB8-H2 and *S. islandicus* REY15A genomes show a highly-similar order, with no significant rearrangements, except for a few regions that are inverted (reverse-complement) **(Figure 5.9)**. The MAUVE software performs local multiple alignments to identify highly similar regions across genomes, then groups homologous DNA regions that show no internal sequence rearrangements as identically coloured segments called local colinear blocks (LCBs, **Figure 5.9A&B**). The region of NOB8-H2 between ca. 1.85 Mb and 2 Mb does not show homology to *S. islandicus* REY15A **(Figure 5.9C)**.

This region contains ~180 genes, of which around 30 code for transposases, mainly clustered at the left and right edges of this region, indicating multiple transposition events that have increased the size of the genome. Interestingly, this portion of the genome is 90-100% homologous to that of *S. solfataricus* P1 (**Figure 5.8)**, perhaps indicating that either a shared ancestral region has been lost from other *S. islandicus* strains, or conversely, incorporated into the NOB8-H2 and P1 chromosomes by large-scale transpositional gene transfer.

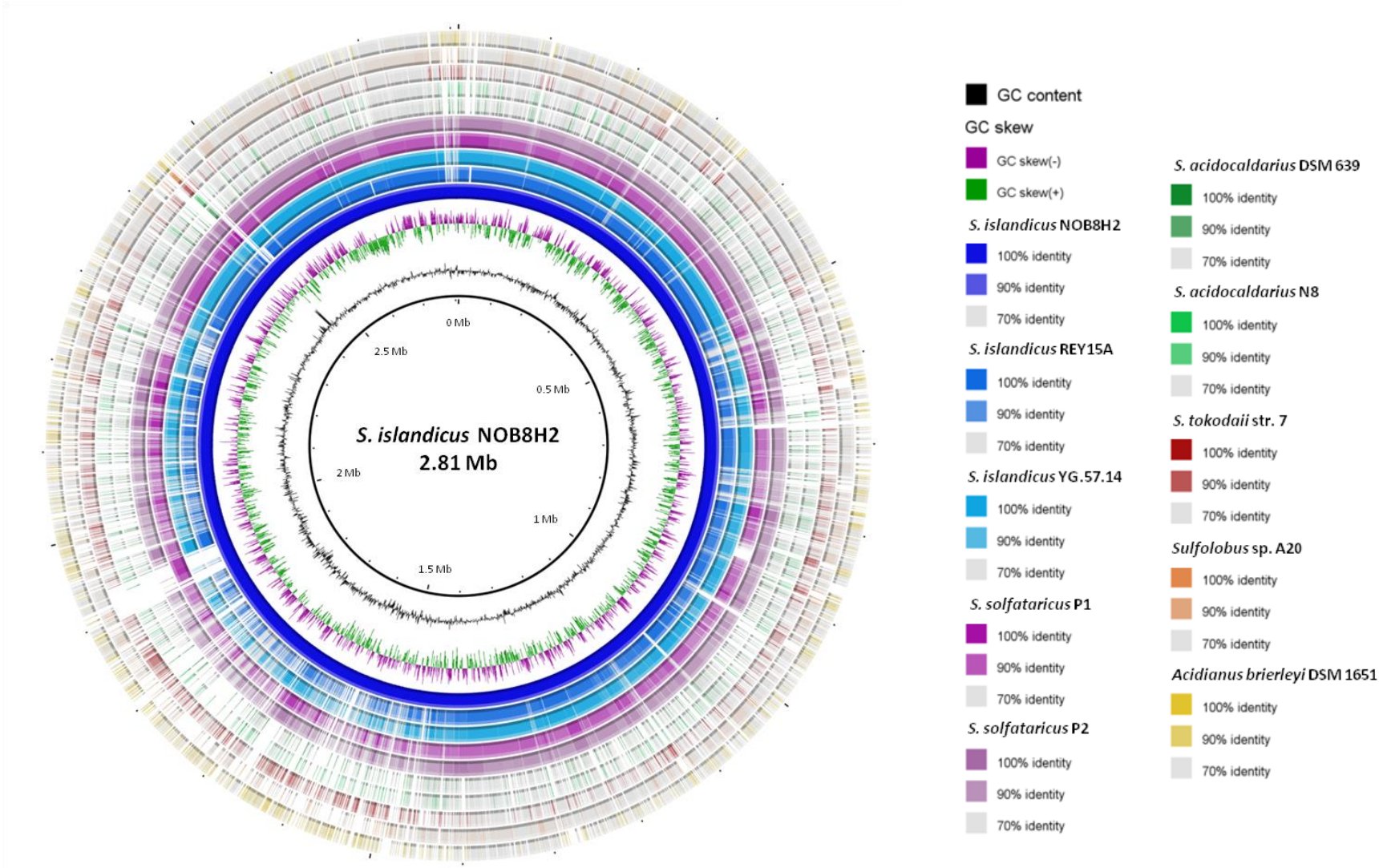


Figure 5.8. Whole genome comparison of *S. islandicus* NOB8-H2 and other complete Sulfolobaceae genomes. Blast Ring Image Generator (BRIG) was used to visualise the genomes. The inner black circle shows the coordinates of the circular NOB8-H2 chromosome, then (inner to outer) G+C%, GC skew, NOB8-H2 chromosome (dark blue). The remaining rings show homology of NOB8-H2 to other Sulfolobaceae complete genomes, based on BLASTn nucleotide sequence identity. Regions of 100%, 90% and 70% homology are indicated by different shadings (see legend). Genomes are ordered from most closely related to least closely related (inner to outer), based on the phylogenetic relationships represented in **Figure 5.6**.

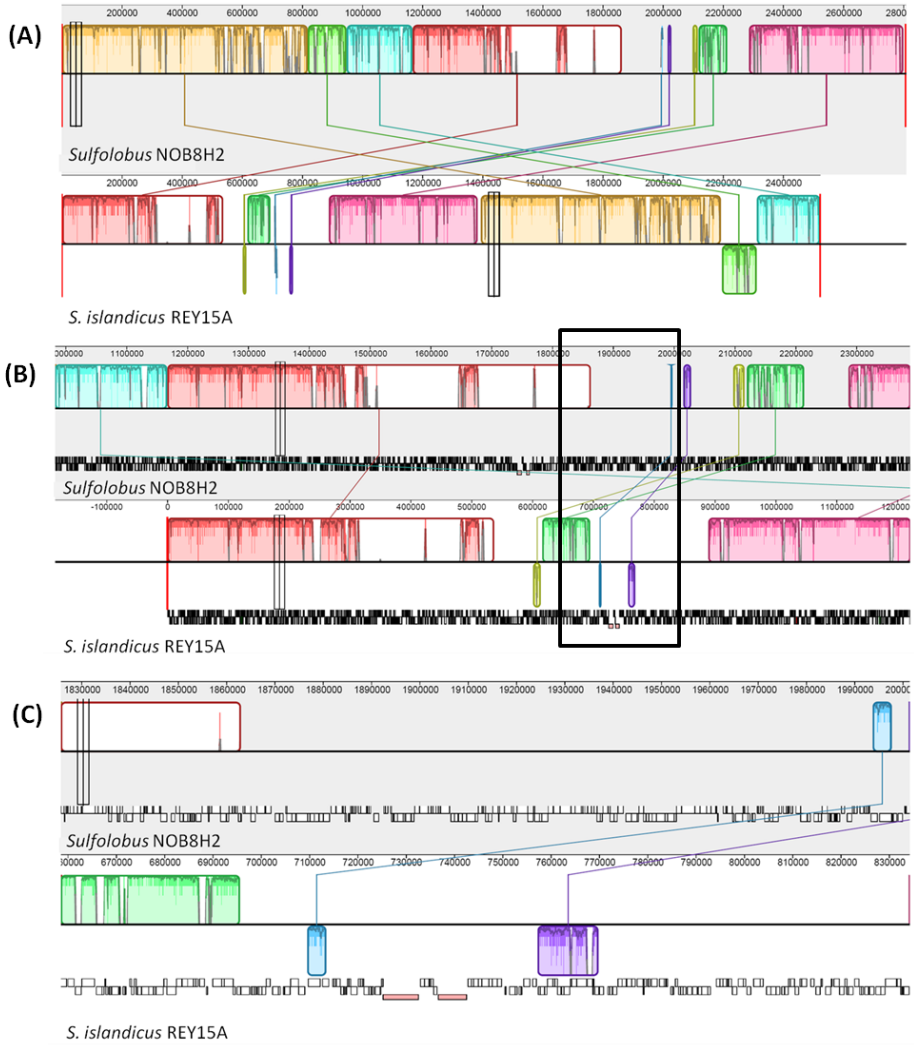


Figure 5.9. Visualisation of genome alignments using MAUVE. The genomes of *Sulfolobus* NOB8-H2 and *S. islandicus* REY15A were aligned using MAUVE. Boxes of identical colour, and connected by lines are local colinear blocks (LCBs), homologous portions of the genomes that align without internal rearrangement, based on a minimum weighting criterion. NOB8-H2 has been set as the reference sequence. LCBs below the black line indicate a reverse complement orientation of that block. The long red vertical lines indicates the start and end points of the chromosome. **(A)** The ordering of the LCBs is consistent between the two strains. **(B)** The red LBCs have now been aligned, and the region from 1.85 Mb to 2 Mb is shown by the black box. The black and white tracks below each genome represent individual genes. **(C)** A close-up view of the region highlighted in **(B)**, showing a region of non-homology between the two strains.

5.2.7 *Sulfolobus* NOB8-H2 genome COG analysis

To further characterise the genome of a newly-sequenced organism, protein-coding genes can be assigned to various classes, or groups, using the Clusters of Orthologous Genes (COG) database (Tatsuov *et al* 2000). This method groups coding sequences (CDS) into COG groups based on orthology, i.e. genes that share a common evolutionary origin, allowing functional characterisation of the translated proteins to be inferred by comparison to characterised orthologues. The database has grown since its inception, expanding the number of COG groups from 2,091 to over 5,000 at the time of writing. COG groups have been constructed using complete genomes from bacteria, archaea and eukaryotes, with an archaeal-specific database (arCOGs) also available (Makarova *et al.* 2015a). There are currently 21 Crenarchaeota genomes in the COG database as of the 2014 update (Galperin *et al.* 2014). The COG groups are further delineated into classes; 26 alphabetised functional categories, e.g. (D) - Cell cycle, cell division and chromosome partitioning, (L) - Replication, recombination and repair. A few classes (B, W, Y and Z) relate to functions that are chiefly found in eukaryotes, and there are two classes where function is unknown: R (generic functional prediction), and S (uncharacterised genes), giving a sense of the current degree of knowledge of protein activities (Galperin *et al*. 2019). Grouping genes into COGs therefore gives a broad-scale assessment of the proportion of the genome dedicated to cellular processes, information storage, metabolism, and the number of genes which remain poorly characterised.

The COG database and associated required software (COGsoft) is available at the NCBI website, however using the software requires knowledge of UNIX and/or database creation, therefore an alternative approach was sought. The web server WebMGA, which includes over 20 tools developed for metagenomic analysis was used, as COG functional annotation is provided (Wu *et al*. 2011). The NOB8-H2 annotation was first converted to a protein FASTA file, then uploaded to the WebMGA server (see Methods 2.7.4). The results obtained from WebMGA showed that out of the 2,339 annotated CDS, 2,271 (97%) were assigned to at least one COG group (proteins can be assigned to multiple COGs). 264 CDS were assigned to cellular processes and signalling, 608 CDS to information storage and processing, 1,057 to metabolism, with 595 CDS assigned as 'poorly characterised' classes R and S **(Figure 5.10)**.

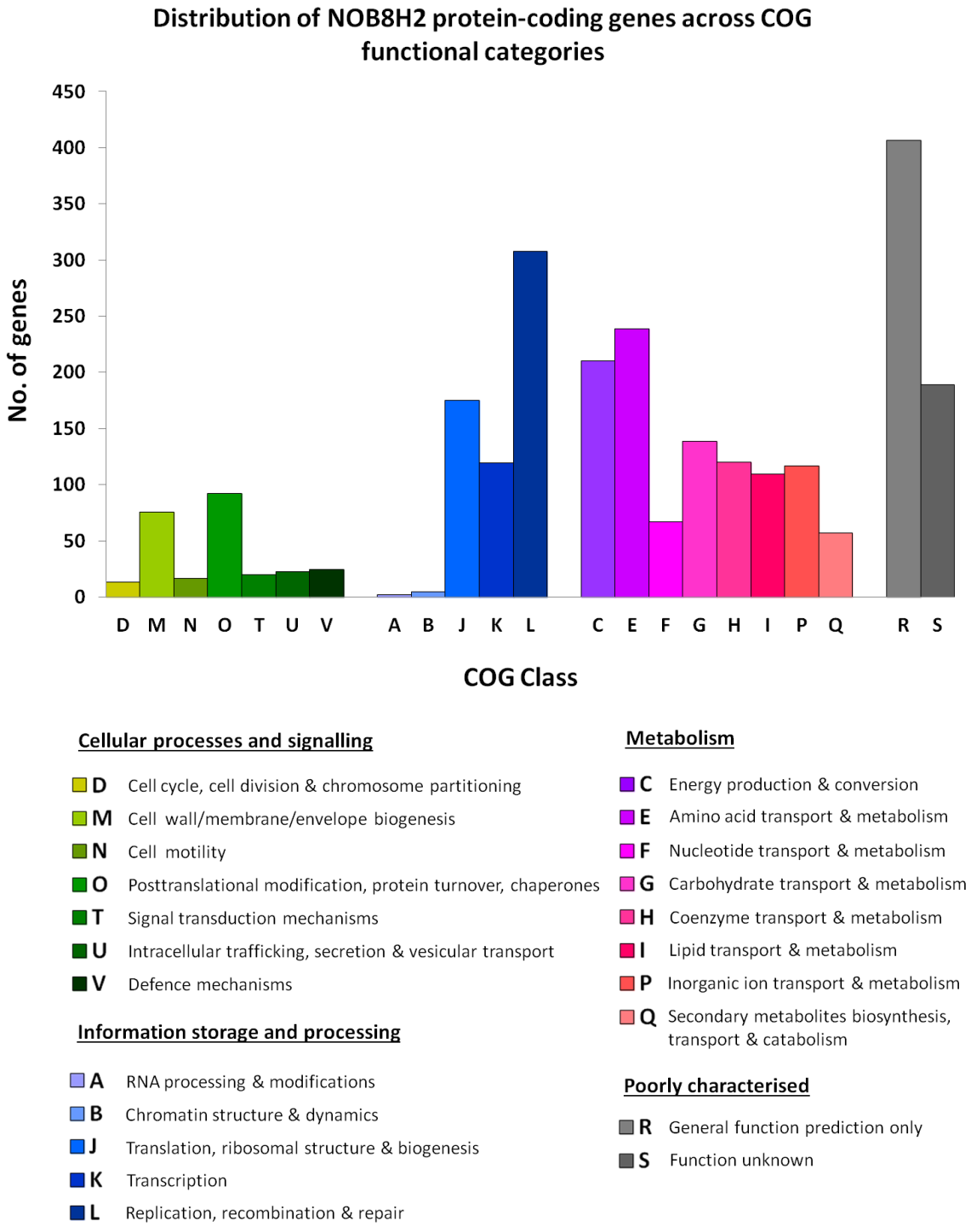


Figure 5.10. Functional classification of *Sulfolobus* NOB8-H2 protein-coding genes. The distribution of *Sulfolobus* NOB8-H2 protein coding sequences based on predicted functional COG (Cluster of Orthologous Genes) classes. COG classes W (Extracellular structures), Y (Nuclear structure) and Z (Cytoskeleton) returned zero hits and are not plotted. The recently added class X (Mobilome) did not appear in the data and is not shown.

COG classes W, Y and Z, representing extracellular structures, nuclear structure and cytoskeleton respectively, which are primarily eukaryotic features, returned zero hits and are therefore not included in **Figure 5.10**. Interestingly, 24 genes are assigned to the recently-added class V (defence mechanisms), and further study of these genes could give an insight into some of the strategies employed by *Sulfolobus* NOB8-H2 when dealing with invasive genetic elements (see CRISPR-Cas section). Unfortunately, another recently added class, X (mobilome), did not appear in the WebMGA output. This mobilome COG class has been shown to comprise genes primarily involved in horizontal gene transfer via mobile genetic elements (transposons, phages and plasmids, Nakamura 2018). In the case of *Sulfolobus* NOB8-H2, this could have provided useful information about genes that are potentially involved in dynamic interactions with pNOB8, therefore it would be interesting to repeat the COG analysis in future studies and further analyse any mobilome class genes.

The number of CDS that are poorly characterised may seem to represent a large percentage of the total, however this is seen elsewhere in prokaryotes. A COG analysis of the genome of the novel bacterial strain *Casimicrobium* SJ-1 assigned 23% of CDS to the poorly characterised classes R and S (Song *et al.* 2020), with a similar percentage seen in an analysis of the bacterium *Bacillus amyloliquefaciens* (Niazi, *et al*. 2014). Excepting classes R and S, perhaps unsurprisingly, the three most numerous COG classes are L, C and E, all of which relate to essential biological functions: DNA replication and repair, energy production, and amino acid transport and metabolism respectively. Genes assigned to metabolism-related COG classes account for the largest proportion (42%) of total genes, a very similar proportion to that of the bacterial strains mentioned above. However, there are a greater proportion of genes involved in information storage and processing in *Sulfolobus* NOB8-H2 (24%) compared with *Casimicrobium* (16%) and *B. amyloliquefaciens* (~15%), perhaps due to the increased complexity of archaeal processes such as transcription when compared to those in bacteria (Kramm *et al.* 2020), being more eukaryotic-like and reflecting the position of archaea in the phylogenetic tree of life**.**

5.2.8 The *Sulfolobus* NOB8-H2 CRISPR-Cas systems

CRISPR-Cas (Clustered Regularly Interspaced Short Palindromic Repeats and CRISPR-associated proteins) systems are defence mechanisms found in 50% bacteria and 80% of archaea (Makarova *et al*. 2015b), and act to protect the host cell from invasive viruses (phage), and other mobile genetic elements such as plasmids (Gudbersdottir *et al.* 2011, Lillestøl *et al*. 2006). An in-depth discussion of the mechanism of CRISPR-Cas adaptive immunity will not be presented here, however the basic action can be summarised as occurring in three stages: 1) adaptation - in which short segments of viral/plasmid DNA is incorporated into the CRISPR array as repeat-spacer units; 2) expression - where this section of repeat-spacers is transcribed into pre-crRNA, then processed into shorter crRNAs; and 3) interference - where the crRNA, complexed with an effector Cas enzyme, targets and cleaves the invading nucleic acid via complementary base-pairing to the crRNA spacers (Makarova *et al*. 2015b).

A number of different CRISPR-Cas systems have been identified, divided into classes, and further subdivided into types, as outlined in abbreviated form in **Table 5.6**. CRISPR-Cas systems have previously been reported in *Sulfolobus* (Deng *et al*. 2013, Zebec *et al*. 2014, Peng *et al*. 2015), therefore it was likely that *Sulfolobus* NOB8-H2 would also contain one or more CRISPR loci. Here the genome browser and annotation software Artemis was used to manually curate the NOB8-H2 genbank annotation, searching for homologues of *cas* genes from the closely-related *S. islandicus* L.S.2.15. Three distinct Cas modules were found; two of Type I-A, and one of Type III-B, along with two CRISPR arrays that contain 24 bp repeats, separated by putative spacer sequences. Where the annotation did not label a particular gene within the CRISPR locus, BLAST was used to identify *cas* homologues, allowing the full Cas module to be annotated. The *Sulfolobus* NOB8-H2 CRISPR-Cas system is shown below (**Figure 5.11** and **Figure 5.12)**.

Table 5.6. Simplified overview of CRISPR-Cas classes and types.

Class	Type	Signature protein	Target nucleic acid	Reference
I	I	Cas3	DNA	Peng *et al*. 2013
	III	Cas10	DNA, RNA	Peng *et al*. 2015
	IV	Csf1	Unknown (likely plasmid targeting)	Pinilla-Redondo *et al*. 2020a
II	II	Cas9	DNA, RNA	Strutt *et al*. 2018
	V	Cas12	DNA, RNA	Yan *et al*. 2019
	VI	Cas13	RNA	O'Connell 2018

The three CRISPR-Cas systems here are almost identical to those found in *S. islandicus* REY-15A (Peng 2013), only differing in the length of the repeat sequence (24 bp *cf.* 23 bp in *S. islandicus* REY-15A), and the presence of a number of transposase genes within the *cas* modules.

It is unsurprising to find two distinct CRISPR-Cas systems within the *Sulfolobus* NOB8-H2 genome. Out of 17 crenarchaeal genomes analysed, 15 were found to harbour type I systems, while 16 contained type III systems, suggesting the vast majority of crenarchaea encode both systems (Makarova *et al.* 2011). Here, the Cascis module **(Figure 5.11, blue)** is involved in the initial CRISPR adaptation stage; namely the acquisition of spacers to incorporate into the CRISPR arrays (Cascis - CRISPR-associated cluster for integration of new spacers), and this is mediated by the Csa1 and Cas4 proteins forming a complex with Cas1 and Cas2 (Peng *et al*. 2013). The interference stage, where the invading nucleic acid is targeted and cleaved, is performed by the Cas3 protein. Cas3 encodes a helicase, that may either be fused to a HD-family nuclease domain, or is adjacent to a separate nuclease-encoding gene (Makarova *et al*. 2015b). Somewhat confusingly, the same *cas* genes often have multiple names (see Table 2 in Makarova *et al.* 2011). The *cas3* gene may be denoted as *cas3'*, with the adjacent nuclease encoded by *cas3''*, and this arrangement is seen in the *Sulfolobus* NOB8-H2 CRISPR system **(Figure 5.11, red)** suggesting that the Type I-A systems is functional in this strain.

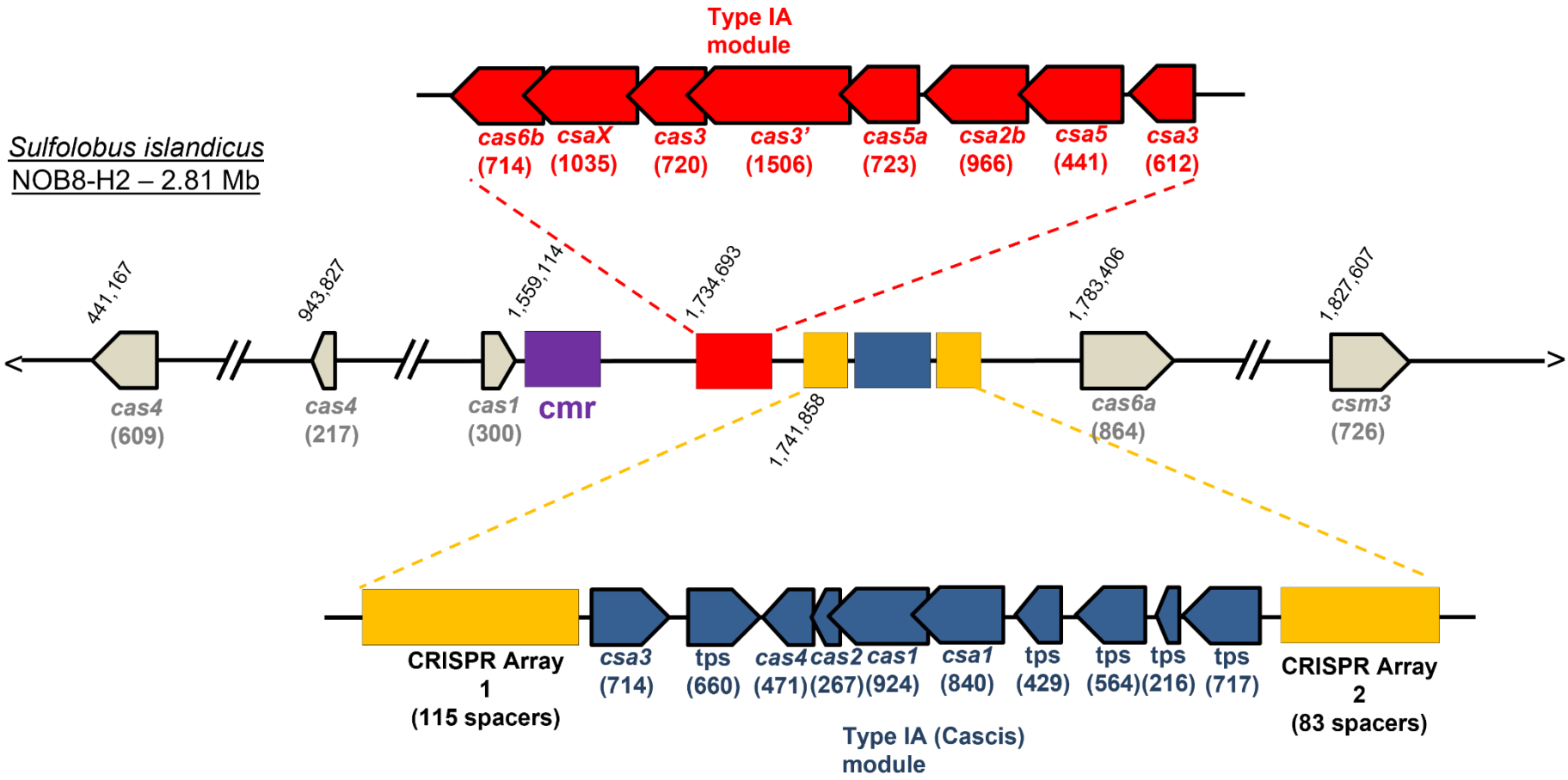


Figure 5.11. *S. islandicus* NOB8-H2 CRISPR-Cas system. Schematic of the CRISPR arrays and Cas modules found in strain NOB8-H2. tps = transposase, number in parentheses indicates gene length in bp. The Type IA interference module is shown in red, and the spacer acquisition module (Cascis) in blue. The yellow boxes are CRISPR arrays that comprise repeat-spacers. The cmr locus (purple) represents a Type IIIB system and is shown in more detail in **Figure 5.12.** Other *cas* genes lying outside of the main CRISPR locus are shown in grey. The number above each genes/module indicates the genome position. Genes drawn to scale.

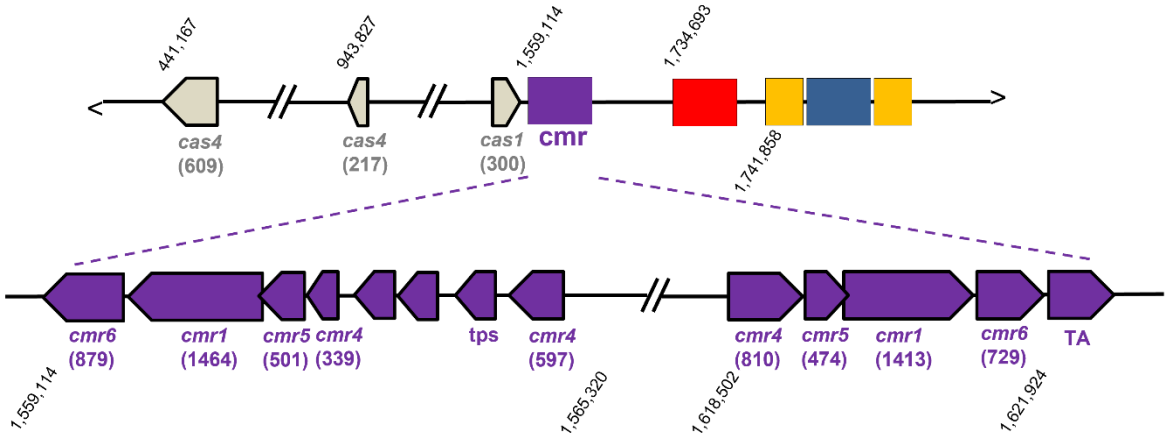


Figure 5.12. *S. islandicus* NOB8-H2 Type III-B (cmr) module. Abbreviations and numbers, along with red, yellow and blue regions are the same as in **Figure 5.11**. The Type IIIB interference locus (cmr) is depicted in purple. Other *cas* genes lying outside of the main CRISPR locus are shown in grey. TA = putative Toxin/Antitoxin gene.

In addition to the Type I-A module, there is also a Type III-B (Cmr) module. Type III CRISPR systems in *Sulfolobus islandicus* exhibit target cleavage against both DNA and RNA substrates (Peng *et al.* 2015, Li *et al.* 2016). In Type III systems, the Cas10 protein combines with other adjacently encoded proteins to form an interference complex. Similar to Cas3, Cas10 is often fused to a HD nuclease domain (Makarova *et al*. 2015b), and is therefore required for DNA cleavage. Here, it appears that the Type III-B Cmr module in *Sulfolobus* NOB8-H2 lacks the *cas10* gene (also called *cmr2*, **Figure 5.12**). As the *Sulfolobus* NOB8-H2 CRISPR system is very similar to that of *S. islandicus* REY15A, the synteny of the Type III-B loci was compared. *S. islandicus* REY15A actually contains two Type III-B Cmr modules, denoted α and β **(Figure 5.13),** which are located upstream and downstream from the Type I-A locus. Both Cmr-α and Cmr-β demonstrated RNA interference activities *in vivo*, whilst displaying distinct mechanistic features and cleavage strength (Peng *et al.* 2015). Comparing the organisation of the Cmr cassettes in *Sulfolobus* NOB8-H2 and REY15A, it appears that the *cmr2*/*cas10* gene has been lost, as otherwise the ordering of *cmr4-6* remains the same **(Figure 5.12)**, raising the possibility that RNA interference is not possible in *Sulfolobus* NOB8-H2 due to the incomplete locus and loss of the effector nuclease Cas10.

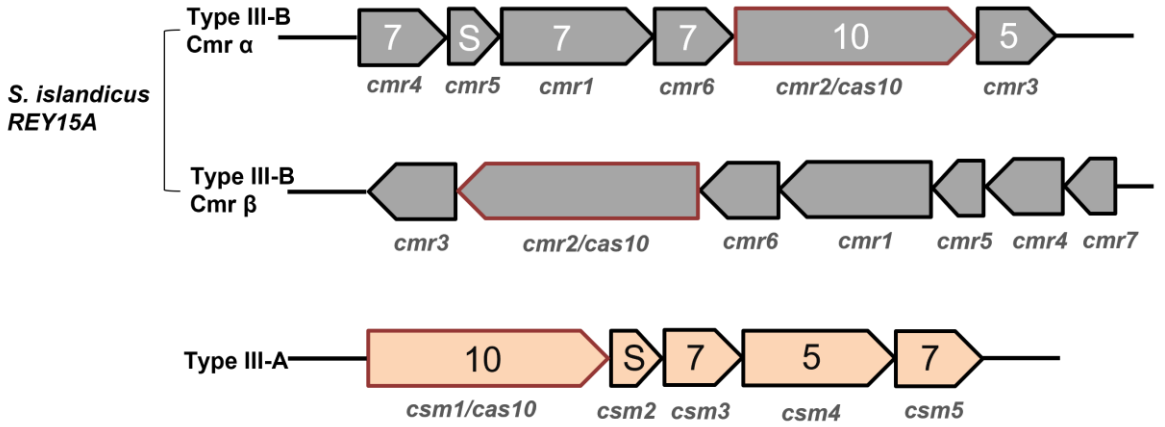


Figure 5.13. *S. islandicus* REY15A Type III-B and typical Sulfolobales Type III-A CRISPR modules. Gene names are shown in gray below genes; *cmr* = Type III-B, *csm* = Type III-A. Numbers/letters inside genes indicate the Cas protein name e.g. *cmr6* and *csm3* both encode Cas7 proteins. The signature *cas10* gene containing a nuclease domain is highlighted in red. Figure adapted from Liu *et al*. 2016, Garrett *et al*. 2017.

It is worth noting that the Type III-A systems (denoted as Csm), though predominantly studied in bacteria so far, have a very similar genetic arrangement to Type III-B systems, also encoding the Cas10 protein **(Figure 5.13)**. In the bacterium *Staphylococcus epidermis*, foreign DNA is cleaved by Cas10, whilst Csm3 subunits destroy the transcribed RNA (Chou-Zheng & Hatoum-Aslan 2019). It is possible that the lack of Cas10 in *Sulfolobus* NOB8-H2 may not affect its ability to degrade RNA, as this function could be carried out by proteins encoded by the adjacent *cmr* genes, or alternatively, Cas10 (or a functional analogue) may be encoded elsewhere on the chromosome. However, it may be that a complete complex of Cas10 plus the accessory subunit proteins is required for both DNA and RNA cleavage, and therefore the nucleic acid targeting activities of *Sulfolobus* NOB8-H2 would need to be experimentally verified.

5.2.9 Analysis of the *Sulfolobus* NOB8-H2 CRISPR-Cas spacers

The first stage of the CRISPR-Cas response is termed the adaptation phase, when sections of foreign invading DNA are incorporated into the host genome as sequences called spacers, thereby providing the organism with a 'memory' of past encounters, and enabling future invasions to be countered (Makarova *et al*. 2015b). To gain insight into the types of viruses and plasmids that *Sulfolobus* NOB8-H2 had previously encountered, an analysis of the source of the CRISPR spacers was conducted. The *Sulfolobus* NOB8-H2 genome contains two CRISPR arrays, regions in which spacers from the invading element are incorporated between repeat sequences **(Figure 5.11, yellow boxes)**. The two arrays contain 115 spacers (Array 1) and 83 spacers (Array 2), each separated by a direct repeat sequence of 24 nt. CRISPR spacers are known to vary in size, ranging from 21-72 nt, though more usually they are 32-38 nt (Barrangou and Marraffini 2014). CRISPR spacers in the Type I-B system found in the euryarchaeon *Haloarcula hispanica* were mainly 35 or 36 nt in length (Li 2017). Here, the *Sulfolobus* NOB8-H2 spacers are predominantly between 38 and 42 nt in length, with one large spacer in Array 1 of 104 nt, and another in Array 2 of 105 nt.

To obtain information on the origin of the spacers, a list of known viruses and plasmids that interact with *Sulfolobus* and the wider crenarchaea was sought (see Li 2015). Ten virus families were represented: the Fuselloviridae*,* Bicaudaviridae*,* Rudiviridae*,* Lipothrixviridae*,* Globuloviridae*,* Ampullaviridae, Guttaviridae*,* Claraviridae, the Monocaudaviruses, and the *Sulfolobus* turreted icosahedral viruses (proposed Turriviridae). 24 plasmids, both cryptic, (pRN type), and the conjugative (pNOB8 type) were also represented. In total, accession numbers of 53 viruses and plasmids were used in a BLAST search against the CRISPR spacers, and significant matches to virus/plasmid types were assigned based on bit score and e-value (see Materials and Methods 2.7.5). A list of the 53 viruses and plasmids used in this analysis can be found in **Appendix 4**.

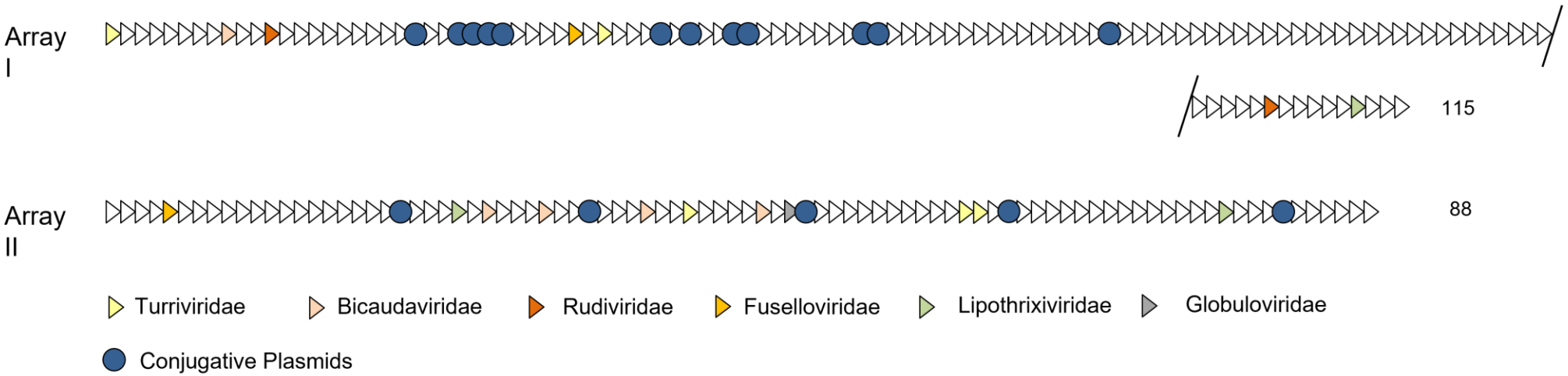


Figure. 5.14. *S. islandicus* NOB8-H2 CRISPR spacer origins. The two CRISPR arrays are shown, with each arrowhead or circle denoting a single spacer-repeat unit. The numbers to the right of each array indicates the total number of spacers. Arrowhead spacers are colour-coded according to the matching virus family; empty arrowheads indicate no matching spacer was found in the database search. Blue circular spacers indicate matches with conjugative plasmids.

A schematic of the two CRIPSR array spacers and their respective viral and plasmid origins is shown in **Figure 5.14**, and a summary of the spacers and their origins in **Table 5.7**. For CRISPR Array 1, 19 out of 115 total spacers produced a significant match (16.5%), and for Array 2, 16 out of 83 (19.3%). These numbers may appear low, but are comparable with those reported in a study of *S. acidocaldarius* spacers, where only 15% of spacers gave significant matches (Lillestøl *et al*. 2009), although this figure is reportedly higher at 40% for other members of the Sulfolobales (Lillestøl *et al*. 2006). The low percentage of spacer matches may also reflect the ongoing coevolutionary arms race between hosts and invading elements, as viruses are known to evade CRISPR-Cas through mutational and recombinational changes to their protospacers (Iranzo *et al*. 2013). An additional explanation lies in the fact that only a small fraction of microbial organisms have been cultured and classified, with the vast majority remaining unidentified 'biological dark matter' (Marcy *et al*. 2007). Therefore CRISPR spacers which do not elicit a match may come from as-yet unidentified plasmids and viruses.

For CRISPR Array 1, five virus families gave significant matches to spacers, with viruses from these families known to infect *Sulfolobus* and the closely related genus *Acidianus*. For CRISPR Array 2, one spacer matched to a virus from the family Globuloviridae, viruses which interact with more distantly-related crenarchaea such as *Pyrobaculum* (Haring *et al*. 2014). Plasmids accounted for approximately two-thirds of the spacer matches identified for Array 1 (63%), and one-third of spacers for Array 2, meaning 49% of the total identified spacers matched plasmid seqeuences **(Figure 5.15)**. In both cases, there were no matches to cryptic plasmids of the pRN type, although these have been found in *S. islandicus* hosts previously (Peng 2008). All plasmid matches were to conjugative plasmids, often called pNOB8 type, and here all conjugative plasmids were those found to be associated with *S. islandicus*, except for two occurrences of the spacer matching plasmid pAH1 from *Acidianus hospitalis* W1.

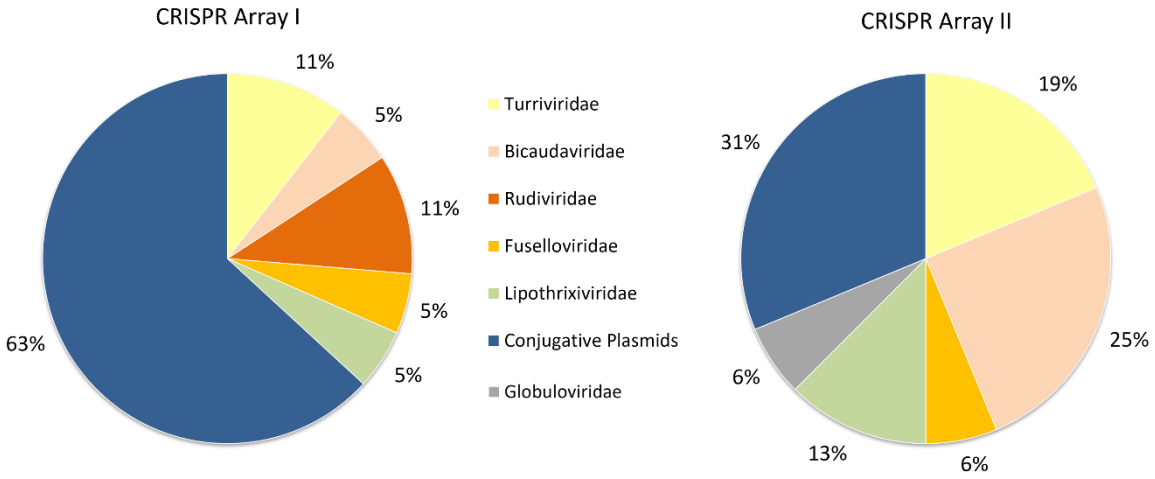


Fig. 5.15. Proportions of virus families and conjugative plasmids matching spacers in each CRISPR array. Spacer matches for *Sulfolobus* NOB8-H2 Array I (19 matches) and Array II (16 matches) against known crenarchaeal viruses and plasmids.

Table 5.7. Summary of *Sulfolobus* NOB8-H2 CRISPR spacer matches to crenarchaeal viruses and plasmids.

CRISPR Array	Spacer no.	Spacer length (bp)	Alignment length (bp)	e-value	Bit score	Virus Family	Plasmid Type
	1	40	30	0.001	34.4	Turriviridae	
	9	38	33	5.56E-07	45.4	Bicaudaviridae	
	12	40	33	0.005	32.5	Rudiviridae	
	22	41	40	3.77E-14	69.4		pAH1
	25	42	43	1.42E-08	51		pSOG1
	26	42	42	3.06E-10	56.5		pSOG2
	27	40	19	3.67E-04	36.2		pLD8501
	28	39	30	0.001	34.4		pAH1
1	33	38	38	0.004	32.5	Fuselloviridae	
	35	39	38	9.68E-10	54.7	Turriviridae	
	39	38	19	3.35E-04	36.2		pLD8501
	41	42	36	0.001	34.4		pYN01
	44	40	40	1.68E-12	63.9		pYN01
	45	38	38	2.00E-06	43.6		pLD8501
	53	42	29	2.38E-06	43.6		pNOB8
	54	104	29	4.00E-06	44.6		pNOB8
	70	39	34	7.54E-06	41.7		pNOB8
	106	40	28	2.83E-05	39.9	Rudiviridae	
	112	42	32	0.005	32.5	Lipothrixiviridae	
	5	40	40	7.88E-06	41.7	Fuselloviridae	
	21	38	29	0.004	32.5		pARN4
	25	40	24	0.005	32.5	Lipothrixiviridae	
	27	42	27	3.08E-05	39.9	Bicaudaviridae	
	31	37	33	2.46E-05	39.9	Bicaudaviridae	
	34	105	25	5.00E-05	41.9		pNOB8
	38	40	17	0.005		Bicaudaviridae	
2	41	39	35	2.08E-11	60.2	Turriviridae	
	46	41	23	0.005	32.5	Bicaudaviridae	
	48	37	20	0.004	32.5	Globuloviridae	
	49	41	35	8.22E-06	41.7		pYN01
	60	40	29	3.67E-04	36.2	Turriviridae	
	61	42	38	9.24E-10	54.7	Turriviridae	
	63	42	20	0.005	32.5		pSOG2
	78	41	24	0.005	32.5	Lipothrixiviridae	
	82	41	41	2.27E-11	60.2		pYN01

5.2.10 *Sulfolobus* NOB8-H2 and pNOB8 interactions

When first analysing the *Sulfolobus* NOB8-H2 genome, BLAST searches for plasmid pNOB8 against the genome were conducted. It has previously been shown that pNOB8 can integrate into and excise from the host chromosome, via an integrase-mediated site-specific recombination mechanism (She *et al*. 2004). This was apparent when sequencing of the initial, presumed NOB8-H2 strain (*S. solfataricus* P1), when multiple insertions of pNOB8 were observed in the host chromosome. Here, initial BLAST results showed 55 matches of pNOB8 to the NOB8-H2 chromosome **(Figure 5.4).** However, it was not apparent if these represented *bona fide* insertions of portions of the plasmid, or simply portions of sequence that were shared by both the plasmid and the host. Almost half of the pNOB8 BLAST matches (25) were clustered around the CRISPR arrays (coordinates 1,742,177 - 1,762,148 bp, **Figure 5.4**), and of these, the majority were between 28 and 34 bp in size. Further investigation showed that these BLAST hits matched the 24 bp CRISPR direct repeat (DR), plus a few bp either side: i.e. half of the pNOB8 BLAST matches were in fact the DR, which comprises the same sequence on both the plasmid and chromosome. When pNOB8 was first sequenced, a region of six 24 bp repeats, separated by ~40 bp was noted, and at the time, this region was thought to be mechanistically involved in plasmid incompatibility/segregation (She *et al.* 1998). However, it is now apparent that this region may represent a mini CRISPR array, comprising the same DR as the NOB8-H2 CRISPR array but potentially containing different spacers. A similar mini CRISPR array is also present on the *S. islandicus* plasmid pKEF9 (Liu *et al.* 2016).

The four *Sulfolobus* NOB8-H2 CRISPR spacers that gave a significant match to pNOB8 **(Table 5.7)** were mapped back to the plasmid **(Figure 5.16)**. On closer inspection, the two larger 'spacers' mentioned previously, of 104 bp (Array 1, spacer 54) and 105 bp (Array 2, spacer 34) match only to the 24 bp CRISPR DR, and have flanking sequences of 40 bp that do not match to the plasmid, and so are not *bona fide* spacers **(Figure 5.16, top)**.

Of the other two spacer matches, the spacer 53 sequence is located in ORF 10 on pNOB8, which encodes for a protein with sequence similarity to the TraG superfamily (She *et al*. 1998). Bacterial TraG homologues are thought to be actively involved in plasmid conjugation and transfer of plasmid DNA from host to recipient (Greve *et al*. 2004, Schröder & Lanka 2003). This leads to the speculation that during the adaptation stage, foreign DNA that 'becomes' a spacer is taken from a gene encoding a protein involved in invasiveness, i.e. a perfect target to prevent future invasions. However, a second spacer (70) matches to an intergenic region between ORFs 12 and 13 **(Figure 5.16, bottom)**. A summary of the NOB8-H2 CRISRPR array matches to pNOB8 regions is shown in **Table 5.8,** with the two presumed *bona fide* CRISPR spacers (53 and 70) highlighted in boldface.

Table 5.8. Summary of NOB8-H2 CRISPR array matches to pNOB8[a]

CRISPR Array	Spacer no.	Spacer length (bp)	Alignment length (bp)	E-value	Bit score	pNOB8 position (bp)	pNOB8 gene/location
1	**53**	**42**	**29**	**2.38E-06**	**43.6**	**7773-7801**	**ORF10 - traG**
1	54	104	29	4.00E-06	44.6	23899-23927	Direct repeats
1	**70**	**39**	**34**	**7.54E-06**	**41.7**	**11342-11375**	**Intergenic region**
2	34	105	25	5.00E-05	41.9	23772-23796	Direct repeats

[a] The two *bona fide* spacers, spacer 53 and 70, are in bold font.

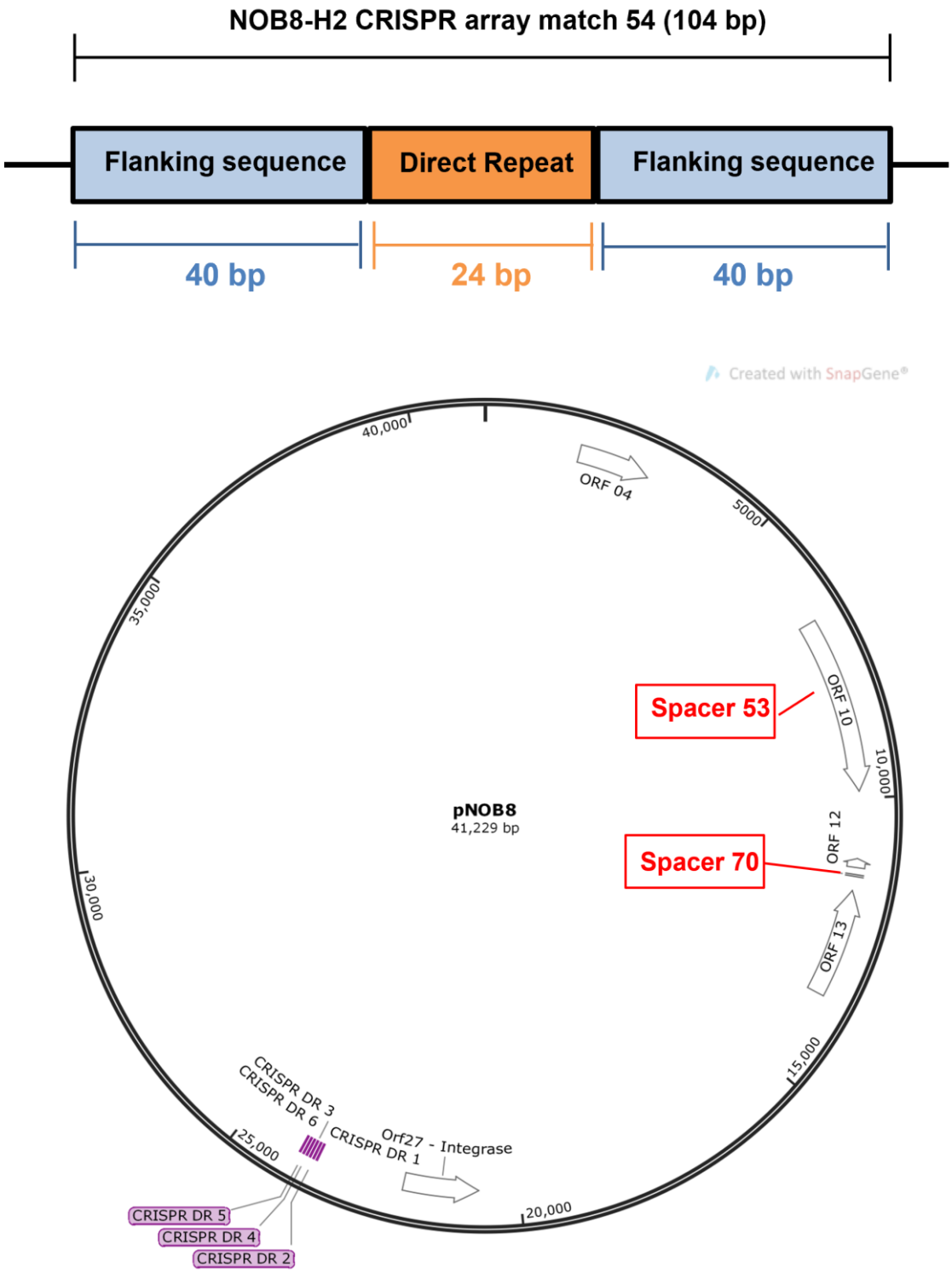


Figure. 5.16. *Sulfolobus* NOB8-H2 pNOB8 spacers. (Top) Example of CRISPR array BLAST match 54. The BLAST match is to the 24 bp DR, and flanking sequences do not match to pNOB8 sequences, therefore this is not a *bona fide* pNOB8 spacer**. (Bottom)** The position of the CRISPR array spacer matches to pNOB8 sequences are shown. Spacers are outlined in red; the pNOB8 CRISPR direct repeats are labelled in purple. Map is simplified, not all pNOB8 genes are shown. Figure created using Snapgene.

Of the rest of the pNOB8 BLAST matches, the majority are ~1,300 bp in length, with this larger-sized region implying that they may represent insertions of complete pNOB8 gene or genes into the *Sulfolobus* NOB8-H2 chromosome. Again, these regions were mapped back pNOB8, and were found to correspond to two genes encoding putative transposases, and one encoding an integrase. The two pNOB8 transposases were previously identified due to their homology to bacterial transposase families; the first (here denoted transposase A) is 406 amino acids and has a homologue in *H. pylori*, the second (transposase B, 413 aa) is homologous to *Mycobacterium* and *Rhizobium* transposases (She *et al.* 1998). The two transposases are present multiple times at different locations in the NOB8-H2 chromosome: transposase A eleven times, and transposase B nine times **(Figure 5.17)**, suggesting that initial transposition was from the chromosome to the plasmid (Stedman *et al.* 2000) .

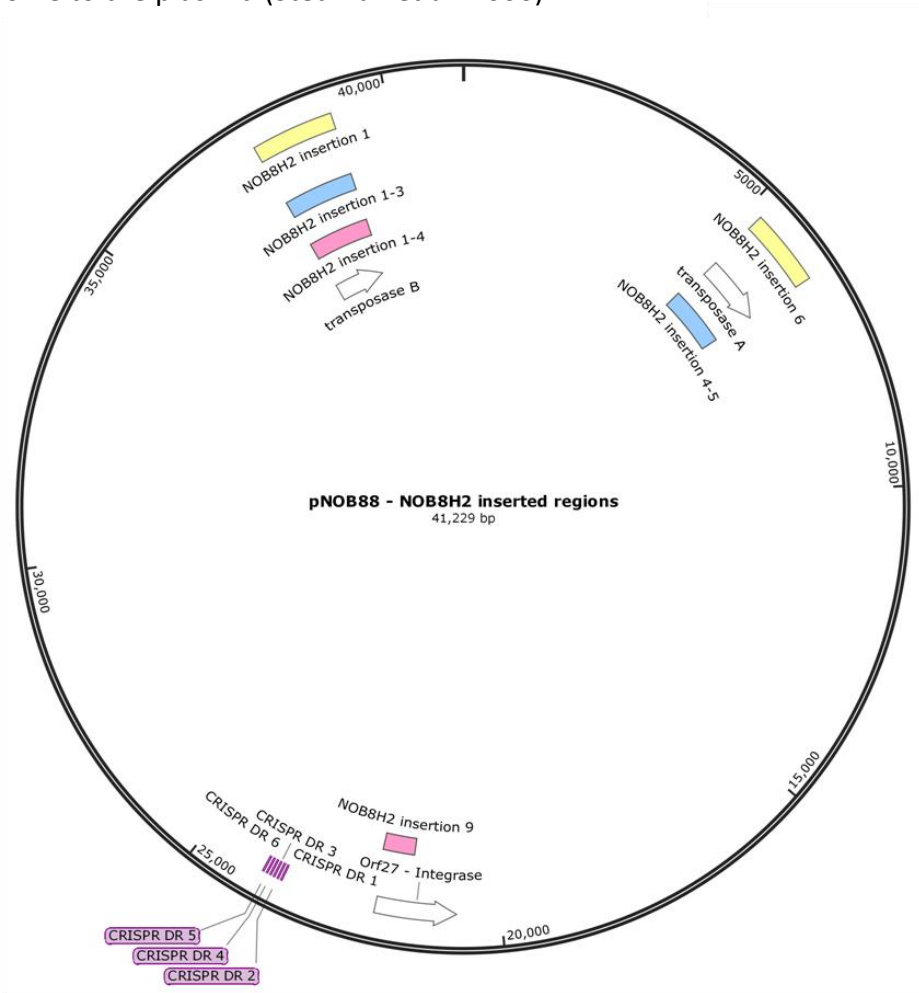


Figure 5.17. Regions of pNOB8 inserted into the *Sulfolobus* NOB8-H2 chromosome. The insertions of the pNOB8 transposase and integrase genes into the NOB8-H2 chromosome are shown. The numbers indicate the different BLAST hits for that particular gene fragment. Colours indicate the portion of the NOB8-H2 chromsome; blue - 1st Mb, yellow - 2nd Mb, pink - 3rd Mb.

The pNOB8 integrase was also identified via homology to a bacteriophage integrase, and was previously demonstrated to mediate complete integration of pNOB8 into the *S. solfataricus* P2 chromosome at a site overlapping a $tRNA^{Glu}$ gene (She *et al.* 2004). Here, a fragment matching part of the integrase is found at the end of the NOBH2 chromosome, just upstream of the $tRNA^{Glu}$ gene, implying that pNOB8 is integrated at this site. The fully sequenced NOB8-H2 chromosome was found to contain multiple pNOB8 repeat sequences at either end, making producing a circularised chromosome difficult. Therefore, when producing the complete polished assembly of the NOB8-H2 chromosome, these pNOB8 repeat regions were removed to enable complete circularisation of the chromosome. There were also other sequenced contigs which comprised tandem repeats of portions of pNOB8, which are also not present in the circularised NOB8-H2 chromosome (Dr John Davey, personal communication). This means that the full pNOB8 insertion is not apparent when viewing the NOB8-H2 chromosome, however the presence of the pNOB8 integrase upstream of $tRNA^{Glu}$ indicates that the plasmid is likely to be integrated here in full, either in single or perhaps multiple instances **(Figure 5.18)**.

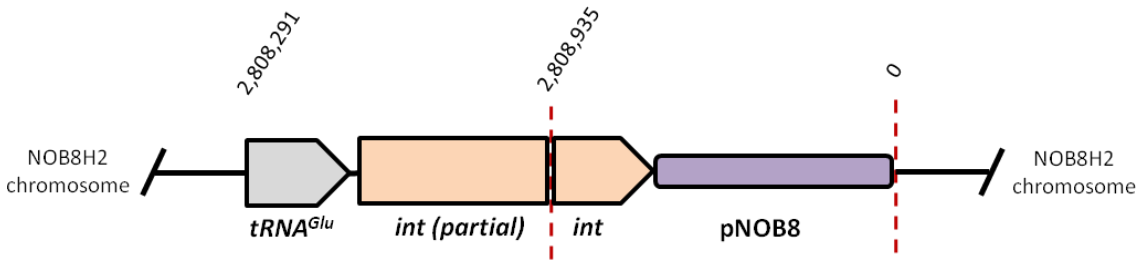


Fig. 5.18. Integration of pNOB8 into the *Sulfolobus* NOB-H2 chromosome. The position of the partial integrase gene (*int*) at the end of the NOB8-H2 chromosome is shown. The numbers above indicate the chromosomal coordinates. The dashed red lines indicate the start and end points of the circular chromosome, with the portion between including the remainder of the *int* gene and pNOB8 that were removed when circularising the chromosome. Figure not to scale.

5.2.11 Genetic analysis of plasmid pNOB8

The conjugative plasmid pNOB8, from *Sulfolobus* NOB8-H2, has previously been sequenced and subjected to genetic analysis (She *et al.* 1998). The plasmid is 41,229 bp in size, and contains 52 genes. Previously, putative functions were assigned to ~20% of the gene products, therefore new database searches were conducted to obtain additional functional information. Each protein sequence was analysed for potential homologues using BLASTp, and the pNOB8 protein FASTA file was also submitted to the WebMGA server for COG analysis as before. The output of these database searches is shown in **Table 5.9**.

Table 5.9. Putative functions of pNOB8 proteins based on BLASTp and COG analyses.

ORF	AA	Previously assigned[a]	Proposed function	COG Hit/Function	Bit score	-e	% Identity
1	116	—	—	—	179	56	75.70%
2	188	—	Txn. Reg.	COG1846 - MarR Txn. Reg.	301	102	87.21%
3	81	—	Txn. Reg.	COG1733 - Predicted Txn. Reg.	157	48	96.30%
4	422	ParB family	—	COG1475 - Spo0J	485	167	63.87%
5	537	Helicase family	DNA Helicase	COG1199 - DinG - Rad3 related helicases	980	0	90.69%
6	72	—	—	—	50.8	7	74.29%
7	50	—	—	—	99.8	26	98.00%
8	406	Transposase	IS200 family transposase	COG0675 - Transposase	703	0	81.66%
9	87	—	—	—	155	47	94.19%
10	1025	TraG family	Conjugation/DNA transfer	COG0433 - HerA helicase COG1321 - Mn-dependent Txn. Reg.	1719	0	80.69%
11	166	—	—	—	337	117	100.00%
12	52	—	Oxidoreductase	—	38.1	0.39	54.29%
13	630	ScdA cell division	—	—	600	0	56.75%
14	620	—	—	COG1938 - Archaeal ATP-grasp superfamily enzymes	271	74	32.20%
15	50	—	—	—	—	—	—
16	246	—	—	—	417	146	80.82%
17	253	—	—	—	251	81	57.75%
18	94	—	Txn. Reg.	COG1846 - MarR Txn. Reg.	58.5	9	39.78%
19	97	—	Txn. Reg.	—	122	34	62.89%
20	108	—	—	—	173	54	82.73%

21	**62**	**—**	**—**	**—**	**106**	**29**	**91.53%**
22	164	—	—	—	266	89	80.00%
23	69	—	—	—	120	34	76.81%
24	72	—	CopG Txn. Reg.	COG0864 - Predicted CopG/Arc/MetJ Txn. Reg.	122	35	86.76%
25	92	—	ZapB cell division	—	140	41	82.02%
26	101	—	—	—	197	64	96.04%
27	439	—	Integrase	COG0582 - XerC integrase	843	0	93.62%
28	80	—	—	—	146	44	91.25%
29	139	—	—	—	146	43	85.39%
30	248	—	—	—	118	29	68.83%
31	630	TrbE family	Conjugation/DNA transfer	COG3451 - VirB4 Type IV secretory pathway, VirB4 components	1238	0	96.84%
32	312	—	—	COG1196 - SMC - Chromosome segregation ATPases	488	172	93.27%
33	778	—	—	—	1398	0	89.13%
34	86	—	—	—	168	52	100.00%
35	109	—	—	—	176	55	90.83%
36	148	—	—	—	256	85	83.11%
37	52	—	—	—	95.5	24	96.15%
38	604	—	—	—	991	0	91.93%
39	165	—	—	—	301	103	91.61%
40	65	—	—	—	97.4	25	80.00%
41	110	—	—	—	184	58	76.85%
42	205	—	GNAT family N-acetylytransferase	—	389	136	91.22%
43	74	—	—	—	70.9	14	90.91%
44	93	—	**AspA**	COG1497 - Predicted transcriptional regulator	168	52	87.10%
45	470	ParB family	**ParB**	COG1475 - Spo0J	640	0	73.94%
46	315	ParA superfamily	**ParA**	COG1192 - Soj - ATPases involved in chromosome partitioning	633	0	98.41%
47	413	Transposase	IS256 family transposase	COG3328 - Transposase and inactivated derivatives	825	0	98.79%
48	142	—	—	—	270	91	90.85%
49	134	—	—	—	—	—	—
50	152	—	—	COG0720 - 6-pyruvoyl-tetrahydropterin synthase	226	73	72.37%
51	83	—	—	—	135	39	81.33%
52	81	—	—	—	138	41	81.48%

The top-matching BLAST hit was usually a closely-related strain e.g. *S. islandicus* and has not been included in the table for brevity. AA, Amino acid; Txn. Reg., Transcriptional regulator; -e, e-value, i.e. 56 = e^{-56}. Bit score, e-value and percentage identity relate to BLASTp output not to COG output. The segregation cassette proteins are in boldface.

[a] ORFS previously assigned putative functions (She *et al*. 1998).

Previously, nine proteins had been assigned putative functions based on homology (She *et al.* 1998). Here, an updated database search plus COG analysis gives hints to the potential function of a few more, although the majority remain 'hypothetical proteins'. ORFs 2,3,18 and 19 are putative transcriptional regulators, whilst ORF 24 was assigned to the COG group corresponding to the CopG/Arc/MetJ family of transcriptional regulators, where CopG is a small protein known to be involved in bacterial plasmid copy number control (del Solar *et al.* 2002). A CopG homologue encoded on the *S. islandicus* cryptic plasmid pRN1 has been demonstrated to bind to the putative *copG-rep* promoter, where *rep* encodes a large replication initiation protein (Lipps *et al.* 2001, Lipps 2009). Here, *orf24* does overlap with *orf25*, however this following gene is small in size compared with pRN1 *rep*, and does not match to any replication initiation proteins. ORFs 20-25 do appear to overlap, and so may represent a single operon, though this awaits further investigation.

Interestingly, three ORFs show similarity to proteins involved in chromosome organisation/segregation and cell division. The previously mentioned ORF 25 is similar to ZapB proteins, which in bacteria are involved in formation of the contractile Z-ring at mid-cell via interactions with FtsZ, mediating cytokinesis (Buss *et al,* 2013). However, *Sulfolobus* and the wider crenarchaea lack FtsZ, instead utilising Cdv (cell division) proteins as part of the cytokinetic machinery (Härtel & Schwille 2014). Here, ORF 25 may perform a different function given that it is plasmid-encoded, though it is an interesting candidate for further study along with ORF 24. ORF 27 matches to the COG class to which the XerC integrases belong. The XerC and XerD family of integrase/recombinases are required for the correct resolution of replicated chromosome dimers in bacteria, and a similar function has been attributed to the Xer homologue in *S. solfataricus* (Duggin *et al.* 2011).

ORF 32 is also interesting as it matches to COG1196 (SMC - Chromosome segregation ATPases). SMC family proteins (Structural Maintenance of Chromosomes) are found across all domains of life and, through interactions with other proteins, form complexes such as condensins and cohesins that mediate the large-scale three-dimensional organisation of the chromosome (Hassler *et al*. 2018). On bacterial chromosomes, SMC condensin complexes interact with ParB bound to *parS* sites near the origin of replication,

structuring the chromosome and allowing its correct segregation post-replication (Wang *et al.* 2017). Again though, whilst SMC condensin proteins are found in many archaeal phyla, they have not yet been detected in crenarchaea (Kamada & Barillà 2017). However, recently a novel SMC-family protein dubbed 'coalescin', which imparts higher-order compartmentalisation to *Sulfolobus* chromosomes, was characterised (Takemata *et al*. 2019). ORF 32 therefore may not perform an analogous function to the canonical SMC proteins found in bacteria, nevertheless, it may be a pNOB8 protein that is a promising candidate for further study.

Homologues of the three genes mentioned above are normally found on chromosomes, given that their encoded products play important roles in chromosome organisation. Here, only ORF 27, encoding the integrase, is also found on the NOB8-H2 chromosome **(Figure 5.18)**, raising the possibility that pNOB8 could have incorporated these genes from the chromosome during cycles of integration and excision, although these genes would have to be functionally characterised before drawing further conclusions.

An updated map of pNOB8, detailing the additional ORFs that now have putative functions based on database searches, is shown below in **Figure 5.19**.

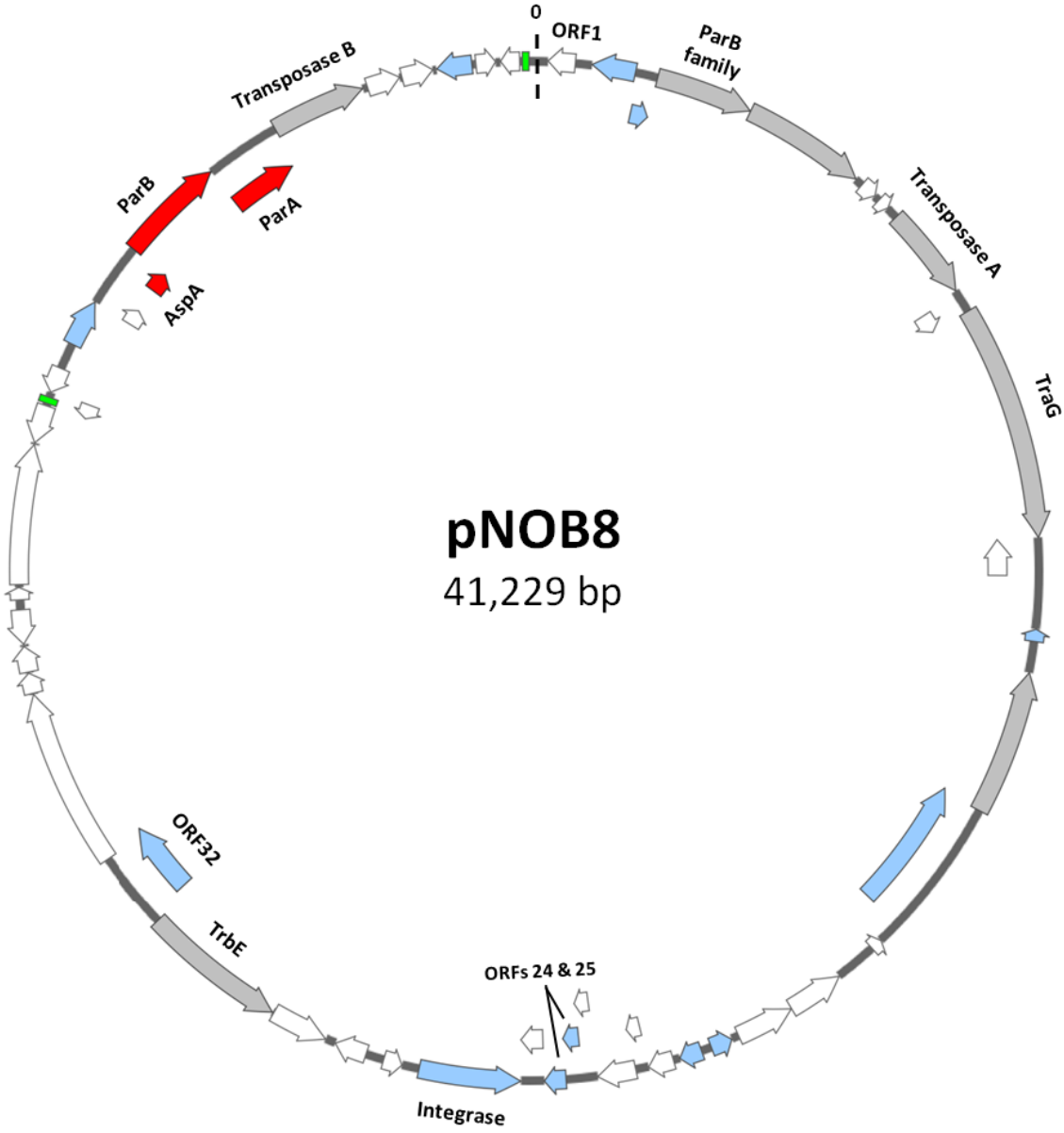


Fig. 5.19. Updated genetic map of pNOB8. The 52 pNOB8 ORFs are shown. Grey, previously assigned function; blue, newly proposed functions based on database homology searches; red, the segregation proteins AspA, ParB and ParA; white, unknown function. The two green segments are 85 bp repeats which form the border of the 8 kb fragment deleted from the variant pNOB8-33. The 0 coordinate is represented by the dashed black line, and some ORFs of interest are labelled.

The 'mini-CRISPR' array of pNOB8 was also investigated, using the same methodology as when ascribing the viral and plasmid sources of the *Sulfolobus* NOB8-H2 CRISPR spacers. This region of the plasmid contains six direct repeats (DR) of 24 bp, which are identical in sequence to the DRs in the NOB8-H2 CRISPR array. It is not the case that this entire region (6 DRs plus five 'spacers') has been 'copied and pasted' from the *Sulfolobus* NOB8-H2

genome, as BLASTing the pNOB8 'spacers' against the NOB-8H2 CRISPR arrays returned zero hits **(Figure 5.20)**. It has recently been demonstrated that the lesser-studied CRISPR Type IV systems are primarily encoded on plasmids, and that their spacers predominantly target other plasmids, suggesting a role in inter-plasmid competition (Pinilla-Redondo *et al*. 2020). Furthermore, some archaeal viruses contain mini-CRISPR arrays whose spacers target closely-related viruses occurring in the same population (Medvedeva *et al*. 2019), suggesting that this might be a common mechanism employed by mobile genetic elements to enable competition amongst close relatives.

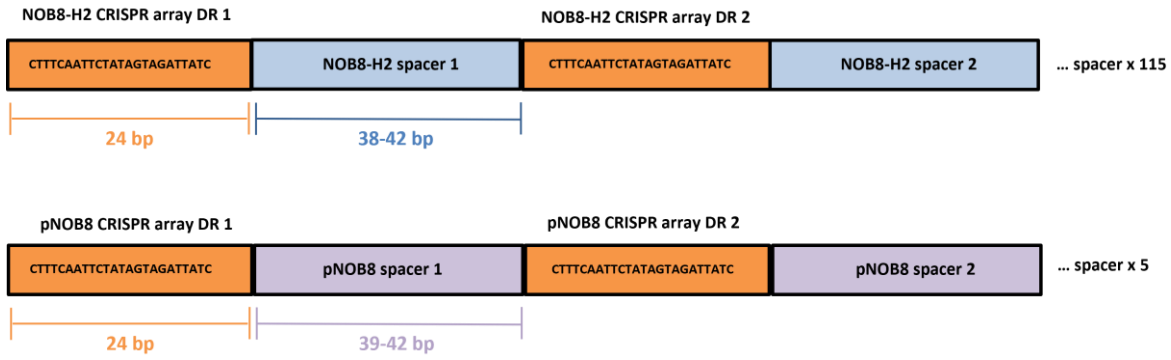


Fig. 5.20. Schematic of NOB8-H2 and pNOB8 CRISPR arrays. The CRISPR direct repeat sequences (DR) are the same between the NOB8-H2 chromosome and the pNOB8 CRISPR arrays. However the spacers are different between the two arrays (represented using different colours) Only the first two repeat/spacers are shown, the number of spacers for each array is indicated to the right (chromosome array 2 with 83 spacers is not depicted).

The five sequences between the DRs, corresponding to spacers, are between 39 and 42 bp in length, similar to the NOB8-H2 spacers. BLASTing the sequences against the 53 accession numbers of crenarchaeal viruses and plasmids as before resulted in zero significant hits. Therefore to widen the search parameters, BLASTn searches were conducted against all Crenarchaeaota. This did produce matches, however, they do not meet the criteria for significance that was used for the NOB8-H2 spacers. For example, spacer 1 matched to the *Staphylothermus marinus* F1 genome (Bit score 38.3, e-value 0.02, 22 bp alignment length), with the other spacers (except for spacer 5) giving similar values. This is interesting as *S. marinus* is a crenarchaeon, though belongs to a different order (Desulfurococcales) to *Sulfolobus*. Here, the e-value is much larger than when assessing the pNOB8 spacers previously (0.02 *cf.* ~$2x10^{-6}$) therefore this may not

represent an actual spacer, however this would be interesting to determine in future studies. Spacer 5 gave a 100% identical match to the *S. islandicus* M.16.27 genome, probably as a result of pNOB8 either being incorporated into the M.16.27 chromosome, or transiting within this closely-related strain at some point.

Finally, the pNOB8 ORFs were also investigated for any homology to known anti-CRISPR proteins (Acrs). Recent work has shown that many bacteriophages encode Acrs to overcome the host cell's CRISPR-Cas defence systems. Acr proteins encoded in phage genomes that inhibit the function of Type I-F and I-E CRISPR-Cas systems have been discovered in *Pseudomonas aeruginosa* (Pawluk 2016), and those inhibiting Type II Cas9 effectors reported in *Streptococcus pyogenes* (Lee 2018). Intriguingly, Acr proteins were recently discovered for the first time in archaeal viruses: two viruses that infect *Sulfolobus islandicus* encoded proteins that conferred protection against the endogenous CRISPR Type I-D system (He 2018).

The set of known anti-CRISPR proteins was searched by uploading the pNOB8 protein FASTA file to http://cefg.uestc.cn/anti-CRISPRdb/, a database constructed by manually screening the literature for referenced anti-CRISPR proteins, and comparing downloaded protein sequence and structural information with that of known Acrs (Dong *et al.* 2017). Interestingly, ORF39 was found to be 40% identical to a known anti-CRISPR protein from a recently characterised bacterium, *Bacteroides ihuae* (bit score 37, e-value 4E-05). The protein, from the AcrIIA9 family, has homologues in the functionally uncharacterised PcfK superfamily, which are found in bacteria and viruses (Bhoobalan-Chitty *et al*. 2019). A second database was also used, paCRISPR, which uses a machine-learning model to give greater accuracy compared to existing homology-based predicters, resulting in significantly higher predictive performance. (Wang *et al*. 2020). Here, due to the fact that Acr proteins possess little sequence similarity, paCRISPR uses an evolutionary-based Position-Specific Scoring Matrix (PSSM) model to predict anti-CRISPRS. The paCRISPR model was trained against a dataset of 98 experimentally verified Acrs alongside 260 non-anti-CRISPR proteins (Wang *et al*. 2020). The paCRISPR server returned 18 predicted anti-CRISPRS from the input of 52 pNOB8 ORFs, scoring each between 0 and 1 and based on a threshold value of 0.5. In this case, the threshold value and thus sensitivity is probably too low, but nevertheless, the top predicted pNOB8 anti-CRISPRs ORFs score quite highly.

ORF22 and ORF16 returned predicted scores of 0.815 and 0.761 respectively, and thus represent candidate anti-CRISPR proteins that require further study.

5.3 Conclusions and discussion

Sulfolobus NOB8-H2 is a hyperthermophilic crenarchaeal strain, originally isolated from hot springs on the Japanese island of Hokkaido (Schleper *et al*. 1995). Different species of *Sulfolobus* have been utilised as models to study fundamental biological processes in archaea (Bernander 2000, 2007), and their evolutionary adaptation to extreme environs means they are a valuable resource for biotechnological and industrial applications (Quehenberger, *et al*. 2017). Over the last few decades, research interest in *Sulfolobus* has resulted in complete genome sequencing of over 20 strains, primarily of species *S. solfataricus*, *S. acidocaldarius* and *S. islandicus*, although *S. solfataricus* has recently been proposed to be reclassified in the novel genus *Saccharolobus* (Sakai & Kurosawa 2018).

Sulfolobus NOB8-H2 also harbours plasmid pNOB8, the first conjugative plasmid isolated from an archaeon (She *et al*. 1998), which has previously been sequenced. The plasmid is known to undergo integration and excision into and from the host chromosome, thus acting as a driver for horizontal gene transfer and evolution of the genome (She *et al.* 2004). Plasmid pNOB8 also encodes a tricistronic operon, *aspA-parB-parA*, reminiscent of the bicistronic segregation cassettes found on bacterial plasmids and chromosomes (Hayes & Barillà 2006b). The mechanism of segregation of pNOB8, and the functional roles played by the AspA, ParB and ParA proteins have been investigated elsewhere (Schumacher *et al*. 2015, Zhang & Schumacher 2017), and in this thesis.

In this chapter, we reported the sequencing of the *Sulfolobus* NOB8-H2 strain, and undertook an analysis of the genome, assessing its place as a novel strain within the *Sulfolobus* phylogeny, alongside characterising some main features of the genome, such as its CRISPR-Cas systems. The interactions of the NOB8-H2 chromosome with pNOB8 were also detailed, and the pNOB8 ORFS were subjected to an updated investigation, to discover any proteins functioning in chromosome organisation or anti-CRISPR activity.

After determining that the 'laboratory strain' of *Sulfolobus* NOB8-H2 was in actual fact a derived strain of *S. solfataricus* P1 that had been used for plasmid conjugation experiments, the 'original' NOB8-H2 isolate was acquired, and sequenced using a combination of MinION and Illumina sequencing. The assembled, polished chromosome, 2.81 Mb in size, was found to be most similar to the species *S. islandicus*, but sufficiently different to be classed as a novel strain **(Table 5.2)**. A phylogenetic tree based on the 16s rRNA positioned NOB8-H2 with *S. solfataricus* strains, but a more robust tree, based on ten concatenated core genes, placed NOB8-H2 as a sister group to the *S. islandicus* strains **(Figure 5.7)**.

These initial analyses involved small groups of genes, or small sections of the genome, therefore *in silico* DNA-DNA hybridisation (DDH) was conducted to assess the whole-genome similarity of NOB8-H2 to other strains. It has been suggested that a DDH score of <70% indicates a novel *species* rather than *strain* (Auch *et al.* 2010a, Tindall *et al*. 2010), and this value was used by Dai and colleagues when proposing *Sulfolobus* sp. A20 as a novel species (Dai *et al*. 2016). In the case of A20, the DDH values were far below 70%, the highest being 23.10%. Here, DDH values for NOB8-H2 and more distantly related species such as *S. solfataricus* and *S. acidocaldarius* are far less than 70%, but for *S. islandicus*, this value is much closer at ~65% **(Table 5.5)**. This suggests that *Sulfolobus* NOB8-H2 is a novel strain of *S. islandicus* rather than a novel species, and the whole genome comparisons seen in **Figures 5.8 & 5.9** support this.

An annotation of the circularised *Sulfolobus* NOB8-H2 chromosome provided many data to analyse. Of particular interest was the endogenous CRIPSR-Cas system, which was found to comprise two types, IA and IIIB, along with a spacer acquisition locus (**Figure 5.11)**. This system was almost identical to that observed in *S. islandicus* REY15A (Peng *et al.* 2013). The *Sulfolobus* NOB8-H2 spacers were analysed and matched to various crenarchaeal viruses and plasmids, as both of these mobile elements are invasive and so induce an immune response in the host (Makarova *et al*. 2011). Therefore, CRISPR spacers represent a history of past conflicts between host and invading element. Only 18% of the spacers matched to known viruses and plasmids of the crenarchaea, but this figure is similar to that of previous studies (Lillestøl *et al*. 2009), and could be explained simply by the size of the database, i.e. many more invasive elements await discovery and

sequencing. However, the spacer analysis could be extended to include the protein sequences in addition to nucleotide sequences, for both the *Sulfolobus* NOB8-H2 and the pNOB8 CRIPSR spacers. Using this approach for a number of crenarchaeal species gave ~30% matches to viruses and plasmids (Shah *et al*. 2009). It would be interesting to further analyse the pNOB8 spacers to see if any matches to crenarchaeal viruses are found. There is evidence of conflict between these two different types of invading element; e.g. *S. solfataricus* infected with the SMV1 virus triggered spacer acquisition, not from the virus, but derived from a co-infecting conjugative plasmid (Erdmann *et al*. 2013). The spacers conferred resistance against the plasmid, not the virus, suggesting complex interactions not only between the host and invasive elements, but between competing invasive elements themselves.

The plasmid pNOB8 has previously been investigated, and putative functions assigned to several of its encoded proteins. Here, pNOB8 was subjected to database searches to expand the predicted functions of its ORFs. Of interest is ORF39, which shares homology with a recently-identified bacterial anti-CRISPR protein from the bacterium *Bacteroides ihuae*. This protein is from the AcrIIA9 family, which has anti-CRISPR activity against Type II-A systems. Given that *Sulfolobus* NOB8-H2 contains Types I and III CRISPR loci, and that Type II systems have so far been found only in bacteria, not archaea (Uribe *et al*. 2019), it is possible that ORF39 has a function unrelated to anti-CRISPR activity. Anti-CRISPR proteins were first discovered, and are primarily encoded by bacteriophages (Bondy-Denomy *et al*. 2013), but genes encoding these proteins have also been found on bacterial plasmids (Mahendra *et al*. 2020).

In archaea, it appears that Acrs have so far been found encoded on viruses only, inhibiting both Type I and Type III CRISPR systems (Bhoobalan-Chitty *et al*. 2019, He *et al*. 2018). Anti-CRISPR proteins are known to use a variety of mechanisms to subvert host CRISPR responses: interference of crRNA loading, prevention of target DNA binding, and inhibition of nuclease activity (Pinilla-Redondo *et al*. 2020b, León *et al*. 2021). The anti-CRISPR protein AcrIF9 was recently shown to induce the Type I-F effector complex to bind non-specifically to DNA lacking spacer sequence complementarity (Lu *et al*. 2021). In archaea, the *Sulfolobus islandicus* rod-shaped virus 2 encodes the anti-CRISPR AcrID1, which interacts with Cas10 and prevents target DNA cleavage, along with AcrIIIB1 which

inhibits Type III-B targeting of viral middle and late-expressed genes (Peng *et al*. 2020). Here, ORF22, ORF16 and possibly other pNOB8 ORFs are worthy of future investigations, as one of these could possibly represent the first plasmid-encoded archaeal anti-CRISPR protein.

Chapter 6

Discussion and Future Work

6.1 Discussion

The accurate dissemination of replicated DNA molecules to daughter cells is a vital process that occurs in organisms from all domains of life. Genetic material must be accurately segregated within the cell volume prior to division, such that cellular progeny inherits the correct amount and ploidy is maintained.

The molecular mechanisms underpinning genome segregation have been extensively studied in prokaryotes using low-copy number bacterial plasmids as model systems, as these were found to encode an active partitioning system to ensure correct maintenance (Bouet & Funnell 2019). The partitioning system (Par) comprises the centromere-like DNA sequence *parS*, a centromere-binding protein ParB, and an NTPase motor protein ParA. Following plasmid replication, the ParB protein binds with specificity to the *parS* site, and this nucleoprotein structure then recruits ParA, resulting in the formation of the partition complex (Hayes & Barillà 2006a, Schumacher 2008). The partition complex componentry acts to position the two replicated plasmids within the cell prior to cell division, and a number of segregation models for different partition systems have been proposed (see Section 1.5). Par systems are not only dedicated plasmid segregation mechanisms, but are extensively encoded on bacterial chromosomes across diverse taxa (Livny *et al*. 2007).

The present study details the partition apparatus encoded on an archaeal plasmid, pNOB8, harboured by the thermophilic strain *Sulfolobus* NOB8-H2, which is of interest as little is known about genome segregation mechanisms in this domain of life. The segregation system comprises three genes, *aspA*, *parB*, and *parA*, and a palindromic centromere-like sequence upstream of the partition cassette (Schumacher *et al*. 2015). The AspA protein performs the role of site-specific centromere-binding protein in this system, binding with high affinity to the palindrome. The plasmid harbours a second identical palindrome, and an initial aim of this work was to characterise the interactions of AspA at this second site and speculate on its functional role here. Moreover, the contribution of specific AspA residues to its activity were assessed, and both these topics will be discussed in the first part of this section.

6.1.1 AspA is a putative transcriptional regulator of two operons

The plasmid pNOB8 harbours two identical 23 bp palindromes, one of which is located immediately upstream of the *aspA-parB-parA* cassette, whilst the other is approximately 1.5 kb away. Typically, the plasmid centromere-like *parS* sequence is a unique site located either upstream or downstream of the partition genes, depending on the partition system type. Some bacterial plasmids however, possess multiple *parS* sites: the *par2* locus of pB171 contains two sites, *parC1* and *parC2*, that are upstream and downstream of the segregation cassette, with both sites being bound with similar affinity by the cognate CBP ParB to form a nucleoprotein complex (Ringgaard *et al*. 2007a). ParB was found to pair molecules containing *parC1* and *parC2* in both a homologous and heterologous fashion, and electron microscopy studies showed the formation of large nucleoprotein complexes comprising several DNA molecules connected by ParB bound to *parC1* (Ringgaard *et al*. 2007b). Recent studies have demonstrated that the *parC2* site is crucial for effective plasmid segregation, as plasmid segregation occurs if at least one single *parC2* repeat is present. The *parC2* site is hypothesised to function as the initial nucleation site for ParB binding, before formation of a second nucleoprotein complex occurs at *parC1*, with both *parC1* and *parC2* then brought together via protein-protein interactions between ParB molecules bound to both sites (Alnaqshabandy thesis 2020).

Of more relevance to the palindrome arrangements on pNOB8, the *Streptococcus pyogenes* plasmid pSM19035 harbours three centromeres: two of which lie in the promoter regions of the genes encoding ParA and ParB homologues (in this case, *parA* and *parB* do not form an operon), however the third centromere lies several kilobases away on the plasmid, in the promoter region of a copy-number control gene, *copS*. The ParB-like CBP in this system, (dubbed Omega) therefore regulates expression of these three genes, and it was demonstrated that the *par* site upstream of gene *w* (omega) was the main centromeric sequence (Dmowski & Jagura-Burdzy 2011, Dmowski & Kern-Zdanowicz 2016). The *E. coli* plasmid prophage N15 harbours four *parS* sites, all located some distance from the *par* operon, each of which demonstrated the ability to act as a centromere, ensuring functional redundancy if one or more sites were lost (Grigoriev & Lobocka 2001).

Bacterial chromosomes also contain multiple *parS* sites, the majority of which are situated in origin-proximal locations (Livny *et al*. 2007). The *P. aeruginosa* chromosome incorporates ten *parS* sites, four of which are close to the origin of replication *oriC*. ParB was found to bind to all sites, but in a hierarchical manner, and had greatest affinity to the four *parS* sites proximal to *oriC*, and moreover, a single *parS* site out of the four closest to *oriC* was necessary and sufficient for accurate chromosome partitioning (Jecz *et al.* 2015). The hierarchy of ParB binding was correlated with *parS* composition; greatest affinity was to perfect palindromes, with affinity decreasing when single mismatches were present. None of the remaining six *parS* sites were able to function as centromere-like sequences in correct chromosome segregation (Kusiak *et al* 2011, Jecz *et al.* 2015). The *parS* sites were also required to be within a certain distance relative to *oriC* in order for correct segregation to occur (Lagage *et al*. 2016). These data indicated that a ParB-DNA complex at a single *parS* site will enable accurate segregation only if the complex forms near to *oriC* (Jecz *et al.* 2015).

Here, the interactions of AspA with the second palindrome on the archaeal plasmid pNOB8 were investigated. AspA was previously reported to bind with high affinity ($K_d$app ~50 nM) to the first palindrome. Here, the first palindrome refers to the sequence upstream of the partition cassette, and is henceforth designated *pal1,* whereas *pal2* refers to the second palindrome under investigation. AspA was found to bind avidly to *pal2 in vitro*, having a slightly greater affinity to that of *pal1* (~20 nM). The comparable affinity between the two sites may suggest that either may be sufficient for correct segregation of the plasmid, as mentioned above with prophage N15 and the chromosomal *parS* sites of *P. aeruginosa* (Grigoriev & Lobocka 2001, Jecz *et al.* 2015). In contrast, some plasmids may contain multiple CBP binding sites, of which only one acts as the main centromere sequence (Kulińska *et al*. 2011, Dmowski & Kern-Zdanowicz 2016). Possessing a similar affinity for both sites could suggest that AspA does not preferentially bind to one over the other *in vivo*, although this, alongside the relative contributions of both pNOB8 palindromes to accurate segregation (i.e. if either one or both are necessary), would require experimental validation. Often, bacterial plasmid *parS* sites contain multiple iterations of the repeat motifs, and the minimal number of repeats required for successful partitioning can be experimentally tested, sometimes with a single repeat being sufficient (Alnaqshabandy thesis 2020).

This is not the case with pNOB8, as both palindromes are single 23 bp sequences and are not arranged into clusters of repeat iterations.

Although the affinity of AspA to the *pal1* and *pal2* sequences is similar, the patterns of binding and spreading at the two sites appear to be different **(Figure 6.1)**. It was previously demonstrated that at higher concentrations, AspA could spread from *pal1*, upstream of the start of the *aspA* gene for over 200 bp, to form an extended nucleoprotein complex (Schumacher *et al.* 2015, Barillá 2016). In this study, we demonstrated that AspA does not spread continuously from *pal2* at higher concentrations *in vitro*, but rather binds at two distinct regions, as evidenced by the clear regions of protection observed in DNase I footprinting assays. One region is the palindrome itself, whereas the other, slightly larger area (ROP2), is adjacent to putative TATA box and Transcription Factor IIB recognition element (BRE) regulatory sequences upstream of *orf41.* The different binding patterns at each palindrome may indicate a different function is performed by AspA at each site (see below), even though the affinity to the DNA is similar at both *pal1* and *pal2*. It is not the case that the amount of spreading from each palindrome is related to intergenic distance, as both palindromes are equidistant from neighbouring genes: *pal1* is 319 bp from *orf43*, whilst *pal2* is 317 bp from *orf42*.

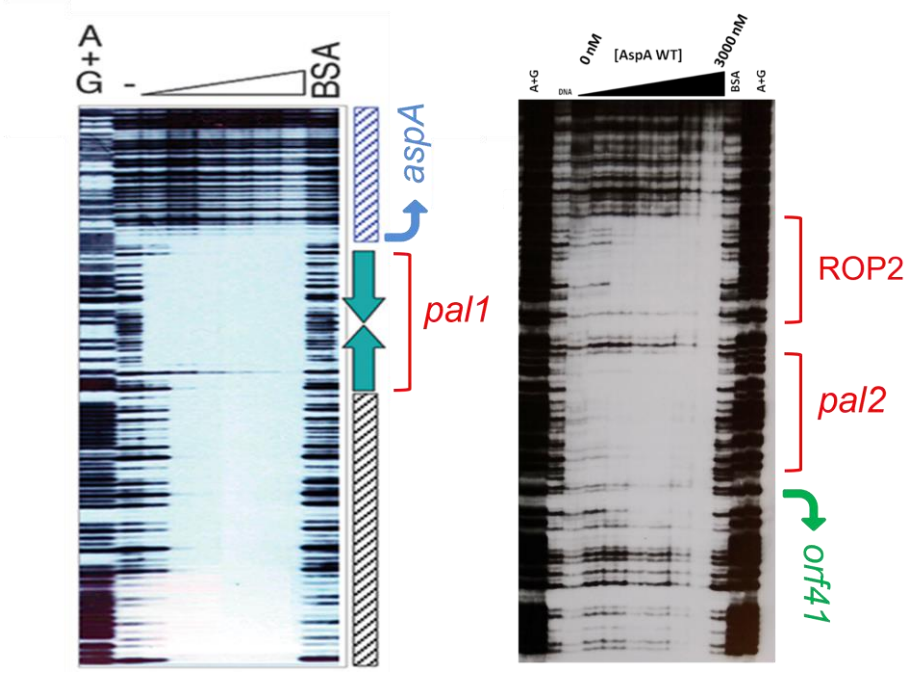


Figure 6.1. Comparison of AspA binding at each palindromic site. (Left) AspA spreads from *pal1* upstream of the start of the *aspA* gene. Adapted from Schumacher *et al*. 2015 **(Right)** AspA does not spread in the same manner from *pal2* upstream of the start of *orf41*, but instead forms two distinct regions of protection.

CBPs of Type Ib, Type II and Type III plasmid partition systems are known to autoregulate expression of the *parAB* operon by binding to upstream promoter regions within the *parS* sites, along with functioning in plasmid segregation (Larsen *et al*. 2007, Schumacher 2008). In Type Ia systems, the autoregulatory role is performed by the ParA motor protein, as here the *parS* site is located downstream of the partition cassette. Given that the palindrome of pNOB8 is upstream of the partition genes, it is therefore reasonable to speculate that AspA performs two functions at *pal1:* autoregulation of expression of partition genes, and formation of the pre-partition complex, which can be extended along the DNA due to the spreading capacity of the protein. At *pal2*, the lack of spreading at higher concentrations may indicate that AspA here only performs the role of transcriptional regulation, although any regulation of gene expression by AspA at either site is yet to be determined.

The *pal2* site is located immediately upstream of three genes: *orf41, orf40* and *orf39*. *Orf41* and *orf40* overlap, whilst there are only 5 bp between *orf40* and *orf39*, therefore it is likely that the three genes may be co-transcribed and part of a single operon. Unfortunately, searching protein databases using the amino acid sequences of the three *orfs* did not provide any clues as to potential function. *Orf41* shares ~75% identity with a number of zinc finger, SWIM-domain containing proteins from species within the Sulfolobaceae, however it is difficult to assign a putative role as these proteins can perform a broad range of cellular functions, and can bind to a wide range of substrates (Krishna *et al*. 2003). Searches against o*rf40* and *orf39* return matches against hypothetical/uncharacterised proteins only. In the example given above of pSM19035, where an additional *parS* site is located a large distance from the partition cassette, the CBP Omega here also regulates expression of a copy-number regulatory gene (Dmowski & Kern-Zdanowicz 2016). In addition, Omega regulates expression of Toxin/Antitoxin (TA) genes that are immediately downstream of its own gene *w*, whereas more usually, expression of TA systems are autoregulated by the TA complex or antitoxin itself (Yamaguchi *et al*. 2011). Although TA modules predominantly comprise two genes, there are examples of three component systems (Unterholzner *et al*. 2013, Gerdes *et al*. 2021). The bacterial broad host-range plasmid pTF-FC2 harbours a TA system encoded by *pasABC*, where, *pasC* encodes a protein that enhances the neutralising ability of the antitoxin (Smith & Rawlings 1998).

It is plausible that the three *orfs* upstream of *pal2* could encode a maintenance mechanism analogous to a three-component TA system, and that its expression is modulated by AspA at the second palindrome. Furthermore, another property of pNOB8 is that it is able to form a genetic variant, pNOB8-33, due to the deletion of a 8 kb fragment of the plasmid. The pNOB8-33 deletion variant appears frequently when transformed into non-host *Sulfolobus* strains, whereas it is not observed in the parental NOB8-H2 strain (Schleper *et al*. 1995, She *et al*. 1998). The 8 kb region includes the *aspA-parB-parA* cassette, meaning that pNOB8-33 does not encode the partition apparatus and so fails to segregate properly in the parental strain (see Figure 5.19). Moreover, one of the borders of the 8kb fragment overlaps with *orf40*, with pNOB8-33 harbouring only *orf39* out of the three genes. If this operon did encode a TA system, with *orf39* encoding the toxin, this could act against cells harbouring segregation-defective pNOB8-33 variants. Here, one speculation is that AspA could act to regulate the expression of two plasmid maintenance systems: the partition apparatus, along with a TA module, by binding to *pal1* and *pal2* respectively.

6.1.2 The role of AspA in pNOB8 segregation

The *in vitro* investigations of AspA binding to the two palindromic sites have shed light on its potential role as a transcriptional regulator. However, a vital function of the CBP is in formation of the nucleoprotein complex at the centromere to mediate segregation. The spreading of AspA from *pal1* to form an extended complex mirrors that of ParB CBPs found on both plasmids and chromosomes (Bartosik *et al*. 2004, Schumacher 2012, Graham *et al*. 2014, Tran *et al.* 2018). Spreading of the CBP serves a number of functions: the extended ParB-DNA architecture can potentially interact with a greater number of motor proteins thus stabilising the partition complex, expression of neighbouring genes can be modulated, and CBPs can interact with each other to organise and condense DNA (Rodionov *et al.* 1999, Kusiak *et al.* 2011, Graham *et al.* 2014, Schumacher *et al*. 2015). AspA has been shown to bind with a stoichiometry of 1:1 with ParB-N, which in turn associates with both ParA and non-specific DNA, therefore the spreading of AspA from *pal1* could result in a more robust partitioning process. However, AspA spreading could

also aid in compaction and condensing of pNOB8 DNA, making transport within the cell more straightforward. The CBP of *B. subtilis* and *H. pylori,* Spo0J (ParB), was demonstrated to bridge together separate sections of DNA via protein-protein interactions between the N-termini of multiple Spo0J dimers (Graham *et al*. 2014, Chen *et al*. 2015).

Here, initial AFM experiments with the CBP AspA, incubated with DNA incorporating one, or both palindromes, appear to show the formation of large protein complexes, with the DNA appearing looped or bridged between them. When just *pal2* was present on a single linear fragment, at increased concentrations of AspA, several DNA fragments appeared to be attached to a large aggregation of AspA molecules. Similarly, when utilising a circular plasmid harbouring both AspA binding sites, greater protein concentrations appeared to induce bridging events, where two distinct sections of the DNA were brought together, presumably due to protein-protein interactions between AspA bound to both specific and non-specific DNA molecules. These experiments were not completed, and so these data should be interpreted with caution, as the condensed appearance of the plasmid may have constituted a natural topology. The CBP of plasmid TP228, ParG has been shown to selectively bind to its cognate site *parH* under AFM using a linear fragment (Wu *et al.* 2011), and post-replicated plasmids bound to ParB of pSM19035 has also been observed (Pratto *et al*. 2009). Although not specifically involved in segregation, proteins functioning to condense, stabilise and repair DNA have also been observed to promote bridging and looping between separate DNA strands by binding to multiple locations on the DNA (Laurens *et al.* 2012, Murugesapillai *et al*. 2014, Andres *et al*. 2019). Therefore, it is feasible that AspA molecules, when bound to both *pal1* and *pal2*, and non-specific DNA, may interact and induce or bolster a plasmid topology which is more conducive to compaction and accurate segregation. A model for the actions of AspA on the DNA, both in terms of transcriptional regulation, and segregation dynamics, is shown in **Figure 6.2**.

Future work to further unpick the role(s) of AspA could include measurement of both the transcript and translated protein levels for those genes thought to be regulated by AspA, a system for which has recently been described for *S. solfataricus* (Lo Gullo 2019). Additionally, further investigations of the products of *orfs39-41*, could include cloning of the genes, purifying the recombinant proteins, and assaying for complex formation

between them. EMSA and DNase footprinting experiments could be performed to see if any of the proteins bind to the upstream regulatory region, as this could indicate a role as co-repressor along with AspA. Finally, it would be useful to acquire more AFM data to further understand the action of AspA at the palindromes. Testing the preference of AspA for either site, for example, could be measured by first introducing a directionality to the DNA fragment, so it is clear where *pal1* and *pal2* are located. This has previously been done by using a DNA molecule labelled with biotin at one end, such that the resultant biotin-streptavidin complex was clearly visible under AFM conditions (Vörös *et al*. 2017).

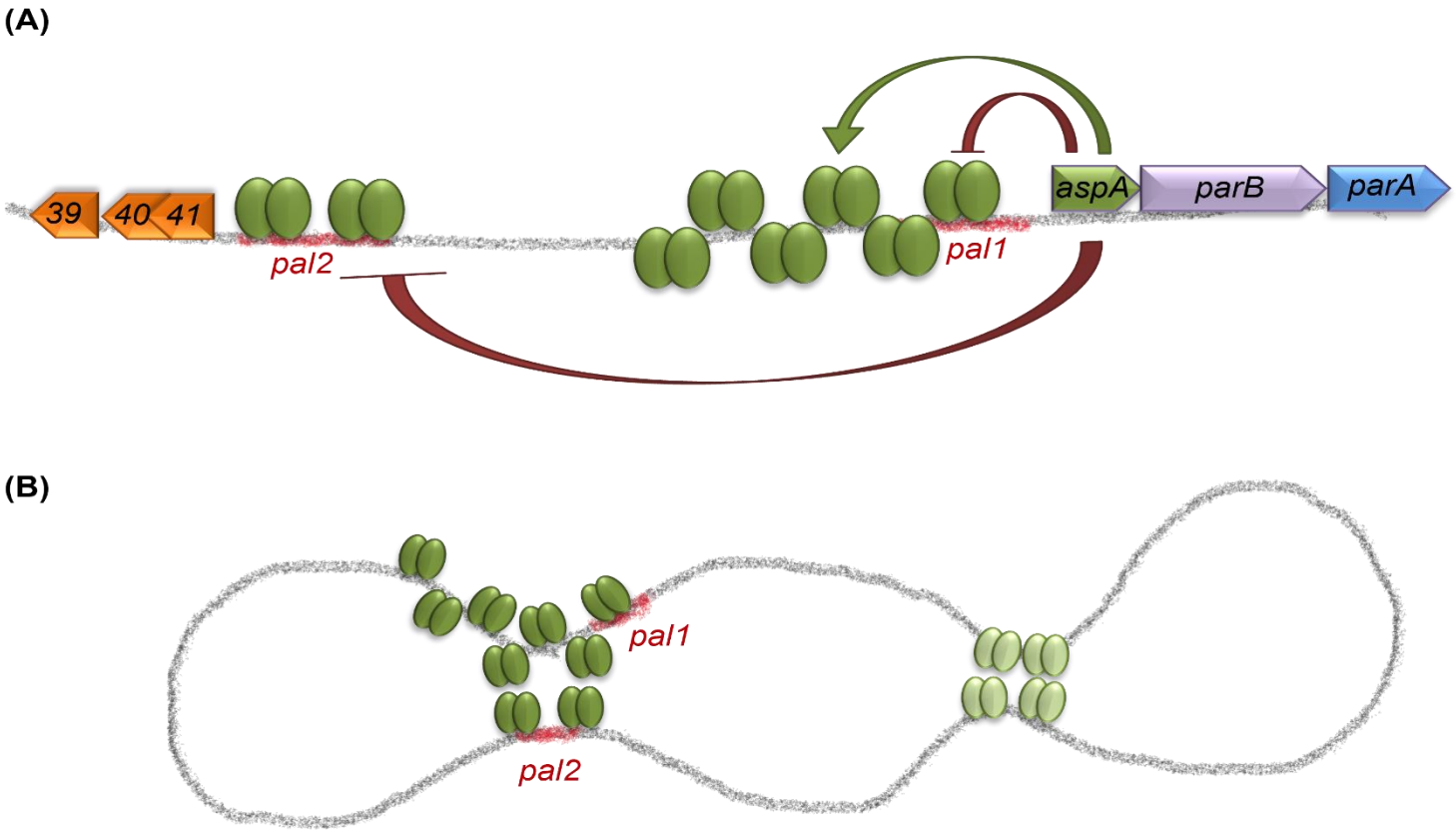


Figure 6.2. Model of AspA functions at the pNOB8 palindromic sites. (A). AspA binds to the first palindrome, *pal1*, upstream of the partition cassette, where it can potentially autoregulate transcription of the *aspA-parB-parA* operon. It can also spread along the DNA at higher concentrations, forming an extended complex that is beneficial for segregation. AspA also binds to the second palindrome, *pal2*, upstream of *orfs41-39*, where it could function to regulate transcription of these genes. **(B)** Cartoon representation of pNOB8. A speculative model for how AspA could act to bolster condensation of the plasmid. The natural plasmid supercoiled topology could be bolstered by AspA-AspA interactions when bound to both palindromes. Non-specific binding by AspA at other locations (light green) could aid DNA organisation at several sites on the plasmid. The ParB and ParA proteins are not shown.

6.1.3 Characterising AspA residues important for function

In order for a centromere-binding protein to function correctly in both DNA segregation and transcriptional regulation, the protein must possess certain properties, such as the ability to dimerise, bind to the DNA, and spread along neighbouring DNA from the initial nucleation site, along with interacting with the cognate motor protein. Some, or all of these functional properties are conserved across a range of CBPs, both chromosomal and plasmid-encoded, in both bacteria and archaea (Lynch & Wang 1995, Rodionov *et al*. 1999, Schumacher & Funnell 2005, Kalliomaa-Sanford *et al*. 2012, Jalal *et al*. 2021). The CBP of pNOB8, AspA, was shown to be dimeric, to bind site-specifically to the palindromic sites, and spread along the DNA forming an extended nucleoprotein complex (Schumacher *et al*. 2015). In this study, we assessed specific amino acids within AspA that may contribute to some of these functions, using mutagenesis followed by different *in vitro* assays such as EMSA. A summary of the AspA mutants produced in this study, and any observed phenotypic effect, is outlined below in **Table 6.1** and described further in the text.

Table 6.1. Summary of AspA mutants produced in this study

AspA mutant	Phenotypic effect	Source
AspA Y41A	Reduction in DNA binding activity	This study
AspA Q42A	Reduction in DNA binding activity	This study
AspA L52K	Reduced spreading due to less dimer-dimer interactions	This study
AspA E54A	Reduced spreading due to less dimer-dimer interactions	This study
AspA L12G	Inability to bind DNA	This study
AspA I85G	Slight reduction in DNA-binding activity	This study
AspA V89G	Slight reduction in DNA-binding activity	This study
AspA I85GV89G	Inability to bind DNA, dimerisation inhibited	This study

As the crystal structures of Asp in both apo- and DNA-bound form (including the 23 bp palindrome) had previously been solved, these were used to guide mutagenesis strategies. A high degree of affinity for the DNA is vital for CBPs to perform their role. AspA is a wHTH CBP, in common with Type Ia plasmid and most chromosomal CBPs. The

majority of the interactions between AspA and the DNA phosphate backbone come from the first two N-terminal alpha-helices, with glutamines 42 and 46 from the 'recognition helix' (α3) providing many DNA base contacts. The conformational flexibility of the N-terminal helices allows them to insert into both major and minor grooves of the DNA (Schumacher 2015). Perhaps surprisingly, given the multitude of DNA-contacting residues (each AspA monomer makes between 12 and 14 contacts), a previously constructed AspA-R49A mutant (Arg-49 makes backbone contacts) completely abolished DNA-binding activity. This indicates that a specific single amino-acid can be crucial for activity, here via the charge interactions between the basic arginine side-chain and the negative charge of the phosphate backbone. The contribution of specific arginine residues to CBP function has previously been assessed. Graham and colleagues replaced several arginines with alanines in conserved residues of *B. subtilis* Spo0J (ParB). Interestingly, mutations of three arginines in close proximity (dubbed the arginine patch) did not negate the DNA-binding activity, instead affecting DNA bridging between different molecules. Mutation of another arginine residue did however produce a band-shift distinct from wild-type as evidenced by EMSA, perhaps as a result of defective *parS* interactions (Graham *et al*. 2014).

Here, additional AspA conserved residues that may be important for DNA binding were mutated: Tyr-41 and Gln-42 within the recognition helix, of which Tyr-41 makes backbone and base contacts, and Gln-42 makes base contacts only (Schumacher 2015). Mutation almost completely abrogated DNA-binding activity for AspA-Y41A and AspA-Q42A as measured by EMSA, with only a slight smearing apparent, indicating a small degree of binding, evident at the highest protein concentration. As both of these residues occur at the start of the recognition helix, it is possible that replacement with alanine may cause local conformational changes at the level of this singular helix, without introducing overall secondary structure alterations.

Jalal and colleagues recently demonstrated how DNA-binding proteins have evolved exquisite specificity to their binding sites. They identified a subset of four amino acids (three within the recognition helix) of *C. crescentus* ParB which were responsible for specific binding to the cognate *parS* site. Alanine scanning mutagenesis demonstrated that here, again, single residue substitutions abolished DNA binding activity. Furthermore, mutation of these four key residues to those found at equivalent positions in a closely-

related DNA-binding protein, Noc, switched binding specificity to that of the Noc binding site (Jalal *et al.* 2020a). Although not a ParB homologue, AspA performs a functionally analogous role, and given that bacterial chromosomal *par* sequences and their CBPs are thought to have arisen early in evolution (Livny *et al.* 2007), it is likely to also be the case with archaeal equivalents that this degree of specificity has been selected for over evolutionary time.

The specificity of the CBP-DNA interaction is not only reliant on specific protein residues; single DNA base mutations, and insertions or deletions of bases have been shown to significantly reduce transcription factor binding affinity (Czerny *et al*. 2013, Liang *et al*. 1996). During this study, a set of four palindrome mutants were constructed, where individual A-T pairs at the palindrome centre were each mutated to G-C, and EMSA used to assess any differences in binding affinity of WT AspA. Chromosomal ParB of *Pseudomonas* was shown to bind preferentially to *parS* sites with perfect palindromes, and with slightly lower affinity to those with mismatched nucleotides (Jecz *et al*. 2015). We hypothesised that AspA would show decreased affinity to DNA fragments containing mutated palindromes. Unfortunately, initial EMSA data were inconclusive and not repeated due to time constraints, and the data not included in this thesis. However, this experiment could be explored and expanded upon further in future work, along with further screening of AspA residues that contact the DNA.

Similarly, the structure of AspA-DNA was used to asses which residues may contribute to another important function, that of spreading along the DNA, which is mediated by the insertion of the recognition helices from two adjacent dimers into the same DNA major groove (Schumacher 2015). Plasmid and chromosomal ParB mutants defective in spreading, and with concomitant defects in partitioning and/or gene silencing, have been previously described (Rodionov *et al*. 1999, Breier & Grossman 2007, Kusiak *et al*. 2011). The two mutants that were hypothesised to be deficient in spreading, AspA-L52K and AspA-E54A, demonstrated a distinct band-shift pattern compared to that of the wild-type protein. Both mutants formed discrete bands on the EMSA films, rather than smearing patterns seen with the wild-type, that is hypothesised to reflect a reduced occupancy of the protein on the DNA. AspA-L52K and AspA E-54A dimers therefore appear unable to

bind the DNA in an adjacent fashion due to decreased dimer-dimer interactions, and so cannot spread and completely cover the DNA until at higher concentrations. For AspA-E54A, the smearing pattern did no become evident until 1000 nM. This is in contrast to the wild-type protein, where smearing becomes evident much earlier at lower concentrations of 200-500 nM. One explanation could be that the mutants have decreased affinity for the DNA, however the pertinent residues do not contact either backbone or bases in the structure, and qualitative estimates of their DNA binding affinities, as measured by constructing a ligand curve based on the one-site specific binding model showed similar apparent dissociation constants to wild-type AspA. A similar observation was made for *B. subtilis* Spo0J (ParB), where a single amino acid substitution did not affect DNA binding *in vitro*, whereas spreading was defective due to decreased dimer-dimer interactions (Breier & Grossman 2007).

The recently-characterised structure of the *S. solfataricus* chromosomal CBP SegB, showed that dimer-dimer interactions were mediated in part by close-range hydrophobic interactions between proline residues at a loop interface between dimers (Yen *et al*. 2021). SegB had previously been shown to spread along the DNA from its cognate site in common with bacterial ParBs (Kalliomaa-Sanford *et al*. 2012). Mutating this proline residue to glycine removed the ability of SegB to spread in DNase footprinting assays, demonstrating that this residue is key in facilitating SegB dimer-dimer interactions and formation of the extended nucleoprotein complex (Yen *et al*. 2021). Similarly, DNase footprinting experiments with AspA-E54A appeared to show a decrease in dimer-dimer interactions at *pal2*. However, the effect is not as obvious as the protein does not spread from this site as it does from *pal1.* Future experiments could validate the importance of these AspA residues in dimer-dimer interactions, and therefore spreading, by performing DNAse footprinting with the mutants and DNA incorporating *pal1*. Performing these experiments in conjunction with WT AspA, which has demonstrated spreading up to 200 bp from *pal1*, should help elucidate the role of E54 further, as any lack of spreading here would produce a more distinct observable footprinting pattern.

Furthermore, microscale thermophoresis (MST) experiments could be performed with wild-type and non-spreading mutants, as in addition to obtaining the dissociation constant, MST can provide a measure of the cooperativity of binding via the derived Hill

coefficient. It would be hypothesised that AspA-L52K and AspA-E54A would show non-cooperative, or less cooperative modes of binding when compared to wild-type. In this study, MST experiments were begun, primarily to obtain a more quantitative value for the dissociation constant, however difficulties in obtaining consistent results meant they were discontinued.

Lastly, the crystal structure of the apo-form of AspA allowed us to speculate which amino acids induce dimerisation, and thus DNA binding, as all CBPs bind to the DNA as dimers (Funnell 2016). The C-terminal domain of AspA has previously been assigned as the dimerisation domain, in common with many other CBPs, in which this domain acts to stabilise the dimeric form to facilitate DNA binding (Delbrück *et al.* 2002, Baxter & Funnell 2014, Oliva 2016). In the AspA structure however, both the N-terminal loop, and C-terminal residues of one monomer are situated in close enough proximity to the equivalent residues in the second AspA monomer to allow hydrogen-bonding and other molecular interactions at distances of less than 4 Å. Of the four dimerisation mutants created, the two C-terminal single mutants plus double mutant resulted in more consistency across the DMP cross-linking and EMSA data. The single mutants, AspA-I85G and AspA-V89G formed less higher-order oligomers and bound to the DNA less avidly than the wild-type, with a concomitant decrease in oligomerisation and binding seen with the AspA-I85GV-89G double mutant. The N-terminal mutant, AspA-L12G however, appeared to form more oligomers when cross-linked, but did not bind to the DNA at all. It is plausible that Leu-12 in the flexible N-terminal loop could make transient contacts with the DNA, thus bolstering the interaction, however these contacts were not observed in the crystal structure. Future work could involve making AspA truncation mutants by deleting portions of both the N- and C-termini, to assess their relative contributions to dimerisation and DNA binding. This was previously done in ParB of plasmid P1: truncation of the final 70 amino acids completely abrogated DNA-binding, showing the importance of the C-terminus in dimer formation (Rodionov *et al*. 1999).

An overview of the various AspA residues which contribute to the functioning of the protein, is shown below in **Figure 6.3**.

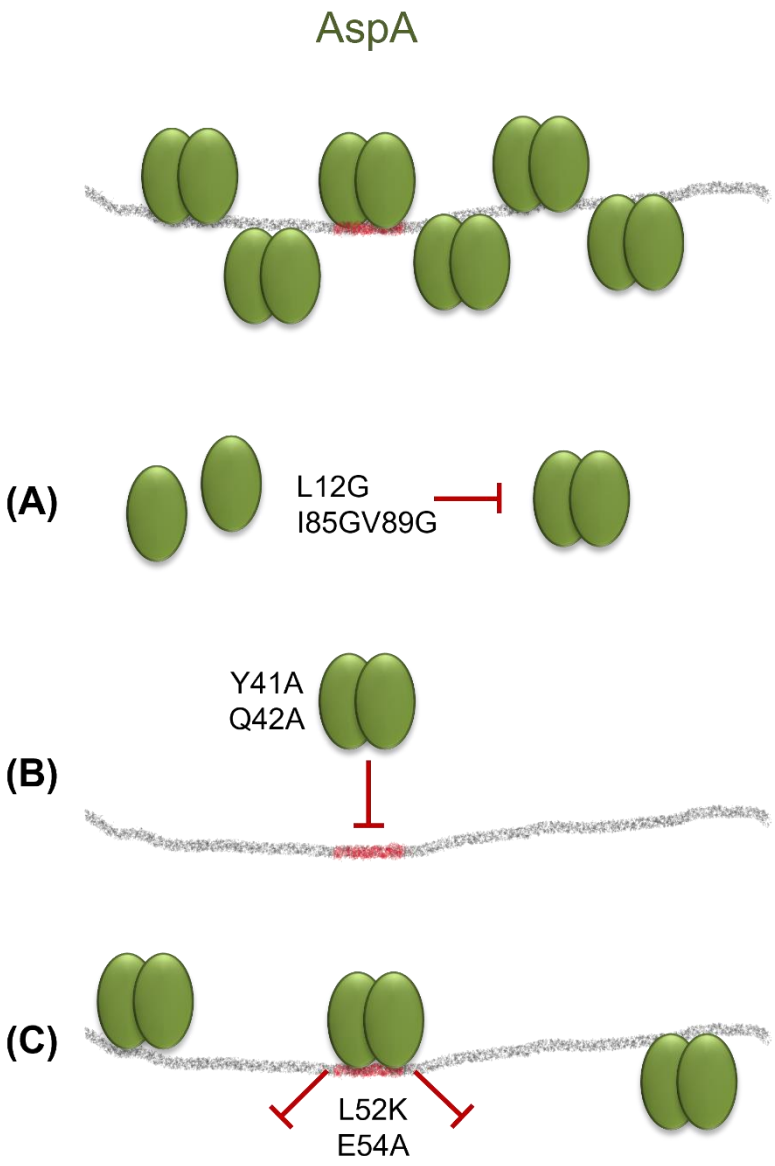


Figure 6.3. Summary of AspA mutations and their effect on function. (Top) WT AspA is shown binding as a dimer to the second palindromic site (red), and spreading along the pNOB8 DNA (grey). The various mutations that provoke a particular loss of function and highlight the importance of that amino acid for that action are shown; **(A)** AspA dimerisation **(B)** DNA-binding, and **(C)** Spreading along the DNA via adjacent dimer-dimer interactions.

6.1.3 Future work involving pNOB8 ParB

The work involving the interactions between the CBP AspA, and the proposed adaptor ParB, unfortunately did not generate any positive data, as *in vitro* AspA:ParB-N complexes were not observed in chemical cross-linking experiments. However, the delineation of the domain boundaries of ParB, and subsequent generation of constructs will allow future experiments to be conducted. It was originally planned to use the ParB constructs in binding/interaction and ATPase activity assays with ParA, to test the hypothesis that it is the flexible linker of ParB that both binds ParA and stimulates its ATP hydrolysis activity. Unfortunately, it was not possible to conduct these experiments due to time constraints.

Recent investigations of the properties of bacterial ParBs have also opened avenues for future studies using pNOB8 ParB. Work by Soh and colleagues on *B. subtilis* ParB demonstrated that the protein had enzymatic properties alongside DNA-binding capabilities, due to the presence of a CTP-binding motif and the subsequent verification of its CTPase activity (Soh *et al*. 2019). The CTP hydrolase activity was subsequently demonstrated to be a property of *M. xanthus* and *C. crescentus* ParBs, and led to models of dissociation from *parS* and spreading along the DNA due to CTP binding and hydrolysis (Osorio-Valeriano *et al*. 2019, Jalal *et al*. 2021). Intriguingly, although pNOB8 ParB does not function as a site-specific DNA binding protein, instead binding to AspA, its N-terminal domain does share homology and structural similarity with several bacterial ParBs, and pNOB8 ParB also harbours the CTP-binding motif (Schumacher *et al*. 2015, Soh *et al*. 2019, Osorio-Valeriano *et al*. 2019). The nucleotide-binding motif of *B. subtilis* ParB is a conserved GxxRxxA, and whilst this amino acid sequence forms an ATP-binding pocket in the eukaryotic enzyme sulfiredoxin, *B. subtilis* ParB did not bind ATP, nor other nucleotides except CTP (Soh *et al*. 2019). In pNOB8 ParB, it appears the motif is not exactly conserved, comprising GxxRxxI, however its presence raises interesting questions. Given that pNOB8 ParB-N performs a distinct function and binds a different substrate compared to bacterial ParBs, this suggests that any nucleotide binding and hydrolysis properties of pNOB8 ParB-N may have evolved to fulfil another role, presumably relating to its interaction with AspA. Although ParB-N was shown to bind to AspA-DNA in the absence of nucleotide, perhaps the addition of CTP binding and hydrolysis would increase

the rate of binding or induce a conformational change in ParB-N that further stabilies the interaction. Interestingly, the other ParB protein located on pNOB8 (ORF4) also contains the same GxxRxxI motif. Characterising any CTP binding and hydrolysis properties of ParB-N would therefore represent an interesting direction for future work.

6.1.4 A novel *S. islandicus* strain and potential pNOB8-encoded anti-CRISPR proteins

The majority of this work has involved studying the segregation system encoded by the *Sulfolobus* conjugative plasmid pNOB8. The plasmid is harboured by the strain *Sulfolobus* NOB8-H2, isolated from Japanese thermal hot springs, and later sequenced (Schleper *et al*. 1995, She *et al*. 1998). Episodes of plasmid insertion into and excision from the NOB8-H2 chromosome have been observed, providing a mechanism for genome evolution (She *et al*. 2004). Therefore, we sequenced the *Sulfolobus* NOB8-H2 genome to further understand the interactions between plasmid and host. A combination of phylogenetic analysis and *in silico* whole genome comparisons demonstrated that NOB8-H2 is a novel strain of *S. islandicus*, although some metrics used to determine novel prokaryotic species, such as DNA-DNA hybridisation (DDH), indicated that NOB8-H2 lies close to the species/strain boundary.

To definitively answer the strain/species question would require additional analysis and possible *in vitro* experiments. The concept of species, particularly when applied to prokaryotes, could almost be viewed as a philosophical question. The 16S rRNA sequence is still used, at least as a starting point, to delineate prokaryotic species, although it is thought to lack the resolving power necessary to guarantee correct species delineation (Rossellò-Mora & Amann 2001). The threshold figure for 16S rRNA similarity of 97% is commonly cited: strains sharing less than this percentage are not thought to be members of the same species (Tindall *et al*. 2010). Additional genetic metrics such as DDH and DNA base ratio (G+C%) are advised to be used when the 16S rRNA alone is not sufficient for species classification. Furthermore, a ‘polyphasic approach’, incorporating both

genomic and phenotypic information is suggested for a reliable classification to be achieved (Rossellò-Mora & Amann 2001). Chan and colleagues used a combination of *in silico* genomic techniques: average nucleotide identity and core genome (>100 genes) phylogenetic analysis to delineate bacterial species (Chan *et al*. 2012). When proposing the novel genus *Saccharolobus* and the reclassification of *Sulfolobus solfataricus* as *Saccharolobus solfataricus*, phylogenetic analysis was supplemented with experimental data on optimal growth conditions, cell morphology and sugar usage to distinguish species (Sakai & Kurosawa 2018). For *Sulfolobus* NOB8-H2, it would be of great interest to derive further phylogenies, based on a larger gene set, and combine this with phenotypic characteristics to help answer the species/strain question, although that is beyond the scope of this thesis. For now, it seems appropriate to consider NOB8-H2 a novel strain of *S. islandicus*. It was also planned to sequence pNOB8 and compare this with the published sequence, to determine if the plasmid had undergone any genetic changes whilst being grown in the laboratory. Unfortunately, despite many efforts to isolate the plasmid, it could not be prepared to the required purity due to the presence of chromosomal DNA, therefore this line of enquiry was not pursued **(Figure 6.4)**.

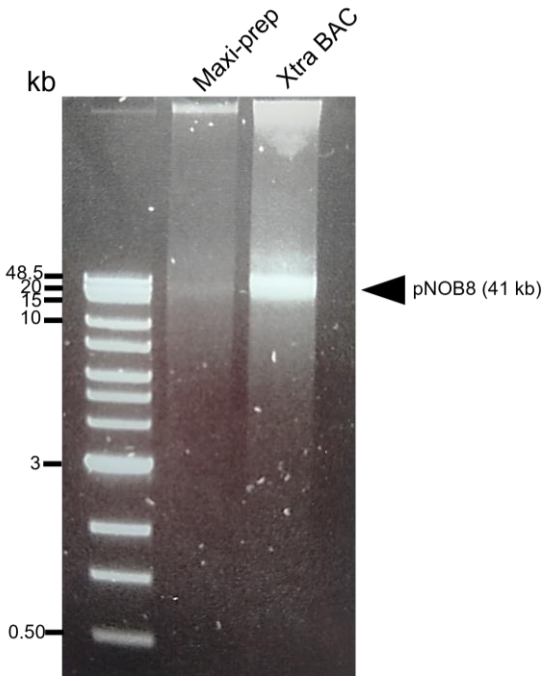


Figure 6.4. Example of attempted pNOB8 isolation. A 0.8% agarose gel showing the isolation of plasmid pNOB8 from culture using either maxi-prep or Xtra BAC plasmid purification methods. The distinct band at the correct location on the gel indicates pNOB8 is present, however there is also smearing representing chromosomal DNA contamination. The marker used was the Quick-Load 1 kb Extend DNA ladder (NEB).

A more in-depth analysis of the NOB8-H2 chromosome showed that it harboured two CRIPSR-Cas modules, plus two CRISPR arrays containing spacers against invasive genetic elements, such as viruses and plasmids, that the strain had previously encountered. Of particular interest to this study, two of the NOB8-H2 CRISPR spacers matched to pNOB8 sequences. Invasion by foreign genetic elements has led to the evolution of a variety of endogenous defence mechanism by the host, including restriction-modification systems, abortive infection, and CRISPR-Cas (Pinilla-Redondo *et al.* 2020b). In turn, plasmids and viruses themselves have developed mechanisms to subvert host defences, and with respect to CRISPR, this evolutionary arms-race has produced anti-CRISPR proteins (Acrs). Acrs have been found encoded on both bacterial viruses and plasmids, but thus far have been discovered on archaeal viruses only, and not plasmids (Bondy-Denomy *et al*. 2013, Bhoobalan-Chitty *et al*. 2019, Mahendra *et al*. 2020). Given that pNOB8 is maintained within NOB8-H2, but that the host CRISPR locus contains two pNOB8 spacers, we speculate that the plasmid may encode an anti-CRISPR protein.

Referencing the 52 pNOB8 ORFs against the paCRISPR online database produced 18 predicted Acrs, using a threshold value of 0.5, therefore it is likely that the threshold here was too low. Nevertheless, a database such as this provides an initial means of screening for Acrs, which is useful as these proteins have diverse sequences and structural motifs, making identification more difficult (Wang *et al*. 2020). The top-scoring predicted Acr from the paCRISPR database, ORF22, is relatively small at 164 amino acids, which is within the size-range of known Acr proteins of <200 aa (Pinilla-Redondo *et al.* 2020b). Two other ORFs out of the top five predicted Acrs, ORFs 48 and 49 are also small proteins of 142 and 134 aa respectively. Interestingly, none of these three ORFs returned any hits when subjecting pNOB8 to an updated BLASTp and COG analysis to ascribe putative functions to the plasmid proteins. Furthermore, ORFs 48 and 49 lie within the 8 kb fragment that is removed from the pNOB8-33 deletion variant, meaning that if either possessed any anti-CRISPR activity, pNOB8-33 would presumably be susceptible to the host CRISPR response, particularly as the two (potentially three) spacers against pNOB8 are located in the remaining ~33 kb of plasmid sequence. The observation that pNOB8-33 is not stably maintained, but is eventually lost, adds weight to this speculation. This suggests that the 8 kb deletion fragment could potentially harbour three separate mechanisms, that when removed, results in the loss of pNOB8-33: (i) the segregation cassette, (ii) a putative TA

operon, (iii) an anti-CRISPR protein. Hypotheses regarding anti-CRISPR proteins would require experimental validation, therefore future work could involve the initial cloning and purification of these ORFs, followed by *in vitro* DNA cleavage assays. These experiments would involve incubating a CRISPR nuclease-sgRNA complex with the purified putative Acrs, and introducing a linear DNA fragment containing one of the pNOB8 spacers sequences. Anti-CRISPR activity would be demonstrated by the DNA remaining uncleaved. These assays have previously validated the anti-CRISPR activities of bacterial mobile genetic elements (Lee *et al*. 2018, Uribe *et al*. 2019). Viral plaque/spot assays have also been used to assess Acr activity, whereby bacterial lawns are infected with phage. The endogenous phage-targeting CRISPR-Cas system can be inhibited by transforming the cells with a plasmid expressing an Acr, leading to virus infectivity as seen by viral plaques (Lu *et al*. 2021). This effect has also lead to the characterisation of AcrID1, an anti-CRISPR encoded by the *S. islandicus* lytic virus SIRV3, which inhibits CRISPR subtype I-D (He *et al*. 2018).

It is also interesting to speculate on the ongoing evolutionary arms-race between plasmid and host that allows continued compatibility between the two. Continuing the CRISPR theme, recent investigations into Type IV systems have shown that not only are they primarily encoded on plasmid-like elements, but 80% of their spacer content matches to other plasmids (Pinilla-Redondo *et al*. 2020a). This suggests that plasmids harbouring these CRISPR systems may mediate inter-plasmid conflict, and thus be stably maintained by virtue of preventing other plasmids from entering the host cell, or by targeting a plasmid already with the host that is competing for metabolic resources (Pinilla-Redondo *et al*. 2022). In this way, a plasmid and host may have aligned goals when facing a common threat such as viruses, however the plasmid may employ its own defensive mechanisms to ward off competing mobile genetic elements and ensure its continued maintenance (Rocha & Bikard 2022). Additional assessment of the pNOB8 CRISPR mini-array may help elucidate the plasmid-host relationship further. Finally, as the phrase suggests, an evolutionary arms-race may result in the selection of additional attack and defence mechanisms. Anti-anti-CRISPR proteins, which repress phage Acr expression, have been discovered in bacteria (Mohanraju *et al*. 2022), suggesting that interactions between invasive genetic elements and hosts are more complex than previously imagined, and therefore providing interesting avenues that await further investigation.

List of Abbreviations

°C – Degrees Celsius

µg – Microgram

µl – Microlitre

µm – Micrometre

µM – Micromolar

mg – Milligram

ml – Millilitre

mM – Millimolar

ng – Nanogram

nm - Nanometre

nM – Nanomolar

pmol – picomole

α – Alpha

β – Beta

Å – Angstrom

aa – amino acid

ATP – Adenosine triphosphate

ADP – Adenosine diphosphate

ATPase – Adenosine triphosphate hydrolase

AP – Alkaline phosphatase

APS – Ammonium persulphate

Amp – Ampicillin

AMPPCP – β,ɣ-methyleneadenosine 5‘-triphosphate

AMPPNP - Adenyl-imidodiphosphate

AFM – Atomic Force Microscopy

BLAST – Basic Local Alignment Search Tool

bp – base pair

BGG – Bovine Gamma Globulin

BS3 - bis[sulfosuccinimidyl] suberate

BSA – Bovine serum albumin

CBP – Centromere-binding protein

CD – Circular Dichroism

COG – Clusters of orthologous genes/groups of proteins

CTP – Cytidine triphosphate

CTPase - Cytidine triphosphate hydrolase

CRISPR – Clustered regularly interspaced short palindromic repeats

CV – Column volume

DDH – DNA-DNA hybridisation

DMP – Dimethyl pimelimidate

DNA – Deoxyribonucleic acid

dNTP – Deoxyribonucleoside triphosphate

EDTA – Ethylenediaminetetraacetic acid

EGTA - Ethylene Glycol-bis(β-aminoethyl ether)-N,N,N’N’-Tetraacetic acid

EMSA – Electrophoretic Mobility Shift Assay

EtOH - Ethanol

GTP – Guanosine triphosphate

GTPase – Guanosine triphosphate hydrolase

His – Histidine

HTH/wHTH – Helix-Turn-Helix/winged Helix-Turn-Helix

IHF – Integration Host Factor

kb – kilobase

kDa - kilodaltons

$K_d$app – Apparent dissociation constant

L – litre

LB – Luria-Bertani medium

M – Molar

Mbp – Mega base pairs

MCS – Multiple Cloning Site

ML – Maximum Likelihood

MST – Microscale Thermophoresis

Mw – Molecular weight

MWCO – Molecular weight cut-off

nt – nucleotide

NEB – New England Biolabs

OD – Optical Density

ORF – Open Reading Frame

P – Promoter

Par - Partition

PCR – Polymerase Chain Reaction

PDB – Protein Data Bank

Poly(dI-dC) – Poly(deoxyinosinic-deoxycytidylic) acid

RE – Restriction Enzyme

RHH – Ribbon-Helix-Helix

RMSD – Root mean square deviation

RPM – Revolutions per minute

SDS – Sodium Dodecyl Sulphate

SDS-PAGE - Sodium Dodecyl Sulphate-Polyacrylamide Gel Electrophoresis

SEC-MALLS – Size Exclusion Chromatography-Multi Angle Laser Light Scattering

TA – Toxin/Antitoxin

TAE – Tris base, Acetic acid, EDTA

TBE – Tris base, Boric acid, EDTA

TEMED - N,N,N’,N’-Tetramethylethylene-1,2-diamine

U – Enzymatic unit

UV - Ultraviolet

v/v – volume by volume

w/v – weight by volume

x g – multiplied by g (Relative Centrifugal Force)

References

Abeles, A. L. *et al*. 1985. Partition of Unit-copy Miniplasmids to Daughter Cells. III. The DNA Sequence and Functional Organisation of the P1 Partition Region. *Journal of Molecular Biology* 185, 261-272.

Alikhan, N. *et al.* 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 12**,** 402.

Alnaqshabandy, I. 2020. Molecular mechanisms and dynamics of the segregation of plasmid pB171 from an enteropathogenic strain of *Eschericia coli*. Unpublished PhD thesis, University of York.

Andes, S. N. *et al*. 2019. Ctp1 protein-DNA filaments promote DNA Bridging and DNA double-strand break repair. *Journal of Biological Chemistry* 294 (9), 3312-3320.

Ao, X. *et al*. 2013. The *Sulfolobus* Initiator Element Is an Important Contributor to Promoter Strength. *Journal of Bacteriology* 195 (22), 5216-5222.

Aravind, L. *et al*. 2005. The many faces of the helix-turn-helix domain: Transcription regulation and beyond. *FEMS Microbiology Reviews* 29, 231-262.

Arora, B. *et al*. 2017. Chemical Crosslinking: Role in Protein and Peptide Science. *Current Protein and Peptide Science* 18 (9), 946-955.

Auch, A.F. *et al*. 2010a. Digital DNA-DNA hybridisation for microbial species delineation by means of genome-to-genome sequence comparison. *Standards in Genomic Sciences* 2:117-134.

Auch, A.F. *et al*. 2010b. Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. *Standards in Genomic Sciences* 2:142-148.

Aylett, C. H. S. *et al*. 2010. Filament structure of bacterial tubulin homologue TubZ. *Proceedings of the National Academy of Sciences* 107 (45), 19766-19771.

Aylett, C. H. S. & Löwe, J. 2012. Superstructure of the centromeric complex of TubZR*C* plasmid partitioning systems. *Proceedings of the National Academy of Sciences* 109 (41), 16522-16527.

Austin, S. & Abeles, A. 1983. Partition of Unit-copy Miniplasmids to Daughter Cells. I. P1 and F Miniplasmids Contain Discrete, Interchangeable Sequences Sufficient to Promote Equipartition. *Journal of Molecular Biology* 169, 353-372.

Badrinarayanan, A. *et al*. 2015. Bacterial Chromosome Organisation and Segregation. *Annual Review of Cell and Developmental Biology* 31, 171-199.

Baek, J. H. *et al*. 2014. Chromosome Segregation Proteins of *Vibrio cholerae* as Transcription Regulators. *mBio* 5 (3), e01061-14.

Balaguer, F. *et al*. 2021. CTP promotes efficient ParB-dependent DNA condensation by facilitating one-dimensional diffusion from *parS*. *eLife* 10:e67554.

Balch, W. E. *et al.* 1977 An ancient divergence among the bacteria. *Journal of Molecular Evolution* 9**,** 305-311.

Barillà, D. and Hayes, F. 2003. Architecture of the ParF-ParG protein complex involved in prokaryotic DNA segregation. *Molecular Microbiology* 49 (2), 487-499.

Barillà, D. *et al*. 2007. The tail of the ParG DNA segregation protein remodels ParF polymers and enhances ATP hydrolysis via an arginine finger-like motif. *Proceedings of the National Academy of Sciences* 104 (6), 1811-1816.

Barillà, D. 2010. One-way ticket to the cell pole: Plasmid transport by the prokaryotic tubulin homolog TubZ. *Proceedings of the National Academy of Sciences* 107 (27), 12061-12062.

Barillà, D. 2016. Driving Apart and Segregating Genomes in Archaea. *Trends in Microbiology* 24 (12), 957-967.

Barrangou, R. and Marraffini, L.A. 2014. CRISPR-Cas systems: prokaryotes upgrade to adaptive immunity. *Molecular Cell* 54 (2), 234-244.

Bartosik, A. A. *et al.* 2004. ParB of *Pseudomonas aeruginosa*: Interactions with Its Partner ParA and Its Target *parS* and Specific Effects on Bacterial Growth. *Journal of Bacteriology* 186 (20), 6983-6998.

Baxter, J.C. & Funnell, B.E. 2014. Plasmid Partition Mechanisms. *Microbiology Spectrum* 2 (6), PLAS-0023-2014.

Barge, M. T. 2015. Role of the unstructured N-terminus of the centromere binding protein ParG in mediating segregation of the multidrug resistance plasmid TP228. Unpublished PhD thesis, University of York.

Bell, S.D. & Jackson, P.J. 2000. The Role of Transcription Factor B in Transcription Initiation and Promoter Clearance in the Archaeon *Sulfolobus acidocaldarius*. *The Journal of Biological Chemistry* 275 (17), 12934-12940.

Bennett, P.M. 2008. Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria. *British Journal of Pharmacology* 153, S347-S357.

Bernander, R. 2000. Chromosome replication, nucleoid segregation and cell division in Archaea. *Trends in Microbiology* 8 (6), 278-283.

Bernander, R. 2007. The archaeal cell cycle: current issues. *Molecular Microbiology* 48 (3), 599-604.

Bharat, T. A. *et al.* 2015. Structures of actin-like ParM filaments show architecture of plasmid-segregating spindles. *Nature* 523, 106-110.

Bouet, J-Y. *et al*. 2007. Polymerisation of SopA partition ATPase: regulation by DNA binding and SopB. *Molecular Microbiology* 63 (2), 468-481.

Bouet, J-Y. & Funnell, B.E. 2019. Plasmid Localisation and Partition in *Enterobacteriaceae*. *EcoSal Plus* 8 (2) doi: 10.1128/ecosalplus.ESP-0003-2019.

Breier, A. M. and Grossman, A. D. 2007. Whole-genome analysis of the chromosome partitioning and sporulation protein Spo0J (ParB) reveals spreading and origin-distal sites on the *Bacillus subtilis* chromosome. *Molecular Microbiology* 64 (3), 703-718.

Breüner, A. *et al.* 1996. The centromere-like *parC* locus of plasmid R1. *Molecular Microbiology* 20 (3), 581-592.

Brock, T. D. *et al*. 1972. *Sulfolobus:* A New Genus of Sulfur-Oxidising Bacteria Living at Low pH and High Temperature. *Archives of Microbiology* 84, 54-68.

Broedersz, C. P. *et al*. 2014. Condensation and localisation of the partitioning protein ParB on the bacterial chromosome. *Proceedings of the National Academy of Sciences* 111 (24), 8809-8814.

Buchan, D.W.A. & Jones, D.T. 2019. The PSIPRED Protein Analysis Workbench:20 years on. *Nucleic Acids Research* 47 (W1), W402-W407.

Campbell, C. S & Mullins, R. D. 2007. In vivo visualisation of type II plasmid segregation: bacterial actin filaments pushing plasmids. *The Journal of Cell Biology* 179 (5), 1059-1066.

Carmelo, E. *et al*. 2005. The Unstructured N-terminal Tail of ParG Modulates Assembly of a Quaternary Nucleoprotein Complex in Transcription Repression. *The Journal of Biological Chemistry* 280 (31), 28683-28691.

Carver, T. *et al*. 2009. DNAPlotter: circular and linear interactive genome visualisation. *Bioinformatics* 25 (1), 119-120.

Chen, B-W. *et al*. 2015. Insights into ParB spreading from the complex structure of Spo0J and *parS*. *Proceedings of the National Academy of Sciences* 112 (21), 6613-6618.

Chou-Zheng, L. and Hatoum-Aslan, A. 2019. A type III-A CRISPR-Cas system employs degradosome nucleases to ensure robust immunity. *eLife* 8:e45393.

Contreras-Martos, S. *et al*. 2018. Quantification of Intrinsically Disordered Proteins: A Problem Not Fully Appreciated. *Frontiers in Molecular Biosciences* 5:83.

Czerny, T. *et al.* 1993. DNA sequence recognition by Pax proteins: bipartite structure of the paired domain and its binding site. *Genes and Development* 7, 2048-2061.

Dai, X. *et al.* 2016. Genome Sequencing of *Sulfolobus* sp. A20 from Costa Rica and Comparative Analyses of the putative Pathways of Carbon, Nitrogen and Sulfur Metabolism in Various *Sulfolobus* Strains. *Frontiers in Microbiology* 7:1902.

Debaugny, R. E. *et al*. 2018. A conserved mechanism drives partition complex assembly on bacterial chromosomes and plasmids. *Molecular Systems Biology* 14: e8516.

DeLano, W. L. 2002. The PyMOL Molecular Graphics System, Version 2.4.1 Schrödinger, LLC.

Delbrück, H. *et al*. 2002. An Src Homology 3-like Domain Is Responsible for Dimerisation of the Repressor Protein KorB Encoded by the Promiscuous IncP Plasmid RP4. *The Journal of Biological Chemistry* 277 (6), 4191-4198.

DeLong, E. F. 1992. Archaea in coastal marine environments. *Proceedings of the National Academy of Sciences* 89, 5685-5689.

Del Solar, G. & Espinosa, M. 2000. Plasmid copy number control: an ever-growing story. *Molecular Microbiology* 37 (3), 492-500.

Del Solar, G. *et al.* 2002. A Genetically Economical Family of Plasmid-Encoded Transcriptional Repressors Involved in Control of Plasmid Copy Number. *Journal of Bacteriology* 184 (18), 4943-4951.

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

Dmowski, M. & Jagura-Burdzy, G. 2013. Active Stable Maintenance Functions in Low Copy-Number Plasmids of Gram-Positive Bacteria I. Partition Systems. *Polish Journal of Microbiology* 62 (1), 3-16.

Dombrowski, N. *et al*. 2019. Genomic diversity, lifestyles and evolutionary origins of DPANN archaea. *FEMS Microbiology Letters* 366, fnz008.

Duggin, I.G. *et al*. 2011. Replication termination and chromosome dimer resolution in the archaeon *Sulfolobus solfataricus*. *The EMBO Journal* 30, 145-153.

Dunham, T. D. *et al*. 2009. Structural basis for ADP-mediated transcriptional regulation by P1 and P7 ParA. *The EMBO Journal* 28, 1792-1802.

Durkacz, B.W. & Sherratt, D.J. 1973. Segregation Kinetics of Colicinogenic Factor Col E1 from a Bacterial Population Temperature Sensitive for DNA Polymerase I. *Molecular and General Genetics* 121, 71-75.

Dyson, H. J. 2016. Making Sense of Intrinsically Disordered Proteins. *Biophysical Journal* 110, 1013-1016.

Ebersbach, G. & Gerdes, K. 2004. Bacterial mitosis: partitioning protein ParA oscillates in spiral-shaped structures and positions plasmids at mid-cell. *Molecular Microbiology* 52 (2), 385-398.

Erdmann, N. *et al*. 1999. Intracellular localisation of P1 ParB protein depends on ParA and *parS*. *Proceedings of the National Academy of Sciences* 96 (26), 14905-14910.

Erdös, G. & Dosztányi, Z. 2020. Analysing protein disorder with IUPred2A. *Current Protocols in Bioinformatics*. 70, e99. doi: 10.1002/cpbi.99

Feng, X. *et* al. 2018. A transcriptional factor B paralog functions as an activator to DNA damage-responsive expression in archaea. *Nucleic Acids Research* 46 (14), 7085-7096.

Fink, G. & Löwe, J. 2015. Reconstitution of a prokaryotic minus end-tracking system using TubRC centromeric complexes and tubulin-like protein TubZ filaments. *Proceedings of the National Academy of Sciences*, E1845-E1850.

Firth, N. & Skurray, R. 1992. Characterisation of the F plasmid bifunctional conjugation gene, *traG*. *Molecular and General Genetics* 232, 145-153.

Firth, N *et al*. 2000. Replication of Staphylococcal Multiresistance Plasmids. *Journal of Bacteriology* 182 (8), 2170-2178.

Fisher, G. L. M. *et al*. 2017. The structural basis for dynamic DNA binding and bridging interactions which condense the bacterial centromere. *eLIFE* 6:28086.

Folta-Stogniew, E. & Williams, K. R. 1999. Determination of Molecular Masses of Proteins in Solution: Implementation of an HPLC Size Exclusion Chromatography and Laser Light Scattering Service in a Core Laboratory. *Journal of Biomolecular Techniques* 10 (2), 51-63.

Fox, G.E. *et al*. 1977 Classification of methanogenic bacteria by 16S ribosomal RNA characterization. *Proceedings of the National Academy of Sciences of the United States of America* 74 (10), 4537–4541.

Friedman, S.A. & Austin, S. J. 1988. The P1 Plasmid-Partition System Synthesises Two Essential Proteins from an Autoregulated Operon. *Plasmid* 19, 103-112.

Fuentes-Pérez, M. E. *et al*. 2017. TubZ filament assembly dynamics requires the flexible C-terminal tail. *Nature Scientific Reports* 7:43342.

Fuhrman, J. A. *et al*. 1992. Novel major archaebacterial group from marine plankton. *Nature* 356, 148-149.

Funnell, B. E. 1988a. Mini-P1 Plasmid Partitioning: Excess ParB Protein Destabilizes Plasmids Containing the Centromere *parS*. *Journal of Bacteriology* 170 (2), 954-960.

Funnell, B. E. 1988b. Participation of *Escherichia coli* integration host factor in the P1 plasmid partition system. *Proceedings of the National Academy of Sciences of the United States of America* 85, 6657-6661.

Funnell, B. E. 2005. Partition-mediated plasmid pairing. *Plasmid* 53, 119-125.

Funnell, B. E. 2016. ParB Partition Proteins: Complex Formation and Spreading at Bacterial and Plasmid Centromeres. *Frontiers in Molecular Biosciences* 3:44, doi:10.3389.

Gabler, F. *et al*. 2020. Protein sequence analysis using the MPI bioinformatics toolkit. *Current Protocols in Immunology* 72, e108.

Gadagkar, S.R. *et al*. 2005. Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *Journal of Experimental Zoology* 304B:64-74.

Galkin, V. E. *et al.* 2009. Structural Polymorphism of the ParM Filament and Dynamic Instability. *Structure* 17 (9), 1253-1264.

Galperin, M.Y. *et al*. 2014. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Research* 43, D261-D269.

Galperin, M.Y. *et al*. 2019. Microbial genome analysis: the COG approach. *Briefings in Bioinformatics* 20 (4), 1063-1070.

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

Gayathri, P. *et al.* 2012. A Bipolar Spindle of Antiparallel ParM Filaments Drives Bacterial Plasmid Segregation. *Science* 338, 1334-1337.

Gerdes, K. *et al*. 1986. Unique type of plasmid maintenance function: Postsegregational killing of plasmid-free cells. *Proceedings of the National Academy of Sciences* 83, 3116-3120.

Gerdes, K. *et al*. 2005. Prokaryotic Toxin-Antitoxin Stress Response Loci. *Nature Reviews Microbiology* 3, 371-382.

Gerdes, K. & Wagner, E.G.H. 2007. RNA antitoxins. *Current Opinion in Microbiology* 10, 117-124.

Gerdes, K. *et al*. 2021. Phylogeny Reveals Novel HipA-Homologous Kinase Families and Toxin-Antitoxin Gene Organisations. *mBio* 12:e01058-21.

Golovanov, A. P. *et al*. 2003. ParG, a protein required for active partition of bacterial plasmids, has a dimeric ribbon-helix-helix structure. *Molecular Microbiology* 50 (4), 1141-1153.

Gudbergsdottir, S. *et al*. 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), 35-49.

Graham, T. G. W. *et al*. 2014. ParB spreading requires DNA bridging. *Genes & Development* 28 (11), 1228-1238.

Greenfield, N. J. 2006. Using circular dichroism spectra to estimate protein secondary structure. *Nature Protocols* 1 (6), 2876-2890.

Greve, B. *et al*. 2004. Genomic comparison of archaeal conjugative plasmids from *Sulfolobus*. *Archaea* 1, 231-239.

Grigoriev, P. S. & Lobocka, M. B. 2001. Determinants of segregational stability of the linear plasmid-prophage N15 of *Escherichia coli*. *Molecular Microbiology* 42 (2), 355-368.

Gruber, S. & Errington, J. 2009. Recruitment of Condensin to Replication Origin Regions by ParB/Spo0J Promotes Chromosome Segregation in *B. subtilis*. *Cell* 137, 685-696.

Guo, L. *et al.* 2008. Biochemical and structural characterisation of Cren7, a novel chromatin protein conserved among Crenarchaea. *Nucleic Acids Research,* 36 (4), 1129-1137.

Guy, L. & Ettema, T. J. G. 2011. The archaeal 'TACK' superphylum and the origin of eukaryotes. *Trends in Microbiology* 19 (12), 580-587.

Guynet, C. *et al*. 2011. The *stb* Operon Balances the Requirements for Vegetative Stability and Conjugative Transfer of Plasmid R388. *PLoS Genetics* 7 (5), e1002073.

Guynet, C. & de la Cruz, F. 2011. Plasmid segregation without partition. *Mobile Genetic Elements* 1:3, 236-241.

Hall, B. G. 2013. Building Phylogenetic Trees from Molecular Data with MEGA. *Molecular Biology and Evolution* 30 (S):1229-1235.

Handa, H. 2008. Linear plasmids in plant mitochondria: Peaceful coexistence or malicious invasions? *Mitochondrion* 8 (1), 15-25.

Hansen, J.C. *et al*. 2006. Intrinsic Protein Disorder, Amino Acid Composition, and Histone Terminal Domains. *The Journal of Biological Chemistry* 281 (4), 1853-1856.

Haring, M. *et al.* 2004. Morphology and genome organisation of the virus PSV of the hyperthermophilic archaeal genera *Pyrobaculum* and *Thermoproteus*: a novel virus family, the *Globuloviridae*. *Virology* 323 (2), 233-242.

Hassler, M., Shaltiel, I.A., & Haering, C.H. 2018. Towards a Unified Model of SMC Complex Formation. *Current Biology* 28, R1266-R1281.

Hatano, T. & Niki, H. 2010. Partitioning of P1 plasmids by gradual distribution of the ATPase ParA. *Molecular Microbiology* 78 (5), 1182-1198.

Hayes, F. & Austin, S. 1994. Topological Scanning of the P1 Plasmid Partition Site. *Journal of Molecular Biology* 243, 190-198.

Hayes, F. 2000. The partition system of multidrug resistance plasmid TP228 includes a novel protein that epitomizes an evolutionarily distinct subgroup of the ParA superfamily. *Molecular Microbiology* 37 (3), 528-541.

Hayes, F. 2003. Toxins-Antitoxins: Plasmid Maintenance, Programmed Cell Death, and Cell Cycle Arrest. *Science* 301, 1496-1499.

Hayes, F. and Barillà, D. 2006a. The bacterial segrosome: a dynamic nucleoprotein machine for DNA trafficking and segregation. *Nature Reviews Microbiology* 4, 133-143.

Hayes, F. and Barillà, D. 2006b. Assembling the bacterial segrosome. *TRENDS in Biochemical Sciences* 31, 247-250.

He, F. *et al*. 2018. Anti-CRISPR proteins encoded by archaeal lytic viruses inhibit subtype I-D immunity. *Nature Microbiology* 3, 461-469.

Heuer, H. & Smalla, K. 2012. Plasmids foster diversification and adaptation of bacterial populations in soil. *FEMS Microbiology Reviews* 36, 1083-1104.

Hirano, M. *et al.* 1998. Autoregulation of the partition genes of the mini-F plasmid and the intracellular localisation of their products in *Escherichia coli*. *Molecular & General Genetics* 257 (4), 392-403.

Hoischen, C. 2008. *Escherichia coli* low-copy-number plasmid R1 centromere *parC* forms a U-shaped complex with its binding protein ParR. *Nucleic Acids Research* 36 (2), 607-615.

Howlader, M. T. H. *et al*. 2010. Alanine Scanning Analyses of the Three Major Loops in Domain II of *Bacillus thuringiensis* Mosquitocidal Toxin Cry4Aa. *Applied and Environmental Microbiology* 76, 860-865.

Hsu, T-M. & Chang, Y-R. 2019. High-Copy-Number Plasmid Segregation - Single-Molecule Dynamics in Single Cells. *Biophysical Journal* 116, 772-780.

Hülter, N. *et al*. 2017. An evolutionary perspective on plasmid lifestyle modes. *Current Opinion in Microbiology* 38, 74-80.

Hwang, L. C. *et al*. 2013. ParA-mediated plasmid partition driven by protein pattern self-organisation. *The EMBO Journal* 32, 1238-1249.

Iranzo, J. *et al*. 2013. Evolutionary Dynamics of the Prokaryotic Adaptive Immunity System CRISPR-Cas in an Explicit Biological Context. *Journal of Bacteriology* 195 (17), 3834-3844.

Ireton, K. *et al*. 1994. *spo0j* Is Required for Normal Chromosome Segregation as well as the Initiation of Sporulation in *Bacillus subtilis*. *Journal of Bacteriology* 176 (17), 5320-5329.

Jacob, A. E. & Hobbs, S.J. 1974. Conjugal Transfer of Plasmid-Borne Multiple Antibiotic Resistance in *Streptococcus faecalis* var. *zymogenes*. *Journal of Bacteriology* 117 (2), 360-372.

Jalal, A. S. B. *et al*. 2020a. Diversification of DNA-Binding Specificity by Permissive and Specificity-Switching Mutations in the ParB/Noc Protein Family. *Cell Reports* 32, 107928.

Jalal, A. S. B. *et al*. 2020b. ParB spreading on DNA requires cytidine triphosphate in vitro. *eLife* 9:e53515.

Jalal, A. S. B. *et al*. 2021. A CTP-dependent gating mechanism enables ParB spreading on DNA. *eLife* 10:e69676.

Jecz, P. *et al*. 2015. A Single *parS* Sequence from the Cluster of Four Sites Closest to *oriC* Is Necessary and Sufficient for Proper Chromosome Segregation in *Pseudomonas aeruginosa*. *PLoS ONE* 10 (3): e0120867.

Jensen, R. B. *et al*. 1994. Partitioning of Plasmid R1. The *parA* Operon is Autoregulated by ParR and Its Transcription is Highly Stimulated by a Downstream Activating Element. *Journal of Molecular Biology* 236, 1299-1309.

Jensen, R. B. & Gerdes, K. 1997. Partitioning of Plasmid R1. The ParM Protein Exhibits ATPase Activity and Interacts with the Centromere-like ParR-*parC* Complex. *Journal of Molecular Biology* 269, 505-513.

Jensen, R. B. *et al*. 1998. Mechanism of DNA segregation in prokaryotes: Replicon pairing by *parC* of plasmid R1. *Proceedings of the National Academy of Sciences USA* 95, 8550-8555.

Jindal, L. & Emberly, E. 2019. DNA segregation under Par Protein control. *PLoS ONE* 14 (7): e0218520.

Jumper, J. *et al*. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583-589.

Kalliomaa-Sanford, A.K. *et al*. 2012. Chromosome segregation in Archaea mediated by a hybrid DNA partition machine. *Proceedings of the National Academy of Sciences* 109 (10), 3754-3759.

Kamada, K. & Barillà, D. 2018. Combing Chromosomal DNA Mediated by the SMC Complex: Structure and Mechanisms. *BioEssays* 40, 1700166.

Kandler, O. and Hippe, H. 1977 Lack of peptidoglycan in the cell walls of *Methanosarcina barkeri*. *Archives of Microbiology,* 113**,** 57-60.

Kędzierska, B. & Hayes, F. 2016. Emerging Roles of Toxin-Antitoxin Modules in Bacterial Pathogenesis. *Molecules* 21 (6):790.

Kelley, L.A. *et al.* 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols* 10 (6), 845-858.

Könneke, M. *et al*. 2005. Isolation of an autotrophic ammonia-oxidising marine archaeon. *Nature* 437, 543-546.

Kozakov, D. *et al*. 2017. The ClusPro web server for protein-protein docking. *Nature Protocols* 12 (2), 255-278.

Krishna, S. S. *et al*. 2003. Structural classification of zinc fingers. *Nucleic Acids Research* 31 (2), 532-550.

Krissinal, E. & Henrick, K. 2004. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica* D60, 2256-2268.

Kulińska, A. *et al*. 2011. The Centromere Site of the Segregation Cassette of Broad-Host-Range Plasmid RA3 Is Located at the Border of the Maintenance and Conjugative Transfer Modules. *Applied and Environmental Microbiology* 77 (7), 2414-2427.

Kusiak, M. *et al*. 2011. Binding and Spreading of ParB on DNA Determine Its Biological Function in *Pseudomonas aeruginosa*. *Journal of Bacteriology* 193 (13), 3342-3355.

Larsen, R. A. *et al.* 2007. Treadmilling of a prokaryotic tubulin-like protein, TubZ, required for plasmid stability in *Bacillus thuringiensis*. *Genes & Development* 21:1340-1352.

Laurens, N. *et al*. 2012. Alba shapes the archaeal genome using a delicate balance of bridging and stiffening the DNA. *Nature Communications* 3:1328.

Lee, J. *et al*. 2018. Potent Cas9 Inhibition in Bacterial and Human Cells by AcrIIC4 and AcrIIC5 Anti-CRISPR Proteins. *mBio* 9:e02321-18.

Lee, P. S. & Grossman, A. D. 2006. The chromosome partitioning proteins Soj (ParA) and Spo0J (ParB) contribute to accurate chromosome partitioning, separation of replicated sister origins, and regulation of replication initiation in *Bacillus subtilis*. *Molecular Microbiology* 60 (4), 853-869.

León, L. M. *et al*. 2021. Mobile element warfare via CRISPR and anti-CRSIPR in *Pseudomonas aeruginosa*. *Nucleic Acids Research* 49 (4), 2114-2125.

Leonard, T. A. *et al.* 2004. Structural analysis of the chromosome segregation protein Spo0J from *Thermus thermophilus*. *Molecular Microbiology* 53 (2), 419-432.

Leonard, T. A. *et al.* 2005. Bacterial chromosome segregation: structure and DNA binding of the Soj dimer – a conserved biological switch. *The EMBO Journal* 24, 270-282.

Liu, G. 2015. Studying Extrachromosomal Genetic Elements in *Sulfolobus*. PhD thesis, University of Copenhagen.

Liu, Y. *et al.* 2021. Expanded diversity of Asgard archaea and their relationship with eukaryotes. *Nature* 593, 553-579.

Li, M. *et al*. 2017. The spacer size of I-B CRISPR is modulated by the terminal sequence of the protospacer. *Nucleic Acids Research* 45, 4642-4654.

Li, Q. *et al*. 2019. The Role of Plasmids in the Multiple Antibiotic Resistance Transfer in ESBLs-Producing *Escherichia coli* Isolated From Wastewater Treatment Plants. *Frontiers in Microbiology* 10:633, doi:10.3389.

Li, Y. *et al*. 2016. Harnessing Type I and Type III CRISPR-Cas systems for genome editing. *Nucleic Acids Research* 44 (4), e34.

Liang, S.D. *et al*. 1996. DNA Sequence Preferences of GAL4 and PPR1: How a Subset of Zn_2Cys_6 Binuclear Cluster Proteins Recognizes DNA. *Molecular and Cellular Biology* 16 (7), 3773-3780.

Lillestøl, R.K. *et al*. 2006. A putative viral defense mechanism in archaeal cells. *Archaea* 2, 59-72.

Lillestøl, R.K. *et al*. 2009. CRISPR families of the crenarchaeal genus *Sulfolobus*: bidirectional transcription and dynamic properties. *Molecular Microbiology* 72 (1), 259-272.

Lim, H. C. *et al*. 2014. Evidence for a DNA-relay mechanism in ParABS-mediated chromosome segregation. *eLIFE* 3:e02758.

Lin, D. C. & Grossman, A. D. 1998. Identification and Characterisation of a Bacterial Chromosome Partitioning Site. *Cell* 92, 675-685.

Liu, G. *et al.* 2016. Diverse CRISPR-Cas responses and dramatic cellular DNA changes and cell death in pKEF9-conjugated *Sulfolobus* species. *Nucleic Acids Research* 44 (9), 4233-4242.

Livney, J. *et al*. 2007. Distribution of Centromere-Like *parS* Sites in Bacteria: Insights from Comparative Genomics. *Journal of Bacteriology* 189 (23), 8693-8703.

Lu, W. *et al*. 2021. Anti-CRISPR AcrIF9 functions by inducing the CRISPR-Cas complex to bind DNA non-specifically. *Nucleic Acids Research* 49 (6), 3381-3393.

Lundgren, M. *et al*. 2004. Three replication origins in *Sulfolobus* species: Synchronous initiation of chromosome replication and asynchronous termination. *Proceedings of the National Academy of Sciences* 101 (18), 7046-7051.

Luo, X. *et al*. 2007. CC1, a Novel Crenarchaeal DNA Binding Protein. *Journal of Bacteriology* 189 (2), 403-409.

Lynch, A. S. and Wang, J. C. 1995. SopB protein-mediated silencing of genes linked to the *sopC* locus of *Escherichia coli* F plasmid. *Proceedings of the National Academy of Sciences USA* 92, 1896-1900.

Makarova, K. S. *et al*. 2011. Evolution and classification of the CRISPR-Cas systems. *Nature Reviews Microbiology* 9, 467-476.

Makarova, K.S. *et al*. 2015a. Archaeal Clusters of Orthologous Genes (arCOGs): An update and Application for Analysis of Shared Features between Thermococcales, Methanococcales, and Methanobacteriales. *Life* 5, 818-840.

Makarova, K. S. *et al*. 2015b. An updated evolutionary classification of CRISPR-Cas systems. *Nature Reviews Microbiology* 13, 722-736.

Marcy, Y. *et al*. 2007. Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proceedings of the National Academy of Sciences* 104 (29), 11889-11894.

Marston, A. L. & Errington, J. 1999. Dynamic Movement of the ParA-like Soj Protein of *B. subtilis* and Its Dual Role in Nucleoid Organisation and Developmental Regulation. *Molecular Cell* 4, 673-682.

Martin-Garcia, B. *et al*. 2018. The TubR-centromere complex adopts a double-ring segrosome structure in Type III partition systems. *Nucleic Acids Research* 46 (11), 5704-5716.

Maruyama, H. *et al.* 2020. Different Proteins Mediate Step-wise Chromosome Architectures in *Thermoplasma acidophilum* and *Pyrobaculum calidifontis*. *Frontiers in Microbiology* 11:1247.

Maxwell, K. L. *et al*. 1999. A simple in vivo assay for increased protein solubility. *Protein Science* 8, 1908-1911.

McAfee, J. G. *et al*. Equilibrium DNA binding of Sac7d Protein from the Hyperthermophile *Sulfolobus acidocaldarius:* Fluorescence and Circular Dichroism Studies. *Biochemistry* 35, 4034-4045.

McLeod, B. N. *et al*. 2017. A three-dimensional ParF meshwork assembles through the nucleoid to mediate plasmid segregation. *Nucleic Acids Research* 45 (6), 3158-3171.

McNicholas, S. *et al*. 2011. Presenting your structures: the CCP4mg molecular-graphics software. *Acta Crystallographica* D67, 386-394.

Medvedeva, S. *et al*. 2019. Virus-borne mini-CRISPR arrays are involved in interviral conflict. *Nature Communications* 10:5204.

Meier-Kolthoff, J.P. *et al*. 2013 Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 14:60.

Micorescu, M. *et al*. 2008. Archaeal Transcription: Function of an Alternative Transcription Factor B from *Pyrococcus furiosus*. *Journal of Bacteriology* 190 (1), 157-167.

Micsonai, A. *et al*. 2015. Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy. *Proceedings of the National Academy of Sciences* E3095-E3103.

Micsonai, A. *et al*. 2018. BeStSel: a web server for accurate protein secondary structure prediction and fold recognition from the circular dichroism spectra. *Nucleic Acids Research* 46, W315-W322.

Million-Weaver, S. & Camps, M. 2014. Mechanisms of Plasmid Segregation: Have Multicopy Plasmids Been Overlooked? *Plasmid* 75, 27-36.

Minnen, A. *et al*. 2016. Control of Smc Coiled Coil Architecture by the ATPase Heads Facilitates Targeting to Chromosomal ParB/*parS* and Release onto Flanking DNA. *Cell Reports* 14, 2003-2016.

Mintseris, J. and Gygi, S. P. 2020. High-density chemical cross-linking for modeling protein interactions. *Proceedings of the National Academy of Sciences* 117 (1),93-102.

Mirdita, M. *et al*. 2021. ColabFold – Making protein folding accessible to all. bioRxiv doi:10.1101/2021.08.15.456425.

Misra, H. S. *et al*. 2018. Maintenance of multipartite genome system and its functional significance in bacteria. *Journal of Genetics* 97 (4), 1013-1038.

Mohanraju, P. *et al*. 2022. Alternative functions of CRISPR-Cas systems in the evolutionary arms race. *Nature Review Microbiology* 20, 351-364.

Mohl, D. A. & Gober, J. W. 1997. Cell Cycle-Dependent Polar Localisation of Chromosome Partitioning Proteins in Caulobacter Crescentus. *Cell* 88, 675-684.

Møller-Jensen, J. *et al*. 2002. Prokaryotic DNA segregation by an actin-like filament. *The EMBO Journal* 21 (12), 3119-3127.

Møller-Jensen, J. *et al*. 2003. Bacterial Mitosis : ParM of Plasmid R1 Moves Plasmid DNA by an Actin-like Insertional Polymerisation Mechanism. *Molecular Cell* 12, 1477-1487.

Møller-Jensen, J. *et al*. 2007. Structural analysis of the ParR/*parC* plasmid partition complex. *The EMBO Journal* 26, 4413-4422.

Mori, H. *et al*. 1986. Structure and Function of the F Plasmid Genes Essential for Partitioning. *Journal of Molecular Biology* 192, 1-15.

Mori, H. *et al.* 1989. Purification and Characterisation of SopA and SopB Proteins Essential for F Plasmid Partitioning. *The Journal of Biological Chemistry* 26, 15535-15541.

Murray, H. and Errington, J. 2008. Dynamic Control of the DNA Replication Initiation Protein DnaA by Soj/ParA. *Cell* 135, 74-84.

Murugesapillai, D. *et al*. 2014. DNA bridging and looping by HMO1 provides a mechanism for stabilising nucleosome-free chromatin. *Nucleic Acids Research* 42 (14), 8996-9004.

Nakamura, Y. 2018. Prediction of Horizontally and Widely Transferred Genes in Prokaryotes. *Evolutionary Bioinformatics* 14, 1-15.

Nasmyth, K. 2002. Segregating sister genomes: the molecular biology of chromosome separation. *Science* 297, 559-565.

Ni, L. *et al*. 2010. Plasmid protein TubR uses a distinct mode of HTH-DNA binding and recruits the prokaryotic tubulin homolog TubZ to effect DNA partition. *Proceedings of the National Academy of Sciences* 107 (26), 11763-11768.

Niazi, A. *et al*. 2014. Genome Analysis of *Bacillus amyloliquefaciens* Subsp. *plantarum* UCMB5113: A Rhizobacterium That Improves Plant Growth and Stress Management. *PLOS ONE* 9 (8), e104651.

Nishida, H. & Oshima, T. 2017. Archaeal histone distribution is associated with archaeal genome base composition. *Journal of General Applied Microbiology* 63, 28-35.

O'Connell, M. R. 2018. Molecular Mechanisms of RNA Targeting by Cas13-containing Type VI CRISPR-Cas Systems. *Journal of Molecular Biology* 431, 66-87.

Ogden, T.H & Rosenberg, M.S. 2006. Multiple sequence alignment accuracy and phylogenetic interference. *Systematic Biology* 55 (2), 314-328.

Ogura, T. & Hiraga, S. 1983. Mini-F plasmid genes that couple host cell division to plasmid proliferation. *Proceedings of the National Academy of Sciences* 80, 4784-4788.

Oliva, M. A. 2016. Segrosome Complex Formation during DNA Trafficking in Bacterial Cell Division. *Frontiers in Molecular Biosciences* 3:51 10.3389.

Ondov, B. D. *et al*. 2016. Mash: fast genome and metagenome estimation using MinHash. *Genome Biology* 17:132.

Osorio-Valeriano, M. *et al*. 2019. ParB-type DNA Segregation Proteins Are CTP-Dependent Molecular Switches. *Cell* 179, 1512-1524.

Osorio-Valeriano, M. *et al*. 2021. The CTPase activity of ParB determines the size and dynamics of prokaryotic DNA partition complexes. *Molecular Cell* 81, 3992-4007.

Page, R. & Peti, W. 2016. Toxin-antitoxin systems in bacterial growth arrest and persistence. *Nature Chemical Biology* 12, 208-214.

Pang, D. *et* al. 2015. DNA studies using atomic force microscopy: capabilities for measurement of short DNA fragments. Frontiers in Molecular Biosciences, 2 (1), 1-7.

Park, S. C. *et al.* 2017. Structural basis of effector and operator recognition by the phenolic acid-responsive transcriptional repressor PadR. *Nucleic Acids Research* 45 (22), 13080-13093.

Peng, N. *et al*. 2011. Archaeal promoter architecture and mechanism of gene activation. *Biochemical Society Transactions* 39, 99-103.

Peng, W. *et al*. 2013. Genetic determinants of PAM-dependent DNA targeting and pre-crRNA processing in *Sulfolobus islandicus*. *RNA Biology* 10:5, 738-748.

Peng, W. *et al*. 2015. An archaeal CRISPR type III-B system exhibiting distinctive RNA targeting features and mediating dual RNA and DNA interference. *Nucleic Acids Research* 43 (1), 406-417.

Peng, X. 2008. Evidence for the horizontal transfer of an integrase gene from a fusellovirus to a pRN-like plasmid within a single strain of *Sulfolobus* and the implications for plasmid survival. *Microbiology* 154, 383-391.

Peng, X. *et al.* 2020. Anti-CRISPR Proteins in Archaea. *Trends in Microbiology* 28 (11), 913-921.

Pilla, G. & Tang, C.M. 2018. Going around in circles: virulence plasmids in enteric pathogens. *Nature Reviews Microbiology* 16, 484-495.

Pinilla-Redondo, R. *et al*. 2020a. Type IV CRISPR-Cas systems are highly diverse and involved in competition between plasmids. *Nucleic Acids Research* 48 (4), 2000-2012.

Pinilla-Redondo, R. *et al*. 2020b. Discovery of multiple anti-CRISPRs highlights anti-defense gene clustering in mobile genetic elements. *Nature Communications* 11:5652.

Pinilla-Redondo, R. *et al*. 2022. CRISPR-Cas systems are widespread accessory elements across bacterial and archaeal plasmids. *Nucleic Acids Research* 6 (50), 4315-4328.

Prangishvili, D. *et al*. 1998. Conjugation in Archaea : Frequent Occurrence of Conjugative Plasmids in *Sulfolobus*. *Plasmid* 40, 190-202.

Pratto, F. *et al*. 2009. Single-molecule Analysis of Protein-DNA Complexes Formed during Partition of Newly Replicated Plasmid Molecules in *Streptococcus pyogenes*. *The Journal of Biological Chemistry* 284 (44), 30298-30306.

Ptacin, J. L. *et al.* 2010. A spindle-like apparatus guides bacterial chromosome segregation. *Nature Cell Biology* 12 (8), 791-798.

Pyne, A. *et al*. 2014. Single-Molecule Reconstruction of Oligonucleotide Secondary Structure by Atomic Force Microscopy. *Small* 10 (16), 3257-3261.

Ramage, H.R. *et al*. 2009. Comprehensive Functional Analysis of *Mycobacterium tuberculosis* Toxin-Antitoxin Systems: Implications for Pathogenesis, Stress Responses, and Evolution. *PLoS Genetics* 5 (12):e1000767.

Ravin, N. V. *et al*. 2003. Mapping of Functional Domains in F Plasmid Partition Proteins Reveals a Bipartite SopB-recognition Domain in SopA. *Journal of Molecular Biology* 329, 875-889.

Reyes-Lamothe, R. *et al*. 2014. High-copy number plasmids diffuse in the nucleoid-free space, replicate stochastically and are randomly partitioned at cell division. *Nucleic Acids Research* 42 (2), 1042-10151.

Ringgaard, S. *et al.* 2007a. Regulatory Cross-talk in the Double *par* Locus of Plasmid pB171. *The Journal of Biological Chemistry* 282 (5), 3134-3145.

Ringgaard, S. *et al*. 2007b. Centromere Pairing by a Plasmid-encoded Type I ParB Protein. *The Journal of Biological Chemistry* 282 (38), 28216-28225.

Ringgaard, S. *et al*. 2009. Movement and equipositioning of plasmids by ParA filament dissembly. *Proceedings of the National Academy of Sciences* 106 (46), 19369-19374.

Rinke, C. *et al*. 2013. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499, 431-437.

Rivera, C. R. *et al.* 2011. Architecture and Assembly of a Divergent Member of the ParM Family of Bacterial Actin-like Proteins. *The Journal of Biological Chemistry* 286 (16), 14282-14290.

Robinson, N.P. *et al*. 2004. Identification of Two Origins of Replication in the Single Chromosome of the Archaeon *Sulfolobus solfataricus*. *Cell* 116, 25-38.

Robinson, N. P. & Bell, S. D. 2007. Extrachromosomal element capture and the evolution of multiple replication origins in archaeal chromosomes. *Proceedings of the National Academy of Sciences* 104 (14), 5806-5811.

Rocha , E. P. C. & Bikard, D. 2022. Microbial defenses against mobile genetic elements and viruses: Who defends whom from what? *PLoS Biol* 20 (1), e3001514.

Rodionov, O. *et al*. 1999. Silencing of Genes Flanking the P1 Plasmid Centromere. *Science* 283, 546-549.

Rutherford, K. *et al*. 2000. Artemis: sequence visualisation and annotation. *Bioinformatics* 16 (10), 944-945.

Sakai, H. D. & Kurosawa, N. 2018. *Saccharolobus caldissimus* gen. nov., sp. nov., a facultatively anaerobic iron-reducing hyperthermophilic archaeon isolated from an acidic terrestrial hot spring, and reclassification of *Sulfolobus solfataricus* as *Saccharolobus solfataricus* comb. nov. and *Sulfolobus shibatae* as *Saccharolobus shibatae* comb. nov. *International Journal of Systematic and Evolutionary Microbiology* 68, 1271-1278.

Saljie, J. & Löwe, J. 2008. Bacterial actin: architecture of the ParMRC DNA partitioning complex. *The EMBO Journal* 27, 2230-2238.

Sanchez, A. *et al*. 2015. Stochastic Self-Assembly of ParB Proteins Builds the Bacterial DNA Segregation Apparatus. *Cell Systems* 1, 163-173.

Schleper, C. *et al*. 1995. A Multicopy Plasmid of the Extremely Thermophilic Archaeon *Sulfolobus* Effects Its Transfer to Recipients by Mating. *Journal of Bacteriology* 177 (15), 4417-4426.

Schreiter, E. R. & Drennan, C. L. 2007. Ribbon-helix-helix transcription factors: variations on a theme. *Nature Reviews Microbiology* 5, 710-720.

Schröder, G. & Lanka, E. 2003. TraG-Like Proteins of Type IV Secretion Systems: Functional Dissection of the Multiple Activities of TraG (RP4) and TrwB(R388). *Journal of Bacteriology* 185 (15), 4371-4381.

Schumacher, M. A. and Funnell, B. 2005. Structures of ParB bound to DNA reveal mechanism of partition complex formation. *Nature* 438, 516-519.

Schumacher, M. A. 2007. Structural biology of plasmid segregation proteins. *Current Opinion in Structural Biology* 17, 103-109.

Schumacher, M. *et al*. 2007. Segrosome structure revealed by a complex of ParR with centromere DNA. *Nature* 450, 1268-1271.

Schumacher, M. A. 2008. Structural biology of plasmid partition: uncovering the molecular mechanisms of DNA segregation. *Biochemical Journal* 412, 1-18.

Schumacher, M. A. *et al*. 2010. Insight into F plasmid DNA segregation complexes revealed by structures of SopB and SopB-DNA complexes. *Nucleic Acids Research* 38 (13), 4514-4526.

Schumacher, M. A. 2012. Bacterial plasmid partition machinery: a minimalist approach to survival. *Current Opinion in Structural Biology* 22, 72-79.

Schumacher, M. A. *et al*. 2015. Structures of archaeal DNA segregation machinery reveal bacterial and eukaryotic linkages. *Science* 349, 1120-1124.

Seeman, T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068-2069.

Sengupta, M. & Austin, S. 2011. Prevalence and Significance of Plasmid Maintenance Functions in the Virulence Plasmids of Pathogenic Bacteria. *Infection and Immunity* 79 (7), 2502-2509.

She, Q. *et al*. 1998. Genetic profile of pNOB8 from *Sulfolobus*: the first conjugative plasmid from an archaeon. *Extremophiles* 2, 417-425.

She, Q. *et al*. 2004. Archaeal integrases and mechanisms of gene capture. *Biochemical Society Transactions* 32 (2), 222-226.

Sheppard, C. *et al*. 2016. Répression of RNA polymerase by the archaeo-viral regulator ORF145/RIP. *Nature Communications* 7:13595.

Sherratt, D. J. 1974. Bacterial Plasmids. *Cell* 3, 189-195.

Shintani, M. *et al*. 2015. Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy. *Frontiers in Microbiology* 6:242, doi:10.3389.

Silvia, G. A. *et al*. 2005. PCR-Induced Sequence Artifacts and Bias: Insights from Comparison of Two 16S rRNA Clone Libraries Constructed from the Same Sample. *Applied and Environmental Microbiology* 71 (12), 8966-8969.

Simpson, A. E. *et al*. 2003. A Single Gene on the Staphylococcal Multiresistance Plasmid pKS1 Encodes a Novel Partitioning System. *Journal of Bacteriology* 185 (7), 2143-2152.

Smith, A. S. G. & Rawlings, D. E. 1998. Autoregulation of the pTF-FC2 Proteic Poison-Antidote Plasmid Addiction System (*pas*) Is Essential for Plasmid Stabilisation. *Journal of Bacteriology* 180 (20), 5463-5465.

Soh, Y-M. *et al*. 2019. Self-organisation of *parS* centromeres by the ParB CTP hydrolase. *Science* 366, 1129-1133.

Some, D. *et al*. 2019. Characterisation of Proteins by Size-Exclusion Chromatography Coupled to Multi-Angle Light Scattering (SEC-MALS). *Journal of Visualised Experiments* 148:e59615.

Song, D. *et al*. 2017. A network of *cis* and *trans* interactions is required for ParB spreading. *Nucleic Acids Research* 45 (12), 7106-7117.

Song, Y. *et al.* 2020. *Casimicrobium huifangae* gen. nov., sp. nov., a Ubiquitous "Most-Wanted" Core Bacterial Taxon from Municipal Wastewater Treatment Plants. *Applied and Environmental Microbiology* 86 (4), e02209-19.

Soppa, J. 1999. Transcription initiation in Archaea: facts, factors and future aspects. *Molecular Microbiology* 3195), 1295-1305.

Soppa, J. 2001. Prokaryotic structural maintenance of chromosomes (SMC) proteins: distribution, phylogeny, and comparison with MukBs and additional prokaryotic and eukaryotic coiled-coil proteins. *Gene* 278, 253-264.

Spang, A. *et al*. 2015. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, 521, 173–179.

Stedman, K.M. *et al*. 2000. pING Family of Conjugative Plasmids from the Extremely Thermophilic Archaeon *Sulfolobus islandicus*: Insights into Recombination and Conjugation in Crenarchaeota. *Journal of Bacteriology* 182, 7014-7020.

Strutt, S. C. *et al*. 2018. RNA-dependent RNA targeting by CRISPR-Cas9. *eLife* 7:e32724.

Sullivan, N. L. 2009. Recruitment of SMC by ParB-*parS* Organises the Origin Region and Promotes Efficient Chromosome Segregation. *Cell* 137, 697-707.

Takemata, N., Samson, R. Y., & Bell, S. D. 2019. Physical and Functional Compartmentalisation of Archaeal Chromosomes. *Cell* 179, 165-179.

Takemata, N. & Bell, S. 2020. Emerging views of genome organisation in Archaea. *Journal of Cell Science* 133, jcs243782.

Tang, M. *et al*. 2006. Minireplicon from pBtoxis of *Bacillus thuringiensis* subsp. *Israelensis*. *Applied and Environmental Microbiology* 72 (11), 6948-6954.

Tetsuov, R.L. *et al*. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research* 28 (1), 33-36.

Thisted, T. & Gerdes, K. 1992. Mechanism of post-segregational killing by the *hok/sok* system of plasmid R1: Sok antisense RNA regulates *hok* gene expression indirectly through the overlapping *mok* gene. *Journal of Molecular Biology* 1 (5), 41-54.

Tindall, B.J. *et al*. 2010. Notes on the characterisation of prokaryotic strains for taxonomic purposes. *International Journal of Systematic and Evolutionary Microbiology* 60,246-266.

Tran, N. T. *et al*. 2017. Permissive zones for the centromere-binding protein ParB on the *Caulobacter crescentus* chromosome. *Nucleic Acids Research* 46 (3), 1196-1209.

Tsang, J. 2017. Bacterial plasmid addiction systems and their implications for antibiotic drug development. *Postdoc Journal* 5 (5), 3-9.

Unterholzner, S. J. *et al*. 2013. Toxin-antitoxin systems. *Mobile Genetic Elements* 3:e26219.

Uribe, R. V. *et al*. 2019. Discovery and Characterisation of Cas9 Inhibitors Disseminated across Seven Bacterial Phyla. *Cell Host & Microbe* 25, 233-241.

Utatsu, I. *et al*. 1987. Yeast plasmids resembling 2 micron DNA: regional similarities and diversities at the molecular level. *Journal of Bacteriology* 169 (12), 5537-5545.

Vecchiarelli, A. G. *et al.* 2010. ATP control of dynamic P1 ParA-DNA interactions: a key role for the nucleoid in plasmid partition. *Molecular Microbiology* 78 (1), 78-91.

Vecchiarelli, A. G. *et al.* 2013a. Cell-free study of F plasmid partition provides evidence for cargo transport by a diffusion-ratchet mechanism. *Proceedings of the National Academy of Sciences* E1390-E1397.

Vecchiarelli, A. G. *et al.* 2013b. Dissection of the ATPase Active Site of P1 ParA Reveals Multiple Active Forms Essential for Plasmid Partition. *The Journal of Biological Chemistry* 288 (24), 17823-17831.

Vörös, Z. *et al*. 2017. Proteins mediating DNA loops effectively block transcription. *Protein Science* 26, 1427-1438.

Walsh, D. A. & Doolittle, W.F. 2005. The real "domains" of life. *Current biology : CB*, 15 (7), R237–R240.

Wang, J. *et al*. 2020. PaCRISPR: a server for predicting and visualising anti-CRISPR proteins. *Nucleic Acids Research* 48 (W1), W348-W357.

Wang, X. *et al.* 2013. Organisation and segregation of bacterial chromosomes. *Nature Reviews Genetics* 14, 191-203.

Wang, X. *et al*. 2014. The SMC Condensin Complex Is Required for Origin Segregation in *Bacillus subtilis*. *Current Biology* 24, 287-292.

Wang, X. *et al.* 2017. *Bacillus subtilis* SMC complexes juxtapose chromosome arms as they travel from origin to terminus. *Science* 355, 524-527.

Wang, X. *et al*. 2018. *In vivo* Evidence for ATPase-Dependent DNA Translocation by the *Bacillus subtilis* SMC Condensin Complex. *Molecular Cell* 71, 841-847.

Wang, Y. *et al*. 2016. Quantitative Localisation Microscopy Reveals a Novel Organisation of a High-Copy Number Plasmid. *Biophysical Journal* 111, 467-479.

Wang, Y. 2017. Spatial distribution of high copy number plasmids in bacteria. *Plasmid* 91, 2-8.

Williams, T. A. *et al.* 2017. Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proceedings of the National Academy of Sciences of the United States of America*, E4602–E4611.

Williams, T. A. *et al*. 2020. Phylogenomics provides robust support for a two-domains tree of life. *Nature Ecology & Evolution* 4, 138-147.

Woese, C.R. & Fox, G.E. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*, 74 (11), 5088–5090.

Woese, C. R., Magrum, L. J. and Fox, G.E. 1978. Archaebacteria. *Journal of Molecular Evolution* 11, 245-252.

Woese, C.R. 1987. Bacterial Evolution. *Microbiology*, 51 (2), 221–271.

Woese, C.R. *et al.* 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences of the United States of America*, 87 (12), 4576–4579.

Wu, F. *et al.* 2022. Unique mobile elements and scalable gene flow at the prokaryote-eukaryote boundary revealed by circularised Asgard archaea genomes. *Nature Microbiology* 7, 200-212.

Wu, M. *et al.* 2011. Segrosome assembly at the pliable *parH* centromere. *Nucleic Acids Research* 39 (12), 5082-5097.

Wu, S., Zhu, Z., Fu, L. *et al.* 2011. WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC Genomics* 12**,** 444 https://doi.org/10.1186/1471-2164-12-444.

Wu, Z. *et al.* 2014. DNA replication origins in archaea. *Frontiers in Microbiology* 5 (179), 1-7.

Xu, M. *et al*. 2016. Core promoter-specific gene regulation: TATA box selectivity and Initiator-dependent bi-directionality of serum-response factor-activated transcription. *Biochim Biophys Acta* 1859 (4), 553-563.

Yamaichi, Y. & Niki, H. 2000. Active segregation by the *Bacillus subtilis* partitioning system in *Escherichia coli*. *Proceedings of the National Academy of Sciences* 97 (26), 14656-14661.

Yan, W. X. *et al*. 2019. Functionally diverse type V CRISPR-Cas systems. *Science* 363, 88-91.

Yen, C-Y. *et al*. 2021. Chromosome segregation in Archaea: SegA- and SegB-DNA complex structures provide insights into segrosome assembly. *Nucleic Acids Research* 49 (22), 13150-13164.

Zampini, M. 2009. Recruitment of the ParG Segregation Protein to Different Affinity DNA Sites. *Journal of Bacteriology* 191 (12), 3832-3841.

Zaremba-Niedzwiedzka, K. *et al*. 2017. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* 541, 353-358.

Zebec, Z. *et al*. 2014. CRISPR-mediated targeted mRNA degradation in the archaeon *Sulfolobus solfataricus*. *Nucleic Acids Research* 42 (8), 5280-5288.

Zhang, C. *et al*. 2018. Revealing S-layer Functions in the Hyperthermophilic Crenarchaeon *Sulfolobus islandicus*. *BioRxiv* doi: 10.1101/444406.

Zhang, H. & Schumacher, M.A. 2017. Structures of partition protein ParA with nonspecific DNA and ParB effector reveal molecular insights into principles governing Walker-box DNA segregation. *Genes & Development* 31, 481-492.

Zhang, Z. *et al.* 2012. Archaeal chromatin proteins. *Science China Life Sciences* 55 (5), 377-385.

Zhang, Z. *et al.* 2019. Architectural roles of Cren7 in folding crenarcheal chromatin filament. *Molecular Microbiology* 111 (3), 556-569.

Zillig, W. *et al*. 1980. The *Sulfolobus*-"*Caldariella"* Group: Taxonomy on the Basis of the Structure of DNA-Dependent RNA Polymerases. *Archives of Microbiology* 125, 259-269.

Zillig, W. *et al*. 1996. Viruses, plasmids and other genetic elements of thermophilic and hyperthermophilic *Archaea*. *FEMS Microbiology Reviews* 18, 225-236.

Zillig, W. *et al*. 1998. Genetic elements in the extremely thermophilic archeaon *Sulofolobus*. *Extremophiles* 2, 131-140.

Ziolkowska, K. *et al*. 2006. Hfq variant with altered RNA binding functions. *Nucleic Acids Research* 34 (2), 709-720.

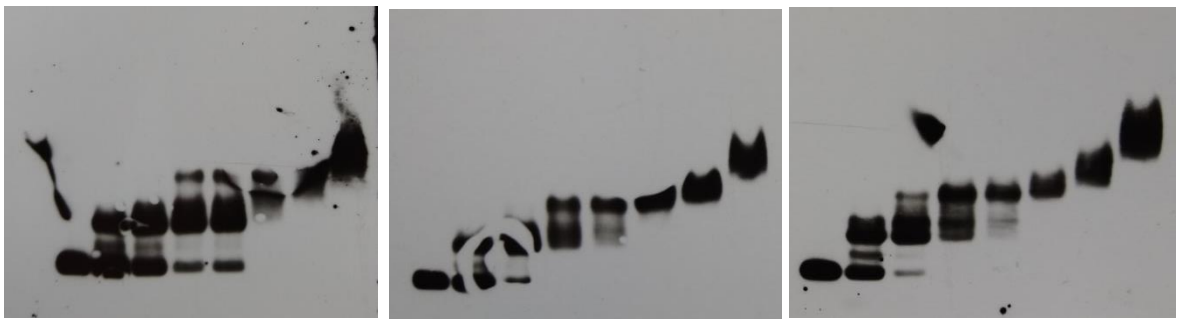
Appendices

Appendix 1 – EMSA replicate data

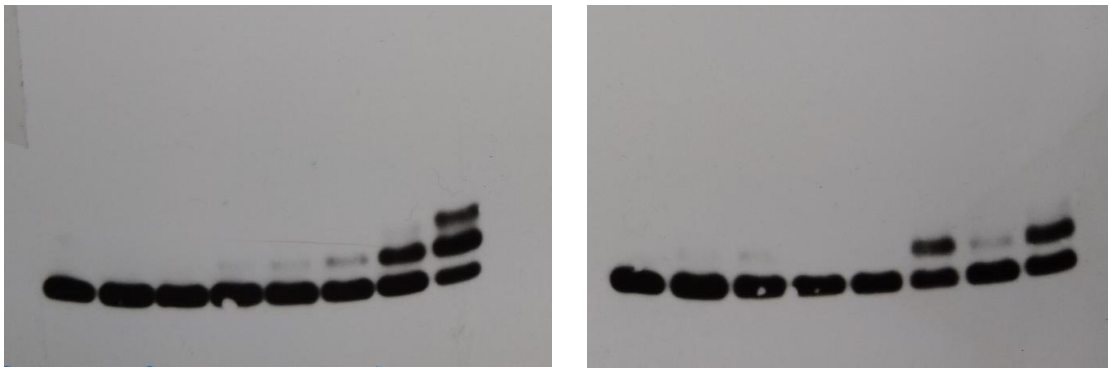
A1.1 AspA WT and AspA-A53K

Protein concentrations are 0,10,20,40,50,100,200,500 nM

WT



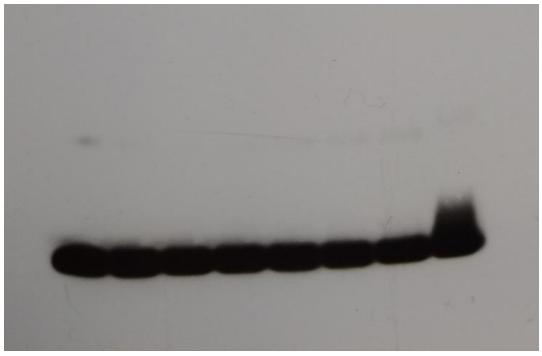
AspA-A53K

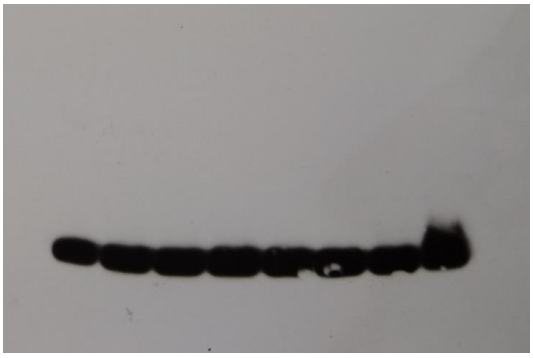


A1.2 AspA-Y41A and AspA-Q42A

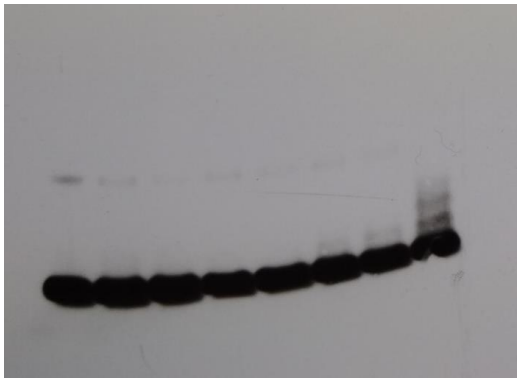
Protein concentrations are 0,10,20,40,50,100,200,500 nM

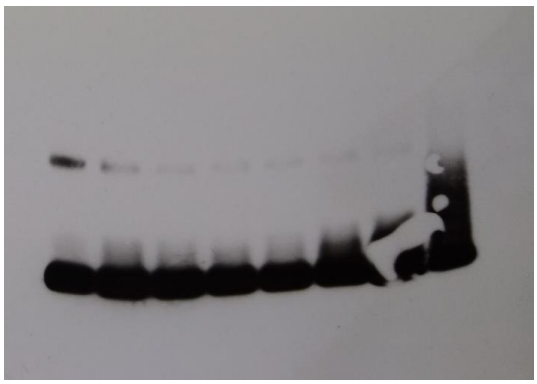
AspA-Y41A





AspA-Q42A

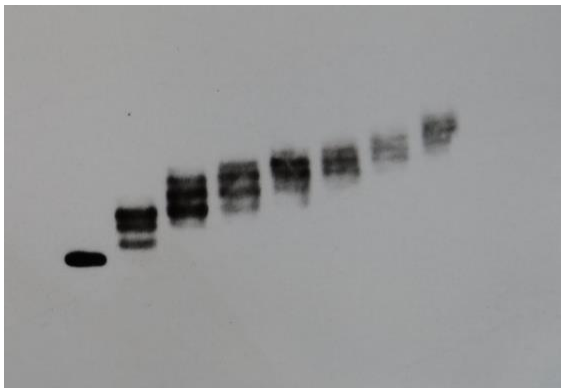


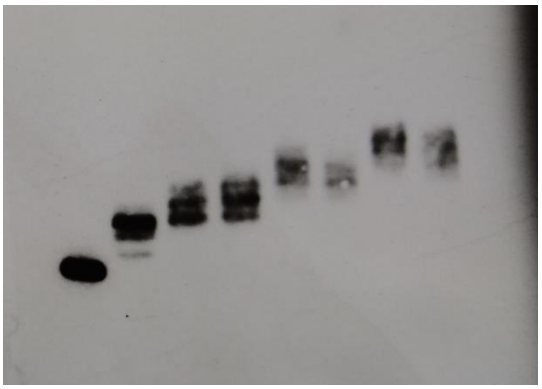


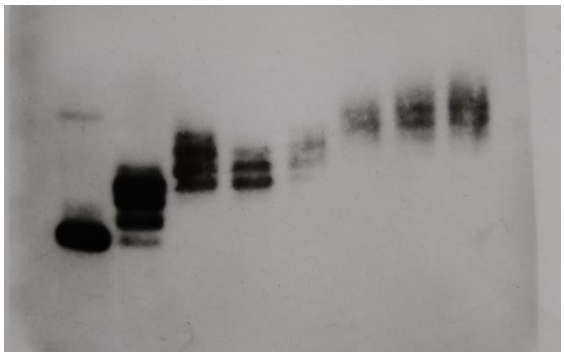
A1.3 AspA-L52K and AspA-E54A

Protein concentrations are 0,100,250,350,500,650,800,1000 nM

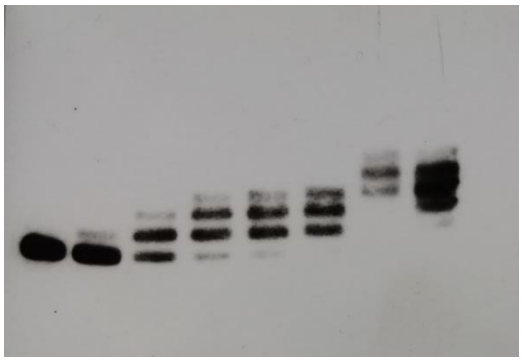
<u>AspA-L52K</u>

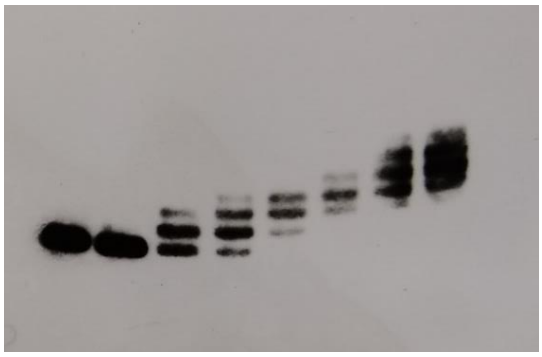


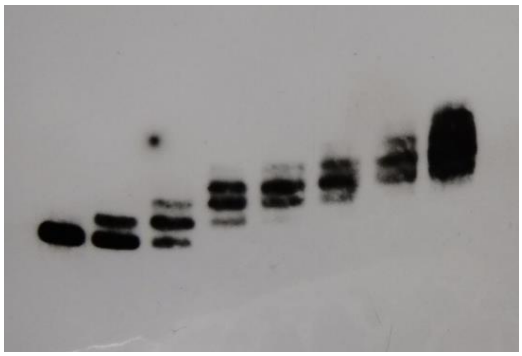


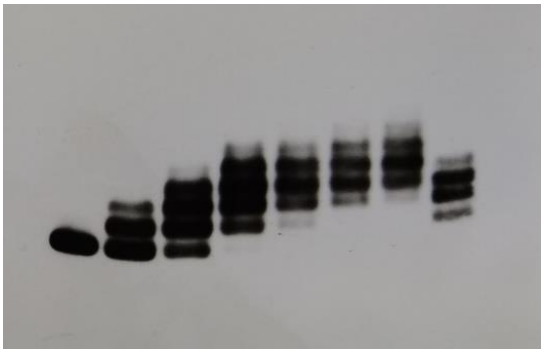


AspA-E54A





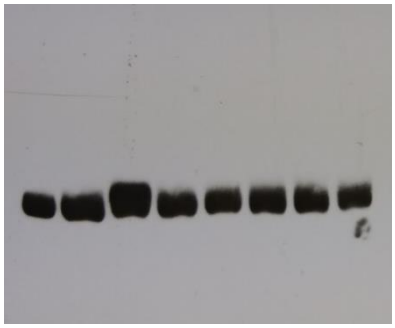


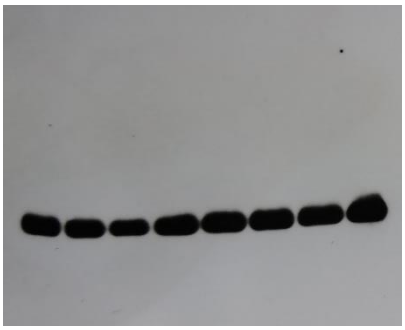


A1.4 AspA L12G, I85G, V89G and AspA I85GV89G

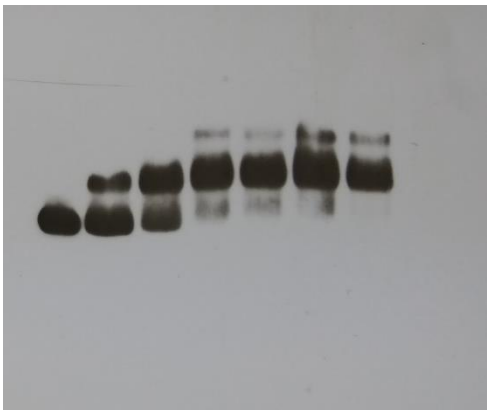
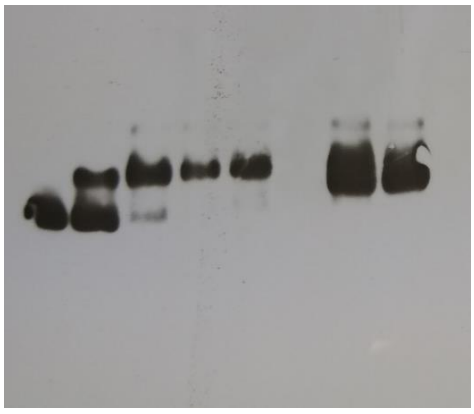
Protein concentrations are 0,100,250,350,500,650,800,1000 nM

AspA-L12G

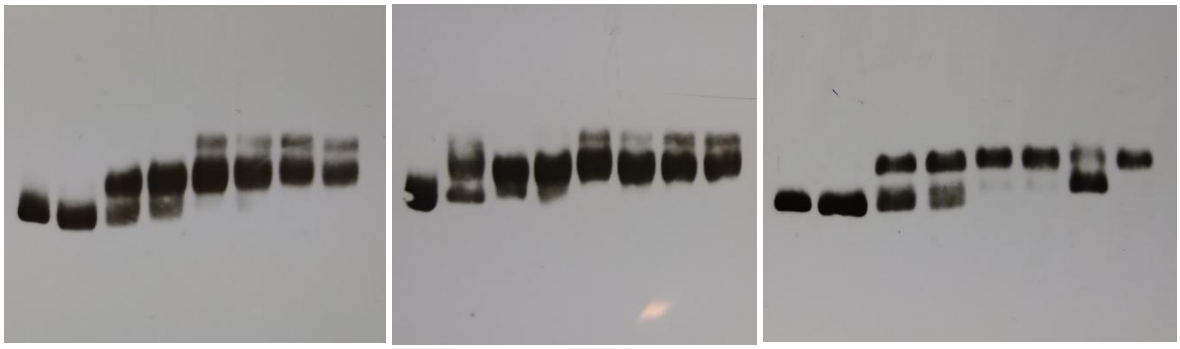




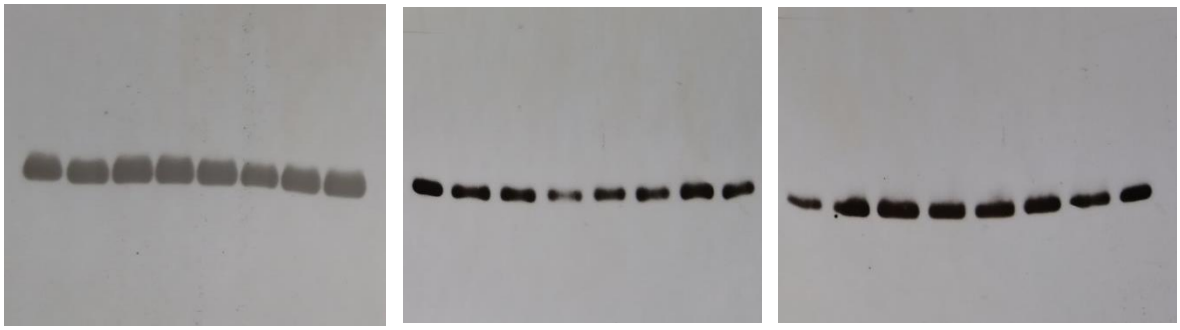
AspA-I85G

AspA-V89G



AspA-I85GV89G



Appendix 2 – DNase I footprinting replicate data

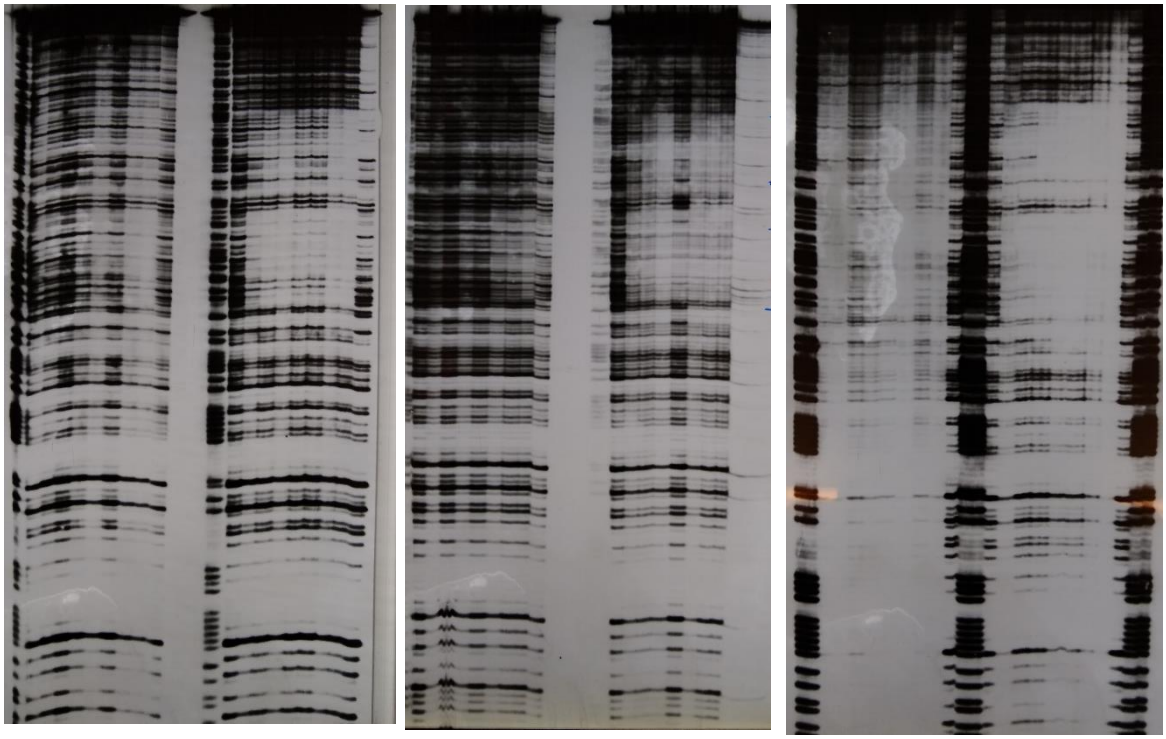
A2.1 AspA WT, AspA-E54A, AspA-Y41A

Protein concentrations are:

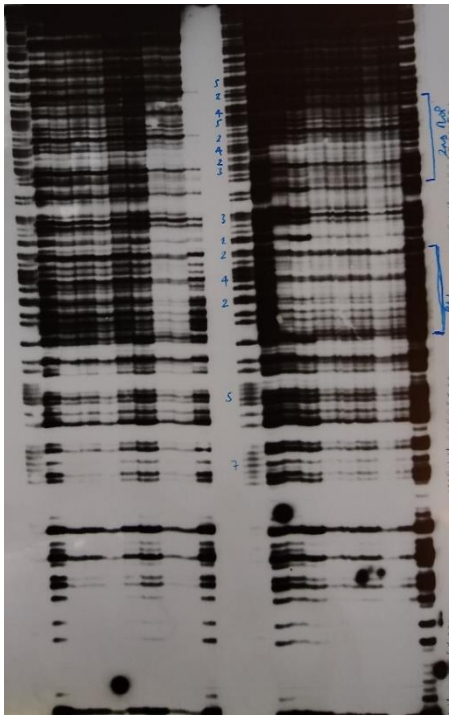
(Left): 0,25,50,100,200,500,750,1000 nM

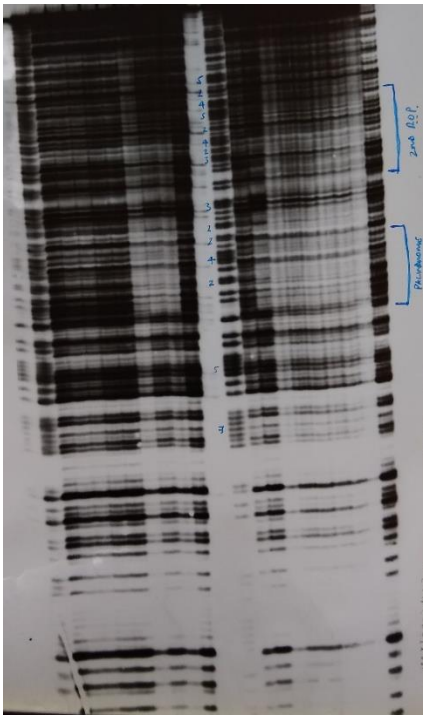
(Right): 0,1000,1250,1500,1750,2000,2500,3000 nM

<u>AspA WT</u>

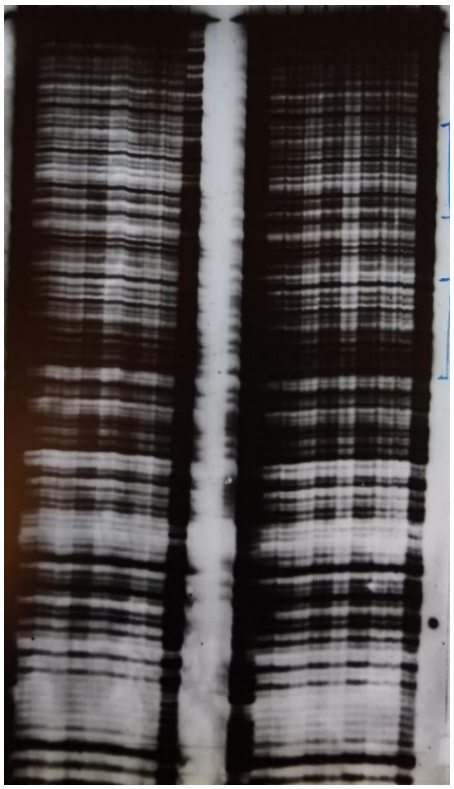


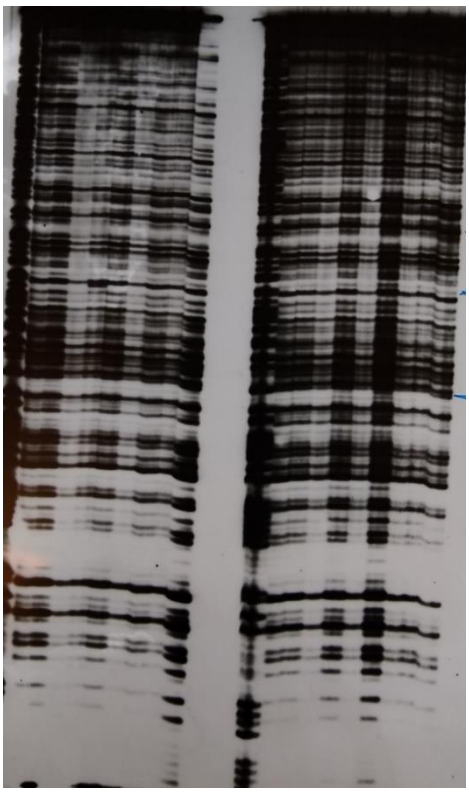
AspA-E54A





AspA-Y41A



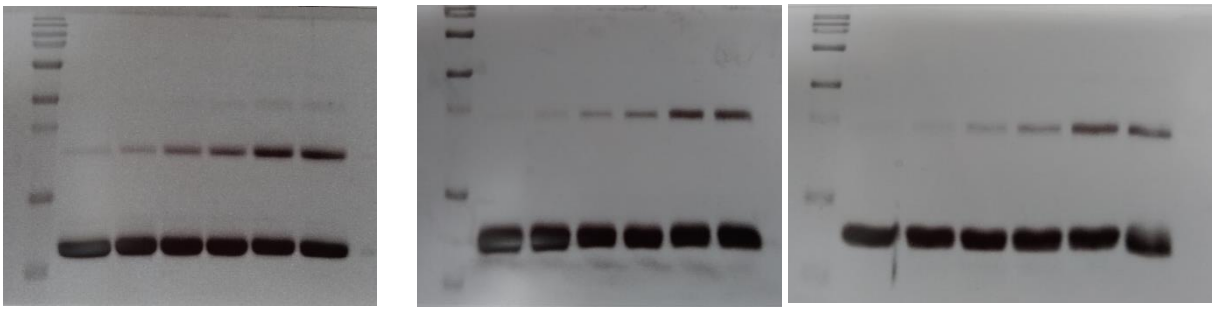


Appendix 3 – DMP cross-linking replicate data

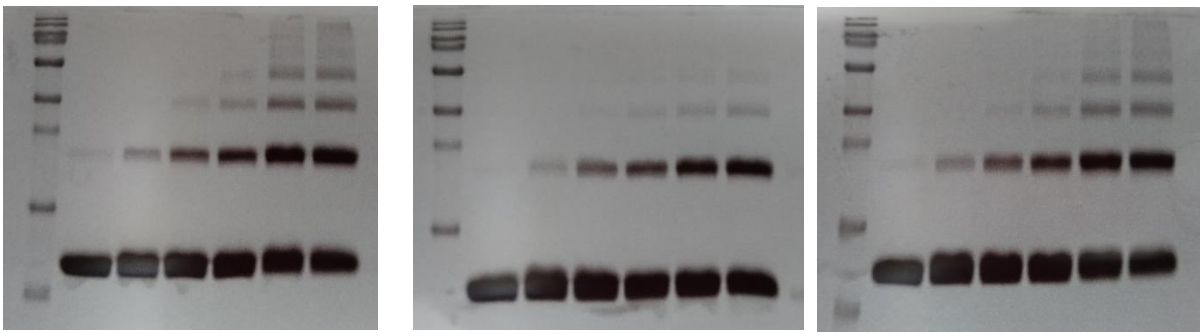
A3.1 AspA WT, AspA-L12G, AspA-I85G, AspA-V89G, AspA-I85GV89G

DMP concentrations are: (Left): 0,0.1,0.5,1,5,10 mM

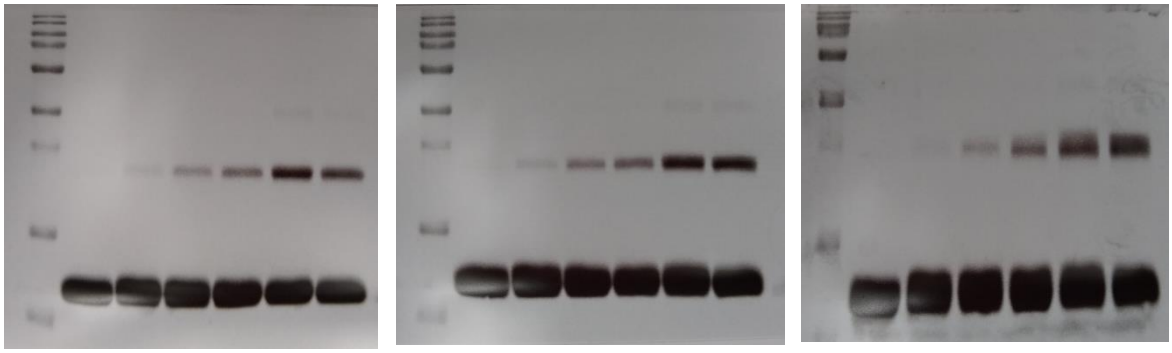
AspA-WT



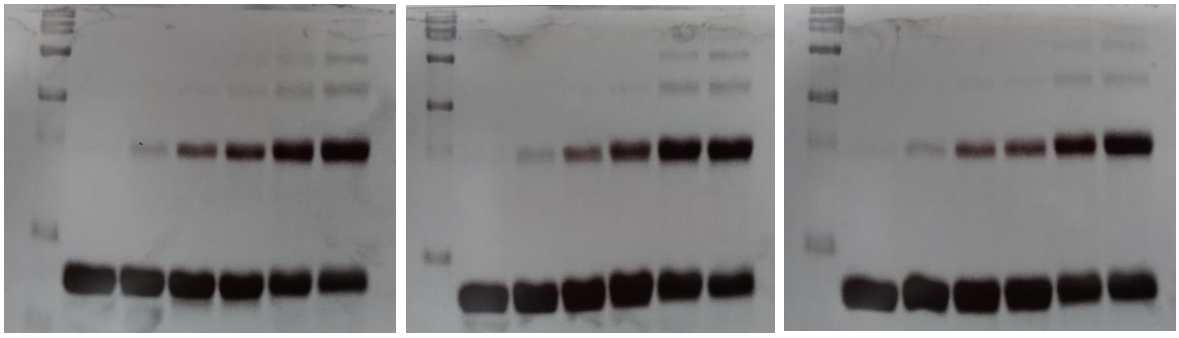
AspA-L12



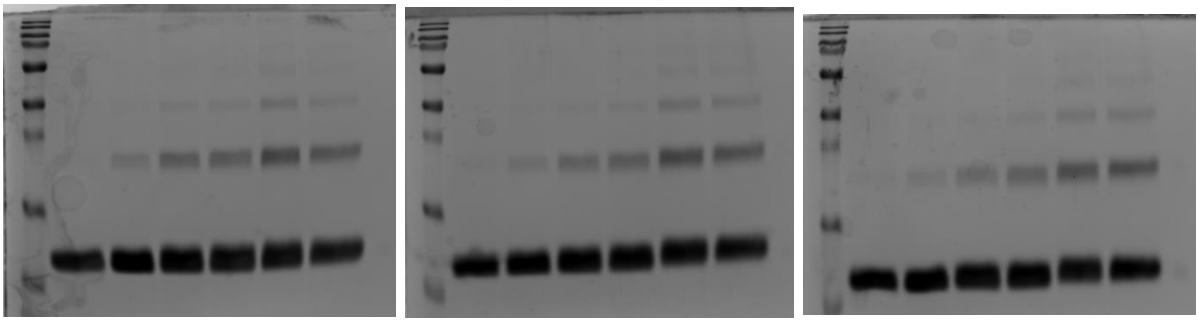
AspA-I85G



AspA-V89G



AspA-I85GV89G



Appendix 4 – List of viruses and plasmids used in CRISPR spacer analysis

Table A1. List of viruses and plasmids know to interact with the crenarchaea[a]

Accession number	Plasmid	Virus	Virus abbreviation
emb|AJ010405	pNOB8		
emb|AJ748324	pHVE14		
emb|AJ748322	pARN3		
emb|AJ748323	pARN4		
ref|NC_004852	pING 1		
gb|DQ335583	pSOG1		
gb|DQ335584	pSOG2		
gb|CP001405	pYN01		
gb|CP001732	pLD8501		
emb|AJ748321	pKEF9		
ref|NC_021914	pMGB1		
gb|EU881703	pAH1		
gb|AY517480	pTC		
gb|U36383	pRN1		
gb|AY591755	pIT3		
gb|EU030940	pXZ1		
gb|U93082	pRN2		
emb|AJ294536	pHEN7		
emb|AJ225333	pDL10		
ref|NG_036063.1	pTIK4		
ref|NG_036062.1	pTAU4		
ref|NC_006906.1	pORA1		
emb|AJ243537.1	pSSVx		
gb|DQ183185	pSSVi		
emb|X07234		*Sulfolobus* spindle-shaped virus 1	SSV1
gb|AY370762		*Sulfolobus* spindle-shaped virus 2	SSV2
gb|EU030938		*Sulfolobus* spindle-shaped virus 4	SSV4
gb|EU030939		*Sulfolobus* spindle-shaped virus 5	SSV5
gb|FJ870915		*Sulfolobus* spindle-shaped virus 6	SSV6
gb|FJ870916		*Sulfolobus* spindle-shaped virus 7	SSV7
gb|AY423772		*Sulfolobus* virus Kamchatka1	SSVk1
emb|AJ888457		*Acidianus* two-tailed virus	ATV
emb|HG322870		*Sulfolobus* monocaudavirus	SMV1

emb|AJ783769		*Sulfolobus tengchongensis* spindle-shaped virus 1	STSV1
gb|JQ287645		*Sulfolobus tengchongensis* spindle-shaped virus 2	STSV2
emb|AJ414696		*Sulfolobus islandicus* rod-shaped virus 1	SIRV1
emb|AJ344259		*Sulfolobus islandicus* rod-shaped virus 2	SIRV2
emb|AJ875026		*Acidianus* rod-shaped virus 1	ARV1
emb|AJ567472		*Acidianus* filamentous virus 1	AFV1
emb|AJ854042		*Acidianus* filamentous virus 2	AFV2
emb|AM087120		*Acidianus* filamentous virus 3	AFV3
emb|AM087121		*Acidianus* filamentous virus 6	AFV6
emb|AM087122		*Acidianus* filamentous virus 7	AFV7
emb|AM087123		*Acidianus* filamentous virus 8	AFV8
gb|EU545650		*Acidianus* filamentous virus 9	AFV9
gb|AF440571		*Sulfolobus islandicus* filamentous virus	SIFV
emb|X14855		*Thermoproteus tenax* virus 1	TTV1
emb|AJ635161		*Pyrobaculum* spherical virus	PSV
gb|AY722806		*Thermoproteus tenax* spherical virus	TTSV
gb|EF432053		*Acidianus* bottle-shaped virus	ABV
dbj|AB537968		*Aeropyrum pernix* bacilliform virus 1	APBV1
gb|AY569307		*Sulfolobus* turreted icosahedral virus 1	STIV1
gb|GU080336		*Sulfolobus* turreted icosahedral virus 2	STIV2

[a] Adapted from Liu 2015.