Comprehensive Characterization of Archaeal MCMs for use in Novel Nanopore Applications

Oliver William Noble, BSc

Doctor of Philosophy

University of York

Biology

September 2021

Abstract

Replicative helicases are central components of the cell cycle, catalysing strand separation ahead of the of the DNA synthesis machinery. Across the domains of life, this role is fulfilled by toroidal hexameric helicases. Consistent with the notion that eukaryotes evolved from within archaea, both domains share the MCM class of helicase. The homohexameric archaeal MCM has provided a valuable tool as a simplified model of the heterohexameric eukaryotic MCM2-7 complex.

In biotechnology, helicases are central components to Oxford Nanopore Technologies (ONT) sequencing platform, which controllably ratchet DNA through the nanopore sensor. Typically, these have been monomeric helicases, however, there is an interest to examine alternative helicase enzymes that may improve the current sequencing workflow and provide different error profiles. The use of processive, self-loading replicative helicases is largely underexplored.

Conditions within ONTs flow cell demand helicase activity in high concentrations of salt and at room temperature. Historically, studied archaeal MCMs have been from thermophilic organisms which are not expected to be active at room temperature. Work presented in this thesis addresses this long-standing bias and identifies *Mac*MCM from the mesophilic archaeon *M. acidiphilum,* which is highly active at 25 °C. *Mac*MCM exhibits a slow, oligomerization-linked kinetic step which is dependent on ATP hydrolysis, DNA-binding, and the regulatory winged helix domain. This kinetic event, likened to ring closure in MCM2-7, is previously unobserved in archaeal models. Through structural analyses it is further demonstrated that *Mac*MCM interfaces are more eukaryotic-like than interfaces from previously resolved archaeal MCM structures.

Optimal conditions have been established for ONT experiments to ensure optimal binding and activity. Initial experiments within an ONT flow cell suggest several events consistent with MCM-driven translocation, however, further work is required to increase the frequency and consistency of measurements. Using the resolved *Mac*MCM structure, we are beginning to engineer efficient, sequencing tractable MCM motors.

Table of Contents

List of Figures

List of Tables

List of Equations

List of Abbreviations

Abbreviation	Meaning
AAA+	ATPase associated with various cellular activities
$\Delta^{i}G$	Free energy of interface formation
^{31}P-NMR	31 Phosphorus NMR
A_{260}	Absorbance measured at 260 nm
A_{280}	Absorbance measured at 280 nm
A_{290}	Absorbance measured at 290 nm
ACl	Allosteric communication loop
ADP	Adenosinediphosphate
ADP•AlF_4^-	ADP-aluminium fluoride
AEX	Anion exchange chromatography
Afu	*Archaeoglobus fulgidus*
AMP-PCP	β,γ-Methyleneadenosine 5′-triphosphate
Amp^{R}	Ampicillin resistance
Ape	*Aeropyrum pernix*
APS	Ammonium persulfate
ARS	Autonomously replicating sequence
AS	Annealed substrate
ASIC	Application specific integrated circuit
ATP	Adenosinetriphosphate
ATPase	Adenosinetriphosphatase
AUC	Analytical ultracentrifugation
BHQ	Black hole quencher
BSA	Bovine serum albumin
CCD	Charge coupled device
CMG	Cdc45-MCM-GINS
Cmp^{R}	Chloramphenicol resistance
Cryo-EM	Cryogenic electron microscopy
CTD	C-terminal ATPase domain
CTE	C-terminal extension
CV	Column volume

Cy2	Cyanine 2
Cy3	Cyanine 3
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DRI	Differential refractive index
dsDNA	Double-stranded DNA
DTT	Dithiothreitol
E. coli	*Escherichia coli*
E1	Papillomavirus E1 helicase
EMSA	Electrophoretic mobility shift assay
ESP	Eukaryotic-specific signature protein
FAM	6-carboxyfluorescein
FID	Free induction decay
FL	Full-length
FRET	Förster resonance energy transfer
GINS	Go-ichi-ni-san
gp41	Bacteriophage T4 helicase
H2i	Helix-2 insert
His$_{10}$-tag	Decahistidine tag
HPLC	High performance liquid chromatography
Hvo	*Haloferax volcanii*
IMAC	Immobilised metal affinity chromatography
IPTG	Isopropyl β-d-1-thiogalactopyranoside
$K_{cat.app}$	Apparent catalytic constant
Kcr	*Korarchaeum cryptofilum*
K_d	Equilibrium dissociation constant
$K_{d,app}$	Apparent equilibrium dissociation constant
kDa	Kilo Dalton
KEGG	Kyoto Encyclopedia of Genes and Genomes
KGlu	Potassium glutamate
LB	Lysogeny Broth
LLG	Log-likelihood gain
Mac	*Mancarchaeum acidiphilum*

Mba	*Methanosarcina barkeri*
MCM	Minichromosome Maintenance Protein
MCM-BP	MCM binding protein
MF	Maximum fluorescence
Mha	*Methanohalophilus halophilus*
Mka	*Methanopyrus kandleri*
MR	Molecular replacement
MSA	Multiple sequence alignment
Mth	*Methanothermobacter themautotrophicus*
MUSCLE	Multiple sequence alignment by log expectation
MW	Molecular weight
MWCO	Molecular weight cut-off
Nac	*Nanohaloarchaea archaeon SG9*
NanoDSF	Nano differential scanning fluorimetry
NCS	Non crystallographic symmetry
Neq	*Nanoarchaeum equitan*
NH	No helicase
Ni-NTA	Nickel nitrilotriacetic acid
NLS	Nuclear localization signal
Nma	*Nitrosopumilus maritimus*
NMR	Nuclear magnetic resonance
NR-CBBR	Neutral red Coomassie brilliant blue stain
NT-hp	N-terminal hairpin
NTD	N-terminal DNA binding domain
NTE	N-terminal extension
NTI	N-terminal insertion
NTPase	Nucleotidetriphosphatase
OB-fold	Oligosaccharide/nucleotide binding fold
OD_{600}	Optical density measured at 600 nm
ONT	Oxford Nanopore Technologies
ORC	Origin recognition complex
ORF	Open reading frame
Ori	Origin of replication (Bacterial/archaeal)

PacBio	Pacific Biosciences
PAGE	Poly-acrylamide gel electrophoresis
PDB	Protein Data Bank
Pfu	*Pyrococcus furiosus*
Phyre2	Protein Homology/AnalogY Recognition Engine
P_i	Inorganic phosphate
PISA	Proteins, Interfaces, Structures and Assemblies
PMCC	Product moment correlation coefficient
POPS	Parameter Optimised Solvent accessibility
PP_i	Pyrophosphate
PS1β	Presensor-1 beta hairpin
QELS	Quasi-elastic light scattering
R	R programming language
R.f	Radio frequency
Rh	Radius of hydration
RMSD	Root mean squared deviation
SAD	Single wavelength anomalous dispersion
SASA	Solvent accessible surface area
SCAV	Scavenger strand
SDS	Sodium dodecyl sulfate
SDS-PAGE	Sodium dodecyl sulfate poly-acrylamide gel electrophoresis
SEC	Size exclusion chromatography
SEC-MALLS	SEC with multi-angle laser light scattering
SEM	Standard error of the mean
SEW	Steric exclusion unwinding and wrapping
SOC	Super optimal broth with Catabolite repression
SSB	Single-stranded DNA binding protein
ssDNA	Single-stranded DNA
Sso	*Saccharolobus solfataricus*
TB	Tris borate
TBS-T	Tris buffered saline with Tween
TCEP	Tris(2-carboxyethyl) phosphine
TEMED	Tetramethylethylenediamine

TEV	Tobacco etch virus protease
TFZ	Translation function Z-score
UV	Ultraviolet
WAMW	Weight-average molecular weight
WHD	Winged helix domain
WT	Wild type
XRF	X-ray fluorescence
ZMW	Zero-mode waveguide
ZnF	Zinc Finger
α-HL	α-hemolysin

Author's declaration

I declare that this thesis is a presentation of original work, and I am the sole author, except where clearly stated in the text below. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

I acknowledge Michael Hodgkinson for performing a single experimental repeat for the ATPase assay (section 3.4.3) and Dr Clément Degut for assistance with data collection and analysis of X-ray diffraction experiments (section 4.3).

Acknowledgements

I would first like to thank my supervisors Prof. James Chong and Dr Michael Plevin for choosing me for this PhD project. I am extremely grateful for their support, feedback and guidance which has provided me with the perfect environment to grow as a researcher. I would also like to thank my Thesis Advisory Panel members, Dr Christoph Baumann and Prof. Daniela Barillà for keeping me on course, and to our industrial collaborators at Oxford Nanopore Technologies (ONT), Dr Mark Bruce and Joseph Lloyd for thoughtful discussions and support throughout this project.

My studies were accelerated massively thanks to Dr Clément Degut and Michael Hodgkinson, who shared with me their expertise in many fundamental biochemical and biophysical techniques I required to complete this PhD. I would also like to thank members of the technology faculty for experimental help, notably Dr Andrew Leech for assistance in molecular interaction techniques, Dr Alex Heyam for NMR guidance, and Dr Johan Turkenburg and Sam Hart for X-ray assistance.

Research carried out in this project would not have been possible without the generous financial support supplied by the BBSRC through the iCASE doctoral training programme, and ONT who provided funding and access to equipment. I am also grateful for Diamond Light Source and its staff for providing the facilities and support required for collecting X-ray diffraction datasets.

I would also like to thank many past and present members of the Plevin group, whose support and friendship throughout this PhD both in and out of the lab brightened my days immeasurably, including Ben Rowlinson, Ed Nay, Emily Flack, Rachael Cooper, and Sam Griffiths. Finally, I would like to express my heartfelt gratitude to my partner Claire and both our families for their tremendous support over the past 4 years.

Chapter 1 – Introduction

1.1 Archaea

The molecular characterization of organisms by Carl Woese revolutionized the way all life is classified [1] . Previously, life was ordered based on the morphological feature of cells observed under a microscope [2]. Observation-based taxonomy grouped life into two domains depending on whether a nucleus was present and were defined as the prokaryotes (before nucleus) and eukaryotes (with nucleus). In the 1970s, restriction enzyme fingerprinting of microbial 16S RNA by Woese exposed a new domain of life, which were coined the Archaebacteria (subsequently Archaea) [1]. Although prokaryotic in nature, in evolutionary terms archaea were as distant to bacteria as they were from eukaryotes. The commercial availability of powerful nucleic acid sequencing techniques in the 1980s (section 4.1.1) opened new possibilities for examining the phylogenetic relationships between organisms. Woese later described archaea within a new three domain system for classifying life, which has remained relatively undisputed until recently (Figure 1.1) [3].

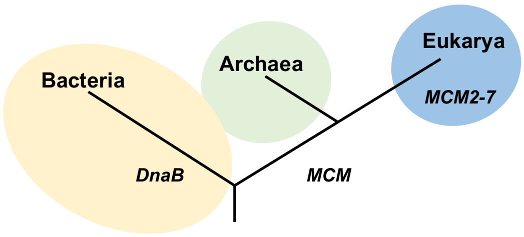


Figure 1.1: The Woesian tree of life.
The three-domain organisation of the tree of life as proposed by Carl Woese[3]. Two replicative helicases, DnaB and MCM, have evolved from a last universal common ancestor.

Within this framework, archaea were initially divided into 2 kingdoms known as Crenarchaeota and Euryarchaeota [3]. The precise organisation of phyla within the archaeal domain has been subject to change in line with advances in sequencing technologies. Genome sequencing of archaea was primarily impeded by difficulties cultivating archaeal species. To this day the difficulties remain, where within the 27 proposed archaeal phyla, only 6 have been successfully cultured[4]. Many archaea are extremely slow growing and require multi-year cultures for the generation of stable isolates [5,6]. In 1996, the initial archaeal genome of the hyperthermophile *Methanococcus janachii* was published [7]. Analysis suggested that many gene pathways including transcription, translation and DNA replication, share greater resemblance to eukaryotes than bacteria [7].

Developments in the generation and analysis of sequencing data has permitted metagenomic sequencing where culturing is no longer essential. Archaea were initially considered extremophilic organisms, however archaea have since been found in virtually all environmental niches [8–10]. To date, the genomes of almost 2,000 archaea have been resolved (https://www.ncbi.nlm.nih.gov/genome/).

Woese's original 2 kingdom archaeal tree has been expanded to accommodate at least 4 superphyla, comprising TACK, DPANN, Asgard and Euryarchaeota. Crenarchaeota has since been relegated to a superphyla including Thaumarchaeota, Aigarchaeota and Korarchaeota (TACK) [11]. Organisms within TACK and Euryarchaeota represent around 83 % of sequenced archaeal genomes (https://www.ncbi.nlm.nih.gov/genome/). Model organisms exist for both superphyla, including *Saccharolobus solfataricus* (TACK) and *Methanothermobacter thermautotrophicus* (Euryarchaota), which are well characterised *in vivo* and *in vitro* [12,13]. DPANN and Asgard are recent additions to the fringes of the archaeal phylogenetic tree with novel cellular properties [14–16].

1.2 DPANN - the origins of life?

The DPANN superphylum was proposed in 2013, although the first member organism was identified back in 2002 [15,17]. DPANN is an acronym of the founding phyla: Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota and Nanohaloarchaota (although many additional phyla have since been identified, including Micrarchaoeta) [15]. Species exist in a range of global habitats including soil, acidic mine drainage, seawater and hydrothermal vents [15,17,18]. DPANN organisms are characterized by exceptionally small cells (<0.5 µm) and reduced genome sizes (<1 Mb) [17,19]. Remarkably, most members lack key respiratory genes and there exist examples where the organism is missing the entirety of the electron transport chain [20,21]. It is suggested that these organisms may rely on fermentation as the primary energy source [14,21]. Many species identified in DPANN are reliant on a symbiont-host relationship for survival. For example, *Nanoarchaeum equitans* is dependent on a relationship with *Ignococcus hospitalis* [22], whilst *Mancarchaeum acidiphilum* is reliant on *Cuniculiplasma divulgatum* [18]. It is suggested that the symbiotic relationship is sustained through reciprocal transfer of metabolites. No DPANN archaea have successfully been cultured in isolation, however *M. acidiphilum* has successfully been maintained in a culture with *C. divulgatum* for 2 years [18]. Notably, addition of rich media to the culture is sufficient to eliminate *M. acidiphilum*, implying that its metabolic function is no longer required. Alternatively, some researchers argue that the majority of DPANN are free living organisms and that their unique features are not dependent on symbiont driven evolution. Instead these organisms may represent an insight to a primitive form of life that pre-dates the electron transport chain [20]. It may be possible that communities of auxotrophic DPANN organisms are sustained through shared nutrient cycling [23].

1.3 Asgard - the origin of eukaryotes?

In 2015, a new archaeal phylum named Lokiarchaeota was reported from a metagenomic analysis of marine sediment extracted from the Loki's castle hydrothermal vent in the Mid-Atlantic [24]. Later, related organisms were identified in freshwater sediments [25], hot springs and groundwater [16]. Organisms were given names according to Norse mythology (e.g., Thorarchaoeta). Phylogenetic analysis grouped the species into a new superphylum, aptly named the Asgard archaea [16].

Asgard archaea are generally reported in very low abundance in environmental samples, comprising less than 1% of the microbial community [4]. To date it has been possible to isolate and grow only a single species from this lineage, a feat spanning almost 10 years [5]. Initial genome analysis suggests that Lokiarchaeota possess a remarkable number (~3.3 %) of eukaryotic-specific signature proteins (ESPs) [24]. These homologues are largely implicated in the regulation of membrane shape, including: endosomal sorting complex required for trafficking (ESCRT) [16,26], actin regulators [27], tubulin [16,28] and SNARE proteins [29]. Other ESPs include a ubiquitin-like modifier system [16].

It is widely regarded that evolution of the eukaryotic cell required endogenization of an alpha-proteobacterium to form a mitochondrion [30]. Endogenization may occur through controlled phagocytosis which demands control of cell membrane shape. Excitingly an Asgard species was recently observed interacting with a bacterium trapped through membrane appendages [5]. The authors conclude that the mechanism of eukaryogenesis is possibly independent of phagocytosis and propose the transient entangle, engulf, endogenize (E3) model of eukaryogenesis [5].

Genetically, Asgard archaea are currently regarded as the lineage with highest similarities to eukaryotes. Although contentious, data support a departure from the Woesian three domain hypothesis and it seems likely that eukaryotes evolved directly from within Asgard archaea (Figure 1.2) [31–34]. Controversially, this 'Eocyte model' suggests that life evolved as only two domains [33,35]. There remain issues with the Eocyte model, in particular with the

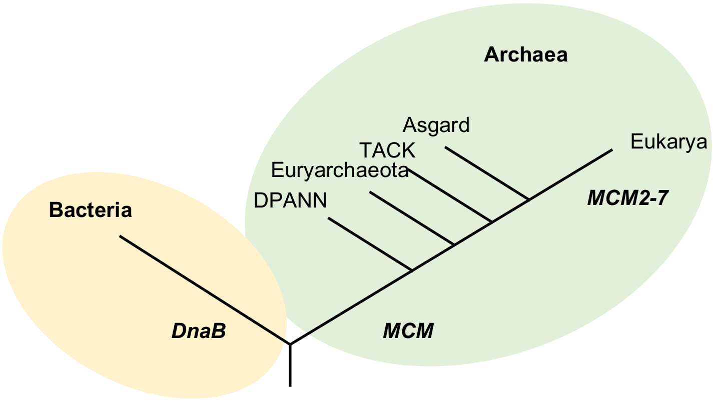


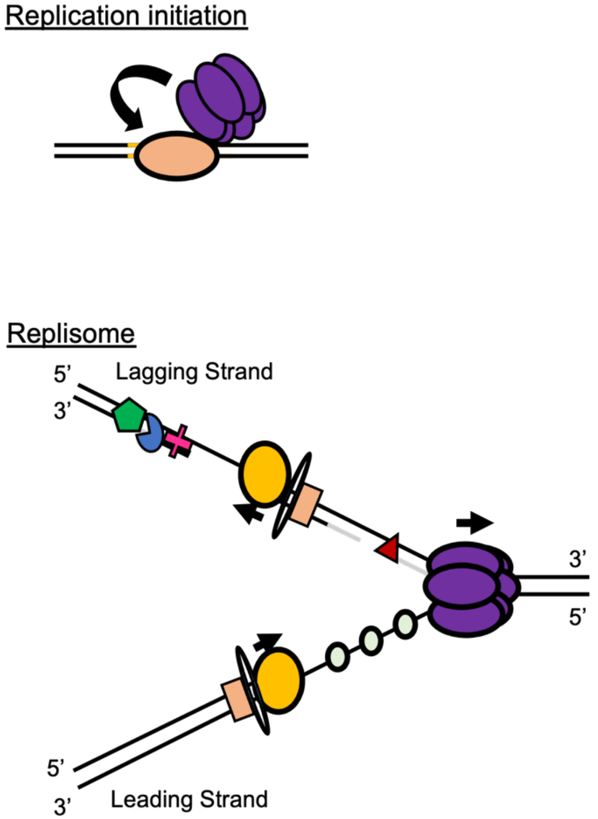
Figure 1.2: The Eocyte tree of life.
The two-domain organization of the tree of life, as proposed by James Lake[34]. Eukaryotes evolved directly within archaea.

evolution of the archaeal membrane. The eukaryotic and bacterial cell membrane fatty acids are linked with an ester bond, whilst archaeal fatty acids are linked with an ether bond [36]. It is however argued that the large differences in diversity between archaea and eukaryotes evolved after eukaryogenesis. Notably the significant energy benefit provided by a mitochondrion is cited to support rapid genetic diversification [37,38].

It is extremely likely in the coming decades we will discover further lineages within Asgard with furthermore diverse eukaryotic features. Nevertheless, archaea have already provided valuable insights into fundamental eukaryotic processes. Owing to their homology, archaeal enzymes have been excellent representatives for understanding the core components of the replisome.

1.4 Archaea and the prototype eukaryotic replisome

The process of DNA replication is central to all living organisms ensuring faithful inheritance of genetic information. Semi-conservative DNA replication is carried out by a macromolecular complex known as the replisome. The replisome performs two fundamental processes. First, strand separation of parental double stranded DNA (dsDNA) is catalysed by a large toroidal hexameric helicase [39,40]. Second, dsDNA is synthesized from the single stranded (ssDNA) templates by DNA polymerases[39]. Two alternative replisome frameworks have evolved and are compared in Figure 1.3. Bacteria form a replisome using DnaB as the replicative helicase and family C polymerases for DNA synthesis [41,42]. In accordance with the two-domain theory, archaea and eukaryotes share a distinct replisome core comprising a Minichromosome Maintenance (MCM) helicase and



Stage	Key	Bacteria	Archaea	Eukaryotes
1. Origin recognition	═	*Ori*	*Ori*	ARS
	⬭	DnaA	Cdc6/Orc1	ORC (Orc1, 2, 3, 4, 5, 6), Cdc6
		DnaC	WhIP (Cdt1)	Cdt1
2. DNA unwinding	⬬	DnaB	MCM GINS (Gins23, Gins51)	MCM (MCM2, 3, 4, 5, 6, 7) GINS (Sld5, Psf1, 2, 3)
	○	SSB	SSB	RPA
3. Primer synthesis	▲	DnaG	DNA primase	Pol α/ primase4.
4. DNA synthesis	●	Family C DNA polymerase Pol (III)	Family D DNA polymerase (Pol D) Family B DNA polymerase (Pol B)	Family B DNA polymerase (Pol δ) Family B DNA polymerase (Pol ε)
			Clamp loader (RFC)	Clamp loader (RFC)
	▭	Clamp loader (γ-complex)	Clamp (PCNA)	Clamp (PCNA)
	⬭	Clamp (β-clamp)		
5. Maturation	⬟		Fen1	Fen1
	◖	RNase H	Dna2	Dna2
	✚	DNA ligase	DNA ligase	DNA ligase

Figure 1.3 Functions and components of DNA replication machinery.
Figure adapted from [40]. The replicative helicase is recruited to the origins of replication (*Ori* /ARS) by a loader enzyme. Accessory proteins may encourage this process. Once the helicase enzyme is activated, a macromolecular complex known as the replisome forms. The helicase is followed by DNA polymerases which catalyse strand extension of the single stranded DNA substrate. Single stranded DNA binding proteins (SSB) protects ssDNA from degradation. As polymerases can only add nucleotides to the 3' end, the leading and lagging strands are synthesized in different directions relative to the helicase. The . leading strand is copied towards the helicase, whilst the lagging strand is copied away from the helicase.

family B polymerases [7,43,44]. The archaeal replisome represents a simplified version of its eukaryotic counterpart sharing various homologues for proper function. For example, archaea generally encode a single MCM gene, whilst all eukaryotes encode at least 6.

Particularly in many archaea and eukaryotes which maintain strict ploidy, the function of the replisome is tightly regulated in space and time to ensure that DNA replication occurs once per cell cycle. Recruitment and activation of the replicative helicase to origins of replication (*Ori* or Autonomously Replicating Sequence (ARS) in eukaryotes) is the primary regulation point for replication licensing [45]. Origins are located by the origin recognition complex (ORC) in G_1-phase that subsequently recruit MCM helicases [46]. After proper licencing, the helicases are activated, and cells enter S-phase. During DNA replication, the activity of the helicase and polymerases are tightly coupled to prevent excessive generation of deleterious ssDNA [47]. Single stranded DNA is further protected by single stranded binding protein (SSB/RPA) which is closely monitored by the DNA damage response pathway [48]. At the end of DNA synthesis, the helicase must be successfully disassembled to prevent re-replication [49]. Departure from controlled DNA unwinding by MCM in eukaryotes is linked with genomic instability that supports the formation of cancers [50,51]. Understanding the mechanisms that support proper DNA unwinding are therefore of fundamental importance.

1.5 Core biochemical properties of MCM

1.5.1 Eukaryotic MCM2-7

MCM genes were first identified in yeast screens that were defective for the licencing of DNA replication [52]. At least six of the identified eukaryotic MCM genes share homology and were subsequently numbered from 2-7 (MCM2-7) [53]. It is worth noting that studies often examine MCM2-7 from different organisms such as *Drosophila melanogaster*[54] and *Saccharomyces cerevisiae*[52]. Most of the core replisome features are expected to be conserved and therefore function similarly. From here on, these studies are collated to produce a comprehensive examination of core MCM2-7 properties.

Initial *in vitro* studies suggested that a subset of the MCM genes (4,6,7) were able to unwind DNA through hydrolysis of ATP[55]. ATP hydrolysis by MCM4,6,7 is also enhanced 2-fold by the presence of DNA [56]. Analysis of various forked DNA substrates revealed a preference for 3′ to 5′ translocation along DNA, juxtaposed with the 5′ to 3′ polarity reported for bacterial DnaB [41]. The MCM4,6,7 complex forms a stable toroidal hexamer in solution from a dimer of trimers and can preferentially bind to non-circular ssDNA [56,57]. It was later implied that mixtures of only MCM4 and 7 functioning as a trimer of dimers are required for robust DNA unwinding [56].

Whilst MCM4,6,7 is active *in vitro*, crosslinking experiments performed on *Drosophila* extract demonstrated all six MCM2-7 subunits interacting in an equimolar ratio [54]. Moreover, subunits interact in a precise circular order of 5,3,7,4,6,2 [54]. It therefore became doubtful that MCM4,6,7 represented the true replicative helicase for eukaryotes. Unusually, mixtures containing the full complement of MCM2-7 proteins lacked activity *in vitro*, where inhibition was provided by subunit 2 or a mixture of 3 and 5 [57,58]. Examination of subunit pairs *in vitro* by gel filtration, demonstrated weak interaction between subunits 2 and 5 [58]. The interaction between subunits 2 and 5 was later shown to be mediated by ATP hydrolysis and is further stabilized by replisome components such as Cdt1, Cdc45 and GINS [59–61]. This implies an important regulatory role for subunits 2 and 5.

Gel filtration analyses imply that the MCM2-7 hexamer is transient, however, the stability of the hexamer shows a strong dependency on experimental conditions. Temperature, salt and protein concentration are all able to alter the oligomeric state of MCM2-7 [60]. Importantly, activation of lone MCM2-7 complexes *in vitro* was demonstrated using an optimized buffer system. MCM2-7 exhibits robust DNA unwinding activity when tested in potassium salts containing organic anions such as glutamate [62,63]. Chloride salts typically used in biochemical experiments impose strong inhibition on the activity of MCM2-7. It was determined that the method of chloride inhibition is independent of DNA binding [62].

During the cell cycle, MCM2-7 interacts with both ds and ssDNA, however, *in vitro* the complex interacts strongest with ssDNA [59]. Like other hexameric helicases such as bacteriophage T4 gp41, DNA binding by MCM2-7 is dependent on the presence of nucleotides [64]. Unusually, ATP hydrolysis dependent association of MCM2-7 with DNA takes minutes, and stable association requires at least 5 times longer than the MCM4,6,7 subcomplex[59]. Values are generally too large to be attributed to diffusion alone and must be due to the process of a slow kinetic step.

Co-immunoprecipitation of MCM2-7 bound to stalled replication forks identified Cdc45 and GINS as factors involved in eukaryotic DNA unwinding [65]. Eukaryotic GINS is a tetrameric complex formed of 4 subunits, Sld5, Psf1, Psf2 and Psf3 [66]. The term GINS is derived from the Japanese translation of 5, 1, 2, 3 (Go-Ici-Ni-San). Both GINS and Cdc45 are essential for eukaryotic cell viability [66,67]. Supplementation of the two co-factors *in vitro* vastly stimulates the DNA unwinding activity of MCM2-7 and lead to the proposal that the eukaryotic replicative helicase functions as a macromolecular Cdc45-MCM2-7-GINS (CMG) complex [68]. In this unwinding configuration, MCM2-7 binds to forked DNA as a heterohexamer [68].

In vivo studies of the eukaryotic replisome suggest that replication forks progress at a rate between 1 to 3 kb/min [69]. This implies that MCM2-7 is capable of unwinding DNA at a rate between 16-50 bp/s. Recent single molecule experiments on CMG complex suggest that isolated CMG unwinds at a rate between 0.1-0.5 bp/s, many orders of magnitude lower than the theoretical rate [70]. Notably, CMG complex exhibits long pause states with back-tracking steps. When the pause states are removed from the analysis, the rate of CMG progression is near physiological at around 18-32 bp/s[70]. The duration of pauses is limited by increasing the concentration of ATP and force applied to the DNA substrate [70]. The number of pause events by CMG was also decreased substantially in the presence of the eukaryotic SSB homologue, RPA. This presumably physically prevents backtracking by CMG. With RPA the observed unwinding rate was between 4.5–15 bp/s [70,71]. Values calculated in the presence of RPA are not dissimilar to the rates determined in the presence of DNA-synthesis machinery, between 5bp/s [72] and 7bp/s [73]. There are likely two reasons for the similarities. First, whilst the precise mechanism is unclear, DNA polymerases provide a force on the DNA that improves MCM unwinding [70,74]. It is hypothesized that DNA synthesis induced forces may lower the energy barrier of strand separation at the fork [75]. Second, like RPA dsDNA synthesis prevents reannealing of ssDNA that may destabilise the fork. In the presence of RPA or DNA-synthesis, MCM2-7 unwinding rates approach values comparable with unimpeded CMG translocation along a single stranded DNA substrate of 10 bp/s [73]. Bidirectional replication from ~3-400 origins in a 12 Mb yeast genome, suggests an average of 15–20 kb of unwinding by MCM2-7 before termination [76]. Initial reports demonstrated low levels of processivity, between 0.8–0.9 kb [70,72]. Addition of SSB is cited to improve the processivity markedly to between 2.7–10 kb [71].

1.5.2 Archaeal homohexamer

The study of archaeal MCM has almost exclusively focused on enzymes from hyperthermophilic organisms (> 60 °C). Primarily, the enzymes are from *M. thermautotrophicus* and *Saccharolobus solfataricus* (formerly *Sulfolobus solfataricus*) [77].

From here on enzymes are referred to using the first letter of the genus name and the second two letters of the species (e.g., *Sso*MCM).

Unlike MCM2-7, initial characterization of *Mth*MCM revealed robust ATP-dependent DNA unwinding in the absence of cofactor proteins or specialized buffers [78,79]. Equally, *Mth*MCM exhibited DNA-stimulated ATPase activity [79]. Examination of *Mth*MCM unwinding various DNA substrates confirmed a preference for translocation in a 3′ to 5′ direction, like eukaryotic MCM2-7 [79]. Identical biochemical properties were also reported for archaeal MCM isolated from other thermophilic species, including *Archeoglobus fulgidus*, *Aeropyrum pernix*, *S. solfataricus* and *Pyrococcus furiosus* [80–84].

Archaeal MCM were proven to unwind blunt ended substrates, DNA-RNA hybrids and histone nucleosomes [85–87]. Importantly, these studies highlight an innate ability for MCM to bypass replicative stresses that may prevent fork progression such as bound RNA transcripts and DNA-binding factors. Although tight nucleoprotein complexes such as streptavidin-biotin blocks are sufficient to block fork progression *in vitro* [88]. Inhibition by streptavidin-biotin complexes show a dependency for the strand, where inhibition only occurs when the block is on the leading strand. The diameter of the central MCM hexamer channel is smaller than the diameter of streptavidin (~48 Å). Assuming MCM are unable to displace streptavidin-biotin complexes, this polarity preference suggests a strand exclusion model of DNA unwinding where the lagging strand of the fork does not directly enter the enzyme [88]. Mixed DNA substrates, where one phosphate backbone is replaced for a nonionic phosphorodiamidate linkage (morpholino DNA), does not support DNA unwinding by MCM [88]. This implies MCMs require electrostatic interactions with DNA to generate unwinding. A role for post translational modifications has also been implicated for archaeal MCM, where methylation of surface lysines increases both the activity and stability of the enzyme by around 2-fold [89].

Assuming bidirectional occurs from a single chromosomal origin of replication in archaea, MCM would be expected to unwind ~1 Mb for a single 2 Mb genome. Numerous *in vitro* studies highlight an intrinsic processivity that is many orders of magnitude lower, around 0.5 to 10 kb for *Mth*MCM [79,90,91]. For *Sso*MCM, isolated processivity is almost negligible, measured at ~30 bp [92]. Addition of the *Sulfolobus* replisome components Cdc45 and GINS greatly improves MCM processivity, possibly through improved interactions with the DNA substrate [61,84,92,93]. All archaea encode a single GINS gene with homology to Sld5/Psf1 (Gins51) [93]. A small number of archaea encode a second GINS gene with resemblance to

Psf2/Psf3 (Gins23) [94]. Like eukaryotic MCM2-7, synthesis coupled unwinding by introduction of the DNA polymerase further improves the processivity of *Sso*MCM [92]. Introduction of archaeal SSB also improves unwinding rates of *Sso*MCM *in vitro* [82]. Single molecule studies demonstrate that archaeal MCM unwinds DNA between 52 and 162 bp/s, much faster than reports for MCM2-7 [91]. This may compensate for the larger theoretical processivity required for archaea which possess fewer *Ori*'s per Mb [95].

Unlike MCM2-7 and other replicative helicases, archaeal MCM generally form stable homohexamers in solution without supplementation with ATP. This has been established for numerous MCM, including: *Afu*MCM [81], *Ape*MCM [80], *Mbu*MCM [96], *Pto*MCM [97], *Sso*MCM [82,98–101] and *Tac*MCM [102]. Initially *Mth*MCM was reported as a double hexamer, interacting in a head-to-head fashion [78,79]. However, like the transiency of the MCM2-7 complex, *Mth*MCM oligomeric state is strongly influenced by experimental conditions. Increasing the temperature of the buffer to within the growth range of *M. thermautotrophicus* (~60 °C) shifts the population of double hexamers towards single hexamers [103,104]. Equally, salt and protein concentration also affect the oligomeric state [104]. Analogously by increasing the protein concentration it is possible to reconstitute a population of *Sso*MCM double hexamers [83]. Observations under negative stain EM suggest that at low temperature, archaeal homohexamer MCM form closed rings. When the temperature is increased to near native growth conditions, the enzyme population is dominated by open rings [103,105]. These are all expected physiological conformations of MCM and will be discussed in further detail later. Other oligomeric states, which are not regarded to be physiological have also been reported such as, heptamers, octamers and filaments [104,106,107].

Extensive work on both *Sso*MCM and *Mth*MCM reveal strong DNA binding with a preference for ssDNA over dsDNA [79,98,108,109]. Examination of various mixed length ssDNA-dsDNA substrates demonstrate *Sso*MCM interacts tightest with forked DNA when both 3′ and 5′ ssDNA tails are present [110]. MCM are expected to interact primarily with the 3′ leading strand. Curiously, increasing the length of the 5′ tail decreases the binding affinity of MCM proportionally with length of the tail. At the same time, when competitor strands are added the presence of a 5′ tail reduces the enzyme dissociation rate [88]. Taken together, this suggests that a forked substrate increases the stability of MCM on DNA. It is important to note that MCM-DNA binding affinities should be assessed with caution. When measuring DNA binding affinities for MCM, the protein concentration is generally serially diluted, which likely also shifts the oligomer equilibrium (discussed previously). Furthermore, given the capability of MCM to bind to either ssDNA or dsDNA, when forked substrates are assessed MCM are exposed to multiple possible binding sites.

1.5.3 A historical perspective: Why study archaeal MCM?

At least until the successful optimization of MCM2-7 buffer conditions, elucidation of the mechanistic properties of eukaryotic MCM2-7 in isolation was very difficult. Consider that the determination of mechanistic relationships in molecular biology requires two main components.

First, to determine the active function of a specific amino acid residue or structural motif, an ability to monitor some baseline functional activity of the enzyme is required. In the case of eukaryotic MCM, it took until 2008 to prove robust activity of an isolated MCM2-7 *in vitro* [62]. Further, to isolate clear atomic level mechanisms, it is generally beneficial to limit the number of components in the system to avoid convolution of phenotypes; MCM2-7 requires many co-factors *in vitro* for activity.

Second, a structural knowledge of the enzyme is required. Traditional structural techniques such as X-ray crystallography demand high yield and pure protein alongside homogeneity of structure. These factors are essential for the formation of a regular crystal lattice. The apparent transiency of the MCM2-7 complex, its large size and subunit heterogeneity make crystallization unfeasible [111]. Further, yields from recombinantly expressed MCM2-7 are reported as low as 82.5 ug/L from *E. coli* [63]*.* This also eliminates the possibility of many enlightening biophysical techniques, such as NMR and AUC.

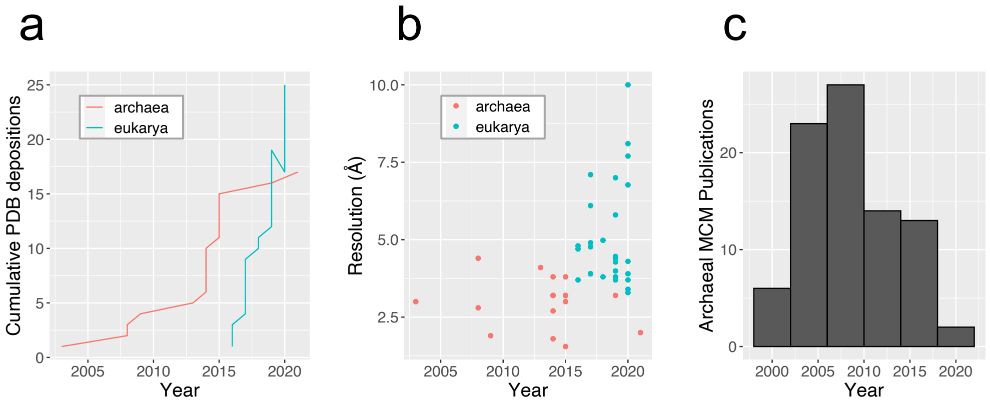


Figure 1.4: Historical perspective of MCM studies and the emergence of cryo-EM. **(a)** Cumulative depositions in the PDB (https://www.rcsb.org) for archaeal and eukaryotic MCM structures. Archaeal structures are derived from X-ray crystallography, eukaryotic structures are derived from cryo-EM. **(b)** Resolution comparison by year of archaeal and eukaryotic MCM structures. **(c)** The number of publications focusing on archaeal MCMs by year.

Therefore, archaeal MCM that could be purified to high homogeneity and yield from recombinant *E. coli* expression were of great interest. Whilst the first high resolution archaeal MCM structure was published in 2003, it would take over a decade and the development of cryo-EM to resolve the atomic structure of MCM2-7 (Figure 1.4a) [112]. The recent ability to examine eukaryotic MCM2-7 by cryo-EM, albeit at a generally lower atomic resolution is likely responsible for the waning interest in studying archaeal MCM within the last decade (Figure 1.4b–c).

1.6 Atomic structure of archaeal MCM

1.6.1 Subunit structure

The structure of MCM subunits can largely be split into 2 domains: The N-terminal DNA-binding and oligomerization domain; the C-terminal ATPase domain (Figure 1.5a–b). The N-terminal domain can be further divided into 3 subdomains. Subdomain A forms a helical bundle that is involved in coordination of the free 5′ strand of DNA; deletion of this

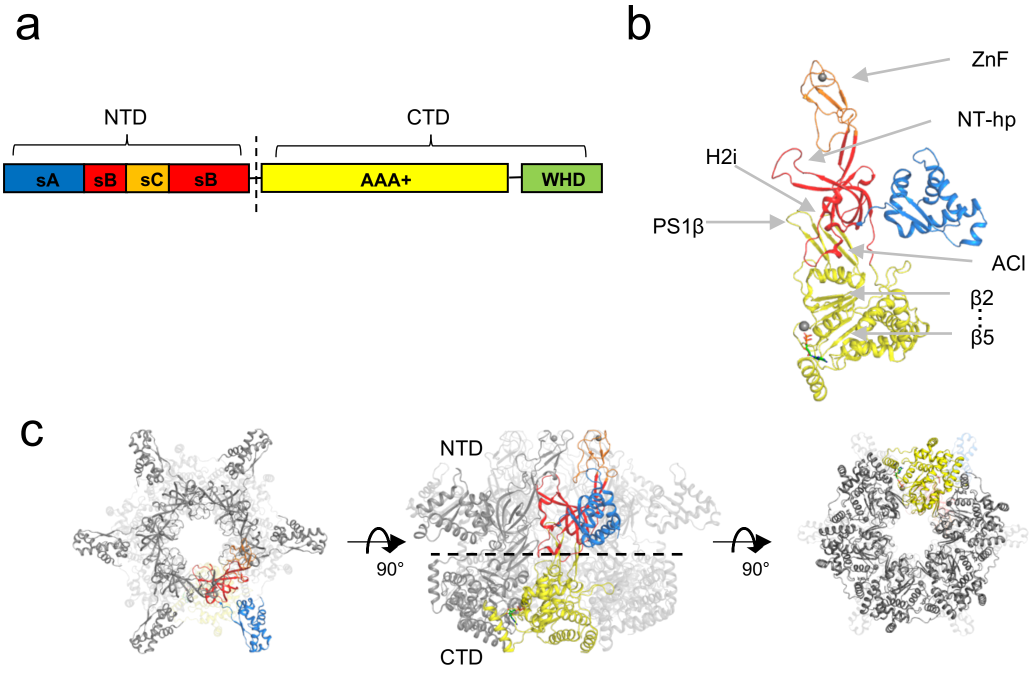


Figure 1.5: Structure of an archaeal MCM.
(a) Relative organization of MCM domains and subdomains. Domains, NTD: N-terminal domain, CTD:C-terminal domain. Subdomains, sA: subdomain A, sB: subdomain B, sC: subdomain C, AAA+: ATPase associated with various cellular activities, WHD: winged helix domain. **(b)** Structural organization of MCM subdomains and functional motifs (PDB:4R7Y)[124]. Protein is visualized as a ribbon diagram.Subdomain coloring is outlined in part (a). Motifs, ZnF: zinc-finger, NT-hp: N-terminal hairpin, ACl: allosteric communication loop, PS1β: pre-sensor 1 β-hairpin, β2…β5: shortened order of beta-sheets, β2:β3:β4:β1:β5. **(c)** Structural organization of subunits within an MCM hexamer (PDB: 4R7Y)[124]. Protein is visualized in the ribbon format, whilst subdomains are coloured as outlined in part (a).

subdomain impedes DNA-binding [110]. Subdomain B contains an essential Zinc Finger Motif (ZnF), which is critical for helicase function [113]. The precise identity of the ZnF motif is generally C4 or C3H-type, where a zinc atom is coordinated by tetrahedral geometry within a zinc ribbon fold [114,115]. Although the ZnF has been implicated in DNA-binding and oligomerization, its precise function is yet to be clearly resolved [98,113,116]. Subdomain C forms an oligosaccharide/nucleotide-binding (OB) fold that harbours residues essential for both MCM hexamerization and DNA-binding. The C-terminal domain contains an ATPase-Associated with various cellular Activities (AAA+) fold, which provides all of the ATP binding and hydrolysing motifs required for helicase function [117]. At the distal end of the C-terminal domain is a Winged Helix Domain (WHD), that appears to restrain ATPase and helicase activity in absence of regulatory co-factors such as ORC [83,105]. The WHD is likely involved in the initial recruitment and loading of MCM double hexamers onto *Ori*'s [105]. No full length archaeal MCM structure has been resolved with the flexible WHD present, although NMR structures have been solved for the domain alone [118].

1.6.2 Oligomerization

Numerous hexamer structures for archaeal MCM have been reported, however only 2 thus far with both the N-terminal and AAA+ domains intact (Figure 1.5c). Studied independently, the N-terminal domain is able to form hexamers *in vitro*, whilst the C-terminal domain cannot [83]. This suggests a critical role of N-terminal residues in MCM oligomerization, where it is likely that multiple motifs are involved.

The ZnF has been implicated in oligomerization, where removal of the subdomain in *Sso*MCM prevents hexamerization of the N-terminal domain [119]. This interaction however does not involve the zinc-ion itself, where mutation of the conserved ZnF cysteine residues to serine maintains the integrity of the hexamer[113]. Instead it is suggested that main chain atoms of the ZnF ribbon form a hydrogen bond network with neighbouring ZnF ribbons to support oligomerization [119]. The ZnFs are also important for the formation of a MCM double hexamer [112]. Removal of a hexamer-hexamer salt bridge in *Mth*MCM ZnF is sufficient to prevent the formation of a double hexamer *in vitro* [120].

The N-terminal OB-fold is also directly involved in oligomerization, where removal of a conserved T-shaped π-interaction (F179A) in *Pfu*MCM prevents hexamerization [119]. Importantly, mutation of the equivalent phenylalanine to isoleucine in mouse MCM4 is associated with an increased incidence of mammary adenocarcinoma, highlighting the importance of MCM oligomer stability for genomic integrity in higher organisms [121,122]. Removal of intersubunit hydrogen bonds in the OB-fold is also cited to reduce oligomerization propensity [123]. DNA-binding to the OB-fold is suggested to further improve hexamer stability, although the precise mechanism is unclear [119]. One hypothesis is that repulsive positive charges between basic residues of neighbouring OB-fold N-terminal-hairpins (NT-hp) are neutralised by the negatively charged phosphate backbone of DNA.

Whilst the C-terminal domain does not hexamerize in isolation, disruption of interactions in the AAA+ fold can affect the oligomeric state of the entire protein. Removal of conserved intersubunit salt bridges on neighbouring subunit alpha-helices in *Sso*MCM disrupts hexamerization [99]. Removal of other conserved residues including surface tyrosines interferes with oligomerization, however, as these residues are not positioned at interfaces the oligomeric differences are likely a result of altered conformational shape of subunits [101,123].

1.6.3 DNA binding

The diameter of the MCM central channel is wide enough to accommodate both single and double stranded DNA (Figure 1.5c) [124]. This is consistent with the notion that MCM encircles dsDNA during initiation of DNA replication and tracks along ssDNA during elongation. It is likely that different portions of the hexamer are involved in coordinating these two discrete events. During elongation, MCM are expected to proceed towards the fork N-terminal first according to a steric exclusion and wrapping model (SEW) [88,125]. Here, the leading strand is encircled by the MCM ring, and separated from the lagging strand. Hydrogen deuterium exchange mass spectrometry has revealed interaction of the lagging strand along a discrete path of the external surface of the MCM, supporting the SEW model [126]. Indeed, atomic force microscopy demonstrated that MCM binds to DNA and generates regularly angled kinks in the strands, consistent with DNA wrapping [127]. So far, atomic structures of archaeal MCM have been reported with ssDNA bound in the central channel of both the N and C-terminal domains [100,128]. Critically, DNA-binding is dependent on the oligomeric state, where hexamerization deficient mutants are unable to interact with DNA [119].

MCM are widely regarded to translocate N-terminus first and is therefore the initial contact point of parental DNA within the replisome. Analysis of the crystal structure of the N-terminal portion of *Pfu*MCM bound to ssDNA, suggests core beta-sheet residues of the OB-fold are strongly involved in DNA binding [128]. As demonstrated for other hexameric helicases, such as E1, DNA is expected to bind perpendicular to the tiers of the ring, however, in *Pfu*MCM ssDNA occupies a planar orientation relative to the OB-folds tiers of the hexamer (Figure 1.6a,c) [128–130]. The orientation of the ssDNA is also perpendicular relative to homologous ssDNA bound OB-fold structures such as SSB [131]. Absence of the C-terminal AAA+ domain from the structure possibly opens the accessibility of the OB-fold and it is therefore unknown whether this represents a true biological conformation. The authors suggest that the *Pfu*MCM-ssDNA structure may reflect a conformation during

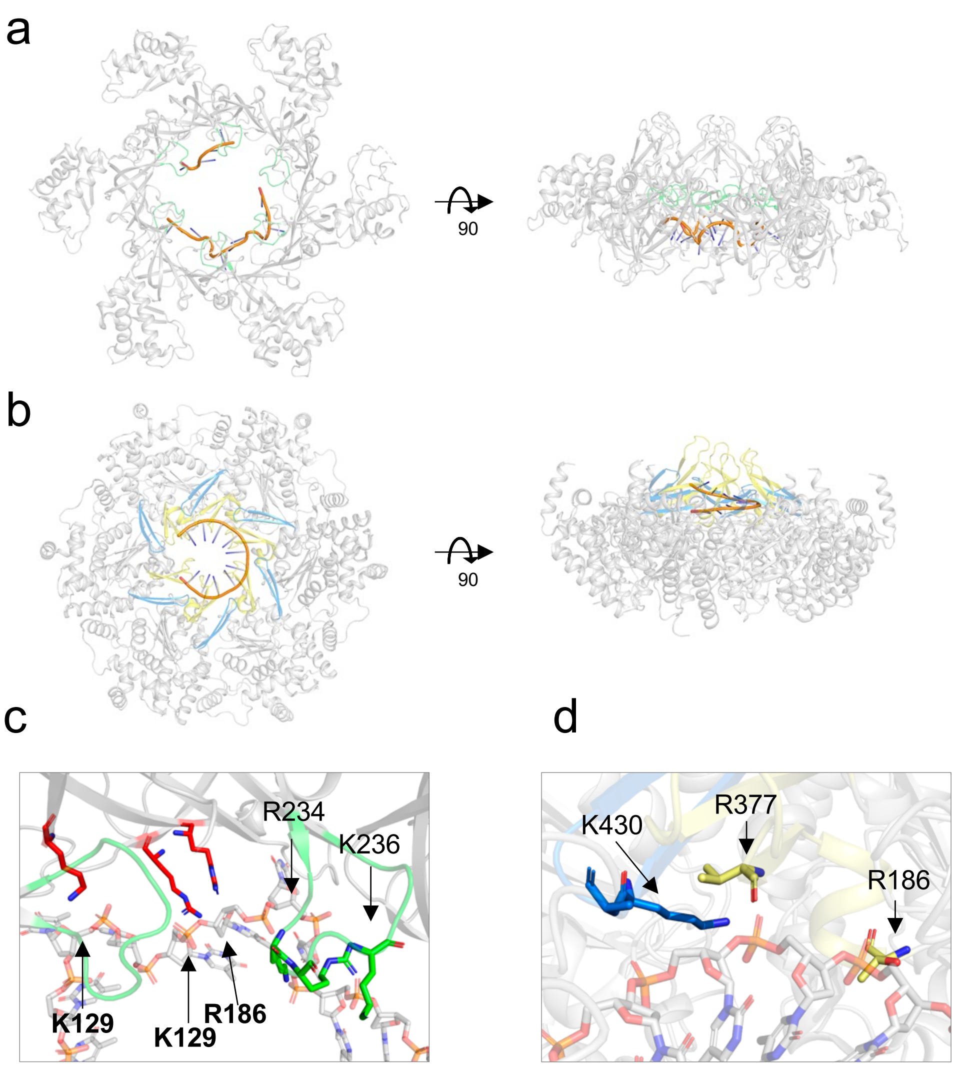


Figure 1.6: Co-ordination of DNA binding by archaeal MCM tiers.
(a) Views of the N-terminal domain of *Pfu*MCM bound to ssDNA (PDB: 4POG)[128]. N-terminal hairpins are colored in green; the DNA-backbone is colored in orange. Protein is visualized as a ribbon diagram. **(b)** Views of the C-terminal domain of *Sso*MCM bound to ssDNA (PDB: 6MII)[100]. Pre-sensor 1 β-hairpin is colored in blue; helix-2 insert hairpin is colored in yellow. Protein is displayed as a ribbon diagram.**(c)** Positioning of important ssDNA-binding residues in *Pfu*MCM N-terminal domain. Core basic OB-fold residues coordinating ssDNA are colored in red. Residues of the N-terminal hairpin that do not appear to co-ordinate DNA in the structure are colored in green. **(d)** Positioning of potential ssDNA binding residues of the C-terminal domain. Coloring as part (b).

replication initiation rather than DNA unwinding, although the role of the OB-fold in dsDNA binding is minimal, as mutants retain binding capacity [128].

Alternatively, numerous publications have demonstrated the importance of the NT-hp for DNA unwinding. Truncation of the entire NT-hp results in complexes that are unable to bind to ssDNA [112,119]. Removal of 2 basic residues in *Sso*MCM NT-hp are sufficient to reduce the unwinding activity 8-fold and ssDNA binding 10-fold [132,133]. Therefore, other structures will be required to further elucidate the functional roles of the N-terminal domain in DNA unwinding.

The recent structure of *Sso*MCM AAA+ domain bound to ssDNA validated the importance of two C-terminal DNA-binding hairpins in MCM activity (Figure 1.6b,d) [100]. These are the pre-sensor-1-beta hairpin (PS1β) and the helix-2 insert (h2i). Removal of a conserved basic lysine residue within the PS1β loop was previously shown to inhibit unwinding by *Sso*MCM [100,132]. Each PS1β hairpin is observed to coordinate 2 bases of ssDNA in a staircase-like fashion around the ring. The lysine residue (K430) on the PS1β coordinates the phosphate backbone of DNA through ionic interaction, whilst the neighbouring alanine (A431) interacts with the phosphate backbone through main-chain hydrogen bonding (Figure 1.6d).

The helix-2 insert is implicated in coupling ATP hydrolysis to DNA binding. Notably, removal of *Mth*MCM h2i results in a construct that binds ssDNA 40-fold tighter than the wild type enzyme but is unable to couple ATP hydrolysis to DNA unwinding [90]. Within the *Sso*MCM structure, threonine, and valine residues of the h2i appear to form hydrogen bonds with the DNA-backbone. Mutation of the *Sso*MCM h2i threonine to alanine is insufficient to reconstitute the phenotype observed in the *Mth*MCM study [100]. It is considered that as the h2i is not well conserved it may perform a purely mechanical function that is independent of precise amino acid interactions [90]. In this case, mutation of the threonine to alanine would not shorten the physical structure of the h2i loop and it would therefore retain mechanical function.

1.6.4 ATP hydrolysis driven hairpin movement

The MCM ATPase domain exhibits a classical AAA+ fold. At its core 5 parallel beta sheets are arranged into the precise order β5-β1-β4-β3-β2 and are flanked by essential motifs [134]. Crucially, ATPase active sites form at the interface between neighbouring subunits, where oligomerization ensures the full complement of motifs required for ATP binding and hydrolysis. The Walker A (GxxGxxK[T/S]), Walker B (DExx) and sensor 1 act in *cis*-, whilst the arginine finger and sensor 2 act in *trans* (Figure 1.7a) [99,124]. Activity of the AAA+ fold is extremely sensitive to mutations of any key motif and hence underpins its high conservation between MCM [99].

The precise mechanism of ATP hydrolysis has largely been inferred from studies of highly conserved hexameric AAA+ family members. A magnesium ion in the active site cleft coordinates the binding of ATP to both the Walker A and B motifs [135]. The arginine finger acts to bind and polarize the γ-phosphate to enhance hydrolysis [135].

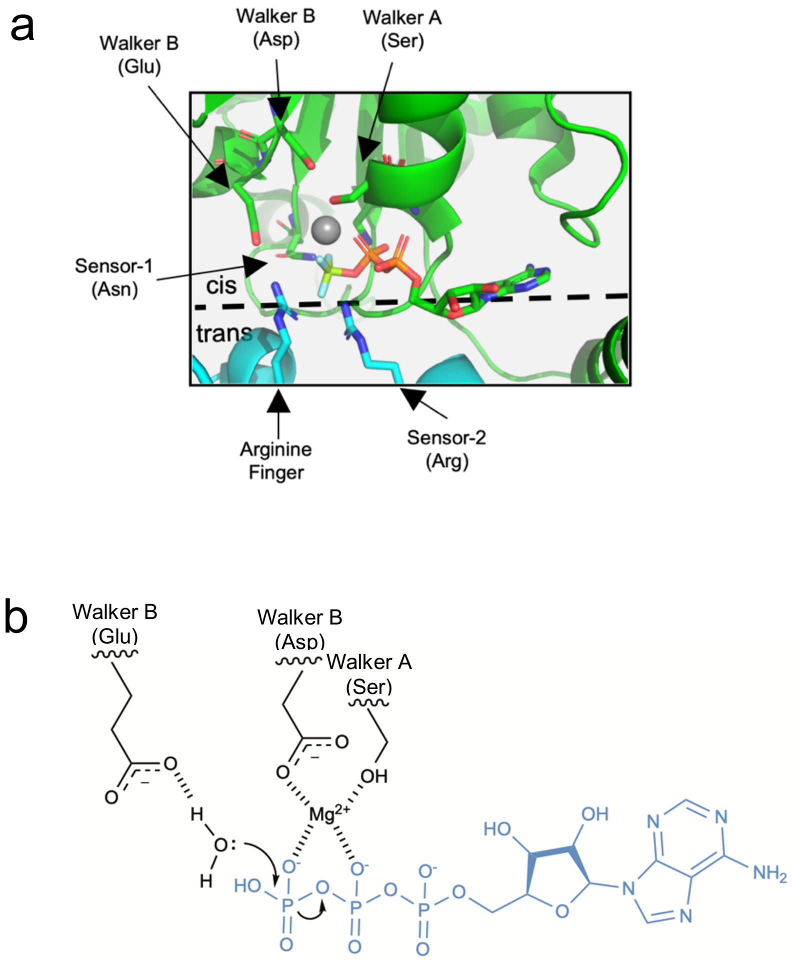


Figure 1.7: ATPase site formation and mechanism of ATP hydrolysis.
Conserved residues from the Walker A (GXXXXGKT/S) and Walker B (DExD) motifs co-ordinate ATP hydrolysis. **(a)** Active site of *Sso*MCM bound to ADP•BeF$_3^-$ (PDB:6MII)[100]. The conserved serine and aspartate residues co-ordinate a magnesium ion (grey sphere) with the β- and γ-phosphates of ATP. The protein is visualized in the cartoon format, where important residues and ligand are highlighted as sticks. **(b)** Proposed mechanism of ATP hydrolysis. The conserved glutamate residue activates a water molecule, which attacks the γ-phosphate of ATP. The density for water is missing in the 6MII, structure.

Subsequently, a molecule of water is activated by the conserved Walker B glutamate for nucleophilic attack on the γ-phosphate of ATP (Figure 1.7b) [136]. The sensor and arginine finger motifs are thought to bind and monitor the changes in the state of the γ-phosphate of ATP to generate conformational change throughout the protein [134].

The allosteric communication loop (ACl) is believed to transmit conformational change from the AAA+ domain to the N-terminal domain. The ACl projects as a surface extension of the OB-fold onto the base of the C-terminal hairpins of the neighbouring AAA+ subunit (Figure 1.5b). Interestingly, removal of the ACl loop in *Sso*MCM inactivates unwinding by the enzyme without inhibition of DNA binding [137]. Subsequent removal of the N-terminal hairpin is sufficient to restore activity of the enzyme, suggesting that the ACl is required to mediate conformational change in the N-terminal domain [137].

Mutant doping studies on *Sso*MCM where catalytically inactive ATPase mutants are titrated into the experiment demonstrate that MCM are tolerant for up to 3 inactive sites within a hexamer [99]. This is in stark contrast to other hexameric helicases, such as bacteriophage T7, which do not tolerate inactive subunits at all [138]. This suggests that unlike bacteriophage T7, which requires sequential hydrolysis of ATP between neighbouring active sites, MCM does not, raising the model of semi-sequential ATP hydrolysis [99]. Whilst this hypothesis suggests that MCM can 'skip' inactive subunits during hydrolysis around the ring, we note that this data may equally point to an intrinsically asymmetric mechanism of DNA unwinding that requires only a portion of MCM subunits to be active.

1.6.5 Structural mechanisms of archaeal MCM hexamer function

The lack of structural studies capturing full length archaeal MCMs in clear functional conformations currently limits understanding of the mechanisms homohexameric MCMs use to unwind DNA. Elucidating archaeal MCM mechanisms currently requires an anecdotal patchwork of structural experiments linked to (generally) bulk biochemistry. Only two, near full-length archaeal hexamer MCM structures exist[100,124]. Of these structures, one displays strict C6 symmetry [124], whilst the other displays asymmetry caused by DNA and nucleotide occupancy in the AAA+ domain[100]. Recent expansion of cryo-EM has allowed atomic insight into the conformational changes leading up to and including DNA replication for MCM2-7. Arguably, we now understand more about the mechanisms of MCM2-7 than the simpler archaeal MCM complex. Importantly, the intrinsic asymmetry of the MCM2-7 complex appears to be central to both regulation and mechanism. This is beginning to raise important questions about the evolution of the complex from archaea to eukaryotes.

1.7 Functional asymmetry within eukaryotic MCM2-7

1.7.1 Unique roles of subunits

The initial atomic structure of MCM2-7 was elucidated in 2015 [139]. The structure of an archaeal MCM was used to initiate model building into the density maps [139]. The atomic structure was later refined again to improve the fit into the density and the geometrical

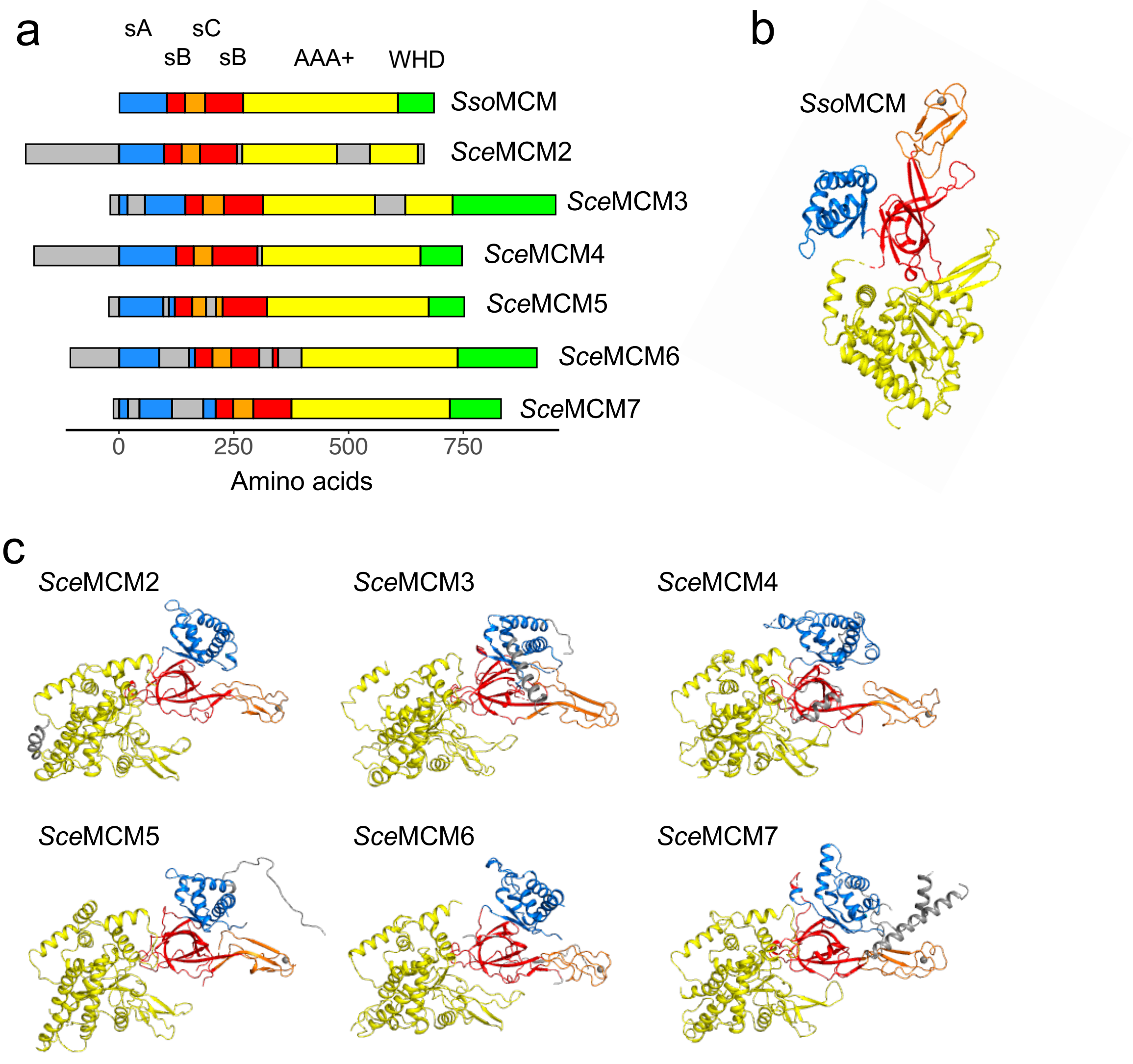


Figure 1.8: Subunit specific features of MCM2-7.
(a) Relative organization of MCM domains and subdomains. Subdomains, sA: subdomain A (blue), sB: subdomain B (red), sC: subdomain C (orange), AAA+: ATPase associated with various cellular activities (yellow), WHD: winged helix domain (green). MCM specific insertions are colored in grey. **(b)** The structure of an archaeal *Sso*MCM subunit (PDB: 6MII)[100] represented in the ribbon format. Subdomain coloring as in part (a). **(c)** The structure of each yeast *Sce*MCM subunit (PDB:6EYC)[140] visualized in the ribbon format. Subdomain coloring as in part (a).

parameters of the model [140]. The core structure of the six MCM2-7 subunits is roughly similar to archaeal MCM, with minor variations between (Figure 1.8a-c). Like archaeal crystallographic studies, the structure of the WHDs is absent from the calculated density maps, implying the subdomain is flexible (except for MCM2 that lacks a WHD). Further, MCM3 does not contain the motifs to form a proper ZnF and appears 'collapsed' in structure [39]. Each of the MCM2-7 subunits contain signature N-terminal (NTE), C-terminal Extensions and N-terminal (NTI), C-terminal insertions (CTI) relative to one another and archaeal MCM. These insertions are all implicated in numerous unique regulatory and mechanistic functions, although most insertions are not observable in the electron density.

1.7.2 Cell cycle regulation

Importantly, many insertions contain target residues for phosphorylation by specific regulatory kinases. For example, the NTE of MCMs 2 ,4 and 6 all contain proven phosphorylation sites that are involved in MCM2-7 activation. For MCM4, site specific phosphorylation promotes the association of the replisome component Cdc45 [141]. Further, phosphorylation of GINS promotes the assembly of an active CMG complex [142]. It is widely regarded that phosphorylation generates conformational change. Mutation of MCM5 proline 83 to leucine (*'Bob1*'), permits replisome licencing in lieu of kinase activity [143]. *Sce*MCM5-P83 is positioned at the intersection of subdomains A and B. Therefore, mutation likely mimics a local phosphorylation event that would reposition the subdomain into a proliferative conformation [144]. The roles of many putative MCM phosphorylation sites involved in regulation are reviewed extensively elsewhere [145].

Aside from phosphorylation, MCM4 CTD has been directly implicated in the DNA-damage checkpoint pathway [146]. Removal of the MCM4 CTD delays recovery after synthesis-mediated replication stress. Both MCM2 and MCM3 are known to encode important nuclear localization signals (NLS) [147]. MCM2 traffics to the nucleus independently of MCM3-7. MCM3-7 requires a further NLS, supplied by MCM binding protein (MCM-BP) [50]. In the absence of suitable NLS or MCM-BP, MCM components are rapidly degraded in the cytoplasm[50]. Taken together, many of the structural NTE, NTI and CTEs of MCM2-7 are implicated in diverse eukaryotic processes.

1.7.3 Ordered heterohexameric assembly

The initial structure of yeast MCM2-7 confirmed the stoichiometry and precise order inferred from previous biochemical studies [54,139] (Figure 1.9). Analysis of individual subunit-subunit interfaces reveals the contacts that support ordered assembly of MCM2-7. As with archaeal MCM, oligomerization motifs are primarily located within the N-terminal domain (Figure 1.9).

Sequence insertions and extensions are core to the asymmetric interactions: MCM5 NTE contacts subdomain A of MCM3; the NTI helix of MCM7 forms a long contact with the collapsed ZnF of MCM3, whilst the NTE of MCM7 forms an association with subdomain A of MCM3; a helical NTI of MCM7 interacts with an NTE of MCM4 subdomain A; an NTE loop from MCM6 forms a structural insertion between the ZnFs of both MCM4 and MCM7 (Figure 1.9) [139]. Other unique interfaces exist without structural insertion, for example, the NT-hp of MCM3 interacts with the OB-fold of MCM5. MCM2 is the only MCM subunit not involved in the formation of a broad unique interface. This structural deficiency of MCM2 may underpin independent trafficking of MCM2 to the nucleus [50].

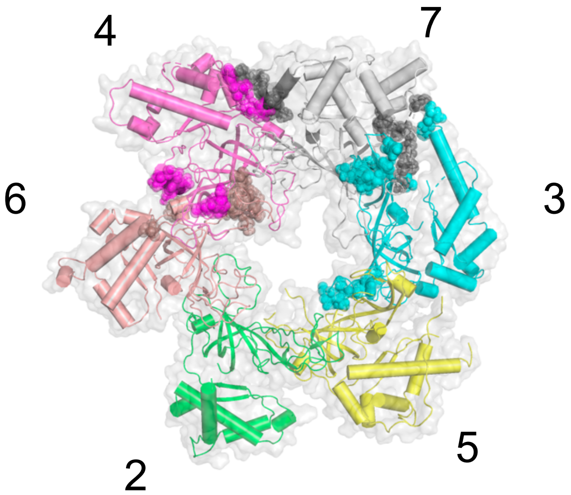


Figure 1.9: Unique interfaces of the MCM2-7 hexamer.
Top-down view of the N-terminal domain of *S. cerevisiae* MCM2-7 heterohexamer (PDB:6EYC)[140]. The presence of unique interfaces between neighboring subunits is highlighted by the representation of spheres between interfaces. Protein is visualized in the cartoon format, where alpha helices are represented by cylinders and the surface is shown in transparent grey.

1.7.4 Specific contributions to DNA binding

All MCM2-7 subunits contain the full complement of DNA-binding motifs observed in the archaeal MCM structures. In comparison with archaeal MCM, there are limited studies examining the biophysical contribution of MCM2-7 DNA-binding motifs through mutagenesis.

Yeast strains are viable when single basic residues in the ssDNA binding OB-fold are substituted to alanine for individual subunits within the MCM2-7 complex [128]. A lethal phenotype is however observed when this mutation is applied to more than one subunit within the heterohexamer [128,148]. An origin loading defective phenotype was also observed when 3 basic residues of the MCM5 OB-fold were mutated to alanine [148].

Mutation of the conserved PS1β basic residue (equivalent to *Sso*MCM K430) in MCM4 to alanine generates no defect in the growth of the strain [149]. Comparably, mutation of the equivalent position in MCM5 yields strains with significant growth defects [149]. Interestingly, neither strain is deficient for the loading of MCM2-7 onto chromatin, suggesting that the loss of fitness phenotype is possibly from DNA unwinding steps that follow replication initiation [149]. Taken together the mechanistic and fitness contributions of each of the DNA binding motifs in eukaryotic MCM2-7 are not necessarily equal.

1.7.5 Specific contributions to ATP Hydrolysis

The ATPase domain of MCM2-7 subunits is conserved, where all sites contain the full complement of motifs required for ATP binding and hydrolysis. However, alongside DNA-binding hairpins, each ATPase active site has distinct roles in DNA unwinding and regulation.

Overexpression of ATPase motif mutant alleles have distinct effects on the viability of cells. Primarily the Walker B motif and arginine finger were focused on that coordinate hydrolysis rather than ATP binding [150]. Both mutants are sufficient to generate inactive ATPase interfaces [151]. Overexpression of mutants at interface 2-6 and 3-5 are viable and *in vitro* analysis suggests that MCM2-7 complexes are active for ATP hydrolysis [151]. Overexpression of mutant subunits of the remaining interfaces exhibits a lethal dominant response when at least one of the *cis*- or *trans*- ATP hydrolysis motifs is deficient[151]. The role of MCM7 appears to be particularly important where overexpression of any MCM7 ATP hydrolysis mutant elicits a lethal phenotype on cells [151]. Further, mutation at either 7-3 or 4-7 interfaces *in vitro* inhibits the ATPase activity of the entire MCM2-7 complex. Consistently, cellular regulators such as retinoblastoma have been demonstrated to tune the ATPase activity of MCM7 to control replisome function [152].

Analysis of the ATP-binding capacity is somewhat more complicated as ATP binding also affects the hydrolysis. This is typically assessed through mutation of the conserved Walker A lysine to alanine (KA). When tested *in vitro*, mutation of ATP binding by the 5-2 or 3-5 interfaces are particularly inhibitory for ATP hydrolysis by entire MCM2-7 complex [153]. Interestingly, overexpression of either MCM5 or MCM3 KA mutants was not lethal *in vivo* [151]. This may be explainable through competitive inhibition. As the interaction between subunits 5 and 2 is dependent on ATP binding and hydrolysis, if MCM5-KA cannot bind to ATP through mutation, it will also be unable to interact with MCM2 [59]. Therefore, native MCM5 will outcompete the overexpressed MCM5-KA mutants and cells will remain viable.

Taken together, these works highlight unique roles for the distinct ATPase active sites in MCM2-7. Importantly this contradicts the semi-sequential ATPase hydrolysis model suggested for archaeal MCM, where up to 3 mutants are tolerated [99]. The discrete roles of specific ATPase sites in the steps leading up to DNA replication have also been elucidated [154]. Of particular focus is the 5-2 ATPase site which is heavily implicated in MCM loading onto dsDNA. The importance of this site has recently been inferred from cryo-EM studies that have captured the various conformational states of MCM2-7.

1.7.6 Structural asymmetry in MCM2-7 function

1.7.6.1 Recruitment of MCM2-7 Cdt1 by ORC1-6 to replication origins

During eukaryotic replication licencing, the Origin Recognition Complex (ORC1-6) identifies and encircles origins of replication [155]. ORC bends the dsDNA and is subsequently activated through interaction with Cdc6 [156,157]. ORC-Cdc6 then recruits and loads two open form Cdt1-MCM2-7 complexes that assemble into a double hexamer [156]. The structure of the open complex was resolved in the presence of non-hydrolyzable ATP analogue AMP-PNP and forms a structurally asymmetric left-handed helical structure (Figure 1.10) [158]. Notably, the ring is cracked by a large ~10 Å gap between subunits 5 and 2. This is believed to be the entry point for dsDNA into the hexamer, since artificial ligand mediated dimerization of the 5-2 interface inhibits loading of MCM2-7 onto dsDNA [159] . In the open conformation, the WHD of MCM5 projects into the central channel of MCM, whilst other WHDs project away from MCM [158]. MCM5 WHD may sterically block interaction of MCM5 and 2 and the passage of dsDNA into the central channel. Cdt1 interacts and encompasses the external surface of subunits 2,6,4. In absence of Cdt1, MCM2-7 is still able to form an open complex, however this interaction improves the stability of the open conformation, and hence the efficiency of loading onto dsDNA [158,160]. Cdt1 possibly also serves to mediate specific

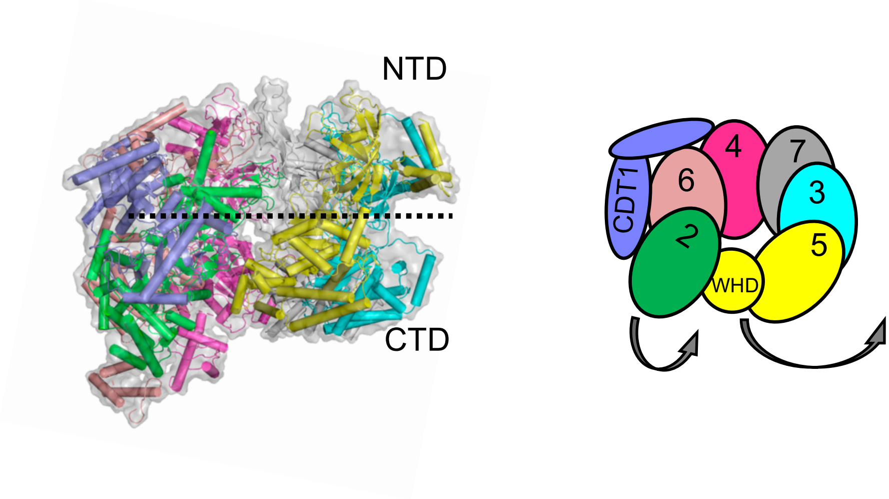


Figure 1.10: Structure of the open complex of MCM2-7.
The open-conformation of Cdt1-bound MCM2-7 (PDB: 5XF8)[158]. The identity and coloring of each subunit is clarified in the righthand panel. Protein is represented in the cartoon format, where alpha helices are represented by cylinders and the surface is shown in transparent grey.

interactions between subunits 2 and 6 (discussed previously) and prevent premature interactions with Cdc45/GINS [158].

Closure of the MCM2-7 hexamer around dsDNA is dependent on the hydrolysis of ATP at the 2-5 gate [60,160,161]. The kinetics of this process is notably slow, taking between 5-30 minutes for stable hexamer formation [59,60,161]. Presumably slow kinetics are dependent on conformational change between subunits 2 and 5. Loading of the complex onto dsDNA by ORC-Cdc6 has been captured using the slowly hydrolysable ATP analogue ATPγS (Figure 1.11) [162]. Extensive subunit specific interactions occur between the 6 Orc subunits and the 5 WHDs of MCM2-7. Biochemical and structural analyses have demonstrated that MCM3 WHD plays a fundamental role in the initial recruitment of MCM2-7 by ORC-Cdc6 [160,163]. The WHD of MCM3 and MCM7 interact primarily with Cdc6, whilst the WHD of MCM 6 and 4 interact directly with ORC [163,164]. MCM6 WHD is able to inhibit ATP hydrolysis dependent ring closure when Cdt1 is absent [161,165]. The MCM6 WHD is also important for complex activation, as truncation also delays dsDNA loading substantially [160]. ORC-Cdc6 then positions and directly threads dsDNA through the MCM2-5 gate [162]. Fundamentally, this remodelling process by ORC-Cdc6 results in movement of the MCM5 WHD from the central channel of MCM and allows closure of the MCM2-5 gate (Figure 1.11) [162,163]. ATP hydrolysis at the MCM2-5 gate generates ring closure and release of Cdt1 and Cdc6. This process occurs twice at two neighbouring inverted sites [157]. Orc6 subsequently mediates the homodimerization of two closed MCM2-7 hexamers in a 'head-to-head' configuration, analogous to the archaeal *Mth*MCM dodecamer [156,157].

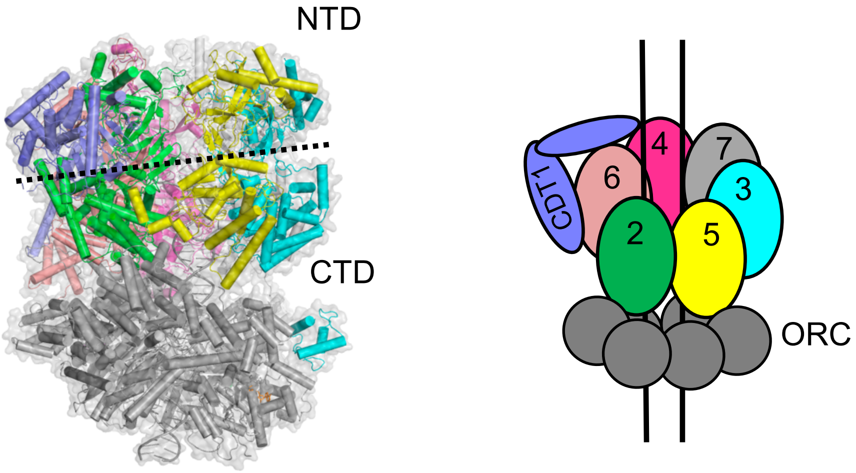


Figure 1.11: MCM2-7 closure around dsDNA.
Structure of the Cdt1-MCM2-7 complex partially loaded onto dsDNA by ORC (grey) (PDB:5V8F)[162]. The identity and coloring of each subunit is clarified in the right-hand panel. Protein is represented in the cartoon format, where alpha helices are represented by cylinders and the surface is shown in transparent grey.

1.7.6.2 Formation of the MCM2-7 Double hexamer

The structure of the MCM2-7 double hexamer has been resolved both with and without dsDNA present in the central channel [139,166]. The double hexamer encircles 62 bases of dsDNA, which is bent by the interaction (Figure 1.12) [166]. Like the precise structure of the MCM2-7 hexamer, the double hexamer structure is ordered and asymmetric. MCM2, 5 and 3 contact subunits MCM6,7 and 7 on the opposing heterohexamer through interdigitation of the ZnFs and NTEs [166]. The process that occurs between double hexamer formation and initiation of DNA synthesis is not well understood. It likely involves phosphorylation of specific sites on MCM2-7 to initiate melting of dsDNA and facilitate recruitment of Cdc45 and GINS, in order to form an active CMG unwinding complex [166,167]. The lagging strand of DNA also must leave the MCM hexamer, possibly via re-opening of the 2-5 gate.

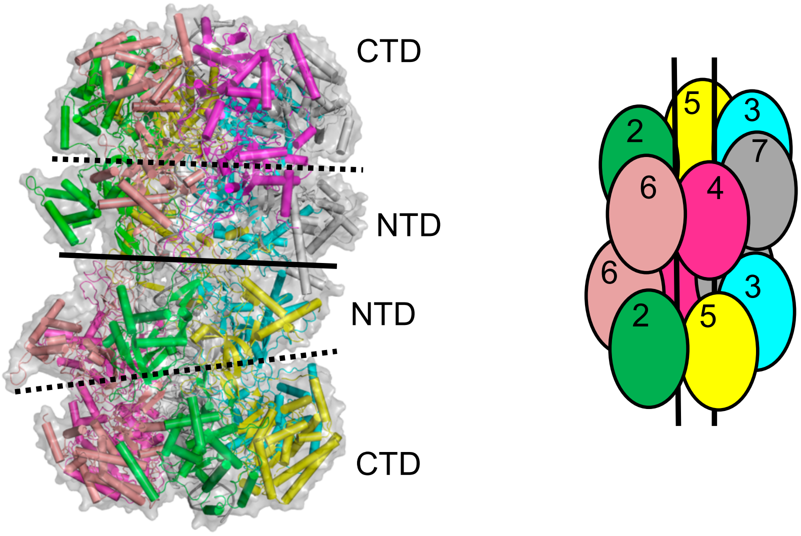


Figure 1.12: Structure of an MCM2-7 double hexamer on dsDNA.
The structure of MCM loaded onto Origin DNA as the pre-replication complex (PDB: 6F0L)[166]. Protein is visualized in the cartoon format, where alpha helices are represented as cylinders and the surface is shown in transparent grey. The identity and coloring of each subunit is clarified in the right-hand panel. DNA bound MCM are interacting in a head-to-head interaction, hence one MCM2-7 is rotated a full 180°.

1.7.6.3 Mechanism of DNA unwinding by MCM2-7

Numerous studies have resolved structures of the CMG complex bound to forked and ssDNA [168–171]. The Cdc45 and GINS complex position on the N-terminal domains of subunits 2,5 and 3 and stabilize the hexamer structure (Figure 1.13a) [172–174]. Recently, solution of CMG complexes bound to forked DNA has permitted accurate localization of the origin of dsDNA strand separation [170,171]. Parental dsDNA is observed to enter a short distance within the ZnF ring of MCM2-7 (Figure 1.13b). This suggests a slightly alternative model of DNA unwinding, whereby the mechanism of strand separation is a partial steric exclusion model. This is in agreement with biochemical studies that propose partial internal unwinding, followed by lagging strand exit [175].

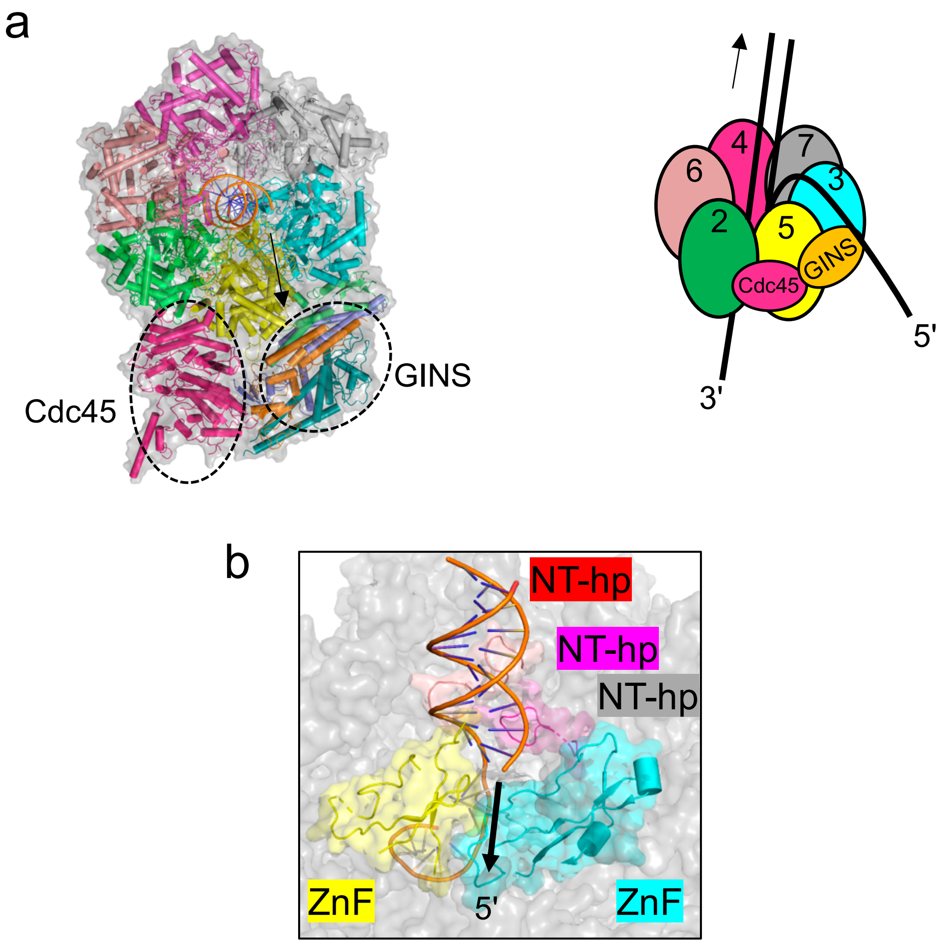


Figure 1.13: Structure of CMG complex bound to forked DNA.
(a) MCM2-7 complex was stalled on forked DNA with Cdc45 and GINS complex (PDB:6U0M)[170]. Protein is represented in the cartoon format, where alpha helices are shown as cylinders and the surface is shown in transparent grey. The lagging strand is directed towards a gap between subunits 3 and 5. The identity and coloring of each subunit is clarified in the righthand panel. The arrow represents the direction of MCM translocation. **(b)** Focused view of the site of strand separation and subsequent exclusion. The arrow represents the direction of the 5' lagging strand through the ZnF tier.

Hairpins of the OB-fold ring form a blockade that prevents progression of the lagging strand through the channel. The hairpins involved in this barrier formation are from the neighbouring subunits, 4,6 and 7 [170,171]. Interestingly, this fraction of subunits was previously identified to form an active complex *in vitro* [55]. A fourth hairpin supplied from MCM3 partially contacts the lagging strand. These hairpins direct the lagging strand of DNA towards a wide gap between the base of the ZnFs of subunits 3 and 5 [170,171]. Density corresponding to the lagging strand is subsequently lost before ssDNA exits the channel. One model suggests that the mechanism of strand separation is purely physical [170]. The authors are supported by the observation that there is no obvious conserved DNA separation site shared between the amino acid sequence of MCMs [170]. Like the h2i, a purely structural element does not necessarily require a precise amino acid sequence [90]. An alternative mechanism implies the role of a phenylalanine on MCM7 NT-hp to function as the separation pin for DNA unwinding [171].

The leading strand continues passage through the hexamer into the AAA+ hexamer ring and interacts with the h2i and PS1β in similar fashion observed for the DNA-bound archaeal MCM structure [100]. Although in the archaeal structure the ssDNA is more compact where all 6 subunits contact DNA within the AAA+ central channel. In the fork bound eukaryotic structure, only 4 subunits in the AAA+ channel contact the ssDNA [168,170]. This may be due to the difference in ATPase active site occupancy. In the MCM2-7 fork bound structure, nucleotide occupancy is only observed in 3 sites, whilst in the archaeal structure all sites are occupied. ATP mediated staircase-style translocation along the single stranded leading strand would force the fork against the OB-fold blockade and/or separation pin [168].

1.7.7 Importance of asymmetry and evolution of MCM function

MCM proteins in archaea and eukaryotes are essential components of the cell cycle regulating DNA synthesis. The archaeal MCM homohexamer represents a simplified version of the eukaryotic MCM2-7 heterohexamer. The recent discovery of distant archaeal lineages has shed light on the evolution of life and eukaryotes. This suggests that archaeal MCM is not just related to eukaryotic MCM2-7, it is the direct ancestor. Importantly, no eukaryote has been identified with fewer than 6 MCM genes, suggesting the evolution of heterohexamer occurred early in the transition from archaea to eukaryotes. Shedding light on this evolution, some archaeal species have been identified with multiple MCM copies [176] (fewer than 6), which are involved in important regulatory processes such as the DNA-damage response [177].

We have seen how the unique structure of a eukaryotic heterohexamer supplies a precise order within the heterohexamer via an evolved interface. This order provides distinctive roles of subunits and interfaces in terms of both the mechanism and regulation of MCM2-7. It has become somewhat puzzling how this complexity relates to the mechanistic properties of a simpler homohexamer.

All MCMs adopt a variety of conformations that relate directly to the function of the enzyme (summarised in Figure 1.14). In MCM2-7 a central regulator is the MCM2-5 gate which is a structural feature for controlled loading onto DNA. For archaeal MCM which is derived from 6 identical subunits, it remains relatively unexplored how an equivalent opening can form in an apparently symmetrical system. Furthermore, open complex stabilizing homologs such as Cdt1 are usually absent in archaea [178]. This suggests that homohexameric MCM can stabilize an open ring independently or is loaded by an alternative mechanism. Negative stain EM studies suggest that an open complex of *Sso*MCM is preferentially loaded onto DNA by ORC [105]. Alternatively, archaeal MCM may adopt an approach akin to the bacterial DnaB homohexamer which is symmetrical and closed until the ring is disrupted by a second component, DnaC [179]. Indeed, the archaeal Cdc6 homologue was previously shown to induce dissociation of closed *Mth*MCM hexamers [180].

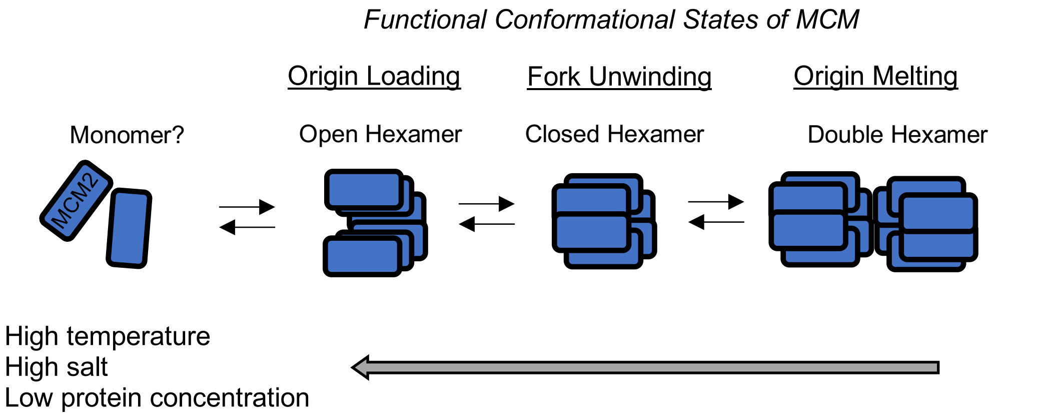


Figure 1.14: Summary of functionally important MCM conformations.
In solution MCM exist in an equilibrium of conformations that reflect important functional states of the enzyme. The equilibrium position may be manipulated by adjusting the conditions which the MCM is subjected to.

Complicating the issue of MCM conformation is the apparent sensitivity of MCM oligomers on experimental conditions. For both archaeal and eukaryotic MCM, conformational stability is influenced by salt, protein concentration and temperature. Of particular importance here is temperature. Virtually all studied archaeal MCM are from thermophilic organisms, and the conformational stability of the enzymes reflect this equilibrium: open rings are observed near the optimal growth temperature of the organism (alike eukaryotic MCM2-7); closed rings are observed towards room temperature [103,105]. Thermophilic Crenarchaota and Euryarchaeota represent only a small portion of archaea, whilst the molecular properties of enzymes in recently discovered archaeal lineages are unexplored. Further, many biophysical techniques are not tractable under the high temperatures required to generate open MCM hexamers. In line with other enzyme studies, we expect MCM from thermophiles (>50 °C) to be inactive under the room temperature conditions required for many biophysical analyses [181].

1.8 Aims

Detailed aims will be provided at the start of each chapter. Broadly, to make better comparisons between eukaryotic and archaeal MCM models, this study aims to:

- Identify and characterize novel archaeal MCMs that are active under conditions used in eukaryotic studies at room temperature (Chapter 3);

- Investigate the structural relationship between identified, active MCMs and pre-existing MCM structures (Chapter 4).

The collective understanding of MCM structure and function also supports a biotechnological application. Helicase enzymes are an essential component of the Oxford Nanopore Technologies (ONT) DNA sequencing platform (detailed in Chapter 5). Here this work aims to:

- Investigate the feasibility of using archaeal MCM as a helicase enzyme in ONT sequencing (Chapter 5).

Chapter 2 – Materials and Methods

2.1 Materials

2.1.1 *Escherichia coli* (*E. coli*) strains

Two *E. coli* strains were used in this study. XL10-Gold cells were used for plasmid production, BL21(DE3)pLysS were used for overproduction of recombinant MCMs. Both strains possess a chloramphenicol resistance gene (*Cmp*^{R}). All strains were obtained commercially from Agilent Technologies.

Table 2.1: *E. coli* strains.

Strain	Genotype
BL21(DE3)pLysS	B F^{-} *ompT gal dcm lon hsdS*$_{B}$(*r*$_{B}^{-}$*m*$_{B}^{-}$) λ(DE3 [*lacI lacUV5-T7p07 ind1 sam7 nin5*]) [*malB*$^{+}$]$_{\text{K-12}}$(λ^{S}) pLysS[*T7p20 ori*$_{\text{p15A}}$](**Cmp^{R}**)
XL10-Gold	endA1 glnV44 recA1 thi-1 gyrA96 relA1 lac Hte Δ(mcrA)183 Δ(mcrCB-hsdSMR-mrr)173 tet^{R} F'[proAB lacI^{q}ZΔM15 Tn10(Tet^{R} Amy **Cmp^{R}**)]

The BL21 strain is deficient in both the *lon* and *omp-T* proteases, which can degrade overexpressed recombinant proteins[182]. The DE3 lysogen possesses a T7 RNA polymerase, which is under the control of a *lac*UV5 promoter. When an analogue of lactose, such as Isopropyl β-d-1-thiogalactopyranoside (IPTG) is supplied, T7 RNA polymerase is produced and induces expression of recombinant genes under the control of a T7 promoter. The pLysS plasmid produces the T7 lysozyme to reduce the basal levels of gene expression. In turn, this improves the growth of *E. coli* cells with particularly toxic proteins.

2.1.2 Bacterial expression vectors

Codon optimised gene sequences were synthesized and cloned into Oxford Nanopore Technologies' proprietary expression vector (pONT) by Genscript. Genes were positioned downstream of a T7 promoter, which controls the expression of an open reading frame (ORF). The ORF begins with a His$_{10}$-tag, followed by a short linker (SGGSGG) and a Tobacco Etch Virus (TEV) protease cleavage site (ENLYFQ |G). As the cutting mechanism targets between the glutamine and glycine residues, a non-native glycine is left at the N-terminus of protein constructs. The ORF is terminated by a double stop codon. The pONT plasmid encodes for an ampicillin resistance gene (Amp^{R}). All sequences can be found in Supplementary Information 7.1.

2.1.3 Bacterial culture media

All media used in this study are outlined in Table 2.2.

Table 2.2: *E. coli* culture media.

Media	Use	Composition	Antibiotics
Lysogeny broth (LB)	Molecular cloning, Protein expression	1 % (*w/v*) Tryptone, 0.5 % (*w/v*) Yeast extract, 171 mM NaCl 1 % (*w/v*) Glucose	100 µg/mL Ampicillin 34 µg/mL Chloramphenicol
LB-agar	Molecular cloning (transformation)	1 % (*w/v*) Tryptone, 0.5 % (*w/v*) Yeast extract, 171 mM NaCl 1.5 % (*w/v*) Agar 1 % (*w/v*) Glucose	100 µg/mL Ampicillin 34 µg/mL Chloramphenicol
Super optimal broth with Catabolite repression (SOC)	Molecular cloning (transformation)	2% (*w/v*) Tryptone, 0.5% (*w/v*) Yeast extract, 10 mM NaCl 2.5 mM KCl 10 mM $MgCl_2$ 10 mM $MgSO_4$ 20 mM Glucose	--*none*--

2.1.4 Buffer solutions

Buffers and solutions used in this study are listed in Table 2.3. Generally, all solutions were passed through a 0.2 µm filter and degassed. The stated pH is determined at room temperature.

Table 2.3: Buffer composition.

Buffer Identity	Use	Composition
Buffer EB	Oligo resuspension	10 mM Tris-Cl pH 8.5
Buffer A	Protein purification	500 mM NaCl, 20 mM Tris-Cl pH 8.0, 20 mM Imidazole, 5 % (*v/v*) Glycerol
High salt wash	Protein purification	2 M NaCl, 20 mM Tris-Cl pH 8.0, 20 mM Imidazole, 5 % (*v/v*) Glycerol
Buffer B	Protein purification, Affinity capture assay	500 mM NaCl, 20 mM Tris-Cl pH 8.0, 500 mM Imidazole, 5 % (*v/v*) Glycerol
TEV-buffer	Protein purification	500 mM NaCl, 20 mM Tris-Cl pH 8.0, 5 % (*v/v*) Glycerol, 0.5 mM DTT
AEX-low	Protein purification	50 mM NaCl, 20 mM Tris-Cl pH 8.0, 5 % (*v/v*) Glycerol, 0.5 mM TCEP
AEX-high	Protein purification	1 M NaCl, 20 mM Tris-Cl pH 8.0, 5 % (*v/v*) Glycerol, 0.5 mM TCEP

SEC buffer	Protein purification	500 mM NaCl, 20 mM Tris-Cl pH 8.0, 5 % (*v/v*) Glycerol, 0.5 mM TCEP
Annealing buffer	Molecular biology	50 mM KCl, 20 mM Tris-Cl pH 8.0
TB buffer (1 x)	EMSA	90 mM Tris, 90 mM Borate, pH 8.3
EMSA loading dye (1 x)	EMSA	1 x TB, 12.5 % (*v/v*) glycerol
XTAL-buffer	Crystallography	100 mM NaCl, 20 mM Tris-Cl pH 8.0, 5 % (*v/v*) Glycerol, 0.5 mM TCEP
Analytical SEC buffer	Molecular biology	200 mM NaCl, 20 mM Tris-Cl pH 8.0, 5 % (*v/v*) Glycerol
Affinity capture buffer	Molecular biology	100 mM NaCl, 20 mM Tris-Cl pH 8.0, 5 % (*v/v*) Glycerol, 1 mM ATP, 10 mM $MgCl_2$
SEC-MALLS buffer	SEC-MALLS, AUC	200 mM KCl, 50 mM Tris-Cl pH 8.0, 5 % (*v/v*) Glycerol, 0.5 mM DTT
SDS-PAGE loading dye (1 x)	Molecular biology	100 mM DTT, 50 mM Tris, 20% (*w/v*) glycerol, 2% (*w/v*) SDS, 0.1% (*w/v*) bromophenol blue
TGS (1 x)	Molecular biology	3 % (*w/v*) Tris, 14 % (*w/v*) glycine, 1 % (*w/v*) SDS
NR-CBBR	Molecular biology	0.005% (*v/v*) Coomassie brilliant blue R 0.0005% (*v/v*) Neutral Red 40% (*v/v*) methanol 7% (*v/v*) acetic acid
TBS-T	Molecular biology	150 mM NaCl, 20 mM Tris-Cl pH 7.5, 0.1 % (*v/v*) Tween 20

2.1.5 Oligonucleotides and preparation

All DNA substrates were prepared as 100 µM stocks in buffer EB and stored at - 20 °C.

Table 2.4: Oligonucleotide sequences.
Chemical modifications are highlighted in **bold.**

Substrate	Sequence (5'-3')	Use
Hel3	TTTGTTTGTTTGTTTGTTTGTTTGTTTGTTTGCCGACGTGCCAGGCCGACGCGTCCC	ATPase assay
Hel5	GGGACGCGTCGGCCTGGCACGTCGGCCGCTGCGGCCAGGCACCCGATGGCGTTTGTTTGTTTGTTTGTTTGTTT	ATPase assay/ Binding assays
Hel5-Cy3	**[Cy3]**GGGACGCGTCGGCCTGGCACGTCGGCCGCTGCGGCCAGGCACCCGATGGCGTTTGTTTGTTTGTTTGTTTGTTT	Short helicase assay
Hel5-FAM	**[FAM]**GGGACGCGTCGGCCTGGCACGTCGGCCGCTGCGGCCAGGCACCCGATGGCGTTTGTTTGTTTGTTTGTTTGTTT	Binding assays
Hel3-BHQ2	TTTGTTTGTTTGTTTGTTTGTTTGTTTGTTTGCCGACGTGCCAGGCCGACGCGTCCC**[BHQ2]**	Short helicase assay
SCAV5	GGGACGCGTCGGCCTGGC	Short helicase assay
ss34-FAM	CCCTGCGCAGCCGGACCGTGCAGCCGTTTGTTTG**[FAM]**	Binding assays
Hel5-F10	GGGACGCGTCGGCCTGGCACGTCGGCCGCTGCGGCC	Binding assays
Hel5-F20	GGGACGCGTCGGCCTGGCACGTCGGCCGCTGCGGCCAGGCACCCGA	Binding assays

Hel5-F40	GGGACGCGTCGGCCTGGCACGTCGGCCGCTGCGGCCAGGCACCCGATGGCGTTTGTTTGTTTGTTT	Binding assays
Hel5-F0	CAAACAAACGGCTGCACGGTCCGGCTGCGCAGGG	Binding assays
PolyT$_{50}$-FAM	**[FAM]**TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	Binding assays/SEC
(ACTG)$_{16}$	ACTGACTGACTGACTG	Binding assays
(ACTG)$_{32}$	ACTGACTGACTGACTGACTGACTGACTGACTG	Binding assays
(ACTG)$_{48}$	ACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTG	Binding assays
(ACTG)$_{64}$	ACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTG	Binding assays
(ACTG)$_{80}$	ACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTG	Binding assays
(ACTG)$_{96}$	ACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTG	Binding assays
Hel80-Cy3	GTTGTTGTTGGCCATACAATGTGCCCACTTGTATGATTGTGCGAAGGAAGGGAGTGTGCATCGATTCGTTCAGTATTGGC**[Cy3]**	Long helicase assay
Hel15-BHQ2	**[BHQ2]**CGTCGCCCATCCAGG	Long helicase assay
Hel15-SCAV	CCTGGATGGGCGACG	Long helicase assay
Hel101	CCTGGATGGGCGACGAGCCAATACTGAACGAATCGATGCACACTCCCTTCCTTCGCACAATCATACAAGTGGGCACATTGTATGGCTGTTGTTGTTGTTGT	Long helicase assay
Hel106	CCTGGATGGGCGACGAGCCAATACTGAACGAATCGATGCACACTCCCTTCCTTCGCACAATCATACAAGTGGGCACATTGTATGGCCAACAACAAC	Long helicase assay
Hel116	CCTGGATGGGCGACGAGCCAATACTGAACGAATCGATGCACACTCCCTTCCTTCGCACAATCATACAAGTGGGCACATTGTATGGCTGTTGTTGTTGTTGTTGTTGTTGTTGT	Long helicase assay
Hel101-1xSp18	CCTGGATGGGCGACGAGCCAATACTGAACGAATCGATGCACACTCCCTTC**/iSp18/**CGCACAATCATACAAGTGGGCACATTGTATGGCTGTTGTTGTTGTTGT	Long helicase assay
Hel101-2xSp18	CCTGGATGGGCGACGAGCCAATACTGAACGAATCGATGCACACTCCC**/iSp18//iSp18/**CGCACAATCATACAAGTGGGCACATTGTATGGCTGTTGTTGTTGTTGT	Long helicase assay
Hel101-4xSp18	CCTGGATGGGCGACGAGCCAATACTGAACGAATCGATGCACACT**/iSp18//iSp18//iSp18//iSp18/**ACAATCATACAAGTGGGCACATTGTATGGCTGTTGTTGTTGTTGT	Long helicase assay
Hel101-6xSp18	CCTGGATGGGCGACGAGCCAATACTGAACGAATCGATGCAC**/iSp18//iSp18//iSp18//iSp18//iSp18//iSp18/**ATCATACAAGTGGGCACATTGTATGGCTGTTGTTGTTGTTGT	Long helicase assay
MCM-ONT-top	**/5phos/**GCCAATACTGAACGAATCGATGCACACTCCC**/iSp18//iSp18/**CGCACAATCATACAAGTGGGCACATTGTATGGCTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT**/3sp3/**	Nanopore experiment
MCM-ONT-bottom	GAGGCGAGCGGTCAATTTGCCATACAATGTGCCCACTTGTATGATTGTGCGAAGGAAGGGAGTGTGCATCGATTCGTTCAGTATTGGCT	Nanopore experiment

2.1.6 Annealed substrates and preparation

Mixtures were combined at an equimolar ratio to a final concentration of 1 µM annealed substrate in Annealing buffer. Samples were placed in a heat block set at 85 °C for 5 minutes. The heat block was then removed, and samples are allowed to cool slowly to room temperature (~2 hours). Samples were briefly centrifuged after cooling to collect residual condensation. The default substrate has a 25 base-pair duplex region, a 49 base 3' overhang and a 31 base 5' overhang (see Figure 2.1).

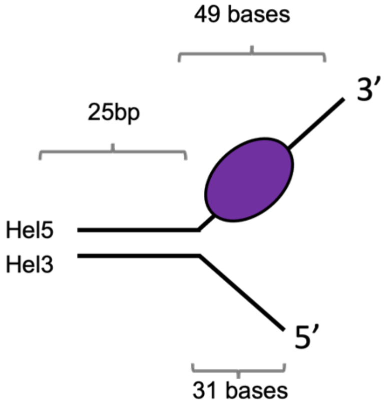


Figure 2.1: Standard fork substrate.
Schematic of the standard fork substrate used in this thesis. MCM (purple) are assumed to load favourably onto the 49 base 3' arm.

Table 2.5: Duplexed DNA substrates.
The default substrates used in chapters 3 and 4 are highlighted in grey.

Name	Combination	Use
Fork	Hel3 Hel5	ATPase
FAM-fork	Hel3 Hel5-6FAM	Binding assays
AS	Hel5-Cy3 Hel3-BHQ2	Helicase
MF	SCAV5 Hel3-BHQ2 ... *Add* Hel5-Cy3 *after cooling*	Helicase
AS-long	Hel101 Hel80-Cy3 Hel15-BHQ2	Long helicase
MF-long	Hel101 Hel80-Cy3	Long helicase
AS-long-F30	Hel116 Hel80-Cy3 Hel15-BHQ2	Long helicase
AS-long-blunt	Hel106 Hel80-Cy3	Long helicase

	Hel15-BHQ2	
1xiSp18	Hel101-1xSp18 Hel80-Cy3 Hel15-BHQ2	Long helicase
2xiSp18	Hel101-2xSp18 Hel80-Cy3 Hel15-BHQ2	Long helicase
4xiSp18	Hel101-4xSp18 Hel80-Cy3 Hel15-BHQ2	Long helicase
6xiSp18	Hel101-6xSp18 Hel80-Cy3 Hel15-BHQ2	Long helicase
F10	Hel5-F10 ss34-FAM	Binding Assays
F20	Hel5-F20 ss34-FAM	Binding Assays
F40	Hel5-F40 ss34-FAM	Binding Assays
dsDNA-36	Hel5-F0 ss34-FAM	Binding Assays
MCM-adapter	MCM-ONT-top MCM-ONT-bottom	ONT

2.1.6 SDS-PAGE gels

1 mm thick SDS-PAGE were prepared at the desired concentration of resolving gel. All SDS-PAGE use a 4 % stacking region. Typically, a 15-well comb was used.

Table 2.6: Composition of SDS-PAGE gels

Component	**4 % stacking**	**12 % resolving**	**18% resolving**
30 % (*w/v*) acrylamide, 0.8 % (*w/v*) bis-acrylamide	4 %	12 %	18 %
1.5 M Tris-Cl pH 8.8	-	0.37 M	0.37 M
0.5 M Tris-Cl pH 6.8	0.13 M	-	-
10 % (*w/v*) Sodium dodecyl sulfate (SDS)	0.1 %	0.1 %	0.1 %
10 % (*w/v*) Ammonium persulfate (APS)	0.1 %	0.1 %	0.1 %
N,N,N’,N’-tetramethylethylenediamene (TEMED)	0.1 %	0.1 %	0.1 %

2.1.7 Native-PAGE gels

1 mm thick native 1 x TB PAGE was prepared with no stacking region.

Table 2.7: Composition of a 5 % native PAGE gel

Component	Final Concentration
TB (see table 2.3)	1 x
30 % (w/v) acrylamide, 0.8 % (w/v) bis-acrylamide	5 %
10 % APS	0.1 %
TEMED	0.1 %

2.1.8 Nucleotide preparation

ATP (Bio Basic) was prepared as a 50 mL stock at a concentration of 100 mM. The pH was adjusted to 7 by addition of 4 M NaOH then stored in 1 mL aliquots at -20 °C until further use. Non-hydrolysable analogues ADP and AMP-PCP were prepared as 100 mM stocks, where lyophilised nucleotides are diluted in 100 mM Tris-Cl pH 7.0 immediately before use. ADP•AlF_4^- is prepared at the stated concentration in the reaction buffer with a ratio of 1:1:5 (ADP: AlCl_3: NaF) and 1 % (*v/v*) DMSO.

2.2 Protein production methods

2.2.1 Site directed mutagenesis

Primers for site directed mutagenesis were designed using the QuickChange Primer Design tool (Agilent) using to the 'QuickChange Lightning' option. Primers can be found in Supplementary Table 7.1. Following the manufacturers' guidelines, QuickChange Lightning mutagenesis (Agilent Technologies) was performed on (100 ng) plasmid DNA.

2.2.2 Extraction of plasmid DNA and sequencing of mutants

Individual colonies were picked from selective LB agar plates and used to inoculate a sterile 50 mL tube containing 5 mL selective LB-media. Samples were left for 16 hours at 37 °C on a shaker set to 225 rpm. Cultures were then centrifuged at 4,000 x *g,* 4 °C and the supernatant is discarded. Plasmid DNA was extracted from the cell pellet using a QIAprep Spin Miniprep kit (Qiagen) according to the manufacturer's guidelines. An aliquot of 15 µL plasmid DNA at 50 -100 ng/ µL was then sent for Sanger sequencing (Eurofins genomics). Where possible, sequencing was performed using either T7 promoter or terminator sequencing primers. Due to the large size of the gene inserts (~2,100 bp), mutations in the middle genes were validated using custom sequencing primers (Eurofins genomics). Sequenced plasmids were then compared against the wild-type sequence in 'SnapGene' software (Insightful Science). Validation of sequencing quality was assessed in 'A plasmid Editor' software (RRID:SCR_014266).

2.2.3 Transformation into *E. coli*

Vectors were transformed into the required strain of *E. coli* (see table 2.1). All chemically competent cells are transformed using the heat shock method according to the manufacturer's guidelines (Agilent Technologies). Following the heat shock protocol, the cell mixture was spread onto selective LB-agar plates and placed for 20 hours at 37 °C. Plates were then stored at 4 °C for up to a week.

2.2.4 Preparation of *E. coli* glycerol stocks

A single colony was taken from BL21(DE3)pLysS transformation plates and used to inoculate 50 mL LB. Cells were then placed for 16 hours at 37 °C with shaking at 120 rpm. Optical density was measured (expected OD_{600} = 2-5) and 500 µL bacterial culture is added to 500 µL sterile 50 % (*v/v*) glycerol in a 1.5 mL Eppendorf tube. The tube is then plunged into liquid nitrogen and stored at - 70 °C until further use.

2.2.5 Overproduction of recombinant MCM for 'crude' protein preparations

Protocol as 2.2.6, however, the volumes for starter and expression cultures were adjusted to 10 and 200 mL respectively.

2.2.6 Overproduction of recombinant MCM for 'pure' protein preparations

A pre-culture of 50 mL of selective LB-media was inoculated with *E. coli* glycerol stock and grown for 16 hours at 37 °C with shaking at 120 rpm. The optical density (OD_{600}) of the overnight culture was determined and the volume of culture required to inoculate 1 L of selective LB-media at an OD_{600} of 0.05 was transferred. The culture was then grown until the OD_{600} reached a value between 0.6 and 0.8. To assess pre-induction protein expression, a volume of culture to resuspend in 0.5 mL lysis buffer was removed according to the equation:

Equation 2.1: OD standardization of cell expression gels.

$$\mathrm{V_{culture}} = \frac{20 \times \mathrm{V_{lysis}}}{\mathrm{OD_{600}}}$$

Where *V* is the volume of culture and lysis buffer in *ml*, and OD_{600} is the optical density of the cells. The sample was then centrifuged at 5,000 x *g* for 10 minutes and the supernatant was removed. The pellet was then placed at -20 °C until further use. Protein expression was then induced in the 1 L culture by addition of IPTG to a final concentration of 1 mM. Cells were then grown in an incubator at 20 °C for 20 hours, with shaking at 170 rpm. The final OD_{600} is then measured (generally between 2-5) and an aliquot was removed according to equation 2.1 to assess the relative level of protein expression. Remaining cells were then isolated by centrifugation at 4,000 x *g.* Cell pellets were then transferred into a 50 mL tube and stored at -20 °C until further use.

2.2.7 Assessment of protein expression and solubility

Expression pellets (section 2.2.6) were thawed and resuspended in 0.5 mL buffer A. Cells were then lysed using the sonication method (30s: 3s on, 7s off). To assess protein expression in the soluble fraction, 50 µL of the lysed post-expression sample was removed and centrifuged at 13,900 x *g* for 10 minutes. The supernatant was then extracted to assess the level of recombinant protein in the soluble fraction. Loading dye was then added to 50 µL of each sample. 6 µL of each sample was then run on a 12 % SDS-PAGE. Gels were fixed using a solution of 10 % acetic acid, 40 % ethanol for 1 hour, then thoroughly washed using deionised water. To determine the presence of His_{10}-bands, gels were then stained with a solution of TBS-T, Ni-NTA-$ATTO_{488}$ (1:1000 dilution of stock, Merck) for 1 hour and imaged on a Typhoon gel scanner (GE healthcare). After imaging, gels were stained with NR-CBBR dye to detect protein bands.

2.2.8 Purification of 'crude' protein

E. coli cell pellets produced from 200 mL of expression culture were thawed and re-suspended in buffer A to an OD_{600} of 100. Buffer A was supplemented with 20 µg/µL DNase, 20 µg/µL RNase and cOmplete protease inhibitor tablets (Roche). Cells were lysed using the sonication method at (3 s on 7 s off for 1 min/100 mL culture). Cell extract was then centrifuged for 45 minutes at 30,000 x *g*, 4 °C. The resulting supernatant was then loaded onto a 1 mL HisTrap FF column pre-equilibrated in buffer A. The column was then washed with buffer A and 2 column volumes (CV) high salt wash, before being re-equilibrated into buffer A. Bound proteins were then eluted from the column using an isocratic elution of buffer B. Fractions were then analysed by SDS-PAGE, pooled, and dialysed against TEV buffer overnight at 4 °C, using a dialysis membrane (Spectrum Labs) with a 3.5 kDa molecular weight cut-off (MWCO). After dialysis, protein was concentrated in an Amicon Ultra 50,000 molecular weight cut-off (MWCO) spin concentrator (Merck). Protein was typically concentrated to the desired concentration (>10 µM hexameric MCM) and divided into 50-100 µL aliquots. Samples were snap-frozen in liquid nitrogen before placing at -70 °C.

2.2.9 Purification of ‘pure’ protein

E. coli cell pellets produced from 1 L of expression culture, were thawed and re-suspended in buffer A, to an OD_{600} of 100. Buffer A was supplemented with 20 µg/µL DNase, 20 µg/µL RNase, and cOmplete protease inhibitor tablets (Roche). Cells were lysed by sonication at 70 W (3 s on 7 s off for 1 min/100 mL culture). Cell extract was then centrifuged for 45 minutes at 30,000 x *g*, 4 °C. The resulting supernatant was then loaded onto a 5 mL HisTrap HP column pre-equilibrated in buffer A. The column was then washed with buffer A and 2 CV high salt wash, before being re-equilibrated into buffer A. Bound proteins were then eluted from the column with buffer B. Fraction selection was based on SDS-PAGE analysis. Selected fractions were collected and then TEV protease (a generous gift from Mike Hodgkinson) was added at a ratio of 1 mg TEV: 50 mg His_{10}-tagged protein. Sample was then dialysed against TEV buffer overnight at 4 °C using a 3.5 kDa MWCO dialysis membrane (Spectrum Labs). Tag cleavage was confirmed by SDS-PAGE analysis. Dialysate was then loaded onto a 5 mL HisTrap FF column and the untagged recombinant MCM was collected from the flow through. Where an extra polishing step was required, samples were subjected to anion exchange chromatography (AEX). Protein and buffer were diluted to final concentration of 50 mM NaCl, 20 mM Tris-Cl pH 8.0, 5 % (v/v) Glycerol, 0.5 mM TCEP and loaded onto a 50 mL Source 15Q column (Merck), pre-equilibrated in AEX-low buffer. Protein was then eluted by applying a 10 CV gradient elution of AEX-high buffer. Fractions were analysed by SDS-PAGE, then concentrated in an Amicon Ultra 50,000 MWCO spin concentrator (Merck) to 10-20 mg/mL. Samples were then loaded onto either a HiLoad 16/60 S-300 or a HiPrep 26/60 S-200 Size Exclusion Column (GE Healthcare), equilibrated in SEC buffer. Fractions were collected, and spin concentrated as before to a final concentration ~7-20 mg/mL. Samples were then snap frozen in 50-100 µL aliquots and stored at -70 °C.

2.2.10 Purification of ‘pure’ protein for X-ray crystallography

Protocol as 2.2.9. however, there is an emphasis on achieving the highest purity possible. This meant being extremely stringent with fraction selection over yield. Protein was always purified in the minimum time possible, usually <30 hours. Protein was never frozen before setting up crystallization screens. After the final SEC, protein was immediately dialysed into the XTAL-buffer overnight at 20 °C.

2.2.11 Preparation of fluorescently labelled MCM

*Mac*MCM$^{FL\text{-}GSC}$ was extensively dialysed overnight at room temperature into SEC buffer, where DTT was substituted for 1 mM TCEP. Subsequently, in a 600 µL reaction volume, a 2-fold excess of fluorescein maleimide (6.2 µM; Vector Labs; stock prepared at 75 µM in 10 % DMSO) to MCM monomer was added in three steps separated by a 10-minute gap at 37 °C. The reaction was quenched and DMSO was removed by dialysis against 2 L SEC buffer (with 1 mM DTT) in a 3.5 kDa MWCO D-Tube Midi (Millipore) for 4 hours at room temperature. Labelled protein was separated from unreacted fluorophore by SEC (see section 2.6.2) using a Superose 6 100/300 GL column equilibrated in 100 mM NaCl, 20 mM Tris-Cl pH 8.0, 5 % (*v/v*) glycerol, 1 mM DTT. 400 µL fractions were collected and analysed via SDS-PAGE. Selected fractions are pooled, then snap frozen in liquid nitrogen before storage at -70 °C.

2.2.12 Preparation of cross-linked MCM

Site specific, inter subunit disulphide cross-links were formed by allowing the reaction of proximally engineered cysteine residues at hexamerization interfaces. Storage buffer reducing agent was removed by spin concentrator dialysis in an Amicon Ultra MWCO 50,000 spin concentrator (Merck). Complex formation was initiated by diluting MacMCM$^{\Delta WH\text{-}DC}$ into a solution of 15 mL 100 mM NaCl, 20 mM Tris-Cl pH 8.0, 5 % (*v/v*) Glycerol, 1/10 mM ATP/$MgCl_2$ and an equimolar ratio of $(ACTG)_{16}$. Protein was then concentrated to ~500 µL, after which the dilution/concentration procedure was repeated a further 3 times. After the final concentration step, the sample was left overnight to incubate at 4 °C. Cross-linked products were then separated by SEC on a Superose 6 100/300 GL equilibrated in 100 mM NaCl, 20 mM Tris-Cl pH 8.0, 5 % glycerol. Fractions of cross-linked products were assessed using SDS-PAGE, on a 12 % polyacrylamide gel. Samples were run twice, with or without DTT in the loading dye.

2.3 Electrophoresis techniques

2.3.1 SDS-PAGE

Samples were prepared by addition of loading dye to a final concentration 1 x and denatured at 95 °C for at least 5 minutes. Samples were briefly centrifuged (13,900 x *g*, 1 minute) and loaded into wells on the stated percentage SDS-PAGE. Usually, 6 µL Precision Plus Protein™ All Blue Pre-stained ladder (Bio-Rad) was used to estimate the MW of resolved bands. Electrophoresis was performed in a 1 x TGS running buffer at 200 V for 45 minutes. If testing for the migration of fluorescently labelled moieties, gels were imaged on a Typhoon scanner (GE Healthcare) at an appropriate wavelength. Unless stated otherwise, protein bands were then identified using NR-CBBR. Gels were generally imaged after 24 hrs staining.

2.3.2 Electrophoretic mobility shift assay (EMSA)

2.3.2.1 EMSA Overview

EMSA is a technique used to determine protein-nucleic acid interactions in non-denaturing conditions[183]. A solution of protein and nucleic acids are resolved by electrophoresis on either agarose or polyacrylamide gels (PAGE). Migration of nucleic acids through the gel may then be determined using labelled substrates or post-staining procedures. Free nucleic acids generally migrate faster than when bound to protein. By examining the nucleic acid binding at defined protein concentrations, binding affinity can be examined[184]. The stoichiometry and cooperativity of protein-nucleic acid interactions can also be assessed[184].

2.3.2.2 Native PAGE EMSA

Protein and buffer were prepared in the stated buffer (200 mM KCl, 20 mM Tris-Cl pH 8.0). Samples were then serially diluted in 10 µL volumes, then mixed with 10 µL 500 nM DNA substrate. Samples were incubated at room temperature for 30 minutes, then 20 µL 2 x TB and 25 % Glycerol was added to each sample. 10 µL of each sample was then run on a 5 % Native 1 x TB PAGE (see Table 2.7), 4 °C for the stated length of time at 80 V. Gels were then stained with SYBR-GOLD (Invitrogen) in 1 x TB and imaged on a Typhoon scanner (GE Healthcare) using the SYBR-GOLD protocol with a 100 µm imaging pixel size. After imaging, gels were stained using NR-CBBR to resolve protein bands.

2.3.2.3 Agarose gel EMSA

0.8 % (*w/v*) agarose gels were prepared in 1 x TB-buffer. Protein and buffer were prepared in the stated buffer (default: 250 mM KGlu, 20 mM Tris-Cl pH 8.0). MCM samples were serially diluted in the buffer in 10 µL volumes and added to 10 µL of 20 nM FAM-labelled DNA (see Table 2.5; default substrate: FAM-fork). A DNA-only control is also prepared to determine the motility of free DNA. Samples were then incubated at room temperature for 30 minutes. Before loading, 20 µL 2 x TB and 25 % (v/v) glycerol was added to each sample. 10 µL each sample was then run on the agarose gel for 20 minutes at 150 V. Gels were imaged on a Typhoon scanner (GE Healthcare) using Cy2 filters with a 100 µm imaging pixel size. To estimate the equilibrium dissociation constant (K_d) between protein and DNA, the protein concentration is identified where the DNA motility is distributed equally between free and protein-bound states.

2.4 Absorbance-based methods

2.4.1 Assessment of biomolecule concentration

The concentration of biomolecules in a given solution was assessed by a UV absorbance on a Genesys BioMate 150 spectrophotometer (Thermo Scientific). Typically, a continuous range of absorbances between 220-320 nm were measured. Baseline absorbance was normalized using an appropriate buffer only blank. Using the Beer-Lambert law, the concentration of protein was calculated from the absorbance (A) measured at 280 nm[185]:

Equation 2.2: Beer-Lambert law.

$$c = \frac{A}{\varepsilon l}$$

where *ε* is the extinction coefficient in M^{-1} cm^{-1}, *c* is the concentration in mol dm^{-3}, and *l* is the path length of the measurement in cm. Extinction coefficients were calculated from the amino acid sequence, using the Expasy ProtParam tool (https://web.expasy.org/protparam/). Absorbance at 260 nm and 280 nm was compared to assess DNA contamination within a sample. For example, a 260/280 ratio of 0.6 and below indicates a sample with minimal nucleic acid contamination. As the 260/280 ratio increases the sample is expected to contain more DNA, where a ratio of 1.0, would suggest a DNA mass contamination of 5 %.

2.4.2 Malachite-based ATPase assay

Hydrolysis of ATP was measured using a colorimetric phosphate detection kit. Briefly, MCM ATPase hydrolysis converts ATP into ADP and inorganic phosphate (P_i). P_i forms a complex with malachite green molybdate, generating a colour change from gold to green[186]. The formation of the green complex can be determined using a spectrophotometer measuring absorbance between 600-660 nm. Absorbance values are calibrated against a known concentration range of P_i.

ATPase activity was determined using the high throughput colorimetric ATPase assay kit (Abcam) in the 96-well plate format. Crude MCM was diluted into a 100 µL reaction mix containing a final concentration of 100 nM MCM (based on hexameric MW), 50 mM Tris-Cl and either 0/100 nM forked DNA substrate. After equilibration at the stated temperature for 30 minutes, 100 µL SB mix was added yielding final concentrations 0.05 mM ATP, 2.5 mM $MgCl_2$. After 5 minutes, reactions were stopped and monitored through addition of 50 µL of PiColorLock^{TM} gold mix with 1:100 dilution of 'Accelerator'. After 2 minutes, 20 µL 'Stabilizer' was added. Dye complexes were allowed to develop for 30 minutes at room temperature, after which, absorbance was measured at 611 nm on a Polarstar plate reader (BMG Labtech). A control without helicase was subtracted from absorbance values.

To relate experimental absorbance values to known quantities of P_i, a standard curve was generated. A serial dilution of a P_i stock solution was performed in 200 µL volumes without helicase. Each dilution of P_i was then monitored by addition of 50 µL of PiColorLock^{TM} gold mix with 1:100 dilution of 'Accelerator'. After 2 minutes, 20 µL 'Stabilizer' was added. Dye complexes were allowed to develop for 30 minutes at room temperature, after which, absorbance was measured at 611 nm on a Polarstar plate reader (BMG Labtech). Absorbance values for the known P_i standards were then fitted to a linear regression model in R. The quantity of P_i formed in each experimental well can then be calculated from the linear regression model.

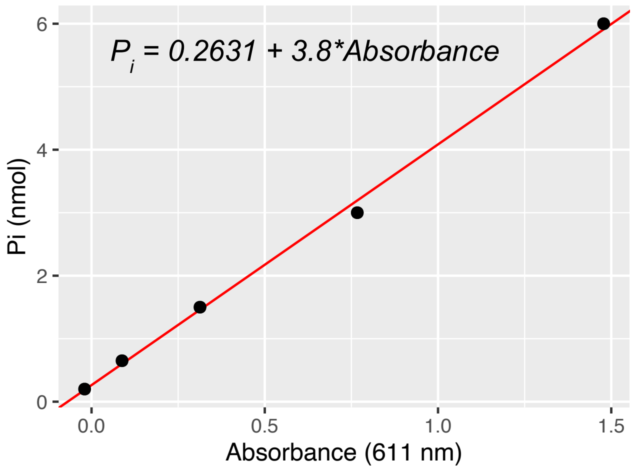


Figure 2.2: ATPase assay calibration.
An exemplary linear regression model relating absorbance readings to known quantities of P_I.

2.5 Fluorescence-based methods

2.5.1 Overview of fluorescence

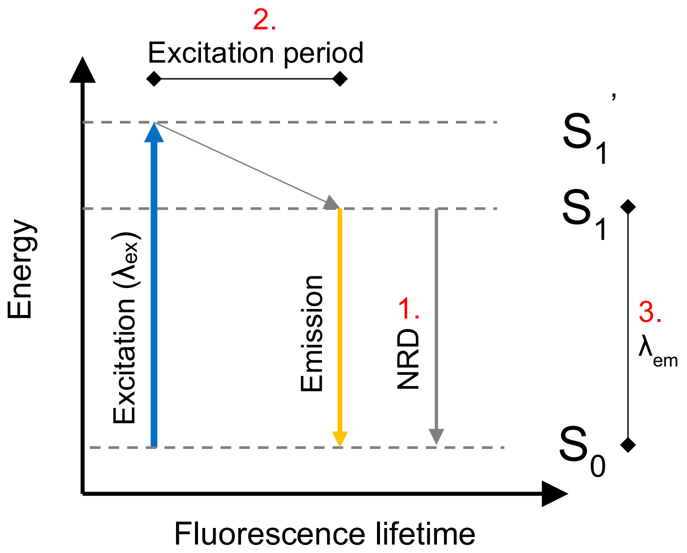


Figure 2.3: Exploitable properties of fluorescence.
λ_{ex}: excitation wavelength. λ_{em}: emission wavelength. S_1 : excited state. S_0 : ground state. NRD: Non-radiative decay.

Many techniques used in modern molecular biology exploit key properties of fluorescence (Figure 2.3)[187]. Photons of energy (generally provided by a laser) are absorbed by fluorescent molecules promoting electrons from the ground (S_0) to an excited state (S_1'). Excited electrons then undergo vibrational relaxation to the lowest excited energy state (S_1). When electrons return to the ground state (S_0), under certain conditions a photon of energy may be released. As the speed of electromagnetic radiation *c*, is constant, the loss of energy (ΔE) during vibrational relaxation ($S_1' \rightarrow S_1$) results in the emission of a photon with a longer wavelength (λ_{em}) than the excitatory photon (λ_{ex}). This is in accordance with the Planck-Einstein relation:

Equation 2.3: Calculation of emission wavelength by the Planck-Einstein relation.

$$\lambda_{em} = \frac{hc}{\Delta E}$$

where *h* is Planck's constant. Many molecules return to the ground state by non-radiative decay, whereby the energy is lost thermally. Non-radiative transition routes between S_1 and S_0 are often referred to as quenching [188]. Polycyclic aromatic compounds have delocalised electrons that are easily excited, whilst the restricted rotation limits the efficiency of vibrational relaxation. These compounds form the basis of many modern fluorophores such as fluorescein.

2.5.2 Fluorescent helicase assay

2.5.2.1 Background

The distance between two light sensitive moieties can be monitored using Förster resonance energy transfer (FRET)[189]. FRET occurs when the fluorescent emission of one donor moiety overlaps with the excitation wavelength of a secondary acceptor moiety. Importantly, FRET efficiency is highly dependent on the two molecules being in close spatial proximity[189]. If the acceptor moiety does not emit fluorescence (e.g., a quencher), then the coupled system will return to the ground state through non-radiative decay. This principle has been adapted to monitor unwinding kinetics of dsDNA by helicase enzymes in real time (Figure 2.4)[190].

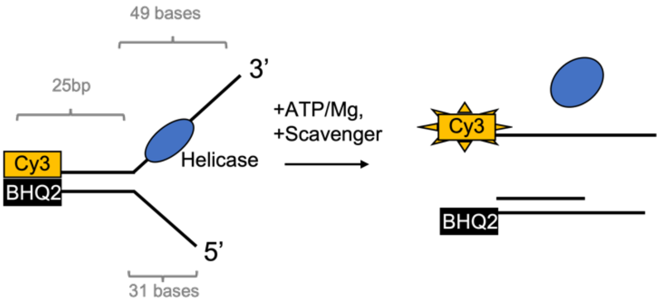


Figure 2.4: Basis of the fluorescent helicase assay.
A helicase enzyme (blue) unwinds a forked DNA substrate on addition of ATP/Mg^{2+}. Unwinding spatially separates a fluorophore (Cy3) and quencher (BHQ2) and allows detection of fluorescence. A scavenger strand (SCAV) prevents reannealing.

Briefly two complementary strands of DNA are chemically labelled: one with a fluorophore (e.g. Cy3); one with a compatible quencher (e.g. BHQ2). When annealed the pair are maintained in close spatial proximity ensuring high FRET efficiency**.** No emission will be observed when excited by the laser. When the helicase is activated, such as by the addition of ATP/Mg^{2+}, dsDNA is unwound, and the FRET pair are spatially separated. A scavenger strand present in excess prevents reannealing and re-localization of the FRET pair. Therefore, unwinding decouples FRET quenching and permits observation of Cy3 emission that is proportional to the amount of dsDNA unwound.

2.5.2.2 Data collection

Outlined are the default reaction parameters for the fluorescent helicase assay, alternative experiments will adjust this protocol accordingly. Reactions are performed in a black, flat bottomed 96-well plate (Thermo Scientific). Initially, a 180 µL mixture was prepared, to which 20 µL of 10 x ATP/Mg^{2+} mix will later be added. The stated concentrations in Table 2.8 represent the final reaction concentration in 200 µL. ATP/Mg^{2+} is added to the control wells in advance to account for slight difference in focusing height when calibrating the plate reader instrument. When different salts and concentrations of salt are used, buffer conditions of the controls are adjusted appropriately to account for salt-dependent effects on the fluorophore. In the 'pre-incubation' experiment (see section 3.8.5), where protein-ATP are pre-equilibrated, 20 µL of either MCM/ATP/Mg^{2+} (10 µM/4 mM/10 mM) was added to a 180 µL reaction containing all the components except the MCM (this 180 µL includes additional ATP/Mg^{2+} at 4/10 mM).

Table 2.8: Standard helicase assay composition for a 200 µL reaction.

		Controls	
Component	**Reaction**	**Maximum**	**No helicase**
MCM	1000 nM (hexamer)	1000 nM (hexamer)	-
KGlu	250 mM	250 mM	250 mM
KPhos pH 8.0	20 mM	20 mM	20 mM
Glycerol	1 % (*v/v*)	1 % (*v/v*)	1 % (*v/v*)
DNA substrate (see table 2.5)	50 nM (AS)	50 nM (MF)	50 nM (AS)
SCAV (see table 2.5)	500 nM	500 nM	500 nM
Add ATP/Mg^{2+} before plate reader? (Y/N)	*N*	*Y*	*Y*
ATP	4 mM	4 mM	4 mM
$MgCl_2$	10 mM	10 mM	10 mM

Reaction plates are placed in a Clariostar microplate reader (BMG Labtech) and allowed to equilibrate at 25 °C for 30 minutes. Cy3 fluorophores are excited using an excitation wavelength of 550 nm, and emission is detected at a wavelength of 570 nm. The gain and focus height are set to the highest valued maximum fluorescence (MF) control well on the plate, the time course is then started. Baseline measurements are recorded every minute for ~10 minutes, after which 20 µL ATP/Mg^{2+}is added to the reaction wells. Recordings then proceed for a further 30 minutes at which point the data are exported from the MARS analysis software (BMG Labtech).

2.5.2.3 Data analysis

Data are analysed using a custom R script (Supplementary Information 7.4). Average values are calculated for both the maximum and no helicase controls. The no helicase control is then subtracted from every reaction and time point in the dataset to remove background noise caused by non-perfect annealing. Each adjusted reaction time point is then divided by the MF control and multiplied by 100, to give the proportion of substrate unwound in %. Where possible, values are then averaged over a series of wells and experiments. Typically, the net unwinding plot represents the average proportion of substrate unwound in 30 minutes for at least 4 experimental wells. To calculate 'lag time', the first derivative of each experimental unwinding curve is calculated. The time at which the maximum gradient is observed, is subsequently extracted as the lag time. The relationship between protein concentration [MCM] and lag time may be described by an exponential decay equation, where:

$$\text{Lag Time} = \text{A}e^{\beta[MCM]} + \theta$$

A, β and Θ are constants.

2.5.3 Fluorescence anisotropy

2.5.3.1 Background

When fluorophores are illuminated with linearly polarized light, specific orientations of the fluorophore are favourably excited by the electrical field of the photon. If fluorophores were static in solution (or there was no delay between excitation and emission), the emitted photons would be equally polarized. However, during the excitation period, fluorophores tumble randomly in solution leading to depolarization of the emitted photons[189]. As the period of excitation is relatively constant, the degree of depolarization directly relates to the tumbling rate of the fluorophore. To this extent, fluorescence polarization can be exploited to monitor the tumbling rates of fluorescently labelled ligands. Importantly increasing the concentrations of a ligand binding partner is expected to change polarization through tumbling[191]. This is commonly used to calculate the binding affinity of protein-nucleic acid interactions[192]. Anisotropy relates to the degree that emission polarization is retained.

2.5.3.2 Data collection

Both protein and DNA were prepared in the stated buffer (default: 250 mM KGlu, 20 mM, Tris-Cl pH 8.0). MCM samples were serially diluted in 50 µL volumes then mixed with 50 µL of 2 nM FAM-labelled substrate (default: FAM-fork; see Table 2.5) on a black flat-bottom 96-well plate (Thermo Scientific). Where present, nucleotide was added to a final concentration of 4 mM supplemented with 10 mM $MgCl_2$. Samples were incubated at room temperature for 30 minutes, then briefly centrifuged (1,000 x *g*, 1 minute, 20 °C). Fluorescence readings were taken on a Clariostar microplate reader, where 6-FAM labelled DNA is excited at 495 nm and emission is recorded at 517 nm (BMG Labtech). The gain was set to a control well containing 100 µL of 1 nM FAM-labelled substrate in the appropriate buffer. Anisotropy was calculated within the MARS software (BMG Labtech).

2.5.3.3 Data analysis

Calculation of binding affinity was performed using a custom R script. Briefly, the change in anisotropy is calculated for each experimental dilution series by subtracting the anisotropy of the no protein control well. The average change in anisotropy (*ΔA*) was calculated for 3 experimental repeats, then fit to a Langmuir single binding isotherm, where:

Equation 2.4: Langmuir isotherm with Hill coefficient.

$$\Delta \mathrm{A} = \frac{(\mathrm{B_{max}}[\mathrm{MCM}])^{\mathrm{n}}}{[\mathrm{MCM}]^{\mathrm{n}} + [\mathrm{K_d}]^{\mathrm{n}}}$$

B_{max} is the maximum change in anisotropy, K_d is the equilibrium dissociation constant, and *n* is a Hill coefficient. The presence of a Hill coefficient is justified by the observation of multiple MCM binding stoichiometries in EMSA here and in previous studies[101]. Model fitting was performed using a self-starting non-linear least squares function from the minpack.lm package[193].

2.5.4 Nano Differential Scanning Fluorimetry (NanoDSF)

2.5.4.1 Background

The energy of the emitted wavelength (λ_{em}) is determined by the difference between the ground (S_0) and excited energy state (S_1). Numerous factors can influence λ_{em} including pH, temperature and solvent polarity[194]. For solvatochromic fluorophores such as tryptophan and tyrosine, increasing solvent polarity results in lower excited energy states through solvent relaxation[194]. Smaller energy differences between the ground and excited states results in a longer emission wavelength. Tryptophan is a hydrophobic amino acid and is typically buried within the cores of proteins. When a protein is denatured through either chemical or thermal treatment, the tryptophan environment is subjected to changes in solvent polarity. This is the basis of NanoDSF, which monitors the differential change in fluorescent signal at 330 nm (non-polar) and 350 nm (polar) due to unfolding[195]. Amino acids such as phenylalanine also contribute to this signal, however this contribution is much smaller. NanoDSF equipment often includes additional detectors that detect light scattering to determine protein aggregation.

2.5.4.2 Data collection and analysis

MCM samples were prepared in 25 µL volumes in the stated buffer to a final concentration of 1.25 mg/mL. Samples were then loaded into NT.48 capillaries via capillary action and placed into the Prometheus NT.48 Nano-DSF (Nanotemper Technologies) instrument. A single thermal denaturation ramp was performed between 20–95 °C using a gradient of 1.0 °C /min. The laser excitation power was set to 10 %. Light scattering was also recorded.

Because the tested MCM possesses 3 tryptophan residues, the fluorescence signal is a convolution of oligomerization and individual subdomain unfolding events. Therefore, aggregation which occurs once per unfolding cycle is chosen as a proxy for stability. The T_{agg} was calculated by extracting the inflection point from the first derivative of the scattering data. To estimate binding affinity, the T_{agg} from a no-ATP control was removed from each experimental well to yield ΔT_{agg}. To estimate the $K_{d,app}$, data were then fit to a model, where:

Equation 2.5: Estimation of ATP affinity.

$$\Delta T_{agg} = \frac{(B_{max}[ATP])^n}{[ATP]^n + \left[K_{d,app}\right]^n}$$

B_{max} is the maximum change in ΔT_{agg}, K_d is the dissociation constant of ATP and *n* is a Hill coefficient.

2.6 Particle sizing methods

2.6.1 Size exclusion chromatography (SEC)

Size exclusion chromatography (SEC) is a useful technique for determining the size and interaction of particles in solution[196]. SEC columns are generally packed with a porous matrix of crosslinked carbohydrates, such as agarose or dextran[196]. Columns are graded depending on the size range of cavities within the matrix. When a sample is passed through an appropriate column, smaller molecules are allowed to pass through a larger number of pores, thereby increasing the overall path length taken by the particle. Consequently, the size of the particle determines the retention time on the column. The composition of the eluate can be further analysed by coupling SEC to techniques such as UV-spectroscopy and multiangle laser light scattering (MALLS).

2.6.2 Analytical SEC (with UV-spectroscopy)

2.6.2.1 Background

Calibration of a SEC column with molecular weight standards allows estimation of the assembly of biological molecules with unknown shape and size. Many liquid chromatography systems such as ÄKTA Pure (GE healthcare) possess in-line UV-detectors that can monitor multiple wavelengths simultaneously. This allows discrimination of biomolecules with distinct absorbance properties in the UV spectrum.

2.6.2.2 Data collection and analysis

A Superose S6 Increase 10/300 GL analytical column (GE Healthcare) was equilibrated by passing through 1.2 CV of the analytical SEC buffer on an ÄKTA Pure (GE Healthcare). Where nucleotide was present, the analytical SEC buffer was adjusted by addition of 1 mM ATP, 10 mM $MgCl_2$. MCM were dialysed and diluted to 60 µM hexamer in the running buffer. Where ligands were present, 5 mM ATP, 10 mM $MgCl_2$ or 10 µM ssDNA-pT_{50}-FAM was added to the sample. Briefly, 200 µL of sample was loaded onto a 100 µL loop (an excess was used to account for system void volumes). Exactly 100 µL samples was then passed over the column at a flow rate of 0.5 mL/min. Protein elution was monitored through absorbance at 290 nm, whilst presence of FAM-labelled DNA was monitored through absorbance at 495 nm. The elution was collected in 400 µL fractions, for analysis on 18 % SDS-PAGE. Column calibration was performed using Thyroglobulin, β-amylase, Carbonic anhydrase, Cytochrome C and Alcohol dehydrogenase (Merck). The estimated molecular weight (MW) is related to the elution volume via linear regression.

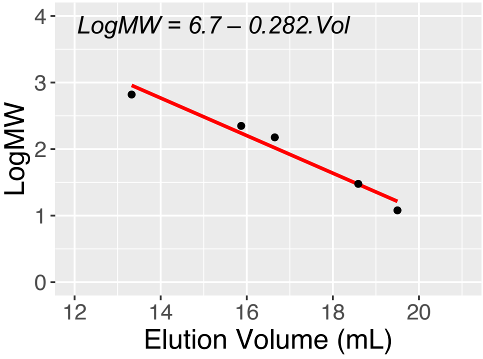


Figure 2.5: Gel filtration column calibration.
Calibration was performed on a Superose 6 Increase 10/300 GL. MW standards used were: Thyroglobulin (660 kDa), β-amylase (223 kDa), Alcohol dehydrogenase (150 kDa), Carbonic anhydrase (30 kDa) and Cytochrome C (12 kDa).

2.6.3 Size exclusion chromatography with multi-angle laser light scattering (SEC-MALLS)

2.6.3.1 Background

The intensity of scattered light relates directly to both the molecular weight (MW) and concentration (*c*) of a population of biomolecules in solution[197]. Therefore, if the intensity of scattered light can be determined for a known concentration of biomolecules, one may derive the molecular weight of the particles. However, in a population of mixed polymers light scattering measurements are biased by large particles (e.g., soluble aggregates). Coupling light scattering with SEC helps to improve the accuracy of light scattering measurements by providing a continuous stream of (ideally) monodisperse analyte. In SEC-MALLS, the biophysical properties of the elution are measured using a series of detectors.

Whilst a known quantity of protein is applied to the SEC column, the concentration in the elution at any one point of time must be determined. SEC-MALLS apparatus consists of both UV and differential refractive index (DRI) detectors to evaluate protein concentration of the elution. DRI detectors are generally more sensitive than UV, which can be easily saturated and are dependent on a reliable extinction coefficient[198]. As the RI of most proteins is very similar, calculation of protein concentration from DRI in each buffer is normalised using a bovine serum albumin (BSA) standard. Integration of a given elution peak provides the quantity of protein.

Light scattering is measured by 16 detectors arranged at discrete angles around the cell. When molecules are particularly large, >10 nm, scattering intensity at the detectors is anisotropic[197]. Determining the angular dependence of scattering can permit calculation of the radius of gyration (R_g)[197]. The scattered intensity determined by the MALLS detectors is examined as the excess Rayleigh ratio (*RΘ*). *RΘ* compares the excess scattering from the analyte versus the scattering of pure solvent with respect to the incident light. The M_W can then be calculated using a Zimm model, that uses the measured *RΘ* and protein concentration, *c* in g/mL [199]:

Equation 2.6: The Zimm model.

$$\frac{K^{*}c}{R\,(\theta,\,c)} = \frac{1}{M_W\,P(\theta)} + 2A_2c$$

Where *K** is a constant that is calculated from the refractive indexes of the solvent and the polymer, *R(Θ, c)* expresses the Rayleigh ratio of the solute as a function of the scattering angle and concentration, *P(Θ)* is the angular dependence of scattered light, A_2 is the second viral coefficient.

2.6.3.2 Data collection and analysis

A Superose S6 Increase 10/300 GL analytical column (GE Healthcare) was equilibrated overnight with SEC-MALLS buffer on a Shimadzu HPLC system. A total of 100 µL protein at 1-10 mg/mL was passed over the column at a flow rate of 0.5 mL/min. Light scattering was determined using a Wyatt HELEOS-II multi angle light scattering detector. DRI was determined using a Wyatt rEX refractive index detector. Data were analysed using Astra 7 (Wyatt) software, where the MW is calculated from a Zimm model. BSA run at 2.5 mg/mL was used to normalize the DRI signal. The *dn/dc* value was adjusted until the expected MW of BSA (66 kDa) was obtained.

2.6.4 Sedimentation Velocity Analytical Ultracentrifugation (SV-AUC)

2.6.4.1 Background

SV-AUC allows the determination of the shape and homogeneity of biomolecules in a matrix free environment[200]. Molecules denser than a surrounding solvent will sediment when subjected to a strong gravitational force provided by a high-speed centrifuge[200]. The centrifugal force *F*, experienced by a biomolecule is given by:

Equation 2.7: Calculation of centrifugal force.

$$F = M_p \omega^2 r$$

where M_p is the mass of the object rotating at a distance *r*, at an angular velocity of ω[201]. However, the centrifugal force is opposed by the buoyant force of the surrounding solvent. Therefore, the true force experienced is dependent on the particle's buoyant mass, where M_p is substituted for M_b. The M_b of a particle is dependent on the mass of solvent displaced, which can be calculated by taking into consideration the partial specific volume of the particle $\bar{v}$, multiplied by the density of the solvent displaced ρ, thus:

Equation 2.8: Calculation of a particles buoyant mass.

$$M_b = M_p(1 - \bar{v}\rho)$$

Hence, in a lower density solution a particle will have a larger buoyant mass and will experience a greater centrifugal force. As a particle moves through solution, it is also subjected to frictional drag forces that are proportional to the velocity of the particle[201]. The frictional drag force F, is derived simply by multiplying the frictional drag coefficient *f*, by the molecule's velocity *v*:

Equation 2.9: Calculation of frictional force.

$$\mathrm{F} = f\mathrm{v}$$

The frictional drag coefficient is dependent on the size and the shape of the particle in solution, and the temperature and viscosity of the solvent[200]. In equilibrium, the forces cancel, such that:

$$\mathrm{s} = \frac{v}{\omega^2 r} = \frac{M_b}{f}$$

This allows derivation of the sedimentation coefficient, which describes the time taken (*s*) for a particle to reach terminal velocity in the absence of drag under a given acceleration[200].

In a typical SV-AUC experiment, the sample is contained within a rectangular cell, directed along the radius of the centrifugal force. The concentration distribution of particles *c*, with respect to time *t*, is determined through either UV or interference readings at defined radial increments *r*. Particle's transition through the cell at a rate proportional to the sedimentation coefficient, however diffusion attempts to redistribute the concentration of particles throughout the cell. The sedimentation coefficient determines the motion of the distribution with respect to time, whilst the size of the diffusion coefficient *D*, determines the shape of the boundary. Consequently, the radial concentration distribution and shape with respect to time, t, may be mathematically described by the Lamm equation[202]:

Equation 2.10: The Lamm equation.

$$\frac{\partial c}{\partial t} = D\left[\left(\frac{\partial^2 c}{\partial r^2}\right) + \ \frac{1}{r}\left(\frac{\partial c}{\partial r}\right)\right] - s\omega^2\left[r\left(\frac{\partial c}{\partial r}\right) + 2c\right]$$

It is important to note that there is no exact solution to the Lamm equation, only approximates[203]. To properly fit this equation to SV-AUC data, prior knowledge may be provided. For example, the diffusion coefficient may be calculated using the Stokes-Einstein equation:

Equation 2.11: The Stokes-Einstein equation.

$$D = \frac{RT}{Nf}$$

where *R* is the molar gas constant, *T* is the temperature in *K*, *N* is Avogadro's number and *f* is the frictional coefficient. The frictional coefficient can be calculated *in silico*, by making assumptions about the partial specific volume of the molecule and the viscosity of the solution[204]. The distribution of sedimentation coefficients in a polydisperse sample may then be evaluated by directly fitting data to a continuous c(s) distribution. This method fits a sum of Lamm solutions to determine the relative population of sedimenting species within a sample[203]. If desired, the Svedberg equation may then be used to calculate the molar mass from sedimentation and diffusion coefficients:

Equation 2.12: The Svedberg equation.

$$M_r = \frac{\mathrm{sRT}}{DM_b}$$

2.6.4.2 Data collection and analysis

Experiments were performed on an Optima XL/I ultracentrifuge (Beckman Coulter). All samples were dialysed overnight into SEC-MALLS buffer, to ensure sample-reference buffer homogeneity. The density and viscosity of the buffers was calculated using the software SEDNTERP[204], yielding values of 1.0234 g.ml^{-1}, and 0.011787 g.cm^{-1}.s^{-1} respectively. Sedimentation velocity data were recorded for a period of 7 hours whilst centrifuging at 30,000 rpm, 20 °C in an AN50Ti rotor (Beckman Coulter). Sedimentation was monitored through absorbance at 280 nm. The partial specific volumes used in analysis were determined based on amino acid composition and were 0.7436 ml.g^{-1} for *Mac*MCM and 0.7502 ml.g^{-1} for *SsoPfu*MCM. Data analysis was performed using the program SedFit[203]. Data was fit to a continuous c(s) distribution with prior knowledge.

2.6.5 Fluorescent-based affinity capture assay

2.6.5.1 Background

Affinity capture assays are simple biochemical experiments for determining a physical interaction between particles[205]. Briefly, a bait protein possessing an affinity tag is immobilized onto a column. An untagged partner protein is mixed with the bait protein that under certain conditions may interact. The interaction can be assessed by analysing the components in the elution.

2.6.5.2 Data collection

1 mg of MCM sample was loaded onto a 1 mL HP Ni-NTA column (GE Healthcare) equilibrated in affinity capture buffer using an ÄKTA Pure (GE healthcare). Mixed 1 mg samples contain an equimolar ratio of MCM partner: bait. Bound proteins were washed with 10 CV affinity capture buffer, then eluted using an isocratic wash of 10 CV buffer B. The elution was collected in 1 mL fractions on a 96-well deep well plate. 5 µL of selected fractions were assessed on 12 % SDS-PAGE. To determine the presence of fluorophores, gels were imaged on a Typhoon gel scanner (GE Healthcare) using the Cy2 method and a

pixel size of 100 µm. After imaging, gels were stained with NR-CBBR dye to detect protein bands.

2.7 NMR spectroscopy

2.7.1 Background

NMR spectroscopy is an essential tool for the elucidation of chemical structures[189,206]. Atomic nuclei possess a quantum-mechanical property called 'spin' that is defined by the number of protons and neutrons of an atomic isotope. Nuclei with spin ½ possess a magnetic dipole moment that can interact with an electromagnetic field. When an external electromagnetic field is applied, ½ spin nuclei assume 2 possible orientations that are parallel or antiparallel to the magnetic field[189]. Energetically speaking, the parallel state is a lower energy conformation than the antiparallel state and hence is adopted by most nuclei. The nuclear magnetic dipole precesses around the magnetic field at its Larmor frequency:

Equation 2.13: Calculation of Larmor frequency.

$$\omega = -\gamma B$$

Where ω is the Larmor frequency, γ is the nuclear gyromagnetic ratio and B is the strength of the magnetic field in T. Therefore, the Larmor frequency of a nucleus is dependent on the nature of the nucleus and the magnetic field that it experiences. For a typical NMR spectrometer with a field strength of 11.7 T the Larmor frequency for many nuclei is equivalent to electromagnetic radio waves (e.g., 1H = 500 MHz; 31P = 202.5 MHz). Importantly, the magnetic field individual nuclei perceive is extremely sensitive to additional magnetic contributions from within the sample. Electrons create additional magnetic fields that oppose the applied field. Nuclei that are better shielded by electrons require a lower resonant frequency (higher energy) and vice versa. Chemical shifts describe the effect of local magnetic contributions on the Larmor frequency of a given atom[189]. The magnetic field of nearby spin 1/2 nuclei may also interact through dipole-dipole coupling.

In an NMR spectroscopy experiment the aim is to determine the chemical shift of nuclei in each sample. Modern instruments apply a broad frequency radio (r.f) pulse 90° to the external magnetic field. Where the Larmor frequency is matched, atoms absorb energy and rotate to align with the r.f pulse. As the nuclear spin relaxes back to the equilibrium position, the rotation of the magnetization vector generates current in a receiver coil. The current signal forms a decaying sine wave as nuclei realign with the field and hence is commonly known as a free induction decay (FID). The FID is a convolution of exponential decays of every perturbed nucleus in the sample and can be converted into its constituent sine waves (unique nuclei) through a Fourier transform. NMR spectroscopy is often deployed to monitor chemical reactions, where the integral of each peak is used to examine the proportional change in chemical environments over time[118]. To this extent NMR spectroscopy was used to monitor NMR active ^{31}P nuclei during enzymatic ATP hydrolysis

2.7.2 Data collection

^{31}P-NMR spectroscopy was used to determine ATPase activity of MCM. Reactions were performed at the stated temperature in 600 µL, containing 50 µM (monomeric) MCM, 50 mM ATP, 2.5 mM $MgCl_2$, 10% (*v/v*) D_2O, 250 mM KGlu and 10 mM Tris-Cl pH 8.0. Where DNA was present, the final concentration was 8.3 µM of 'fork' substrate (see Table 2.4). A time series of 1D ^{31}P-spectra was recorded on a 500 MHz NMR spectrometer (Bruker). Spectra were recorded at 2 minutes intervals for at least 30 minutes.

Individual peaks in the spectra were integrated using TopSpin software (Bruker) and analysed by linear regression analysis performed in R. The change in β-phosphate peak integral was primarily used to estimate $K_{cat\ app}$, where integral values are adjusted to reflect the starting concentration of 50 mM ATP.

2.8 X-ray crystallography

2.8.1 Background

X-ray crystallography is a widely used technique for elucidating the distribution of atoms in a repetitive crystalline lattice[207]. When a crystal is illuminated by an X-ray beam, X-rays are (primarily) scattered elastically by electrons. Scattered X-rays interfere and in most directions the waves interfere destructively. Under reflection conditions outlined by Bragg's law, scattered waves interfere constructively and enhance the signal exponentially to create a measurable reflection [207]. Bragg's law explains that waves will constructively interfere when the pathlength taken by two coplanar waves (*2dsinθ*) of wavelength *λ*, differ by a whole integer *n*:

Equation 2.14: Bragg's law.

$$n\lambda = 2dsin\theta$$

Experimentally, Bragg's law relates the interplanar spacing *d*, to the directions of reflections, and thus the position of a measured signal on a detector, indexed by Miller indices *hkl*. Hence, higher resolution datasets that successfully probe smaller interatomic distances include signal recorded at a larger diffraction angle. The diffraction pattern is also dependent on the dimensions of the unit cell and the symmetry and spatial distribution of electrons contained within. The unit cell is defined as the smallest parallelogram within the lattice that can be used to recreate the entire crystal by translation alone. The intensity and distribution of diffraction pattern is related to the electron density at a point in the unit cell *ρ*(*xyz*), through an inverse Fourier transform[208]:

Equation 2.15: The electron density equation.

$$\rho(xyz) = \frac{1}{V}\sum_{\substack{hkl \\ -\infty}}^{+\infty} |F(hkl)| e^{-2\pi i[hx+ky+lz-\phi(hkl)]}$$

Where V is the volume of the unit cell, *hkl* are the Miller indices of the diffracted beams, F(*hkl*) or 'structure factors' are the amplitudes of the diffracted beams of all atoms in the unit cell in the direction *hkl*, and *Φ*(*hkl*) are the relative phase of the structure factors. Phase cannot be directly measured and must be determined through methods such as single/multiple-wavelength anomalous dispersion (SAD/MAD) or molecular replacement

(MR)[207]. The method that the phase problem is solved will influence the data collection strategy, particularly when SAD or MAD is involved.

SAD/MAD phasing exploits anomalous scattering to determine the position of heavy atoms distributed in a crystal lattice[207]. Friedel's law states that reflections related by inversion through the origin will have equal intensity[208]. Usually the atomic scattering factor *f,* is proportional to the number of electrons possessed by an atom. At certain wavelengths, X-rays are absorbed by atoms causing a change in the atomic scattering factor[208]. The atomic scattering factor is then described by an additional real *f'*, and imaginary component *if''*, that outline the phase shift between the Friedel pair[208]. This phase shift causes a breakdown in Friedel's law and allows the position of anomalously scattering atoms in the lattice to be determined. The phases of the anomalously scattering atoms can then be estimated, providing an initial solution to the electron density equation.

2.8.2 Data collection principles

Many heavy atoms absorb X-rays in the range useful for crystallography. Heavy atoms can be introduced into the crystal lattice through protein labelling strategies (e.g., Selenomethionine substitution)[209]; naturally occurring ion binding sites (e.g., Zn fingers)[210]; post-crystallization metal soaking (e.g. $PbCl_2$)[211]. To maximize the anomalous signal in a SAD dataset it is possible to tune the wavelength of a synchrotron X-ray source to the absorbance peak of the heavy atom. Both the presence of heavy atoms and selection of the appropriate wavelength can be determined through an X-ray fluorescence (XRF) scan.

The 3D reciprocal lattice may be reconstructed by processing the diffraction patterns collected at discrete radial increments. Data is first indexed to retrieve essential metrics for calculating *p(xyz)*[212]. Individual reflections are identified and the position is labelled with Miller indices (*h,k,l*). The geometrical arrangement of reflections is then analysed to disclose information about the unit cell dimensions and the symmetrical point group of the crystal. Next, to determine the structure factors the intensity of individual reflections is measured. Each spot should form a smooth Gaussian curve that is integrated to determine the intensity. The uncertainty of any one reflection can be calculated by comparing the intensity of the reflection (I) against the estimated variance of the background signal (σI)[207]. Spot intensity decreases at higher diffraction angles, hence higher resolution reflections have poor signal to noise ratios. Intensities may be adjusted or 'scaled' to account for inconsistencies during data collection that may arise from non-uniformity of the beam, detector, changes in beam intensity and volume of crystal illuminated[212]. According to the symmetry of the crystal, reflections may be recorded multiple times during data collection. Merging removes outliers and calculates the average intensity for identical reflections[212]. R_{merge} is a statistic that describes the self-consistency of measurements made on identical reflections within a dataset, however, R_{merge} is negatively biased with increasing multiplicity (the number of times an identical reflection is recorded)[213]. Therefore, R_{meas} is typically calculated to adjust R_{merge} for multiplicity, where:

Equation 2.16: $R_{meas.}$

$$R_{meas} = \frac{\sum_{hkl} \sqrt{\frac{n}{n-1}} \sum_{i=1}^{n} | I_i(hkl) - \bar{I}(hkl) |}{\sum_{hkl} \sum_{i=1}^{n} I_i(hkl)}$$

To determine the maximum resolution of a dataset, a correlation coefficient $CC_{1/2}$, is widely used that compares the average intensity of a random half of measurements for each reflection[214], where:

Equation 2.17: $CC_{1/2}$

$$CC_{1/2} = \frac{< I^2 > \; - < I >^2}{< I^2 > \; - < I >^2 + \; \sigma_e^2}$$

At low resolution, $CC_{1/2}$ values are near 1 and drop toward 0 at high resolution. Importantly, $CC_{1/2}$ may be used to determine the resolution where measured intensities are no-longer statistically significant. CC* attempts to correlate the averaged dataset after merging with the real atomic values that they represent[214]. This is estimated using:

Equation 2.18: CC*

$$CC^* = \sqrt{\frac{2CC_{1/2}}{1 + CC_{1/2}}}$$

This provides a statistic to compare the quality of the data and the calculated model.

2.8.3 Molecular replacement

Whilst SAD directly infers the position of heavy atoms within the unit cell, molecular replacement (MR) uses homologous models to calculate the initial phases[215]. MR exploits the structural similarities between related proteins that may approximate to an unknown asymmetric unit. Modern MR programmes use a likelihood function to calculate the probability of observing the experimentally derived structure factors from a given model[216]. Likelihood is maximised by optimization of model placement in an asymmetric unit through rotation and then translation functions. Importantly, likelihood methods account for the true atomic coordinate errors between the model and the unknown target, calculated as a C_α root mean squared deviation (RMSD). The C_α RMSD is roughly proportional to sequence identity as a function of the number of residues, hence the RMSD for an unknown structure can be estimated[217]. The RMSD agreement may be improved by removal of flexible loop regions. Generally, MR is able to perform when the RMSD is < 2.5 Å[218].

In Phaser, MR solutions are assessed by two statistics: log-likelihood gain (LLG) and translation function Z-score (TFZ)[218]. The LLG score is a measure of how well the model agrees with the recorded data. This is calculated by comparing the likelihood of agreement between the dataset and the model versus a set of identical atoms which are randomly distributed. The TFZ score reports the standard deviation of a specific translation function LLG from the mean of all the TF LLGs. It can be considered as the signal to noise of a given translation function. Generally, LLG of >120 and TFZ >8 indicate definitive MR solutions. To calculate the electron density, phases are calculated from the model and combined with the structure factors of the experimentally derived data.

2.8.4 Phase improvement, model building and refinement

Crystallographic models may then be improved by examining the model in the electron density map. In many instances, the initial phases and hence the electron density map is poor. Phases can be refined by improving the relationship between the model and the data. Alternative electron density maps may also be calculated, providing additional information to guide model building. Density modification introduces a number of real space constraints for the electron density to improve the phase estimates[219]. The constraints include the shape of the solvent, the shape of atom peaks and non-crystallographic symmetry (NCS)[220]. Anomalous maps are able to detect the location of anomalous scatters in a dataset, even when the signal is weak[221]. Improvement of the electron density map allows better model building.

Refinement of a crystallographic model ensures that a model is faithful to the experimental structure factors and theoretical limits of stereochemistry. These are monitored by several statistics. R factors measure the agreement between observed and calculated structure factors.

Equation 2.19: R-factor

$$R = \frac{\sum_{hkl} | F_{obs}(hkl) - F_{calc}(hkl) |}{\sum_{hkl} | F_{obs}(hkl) |}$$

At the start of refinement, 5 % (or 2,000) random reflections are removed from the refinement dataset; these reflections are used to calculate R_{free}. R_{work} is calculated on the refinement dataset and compared against R_{free} to determine overfitting of the experimental data. In over-fitted data, the gap between R_{work} and R_{free} will be very large. R_{free} is always larger than R_{work} however, the gap is typically less than 20 %. As the resolution gets worse, R_{free} generally increases, hence a modelled 1.5 Å dataset will yield on average an R_{free} of ~0.21, whilst a 3.0 A dataset will yield an R-free of ~0.28[222]. CC* may be used as a refinement target against CC_{work} and CC_{free}, which determine the correlation between the observed and calculated structure factors.

Validity of model stereochemistry is determined through metrics including Ramachandran outliers, clashscore and the RMS for bond angles and lengths. Ramachandran outliers determine amino acids with non-favourable dihedral angles. Clashscore determines the number of serious atomic clashes per 1,000 atoms. RMS bond angles and lengths report how well the geometry of the atomic bonds in a model structure conform to a set of ideal stereochemical values.

2.8.5 Initial crystallization and optimization

Crystallization of proteins is a notoriously unpredictable process[223]. Typically, hundreds of conditions are tested to yield high quality diffracting crystals. Crystallization conditions are examined using commercially available crystallization screens, including Morpheus (Molecular Dimensions), PEG/Ion (Hampton), PACT (Molecular Dimensions) and PDB

(Molecular Dimensions)[224]. Screening is performed on 96-well sitting drop trays. First, 54 µL of crystallization condition is added into each mother liquor cell on the plate. Each cell serves two crystallization areas, allowing assessment of multiple protein concentrations in parallel. Using a Mosquito LCP robot (TTP Labtech), two protein concentrations (~5-9 mg/mL) are mixed with the cell solution at a ratio of 100 nL:100 nL (1:1) in the appropriate crystallization area. ClearVue (Molecular Dimensions) sheets are used to seal the plates. Plates are then placed at either 4 or 20 °C and regularly checked for crystal growth.

Crystals are harvested using a cryo-loop (Crystal Cap HP) and plunged into liquid nitrogen. Crystals are then screened using an in-house rotating anode X-ray source (MicroMax-007 HF), equipped with an image plate detector (Mar345). Images are then viewed using Adxv software. Where diffraction is observed, crystals are optimized by varying the salt, protein, and precipitant concentrations. Optimizations are performed in 24-well trays in either the hanging or sitting drop format over a 1 mL reservoir. When the hanging drop format is used, 2 µL protein solution is mixed with 2 µL reservoir solution on a siliconized glass slide and sealed using grease. When the sitting drop format is used, 10 µL protein is mixed with 10 µL reservoir solution in the well and sealed with a glass slide and grease.

2.8.6 Crystallization of *Mac*MCM$^{\Delta WHD}$

Crystals, grew at 20 °C in the hanging drop format containing 2 µL of 7 mg/mL protein solution and 2 µL of reservoir solution containing 400 mM $(NH_4)_2SO_4$, 0.1 M Bis-Tris-Cl pH 6.5, 25 % (*w/v*) PEG 3,350. Crystals were harvested with a cryo-loop (Crystal Cap HP) and flash frozen in liquid nitrogen. Data were collected at Diamond Light Source, at a wavelength of 0.9795 Å at a temperature of 100 K. Data were indexed, scaled and merged using Xia2-DIALS software package to a resolution of 4.02 Å[225].

2.8.7 Crystallization of *Mac*MCM$^{\Delta WHD\text{-}E418Q}$

Hexamer formation was induced 10 minutes before setting up the plates by addition of 10 mM ATP and 10 mM $MgCl_2$. Long plate-shaped crystals, grew at 20 °C in a sitting drop containing 10 µL of 7 mg/mL protein solution and 10 µL of reservoir solution containing 0.03 M NPS, 0.1 M MOPS/HEPES pH 7.5, 10 % (*w/v*) PEG 20,000, 20 % (*v/v*) PEG MME 550. Crystals were harvested with a cryo-loop (Crystal Cap HP) and flash frozen in liquid nitrogen. Data were collected at Diamond Light Source, at a wavelength of either 0.9763 Å or 1.2828 Å at a temperature of 100 K. Data were indexed, scaled and merged using Xia2-DIALS software package to 2.59 Å resolution[225].

2.8.8 Structure determination and refinement of *Mac*MCM$^{\Delta\text{WHD-E418Q}}$

A custom search model from the AAA+ domain of a single *Pfu*MCM (PDB:4R7Y)[124] subunit was generated. In total, this model shares 50 % sequence identity with the AAA+ domain of *Mac*MCM. The model was prepared by removing loops manually in PyMol, then the residues were truncated to polyalanine, using the 'PDB Tools' editor in Phenix[219]. Initial phases were calculated using Phaser molecular replacement software[218], which placed six copies of the search model into a ring (PDB: 4R7Y[124]). Following the placement of this model, the electron density map was improved using RESOLVE density modification[226]. Loops and the entire N-terminal domain were then built iteratively using manual building in Coot[227] and automated model building software, Autobuild[219] and Buccaneer[227]. Refinement was carried out using phenix.refine[228]. Isotropic B-factors were refined, using a single B-factor per residue. Secondary structure restraints were applied to the model during refinement. After building the N-terminal domain, Zn^{2+} ions were placed through observation of weak anomalous data. Cysteine co-ordination and ligand restraints were then generated using ReadySet![229]. Where present, nucleotide in the active sites was modelled as ADP. Seven phosphate molecules were also added into the model density.

2.8.9 Crystallization of *Mac*MCM$^{\Delta\text{WHD-E418Q}}$ and ss(ACTG)$_{16}$

Complex formation was induced 10 minutes before setting up the plates by addition of 10 mM ATP, 10 mM $MgCl_2$, and an equimolar ratio of (ACTG)$_{16}$. Crystals, grew at 20 °C in a sitting drop containing 10 µL of 7 mg/mL protein solution and 10 µL of reservoir solution containing 0.2 Sodium Malonate, 0.1 M Bis-Tris Propane-Cl pH 6.5, 20 % (*w/v*) PEG 3,350. Crystals were harvested with a cryo-loop (Crystal Cap HP) and flash frozen in liquid nitrogen. Data were collected at Diamond Light Source, at a wavelength of either 0.9795 Å at a temperature of 100 K. Data were indexed, scaled and merged using Xia2-DIALS software package to a resolution of 2.64 Å[225].

2.8.10 *In silico* structure analysis

The solvent accessible surface area (SASA) can be used to determine residues involved in interface formation. SASA is initially calculated for residues in each MCM subunit monomer ($SASA_{residue\ monomer}$) using Parameter OPtimised Surfaces (POPS)[230]. Hexameric structures are then split into 6 dimer pairs. Subunits in each pair are defined based on the motifs contributed to the ATPase active site (*cis*- or *trans*-acting). SASA is then calculated for each residue in the dimer ($SASA_{residue\ dimer}$). To determine the residues buried in the interface, the change in SASA (ΔSASA) is calculated between the residues of a chain in a monomer and a dimer conformation.

Equation 2.20: Calculation of ΔSASA for a given residue

$$\Delta SASA_{residue} = SASA_{residue\ monomer} - SASA_{residue\ dimer}$$

Where ΔSASA > 0, the residue is expected to be buried to a degree by the formation of a subunit-subunit interface. For clear comparison, residues are then renumbered from the PDB file to the position in a multiple sequence alignment (MSA) generated by MUSCLE[231]. The average ΔSASA at each MSA position is then calculated for residues across all 6 subunits. As the primary focus is on the evolutionary conserved interfaces, the ΔSASA occurring at <2 interfaces are removed from the analysis to correct for subunit-specific interfaces formed in eukaryotic MCM2-7. To determine interface relatedness, $CC_{interface}$ was calculated. This compares the ΔSASA at each position in the MSA between two structures using a Pearson correlation coefficient.

2.9 Oxford Nanopore Technologies (ONT) flow cell experiments

The theory of ONT sequencing will be discussed extensively in Chapter 5. The DNA-sequencing library is prepared according to ONTs Lambda control expansion kit. Where MCM are tested, ONTs helicase-adapter mix (AMX) was substituted for 1 µM annealed MCM-adapter. Loading samples are then prepared in 75 µL containing 450 mM KGlu, 25 mM HEPES pH 8.0, 6 mM ATP, 10 mM $MgCl_2$. MCM are added at an equimolar ratio of hexamer:DNA, where the final concentration of each is 0.08 nM.

MinION devices are controlled using a developer version of ONTs MinKNOW software, version 1.10.16. The platform QC script was used to determine the number of available pores in each flow cell. MinION R9.4 flow cells were then prepared using ONTs flow cell priming mix. After 5 minutes, the loading sample was then added to the flow cell. Experiments were set to run at 34 °C for 6 hours at a voltage of -180 mV. Static flips were performed every 5 minutes to clear blocked pores. Data analysis was performed in ONTs in-house ‘TraceViewer’ software within LabVIEW (National Instruments).

Chapter 3 – Screening and Biochemical Characterization of a Unique Minichromosome Maintenance Protein from the Mesophilic Archaeon, *Mancarchaeum acidiphilum*

3.1 Introduction

Archaea have long provided a valuable model for understanding the core mechanisms of the Minichromosome Maintenance (MCM) family of helicases. Since the primary characterization of the archaeal *Mth*MCM[78,79], studies have largely focused on MCM from thermophilic organisms. For example, 95 % of archaeal MCM publications examine homologs from organisms that live over 50 °C, with 78 % of these publications focusing on only 3 enzymes (*Mth*MCM, *Sso*MCM and *Tac*MCM) (Figure 3.1a–b). This focus was initially driven by biotechnological interest that investigated thermostable biocatalysts, such as the *Pfu* polymerase[232,233] . Owing to the high stability of thermostable enzymes, this has also supported the success of high-resolution crystal structures of archaeal MCM. All preliminary *in vitro* studies on mesophilic archaeal MCM varieties have been unsuccessful with insolubility and poor yields cited[96,234].

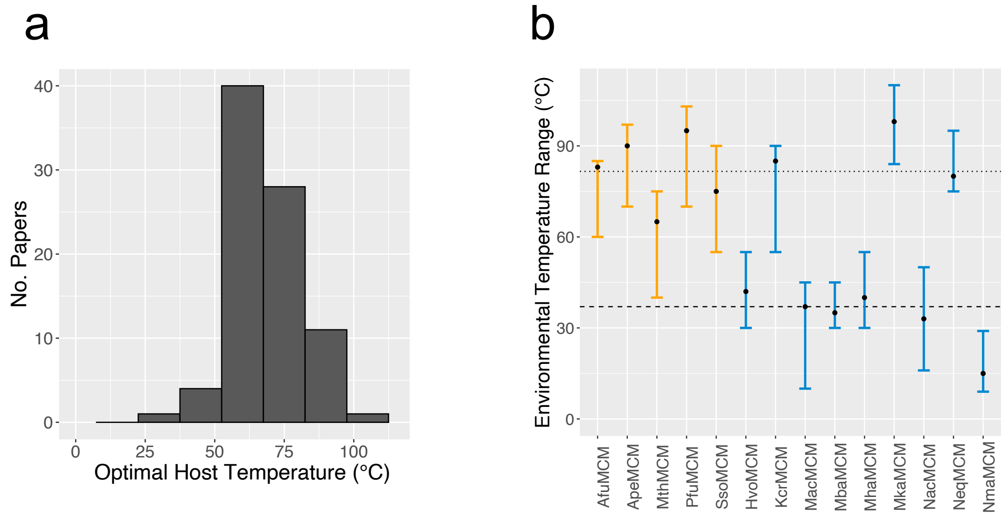


Figure 3.1: Perspective of MCM models based on organism environment.
(a) The number of archaeal MCM publications is compared against the environmental temperature that the MCM would typically be exposed to *in vivo.* **(b)** Environmental temperature range of the MCMs used in this study would be subjected to *in vivo*. Points represent the optimal growth temperature of the organism. Error bars represent the growth range. Orange error bars represent MCM which have been studied previously. Blue error bars represent new, biochemically uncharacterized MCM. References are outlined in Supplementary Table 7.3.

Steady advances in protein purification and structural techniques with low sample requirements, such as cryogenic electron microscopy (Cryo-EM), have introduced discrepancies in the relationship between the archaeal and eukaryotic MCM models. In the absence of ATP or DNA, eukaryotic MCM does not form a stable heterohexamer under physiological conditions [60,157,161]. Most of our understanding of archaeal MCM oligomeric states comes from gel filtration analyses that suggest MCM form stable hexamers in absence of ATP or DNA. However, these experiments have generally been performed well below the optimal growth temperature of the organism due to column and molecular weight standard limits. Gel filtration analyses suggest that at 20 °C *Mth*MCM forms a double hexamer, but increasing the temperature to 50 °C (near the optimal growth temperature), *Mth*MCM forms a hexamer[104]. Indeed, negative stain-EM where samples were incubated at a range of temperatures before grids are prepared, suggest that at the optimal growth temperature *Mth*MCM forms open hexamer[235]. Below the optimal growth temperature, the proportion of higher molecular weight complexes, such as closed hexamers, heptamers and double hexamers increases significantly[235]. Equally, 4 °C is sufficient to support eukaryotic MCM2-7 hexamerization in the absence of ATP or DNA [60].

Therefore, measurement outside native temperatures may promote unnatural oligomeric complexes. Furthermore, at temperatures outside the physiological range, the activity of many enzymes falls exponentially[236]. There is a demand for discovering a more convenient archaeal MCM model through which better comparisons with eukaryotic MCM can be made. Finally, we have an interest in discovering MCMs that are active at room temperature for use in Nanopore sequencing.

3.2 Aims

This chapter contains biochemical screening experiments performed on 16 archaeal MCM homologues and the in-depth characterization of two of these enzymes. Here we aim to:

- Identify MCM that are active in conditions equivalent to eukaryotic MCM studies. These conditions also support our biotechnological applications;
- Identify MCM that can be directly compared against a thermophilic model MCM;
- Assess whether a mesophilic enzyme can be used as a better model for MCM;
- Assess whether temperature can explain the discrepancies between archaeal and eukaryotic MCM models;
- Identify a mesophilic MCM for structural studies.

3.3 Identification of MCM homologues

3.3.1 Selection rationale

Since the first archaeal genome sequence of *Methanococcus janachaii* [7]over 20 years ago, the number of genomes deposited online worldwide has increased rapidly and now includes many uncultured archaeons[237]. To this end, we selected three well-studied archaeal orthologues with structures published in the protein data bank (PDB), to allow for structural comparisons (*Mth*MCM, *Pfu*MCM, *Sso*MCM)[100,112,124,128,238,239]. For further comparison, we included two previously studied but structurally uncharacterised MCMs from other thermophilic organisms (*Ape*MCM, *Afu*MCM)[81,240].

Organisms are uniquely adapted to their environment. This extends to the protein level, where enzymes must be adapted to work efficiently under specific conditions. Therefore, to extend the screen further, we selected nine MCM from distant archaeal phylogenies and environmental niches to investigate the resultant enzymatic activity from environmental selection pressures (Figure 3.1b). Of these, a further three thermophilic MCMs were selected that are found in archaea from higher temperature environments, or from more distant archaeal lineages than previously tested (*Kcr*MCM, *Mka*MCM, *Neq*MCM)[17,241,242]. The remaining six MCMs were selected from genomes of mesophilic archaea (20 °C - 45 °C), which are adapted for life in various distinguishable environments such as: saline (*N. maritimus*)[243], hypersaline (*H. volcanii*, *Nanohaloarchaeota. SG9, M. halophilus*)[244–246], anaerobic (*M. barkeri*)[247] and acidic (*M. acidiphilum*)[18]. Of these fourteen MCM, three are encoded by the genomes of parasitic/symbiont archaea (*Nac*MCM, *Neq*MCM, *Mac*MCM), all of which have extremely small genomes <1 Mb [17,18,246]**.** Finally, we included an active chimeric MCM enzyme with resolved structure: *SsoPfu*MCM and a newly engineered reverse domain swap *PfuSso*MCM[124]. This gave us a total of 16 MCM proteins (Figure 3.1b).

3.3.2. MCM identification and *in silico* analysis

MCM homologue sequences were obtained following a genome search of the selected archaeal organism on the Kyoto Encyclopedia of Genes and Genomes (KEGG) database[248]. All but two archaeal genomes encode a single MCM copy. Firstly, *M. kandleri*, which contains a known inactive orthologue with resolved structure (*Mka*2)[249] . In this instance the longer, expected active *Mka1*MCM was chosen. Secondly in the case of *Nanohaloarchaeota. SG9*, which encodes 2 homologues, the longer, more divergent MCM was selected for variety. A known 368 amino acid intein sequence of was identified and removed from the *Pfu*MCM sequence.

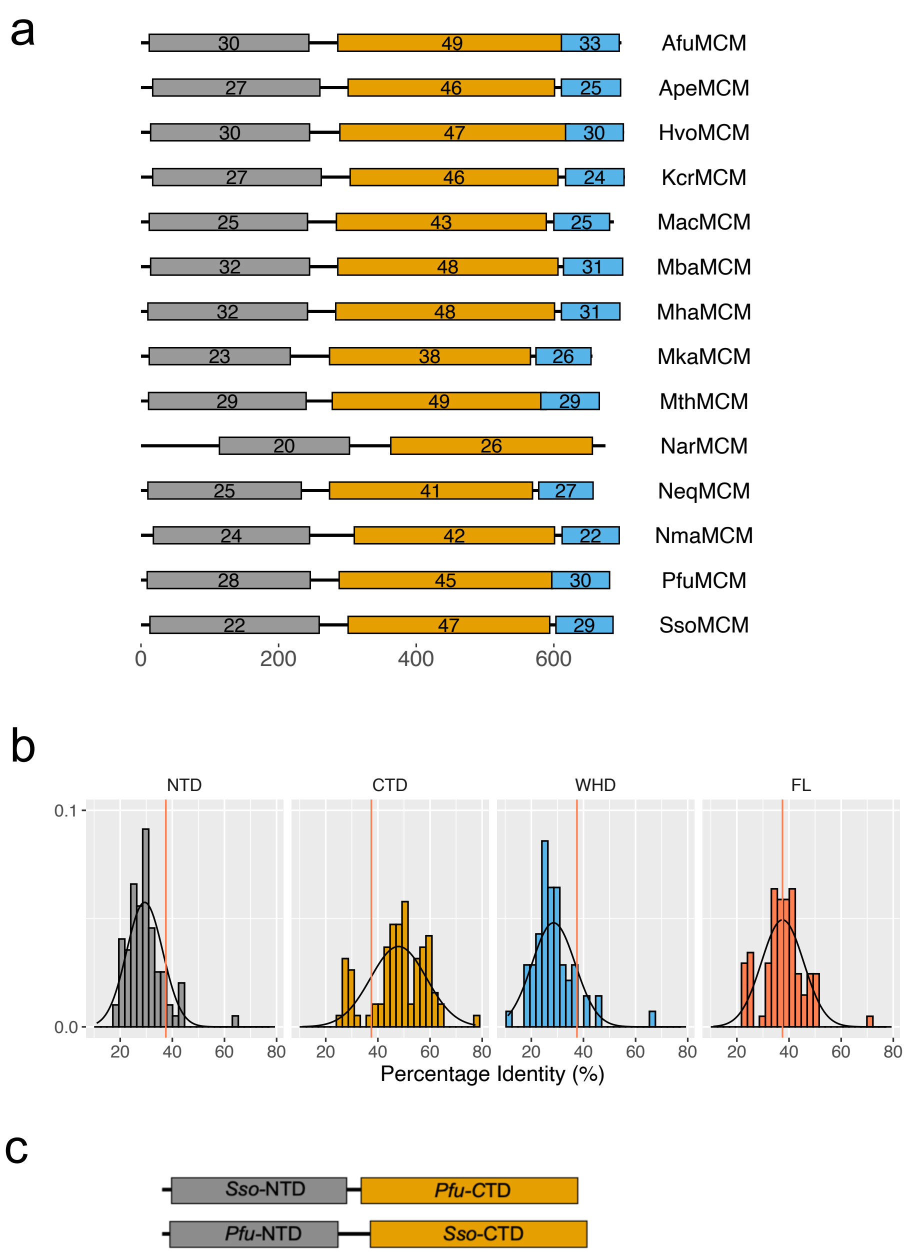
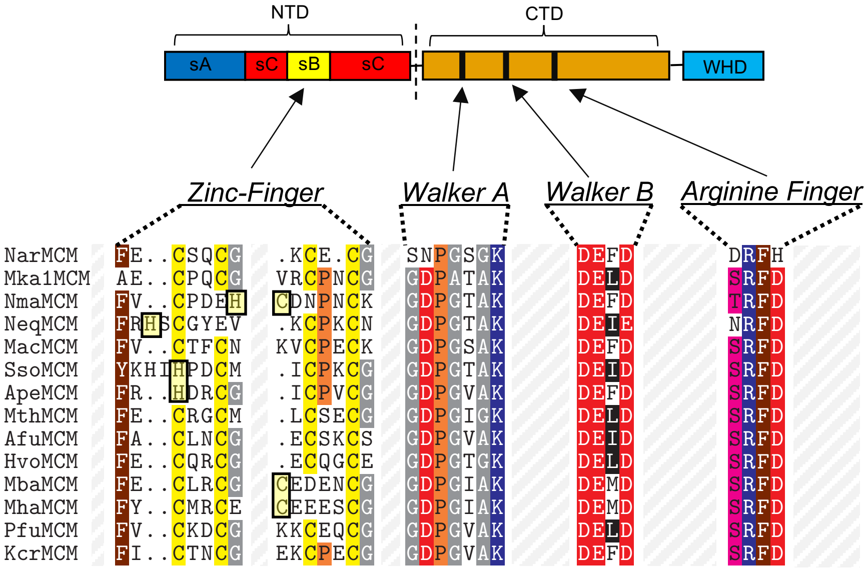


Figure 3.2: Schematic and conservation analysis of MCM used in this study.
(a) Conserved MCM subdomains domains were identified using the InterProScan tool (https://www.ebi.ac.uk/interpro/search/sequence/). Coloured boxes indicate the identified domain and length (amino acids), where grey is a N-terminal DNA binding domain, orange is a P-type NTPase and blue is a DNA binding winged helix domain. Numbers inside the box indicate the average percentage identity score for each MCM subdomain when aligned to all MCM studied here (https://www.ebi.ac.uk/Tools/msa/clustalo/). **(b)** The distribution of percentage identity scores for all MCM and subdomains. The black line represents the normal distribution, with a standard deviation σ and mean μ. The red line represents the mean value for the full-length enzyme. **NTD**: μ=29.4, σ=6.9; **CTD**: μ=47.9, σ=10.7; **WHD**: μ=28.5, σ=8.3; **FL**: μ=37.5, σ=8.1). **(c)** Schematic outlining the design and organization of the chimeric domain-swap enzymes.

Sequences were then analysed using InterProScan[250] to identify conserved domains within each MCM. All MCM are >600 amino acids in length and are predicted to have a N-terminal DNA binding (NTD) domain connected to an AAA+-domain through a linker. A second short linker joins the AAA+-domain to the C-terminal WHD. (Figure 3.2a). *Nar*MCM was identified to lack a C-terminal WHD, however, a Phyre2 search of the first 111 residues predicts a novel N-terminal WHD fold. Conservation of the full protein and each domain was then determined through Clustal Omega which measures the conservation between each aligned protein. Whilst the enzymes share on average 37.5 % conservation, this is skewed by the large AAA+-domain that share on average 48 % identity (Figure 3.2b). Comparatively, the NTD and WHD share more similar conservation of 29.4 % and 28.5 %, respectively. The engineered chimera enzymes were generated by swapping the N- and C-terminal domains of *Sso*- and *Pfu*MCM (Figure 3.2c).



	Zinc-Finger		Walker A	Walker B	Arginine Finger
NarMCM	FE..CSQCG	.KCE.CG	SNPGSGK	DEFD	DRFH
Mka1MCM	AE..CPQCG	VRCPNCG	GDPATAK	DELD	SRFD
NmaMCM	FV..CPDEH	CDNPNCK	GDPGTAK	DEFD	TRFD
NeqMCM	FRHSCGYEV	.KCPKCN	GDPGTAK	DEIE	NRFD
MacMCM	FV..CTFCN	KVCPECK	GDPGSAK	DEFD	SRFD
SsoMCM	YKHIHPDCM	.ICPKCG	GDPGTAK	DEID	SRFD
ApeMCM	FR..HDRCG	.ICPVCG	GDPGVAK	DEFD	SRFD
MthMCM	FE..CRGCM	.LCSECG	GDPGIGK	DELD	SRFD
AfuMCM	FA..CLNCG	.ECSKCS	GDPGVAK	DEID	SRFD
HvoMCM	FE..CQRCG	.ECQGCE	GDPGTGK	DELD	SRFD
MbaMCM	FE..CLRCG	CEDENCG	GDPGIAK	DEMD	SRFD
MhaMCM	FY..CMRCE	CEEESCG	GDPGIAK	DEMD	SRFD
PfuMCM	FV..CKDCG	KKCEQCG	GDPGVAK	DELD	SRFD
KcrMCM	FI..CTNCG	EKCPECG	GDPGVAK	DEFD	SRFD

Figure 3.3: Sequence alignment of core MCM motifs.
MSA of MCM used in this study. Sequences were aligned using Clustal Omega and visualized using the TexShade package in LaTeX. Amino acids are coloured based on the chemical properties of the conserved functional side chain group. The conserved glutamate (E) residue in the Walker B motif is mutated to glutamine (Q) for *Mac*MCM^{E418Q}. Brown: aromatic, Yellow: sulfur, Orange: imino, Grey: aliphatic (small), Red: acidic, Blue: basic, Black: aliphatic, Purple: hydroxyl.

Importantly, it appears that all MCM contain the key conserved motifs required for ATP binding and hydrolysis and are therefore expected to be active. All enzymes are also expected to coordinate a functional zinc ion through either a C_4 or C_3H-type zinc finger [115] (Figure 3.3). Variations or loop insertions are not uncommon for MCM zinc-fingers. Due to its strong conservation and likely intolerance to mutations, the C-terminal domains were then used to predict the phylogeny of the selected MCM sequences versus human MCM. Interestingly, all previously studied MCM are located very closely within the tree, suggesting our new MCM are of reasonable divergence (Figure 3.4). Furthermore, at least 6 of the new MCMs appear to be more closely related to the eukaryotic MCM subunits than previously studied MCM. Eukaryotic MCMs form a subgroup as expected.

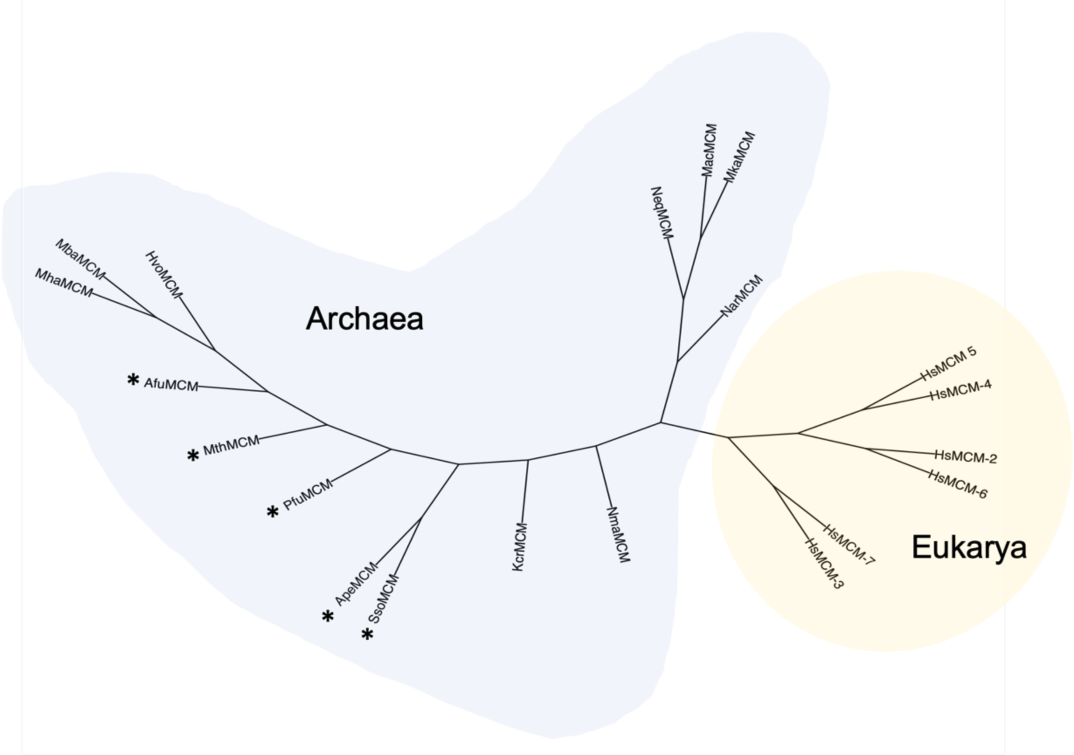


Figure 3.4: Phylogenetic analysis of MCM C-terminal domains.
The evolutionary relationship of MCM C-terminal domains were analyzed using Clustal Omega. An unrooted phylogenetic tree was visualized using the ‘ape’ package within R. Asterixis denote MCM which have been extensively biochemically studied previously.

3.4 Preliminary molecular biology of MCM orthologues

3.4.1 Protein expression and solubility tests

Selected MCM genes were synthetically produced with a removable His$_{10}$-affinity tag and cloned into appropriate *E. coli* expression vectors. Recombinant MCM was then overproduced. The parameters of each recombinant His$_{10}$-labelled MCM can be seen in Table 3.1. Protein expression and solubility was assessed using gel electrophoresis (see methods 2.2.7). Fourteen of the MCM were expressed strongly and at least eleven have a clear signal for the related recombinant MCM in the soluble fraction (Table 1, Figure 3.5a-d). Since the recombinant MCM contains a His$_{10}$ tag, a fluorescent Ni-NTA conjugate was used to confirm the identity of bands. In general, MCMs from thermophilic organisms appear to be better expressed and more soluble than their mesophilic counterparts. As the sample requirements for our initial downstream assays is low, all samples at this stage were taken forward for further analysis. Furthermore, the band signals of any lowly expressed protein may be hidden in the signal of native *E. coli* protein bands. Additional purification could resolve the bands. Although it was not attempted here, optimization of *E. coli* cell-lines, expression and lysis conditions may yield better expression and solubility of recombinant MCMs.

Table 3.1: Parameters of recombinant His$_{10}$-tagged MCM constructs.
MW: Molecular weight. pI: Isoelectric point. **ε**: extinction coefficient. Parameters were calculated using ExPASY ProtParam (https://web.expasy.org/protparam/)

		+His$_{10}$			
Host	**Identifier**	**Length**	**MW (kDa)**	**pI**	**ε (M^{-1} .cm^{-1})**
A. fulgidus	*Afu*MCM	725	81.77	5.93	46760
A. pernix	*Ape*MCM	724	81.46	6.00	55810
H. volcanii	*Hvo*MCM	729	81.84	4.73	45730
K. cryptophilum	*Kcr*MCM	730	82.36	5.60	63260
M. barkeri	*Mba*MCM	727	81.89	5.50	43320
M. halophilus	*Mha*MCM	723	81.14	5.59	51800
M. acidiphilum	*Mac*MCM	714	79.11	5.41	46300
M. kandleri	*Mka*MCM	683	77.05	5.15	45380
M. thermautotrophicus	*Mth*MCM	693	78.54	5.49	33810
N. Archaeon SG9	*Nac*MCM	702	79.06	5.08	52260
N. equitans	*Neq*MCM	684	77.10	5.97	49280
N. maritimus	*Nma*MCM	722	81.12	5.87	29800
P. furiosus	*Pfu*MCM	708	79.70	5.88	37820
Engineered hybrid	*PfuSso*MCM	616	69.09	5.65	38850
Engineered hybrid	*SsoPfu*MCM	640	72.36	6.50	46300
S. solfataricus	*Sso*MCM	713	80.40	6.28	54780

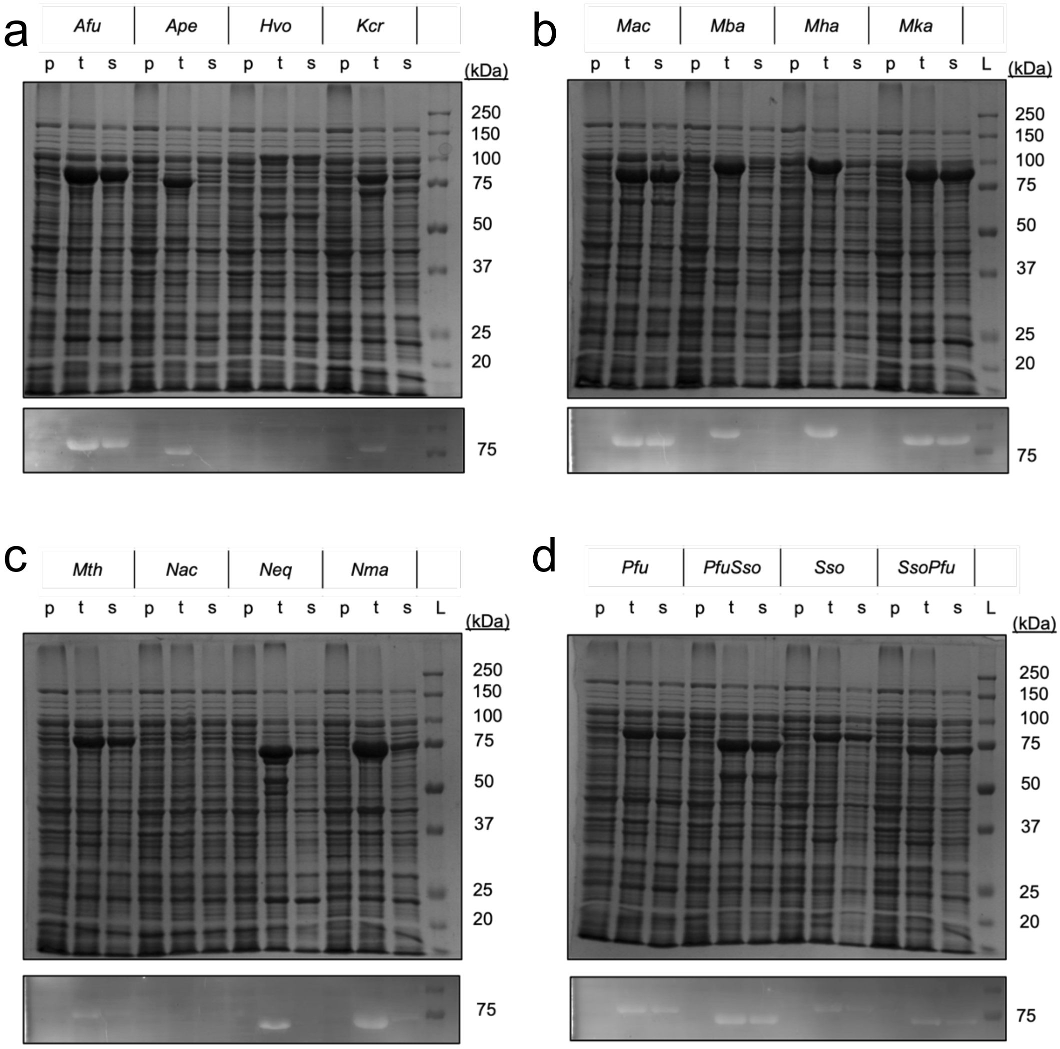
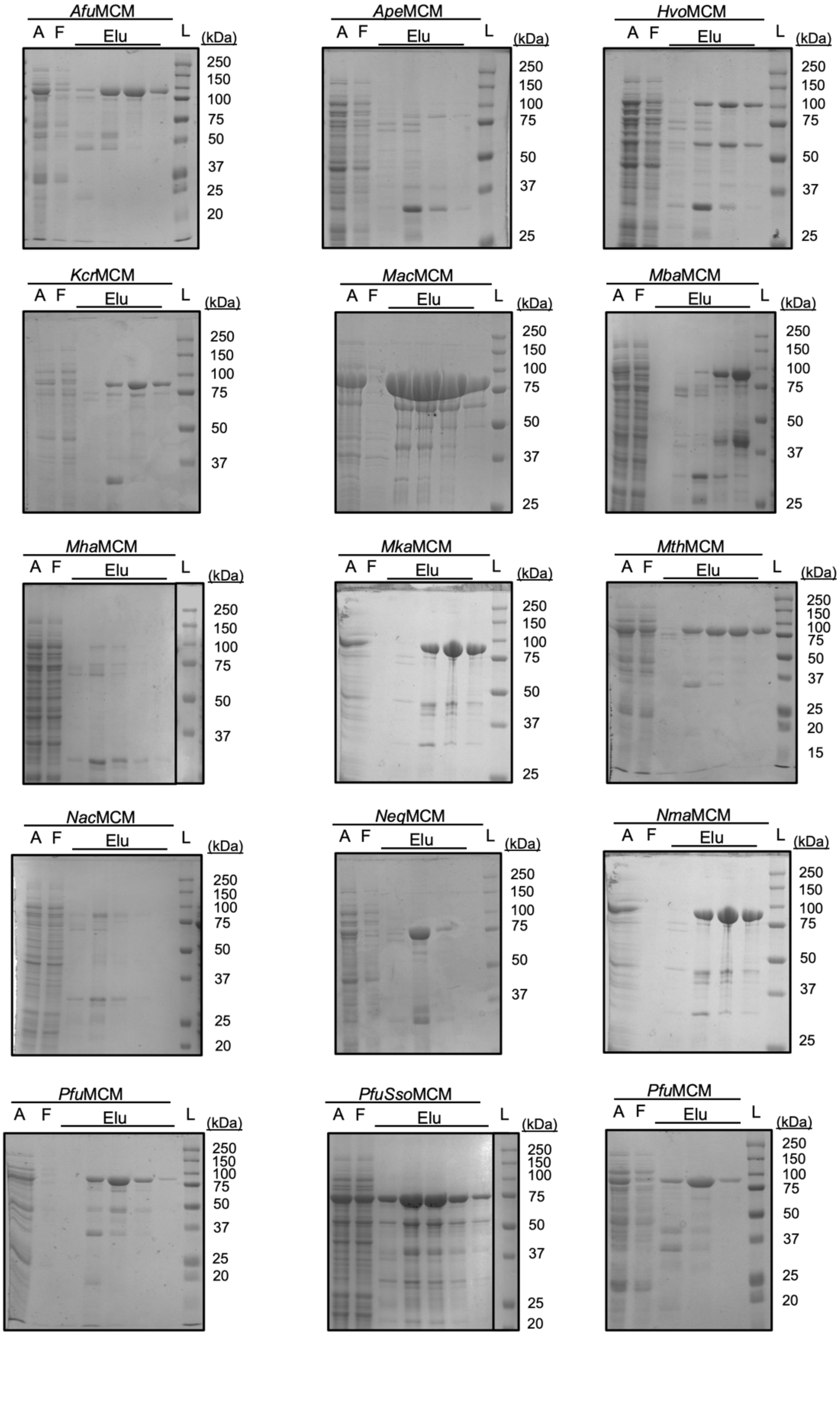


Figure 3.5: Test for expression and solubility of 16 archaeal MCM orthologues.
(a-d) All samples were analysed on a 12 % (*w/v*) SDS-PAGE. Samples represent pre-induction of expression (p), then total (t) and soluble (s) fractions which are collected 20-hours after the IPTG-induction at 20 °C. Top panels represent gels where proteins are staining through a Coomassie-based approach. Bottom panels represent gels where proteins are stained using a fluorescent Ni-NTA conjugate that test for the presence of poly-histidine tags. Reference (L) lanes represent MW standard marker (Precision Plus Protein™ All Blue Pre-Stained Protein Standards).

3.4.2 Crude purification of His$_{10}$-labelled MCM

Immobilised metal affinity chromatography (IMAC) was used to purify recombinant His$_{10}$ labelled MCMs from the soluble cell extract. To remove potentially protein-bound DNA, immobilised proteins were washed on the nickel resin using an isocratic wash of high salt buffer containing a NaCl concentration of 2 M. Bound proteins were then eluted using an isocratic elution with a constant concentration of buffer B, containing an imidazole concentration of 500 mM. For all constructs, some protein was present at the expected molecular weight (Figure 3.6). A range of purity and sample quality was observed in the elution between MCM constructs. For example, *Mth*MCM, looks >90% pure following a single affinity column, whilst *PfuSso*MCM appears less than 80% pure. It is possible that contaminants observed here are either host *E. coli* proteins, degradation products or incompletely translated MCMs[251,252]. A large proportion of contaminants are <50 kDa. Assuming they do not form tight complexes, contaminants may be further separated during the spin concentration step that utilises a 50 kDa MWCO filter.



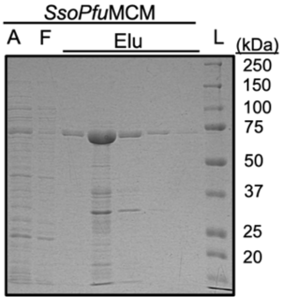


Figure 3.6: Purification of His$_{10}$-labelled recombinant MCM.
Fractions from IMAC purification were analyzed by SDS-PAGE on 12 % (*w/v*) polyacrylamide gels. Fractions include the sample applied to the Ni-NTA column (A), the column flow through (F) and select elution's (Elu). Reference (L) lanes represent MW standard marker (Precision Plus Protein™ All Blue Pre-Stained Protein Standards).

Fractions were selected based on SDS-PAGE analysis and dialysed to remove imidazole (Figure 3.6). Proteins were subsequently concentrated to ideally ≥10 µM using a spin concentrator (MWCO 50 kDa, Merck), then stored at - 80°C. Final purity was assessed further by SDS-PAGE (Figure 3.7). All but 2 samples have protein at the expected molecular weight. Differences in the main band intensities are likely caused by inaccuracies in absorbance-based measurements. For example, His-tagged protein truncations may still contain tryptophan residues that can make measurements inaccurate. Equally, common co-contaminants such as DNA and chaperones may also pollute the absorbance signal. Many of these issues could be resolved with further purification steps. The estimated yield of ‘crude’ preps is displayed in Table 3.2. Yields for *Mha*MCM and *Nac*MCM were insufficient

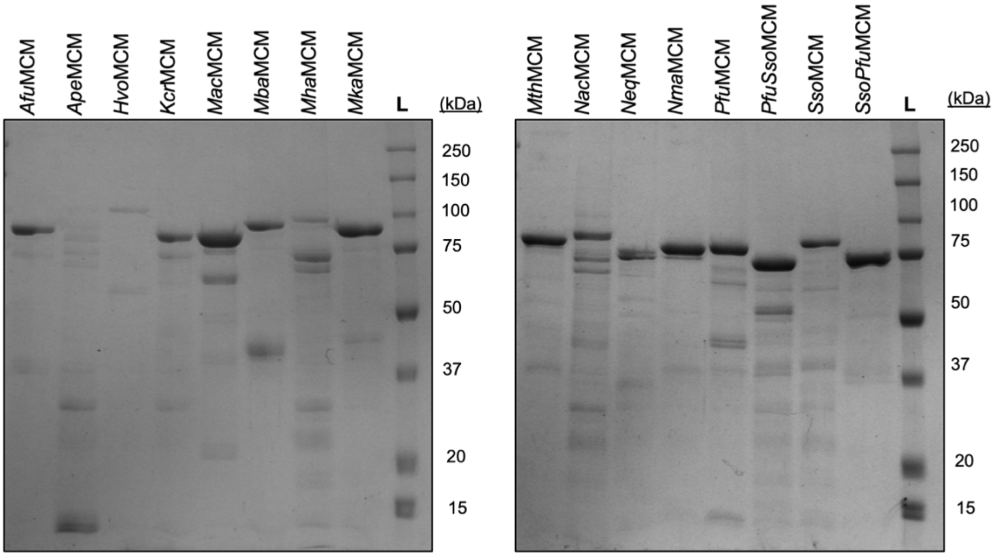


Figure 3.7: Purity analysis of His$_{10}$-labelled recombinant MCM.
The homogeneity of purified MCM was assessed through SDS-PAGE analysis. An estimated 3 µg of each purified MCM was run on a 12 % (*w/v*) poly acrylamide gel. Gels were stained using a Coomassie-based dye. Reference (L) lanes represent MW standard marker (Precision Plus Protein™ All Blue Pre-Stained Protein Standards).

for downstream biochemical analyses. Since MCMs are DNA binding proteins, A_{260}/A_{280} ratios were measured to assess potential DNA contamination (Table 3.2). Both *Mac*MCM and *SsoPfu*MCM exhibit the smallest amount of DNA contamination, where a ratio of 0.6 is assumed to be ~0 % DNA contamination[253]. All samples contain <10 % DNA (ratio = 1.3) contamination, however it is difficult to ascertain the effect of contamination particularly when the size of DNA molecules is unknown.

Table 3.2: Final yields of recombinant His_{10}-tagged MCM samples.

Identifier	Raw Yield (mg)	Adjusted Yield ($mg.L^{-1}$)	$A_{260/280}$
*Afu*MCM	4.14	20.7	0.83
*Ape*MCM	0.99	4.95	1.01
*Hvo*MCM	1.17	5.85	1.03
*Kcr*MCM	3.46	17.3	0.95
*Mba*MCM	1.20	6.00	0.92
*Mha*MCM	0.09	0.45	0.87
*Mac*MCM	18.4	91.8	0.63
*Mka*MCM	3.15	15.8	0.76
*Mth*MCM	5.65	28.3	0.82
*Nac*MCM	0.05	0.27	1.24
*Neq*MCM	1.54	7.70	0.76
*Nma*MCM	3.43	17.2	0.86
*Pfu*MCM	2.30	11.5	0.82
Pfu_NSso_CMCM	2.39	11.9	0.71
Sso_NPfu_CMCM	2.25	11.3	0.63
*Sso*MCM	2.80	14.0	0.88

3.4.3 Assessment of ATPase activity of ‘crude’ MCM samples

MCM helicases must be able to hydrolyse ATP to generate translocation along a DNA substrate. Whilst some basal ATPase activity is expected, MCM have frequently been shown to exhibit DNA-stimulated activity in both archaea and eukaryotes [79,90,99,116,132]. Here, we tested the ATPase activity of crude MCM purifications through use of a malachite-based absorbance assay to detect the production of inorganic phosphate after ATP hydrolysis. This was carried out in the presence or absence of a forked DNA substrate (see methods 2.4.2)

Samples were incubated in the presence of ATP for 5 minutes at 25 °C. To minimize intrinsic differences in DNA binding capacity, a low salt buffer was used. MCM-DNA interactions are likely mediated through electrostatic interactions that a high salt concentration may inhibit[101]. After 5 minutes, the presence of inorganic phosphate was detected using malachite dye and absorbance readings were taken at 611 nm. Absorbance values were then quantified using an inorganic phosphate standardization curve.

All native MCMs exhibit comparable low levels of ATP hydrolysis that is stimulated by addition of DNA (Figure 3.8). Both engineered chimera enzymes exhibit elevated basal (–DNA) ATPase activities relative to the +DNA. This is expected since they lack the regulatory winged helix domain. Removal of the winged helix domain is speculated to increase basal and DNA-stimulated ATP hydrolysis[83]. The comparable hydrolysis levels of the native orthologues may be indicative of the high conservation levels of the ATPase fold (mean identity = 47.9 %; Figure 3.2b). To investigate temperature dependency of ATP hydrolysis, 4 thermophilic MCM were assessed for ATPase activity at 50 °C (Figure 3.8). It is speculated that non-thermostable MCM would precipitate at high temperature and thus may interfere with absorbance readings. In all instances, increasing the temperature to 50 °C, resulted in an approximate 2-fold increase in ATPase activity.

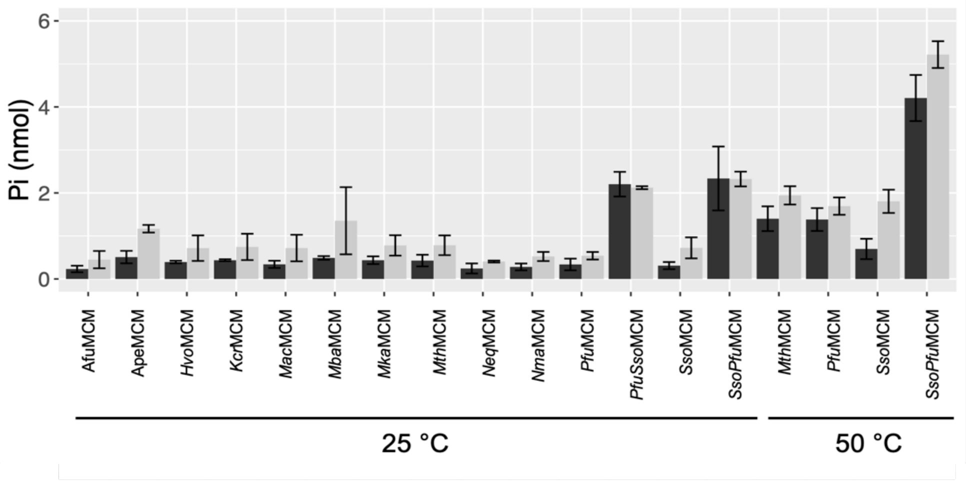


Figure 3.8: ATPase hydrolysis of recombinant His$_{10}$-tagged MCM samples.
Hydrolysis of ATP by MCM was determined through a 96-well plate-based Malachite green assay (AbCam). Reactions were performed at both 25 °C and 50 °C for the stated MCMs, in the presence (grey) or absence (black) of DNA. Production inorganic phosphate was determined through absorbance of malachite green dye at 611 nm. Reactions were performed in absence and presence of a forked DNA substrate added at an equimolar ratio. Error bars represent +/- 1 standard error of the mean, where n = 3. Biological repeats were collected in collaboration with Mike Hodgkinson (York).

3.4.4 Assessment of DNA unwinding activity by MCM samples

Whilst ATP hydrolysis is limited by the AAA+-domain and an MCMs ability to interact with DNA, DNA unwinding is a cumulation of ATP hydrolysis coupled to conformational change that directs efficient DNA binding and unbinding. Studies on the heterohexameric CMG complex are typically performed in 2-300 mM potassium salt, where the anion is typically an organic acid, such as glutamate or acetate[63]. Therefore, we decided to examine the activity of the MCMs when tested against 250 mM KGlutamate at 25 °C.

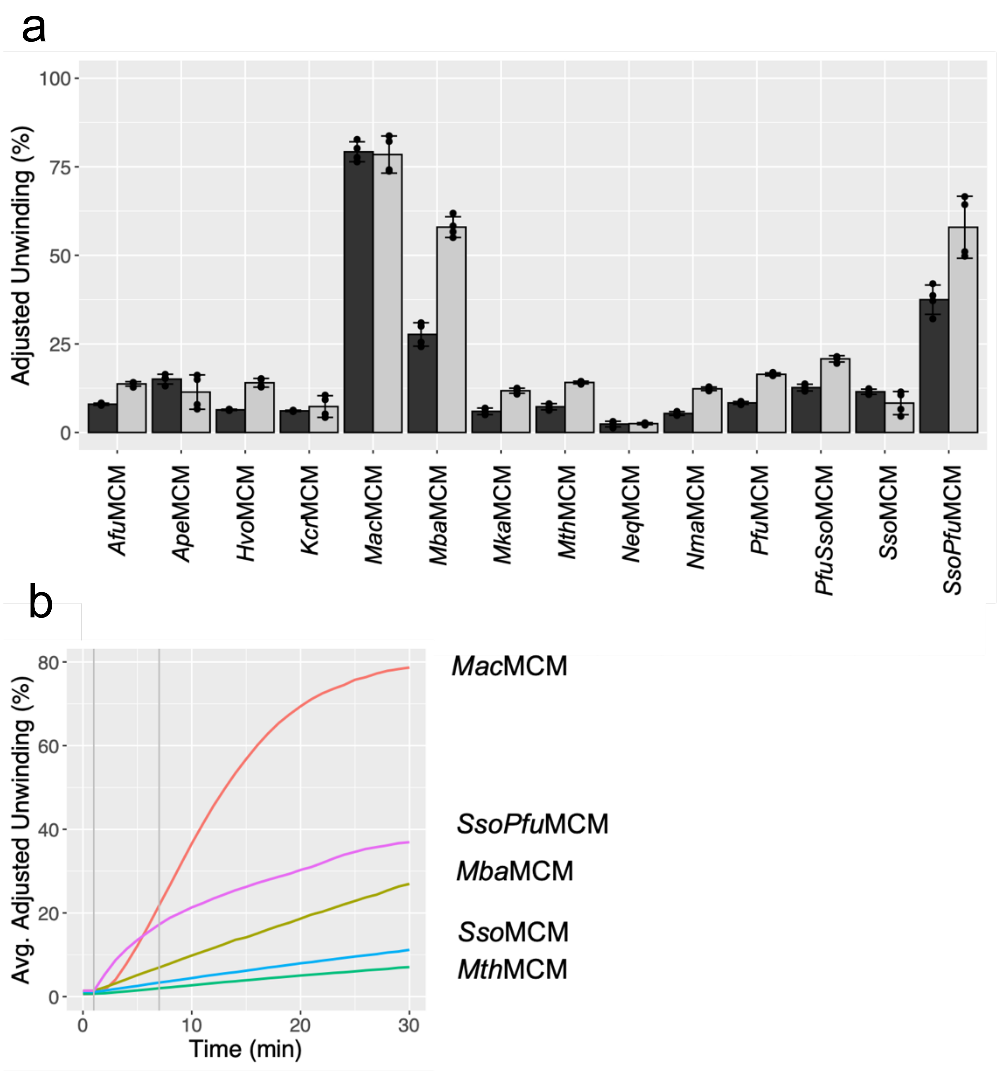


Figure 3.9: DNA unwinding by His$_{10}$-tagged MCM samples.
(a) DNA turnover of a forked DNA substrate was determined through a 96-well plate-based fluorescent helicase assay. Samples were incubated at either 25 °C (black) or 45 °C (light grey) After 30 minutes, ATP/Mg^{2+} was added to a final concentration of 4/10 mM. DNA turnover was monitored through changes in fluorescence over 30 minutes on a Clariostar plate reader (BMG Labtech). Data points were standardized to a maximum fluorescence control substrate, where the fluorescent Cy3 strand is not annealed to the BHQ2 quencher strand. Data points were then adjusted to 50 nM DNA unwinding per 1000 nM hexamer. This was to account for *Mba*MCM, for which protein concentration was an issue and therefore samples contained half the concentration of other experimental wells. Error bars represent +/- 1 standard error of the mean, where n = 4. **(b)** Exemplary real time data traces for 5 MCMs measured at 25 °C. The orange vertical line represents the point of maximum velocity (8 minutes) for the sigmoidal unwinding kinetics of *Mac*MCM. All other MCM reach maximum velocity near instantly after ATP is added.

Generally speaking, all MCMs exhibited at least a small degree of unwinding under the assay conditions (Figure 3.9a). Two of the MCMs, *SsoPfu*MCM and *Mac*MCM unwind >40% of the substrate. Unusually, the most active enzyme in the screen exhibits an atypical sigmoidal unwinding profile, rather than an expected hyperbolic profile (Figure 3.9b). From here on, we refer to these sigmoidal kinetics as 'lag time', which is equal to the time taken to reach maximum velocity. All other MCMs appear to reach maximum velocity almost instantly. Increasing the temperature to 45 °C results in a clear increase in DNA unwinding activity for all but 4 MCM (Figure 3.9a). *Mba*MCM exhibits the largest increase in net DNA unwinding amongst the MCM that correlates well with the environmental range of the organism (30 - 50 °C)[247].

Based on its exceptional yield, high activity, and unusual kinetics, *Mac*MCM was taken forward for further study. As *SsoPfu*MCM unwinding is the most comparable in terms of activity, we elected to use this enzyme as a contrast point for 'traditional' archaeal MCM. At 25 °C, *SsoPfu*MCM exhibits the same kinetic profile and activity as the traditionally studied thermophilic *Mth*MCM enzyme measured at 60 °C (Mike Hodgkinson, personal communication). Understanding why the majority of MCMs exhibit low *in vitro* activity, is only possible by deconvoluting a quagmire of factors such as protein stability, DNA binding, oligomeric state, and protein flexibility.

3.5 Purification of 'pure' MCM samples

Since the fusion protein *SsoPfu*MCM lacks the presence of a regulatory winged-helix domain, a truncation mutant of *Mac*MCM (*Mac*MCM$^{\Delta WHD}$) was also designed. Sequence alignment and structural analysis was used to identify a conserved domain boundary flanking the AAA+-WHD linker. A premature double stop codon was incorporated into the *Mac*MCM^{FL} (full-length) sequence, and after expression should produce an enzyme 87 amino acids shorter than *Mac*MCM^{FL}.

Techniques for analysing proteins biophysically regularly demand high yields and purity. Therefore, 'crude' samples required further purification from larger scale *E. coli* cultures. To this end, MCMs were overproduced in 1 L cultures. To further improve the IMAC procedure, the isocratic imidazole elution was replaced with an imidazole gradient, starting at 20 mM (0 %) and finishing at 500 mM (100 %) (Figure 3.10a) (see methods 2.2.9). Differential binding capacities of proteins to the resin increased separation of the MCMs from contaminants when assessed by SDS PAGE (Figure 3.10b).

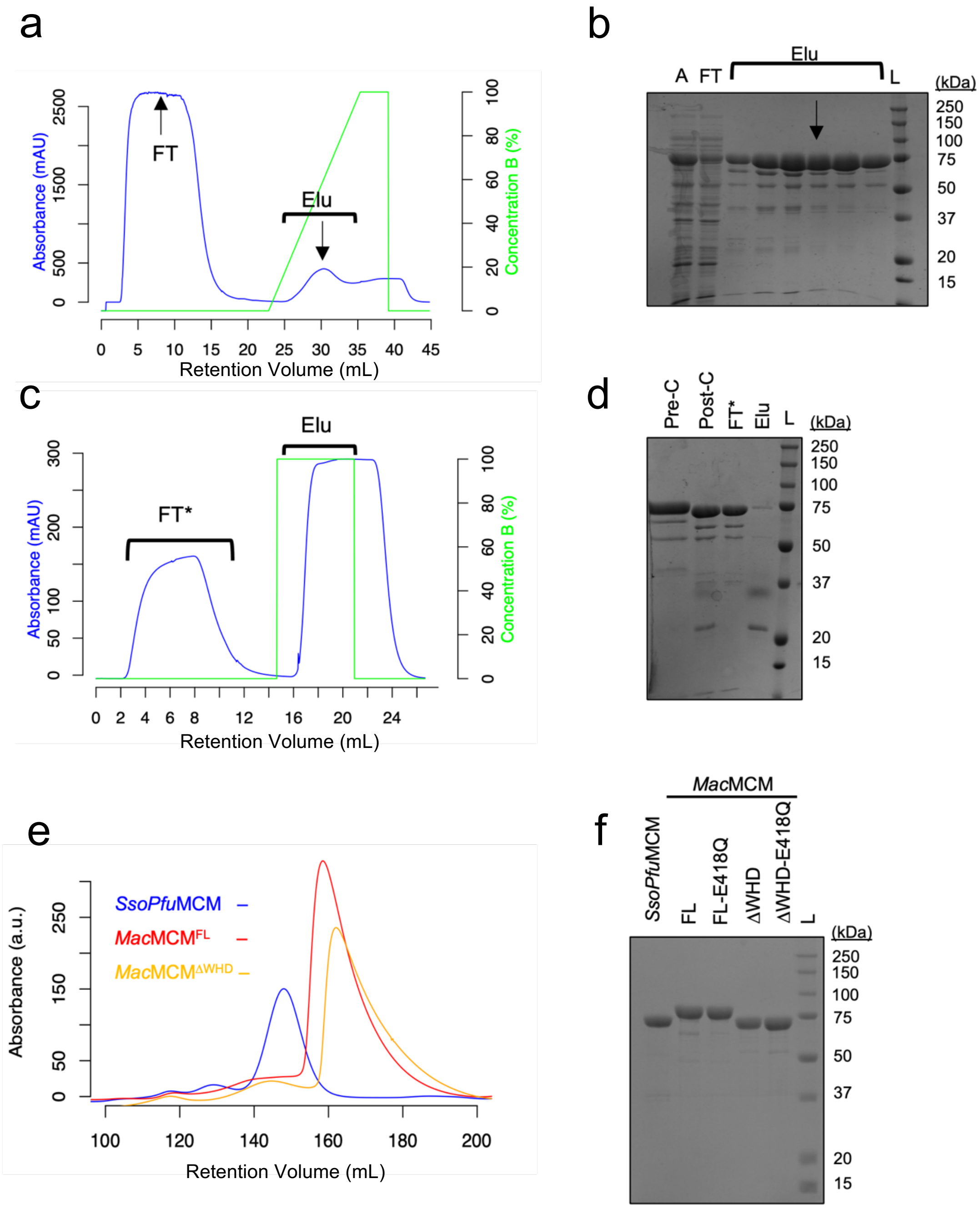


Figure 3.10: Routine purification strategy of 'pure' MCM samples.

All examples in a–d here are for *Mac*MCM$^{\Delta WHD}$. Purification fractions include, Applied (A), Flow through (FT), Elution (Elu), Pre-TEV cleavage (Pre-C), Post-TEV cleavage (Post-C), Gels were stained using a Coomassie-based dye. Reference (L) lanes represent MW standard marker (Precision Plus Protein™ All Blue Pre-Stained Protein Standards) **(a)** Chromatogram of an IMAC purification, where protein is monitored at A_{280} (blue line). Protein is eluted using an increasing concentration gradient of imidazole up to 500 mM (green line). Fractions were eluted into 4 mL aliquots. **(b)** SDS-PAGE analysis of an IMAC purification of His_{10}-MCM. **(c)** Chromatogram of IMAC separation of MCM from His_{10} conjugates after tag cleavage. **(d)** SDS-PAGE analysis of TEV protease cleavage of His_{10}-MCM and subsequent separation by IMAC. **(e)** Exemplary SEC chromatograms for the core enzymes used in this thesis. Samples were separated on a Hi-Load 26/600 S200 gel filtration column (Cytiva). Elution is monitored at A_{280}. **(f)** Purity analysis of purified recombinant proteins used in this chapter. SDS-PAGE analysis of 3 µg each purified MCM.. FL: Full-length, WHD: Winged-helix domain.

Subsequently, the His_{10} affinity tag was cleaved from MCMs using the TEV protease at a mass ratio of 1:50 (TEV: His_{10}-MCM). Tag cleavage was allowed to progress for 18 hours at 4 °C at which point successful TEV cleavage was observed for all MCM constructs (Figure 3.11d). MCM were then separated from the His_{10}-tag and His_6-TEV sample using a second Ni-IMAC column where MCM is collected in the flow-through. The unwanted His-labelled conjugates are observed following an isocratic elution of 500 mM Imidazole (Figure 3.10c-d).

MCMs were then further purified through gel filtration on a HiLoad 26/600 Superdex S200 (GE Healthcare) column. *SsoPfu*MCM elutes at 148 mL, whilst MacMCM^{FL} despite being ~8 kDa larger elutes from the column at 159 mL (Figure 3.10e). Given archaeal MCMs are usually hexameric in solution, it suggests that *Mac*MCM oligomeric state is sub-hexameric or is exhibiting a potential column dependent interaction. *Mac*MCM$^{\Delta WHD}$ is retained further, eluting at a volume of 163 mL (Figure 3.10e). Purity and molecular weights of MCM constructs used in this chapter were confirmed by SDS-PAGE (Figure 3.10f). Absence of DNA contamination was confirmed through measurement of A_{280}/A_{260} ratios (Table 3.3).

Table 3.3: Final yield of 'pure' recombinant MCM samples
pI: Isoelectric point. **ε**: extinction coefficient. Parameters were calculated using ExPASY ProtParam (https://web.expasy.org/protparam/)

MCM	Yield (mg. L^{-1})	pI	ε (M^{-1} . cm^{-1})	A_{260}/A_{280}
*SsoPfu*MCM	12	6.32	44810	0.55
*Mac*MCM^{FL}	38	5.12	44810	0.57
*Mac*MCM$^{E418Q.FL}$	45	5.15	44810	0.63
*Mac*MCM$^{\Delta WHD}$	31	5.19	43320	0.63
*Mac*MCM$^{E418Q.WHD}$	35	5.23	43320	0.63

3.6 Investigations of the oligomeric state of *Mac*MCM and *SsoPfu*MCM

3.6.1 SEC-MALLS

Archaeal MCM have been identified a range of oligomeric complexes, including hexamers[82,235], heptamers[235], octamers[106], dodecamers[79,120,235] and filaments[107]. However, like many other replicative helicases, such as DnaB[41] and E1[254], MCM proteins are widely regarded to function as toroidal hexamers[104] . Eukaryotic MCM2-7 generally form transient complexes, that range from a collection of monomers to a heterohexamer[58,60]. To confirm that the unexpected late elution volume of *Mac*MCM variants are not column related artefacts, the oligomeric state of MCM apoenzymes was assessed using SEC-MALLS (see methods 2.6.3).

MCM constructs were passed over a Superose 6 Increase 10/300 GL gel filtration column equilibrated in 200 mM KCl, 50 mM Tris-Cl pH 8.0, 5 % (v/v) Glycerol and 0.5 mM DTT. Two protein concentrations were chosen (1 and 10 mg/mL) to investigate potential protein concentration dependent oligomerization. Refractive index and light scattering

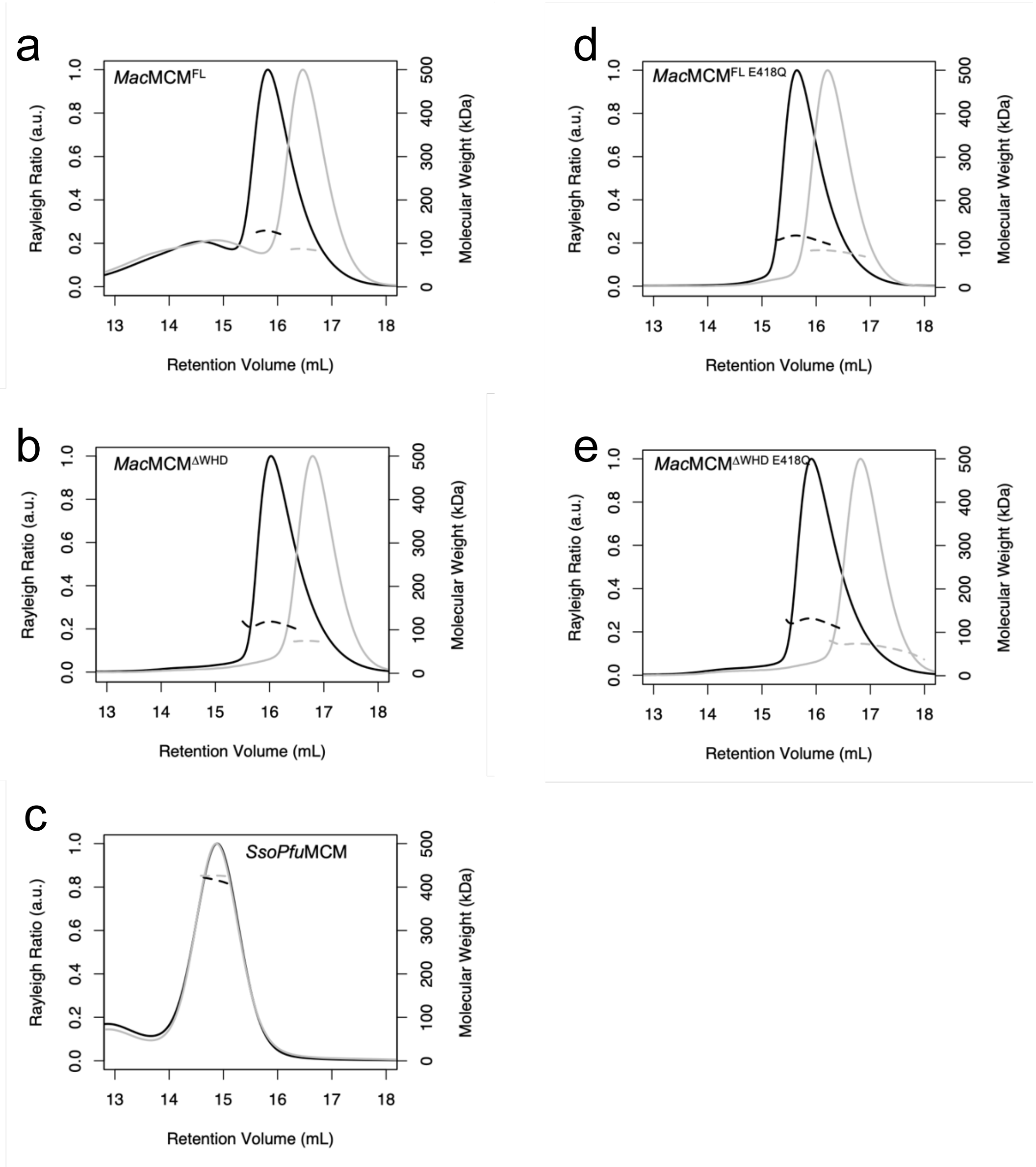


Figure 3.11: SEC-MALLS analysis of recombinant MCM.
(a-e) The elution of 1 (grey) and 10 (black) mg/ml of the stated MCM from a Superose 6 Increase Column was monitored through light scattering (Rayleigh ratio). The Rayleigh ratio was normalised to the height of the main peak. Molar mass estimates (kDa) of the eluate are shown as a dotted line. Experiment was performed in 200 mM KCl, 50 mM Tris pH 8.0, 5 % (v/v) Glycerol and 0.5 mM DTT.

measurements of the elution were used to calculate the molecular weight. Data are plotted in Figure 3.11 and summarised in Table 3.4.

Table 3.4: Molecular weight and size of MCM as calculated by SEC-MALLS and AUC

Protein	**gL^{-1}**	**Da**		**S**
	SEC-MALLS (AUC)	**Theoretical:** Monomer (Hexamer)	**Molar mass moments**	**AUC**
Sso$_N$*Pfu*$_C$MCM	10 (3)	69,429 (416,574)	414,400	10.6
	1 (0.3)		425,500	10.8
MacMCM	10 (3)	76,183 (457,098)	126,500	5.3
	1 (0.3)		86,500	3.9
MacMCM E418Q	10	76,182 (457,092)	113,500	-
	1		84,940	-
MacMCM ∆WHD	10	66,369 (398,214)	110,400	-
	1		72,740	-
MacMCM ∆WHD E418Q	10	66,368 (398,208)	122,800	-
	1		72,270	-

At both concentrations, *SsoPfu*MCM eluted earlier than either *Mac*MCM^{FL} or *Mac*MCM$^{\Delta WHD}$ (Figure 3.11). Calculation of the *in vitro* molecular weights revealed that *SsoPfu*MCM exists as a stable hexamer in solution in absence of ligands. Strikingly, *Mac*MCM^{FL} and *Mac*MCM$^{\Delta WHD}$ display a protein-concentration dependent shift; as protein concentration is lowered, the elution volume increases. Both *Mac*MCM constructs elute with molecular weights (MW) that are consistent with monomeric species. The 'higher' estimated molecular weight of the 10 mg/mL sample is consistent with a population weighted average molecular weight (WAMW). A minor peak is observed at ~14-15 mL for *Mac*MCM^{FL}, which may be consistent with a minor population of hexamers.

3.6.2 AUC

To further examine the oligomeric states of *Mac*MCM^{FL} in absence of any column dependent effects, sedimentation velocity AUC was performed which has the benefit of requiring no separation matrix[200] (see methods 2.6.4). Sedimentation velocity is useful for determination of hydrodynamic radius and the shape of particles [255,256]. For best comparison with SEC-MALLS data, lower protein concentrations were run (0.3 and 3 mg/mL) to take into consideration the column dilution effects before samples reach the detector. Sedimentation was carried out at 30,000 rpm over a period of 7 hours and monitored using both UV absorbance and DRI. Partial specific volumes were calculated based on the amino acid composition of the protein. The density and viscosity of the buffer were calculated using Sednterp[257]. Data were fitted to a continuous c(s) distribution using SedFit[203].

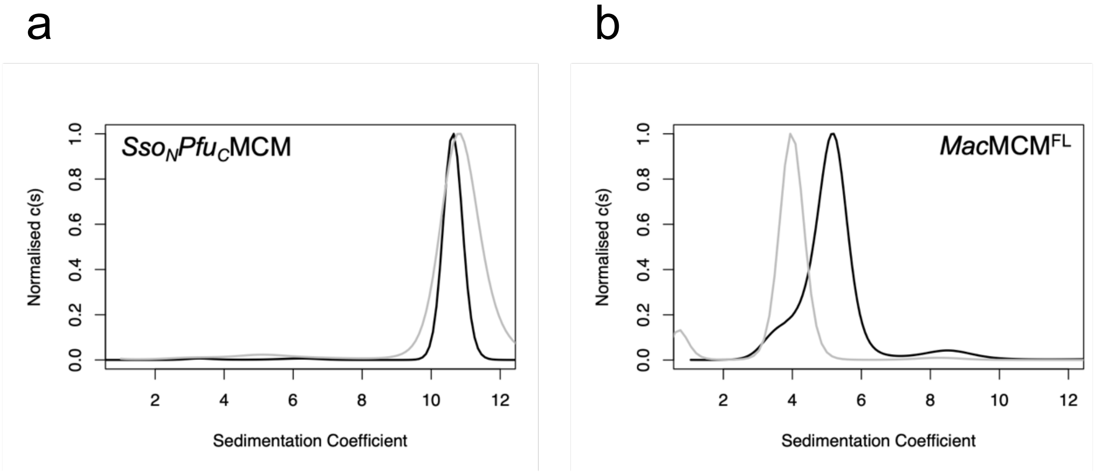


Figure 3.12: Velocity AUC analysis of recombinant MCM.
The sedimentation of 0.3 (grey) and 3 (black) mg/ml MCM was monitored in 200 mM KCl, 50 mM Tris-Cl pH 8.0, 5 % (*w/v*) glycerol and 1 mM DTT. Sedimentation was monitored using UV and DRI detectors whilst centrifuging at 30,000 rpm in an AN50Ti rotor. Analysis was performed in Sednterp and SedFit, where data was fitted to a continuous c(s) distribution. The c(s) distribution was then normalised to the maximum value in each sample. Data were recorded for (**a)** *SsoPfu*MCM and **(b)** *Mac*MCM.

At both concentrations, *SsoPfu*MCM sediments with a coefficient that is consistent with larger sized particles than the *Mac*MCM^{FL} samples (Figure 3.12a-b). In both instances, *SsoPfu*MCM samples sediment as homogeneous single peaks. Consistent with SEC-MALLS data, dilution of *Mac*MCM^{FL} from high to low protein concentrations drives a shift of the population WAMW towards smaller sized oligomeric complexes. This observation is consistent with SEC data and supports the conclusion that *Mac*MCM variants cannot form hexameric complexes in the absence of ligands.

3.6.3 Fluorescent helicase activity measured vs. protein concentration

Eukaryotic MCM exhibits a slow assembly that takes ~5-10 minutes following the addition of ATP and DNA [59,60,157,161]. To date, this observation has never been made for an archaeal MCM. The stability of the eukaryotic MCM2-7 hexamer complex is also influenced by various factors including protein concentration, salt, ATP hydrolysis and temperature[60].

To establish whether these factors also affect the kinetics of *Mac*MCM lag time, a fluorescent helicase assay was then performed, where each MCM was serially diluted. The activity for each protein concentration is then measured independently over 125 minutes. Net unwinding of each MCM was calculated after 30 minutes (Figure 3.13a-c). Over 30 minutes, *Mac*MCM constructs consistently unwind more substrate than *SsoPfu*MCM, and pure MCM samples generally unwind more DNA than crude samples (Figure 3.9a). Furthermore, the *Mac*MCM$^{\Delta WHD}$ enzyme turns over more substrate than the wild-type *Mac*MCM^{FL} at every tested concentration. This is consistent with previous MCM studies, where removal of the winged helix domain improves DNA unwinding[83]. At lower protein concentrations, the activity of *Mac*MCM$^{\Delta WHD}$ does not decline as much as either *SsoPfu*MCM or *Mac*MCM^{FL}.

Alongside net unwinding, lag time was then quantified for each protein concentration. This involves calculating the first derivative of the plot, then extracting the time point where the substrate turnover rate is at a maximum (Figure 3.13d-f). Lag time for each concentration is then plotted as a function of hexamer concentration for each MCM (Figure 3.13g-i). The stable, hexameric *SsoPfu*MCM does not exhibit a relationship between the concentration of the protein and the lag time (Figure 3.13g). *Mac*MCM^{FL} exhibits an exponential relationship between protein concentration and lag time; as protein concentration is decreased, lag time increases exponentially (Figure 3.13h). Considering column dilution effects, the dilution series in this assay are consistent with the dilution series examined in SEC-MALLS and AUC. Taken together, these data suggest that oligomerization propensity is a key determinant of lag time. Removal of the winged-helix domain completely ablates lag time observation under any tested protein concentration (Figure 3.13i). This suggests that either the winged helix domain is the direct cause of lag time, or that the winged helix domain indirectly causes lag time by limiting ATP hydrolysis. It is also plausible that *Mac*MCM$^{\Delta WHD}$ does exhibit a lag time behaviour that is not detectable under the current data sampling rate (1 reading per minute). It is striking that *Mac*MCM^{FL} lag time is on a kinetic time scale that is comparable with eukaryotic MCM2-7 assembly onto DNA.

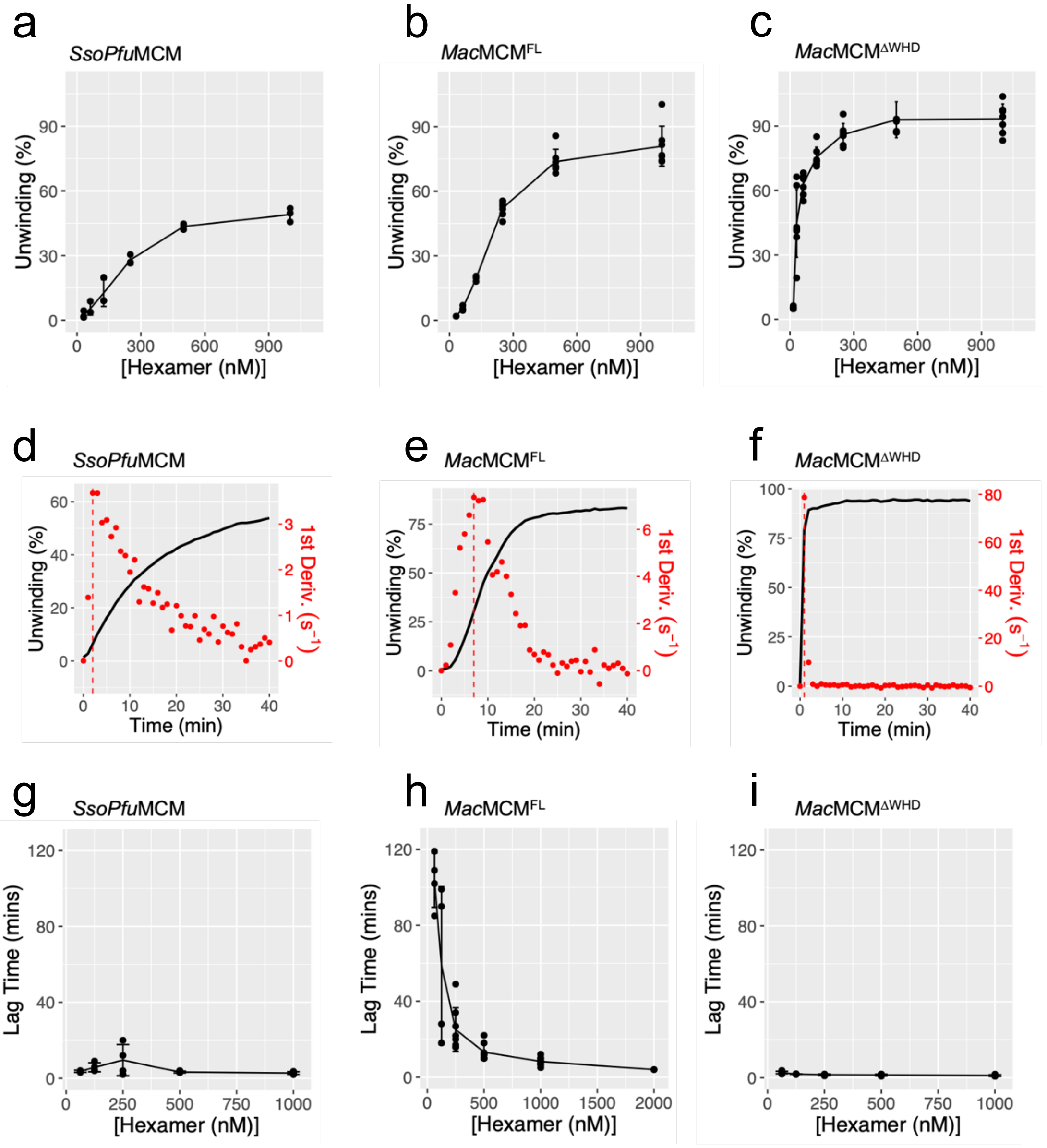


Figure 3.13: Effect of protein concentration on DNA unwinding kinetics.
DNA turnover of a forked 26 base pair DNA substrate was measured in a 96-well plate fluorescent helicase assay. Samples were incubated at 25 °C. After 30 minutes, ATP/Mg^{2+} was added to a final concentration of 4/10 mM. DNA turnover was monitored through changes in fluorescence over 125 minutes on a Clariostar plate reader (BMG Labtech). Data points were standardized to a maximum fluorescence control substrate. **(a-c)** The net DNA turnover was determined for each MCM after 30 minutes for each protein concentration and MCM (Error bar = +/- 1 SEM: n=6). **(d-f)** Exemplary 'lag time' calculation for each MCM measured at 1000 nM Hexamer. The 1st derivative (red points) of each curve is calculated and then time of the maximum 1st derivative value (dotted red line) is extracted from each trace. **(g-i)** Lag time from (d-e) calculations are plotted against hexamer concentration. A line is plotted through the average value for each protein concentration (Error bar = +/- 1 SEM: n=6).

3.7 Investigations into MCM-ATP interactions

3.7.1 Fluorescent helicase activity measured vs. ATP concentration

Loading of the eukaryotic MCM2-7 heterohexamer onto DNA is believed to be dependent on the hydrolysis of ATP [59,60,157]. ATP-dependent ring closure around DNA has been observed for other hexameric helicases, including gp41[258] and E1[254]. However, this observation has not been made for an archaeal MCM. To investigate further whether lag time is influenced by ATP binding and hydrolysis, MCM unwinding, and lag time were measured against various concentrations of ATP.

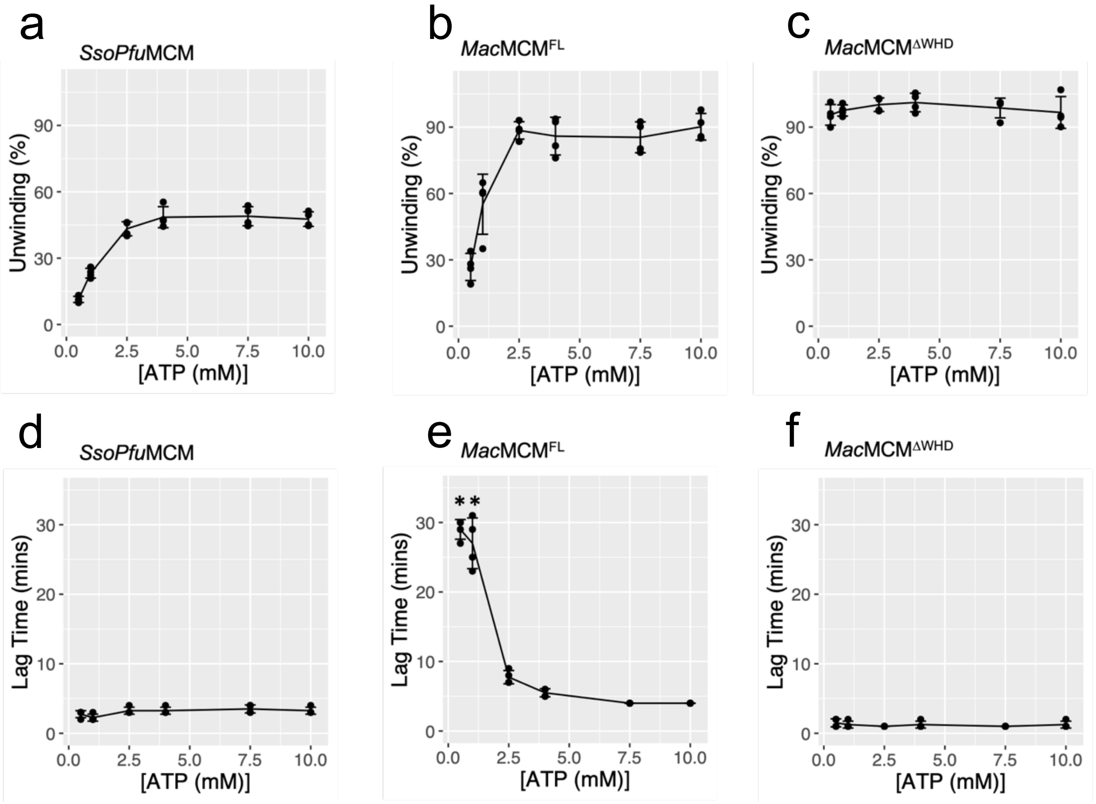


Figure 3.14: Effect of ATP concentration on DNA unwinding kinetics.
DNA turnover of a forked 26 base pair DNA substrate was measured in a 96-well plate fluorescent helicase assay. Samples were incubated at 25 °C. After 30 minutes, ATP/Mg^{2+} was added to a final concentration with the stated concentration of ATP. DNA turnover was monitored through changes in fluorescence over 30 minutes on a Clariostar plate reader (BMG Labtech). Data points were standardized to a maximum fluorescence control substrate. **(a-c)** The net DNA turnover was determined after 30 minutes for each MCM at the stated concentration of ATP (Error bar = +/- 1 SEM: n=4). **(d-f)** Extracted lag times are plotted against ATP concentration. A line is plotted through the average value for each ATP concentration (Error bar = +/- 1 SEM: n=4). The * represent readings for which lag time values are on the boundary of the duration of the time course, and real values may be larger than 30 minutes.

Both *SsoPfu*MCM and *Mac*MCM^{FL} exhibit a plateau in net DNA unwinding capabilities beyond 2.5 mM ATP (Figure 3.14a-b). This activity plateau is consistent with previous work that measured ATP hydrolysis for MCMs in increasing concentrations of nucleotides[81]. *Mac*MCM^{ΔWHD} does not exhibit such a relationship, however, this observation may be obscured by the high activity of *Mac*MCM$^{\Delta WHD}$ at every tested ATP concentration (Figure 3.14c). As before, both *SsoPfu*MCM and *Mac*MCM$^{\Delta WHD}$ do not show a strong relationship between ATP concentration and lag time (Figure 3.14d, f). Comparable with the net unwinding data, *Mac*MCM^{FL} exhibits a plateau in lag time beyond 2.5 mM ATP, whilst decreasing the ATP concentration further increases lag time exponentially. This suggests that ATP is a key factor in the assembly mechanisms of *Mac*MCM^{FL}.

To further determine the biophysical properties that regulate *Mac*MCM behaviour, we need to consider the binding of the enzyme to DNA in absence of translocation and DNA turnover. There are two possible approaches to this: 1) Mutation of key residues in the ATPase active site of MCM; 2) Use of non-hydrolysable ATP analogues.

3.7.2 ATPase inhibition by mutation of catalytic residues

ATP turnover by MCM is a cumulation of ATP binding, followed by activation of a water molecule that hydrolyses the γ–phosphate of ATP. As outlined above, ATP-binding is key for *Mac*MCM assembly, hence, it is important to maintain the affinity of the fold for ATP. Furthermore, there are multiple motifs involved in the correct coordination of a molecule of ATP, meaning several residues may have to be mutated to limit activity. However, only a single glutamate residue of the Walker B motif is required to activate a water molecule that performs hydrolysis of the ATP (see section 1.6.4). The role of glutamate in ATP binding is expected to be limited. Therefore, it was decided to mutate this key residue from a glutamate to a glutamine residue, thereby replacing the reactive sidechain carboxyl group with a non-reactive amide. This mutation was carried out for *Mac*MCM variants only. All mutants were isolated to comparable purity levels, and do not interfere with the oligomeric state of the apoenzyme (Figure 3.10f; Figure 3.11d–e).

3.7.3 ATPase hydrolysis by Walker B motif mutants

Based on our ATP concentration versus helicase data, it is evident that the colorimetric ATPase assays were performed under limiting ATP concentrations (0.05 mM). To identify the ATP turnover rates of the MCMs and confirm that our ATPase mutants were inactive, a ^{31}P-detected NMR experiment was performed [118](see section 2.7.2). This experiment exploits the natural abundance of ^{31}P. Samples were equilibrated to temperature in a 500 MHz NMR spectrometer, then an initial 1D NMR spectrum is recorded for a time zero measurement. MCM were dialysed into the appropriate reaction buffer, then added to the NMR tube at a final concentration 8.3 µM (hexamer). ATP concentrations were used at 50 mM, which is assumed to saturate the MCM. 1D NMR spectra were then recorded every 2

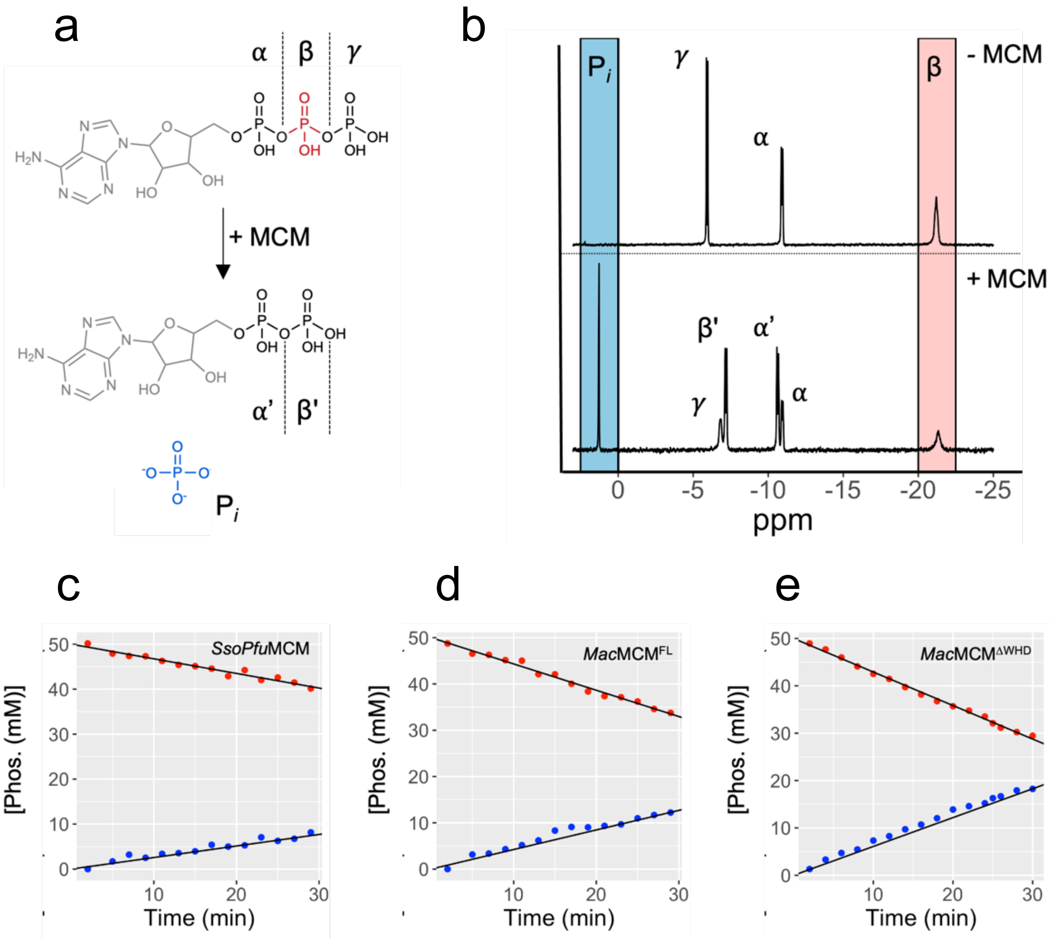


Figure 3.15: Determination of MCM ATP hydrolysis by ^{31}P-NMR
(a) Schematic of ATP hydrolysis by MCM**.** MCM catalyzes the hydrolysis of the γ-phosphate of ATP, resulting in the production of ADP and inorganic phosphate (P_i). **(b)** Experimental ^{31}P-NMR spectra of ATP samples measured on a 500 MHz spectrometer. Top panel represents 50 mM ATP sample without MCM. Bottom panel represents a sample of 50 mM ATP, 30 minutes after the addition of MCM to final concentration 8.3 μM *Mac*MCM$^{\Delta WHD}$ hexamer. Blue region highlights P_i peak, red region highlights β-phosphate peak. Blue and red regions are integrated in TopSpin (Bruker) for regression analysis. **(c-e)** Representative regression analysis for active MCMs in absence of DNA. Red points (β-phosphate), blue points (P_i).

minutes after the addition of ATP for at least 30 minutes. Spectral peaks were assigned according to previous literature values[118]. As ATP is hydrolysed, the population of phosphates in the β–phosphate state is reduced as the β'–phosphate state shifts towards a ɣ–phosphate-like environment (Figure 3.15a–b). A peak is generated down spectra consistent with the production of inorganic phosphate (Figure 3.15b). The β-phosphate and inorganic phosphate peaks are integrated using TopSpin (Bruker). Regression analysis is then performed on the linear region by comparing peak integral versus time. The linear regression is then adjusted using the starting concentrations of ATP and MCM to calculate $K_{cat,app}$ (Figure 3.15c–e)

Table 3.5: ATP hydrolysis rates as determined by ^{31}P-NMR.
K$_{cat.app}$: Apparent catalytic constant. SE: standard error of the fit for K$_{cat.app}$

		(°C)	(Hexamer^{-1}.min^{-1})	
MCM	**Buffer**	**Temp**	**K$_{cat\ app}$**	**SE**
*Sso*MCM	150 mM NaCl	60 °C	7.67	2.08
*Sso$_N$Pfu$_C$*MCM	250 mM KGlu	25 °C	35.76	0.75
*Mac*MCM ^{FL}	250 mM KGlu	25 °C	65.14	2.02
*Mac*MCM ^{FL}	250 mM KGlu + DNA	25 °C	70.13	1.94
*Mac*MCM $^{\Delta WHD}$	250 mM KGlu	25 °C	85.81	1.98
*Mac*MCM $^{FL.E418Q}$	250 mM KGlu	25 °C	0	-
*Mac*MCM $^{\Delta WHD.E418Q}$	250 mM KGlu	25 °C	0	-

An initial standardization experiment was performed using *Sso*MCM, yielding a value in excellent agreement with a previous study (12 hexamer^{-1}.min^{-1})[118]. Consistent with our helicase data, *Mac*MCM^{FL} is ~2-fold more active at ATPase hydrolysis under the conditions than *SsoPfu*MCM (Table 3.5). As observed previously, under highly ATP saturating conditions, MCM do not exhibit a clear increase in DNA-stimulated activity[118]. This is likely due to saturation of the ATPase fold. Similarly, despite *Mac*MCM$^{\Delta WHD}$ exhibiting vastly elevated DNA unwinding activity, the ATPase activity recorded here is very similar between *Mac*MCM$^{\Delta WHD}$ and *Mac*MCM^{FL}. This is likely due to both MCM harbouring identical ATPase folds and ATP concentration no-longer being a limiting factor. As expected, both Walker B E418Q mutants exhibit no ATP hydrolysis over the time course.

3.7.4 Unwinding activity of Walker B motif mutants

Unwinding activity was measured for both MacMCM^{FL} and MacMCM$^{\Delta WHD}$, revealing that both mutations are also catalytically inactive for DNA turnover (Figure 3.16a).

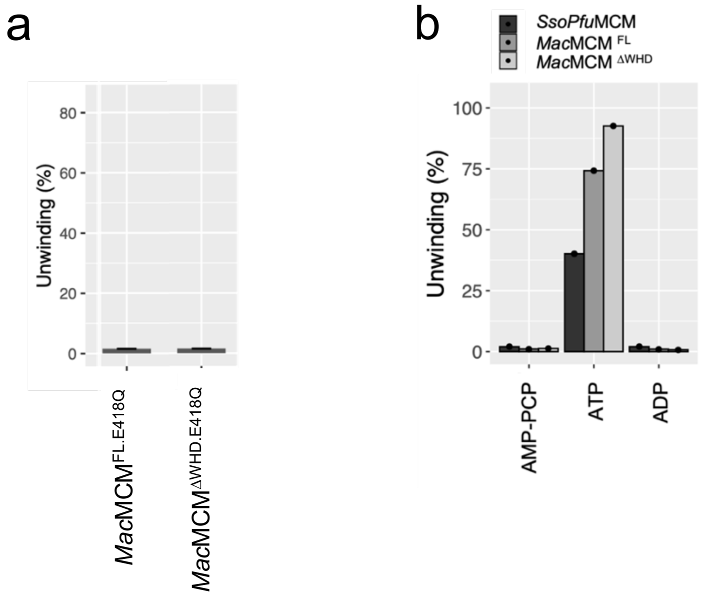


Figure 3.16: Inactivation of MCM activity through ATPase inhibition.
(a) DNA turnover of a forked 26 base pair DNA substrate was measured in a 96-well plate fluorescent helicase assay. Unwinding levels were measured 30 minutes after the addition of ATP/Mg^{2+} at a final concentration of 4/10 mM. **(b)** DNA turnover of a forked 26 base pair DNA substrate was measured in a 96-well plate fluorescent helicase assay. Unwinding levels were measured 30 minutes after the addition of AxP to a final concentration of 4 mM (where AxP is an analogue).

3.7.5 Non-hydrolysable ATP Analogue Fluorescent Helicase Assay

Non-hydrolysable ATP analogues are useful for determining the discrete states of an enzyme during nucleotide hydrolysis. These have successfully been used for understanding the role of ATP-hydrolysis in the conformational change of hexameric helicases [64,259] . In our experiments AMP-PCP is used as a non-hydrolysable ATP analogue, which aims to mimic the pre-hydrolysis ATP-bound state[259]. Other studies in the past have used ATP-γS, however, this analogue is proven to be slowly-hydrolysable[260]. The transient ADP•AlF$_4^-$ is used as a transition state analogue[100]; ADP is used as a post-hydrolysis analogue. In all instances, no analogue other than ATP was capable of eliciting DNA turnover in a fluorescent helicase assay (Figure 3.16b). ADP•AlF$_4^-$ was not tested.

3.8 Investigations of *Mac*MCM DNA loading

3.8.1 EMSA

Electrophoretic mobility shift assay (EMSA) was used to investigate the state of the helicase reaction prior to addition of ATP/Mg^{2+}. A forked, fluorescein labelled helicase substrate was used to determine DNA binding capacity. A forked substrate is chosen, since both the encircled and excluded strands contribute to DNA binding[88]. Tris-Borate buffer (1 x) was chosen without EDTA, to prevent chelation of intrinsic MCM zinc atoms.

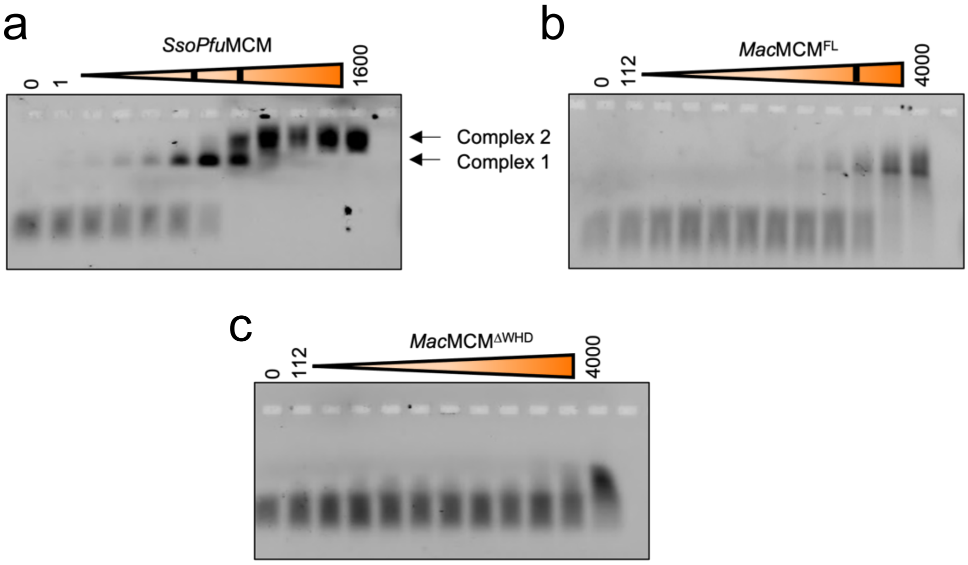


Figure 3.17: Assessment of DNA binding by EMSA.
(a-c) MCM DNA binding was measured by EMSA. The stated MCM was mixed at the stated concentration (µM hexamer) with 10 nM FAM-labelled forked DNA substrate and incubated for 30 minutes at 25 °C. Samples were then resolved on a 1 x TB 0.8 % agarose gel and imaged using a Typhoon gel scanner (GE healthcare).

*SsoPfu*MCM, which forms a hexamer in solution, can bind to forked DNA in absence of ATP/Mg^{2+} with an estimated affinity ~25-50 nM per hexamer. At least two band shifts are observed for *SsoPfu*MCM, which is likely due to the interaction of multiple MCM complexes that the substrate is large enough to accommodate (Figure 3.17a). These multiple band shifts are consistent with EMSA performed in previous MCM studies [101,128]. Comparably, *Mac*MCM^{FL} is not able to bind DNA with high affinity, ~1-2000 nM per hexamer(Figure 3.17b).

Consistent with the lag time data, only a pre-loaded MCM can unwind DNA as soon as it is exposed to ATP/Mg^{2+}. Under the tested conditions, *Mac*MCM$^{\Delta WHD}$ exhibits no DNA binding in the absence of ATP (Figure 3.17c). This suggests that DNA binding by *Mac*MCM^{FL} could in part be mediated by the winged helix domain, in disagreement with previous studies[83].

Trial EMSA with ATP and non-hydrolysable ATP analogues did not generate observable changes in DNA binding affinity (data not shown). It is hypothesized that under EMSA conditions the expected low affinity of MCM for ATP and analogues means that ATP is lost from the complex, thereby causing rapid dissociation of complexes within the gel. Equilibration of buffers and gel in ATP analogues to solve issues associated with low nucleotide affinity issues was unviable.

3.8.2 Fluorescence Anisotropy

Fluorescent anisotropy assays permit observation of DNA binding under matrix-free environments. Furthermore, the volume required in a microplate is ~10-fold less than that required for gel electrophoresis. Samples were serially diluted in buffer on a microplate, and an equal volume of forked DNA was added to final concentration 1 nM. A concentrated stock of ATP analogue is then added to final concentration 4 mM, and the sample is allowed to incubate for 30 minutes. Wells are then recorded on a plate reader (BMG Labtech). Anisotropy values are calculated using Mars software (BMG Labtech) and adjusted by removal of a protein-free control well. Data are then fitted to a Langmuir single-binding isotherm in R using a nonlinear least square method to extract equilibrium dissociation constants (K_d), and Hill (n) values (see equation 2.4; Supplementary Figure 7.5). The observation of multiple complexes in EMSA suggest deviation from 1:1 binding and supports the choice of a Hill coefficient.

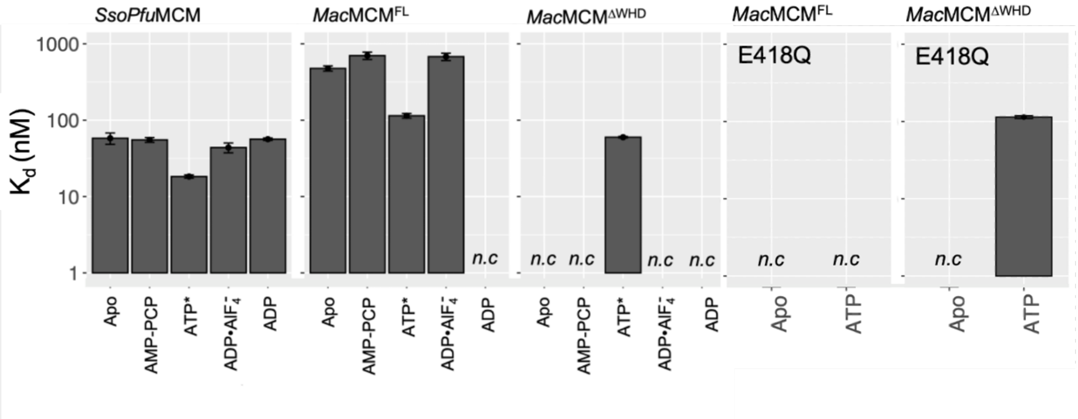


Figure 3.18: Assessment of DNA binding by fluorescence anisotropy.
DNA binding by MCM was measured through fluorescence anisotropy. MCM was mixed with 1 nM FAM-labelled forked DNA substrate and incubated for 30 minutes at 25 °C. Measurements were performed on a Clariostar plate reader (BMG Labtech). Anisotropy values were standardized by removal of a no protein control well, then fit to Langmuir binding isotherm with Hill coefficient. Error bars represent +/-1 standard error of the calculated K_d value. Each model is based on the fit to the average of 3 independent experiments. N.c: not calculated due to insufficient data/quality of fit.

The K_d for apo MCMs are consistent with values observed in EMSA, although it is important to note that there is no dilution occurring here in the microplate. *SsoPfu*MCM shows tight binding to DNA with all tested analogues, where the strongest binding occurs in the presence of ATP (note, these values also include DNA turnover) (Figure 3.18, Table 3.6). The tightest binding is also observed for *Mac*MCM constructs only in the presence of ATP hydrolysis. Non-hydrolysable analogues in all instances reduce DNA binding for *Mac*MCM^{FL}. No non-hydrolysable analogue can induce DNA binding for *Mac*MCM$^{\Delta WHD}$. *Mac*MCM$^{\Delta WHD.E418Q}$ only binds to DNA tightly where ATP is present, while *Mac*MCM$^{FL.E418Q}$ is unable to interact with DNA under any circumstance. This suggests that the winged helix domain inhibits DNA interactions in the absence of ATP hydrolysis.

Table 3.6: Calculated binding affinities based on fluorescence anisotropy.
K_d: equilibrium dissociation constant, SEM: standard error of the mean, *n.c.*: not calculated (experimental data is insufficient for a reasonable fit)

	K_d (nM) ± SEM		
State	***SsoPfu*MCM**	***Mac*MCM^{FL}**	***Mac*MCM$^{\Delta WHD}$**
Apo	58.1 ± 9.81	476 ± 36.6	*n.c.*
AMP-PCP	55.2 ± 3.94	700 ± 77.4	*n.c.*
ATP	18.2 ± 0.90	114 ± 8.11	59.83 ± 1.46
ADP•AlF_4^-	43.8 ± 6.46	678 ± 74.4	*n.c.*
ADP	56.4 ± 2.07	*n.c.*	*n.c.*
Apo-E418Q	-	*n.c.*	*n.c.*
ATP-E418Q	-	114 ± 4.99	*n.c.*

3.8.3 SEC-MALLS

SEC-MALLS was evaluated to determine whether addition of ATP to the loading sample can support formation of a hexamer. The experiment was carried out as described previously, however with the addition of ATP/Mg^{2+} or just ATP in the loading samples. The resulting eluate does not differ between either ATP sample here and the apo sample measured previously (Figure 3.19a–b, Figure 3.12a). It was hypothesized that due to the low expected binding affinity of MCM to ATP, it is required for the entire column to be equilibrated in ATP/Mg^{2+} to prevent dissociation of the complex. Furthermore, as the MCM consumes ATP, the presence of nucleotide in the buffer instead would provide a continuous stream of available ATP.

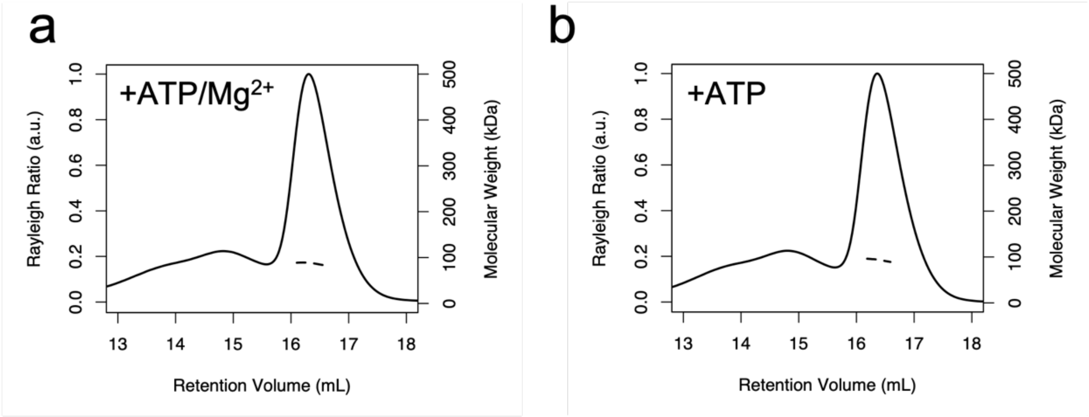


Figure 3.19: SEC-MALLS analysis of *Mac*MCM with ATP in the loading buffer.
(a-b) The elution of 1 mg/ml MCM from a Superose 6 Increase 10/300 GL column was monitored through light scattering (Rayleigh ratio). Sample was pre-equilibrated with ATP/Mg^{2+} (4/10 mM) before loading onto the column. The scattering signal was normalised to the height of the main peak. Molar mass estimates (kDa) of the eluate are shown as a dotted line.

3.8.4 Analytical SEC

Performing analytical SEC in the presence of mM concentrations of ATP is difficult due to the strong absorbance of ATP. ATP has a broad absorption peak, with a maximum absorbance at 260 nm (Figure 3.20a). When the concentration of ATP is increased there is significant overlap with the peak maxima of protein at 280 nm. Mixing protein and ATP in the same cuvette reveals that protein can clearly be distinguished from <2 mM ATP at wavelengths above 290 nm (Figure 3.20b). Many HPLC systems, such as ÄKTA Pure are equipped with multi-wavelength absorbance detectors, allowing observation of other distinct biomolecules. Since DNA also shares the same spectral region as ATP (~260 nm), it was decided to use a fluorescently labelled DNA substrate. To remove DNA turnover, a single stranded substrate of polyT$_{50}$ was selected, with a 5′-6-FAM label. Fluorescein absorbance was monitored at 495 nm.

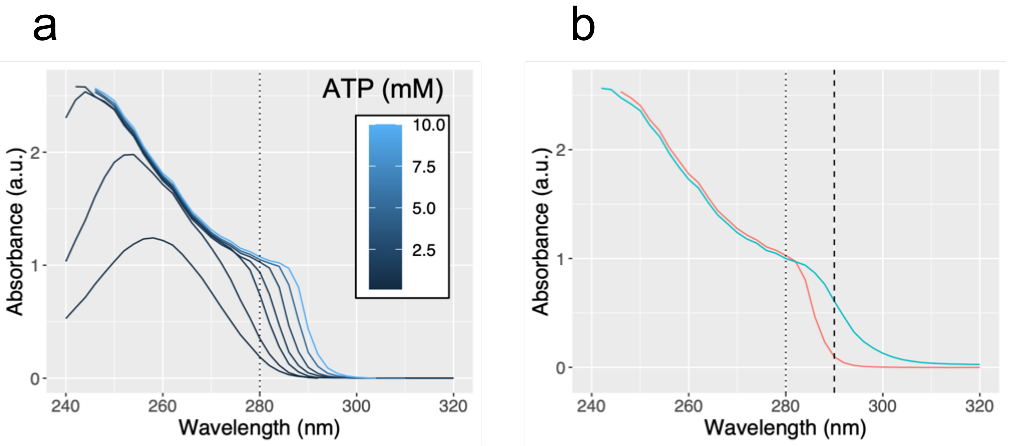


Figure 3.20: Spectrophotometry of ATP-protein containing solutions.
(a) A UV wavelength absorbance scan was performed on various ATP concentrations diluted in 200 mM NaCl, 20 mM Tris-Cl pH 8.0, 5 %(v/v) glycerol. Dotted line (280 nm) represents the absorbance maxima of tryptophan in protein. **(b)** A UV spectrum was performed on 2 mM ATP diluted in buffer in the presence/absence of 2 mg/mL protein (1 A.U). Dashed line (290 nm) represents the wavelength chosen for SEC analyses.

Briefly, 10 µM of MCM samples were pre-incubated in the sample buffer for 10 minutes. Where present, ATP/Mg^{2+} and pT$_{50}$-(6-FAM) was added to the sample at concentration 5/10 mM and 10 µM respectively. Samples were eluted over a Superose 6 Increase 10/300 GL column (GE Healthcare) pre-equilibrated in ATP/Mg^{2+}, 1/10 mM. The elution was monitored by a triple wavelength in-line spectrophotometer and fractionated into 400 µL aliquots. Absorbance profiles were normalised to the maximum absorbance value of each trace. The molecular weight of elution species was estimated through a column calibration curve (Figure 2.5). Fractions were subsequently analysed on SDS-PAGE.

In the absence of ATP or DNA, no *Mac*MCM variant can elute with a volume consistent with a hexamer (Figure 3.21a–d). Addition of ATP drives the equilibrium WAMW towards higher molecular weight species. Out of these samples, only *Mac*MCM$^{\Delta WHD\ E418Q}$ exhibits an elution volume consistent with a hexamer. Addition of DNA alone to *Mac*MCM^{FL} samples results in two peaks that are consistent with non-associating DNA and protein.

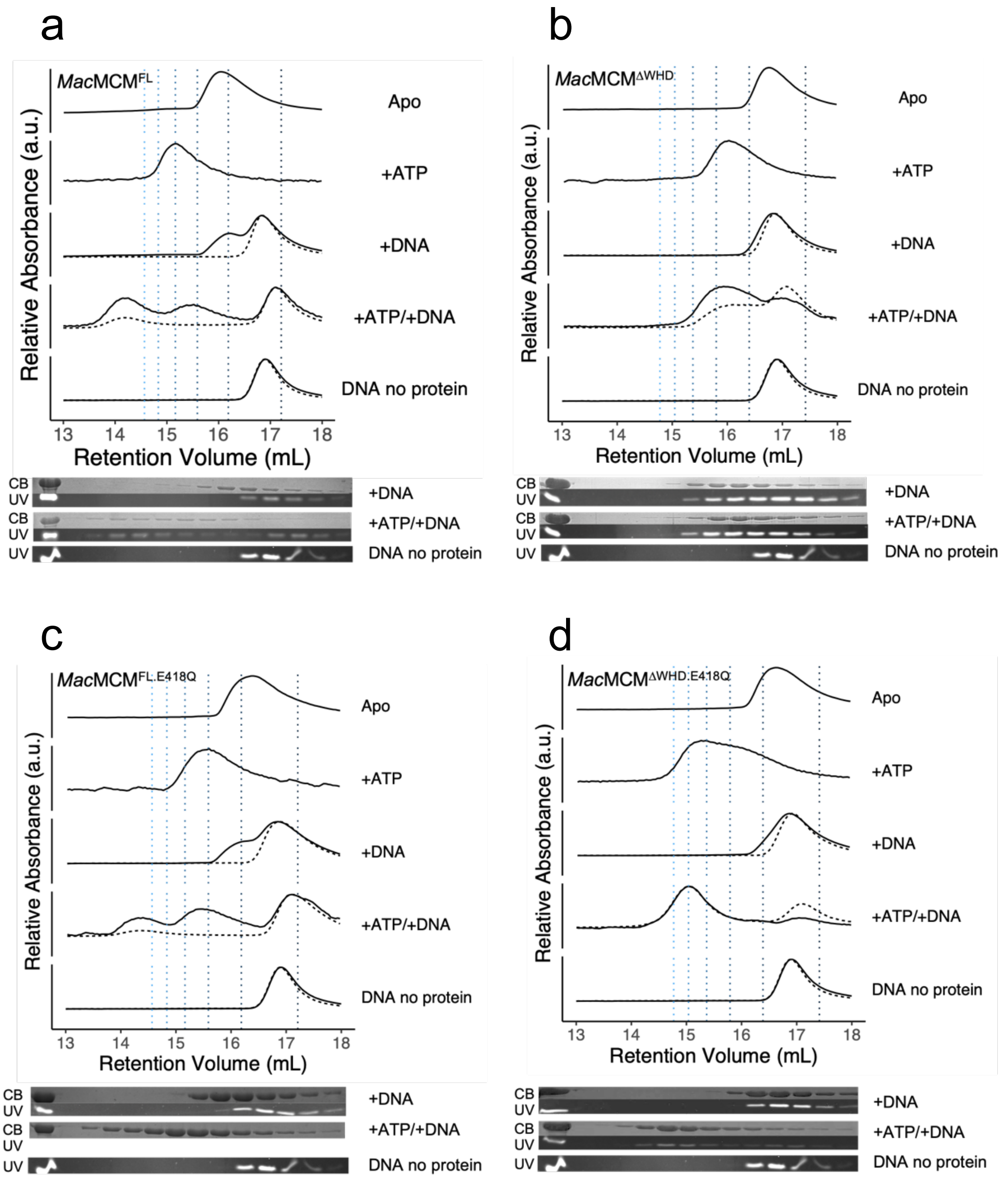


Figure 3.21: Analytical SEC investigations of *Mac*MCM loading onto DNA.
(a-d) 10 µM each stated MCM sample was passed over a S6 10/300 GL Increase column. Where present, equimolar ratios of FAM-pT$_{50}$ DNA were added to the sample. Columns were pre-equilibrated in 200 mM NaCl, 20 mM Tris-Cl pH 8.0, 5 % (v/v) glycerol. Where stated, ATP/Mg^{2+} was added to the column at 1/10 mM. Absorbance was measured at both 290 nm (solid trace) and 495 nm (dotted trace). 400 µL fractions were collected and analysed on a 18 % poly acrylamide gel. DNA was then visualized using a UV transilluminator (UV), then stained with a Coomassie-based dye (CB) to detect protein. The expected elution volumes of hexamers to monomers (left to right) are denoted with a vertical dotted line.

When ATP is not present, *Mac*MCM$^{\Delta WHD}$ variants elute at the same volume as DNA (Figure 3.21b, d). Based on the anisotropy data, they are not expected to interact, however it is likely they share a similar hydrodynamic radius. Addition of both ATP and DNA results in distinct traces for each sample. For *Mac*MCM^{FL} and *Mac*MCM$^{FL.E418Q}$, the enzyme co-elutes with DNA only when there is a peak consistent with the elution volume of a hexamer (Figure 3.21a, c). In the presence of ATP hydrolysis, the hexamer peak for *Mac*MCM^{FL} is notably larger than when ATP hydrolysis is inhibited, consistent with the anisotropy binding data. Comparably, for *Mac*MCM$^{\Delta WHD}$, addition of ATP generates co-elution of protein and DNA at a sub-hexameric oligomeric state (Figure 3.21b). When ATP hydrolysis is removed by mutation, only then is *Mac*MCM$^{\Delta WHD}$ able to form a hexamer that co-elutes with DNA (Figure 3.21d).

*SsoPfu*MCM can form a hexamer under all the tested conditions (Figure 3.22). However, under the tested salt conditions (200 mM NaCl), *SsoPfu*MCM does not interact with DNA (Figure 3.22). Reduction of salt concentration to 100 mM NaCl permits interaction of *SsoPfu*MCM in the absence of ATP. This is consistent with our salt versus DNA-binding experiments discussed in chapter 5. This shift in elution peak for unbound DNA in the low salt condition is likely due to salt-dependent effects on the hydrodynamic radius of DNA[261].

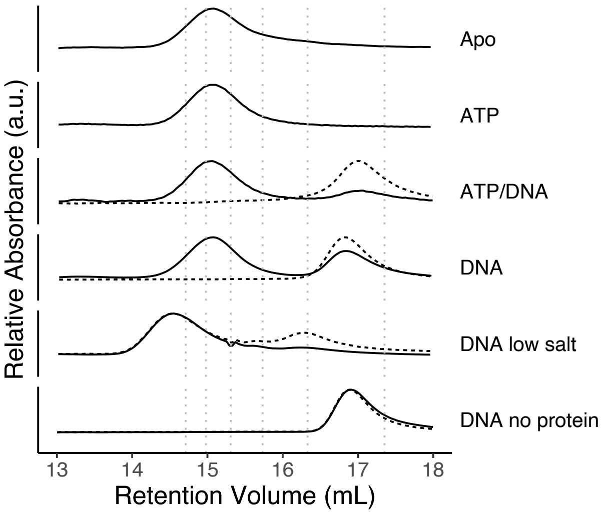


Figure 3.22: Analytical SEC investigations of *SsoPfu*MCM loading onto DNA.
10 µM MCM sample was passed over a S6 10/300 GL Increase column. Where present, equimolar ratios of FAM-pT_{50} DNA were added to the sample. The column was pre-equilibrated in either 100 mM (low salt), or 200 mM NaCl, 20 mM Tris-Cl pH 8.0, 5 % glycerol. Where stated, ATP/Mg^{2+} was added to the column at 1/10 mM. Absorbance was measured at both 290 nm (solid trace) and 495 nm (dotted trace). The expected elution volume of hexamers to monomers (left to right) are denoted with a vertical dotted line.

3.8.5 Pre-incubation fluorescent helicase assays

To examine whether the shift peak observed in the addition of ATP alone for *Mac*MCM^{FL} impacts lag time, a helicase assay was performed. In a standard fluorescent helicase assay, MCM are pre-incubated in the presence of DNA without ATP. Following incubation for ~10 minutes, ATP is added to the sample. To examine whether pre-incubation ATP instead reduces the period of lag time, *Mac*MCM was incubated with ATP for 10 minutes before addition to the DNA sample.

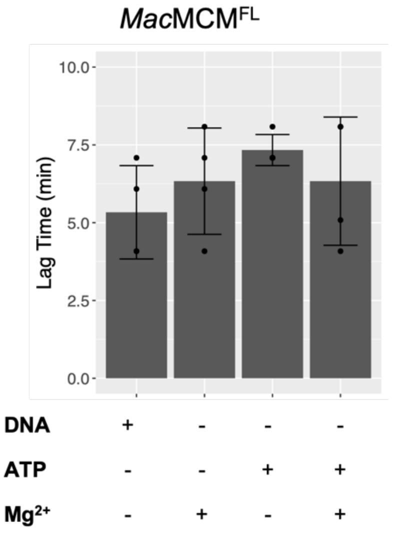


Figure 3.23: Analysis of *Mac*MCM loading by a fluorescent helicase assay.
DNA turnover of a forked 26 base pair DNA substrate was measured in a 96-well plate fluorescent helicase assay. Samples were pre-incubated wither with DNA, ATP, $MgCl_2$ or ATP/$MgCl_2$. For each sample, lag time was calculated from a 30-minute time course experiment, where the time taken to reach the maximum rate of unwinding is extracted. Error bars represent +/-1 standard error of the mean, where n=4.

Under the tested conditions, there is little evidence to suggest that pre-incubation of *Mac*MCM^{FL} with either DNA or ATP alone supports reduction of the lag time (Figure 3.23). This suggests that both DNA and ATP hydrolysis are important together during the kinetics of lag time.

3.9 Conclusions for this chapter

This chapter addresses a long-standing bias with archaeal MCM studies and attempts to relink comparisons with eukaryotic MCM. Historic archaeal MCM studies have largely focused on the biophysical characterisation of MCM from archaeal thermophiles at room temperature. It was hypothesized that since thermophilic enzymes are uniquely adapted to high temperatures, comparison with mesophilic eukaryotic enzymes is weakened. Traditionally, MCM from archaeal mesophiles have been difficult to work with. Computational analyses identified 9 unstudied archaeal MCM homologues from divergent phylogenies and environmental niches. Biochemical analyses of these homologues revealed that almost all MCM from thermophiles are poorly active at room temperature, meanwhile, 2 MCMs from mesophilic organisms and an engineered hybrid exhibit strong activity. Interestingly, the real-time helicase assay permitted observation of a previously unseen kinetic step that has not been seen for a homomeric MCM.

In depth biochemical biophysical characterization of *Mac*MCM suggests that the slow kinetic step is limited by protein-protein, protein-ATP, and protein-DNA interactions that all contribute equally. This agrees with the slow kinetics observed for eukaryotic MCM hexameric ring closure, and to our knowledge has never been demonstrated for an archaeal MCM. Further investigation implies a novel role for the WHD in regulating the slow kinetics *Mac*MCM assembly into a DNA-bound hexamer. It is likely ATP-hydrolysis is required for the remodelling of the position of the WHD.

Chapter 4 – Structural Characterization of *Mac*MCM

4.1 Introduction

The asymmetry provided by the six unique eukaryotic MCM2-7 subunits is central to its function in both the regulation and mechanism of DNA unwinding. When MCM2-7 is loaded onto chromatin in G_1 phase, DNA is threaded through a specific opening between the subunits 2 and 5 [157–159,162,163,165]. When DNA is unwound by the MCM2-7 complex, the lagging strand is believed to exit through a gap between the base of the zinc fingers of neighbouring subunits 3 and 7 [168,170,171].

Evolution of functional asymmetry from an archaeal homohexameric ancestor is unclear. The lack of an ancestral MCM2-5 gate raises fundamental questions about how archaeal MCM is loaded onto archaeal origins of replication. To date, 18 structures of archaeal MCM have been deposited into the PDB utilizing both NMR (2) and X-ray diffraction (16) data [100,107,112,115,118,119,123,124,128,239,249,262]. Uniquely, all these structures focus on MCM from 5 thermophilic archaeal species (*M. kandleri, M. thermautotrophicus, P. furiosus, S. solfataricus, T. acidophilum*). Thus far, the structure of a complete full-length archaeal MCM remains unresolved. The flexibility of the C-terminal winged helix domain is detrimental to the success of crystallography. Four structures exist of archaeal MCM without the winged helix domain, however, only one of these structures *Sso*MCM (PDB:6MII) captures a near native homohexameric form [100]. Whilst *SsoPfu*MCM (PDB:4R7Y) forms an active hexamer, it is an engineered chimeric fusion protein and therefore does not represent a true biological MCM.

Comparatively, the majority of all eukaryotic MCM structures have been resolved using cryo-EM. Density corresponding to the winged-helix domain is also almost always absent, highlighting its expected flexibility. Thus far we have suggested that the sole MCM from *M. acidiphilum* is more biochemically like eukaryotic MCMs than previously studied enzymes. Therefore, there is an interest to discover whether similarities also extend to the structural level.

4.2. Aims

Within this chapter, experiments are performed to resolve the 3D structure of *Mac*MCM to elucidate the structural basis for lag time and its potential link to eukaryotic MCM. To achieve this goal, the following aims were set:

- Determine conditions and constructs suitable for *Mac*MCM crystallization in either the Apo, nucleotide or both a nucleotide and DNA bound state;
- Resolve the 3D crystal structure of *Mac*MCM via X-ray diffraction;
- Assess the link between *Mac*MCM and previously solved MCM structures;
- Investigate the interface composition of MCM based on thermal environmental pressures;
- Investigate the structural mechanisms of lag time through analysis of putative DNA binding residues.

4.3 Crystallization and structure determination of *Mac*MCM

4.3.1 Crystallization of Apo *Mac*MCM

X-ray diffraction has historical success for solving archaeal MCM structures and it was therefore decided to use it for resolving *Mac*MCM [100,107,112,115,119,123,124,128,239,249,262]. As outlined in chapter 3, *Mac*MCM constructs cannot assemble into a hexamer in the absence of ligands. Initial crystallization screens were set up to crystallize the protein in the apo, unassembled form. This would give an insight into the state of the protein that limits assembly before ATP or DNA is added.

Protein homogeneity is a major determinant of crystallography success hence, it was decided to include an anion exchange polishing step for crystallography constructs [263]. Following the second nickel column, samples were passed over a source 15Q anion exchange column and subject to a 20 CV gradient elution from 50 mM NaCl to 1 M NaCl (Figure 4.1a-b). The target protein eluted around 420 mM NaCl, and further increased the purity before spin concentration and SEC (Figure 4.1c-d). A low salt buffer was selected for dialysis before crystallization and the reducing agent DTT was substituted for TCEP, which has a longer half-life. An effort was made to purify the protein as quickly as possible, where protein was prepared immediately before setting up crystallization screens. Protein was routinely dispensed into screens within ~30 hours of cell lysis, where the limiting factor was the duration of tag cleavage. It was noted that in the crystallography buffer, *Mac*MCM was limited to ~11 mg/ml when concentrated by centrifugation. Beyond this point the solution became saturated and a mild, hairy precipitate formed.

Crystallization conditions were screened for both *Mac*MCM^{FL} and *Mac*MCM$^{\Delta WHD}$. A wide range of crystal screens were set up, carried out in parallel at 4 and 20 °C. No crystals were observed for the *Mac*MCM^{FL} screens, which is consistent with studies that have struggled to generate high resolution diffracting crystals when the flexible winged helix is

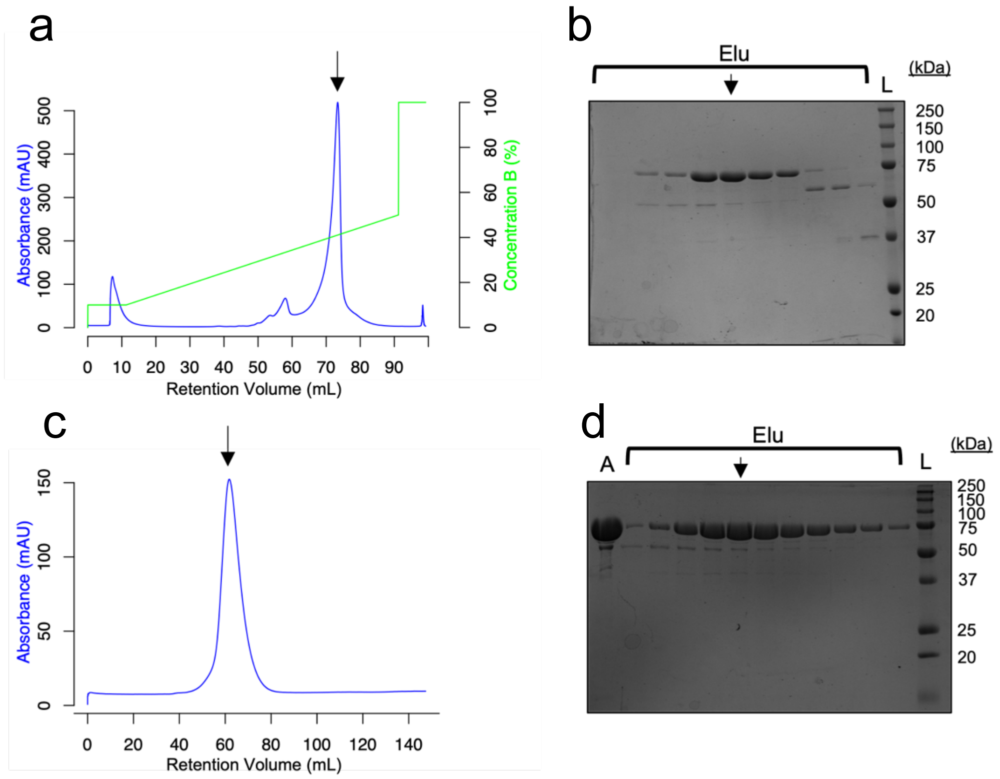


Figure 4.1: An example purification strategy for crystallography constructs.
Purification strategy was selected to maximize purity over yield. Protein is routinely purified within 30-hrs post lysis. **(a)** Anion exchange chromatogram recorded at A_{280}, eluted with an NaCl gradient (Green). 100 % = 1 M NaCl. **(b)** SDS-PAGE of *Mac*MCM (MW = 69.3 kDa), anion exchange fractions on a 12 % (*w/v*) polyacrylamide gel. **(c)** Size exclusion chromatogram recorded at A_{280}. **(d)** Fractions of size exclusion chromatography analysed by SDS-PAGE on a 12 % (w/v) polyacrylamide gel. A: applied fraction; Elu: elutions; L: MW ladder (BioRad); Arrows: equivalent elution fractions on chromatogram/gel.

present[107,123]. Removal of flexible domains that increase the conformational heterogeneity is a well-established route for improving crystallography success [223]. Crystals were observed in at least 6 conditions for *Mac*MCM$^{\Delta WHD}$, however only at 20 °C. Of the 6 crystals, only one of the conditions (0.2 M $(NH_4)_2SO_4$, 0.1 M Bis-Tris-Cl pH 8.0, 25 % PEG 3350) from the PDB screen exhibited diffraction. In this instance, a large plate shaped crystal yielded weak ~10 Å diffraction (Figure 4.2a–b). Extensive optimization of this condition was performed in the hanging-drop format, where increasing the concentration of $(NH_4)_2SO_4$ yielded an increase in the intensity and resolution of diffraction to ~8 Å (Figure 4.2d). Unexpectedly, the increase in salt concentration altered the morphology of the crystal from plate to triclinic (Figure 4.2c). Despite improvements in diffraction quality, crystal handling deteriorated substantially, becoming brittle and sensitive to air. Typically, 1 minute after exposure to air crystals would rapidly phase separate. Extensive optimization was carried out on this condition based on established techniques including: in drop dehydration, cryoprotection, protease treatment with trypsin and chymotrypsin (for removal of unstable surface loops) and additive screens [264–267].

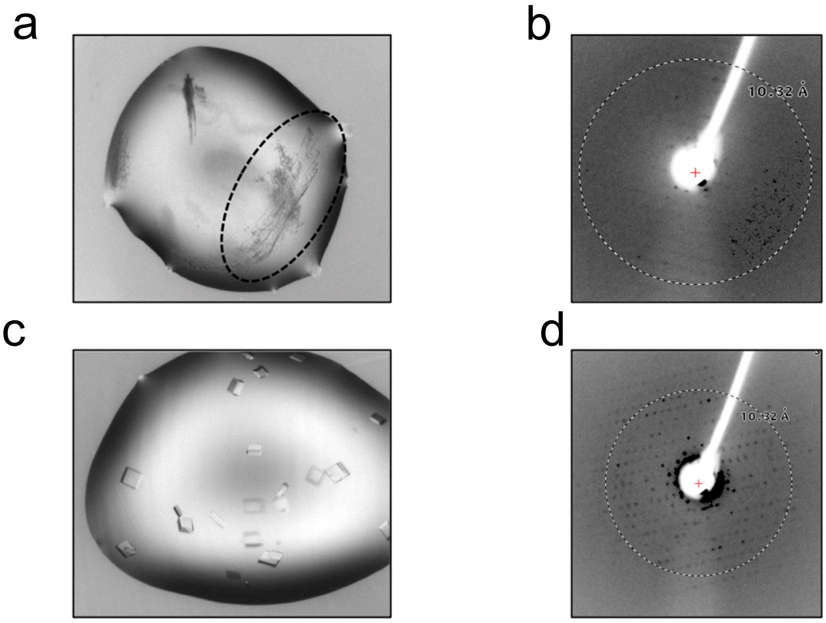


Figure 4.2: Crystals of apo *Mac*MCM$^{\Delta WHD}$.
(a) *Mac*MCM$^{\Delta WHD}$ crystals grown in in 200 mM $(NH_4)_2SO_4$, 0.1 M Bis-Tris-Cl pH 6.5, 25 % (*w/v*) PEG 3,350 in the sitting drop format. **(b)** Diffraction pattern of crystals in part (a). **(c)** *Mac*MCM$^{\Delta WHD}$ crystals grown in in 400 mM $(NH_4)_2SO_4$, 0.1 M Bis-Tris-Cl pH 6.5, 25 % (*w/v*) PEG 3,350 in the hanging drop format. **(d)** The measured diffraction pattern of crystals in part (c).

Performance of an additive screen yielded a broad spectrum of crystal morphologies and resulted in diffraction testing of 106 crystals. Two additives increased diffraction notably: 10 mM $MgCl_2$ (a known metal cofactor of MCM), which improved the intensity of diffraction spots and 10 mM taurine, which reduced smearing of the diffraction pattern. Based on this, the two conditions were combined, yielding large, plate-shaped crystals that generated clean 6 Å diffraction in house. This crystal was subsequently sent to Diamond Light Source for further analysis, where it produced a 4.2 Å dataset with poor merging statistics (Table 4.1). Further optimizations failed to improve data quality and it was therefore not possible to solve a structure from this dataset.

Table 4.1: Data collection statistics for *Mac*MCM$^{\Delta WHD}$ (apo).
Values in parentheses outline equivalent data for the highest resolution shell. DLS: Diamond Light Source. CC: correlation coefficient. For the formulas of R_{meas} and $CC_{1/2}$, see section 2.8.

Parameters and statistics	
Data collection	
Beamline	I04, DLS
Wavelength (Å)	0.9795
Resolution (Å)	4.02 – 119.55 (4.02 – 4.51)
Cell dimensions	
Space group	P1
Unit-cell a,b,c (Å)	91.64 132.57 151.93
Unit-cell α, β, γ (˚)	102.88, 100.43, 108.89
Data merging statistics	
Unique reflections	45,356
Completeness (%)	86.5 (65.2)
Multiplicity	3.6 (3.6)
I/σ <I>	5.0 (1.7)
R_{meas} (%)	0.130 (0.770)
$CC_{1/2}$	1.0 (0.7)
Wilson B-factor	141.7

4.3.2 Crystallization of nucleotide bound *Mac*MCM

Co-crystallization of protein with ligands is a routine strategy for increasing success of a crystallography project through increased structure stabilization and homogeneity [268–270]. In many cases, thousands of ligands must be screened, however, if a known natural ligand is known the process can be accelerated markedly. For MCM, these ligands may include DNA, ATP/Mg^{2+} and other nucleotide analogues. Many of the ATP analogues tested in chapter 3 do not elicit a strong response from MCMs when DNA binding affinity is assessed. Comparatively, the inactive *Mac*MCM$^{\Delta WHD.E418Q}$ mutant exhibited a strong oligomerization and DNA binding response in the presence of ATP/Mg^{2+}. It was therefore decided to perform crystallization of *Mac*MCM$^{\Delta WHD.E418Q}$ in the presence of ATP/Mg^{2+}.

4.3.3. Determination of protein-ATP affinity by NanoDSF and static light scattering

To this point, it had been assumed that *Mac*MCM$^{\Delta WHD.E418Q}$ is able to interact with ATP as despite being functionally inactive (see section 3.7.3), it is able to hexamerize and bind to ssDNA when ATP is present. To examine whether ATP interacts with *Mac*MCM$^{\Delta WHD\text{-}E418Q}$, and establish at what point the enzyme becomes saturated, NanoDSF was performed (see section 2.5.4). NanoDSF primarily monitors the solvent accessibility of the 3 native tryptophan residues in *Mac*MCM$^{\Delta WHD.E418Q}$ monomers. Changes in solvent accessibility and hence fluorescence may be caused by either oligomeric state, thermal denaturation, or aggregation.

In short, *Mac*MCM$^{\Delta WHD.E418Q}$ was dialysed into the crystallization buffer and mixed to a final concentration of 2.5 mg/mL with a series of ligands, including 10 mM $MgCl_2$, 6.3 µM ssDNA ($(ACTG)_{16}$) and 2.5 mM ATP. Thermal denaturation of samples was performed from 20 °C to 95 °C at a rate of 1 °C/min. Unfolding was monitored through dual detection of tryptophan fluorescence emission at 350 and 330 nm. Aggregation was also monitored through static light scattering.

In the absence of ligands, *Mac*MCM$^{\Delta WHD.E418Q}$ is well folded up to 54 °C. This is consistent with the expected environmental temperature of *M. acidiphilum* (maximum ~45 °C). Addition of $MgCl_2$ or ssDNA alone has no effect on the stability of *Mac*MCM$^{\Delta WHD.E418Q}$, suggesting minimal structural interaction. Interestingly, as the enzyme unfolds, the tryptophan environment becomes more hydrophobic indicated by the decrease in the F_{350}/F_{330} ratio (Figure 4.3a). Decrease in solvent accessibility may occur through association of hydrophobic material during aggregation. For *Mac*MCM$^{\Delta WHD.E418Q}$, the decrease in solvent accessibility coincides with onset of aggregation, as measured through static light scattering (Figure 4.3b). The unfolding profile of *Mac*MCM$^{\Delta WHD.E418Q}$ only changes where ATP is added alongside magnesium (Figure 4.3a). Magnesium is assumed to be essential for coordinating ATP in the active site of MCM. Where both ATP and magnesium are present, the fluorescence profile changes from a single peak transition to multi peaked transition (Figure 4.3a). This is also the case when DNA, ATP and $MgCl_2$ are all present. Based on our previous SEC data (see section 3.8.4), the single transition likely corresponds to denaturation of a monomer in the absence of oligomerization. The multi-peak transition likely corresponds to the dissociation of higher molecular weight species (i.e., hexamers) followed by the denaturation of a monomer.

Because each *Mac*MCM$^{\Delta WHD.E418Q}$ subunit contains 3 tryptophan residues in different positions, the fluorescence data is a complicated signal reporting on multiple local environments simultaneously. Alternatively, for a homogenous solution of protein, it is only possible for the sample to aggregate once during denaturation. Therefore, protein aggregation measured from the scattering signal can represent a simplified description of protein stability in response to ligands. Here, only where ATP and Magnesium are present does the onset of aggregation increase from 55 to 75 °C (Figure 4.3b). In all instances, aggregation occurs as a single event/peak.

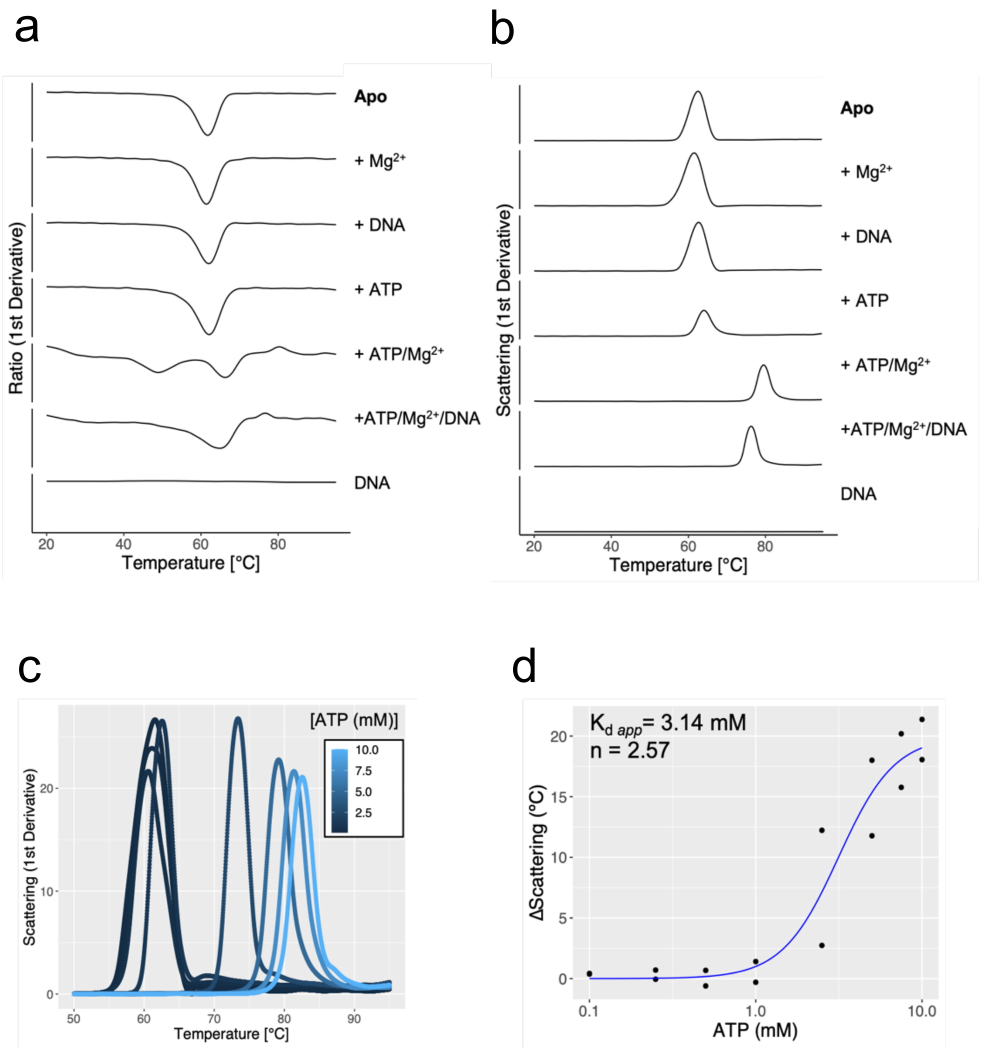


Figure 4.3: Thermal denaturation to elucidate *Mac*MCM$^{\Delta WHD.E418Q}$-ligand interactions.
(a) Thermal denaturation is measured by F_{350}/F_{330} and converted to a 1st derivative to highlight the points of maximal transition. **(b)** Heat-induced aggregation is determined by static light scattering and converted to a 1st derivative to examine the points of maximal aggregation onset. **(c)** Heat-induced aggregation is determined by static light scattering as a function of ATP concentration and converted to a 1st derivative to examine the points of maximal aggregation onset. **(d)** The change in aggregation propensity in the presence of increasing concentrations of ATP (from **(c)**) is standardized by subtracting the T_m of the scattering peak for +Mg^{2+} only from each ATP concentration. Data is fit to a Langmuir binding isotherm where $K_{d,\ app}$ = 3.14 mM and the Hill coefficient is equal to 2.57 (please refer to section 2.5.4 for the equation).

To achieve complex saturation, it is essential in crystallography to add a concentration of ligand that is ideally in at least 10-fold excess of the equilibrium dissociation constant (K_d) between the protein and the ligand [271]. To determine enzyme saturation, ATP was added to samples at increasing concentrations ranging from 0.1–10 mM and thermal denaturation was repeated. Given the complexity of the tryptophan fluorescence data, only the scattering data was analysed. The temperature of the maximal rate change of scattering for each ATP concentration was extracted (Figure 4.3c). This temperature was then standardized by measuring the temperature change compared to a no ATP (Mg^{2+} only) control. Data was then fit to a Langmuir binding isotherm, yielding an estimated K_d ($K_{d,\,app}$) of ~3.14 mM (Figure 4.4d). It is important to note that this calculation of $K_{d,\,app}$ is a best estimate of the K_d and measurements are assumed to be a proxy for ATP binding to the active site. For example, ATP is a known hydrotrope and hence may natively increase protein stability independent of specific protein-ligand interactions [272]. A Hill-coefficient was used to improve the quality of the model fit. This is justified by the observation that hexameric AAA+-enzymes exhibit intersubunit allostery between active sites [273,274]. Saturation of the enzyme was achieved around 10 mM, and this was therefore selected for crystallization screening. Increasing the concentration of ATP an order of magnitude larger than 10 mM is not viable, as the maximum solubility of ATP stocks is ~100 mM.

4.3.4 Confirmation of oligomeric state by analytical SEC

Importantly, PDB depositions contain information about the suggested biological assembly of the enzyme. Analytical SEC was deployed to confirm the oligomeric state of *Mac*MCM in the crystallography buffer with the presence of ligands (Figure 4.4a). As observed previously, *Mac*MCM$^{\Delta WHD.E418Q}$ does not assemble into a hexamer in the absence of ligands. In the presence of 1/10 mM of ATP/Mg^{2+}, *Mac*MCM$^{\Delta WHD.E418Q}$ elutes as a broad peak with a population that is consistent with a hexamer. The previous ssDNA SEC substrate was possibly too long (50 nt) for crystallization, where free, flexible macromolecules are detrimental (see section 5.8.3 for estimates of MCM DNA binding footprint). Therefore, we opted for a short, unlabelled 16 nt ssDNA molecule. As before, ssDNA can increase the stability of the hexamer complex, indicated by the decrease in elution volume in the presence of ATP. As SDS interferes with SYBR-gold labelling efficiency, DNA was detected on a secondary 8% 1 x TB native PAGE gel. As previously, *Mac*MCM$^{\Delta WHD.E418Q}$ coelutes with the short ssDNA in the presence of ATP (Figure 4.4b).

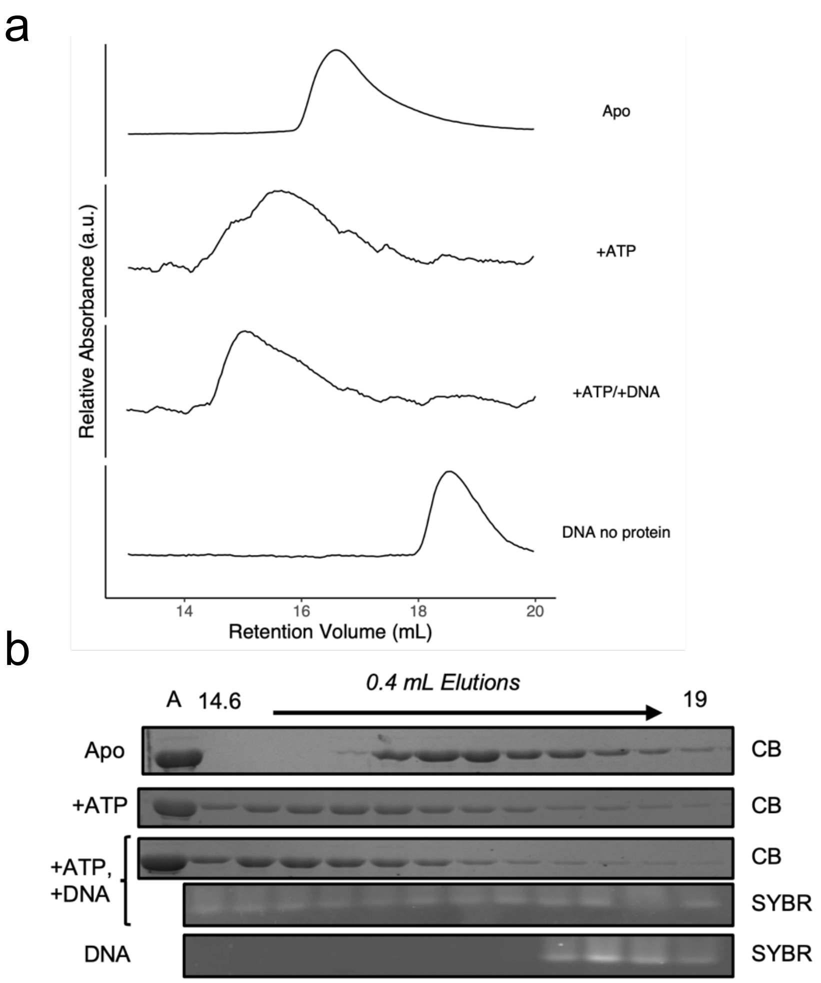


Figure 4.4: Analytical SEC confirms *Mac*MCM$^{\Delta\text{WHD.E418Q}}$ hexamerization in 1/10 mM ATP/Mg^{2+}. **(a)** 10 µM each stated MCM sample was passed over a S6 10/300 GL Increase column equilibrated in crystallography buffer. Where present, equimolar ratios of (ACTG)$_{16}$ were added to the sample. Where stated, ATP/Mg^{2+} was added to the column at 1/10 mM. Absorbance was measured at A$_{290}$. **(b)** Fractions were collected and analyzed on polyacrylamide gels, either: 4 % native 1 x TB PAGE (ssDNA) or 12 % SDS-PAGE (protein). Gels were then post-stained and imaged based on the macromolecular target, either SYBR-GOLD (ssDNA) or Coomassie (protein) dye (CB).

4.3.5 Initial screens and optimization of ATP bound *Mac*MCM$^{\Delta\text{WHD-E418Q}}$

Subsequently, protein-ATP complex was prepared by addition of both ATP and MgCl$_2$ to final concentrations of 10 mM. Protein was dispensed as mentioned previously, however screens were only set up at 20 °C. Crystals were obtained in 10 conditions in both MORPHEUS and PACT screens. Large plate-shaped crystals in condition C5 of the MORPHEUS screen generated the best 6 Å diffraction when tested on the York X-ray facilities (10 % (*w/v*) PEG 20 000, 20 % (*v/v*) PEG MME 550, 0.03 M of each NPS, 0.1 M MOPS/HEPES-NaOH pH 7.5) (Figure 4.5a). Optimization of the buffer components did not improve the diffraction quality of the crystal notably. Instead, growth of crystals in a 20 µL sitting drop format generated giant 0.5 mm long crystals that diffracted to ~3.8 Å on the York X-ray facilities (Figure 4.5b-c). No ice issues were observed which is expected from the Morpheus screen, where all conditions contain small polyols that confer cryoprotection [275].

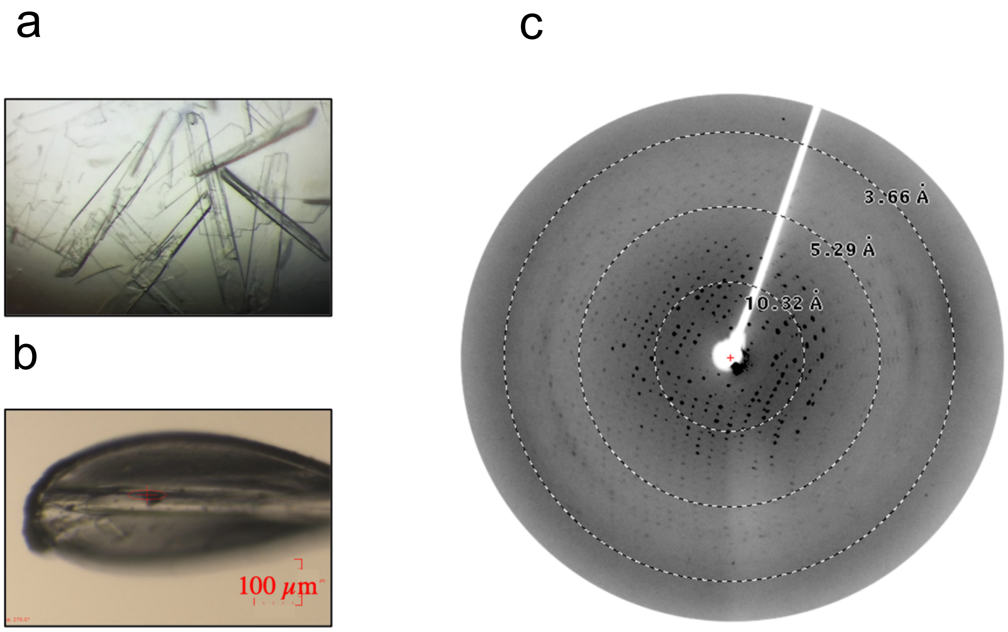


Figure 4.5: Crystals of *Mac*MCM$^{\Delta\text{WHD-E418Q}}$ + ATP/Mg^{2+} diffract.
(a) Large crystals grown in 10% (*w/v*) PEG 20 000, 20% (*v/v*) PEG MME 550 0.03 M of each NPS 0.1 M MOPS/HEPES-NaOH pH 7.5 at 20 °C, in the sitting drop format**.** **(b)** Scaled image of a large single crystal from (a) immobilised in a cryo-loop without cryoprotection. **(c)** Diffraction pattern generated from the crystal in (b).

4.3.6 Structure determination of *Mac*MCM$^{\Delta WHD-E418Q}$

As zinc is intrinsically present within the zinc fingers of MCM, the feasibility of using native heavy atoms to solve the structure through SAD phasing was assessed. An X-ray fluorescence (XRF) scan was collected on a crystal, revealing a spectrum consistent with the presence of zinc (http://www.xrfresearch.com/xrf-spectrum-zinc/) (Figure 4.6a). Notably, zinc is never supplemented exogenously to MCM during purification. Analysis of the XRF scan was then performed to determine the optimal wavelength for Zn-SAD dataset collection (Figure 4.6b). Zn-SAD collection was then performed at a wavelength of 1.2828 Å and generated a dataset with a maximum resolution of 3.1 Å (Table 4.2). Due to the weak nature of the anomalous zinc signal, it was not possible to phase the dataset via SAD.

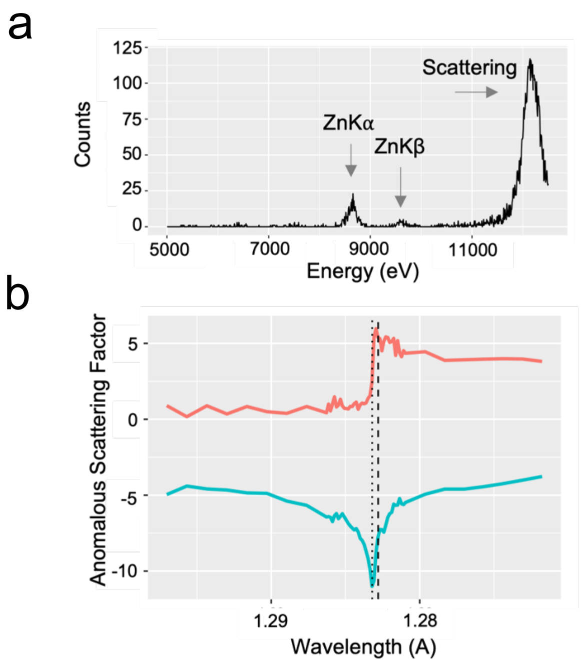


Figure 4.6: *Mac*MCM$^{\Delta WHD.E418Q}$ crystals contain zinc, without exogeneous supplementation. **(a)** X-ray fluorescence spectrum for a MacMCM $^{\Delta WHD.E418Q}$ +ATP/Mg^{2+} crystal reveals a characteristic trace for the presence of zinc. Pure 99% zinc exhibits 2 characteristic peaks: 1 main peak (ZnKα) at 8.64 keV and a smaller secondary peak (ZnKβ) at 9.57 keV. The spectra suggest that crystals may be solvable through SAD phasing that exploits the intrinsic heavy atom binding sites (zinc fingers) in MCM. **(b)** Determination of optimal wavelength for Zn-SAD dataset collection at Diamond Light Source. Data from (a) are converted into a wavelength using E = hc/λ, then the first (f') and second (f'') derivatives are calculated to determine the absorption edge.

Molecular replacement (MR) was then performed on the Zn-SAD dataset using the program Phaser within Phenix [218,229]. MR successfully placed a custom search model of the C-terminal domain of *Pfu*MCM (50% identity) as a ring-shaped hexamer within the asymmetric unit (Figure 4.7a-b) [124]. The log-likelihood gain (LLG) and translation function Z-scores are commonly used to assess how well an MR solution matches the experimental data. The LLG and TFZ scores were 234 and 14.5 respectively, which are indicative of definitive MR solutions [218]. Subsequent attempts to automatically place models of the N-terminal domain by Phaser failed, suggesting structural differences exist between *Mac*MCM and MR models. If the MR N-terminal domain models were similar in structure to *Mac*MCM, Phaser would have been expected to place the models into the asymmetric unit. In accordance with Bragg's law (see section 2.8.1), the wavelength selected for the Zn-SAD limited the maximum resolution generated by the crystal. As it was possible to solve the SAD-dataset by MR, a lower wavelength dataset was recorded at 0.9763 Å to increase the angle of diffraction. This yielded an improved dataset with a maximal resolution of 2.59 Å (Table 4.2).

Table 4.2: Data collection statistics for *Mac*MCM$^{\Delta WHD.E418Q}$.
Values in parentheses outline equivalent data for the highest resolution shell. DLS: Diamond Light Source. CC: correlation coefficient. For the formulas of R_{meas} and $CC_{1/2}$, see section 2.8.

Parameters and statistics		
Data collection	λ_1 Zn-SAD	λ_2
Beamline	I04, DLS	I03, DLS
Wavelength (Å)	1.2828	0.9763
Number images	12000	3600
Rotation per image (°)	0.1	0.1
Resolution (Å)	3.13 – 97.23 (3.13 – 3.19)	2.59 – 57.15 (2.59 – 2.64)
Cell dimensions		
Space group	C 1 2 1	C 1 2 1
Unit-cell a,b,c (Å)	228.22 127.95 176.87	228.42 127.29 176.75
Unit-cell α, β, γ (˚)	90.00 91.79 90.00	90.00 91.71 90.00
Data merging statistics		
Unique reflections	1475427 (63606)	157088 (15547)
Completeness (%)	100.0 (99.1)	99.99 (98.2)
Multiplicity	22.6 (21.1)	6.8 (7.0)
I/σ <I>	9.2 (0.3)	11.3 (0.3)
R_{meas} (%)	0.2 (5.868)	0.115 (4.564)
$CC_{1/2}$	1.0 (0.3)	1.0 (0.5)
Wilson B-factor	123.6	84.4

MR was performed, and phases were improved using the RESOLVE density modification package in Phenix (Figure 4.7a–c) [219]. Subsequently, cycles of manual building in Coot, chain extension in AutoBuild and Buccaneer were performed to build a model of the N-terminal domain into the electron density. Between cycles of model building, refinement of the model and phases was performed using Phenix.refine until improvements ceased [219,228,276]. Using unmerged data, anomalous difference maps were calculated in Phenix.maps to reveal the precise positioning of the metal zinc ions within the N-terminal domain (Figure 4.7d). Tetrahedral, zinc coordination restraints were then generated for the 4 cysteine residues using Phenix ReadySet! (Figure 4.8a). Strong density was observed at the end of a cluster of basic residues in the central channel (Figure 4.8b). As this patch is expected to bind the phosphate backbone of DNA, the density was modelled as phosphate, which was present in the crystallization solution [128]. It is of note that the buffer also contained a cocktail of similar anions, including sulfate and nitrate. Because of the strong density, the molecule is likely a phosphate or sulfate, however it is difficult to definitively conclude the absolute identity. Applying prior knowledge represents the best approach.

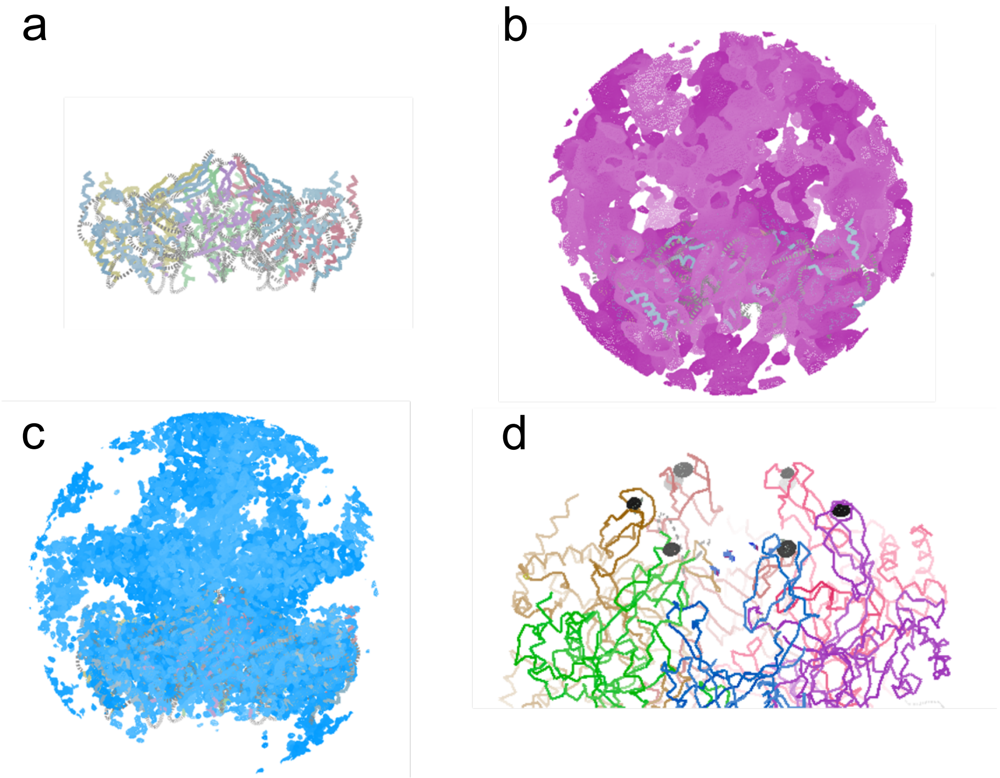


Figure 4.7: Solving the structure of *Mac*MCM$^{\Delta WHD.E418Q}$
(a) Placement of 6 copies of a polyalanine model of *Pfu*MCM CTD into a planar hexamer by Phaser-MR. **(b)** Initial electron density map output by Phaser for (a), where σ = 0.7, and the map radius = 75 °A. **(c)** Improvement in map quality following Phenix density modification, where σ = 2.0, and the map radius = 75 °A. **(d)** Calculated anomalous difference map for *Mac*MCM$^{\Delta WHD.E418Q}$ after model building reveals precise positioning of zinc ions in the Zinc fingers (σ = 3.0). All images were created in Coot, where the protein C-alpha trace is displayed.

Of the 6 ATPase active sites, 5 contained density that is consistent with the presence of ADP. Despite the proven enzymatic inactivity, it is likely that ATP breaks down in the sample over multiple days at 20 °C (Figure 4.8c). The remaining active site (chain D) contained density consistent with a phosphate ion (Figure 8d). Density consistent with Mg^{2+} was not observed in any active site, expected at low resolution. Observation of coordination geometry is required to distinguish between water and small ions. This is only achievable through collection of high resolution datasets (<2 Å) [277]. Ligand refinement was performed through extensive cycles of editing in Coot and Phenix.refine until improvements ceased and a final R_{work}/R_{free} of 0.23/0.25 was achieved (Table 4.3) [228,278].

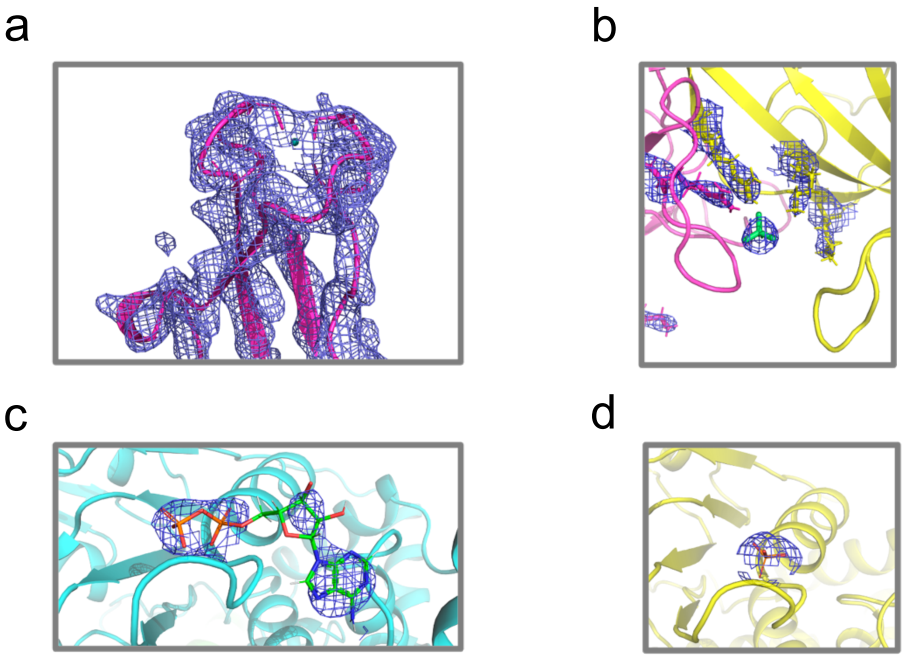


Figure 4.8: Example ligands in *Mac*MCM$^{\Delta WHD.E418Q}$.
All images were created in PyMol, where the respective map is contoured to σ = 2. Protein is visualised in the ribbon format. **(a)** Rich electron density ($2mF_o – DF_c$) around the zinc finger due to the presence of strong scatterers (S/Zn). Zn^{2+} is represented by a sphere. **(b)** 1 of the 6 phosphate (sticks) residues in the central channel of the N-terminal domain, co-ordinated by basic residues of the OB-fold (Omit-map). **(c)** 1 of the 5 ADPs (sticks) modelled in an ATP active site of the C-terminal domain (Omit-map). **(d)** The sole phosphate (sticks) residue modelled in an ATP active site of the C-terminal domain.

Table 4.3: Refinement statistics for *Mac*MCM$^{\Delta WHD.E418Q}$.
Values in parentheses outline equivalent data for the highest resolution shell. CC: correlation coefficient. RMSD: root mean square deviation. For the formulas of CC*, CC$_{work}$, CC$_{free}$, R$_{work}$ and R$_{free}$, see section 2.8.

Refinement	**Standard**
Reflections used in refinement	154138 (12791)
Reflections used for R-free	1983 (169)
R$_{work}$	0.230 (0.419)
R$_{free}$	0.253 (0.430)
CC*	0.999 (0.8)
CC$_{work}$	0.951 (0.359)
CC$_{free}$	0.936 (0.166)
Protein residues	3439
Number of non-hydrogen atoms	26948
Macromolecules	26750
Ligands	198
RMSD bonds (Å)	0.010
RMSD angles (°)	1.70
Average B-factor (Å^2)	98.4
Macromolecules	98.2
Ligands	124.4
Ramachandran plot (%)	
Favored	93.6
Allowed	5.6
Outliers	0.8

4.3.7 Structural overview of ADP bound *Mac*MCM$^{\Delta WHD\text{-}E418Q}$

ADP bound *Mac*MCM forms a large ring-shaped hexamer (Figure 4.9). The diameter and height of the enzyme are 115 Å and 98 Å respectively. Measured at its narrowest point, the central channel is 20 Å wide, making it wide enough to accommodate dsDNA. In general, all the key features of MCM are structurally present, except the winged helix domain which was not included in this construct. The N and C-terminal domains are joined by a short linker with well resolved density. Previous archaeal MCM hexamer structures have relied on engineered, non-native linker regions to reduce inter-domain flexibility and improve crystallization [100,124] . In the final model, out of 3,600 amino acids in total, only 4 % are unresolved. Unresolved residues are expected to be at termini and within flexible loop regions in the C-terminal domain.

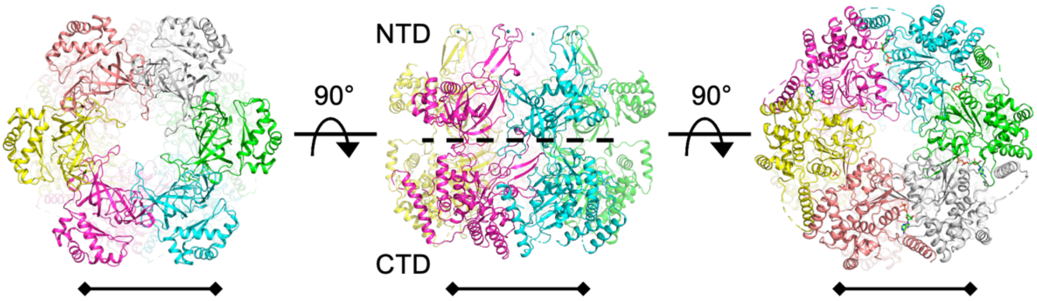


Figure 4.9: *Mac*MCM$^{\Delta WHD.E418Q}$ structure overview.
A cartoon visualization of *Mac*MCM$^{\Delta WHD.E418Q}$ in the ribbon format, where colours represent each chain of the model. The scale bar represents 50 Å The diameter of the enzyme is 115 Å, whilst the height of the enzyme is 98 Å. All images were created in PyMol. NTD: N-terminal domain, CTD: C-terminal domain.

4.4 Investigations of *Mac*MCM interface interactions

Comparisons of *Mac*MCM interfaces will primarily be made with *Sso*MCM (PDB: 6MII)[100] and *Sce*MCM (PDB:6EYC)[140]. Importantly, all structures represent hexameric complexes with 6 structurally distinct chains. *Sce*MCM represents a near complete ADP-bound eukaryotic MCM2-7 structure in absence of other protein cofactors. This structure also represents a well refined eukaryotic MCM2-7 structure. Many eukaryotic cryo-EM structures contain large geometrical errors, such as cis-peptides and shift registers [140]. Although the *Sso*MCM structure is bound to ssDNA and ADP, ADP-BeF_3, it represents the only possible comparison with a native hexameric archaeal MCM structure. Other archaeal MCM hexamer structures lack both N- and C-terminal domains together and since there is domain overlap, may not reflect the true nature of the interface.

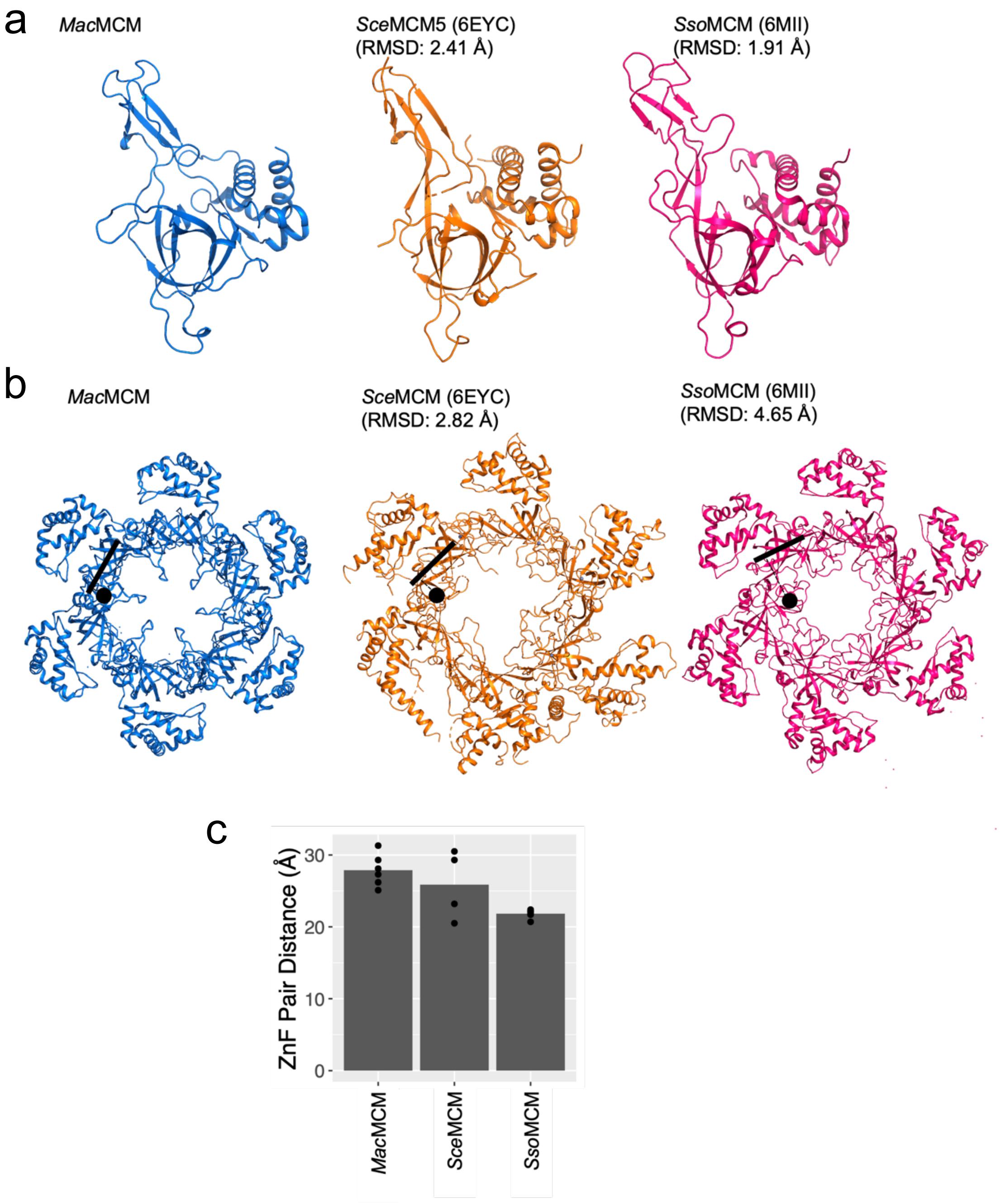


Figure 4.10: Comparison of MCM N-terminal domains.
N-terminal domains of *Mac*MCM (this study, blue), *Sce*MCM (PDB:6EYC, orange)[140] and *Sso*MCM (PDB: 6MII, magenta)[100] are represented in the ribbon format **(a)** Single N-terminal domains were compared using the PyMol sequence-based structural alignment command, align. Values in parentheses represent the all-atom RMSD between the respective structure and *Mac*MCM. **(b)** N-terminal domain hexamers were compared using the PyMol structure-based alignment command (cealign). Black dots represent the position of a zinc atom, whist the black line represents the direction of the core OB-fold. Values in parentheses represent the alpha-carbon RMSD between the respective structure and *Mac*MCM. **(c)** Distances were measured between neighbouring zinc finger pairs for each MCM structure. Note: *Sce*MCM subunit 3 lacks a functional zinc atom, and therefore *Sce*MCM2-7 has 2 fewer measurements.

4.4.1 *Mac*MCM zinc-finger asymmetry

The largest topological difference between *Mac*MCM and previously elucidated archaeal MCM structures are the N-terminal Zinc Fingers (ZnFs). The structure of the N-terminal domains align well when analysis is performed on single protomers (Figure 4.10a). However, when hexamers are compared through structural alignment, *Mac*MCM more closely resembles *Sce*MCM2-7 than *Sso*MCM (Figure 4.10b). Structural comparison of the N-terminal rings suggests *Sso*MCM has a smaller central channel. Measured at the ZnFs, the average diameter of the central channel is 56 Å for *Mac*MCM, 55 Å for *Sce*MCM, but only 42 Å for *Sso*MCM. However, when measured at the N-terminal DNA-binding hairpins, the diameter is 20 Å for *Mac*MCM, 28 Å for *Sce*MCM and 25 Å for *Sso*MCM. Inspection of hexamers suggest that *Sso*MCM ZnFs tilt more into neighbouring subunits and the central channel than *Mac*MCM or *Sce*MCM2-7 (Figure 4.10c). ZnFs of both *Sce*MCM and *Mac*MCM sit more directly over the OB-fold of the same subunit.

On average, the distance between neighbouring ZnF pairs is largest for *Mac*MCM at 28 Å. For *Sce*MCM, ZnF pairs are on average ~2 Å closer, whilst *Sso*MCM ZnFs are 6 Å closer. Moreover, like *Sce*MCM, the distance between neighbouring ZnFs is not consistent in *Mac*MCM, where distances range from 23–30 and 25–31 Å respectively (Figure 4.10c). For *Sso*MCM, the distance between neighbouring ZnFs is much more consistent, ranging from 21-22 Å. In *Sce*MCM2-7, asymmetric differences are generated through MCM subunit specific structural insertions between ZnFs[140]. For *Mac*MCM it is likely these differences are caused by asymmetric crystal contacts. Reconstruction of the *Mac*MCM crystal lattice in PyMol reveals asymmetrical contacts between the 6 ZnFs of one MCM hexamer and chain A of a neighbouring MCM hexamer in the crystal (Figure 4.11). The ZnFs are connected to

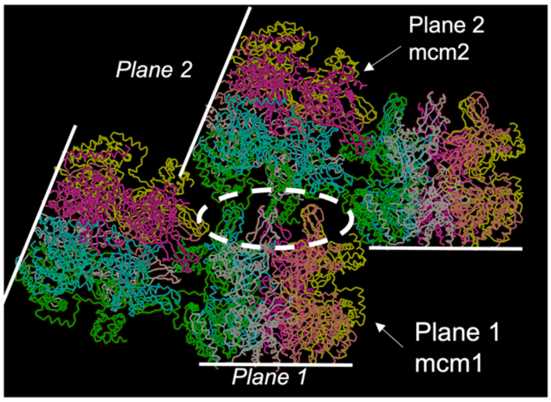


Figure 4.11: *Mac*MCM makes asymmetric crystal contacts at the zinc fingers.
*Mac*MCM$^{\Delta WHD.E418Q}$ crystal packing is visualized in PyMol, using the 'generate symmetry mates' command. The C-alpha trace is visualized and coloured by chain identity. The crystal lattice is organized into 2 intersecting planes. An MCM (e.g., mcm1) contacts the chain A of a neighbouring MCM (e.g., mcm2). The chain A of the neighbouring mcm2 directly enters the ring of mcm1, making asymmetric contacts with residues of the zinc fingers. This is repeated throughout the lattice.

the OB-fold by a short linker, and therefore the precise positioning of the ZnF subdomain is likely flexible. It is possible that asymmetric crystal contacts will slightly change the positioning of each ZnF within an asymmetric unit. Therefore, the protein crystal contacts limit the extent which conclusions may be drawn about the mechanistic functions of ZnF asymmetry in *Mac*MCM.

4.4.2 Comparison of *Mac*MCM AAA+ fold with archaeal and eukaryotic homologues

Structural alignment of C-terminal domain hexamers, suggest *Mac*MCM shares equivalent structural similarities to both *Sso*MCM and *Sce*MCM (Figure 4.12a). Despite slight differences in quaternary structure, all-atom RMSD structure alignments between individual C-terminal domain fold, reveals that excellent structural homology exists between MCM of different phylogenetic domains (Figure 4.12b). The high all-atom RMSD is expected with the high degree of shared sequence conservation.

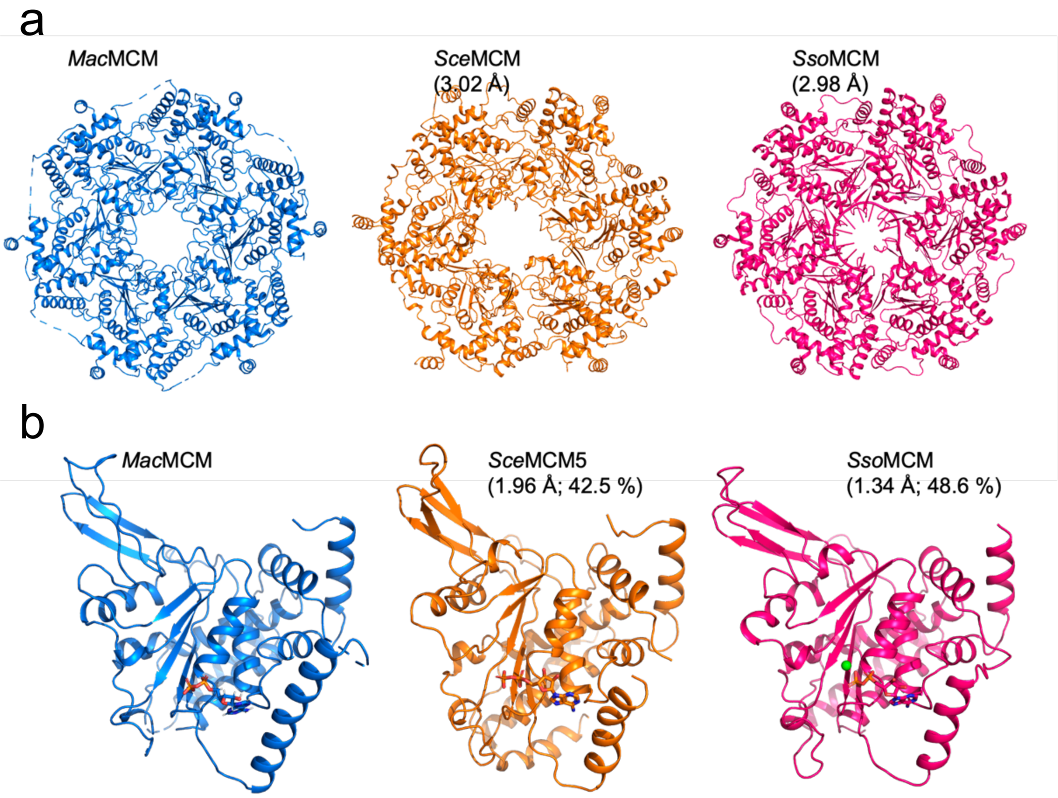


Figure 4.12: Comparison of MCM C-terminal domains.
C-terminal domains of *Mac*MCM (this study, blue), *Sce*MCM (PDB:6EYC, orange)[140] and *Sso*MCM (PDB: 6MII, magenta)[100], are represented in the ribbon format. **(a)** C-terminal domain hexamers were compared using the PyMol structure-based alignment command (cealign). Values in parentheses represent alpha-carbon RMSD to *Mac*MCM. **(b)** Single C-terminal domains were compared using the PyMol sequence-based structural alignment command, 'align'. Values in parentheses represent all-atom RMSD (Å) and sequence identity (%: Clustal Omega) to *Mac*MCM.

4.4.3 Comparison of conserved interfaces in *Mac*MCM with archaeal and eukaryotic homologues

4.4.3.1 Buried surface area; POPS analysis

Conservation of interface residues was assessed using Parameter OPtimised Solvent accessibility (POPS), MUSCLE sequence alignment, and a custom R script that utilises the Bio3D package [230,231,279]. Briefly, solvent accessibility was calculated for a 1.4 Å probe around atoms on each subunit alone, then again in the presence of its neighbouring subunit. Due to the low resolution of selected structures (*Sso*MCM = 3.15 Å; *Sce*MCM = 3.8 Å), solvent accessibility calculations were performed on residues, instead of individual atoms. The difference in solvent accessibility of residues was then calculated by subtracting the solvent accessibility when subunits are interfaced from each subunit measured alone. Accessibilities were filtered to keep residues where change is observed at >1 interfaces. When compared against eukaryotic MCM, this filter helps to remove subunit specific interfaces (such as insertions) and allows a focus on core, shared residues. Residues are then labelled based on whether the residue belongs to the *cis*- or *trans*- side of the subunit interface. The '*cis*-' subunit contributes Walker A, B and sensor-1 to the ATPase active site, whilst the '*trans*-' subunit contributes sensor-2 and the arginine finger [99]. Residue numbers are then adjusted to their equivalent position in a sequence alignment generated by MUSCLE.

Comparison and alignment between *Mac*MCM, *Sso*MCM and *Sce*MCM reveals reasonable visual conservation between the regions where residues are buried at interfaces (Figure 4.13a). The absence of buried ZnF residues for *Mac*MCM is again highlighted relative to both *Sso* and *Sce*MCM. Across the three structures, *Mac*MCM and *Sce*MCM share the largest number of buried residues at equivalent positions in the MSA (45.6 %), this is followed by: *Mac*MCM and *Sso*MCM (25.2 %); *Sce*MCM and *Sso*MCM (20.7%). This compares reasonably with the values for solvent accessible interface correlation (Figure 4.13b). In this calculation, the solvent accessibility of each residue in an interface is compared at each position in the MSA and correlation analysis is performed. Values where solvent accessibility is observed in only one of the two proteins are retained (i.e., correlation points that lie on the x and y-axes) to avoid biasing correlation values (Supplementary Figure 7.6). Residues in *Mac*MCM and *Sce*MCM share the largest correlation (PMCC=0.60), followed by: *Sso*MCM and *Sce*MCM (PMCC=0.37); *Mac*MCM and *Sso*MCM (PMCC=0.29). *Mac*MCM was then compared against all individual *Sce*MCM interfaces, however, there is no clear interface with a highest shared similarity, with all correlations ranging between 0.54 and 0.59 (Supplementary Figure 7.7).

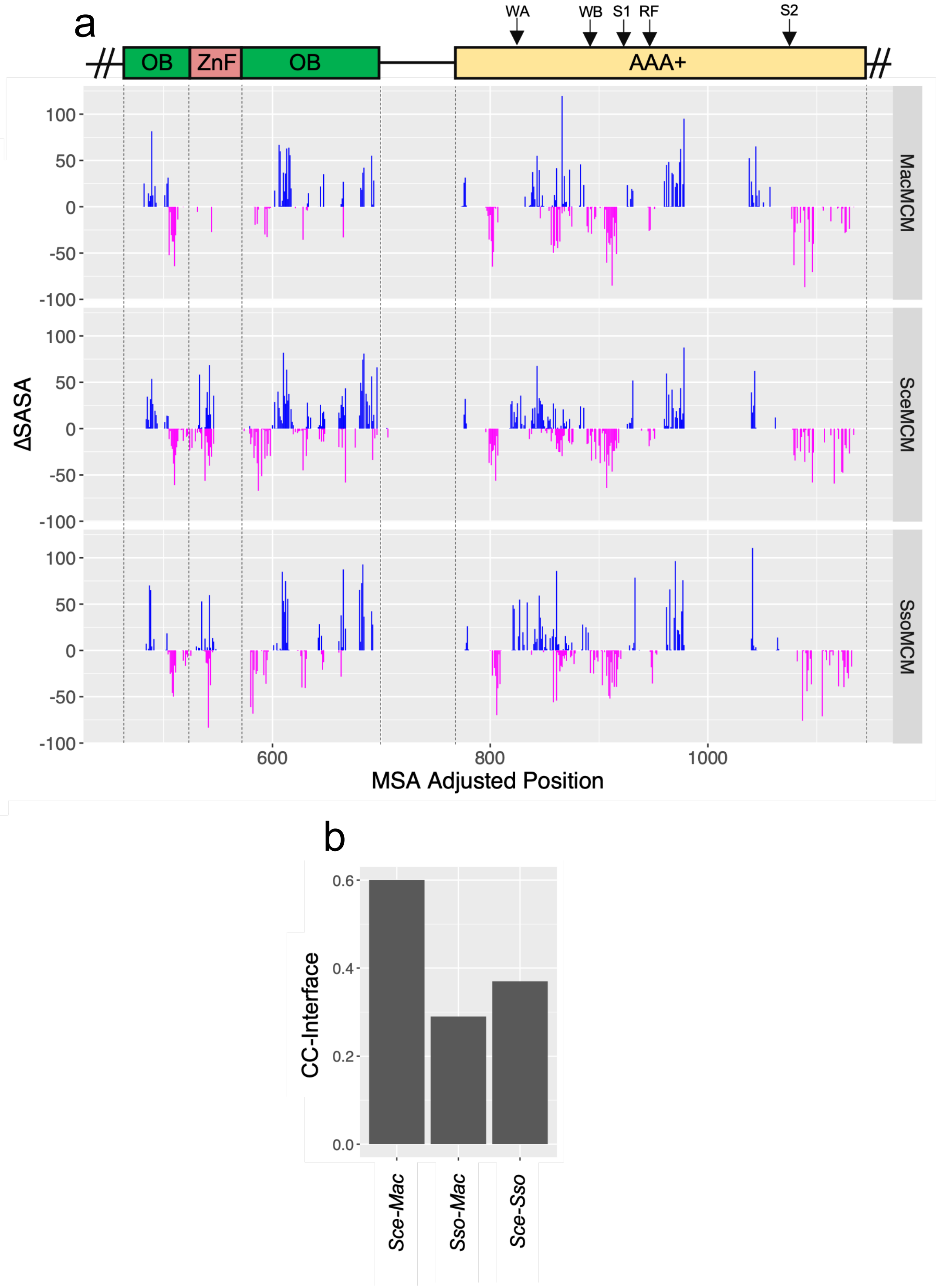


Figure 4.13: Correlation of buried subunit interfaces.
(a) The change in solvent accessible surface area (ΔSASA) was determined for residues in each MCM: *Mac*MCM (this study), *Sce*MCM (PDB: 6EYC)[140] and *Sso*MCM (PDB: 6MII)[100]. Residues were then determined to be either *cis*- (+ve, blue) or *trans*-acting ('-ve', magenta), based on the motifs supplied by the subunit active site. Residues that occur <2 times were removed and the average ΔSASA was calculated at each position. The SASA profile for each protein was then visualised using ggplot2. **(b)** The similarities of two SASA profiles were then directly compared by performing a Pearson correlation coefficient (PMCC) analysis between ΔSASA for residues at equivalent positions in the MSA.

4.4.3.2 Interface residue composition; PISA analysis

Whilst clear disparities lie between the interfaces of the enzymes, it was decided to determine the composition of interfaces within each of the hexameric MCM structures. Mesophilic enzymes are widely regarded to form assemblies with more hydrophobic interfaces than thermophilic homologues [280]. As a trade-off, thermophilic enzymes generally form more salt-bridges [281].

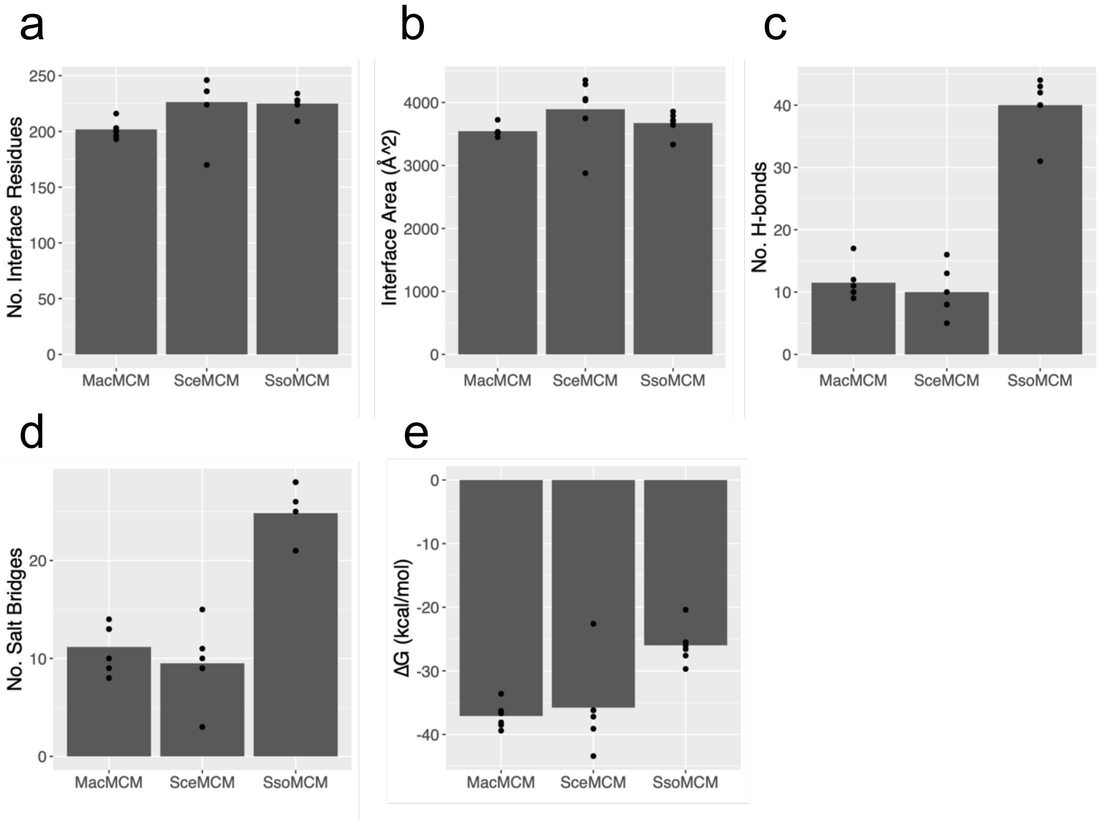


Figure 4.14: PISA analysis of interface composition.
Structures were imported into the PDBePISA (Proteins, Interfaces, Structures and Assemblies) portal and analysed. Data were then exported and averaged for each metric and MCM. *Mac*MCM (this study), *Sso*MCM (PDB: 6MII)[100], *Sce*MCM (PDB:6EYC)[140]. Black points represent raw subunit-subunit values. **(a)** Average number of interface residues between MCM interfaces. **(b)** Average buried surface area between MCM interfaces. **(c)** Average number of hydrogen bonds formed between MCM interfaces. **(d)** Average number of salt bridges formed between MCM interfaces. **(e)** Estimated average solvation free energy (kcal/mol) gain of interface formation (this only includes the contribution of hydrophobic residues).

Analysis was performed on each enzyme using the PISA server (https://www.ebi.ac.uk/pdbe/pisa/), which suggests that all 3 MCM have equivalent numbers of residues within interfaces (Figure 4.14a). Equally, the total interface sizes are equivalent despite eukaryotic MCM being composed of larger subunits (Figure 4.14b).

However, investigation into the composition of interfaces reveals inconsistencies between MCM from organisms of different environments. Both *Mac*MCM and *Sce*MCM interfaces are composed of fewer hydrogen bonds and salt-bridges than *Sso*MCM (Figures 4.14c-d). Comparatively both *Mac*MCM and *Sce*MCM interfaces are composed of more hydrophobic interactions, as highlighted by their low estimated solvation free energy of interface formation (Δ^{i}G) values (Figure 4.14e). Where interfaces are hydrophobic, the bound state is much lower energy than the unbound state, hence Δ^{i}G is more negative for hydrophobic interfaces. It is difficult to split interface roles of residues further through PISA analysis between N-terminal and C-terminal domains since there is significant structural overlap between domains.

4.5 Analysis of *Mac*MCM DNA binding interface

In the previous chapter, we observed that both ATP hydrolysis and DNA binding are important for *Mac*MCM assembly into a hexamer. The coordination of a phosphate ion by 3 basic residues within the OB-fold is significant as it is likely analogous to interactions made with the phosphate backbone of DNA. It is possible that DNA neutralises the local positive charge and stabilises *Mac*MCM hexamer formation. Furthermore, the 3-phosphate co-ordinating basic residues (R124, K176, R224) share structural homology with residues that were implicated in ssDNA binding by *Pfu*MCM (K129, K186, R233) [128] (Figure 4.15a–b). Mutation of just two of the homologous basic residues *Sce*MCM is lethal for cells and when tested *in vitro*, enzymes exhibit severe DNA loading defects [128]. Since we hypothesize that

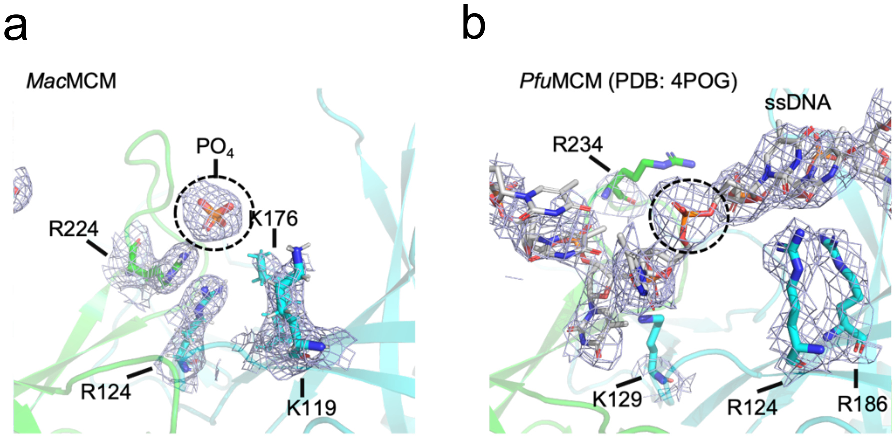


Figure 4.15: Analysis of N-terminal domain (NTD) DNA-binding residues.
N-terminal domains as viewed from below the NTD hexamer, visualized as in the cartoon format in PyMol. Colours represent neighbouring subunits. Phosphate/DNA/DNA-binding residues are visualized with sticks, surrounded by the electron density map (2F$_o$ -F$_c$), contoured at σ = 2.0. Core/equivalent residues for **(a)** *Mac*MCM and **(b)** *Pfu*MCM (PDB:4POG)[128] are annotated. The position of equivalent phosphate residues are highlighted with a circle (dotted line). DNA/phosphate sticks are colored by element.

*Mac*MCM lag time is a proxy for loading onto DNA, it was decided to perform mutagenesis of homologous basic residues and measure activity.

4.5.1 Selection of putative *Mac*MCM DNA-loading mutants

The three basic phosphate coordinating residues (R124, K176 and R224) were selected for mutagenesis. A double arginine mutant (R124-R224) was also selected to investigate the contribution of basic residues from neighbouring subunits for activity. Aside from the phosphate coordinating residues, a further basic residue within the OB-fold (K119) was also chosen for mutagenesis. Although *Mac*MCM K119 does not co-ordinate the phosphate ion in our structure, the structurally equivalent R124 in *Pfu*MCM was demonstrated to coordinate ssDNA [128] (Figure 4.15). For further comparison, *SsoPfu*MCM K130 (homologous to *Mac*MCM K119) was also chosen to examine the impact of DNA-binding mutations on an obligate hexameric MCM. All amino acids were mutated to alanine to neutralise the positive charge.

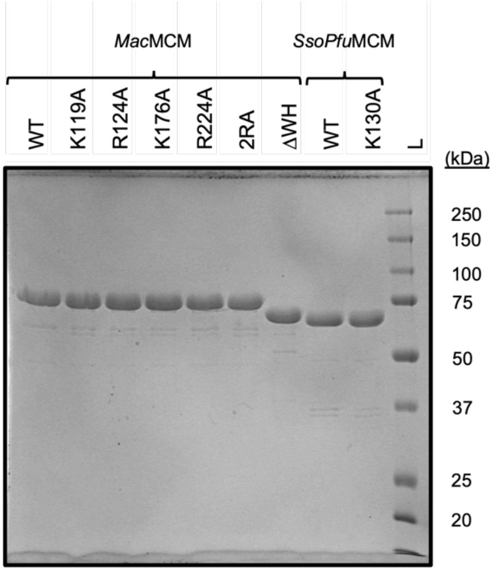


Figure 4.16: *Mac*MCM DNA-binding mutants are pure.
Purity analysis of DNA-binding mutants. 3 µg of each protein was run on a 12 % (*w/v*) SDS-PAGE and visualized with a Coomassie stain. L: molecular weight ladder.

4.5.2 Analysis of Mutants

Mutagenesis was performed using a Lightning QuikChange kit (Agilent) and validated through DNA sequencing. All proteins were overproduced and purified as before to near homogeneity. Purity was assessed by SDS-PAGE (Figure 4.16) and the absence of DNA-contamination was confirmed through measurement of A_{260}:A_{280} ratios. As charged, basic amino acids have been mutated proximal to subunit interfaces, all mutants were assessed for changes in oligomeric state. It is reasonable to consider that charge neutralisation near subunit interfaces may alter inter-subunit interactions. However, assessment of oligomerization by SEC confirms that substitution of these basic amino acids does not alter the elution volume for either *Mac*MCM or *SsoPfu*MCM variants (Figure 4.17). Based on data outlined in chapter 3, this implies that all *Mac*MCM mutants are monomeric, whilst the *SsoPfu*MCM mutant is a hexamer.

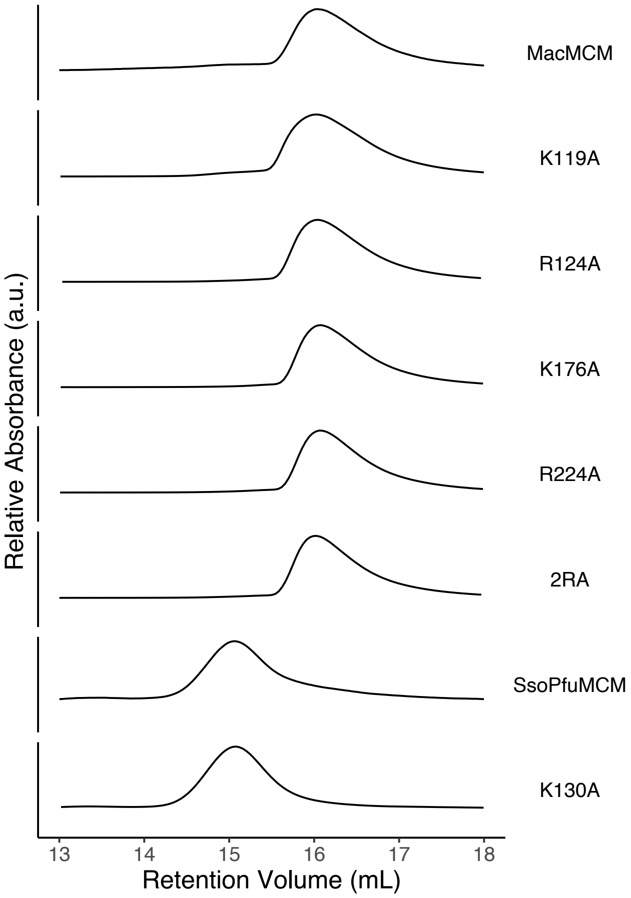


Figure 4.17: *Mac*MCM DNA-binding mutants do not have altered oligomeric state.
100 µL at 10 µM (hexamer) of each protein sample was passed over a Superose 6 10/300 GL Increase chromatography column. The column was pre-equilibrated in a buffer containing 200 mM NaCl, 20 mM Tris-Cl pH 8.0, 5 % (*v/v*) glycerol. Absorbance was measured at A_{280}. and standardized to the maximum absorbance of each experimental trace.

4.5.3 Assessment of DNA-binding mutants; EMSA

The ability of each MCM construct to bind to a forked DNA substrate was then assessed by EMSA (see section 2.3.2). For each tested *Mac*MCM variant, no DNA-binding could be observed at any of the tested concentrations of protein (Supplementary Figure 7.8). In all instances this represents a reduction in DNA-binding relative to the wild-type enzyme in absence of ligands. Equally, the *SsoPfu*MCM K130A mutation, results in a ~2-fold reduction in DNA binding with respect to the wild-type enzyme (Figure 4.18a–b). Whilst the OB-fold is primarily a ssDNA binding motif, the forked substrate also contains dsDNA that the enzyme may be able to interact with.

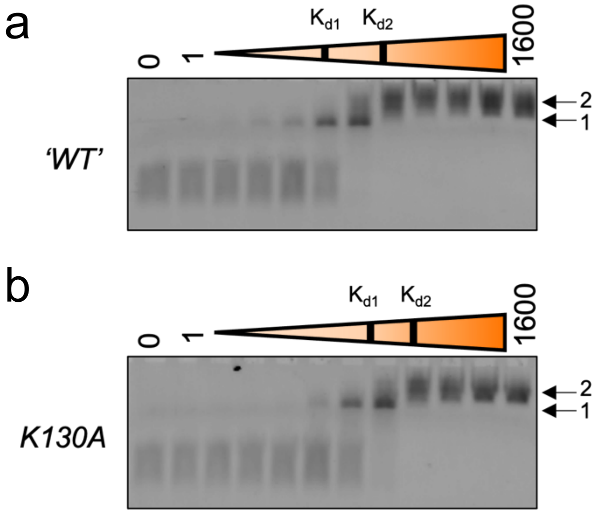


Figure 4.18: DNA binding affinity of *SsoPfu*MCM OB-fold mutants.
The binding affinity of selected MCM constructs was determined by EMSA. MCM were serially diluted and mixed with 10 nM forked DNA to the stated concentration. Samples were then run on a 0.8 % (*w/v*) TB agarose gel. Gels were visualized on a Typhoon gel scanner (GE healthcare). Binding affinity was then estimated for each complex for **(a)** WT *SsoPfu*MCM: K_{d1} = 25 nM, K_{d2} = 75 nM, **(b)** *SsoPfu*MCM^{K130A}: K_{d1} = 50 nM, K_{d2} = 150 nM. K_{d1} represents the estimated association constant of the MCM-DNA complex with motility '1'. K_{d2} represents the estimated association constant of the MCM-DNA complex with motility '2'.

4.5.4 Assessment of DNA-binding mutants; Helicase unwinding

The activity of each MCM variant was then assessed through a fluorescent helicase assay. The capacity of each MCM variant to unwind a forked DNA substrate was assessed at the 30 minute time point (Figure 4.19a–b). Single point mutation of residues R124A, K176A and R224A in the *Mac*MCM structure exhibit a slight (~10%) reduction in activity over the time course. The double arginine mutant exhibits a larger ~30% reduction (Figure 4.19a). The most drastic reductions in activity were observed for the analogous K119A (*Mac*MCM) and K130A (*SsoPfu*MCM) mutations. In yeast, residues with structural homology to K119 were demonstrated to mediate MCM2-7 loading onto DNA [128]. MCM5 is the only yeast subunit not to contain a basic residue at this position [128].

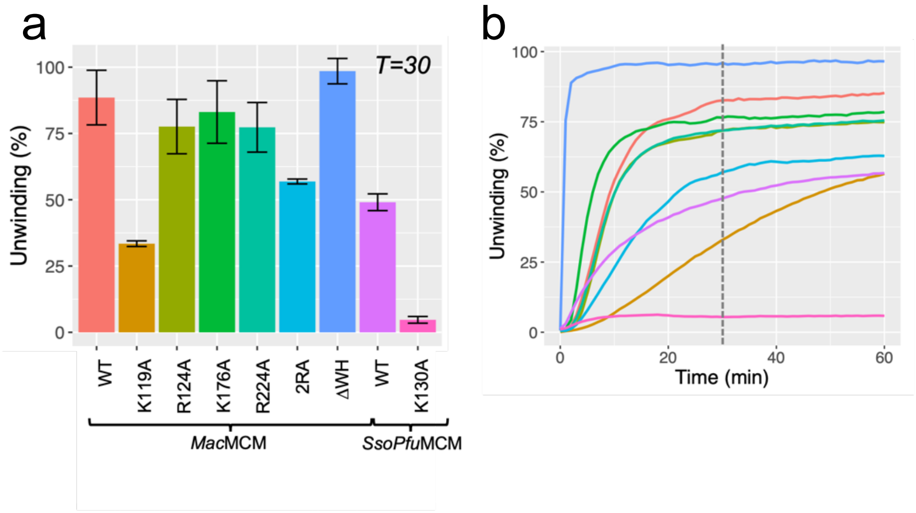


Figure 4.19: Forked DNA unwinding ability of OB-fold mutants.
DNA turnover of a forked 26 base pair DNA substrate was measured in a 96-well plate fluorescent helicase assay. After 30 minutes of incubation, ATP/Mg^{2+} was added to a final concentration of 4/10 mM. DNA turnover was monitored through changes in fluorescence over the stated concentration of time on a Clariostar plate reader (BMG Labtech). Data points were standardized to a maximum fluorescence control substrate. **(a)** The net DNA unwinding for 1000 nM MCM (Hexamer) was determined after 30 minutes. Error bars represent +/- 1 SEM, n=3. **(b)** The real time unwinding data for 1 experimental repeat. Dotted line represents T=30. Colours are as part (a).

4.5.5 Assessment of DNA-binding mutants; lag time, a proxy for *Mac*MCM DNA loading?

We consider that lag time observed in *Mac*MCM^{FL} helicase assays is a proxy for the kinetics of MCM-DNA loading. This is consistent with a growing body of research that suggests eukaryotic MCM takes ~5-10 minutes to load onto DNA [59,60,157]. Such an observation has never been made for an archaeal MCM. Previous work implicates basic amino acids within the central channel to regulate loading of eukaryotic MCM onto DNA [128]. One dimensional net unwinding data only suggests how active an enzyme is over a given time. However, net unwinding for MCM is derived from both assembly of the enzyme onto the substrate followed by ATP dependent unwinding. For enzymes such as *Mac*MCM$^{\Delta WHD}$ and

*SsoPfu*MCM, contribution of assembly time is assumed to be minimal under any of the conditions tested in this thesis; lag time is not observed under any circumstance. However, as demonstrated in chapter 3, when the ATP or protein concentration is decreased for *Mac*MCM^{FL}, lag time increases exponentially. It was suggested in the order of addition experiments and analytical SEC, that ATP and DNA binding are required to cooperatively load an active *Mac*MCM^{FL} hexamer onto DNA, like eukaryotic MCM2-7. Pre-incubation of the enzyme with either ATP or DNA alone has no effect on the lag time. To further study the role of DNA binding, it was decided to investigate:

- Whether loss of DNA-binding infringes on the lag time;
- Whether assembly kinetics relate directly to the net activity of the enzyme.

The activity of serially diluted *Mac*MCM^{FL} and DNA binding mutants was then investigated over 120 minutes. Lag time was calculated for each protein concentration as previously outlined (section 3.6.3). The protein concentration dependence of lag time for each mutant was then fit to a standard exponential decay model (Figure 4.20a-f) (see section 2.5.5.3). Within this model, the decay coefficient β, describes the dependency of lag time on protein concentration. A large β describes an enzyme where lag time has a lower dependency on the protein concentration. A small β describes an enzyme where lag time has a higher dependency on the protein concentration.

Relative to the wild type *Mac*MCM^{FL} enzyme, each DNA binding mutant exhibits a decrease in the size of β (Figure 4.20g). The largest decreases in β are observed for 2RA and K119A, implying that lag time measured in these enzymes has a greater dependency on protein concentration. This suggests that interactions with DNA by these residues are particularly important in the association equilibrium of *Mac*MCM^{FL} DNA-bound hexamers.

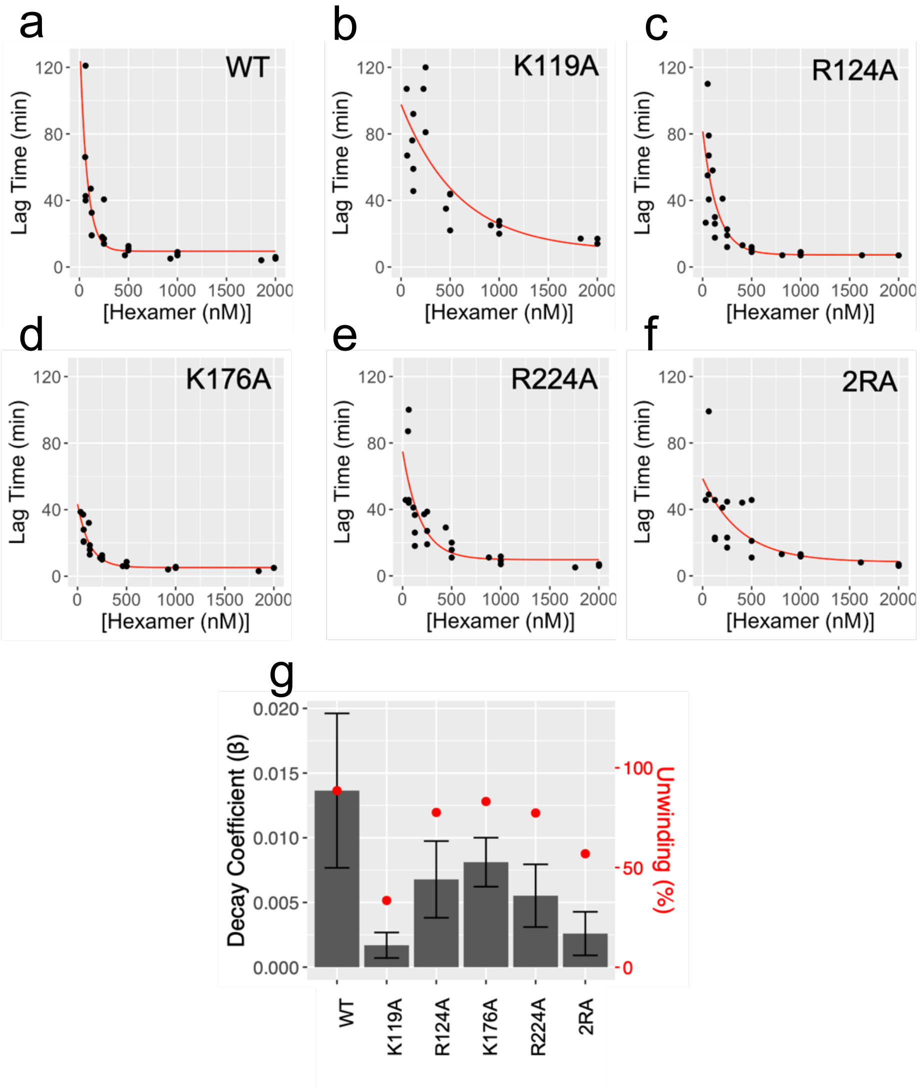


Figure 4.20: Exponential fit analysis of *Mac*MCM^{FL} mutant lag time.
(a-f) DNA turnover of a forked 26 base pair DNA substrate was measured in a 96-well plate fluorescent helicase assay. After 30 minutes incubation at 25 °C, ATP/Mg^{2+} was added to a final concentration of 4/10 mM. DNA turnover was monitored through changes in fluorescence over 120 minutes on a Clariostar plate reader (BMG Labtech). The first derivative of each experimental unwinding trace is calculated and then the time taken to reach the maximum first derivative is extracted. Extracted lag time is then plotted against hexamer concentration for each mutant. Data were then fit to an exponential decay model: Lag Time = $Ae^{\beta[MCM]} + \Theta$. **(g)** The exponential decay coefficient β compared against the net unwinding data from Figure 4.19a. Error bars represent +/-1 standard error of the fit.

The data trend for β coefficients across variants matches the trend measured for net unwinding activity (e.g., the largest difference is observed between *Mac*MCM^{FL} and K119A). Therefore, β coefficients roughly define how active a *Mac*MCM^{FL} variant will be (Figure 4.20g). This is not perhaps surprising as the proportion of *Mac*MCM^{FL} successfully loaded onto the substrate at a given protein concentration will define the number of active complexes. It is difficult to speculate a *Mac*MCM^{FL} mutation where activity is reduced and the β coefficients is unaffected, since DNA-binding, ATP hydrolysis and oligomerization all appear to be critical for defining *Mac*MCM^{FL} assembly. ATP-dependent *Mac*MCM^{FL} hexamer assembly onto DNA is clearly global and cooperative process.

4.6 Crystallization of ssDNA bound *Mac*MCM$^{\Delta WHD.E418Q}$

Protein-ATP-DNA complex was prepared by addition of ATP and $MgCl_2$ to final concentrations of 10 mM. Single-stranded DNA $(ACTG)_{16}$ was added at a final molar ratio of 1:1.2 (MCM hexamer: DNA). Screens were set up as previously described. Crystals were obtained in 4 conditions in both PACT and PDB screens. Small crystals were observed in condition F10 (0.2 M Sodium Malonate, 0.1 M Bis-Tris Propane pH 6.5, 20 % (*w/v*) PEG 3,350) of the PACT screen (Figure 4.21a). These crystals generated 5 Å diffraction at Diamond with minimal ice issues. Preliminary MR suggested presence of an MCM double-hexamer within 1 ASU, with promising statistics. Crystals were further optimized to improve diffraction by buffer optimization and increasing the drop size to 20 µL. Large crystals in the unoptimized condition generated the best ~2.7 Å diffraction. Analysis of this dataset revealed that the crystal isoform was identical to the *Mac*MCM$^{\Delta WHD-E418Q}$ ADP complex solved previously (Table 4.4). Phasing performed on this dataset using the solved *Mac*MCM$^{\Delta WHD-E418Q}$ structure yielded a near perfect MR solution. Further refinement of this MR solution suggested that DNA was absent from the complex (Figure 4.21b).

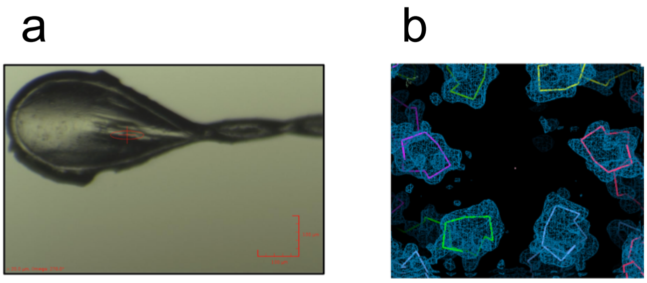


Figure 4.21: Crystal attempts of *Mac*MCM$^{\Delta WHD-E418Q}$ + ATP/Mg^{2+} + ssDNA.
(a) Large crystals grown in 0.2 M Sodium Malonate, 0.1 M Bis-Tris Propane pH 6.5, 20 % (*w/v*) PEG 3,350 at 20 °C, immobilized in a cryo-loop without cryoprotection. **(b)** Electron density of a vacant channel of hexameric *Mac*MCM$^{\Delta WHD-E418Q}$. Data were phased perfectly by molecular replacement, using a search model of 1 copy of the previously solved hexameric *Mac*MCM$^{\Delta WHD-E418Q}$-ADP structure. Coloured chains are visualized as a C-alpha trace in Coot.

Table 4.4: Data collection statistics for *Mac*MCM$^{\Delta WHD\text{-}E418Q\text{-}ssDNA}$.
Values in parentheses outline equivalent data for the highest resolution shell. DLS: Diamond Light Source. CC: correlation coefficient. For the formulas of R_{meas} and $CC_{1/2}$, see section 2.8.

Parameters and statistics	
Data collection	
Beamline	I04, DLS
Wavelength (Å)	0.9795
Resolution (Å)	2.64 – 56.65 (2.64– 2.69)
Cell dimensions	
Space group	C 1 2 1
Unit-cell a,b,c (Å)	230.13, 126.59, 168,69
Unit-cell α, β, γ (°)	90.00, 104.35, 90.00
Data merging statistics	
Unique reflections	947,618 (51,238)
Completeness (%)	99.9 (99.0)
Multiplicity	6.9 (7.3)
I/σ <I>	1.5 (1.5)
R_{meas} (%)	1.27 (1.72)
$CC_{1/2}$	0.8 (0.5)
Wilson B-factor	141.7

4.7 Conclusions for this chapter

This chapter addresses the first structural characterization of a near full length archaeal MCM from mesophilic archaea. Preliminary crystallization trials suggested that solving the structure of the apo full-length enzyme would be difficult. Logical optimization of crystallization using well established routes including truncation of flexible regions and addition of ligands, allowed the determination of a 2.6 Å resolution crystal structure of homohexameric *Mac*MCM. In depth analyses of the structure with respect to previously solved MCMs, implies that the intersubunit interfaces of *Mac*MCM are possibly more like eukaryotic MCM than MCM from archaea thermophiles. Observed similarities at interfaces extend in terms of ZnF asymmetry, residue accessibility and chemical composition. However, definitive conclusions are somewhat limited by the low number of comparable MCM structures available in the PDB. Future studies may deploy dynamic structural techniques such as cryo-EM to permit elucidation of the apo and DNA-bound forms of the enzyme. Already, such studies will benefit greatly from the high-resolution crystal structure and will be discussed in further detail in the discussion.

Real time helicase assays performed on *Mac*MCM suggest an important role of DNA binding residues in the slow kinetic ‘lag time’ step observed for *Mac*MCM. Removal of key basic residues within the central channel of the N-terminal domain increases the dependency of protein concentration in lag time. It would also be interesting to determine the affinity of each mutant after the addition of ATP, for example, does it limit the complexes binding affinity, rate of complex formation or both? It is possible that co-operative DNA binding by neighbouring subunits contributes to the stability of a hexamer.

Chapter 5 – Biotechnological Applications of MCM in Nanopore Sequencing

5.1 Introduction

The order of just 4 nucleic acids (A, C, T and G) provides the genetic information for understanding the diversity of all life on earth. And yet despite the simplicity of DNA, efficiently decoding the simple genetic sequence has required decades of research. Today DNA sequencing is a central technology to multiple fields, including: medicine, biotechnology, ecology, food safety, forensics and archaeology [282–286]. In 2020 we saw DNA sequencing become a central tool in the COVID-19 response, permitting simultaneous pandemic monitoring at the level of molecules, individuals and populations [287]. Sequence level understanding of diseases will enhance the development of emerging personalised medicines, such as mRNA therapeutics [288,289]. In the future we will see expansion of sequencing to facilitate new DNA-based technologies, including information storage and anti-fraud tags [290,291].

This chapter will outline the brief technical history of DNA sequencing technologies and their limitations, the emergence of Nanopore sequencing and its commercialization by Oxford Nanopore Technologies (ONT). I will then discuss how alternative enzymes such as MCMs could be used to further streamline ONTs sequencing platform and describe preliminary experiments to assess viability.

5.2 The dawn of DNA sequencing

Two emergent techniques in the 1970s would demonstrate the feasibility and lay the foundation for decades of DNA sequencing technologies. These techniques were the Maxam-Gilbert and Sanger sequencing methods [292,293]. Gilbert sequencing requires chemical modification of bases to generate base specific cleavages of a radiolabelled DNA-restriction fragment. Sanger sequencing utilises a DNA polymerase to perform 4 independent extension reactions of a primer in the presence of an inhibitory, chain-terminating nucleotide. Chain terminators lack a reactive 3′ hydroxyl group, and hence cannot form a phosphodiester bond with subsequent native nucleotides [294]. When chain terminating nucleotides are present in trace amounts, they are incorporated stochastically but importantly in a base specific fashion thereby producing fragments of discrete lengths. Both techniques generate sequence-specific fragments that can be electrophoretically separated by size on a polyacrylamide gel. Either the primers or nucleotides are radiolabelled to allow visualization of fragments by autoradiography. These techniques define the first generation of DNA sequencing.

The Sanger sequencing method was substantially developed and commercialised over the subsequent decade. Advances in nucleotide chemistry replaced radiolabelling strategies with base-specific fluorescent terminators that allowed discrimination of all nucleobases in a single reaction vessel (Figure 5.1) [294–297]. This led to the development of the first commercially available DNA sequencer by Applied Biosystems in 1986. Time consuming gel-based electrophoretic techniques were eventually replaced by capillary electrophoresis, allowing sequential detection of fluorescent fragments by size [298,299]. The Sanger method is limited to sequencing fragments <1 kb in size; beyond 1 kb, the spatial resolution of bases becomes poor [300]. To sequence large genomes in excess of this limit (i.e., Gb) the ‘Shotgun’ approach was devised. Here fragments or contigs of a genome are sequenced then reassembled *in silico* using complex algorithms to detect overlapping sequences [301]. The Sanger method demands high concentrations of starting material and benefited from the discovery of PCR, where high quantities of specific DNA can be produced *in vitro* [302]. Whilst highly accurate (>99.9%), Sanger sequencing is expensive, low-throughput and not particularly suited for large scale genome projects [303]. It is still widely used in many laboratories when high throughput is not required, for example as a validation step in molecular cloning (see section 2.2.2).

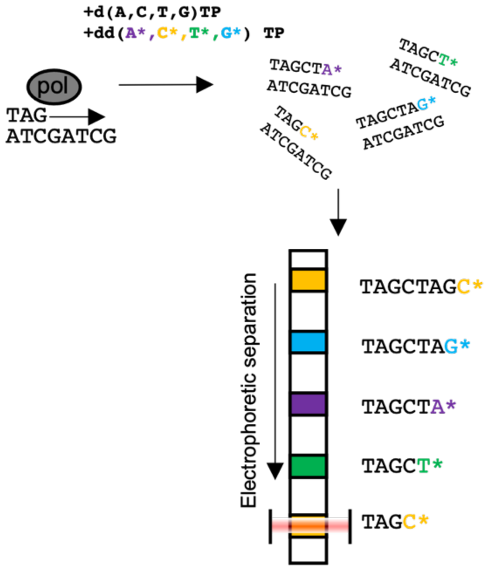


Figure 5.1: Core chemistry of commercial Sanger sequencing (Applied Biosciences)
Chain extension of a primed DNA sequence is catalyzed by a DNA polymerase in the presence of trace amounts of inhibitory fluorescent nucleotide analogues. Sanger sequencing generates fragments of increasing sequence length, which are separable by electrophoretic techniques.

5.3 Second generation high throughput sequencing technologies

Second generation sequencing technologies were primarily designed to address throughput issues. Various strategies were devised, all commonly replacing electrophoresis-based detection. Technologies substituted single contig reactions for mass parallelization, where multiple fragments are sequenced simultaneously in the same vessel. Pyrosequencing (licensed initially by 454) largely dominated the early years of this generation but was later surpassed by the Illumina (formerly Sodexa) method.

5.3.1 454 pyrosequencing

In pyrosequencing, a single DNA template is immobilised onto a bead. Emulsion PCR is then performed, amplifying the copies of the immobilized template on the bead [304]. This process is massively parallel, producing millions of monoclonal beads. Beads are then loaded into individual wells in a PicoTiterPlate; well sizing does not permit more than one bead per well [305] (Figure 5.2). Each nucleotide is isocratically washed over the entire plate in turn [306]. When the complementary nucleotide is supplied, a polymerase catalyses strand extension and in turn releases pyrophosphate. ATP sulfurylase rapidly converts pyrophosphate into ATP, which in turn is converted into light by a luciferase enzyme[306]. The intensity of light generated in each well is proportional to the number of bases incorporated to the sequence. A charge coupled device (CCD) determines the emission of light generated in each well. Typically read lengths generated by this technique are between 1-100 bases long, however reads are generated for >1 million wells simultaneously, representing a vast

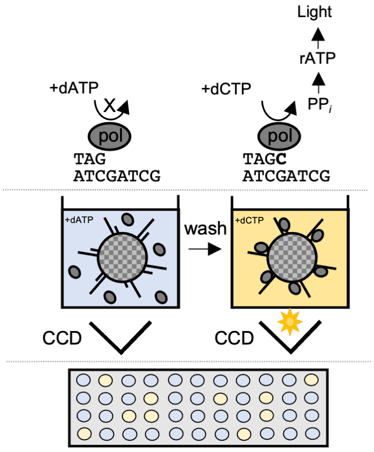


Figure 5.2: Core chemistry of pyrosequencing (454).
Fragmented DNA libraries are immobilized onto beads and clonally amplified through emulsion PCR. Monoclonal beads are isolated in single wells on a PicoTitre plate. Solution containing individual nucleotides are washed over the plate in turn. When the correct nucleotide is suppled, strand extension is catalyzed by a DNA polymerase, releasing pyrophosphate (PP_i). PP_i production is quantified by an enzyme-coupled luciferase assay that produces light, detectably by a charge-coupled device (CCD).

improvement of the Sanger technique [307]. Like sanger sequencing, single read accuracy is high, around 99.5 % [308]. This methodology was eventually commercialised by 454.

5.3.2 Illumina sequencing

The Illumina platform to an extent represents an advancement on the chain termination chemistry of Sanger sequencing. First, defined clusters (~2,000 strands) of immobilised clonal DNA are coated onto the surface of a glass flow cell via isothermal amplification. Strand extension of DNA immobilised on a glass slide is performed by a polymerase which catalyses the addition of a single nucleotide analogue (Figure 5.3)[309]. Added nucleotides are reversibly labelled with a unique fluorophore that occupies the position of the reactive 3' hydroxyl group [310,311]. After each extension cycle, nucleotides are washed away, and a CCD is used to determine the base incorporation at each cluster within a flow cell. The fluorophore is then cleaved, exposing the 3' hydroxyl for subsequent extension cycles. This cyclic process is repeated across ~200 million clusters in a flow cell, generating reads of 50-300 bases long. Like the previously discussed methods, read accuracies are very high (99.8%) [303].

Except for the Maxam-Gilbert method, all technologies mentioned so far are dependent on the synthesis of DNA molecules to generate identifiable signals. However, there are drawbacks to synthesis-based approaches. Read lengths of all discussed techniques are low (<1 kbp), and therefore when performing genomic analysis require extensive reassembly *in silico* [312]. Long read lengths favour ease of reassembly and require lower computational power. Many portions within genomes are also unmappable with short read technologies,

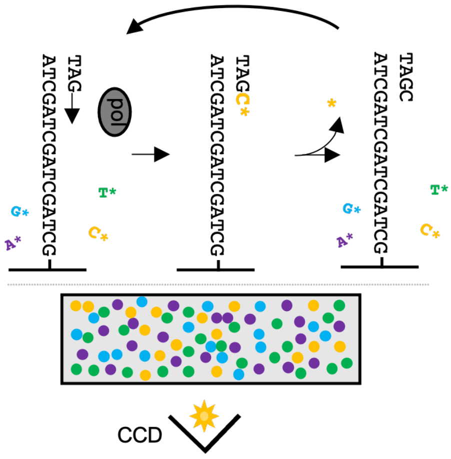


Figure 5.3: Core chemistry of Illumina sequencing.
Clusters of clonal DNA are generated onto the surface of a proprietary flow cell. A DNA polymerase catalyzes single-base strand extension through addition of a distinct fluorescent nucleotide. Fluorescence of each cluster is detected with a charge-coupled device (CCD) and the fluorophore is removed. This permits further cycles of strand extension.

this includes regions of large structural variation, high repeat and high GC contents [313]. Amplification of DNA can also incorporate biases and polymerase mediated errors into the sequence [314–316]. Information about epigenetic modifications such as cytosine methylation is lost without targeted bisulfite preparation [317]. Next generation sequencing technologies aim to address these issues.

5.4 Third generation sequencing technologies

Two single molecule-based long read technologies have emerged at the forefront of DNA sequencing in the last decade: Pacific Bioscience (PacBio) and Nanopore (ONT) sequencing.

5.4.1 Pacific Biosciences (PacBio) sequencing

In Pacific Biosciences (PacBio) sequencing, a single polymerase enzyme is immobilised into wells on a commercial 'SMRT™ chip' (Figure 5.4). After addition of DNA and discretely labelled fluorescent nucleotide analogues, a zero-mode waveguide (ZMWs) is able to directly monitor strand extension by the polymerase [318,319]. ZMWs can illuminate a minute volume that only contains the polymerase and its substrate. Whilst free nucleotides diffuse rapidly, polymerase-bound fluorescent nucleotides are held within the detection volume long enough for base discrimination [320]. As with Illumina sequencing, the terminator group is cleaved, and strand extension can continue. For PacBio sequencing, read lengths are generally <60 kb, where the key limitation is the lifetime of the immobilised polymerase [321]. Whilst the accuracy of a single read is relatively poor (~90 %), the error is believed to be random [322]. It is important to note that the accuracy of previous reads was determined from a population weighted signal through which random error is averaged out. In single molecule sequencing, addition of multiple single reads can also reduce random error

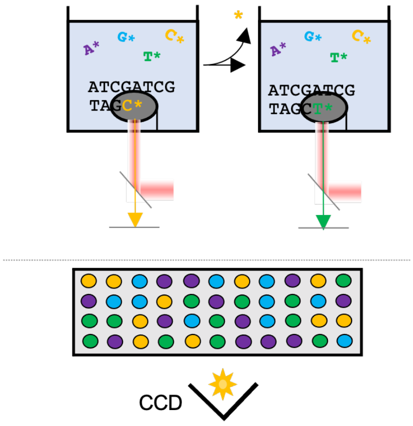


Figure 5.4: Core chemistry of Pacific Bioscience (PacBio) sequencing.
A DNA polymerase is immobilized to the bottom of a well. Strand extension through addition of fluorescent nucleotides is directly monitored by a Zero Mode Waveguide (ZMW).

through averaging. Each SMRT™ chip contains ~150,000 wells, of which up to half will generate data [320].

5.4.2 Nanopore sequencing

Each discussed technology so far has been reliant on an optical measurement of DNA synthesis. In the 1980s, David Deamer outlined an alternative framework to sequence DNA based on the use of a Coulter device [323]. A Coulter device contains an electrolytic solution with two reservoirs connected by a narrow aperture over which a constant electric field is applied. Coulter devices were initially designed for counting erythrocytes [324]. As particles such as erythrocytes pass through and occlude the aperture, electrolyte is displaced generating detectable changes in electrical resistance proportional to the size of the molecule (Figure 5.5a-b). Deamer hypothesized that fabrication of a small enough aperture or a 'Nanopore', may allow discrimination of electrophoretically motile nucleobases. Discrimination is partially limited by the 'sensing region' [325]. The sensing region of a nanopore outlines the number of bases that fit within the aperture to contribute to the signal. This is where the largest relative proportion of electrolyte is displaced in the current flow and hence dominates the signal. A long sensing region may incorporate too many bases for clear base discrimination. The use of nanopores in sequencing has been explored using two main strategies that apply solid-state and biologically fabricated nanopores as an aperture.

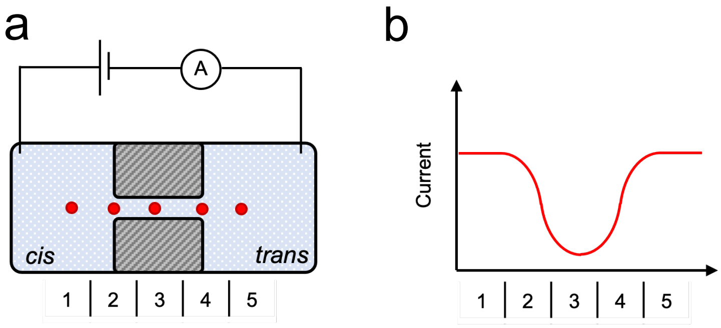


Figure 5.5: Theoretical basis of Coulter devices.
(a) Schematic of a Coulter device. An electrical current is provided between two reservoirs of electrolyte connected by a small aperture. As particles (red) pass through the aperture, current is blocked providing resistance under a constant voltage bias. **(b)** Expected current trace observed at the ammeter in (a), for each of the stated particle positions.

5.4.2.1 Solid state sequencing

Solid state nanopores are engineered into an impermeable solid-state membrane, such as Silicon Nitride, via transmission electron microscopes or ion beam milling [326,327]. These procedures generally create hourglass shaped nanopores ~1-2 nm in diameter, where around 10% manufacturing error can be expected [328,329]. Measurement of ssDNA homopolymers has been observed through solid state nanopores, however, high translocation rates (1-100 Mb/s) do not currently permit nucleobase discrimination [330]. Various solutions have been suggested to improve resolution, including increased data sampling (> 1 MHz)[328] and direct reduction of DNA movement. Retardation of DNA translocation has been explored through alterations that increase the viscosity of the electrolyte [331,332] and interactions between the DNA around the nanopore surface [333–335]. Currently, solid-state sequencing is not commercially viable, however, improved mechanical, thermal and chemical stability over biological nanopores make them a promising alternative for the future.

5.4.2.2 Biological nanopore sequencing

Biological nanopore sequencing exploits pores found in nature that spontaneously insert into lipid bilayers. Proteins such as α-hemolysin (α-HL) from *Staphylococcus aureus* form pores with diameter ~1.4 nm that are suitable only for passage of single stranded DNA [336]. In 1996, passage and detection of ssDNA through α-HL pores was reported. DNA translocation was slower than reported for solid-state nanopores (~0.1-1 Mb/s) although still too fast for single base discrimination [337]. Inclusion of DNA polymerase enzymes was

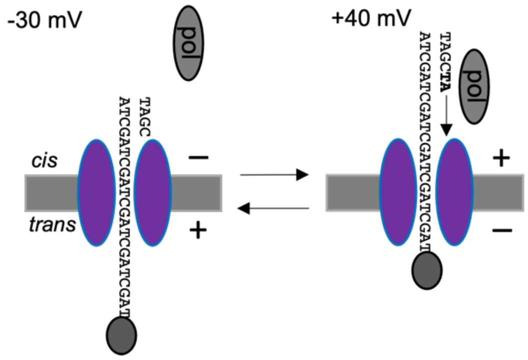


Figure 5.6: Proof of concept nanopore sequencing.
Experimental set up of early nanopore experiments demonstrated controlled translocation of ssDNA through a pore using a synthesis-based approach. In the left-hand panel, the electrical field pulls DNA taught, close to the nanopore, preventing accessibility of a polymerase. When the voltage is flipped, the DNA becomes accessible to a polymerase (right) and the dsDNA is elongated. The current is then reverse, the polymerase dissociates, and the DNA is held taught against the pore again (left). Because dsDNA is occluded from the pore, a new region of ssDNA will be placed in the aperture of the pore when the current is reversed. The process of current reversal and synthesis is repeated iteratively.

the first effective tool deployed to limit translocation to rates permissive for base discrimination [338,339] (Figure 5.6). Here the voltage is sequentially switched between +40 and -30 mV. When the voltage is + 40 mV, DNA is held out of the α-HL nanopore and primed ssDNA is exposed for extension by a DNA polymerase. When the voltage is flipped to -30 mV, DNA is pulled taut close to the α-HL pore and ssDNA is no longer exposed for elongation by the DNA polymerase [338,339]. DNA is ratcheted from *trans* to *cis* as the dsDNA section is sequentially elongated. Importantly, this iterative process discreetly placed nucleotides within the sensing region of the nanopore in a controlled manner and demonstrated the feasibility of using biological nanopores for sequencing.

5.4.2.3 Oxford Nanopore Technologies (ONT) sequencing

Oxford Nanopore Technologies (ONT) was initially set up by Hagan Bayley in 2005. Since foundation, ONT have produced and marketed diverse workflows and devices for native DNA (and RNA) sequencing using biological nanopores. ONTs core MinION sequencer is substantially smaller than previous devices allowing for sequencing at the point of sample collection [340]. ONT workflows can be split into 3 phases: sample preparation, data collection and data analysis.

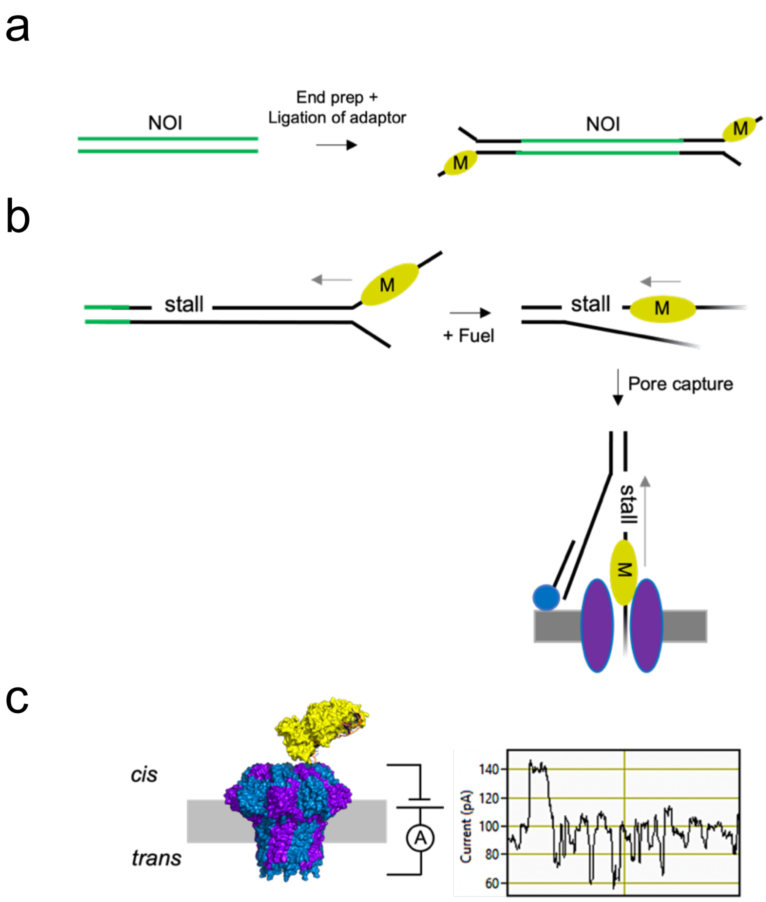


Figure 5.7: Oxford Nanopore Technologies (ONT) sequencing workflow.
(a) A standard library preparation workflow. A nucleic acid of interest (NOI) is prepared through a dA tailing step to generate sticky ends. An adapter, containing a stalled motor (M) is then ligated to the NOI. **(b)** Molecular basis for the proprietary adapters. When fuel (e.g., ATP) is added, the motor protein (M) can translocate through a section of DNA up to a stall, which it cannot pass through. This exposes the 5' end of DNA for capture by the pore. Capture of the DNA imparts a force on the motor, allowing the enzyme to pass through the stall and carry-on unwinding through the NOI. A cholesterol tether (blue circle) limits the diffusion of the DNA. **(c)** Visualized example of unwinding at the pore and exemplary data. The motor protein (e.g., Hel308, PDB: 2P6R), is shown in yellow, the pore (CsgG, PDB: 6LQH)[346] is shown in purple/blue.

First the nucleic acid of interest (NOI) is prepared for sequencing. This generally involves a dA-tailing step to incorporate sticky ends to each end of a DNA strand (see section 2.9). Proprietary 'adapter' sequences are then ligated to both ends of the DNA library (Figure 5.7a). Each adapter contains a tether sequence, a sticky dT overhang and a stalled motor protein. Together, this adapter is also referred to as 'AMX'. Prepared DNA libraries can then be loaded onto an ONT flow cell to initiate data collection. Central to ONTs platform is a processive motor protein [341]. Instead of a sequencing by synthesis approach described previously, ONT sequencing uses a heavily engineered helicase motor to unwind DNA directly above the pore [341]. The helicase provides ssDNA that is pulled from *cis* to *trans* through the pore under constant voltage. Translocation is limited to ~450 bases per second, where base discrimination is monitored using a sampling rate of 4 kHz [342]. Published examples of helicase mediated DNA translocation through biological nanopores have been examined elsewhere [343,344]. To limit helicase unwinding in bulk solution, a motor-specific stall is incorporated within the adapter[341]. Capture of the free adapter end by a nanopore imparts a force on the helicase, pushing the enzyme across the stall to initiate sequencing (Figure 5.7b). Annealing of a complementary membrane embedded tether to the adapter limits diffusion of DNA to 2-dimensions and enhances the efficiency of pore capture (Figure 5.7b).

Platform quality control is performed before the start of sequencing. For a MinION flow cell, the current of each nanopore is monitored by a 'mux scan' that ranks 4 nanopores that share 1 common electrode, across 512 groups [345]. Ranking also identifies well or channel pathologies, such as pore blockages, multiple pore insertions and membrane damage. As only 1 nanopore per electrode can be monitored at a time, the 'mux scan' ensures that the best pores are sequenced first. [345] A constant voltage of -180 mV is applied, and the signal is recorded by an application specific integrated circuit (ASIC) (Figure 5.7c)[345]. The magnitude of the signal at any one point is dominated by the number of nucleotides within the sensing region of the nanopore. For ONTs R9.4 nanopore, 5 nucleotides fit within the sensing region of CsgG, giving rise to at least 4^5 possible states before consideration of epigenetic modifications [346]. Recorded electrical signals (Figure 5.7c) are translated into DNA sequence *in silico* by use of trained neural network basecalling software such as Guppy and Bonito [347].

ONT sequencing produces remarkably long read lengths, where single sequencing events of over 2 Mb have been reported [348]. The primary limitation here is the physical properties of DNA, for example: narrow pipette tips physically shear long DNA [349]. At 450 bp/s, reads this long are expected to take > 1 hr to record. With only 512 active recording channels on a MinION device (extremely small relative to Illumina/Pyrosequencing), it is important to ensure that recording time is spent efficiently, sequencing primarily regions of interest. This can be ensured during sample preparations through enrichment of native samples via CRISPR-Cas9 methods [350]. Alternatively, recent software developments allow selective rejection of strands from a nanopore [351]. Here data is analysed in real time based on a defined selection of DNA. If the captured strand is determined as not interesting, it can be expelled from the pore via reversal of voltage [352,353].

The accuracy of single nanopore reads were initially poor at ~85-90% in the initial platform release, however alterations have improved the accuracy of reads dramatically to >98 %[345]. Firstly, alternative chemistries have been explored within the nanopore device. Various adapter formats have been explored to ensure that the complement strand is sequenced in succession to the primary strand[345]. This generates 2-fold coverage for a single molecule of DNA. Adaptations have also been made to the biological pore to reduce noise and increase the discrimination homopolymeric regions of DNA [354,355]. Further, large improvements have been achieved with updates to base calling algorithms [347] . Recently, new software has been developed to determine sequence accuracy in real time. This can be coupled with voltage mediated back tracking of the motor allowing indefinite re-reading of a sequencing until high enough accuracy is achieved [351].

5.5 Novel roles for MCM in ONT sequencing?

ONT sequencing has rapidly become a well-established technology within genomics, however, there remains scope to further improve current limitations of the system (Figure 5.8). These include:

1) Error rates;
2) Sample preparation time.

On the basis of 1), unique motor proteins rely on alternative mechanisms for translocation along DNA. For example, helicase enzymes exhibit different unwinding rates dependent on the %GC composition of DNA [75,356]. Unique translocation patterns would generate distinct signals when DNA is captured by a nanopore. These unique patterns contain an intrinsic error rate that is dependent on the motor enzyme[351]. The use of multiple motor proteins may therefore be beneficial to remove poor regions of coverage that are dependent on the motor. In this instance, we would like to examine whether an archaeal MCM can be used as a motor within an ONT flow cell. Success of this would also allow single molecule measurement to be made about an MCM translocating along DNA (Figure 5.8b). This may address longstanding questions, including: how many bases are translocated per ATP? What is the effect of DNA sequence on MCM translocation rate? How processive are MCM?

On the basis of 2), ONT sequencing is entirely reliant on sample preparation steps, where a motor protein stalled on an adapter is ligated onto a sequencing library. The preparation period using commercial kits generally takes between 10 minutes to > 1hr, depending on the scope of the experiment. However, in many real-world situations, such as medicine and forensics, speed of diagnosis is essential. Ideally, this step would be removed entirely. One possibility is to integrate both the pore and a motor. Since MCM is a motor that forms a pore, we hypothesize that MCM could be embedded within the membrane to perform this application (Figure 5.8c). Here, any DNA could be added to the flow cell, and MCM could capture and perform sequencing.

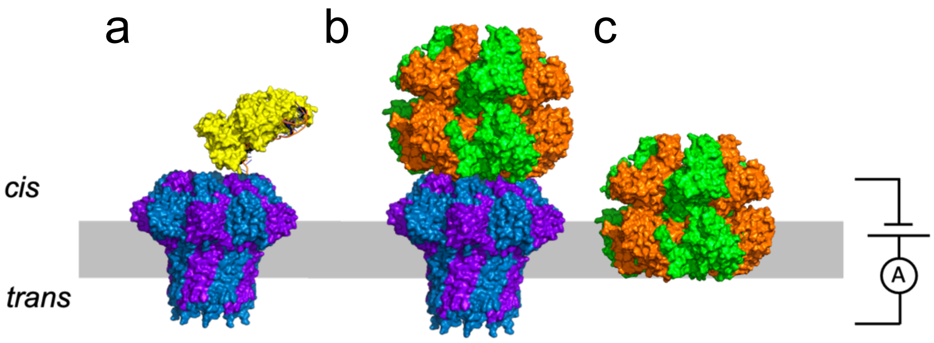


Figure 5.8: Potential uses of MCM within an ONT flow cell.
Proteins are represented in the surface format. **(a)** Example for scale of a monomeric motor in yellow (Hel308, PDB:2P6R) unwinding DNA above an R9.4 ONT nanopore in blue/purple (CsgG, PDB: 6LQH)[346]. **(b)** MCM (green/orange, structure from this thesis), may be used as a motor for nanopore sequencing. **(c)** It may be possible to embed an MCM into a membrane to function as both a motor and a pore.

5.6 Project aims

Based on these broader project aims, it was decided to examine first whether or not an MCM can act as a motor in an ONT flow cell. The following aims for this chapter were set:

- Determine optimal enzyme and buffer conditions for MCM in a flow cell;
- Design an MCM specific adapter substrate for sample preparation;
- Determine an MCM specific stall;
- Perform preliminary ONT experiments;
- Assess optimization/engineering of enzymes to enhance flow cell performance.

5.7 Testing MCM behaviour in potassium salts

5.7.1 Unwinding activity

Generation of electrical current within an ONT flow cell requires high concentrations of salt, where KCl is typically used at concentration 0.2 -1 M [355]. However, high concentrations of salt are detrimental to protein-DNA interactions particularly when electrostatic interactions are involved [357]. Equally in accordance with the Hofmeister series, chloride ions are more destabilising to proteins than organic anions such as glutamate (Glu)[358]. Indeed, studies that evaluated the activity of eukaryotic MCM in various concentrations of KCl and KGlu, suggest a preference for organic glutamate[63] . In nanopore

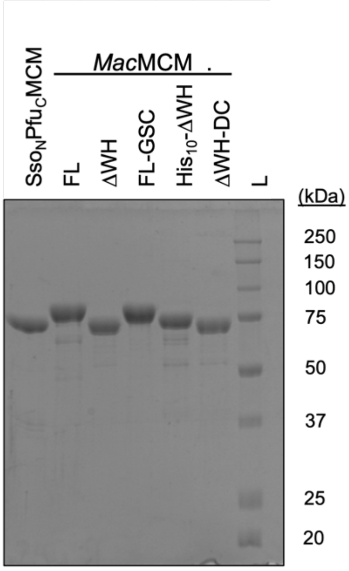


Figure 5.9: Purity of MCM constructs used in this chapter.
(a) The homogeneity of purified MCM was assessed through SDS PAGE analysis. 3 µg each purified MCM was run on a 12 % (*w/v*) polyacrylamide gel. Gels were stained using a Coomassie-based dye. Reference (L) lanes represent the MW standard marker (Precision Plus Protein™ All Blue Prestained Protein Standards).

sequencing it is possible to substitute chloride for glutamate to enhance protein stability, however differences in the conductivity provided by anions means higher concentrations of KGlu are required (>0.4 M) [331]. Potassium salts are generally preferable to sodium salts, which consistently exhibit lower conductivity [331]. High conductivity is important to maintain high DNA capture and a good signal to noise ratio for base discrimination [357,359]. To this extent, we decided to test the performance of 3 MCM with high activity at room temperature: *SsoPfu*MCM, *Mac*MCM^{FL} and *Mac*MCM$^{\Delta WHD}$ against different viable salts. The purity of all MCM used in this chapter was assessed by SDS-PAGE (Figure 5.9).

Briefly, the activity of each MCM was tested in a fluorescent helicase assay (described previously). Unwinding of a 26-base pair forked substrate was measured against increasing concentrations of KCl and KGlu. Net unwinding was quantified after 30 minutes by standardisation against a maximum fluorescence control. Except for 100 mM for *SsoPfu*MCM, MCM were less active in KCl at every tested concentration (Figure 5.10a). *SsoPfu*MCM is particularly intolerant to KCl, where activity is completely inhibited at 250 mM. Comparatively, *Mac*MCM^{FL} exhibits a stimulation in DNA unwinding when the concentration of KCl is increased from 100 to 250 mM. At concentrations of 500 mM KCl and higher, unwinding by both *Mac*MCM variants is nearly completely inhibited. *SsoPfu*MCM is much more tolerant to KGlu, where DNA unwinding reduces only 2-fold when the concentration is increased from 250 to 1000 mM. *Mac*MCM variants exhibit different tolerances to KGlu. Whilst the full-length enzyme is not able to unwind DNA at

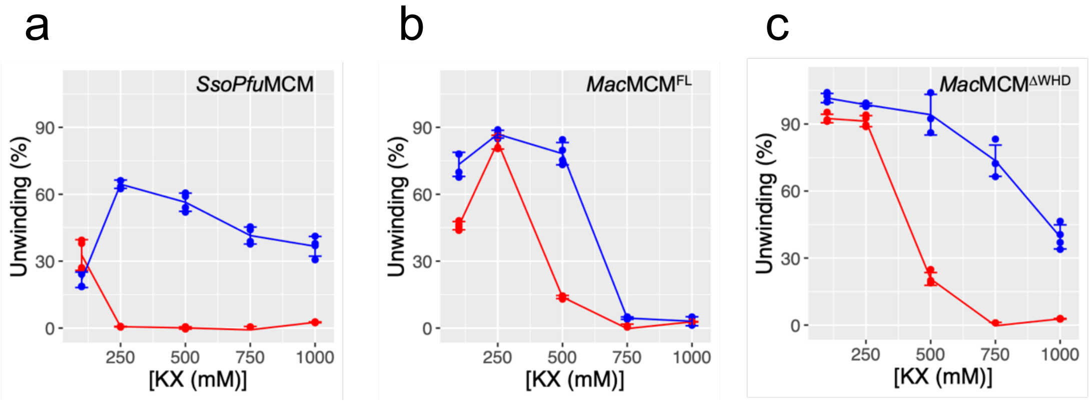


Figure 5.10: Unwinding activity of MCMs measured in different salts.
DNA turnover of a forked 26 base pair duplex was measured in a 96-well plate fluorescent helicase assay. Final salt concentrations are as stated, where red lines represent KCl and blue lines represent KGlu. Samples were incubated at 25 °C. After 30 minutes, ATP/Mg^{2+} was added to a final concentration of 4/10 mM. Substrate turnover was then monitored through changes in fluorescence for 30 minutes on a Clariostar plate reader (BMG Labtech). Data points were standardized to a maximum fluorescence control substrate for each salt concentration. The net DNA turnover was determined after 30 minutes for each salt concentration and MCM, where **(a)** *SsoPfu*MCM, **(b)** *Mac*MCM^{FL} and **(c)** *Mac*MCM$^{\Delta WHD}$. Error bars represent +/- 1 standard error of the mean, where n=4.

500 mM KGlu, the activity of the ΔWHD variant only declines ~2.5-fold as the concentration is increased from 500 to 1000 mM (Figure 5.10b-c). These data demonstrate the importance of salt specificity for the MCM motors tested.

5.7.2 DNA binding assessed by EMSA

Understanding the absolute mechanistic basis of salt tolerance for each enzyme requires a quantification of properties including protein stability, oligomerization and DNA binding under each condition. Interference of electrostatic interactions involved in DNA binding by salt is likely significant in explaining reduced activities. Previously, we have seen that removal of electrostatic interactions through mutation results in reduced DNA-binding and helicase activity (section 4.5). Due to the difficulties measuring *Mac*MCM binding, *SsoPfu*MCM was elected to examine the effects of salt concentration on DNA binding by EMSA and fluorescence anisotropy. Experiments were performed as described previously, however the concentration of the salt used in the binding reaction (before addition of loading dye) was adjusted as stated.

Under the tested conditions, *SsoPfu*MCM generally binds to the forked DNA substrate 2-fold tighter in KGlu versus KCl (Figure 5.11, Table 5.1). In both salts, increasing the concentration from 100 to 500 mM results in 2-fold decrease in binding affinity. Interestingly, there appears to be a dependency on the salt and concentration on the mobility of the free substrate. Increasing the concentration of KCl improves the resolution of the free band, whilst increasing the concentration of KGlu appears to decrease the resolution of the free band. The origin of this artefact is unknown, although it is likely due to salt-dependent conformations of substrates [360]. In EMSA experiments, it is difficult to reach the quoted concentration of salt and is always likely to be considerably lower. Initially, the sample is diluted in the loading buffer to achieve buffer sample homogeneity. Whilst it is possible to adjust for this, the sample is subsequently added to wells, where it is immeasurably diluted through the displacement of the running buffer.

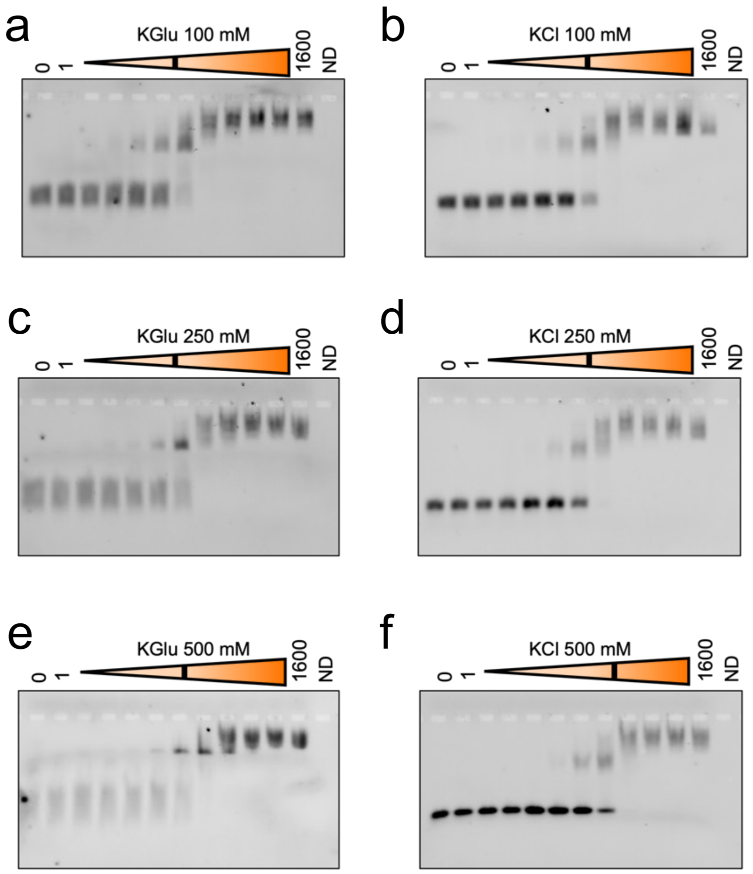


Figure 5.11: Assessment of *SsoPfu*MCM DNA binding to forked DNA by EMSA.
(a-f) The binding of *SsoPfu*MCM to a forked DNA substrate in different salts and concentrations was measured by EMSA. MCM were mixed at the stated concentration (nM hexamer), with 10 nM FAM-labelled DNA substrate. Final concentrations of salt are as stated during incubation, but before the addition of loading dye. Samples were incubated for 30 minutes at 25 °C . Samples were then resolved on a 1 x TB 0.8 % agarose gel and imaged using a Typhoon gel scanner (GE healthcare). A protein only control well, containing no DNA is represented by 'ND'.

5.7.3. DNA binding assessed by fluorescence anisotropy

To better examine the biophysical effects of salt on DNA-binding in absence of dilution effects, experiments were repeated via fluorescence anisotropy. Experiments were carried out as previously described, however, the default 250 mM KGlu salt concentration was replaced with the stated concentration of salt. Observations made under EMSA suggest all salt concentrations induce the formation of multiple MCM shifts/complexes, thus, Hill values are included in the Langmuir binding fit.

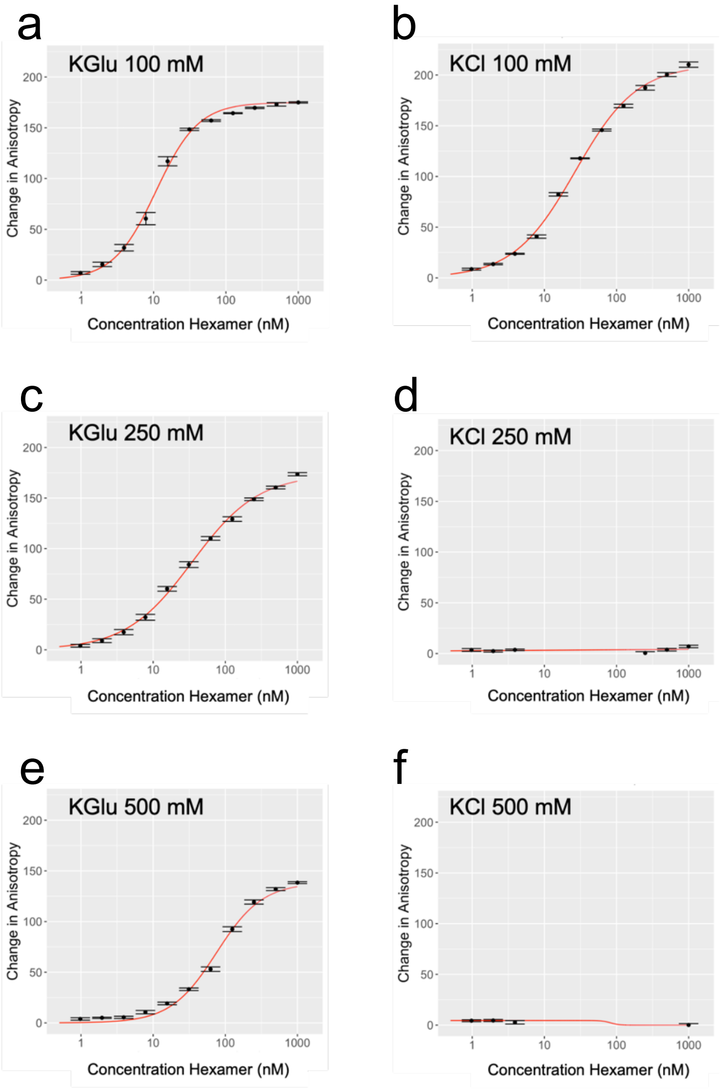


Figure 5.12: Assessment of *SsoPfu*MCM binding to forked DNA in different salts by anisotropy. (a-f) The binding of *SsoPfu*MCM to DNA in various concentrations of salts was determined by fluorescence anisotropy. MCM were mixed at the stated concentration (nM Hexamer), with 1 nM FAM-labelled forked DNA substrate and incubated for 30 minutes at 25 °C. Final concentrations of salts are as stated. Measurements were performed on a Clariostar plate reader (BMG Labtech). Anisotropy measurements were standardized by removal of a no-protein control well then fitted to a Langmuir binding isotherm with Hill coefficient. Error bars represent +/-1 standard error of the mean, where n=3.

Binding constants generated by fluorescent anisotropy are consistent with unwinding data, where MCM are unable to interact with DNA at concentrations of KCl greater than 100 mM (Figure 5.12). Differences in DNA binding capacity in KGlu are more apparent than measured by EMSA, where increasing the concentration of KGlu from 100 to 500 mM results in a decrease in affinity ~7.5 fold. The data outlined above suggest that all enzymes are suitable for study in ONT flow cells however ideally in a KGlu-based buffer.

Table 5.1: Calculated binding affinities of *SsoPfu*MCM DNA interactions in various concentrations and types of salt.

		K_d (nM)	
Concentration	Salt	EMSA	Anisotropy
100	KCl	50	26.8
	KGlu	25	10.8
250	KCl	50	n.c
	KGlu	50	34.7
500	KCl	100	n.c
	KGlu	50	75.5

5.8 MCM adapter design

5.8.1 Adapter evaluation by EMSA

ONT sequencing is reliant on efficiently stalling a motor protein on a forked DNA substrate or ‘adapter’. The motor protein must be in the correct orientation for unwinding and ideally at a 1:1 stoichiometry (protein: adapter). Furthermore, the length of the free end of the motor-bound arm must be long enough to encourage efficient pore capture (Figure 5.7b). Ideally, It must be longer than the binding footprint of the enzyme to allow capture of free ssDNA by the pore. To this end, a variety of different DNA substrates were examined to determine an ideal MCM adapter.

Based on the available structures, MCM are expected to have a total binding footprint of ~28 bases in length (i.e., ~100 Å height/ assuming 3.4 Å length per base). MCM are also expected to load onto a 3′ overhang, N-terminal domain first [78,125,361]. The 5′ overhang of the previously used forked substrate was shortened from 31 bases to 10 bases to prevent inverted binding of MCM. The 5′ excluded strand was not removed entirely, as it is thought to improve MCM-DNA binding through interactions with the external surface [88]. The 5′ excluded strand was labelled with 6-FAM and annealed to various DNA substrates to generate forks with increasing lengths of 3′ overhang. The dsDNA region remained at 26 base pairs in length.

DNA-binding was initially assessed by EMSA (as described previously) in 250 mM KGlu. *SsoPfu*MCM exhibits equivalent binding to ssDNA (36-bases) and forks with 10 and 20-base 3′ overhangs (Figure 5.13a-c). Increasing the 3′ overhang from 20 to 40-bases does not impact on the binding efficiency, however it encourages the formation of a secondary protein-dependent binding shift (Figure 5.13d). This suggests that beyond 36 bases, multiple MCM hexamers can bind to the DNA substrate. Tests with a blunt ended double stranded DNA construct reveals dsDNA binding is ~2-4 fold less efficient than ssDNA (Figure 5.13e).

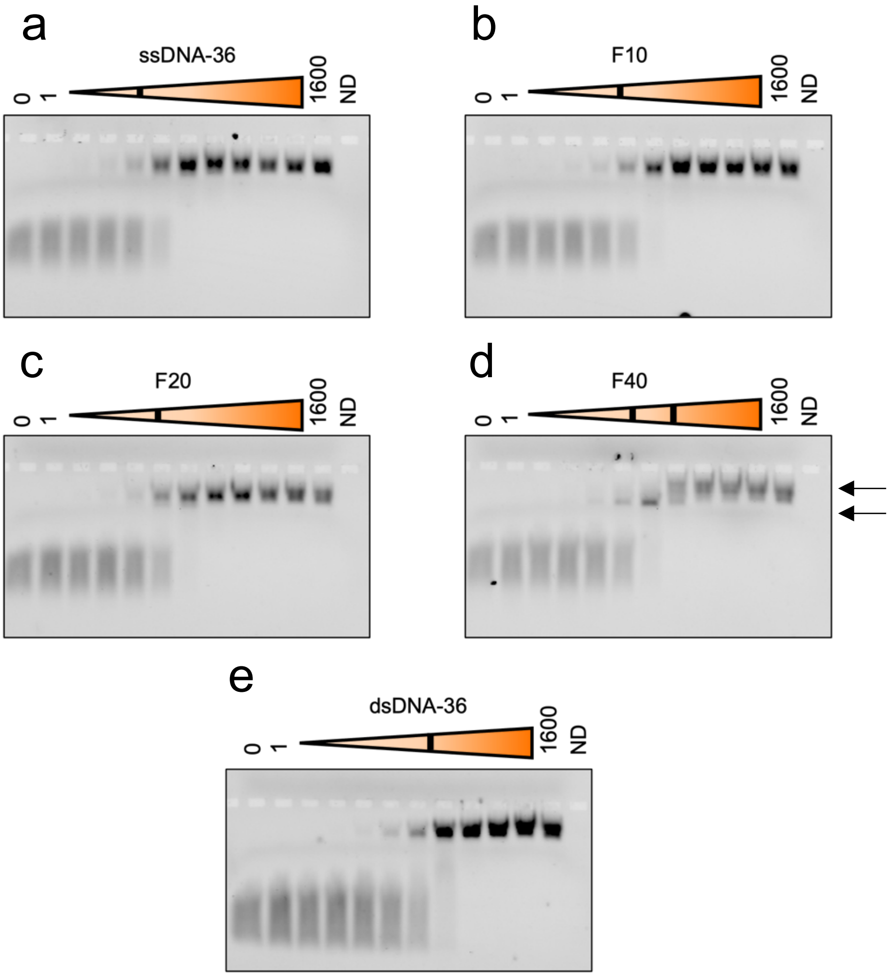


Figure 5.13: Assessment of *SsoPfu*MCM binding to different DNA substrates by EMSA.
(a-e) The binding of *SsoPfu*MCM to different DNA substrates was measured by EMSA. MCM were mixed at the stated concentration (µM hexamer) with 10 nM FAM-labelled forked DNA substrate. The substrate was as stated. Samples were incubated for 30 minutes at 25 °C. Samples were then resolved on a 1 xTB 0.8 % agarose gel and imaged using a Typhoon gel scanner (GE healthcare**).** A protein only control well, containing no DNA is represented by ‘ND’.

5.8.2 Adapter evaluation by fluorescence anisotropy

Experiments were repeated via fluorescence anisotropy, as described previously (Figure 5.14). All data are in reasonable agreement with values from EMSA, with similar magnitude (Table 5.2). Trends between datasets also match, however, K_d estimates calculated by anisotropy show ~2/3-fold tighter affinity. Slight discrepancies are expected as conditions measured through anisotropy do not fully match those used in EMSA, where the complex is subject to dilution effects when running through a gel.

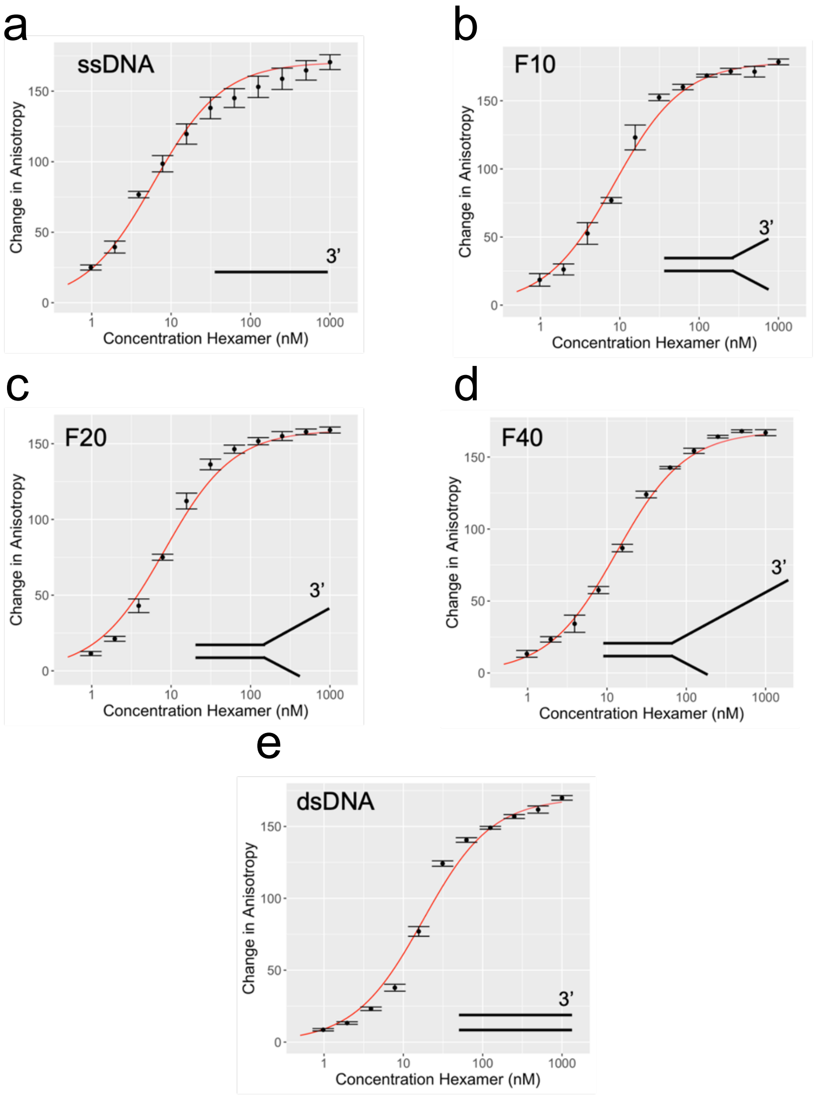


Figure 5.14: Assessment of of *SsoPfu*MCM binding to different DNA substrates by anisotropy. (a-e) The binding of *SsoPfu*MCM to different DNA substrates was measured by fluorescence anisotropy. MCM were mixed at the stated concentration (nM hexamer) with 1 nM FAM-labelled forked DNA substrate and incubated for 30 minutes at 25 °C. Measurements were performed on a Clariostar plate reader (BMG Labtech). Anisotropy values were standardized by removal of a no-protein control well and fitted to Langmuir binding isotherm with Hill coefficient. Error bars represent +/-1 standard error of the mean, where n=3.

Table 5.2: Calculated binding affinities of *SsoPfu*MCM against different DNA substrates.

	K_d (nM)	
Substrate	EMSA	Anisotropy
ssDNA-36	18.5	6.0
F10	18.5	8.2
F20	18.5	8.5
F40	25.0	13.2
dsDNA-36	50.0	17.3

5.8.3 Optimization of the 3' ssDNA overhang

The observation of multiple band shifts in EMSA has been observed using F40 and in our previous forked substrate (Figure 3.17; Figure 5.13). It allows justification of the inclusion of Hill coefficients in binding affinity calculations, however, in ONT sequencing 1:1 binding is desired. To examine the optimal ssDNA length for 1:1 binding, EMSA was performed with a sequentially elongated repeat substrate of (ACTG)$_n$ and *SsoPfu*MCM. The length of ACTG repeats were sequentially increased by 16 bases from 16 to 96 bases. The repeat sequence ensures the absence of secondary structure in the presence of an equal percentage of A, C, T and G nucleotides.

EMSA were performed on a 4% TB (no EDTA) PAGE to permit post-staining of DNA (via SYBR-GOLD) and protein (Coomassie; Supplementary Figure 7.9). A high concentration of DNA (250 nM) was used, as the focus is the number of shifts, rather than the precise affinity. Protein was serially diluted from 16 to 0.5-fold excess of hexamer to DNA.

After 40 minutes of resolving, gels were stained with SYBR-GOLD and visualised on a Typhoon gel scanner (GE Healthcare). In all instances, after 40 minutes, a single band shift was observed (Figure 5.15a). The intensity of the longer substrate is stronger, as more SYBR-GOLD molecules can bind. Gels were then repeated but allowed to progress for an extra 110 minutes. Staining reveals multiple states (Figure 5.15b). Substrates shorter than 33 bases exhibit a sole shift consistent with a single 1:1, hexamer: DNA complex. This is in excellent agreement with the structural estimations of 28 bases per hexamer. Increasing the substrate length to 48 and 64 bases generates 3 band shifts consistent with multiple binding events, or at least a 3:1 hexamer: DNA interaction. Further elongation of the substrate to 80 and 96 bases, generates further complexes, consistent with at least 5:1

protein: DNA binding. Therefore, the extensively used forked substrate, which has 49 and 31 base overhangs are exhibiting multiple DNA binding events with at least every 32 bases of ssDNA. This is also consistent with the extra complex formation between substrates F20 and F40 (Figure 5.13b-d).

It is important to note that these EMSA results are somewhat qualitative. Whilst longer DNA migrates more slowly through the gel, longer DNA also provides more force to pull a complex into the gel. However, longer substrates also provide opportunities for more MCM to bind the substrate and hence slow the migration of the complex further. This data implies that substrate lengths ~16 bases are sufficient for single MCM loading.

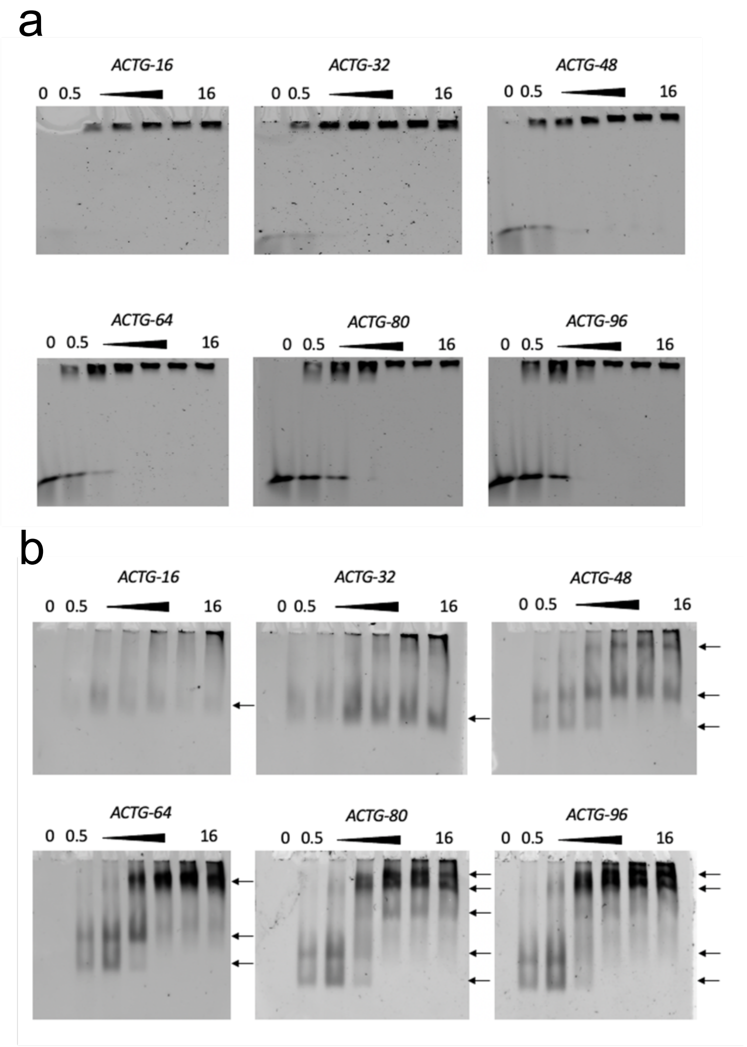


Figure 5.15: Estimation of MCM DNA binding footprint by EMSA.
(a) 250 nM of the stated (ACTG)$_n$ substrate was mixed with the stated molar excess of *SsoPfu*MCM. Samples were run on a 1 x TB 4 % (*w/v*) polyacrylamide gel for 40 minutes. Gels were then stained using SYBR-GOLD and analysed on a Typhoon Imager (GE Healthcare). **(b)** Experiments in (a) were repeated, however gels were allowed to run for 2h 30 before staining. Arrows represent the observed complexes. Samples lining/ trapped in the well are ignored as possible aggregation/artefact.

5.8.4 A background to helicase stalls

To measure DNA unwinding at a nanopore, it is important to determine an efficient motor-specific stalling mechanism. A stall ensures that DNA unwinding occurs primarily when an open pore is available for strand capture, rather than occurring in bulk solution. Motor specific stalls can include a variety of different approaches, including: protein blocks (i.e. streptavidin/biotin); abasic DNA; morpholino DNA, non-native sequences (i.e. spacers, RNA, fluorophores etc) and secondary structure/aptamers [101,170,341] . Stall compatibility depends primarily on the molecular interactions that a motor protein makes. For example: incorporation of abasic sites is unlikely to stall a motor which relies primarily on electrostatic interactions with the phosphate backbone.

Preliminary assumptions can be made about MCM. Firstly, as MCMs unwind the genome, they natively encounter and translocate through protein blocks, RNA, abasic sites and DNA secondary structures [71,87,362]. This limits our stall selection to non-native sequences. Secondly, the footprint of MCMs is particularly large and it is unclear how long a stall would have to be to prevent translocation. For example, one may postulate that it is only the interactions and binding footprint of the active C-terminal domain (~10 bases) that need to be considered to stall an MCM, rather than total footprint of the MCM (28 bases) (Figure 5.16a). Based on the multitude of DNA-binding interactions that an MCM makes, increasing lengths of a non-native, hexaethylene glycol spacer (iSp18) was chosen to act as a stall (Figure 5.16b).

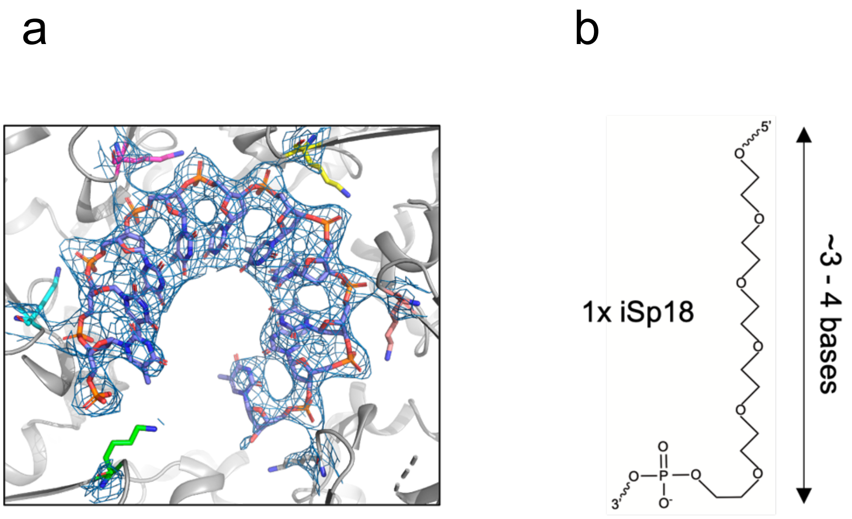


Figure 5.16: Design of MCM specific stalls.
(a) Binding of *Sso*MCM ATPase domain (PDB: 6MII)[100] to 10 bases of ssDNA (sticks). The electron density of the DNA is coloured in blue. The phosphate backbone of DNA is co-ordinated by 4 lyisine residues (K430) through electrostatic interactions Protein is shown in the ribbon format. **(b)** Chemical structure of an iSp18 spacer. The length of 1 iSp18 spacer is equivalent to 3-4 bases.

To test this stall, it was decided to perform a fluorescence helicase assay with a DNA substrate that internally incorporates iSp18 stalls. Our previously utilised DNA substrate however is not suitable to test this hypothesis. The substrate is a duplex of only 26 base pairs and when testing stalls up to 18 bases in length (6x iSp18) will not leave enough bases for stable substrate annealing.

5.8.5 MCM specific stall design

A long 86-base pair fluorescently labelled DNA duplex was designed. This is comprised of 3 oligonucleotides: a top strand with 86 bases of complementary region is annealed to a Cy3 labelled 80 base oligo and a 15 base BHQ2 quencher oligo. (Figure 5.17; see section 2.1.6). It is acknowledged that a 1 base gap should be left between the fluorophore and quencher [363]. Both duplex regions were designed with GC clamps to ensure proper annealing. Based on the EMSA footprinting work, it was decided to use 15 and 10 base overhangs for 3′ and 5′ respectively. This set up permits rapid testing of multiple top strands with different internal stalls present.

Initial tests suggest that both *Mac*MCM^{FL} and *Mac*MCM$^{\Delta WHD}$ can unwind the longer substrate with comparable efficiency to the short substrate (Figure 5.17b; Figure 5.10b-c). Surprisingly, *SsoPfu*MCM was unable to unwind the substrate, suggesting that it is unable

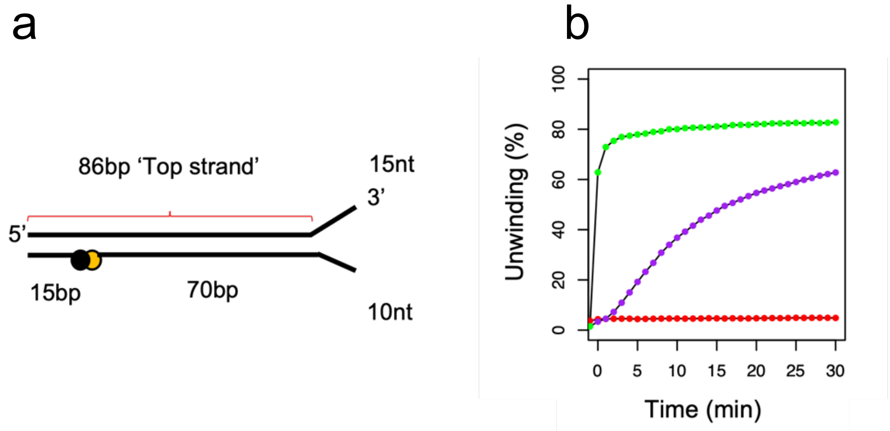


Figure 5.17: Benchmarking a helicase substrate for determination of adapter stall efficiency. **(a)** A long fluorescent helicase substrate was designed for testing new stalls in the top strand region. An 86 base top strand region is designed complementary to a 70 base BHQ2 strand. The Cy3 and BHQ2 moieties are designed with a 1 base gap. The long substrate has a 15 base 3' overhang and a 10 base 5' overhang. **(b)** Real time unwinding profiles for 3 MCM orthologues on the long helicase substrate. DNA turnover of the long substrate was measured in a 96-well plate fluorescent helicase assay. Samples were incubated at 25 ºC. After 30 minutes, ATP/Mg^{2+} was added to a final concentration of 4/10 mM, and 50 nM substrate turnover by 1000 nM MCM (hexamer) was monitored through changes in fluorescence over 30 minutes on a Clariostar plate reader (BMG Labtech). Data points were standardized to a maximum fluorescence control substrate and represent the mean value where n=2. *SsoPfu*MCM: red, *Mac*MCM^{FL}: purple, *Mac*MCM$^{\Delta WHD}$: green.

to unwind processively. It was hypothesized that the loading footprint was too short for *SsoPfu*MCM to load onto DNA properly. Therefore, a longer top strand was tested with a 30 base 3’ overhang (Figure 5.18a). Increasing length of the loading strand was insufficient to increase the measured activity of *SsoPfu*MCM (Figure 5.18b). The activities of both *Mac*MCM^{FL} and *Mac*MCM$^{\Delta WHD}$ unwinding remain unaffected by the extension of the 3’ overhang.

The ability of the enzymes to unwind blunt-ended DNA was also assessed. Understanding basal levels of unwinding from a blunt substrate suggests the activity we may expect from a successful stall (It is assumed that some unwinding will occur from the blunt quencher end of the duplex). A 96 base complement strand was annealed to the fluorescent and quencher strands (Figure 5.18c). Analysis of the blunt dsDNA substrate reveals that over 30 minutes, both *Mac*MCM constructs unwind ~4-fold less than a forked substrate (Figure 5.18d). *SsoPfu*MCM is not able to unwind blunt ended DNA.

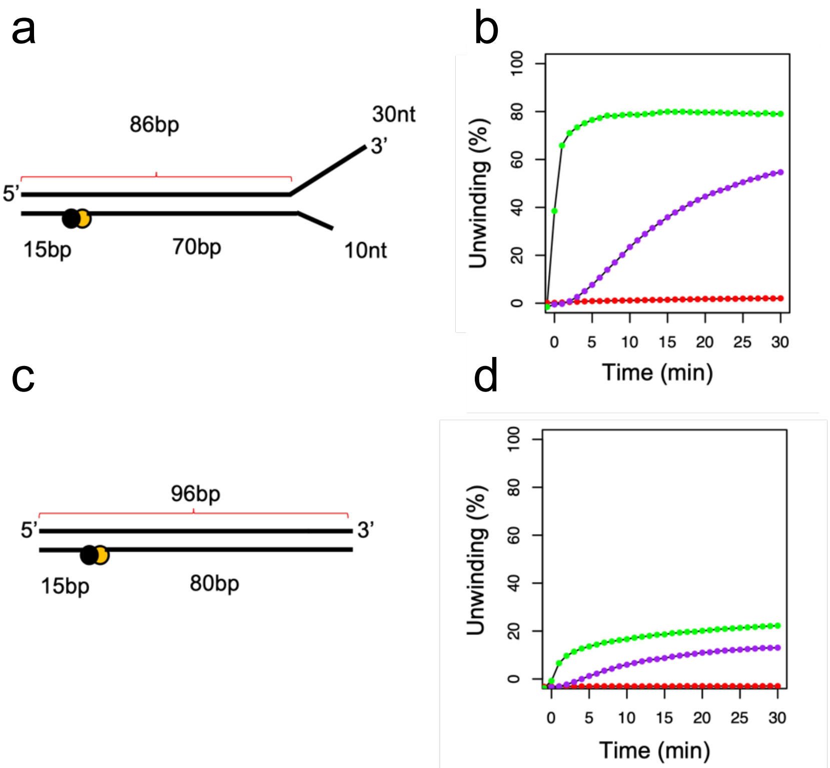


Figure 5.18: Further assessment of ideal 3' MCM loading platforms and blunt unwinding. **(a)** The 3' overhang of the top strand was increased in length to 30 nt to determine whether *SsoPfu*MCM unwinding is affected by loading platform size. **(b)** Real time unwinding profiles for 3 MCM orthologues on the long helicase substrate with a 30 nt 3' overhang. *SsoPfu*MCM: red, *Mac*MCM^{FL}: purple and *Mac*MCM$^{\Delta WHD}$: green. DNA turnover of the long substrate was measured in a 96-well plate fluorescent helicase assay. Samples were incubated at 25 ºC for 30 minutes. ATP/Mg^{2+} was then added to a final concentration of 4/10 mM, and 50 nM substrate turnover by 1000 nM MCM (hexamer) was monitored through changes in fluorescence for 30 minutes on a Clariostar plate reader (BMG Labtech). Data points were standardized to a maximum fluorescence control substrate and represent the mean value for n=2 traces. **(c)** To quantify blunt end unwinding, the 3' overhang of the top strand was shortened in length to 10 nt and made complementary to the 5' overhang to generate a blunt ended duplex. A 96-base top strand is designed complementary to an 80 base Cy3 strand and a 15 base BHQ strand. **(d)** The experiment in part (b) was repeated, however using the blunt substrate outlined in part (c).

5.8.6 Determining the efficiency of iSp18 stalls

As *SsoPfu*MCM was unable to unwind the 86-base duplex, it was omitted from subsequent analysis. Spacers were then incorporated into the centre of the 70-base portion of the 86-base duplex, with equal portions of duplex on each side to maintain annealing (Figure 5.19a). For each iSp18 spacer incorporated, 3 bases of DNA were replaced (see section 2.1.6). The level of unwinding was then assessed 30 minutes after the addition of ATP.

When tested in standard conditions, only *Mac*MCM^{FL} exhibited inhibition of unwinding that is proportional to the number of stalls in the substrate (Figure 5.19b). Beyond four iSp18 spacers, DNA unwinding is reduced to a level consistent with the blunt substrate. By comparison, under the same conditions, *Mac*MCM$^{\Delta WHD}$ does not exhibit any inhibition by iSp18 spacers (Figure 5.19c).

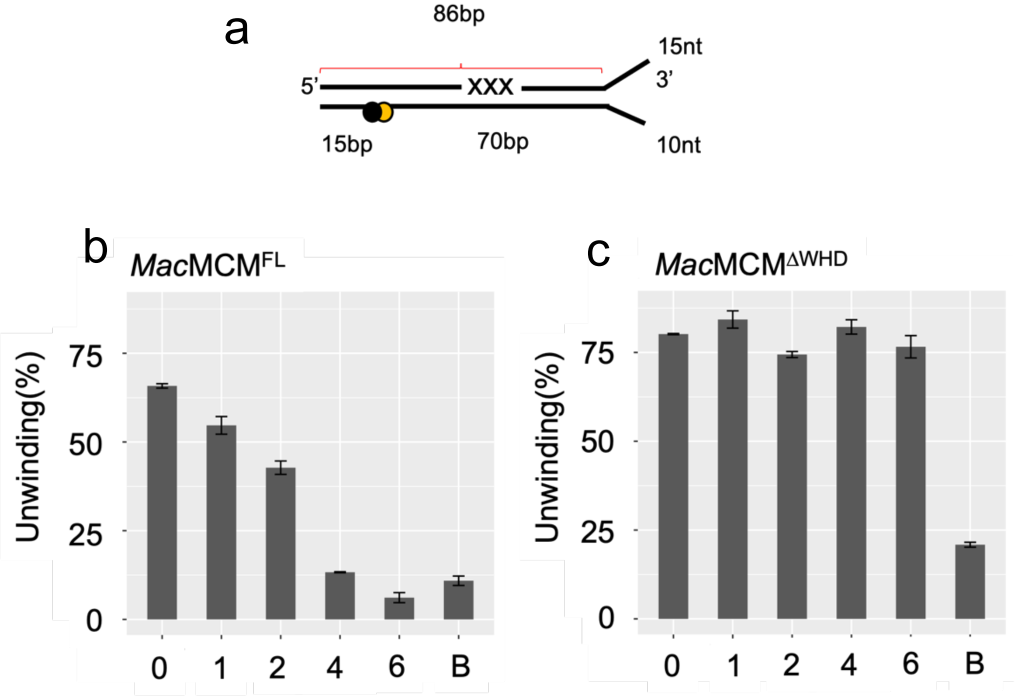


Figure 5.19: Testing iSp18 spacers as an MCM stall.
(a) To determine the effect of iSp18 stalls, the top strand of the long substrate was redesigned to incorporate multiple iSp18 spacers in the centre of the 70-base duplex. The length of each iSp18 spacer was assumed to be 3 bases. The Cy3 and BHQ2 moieties are designed with a 1 base gap. The long substrate has a 15 base 3' overhang and a 10 base 5' overhang. **(b-c)** DNA turnover of the stall substrates was measured in a 96-well plate fluorescent helicase assay. Samples were incubated at 25 ºC for 30 minutes. ATP/Mg^{2+} was then added to a final concentration of 4/10 mM, and 50 nM substrate turnover by 1000 nM MCM (hexamer) was monitored through changes in fluorescence for 30 minutes on a Clariostar plate reader (BMG Labtech). The net DNA turnover was determined after 30 minutes for each spacer length (0-6) and MCM. The letter B denotes experiments performed with the blunt-ended control substrate (see Figure 5.18c), to quantify DNA unwinding that may be possible from the 15 bp end of the duplex. Data points were standardized to a maximum fluorescence control substrate and represent the mean value for n=2. Error bars represent +/- 1 standard error of the mean.

It was hypothesized that when *Mac*MCM$^{\Delta WHD}$ approaches a stall, a secondary MCM follows and pushes the primary MCM through the stall, generating unwinding (Figure 5.20a). This hypothesis is plausible, as the standard helicase protocol utilises an excess ratio of 20:1 hexamer: DNA. Similar hypotheses have been explored elsewhere [341]. It was therefore decided to test unwinding of substrates at smaller protein: DNA ratios.

At both 2:1 and 1:1, *Mac*MCM$^{\Delta WHD}$ can turn over comparable levels of no-stall substrate to 20:1 (~75%) (Figure 5.20b-c). For substrates with greater than one iSp18 spacer, reduction of the protein DNA ratio to 2:1 suggests inhibition of *Mac*MCM$^{\Delta WHD}$ unwinding by ~40-50% (Figure 5.20b). At this ratio it is still possible that secondary MCM can force the primary MCM across the stall. Further reduction of the protein: DNA ratio to 1:1 reveals total inhibition of *Mac*MCM$^{\Delta WHD}$ for substrates with greater than one iSp18 spacer (Figure 5.20c). The amount of unwinding observed is not more than what is expected from the unwinding from a blunt end. From the data outlined above, an MCM nanopore sequencing adapter was designed.

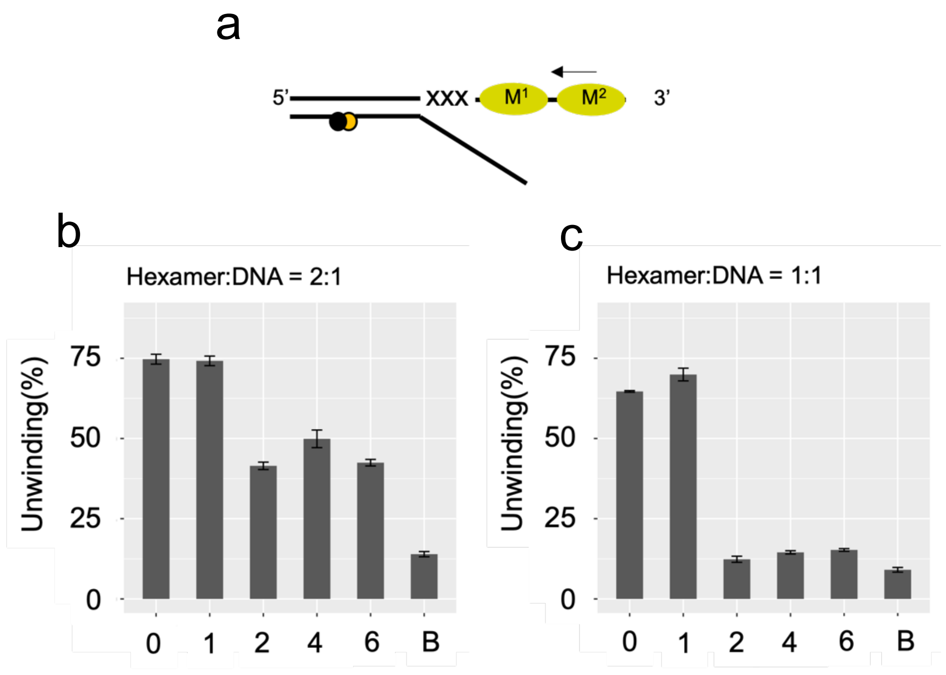


Figure 5.20: *Mac*MCM$^{\Delta WHD}$ stalling is dependent on protein concentration.
(a) To determine whether the primary loaded *Mac*MCM$^{\Delta WHD}$ (M^1) is pushed through iSp18 stalls by secondary *Mac*MCM$^{\Delta WHD}$ (M^2), fluorescent helicase assays were performed with lower concentrations of protein. **(b-c)** DNA turnover of the stall substrates was measured in a 96-well plate fluorescent helicase assay with the stated ratio of protein: DNA. Samples were incubated at 25 ºC for 30 minutes. ATP/Mg^{2+} was then added to a final concentration of 4/10 mM. The turnover of 50 nM DNA substrate was monitored through changes in fluorescence for 30 minutes on a Clariostar plate reader (BMG Labtech). The net DNA turnover was determined after 30 minutes for each spacer and hexamer: DNA ratio. Data points were standardized to a maximum fluorescence control substrate and represent the mean value for n=2.

5.9 Initial ONT sequencing experiments

The MCM-specific adapter was designed to be compatible with the ONTs Control Expansion sequencing kit (Figure 5.21). This involved addition of a single dT overhang to the blunt end of the dsDNA section of the MCM adapter, permitting ligation to a dA-tailed genomic phage lambda DNA substrate. The 5′ duplex end was also phosphorylated. A double iSp18 spacer was incorporated into the adapter, and a 15 base 5′ overhang sequence was added complementary to ONTs cholesterol-based tether strand.

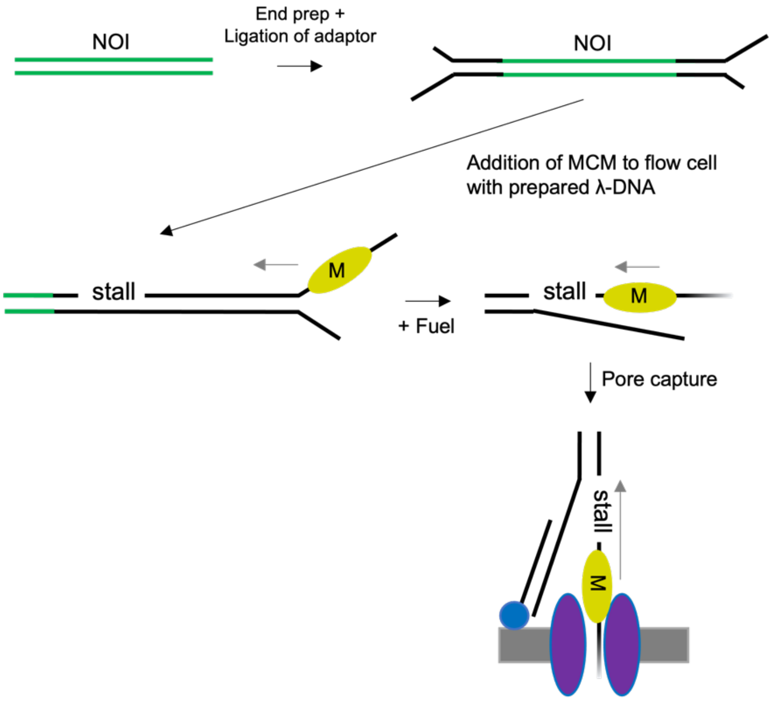


Figure 5.21: Experimental outline for preliminary MCM-ONT experiments.
The nucleic acid of interest (NOI; 48 kb phage λ-DNA) is prepared by dA-tailing and ligation to an MCM specific adapter containing 2x iSp18 spacers. The MCM is mixed with the DNA before addition to the MinION flow cell. ATP is added to the flow cell and the experiment is run.

Briefly, AMX was replaced with a 1 µM MCM adapter. Samples were loaded into an R9.4 MinION flow cell at 1:1, Hexamer: DNA, where the final quantity was 80 pmol. The experiment was set to run for 6 hours at -180 mV, with static flips every 5 minutes. After 30 minutes run time, a second buffer was added to the flow cell, including 450 mM KGlu, 25 mM HEPES pH 8.0, 12 mM ATP, 20 mM $MgCl_2$. Data were analysed in TraceViewer (ONT, personal communication). A control AMX experiment was also performed for internal comparison.

Initial analysis of the ONT AMX motor control gives an example of the performance for a commercial sequencing grade motor protein (Figure 5.22a). Capture of the DNA strand into the pore results in a sharp drop in current by 140 pA. Without helicase translocation, captured DNA moves through the pore very quickly and no discrete states are observed. As AMX passes through the stall, it is placed into the pore, generating a current spike. Afterwards, AMX translocates along the DNA, and nucleobase specific states are observed. Sequencing events are observed with high frequency across the flow cell.

Analysis of initial *Mac*MCM experiments suggest that DNA translocation-like events are very infrequent. Examples of the most movement-like events are shown (Figure 5.22b). There is evidence of some pore capture of DNA, although no evidence of a discrete adapter stall peak. Within the motor-like events there are some suggestions of enzyme-controlled movement of DNA, with multiple discrete levels observed. If this is true enzyme-controlled movement, these MCM data are much noisier than for AMX.

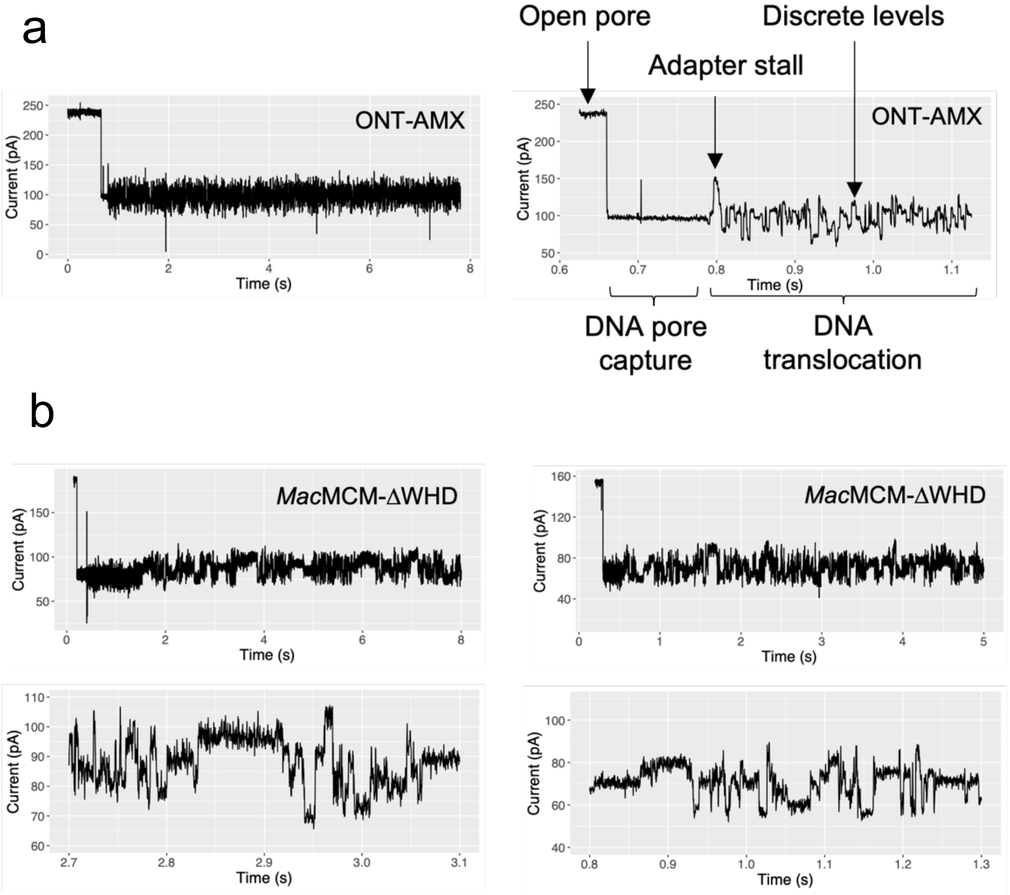


Figure 5.22: Potential signs of *Mac*MCM$^{\Delta WHD}$ unwinding in a MinION flow cell.
(a) Representative current trace of ONTs commercial grade sequencing motor (AMX). Open pore is at ~240 pA. Capture of DNA by the pore causes a drop in current of ~140 pA. The adapter stall is identifiable as a current peak before the main bulk of sequencing occurs. Following the passage of the adapter stall, discrete current levels are observed, coherent with DNA ratcheting through the Nanopore. **(b)** Example current traces for *Mac*MCM$^{\Delta WHD}$ experiments. A sharp drop in current of ~100 pA is observed upon putative pore capture. Close inspection of the current suggests discrete levels of current are achieved, consistent with controlled DNA translocation through the pore.

5.10 MCM engineering

5.10.1 Improvement of noise and frequency of events: Protein-pore tethering

One hypothesis for the noise of potential events is that the MCM protein does not sit stably on top of the pore. Movement of the enzyme on top of the pore can cause stochastic leakage of current through the pore, generating noise (Figure 5.23a-b). Tethering of the enzyme to the pore can limit the orientations the enzyme can assume and reduce the noise. This approach has been used before using a SpyTag to tether a polymerase enzyme onto a single α-HL pore [364]. Maintaining close proximity of an MCM to the pore would also likely increase the number of observed events.

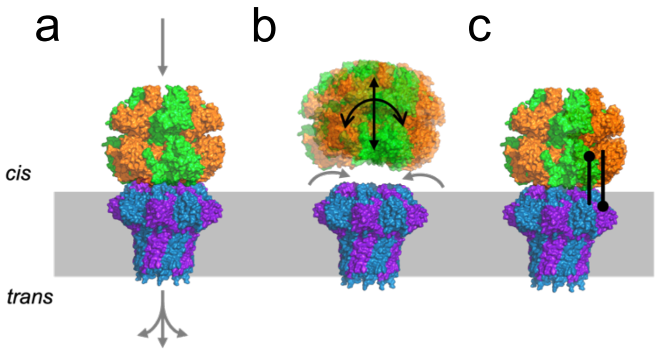


Figure 5.23: Sources of noise and potential fixes.
Protein is visualized in the surface format. **(a)** In an ideal situation, MCM sits stable on top of the CsgG[346] nanopore (purple/blue). Current passes only through the MCM and pore in a defined manner. **(b)** In a non-ideal situation, MCM does not sit stably on top of the nanopore. Here current stochastically leaks into the nanopore, generating random current drops and spikes. **(c)** One method to reduce MCM motion on top of the pore is to physically tether the enzyme. Here we propose the use of a DNA-based tethering system to immobilize the enzyme.

A DNA-based tethering approach was applied, whereby a maleimide-labelled oligonucleotide could be covalently linked to the MCM through a thiol bond. This oligonucleotide would be complementary to a capture oligonucleotide linked to the pore. Annealing of the two nucleotides would achieve tethering of MCM to the pore (Figure 5.23c). A reactive mutant was designed through addition of a cysteine residue to the C-terminal domain. The final 3 residues of *Mac*MCM^{FL} (EEE) were mutated to GSC to minimize electrostatic repulsion with the negatively charged phosphate backbone of a conjugate DNA tether. Incorporated glycine and serine residues further improve accessibility of the reactive cysteine. The cysteine mutant was purified as outlined previously (Figure 5.9), however in the final purification step, 0.5 mM DTT was substituted for 0.5 mM TCEP, which does not interfere with crosslinking reactions.

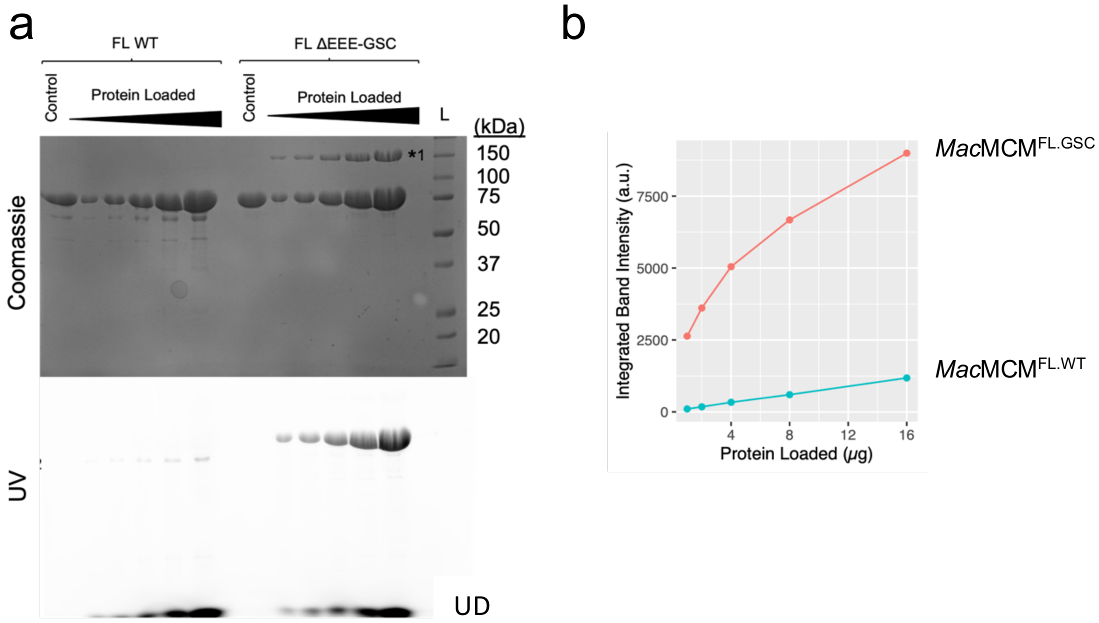


Figure 5.24: Selective labelling of engineered cysteine residues.
(a) To compare the incorporation of fluorescein-maleimide into *Mac*MCM^{F}L and *Mac*MCM$^{FL\Delta EEE.GSC}$, increasing quantities of reaction product were analysed on a polyacrylamide gel. Gels were imaged on a UV transilluminator to determine incorporation of fluorescein and then stained using a Coomassie-based approach to detect protein. UD: Unincorporated dye. **(b)** The intensity of the fluorescent bands in (a) were quantified in ImageJ and plotted as a function of the protein loaded. Lines are coloured for each MCM, where *Mac*MCM^{FL} is cyan and *Mac*MCM^{FL-GSC} is coral.

To determine the reactivity of the engineered cysteine residue, a maleimide-fluorescein labelling strategy was performed (see section 2.2.11). Briefly, a 2-fold molar excess of maleimide-fluorescein was added to *Mac*MCM^{FL} and incubated for 30 minutes at 37 °C. As each MCM subunit contains 4 native cysteine residues that coordinate a zinc atom, it is important to ensure that ZnF integrity is maintained and that the reaction is site specific. Therefore, this experiment was performed in parallel against the wild-type enzyme.

Polyacrylamide gel analysis suggests the reactivity of the cysteine mutant is at least 10-fold more than the native cysteines, suggesting reasonable selectivity (Figure 5.24a-b). Interestingly, in the case of the cysteine mutant a small amount of non-fluorescent dimer was formed, however, this only occurs in the presence of the maleimide dye. Dimerization appears to be mediated by both the dye and the cysteine mutant as it does not occur in the DMSO only control. The absolute identity of this band is yet to be identified. After the reaction, labelled product is dialysed and desalted via SEC to remove unwanted DTT, DTT-dye and DMSO (Figure 5.25). It was not possible to resolve the monomer and covalent dimer species using the chosen column. Calculation of labelling efficiency was not deemed important for this application.

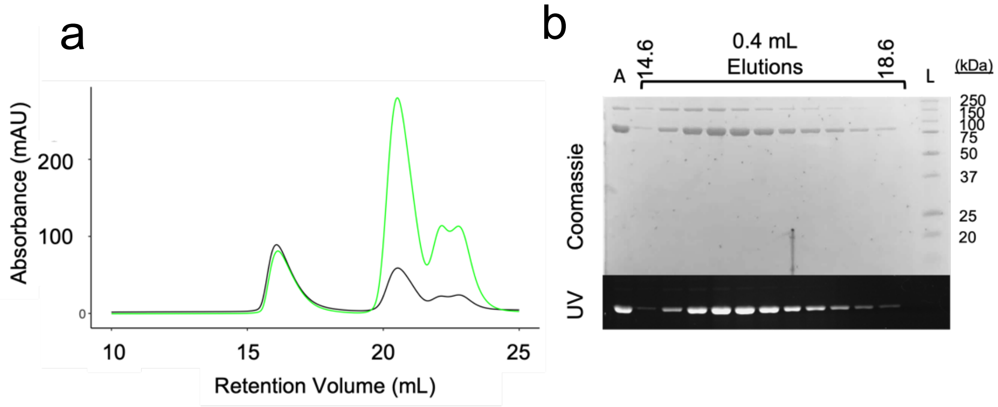


Figure 5.25: Purification of fluorescently labelled *Mac*MCM$^{FL.GSC}$.
(a) Reaction products (DTT, DTT-dye, dye and DMSO), were separated by size exclusion chromatography from MCM. SEC was performed on a Superose 6 Increase 10/300 GL column. Absorbance of the eluate was monitored at 280 (black) and 495 nm (green). **(b)** Polyacrylamide gel analysis of elution fractions from SEC resolved on a 12 % SDS-PAGE. Gels were imaged on a UV transilluminator to determine the incorporation of fluorescein and then stained using a Coomassie-based approach to determing protein. Trace amounts of non-fluorescent covalent dimer species coelute with monomeric *Mac*MCM.

5.10.2 Improvement of noise and frequency of events: Protein-pore tethering - validation of interactions

If the MCM is to act as a tethered motor, it must be able to interact with non-tethered subunits, whilst attached to a bulky conjugate. To investigate whether interactions are possible, a column-based affinity capture assay was performed between *Mac*MCM^{FL-GSC}-FAM and His$_{10}$-*Mac*MCM$^{\Delta WHD}$. The *Mac*MCM^{FL-GSC}-FAM has an intrinsic, bulky WHD attached.

Initially, His$_{10}$-*Mac*MCM$^{\Delta WHD}$ was passed over a 1 mL Ni-NTA column equilibrated in ATP buffer and eluted using a stepwise elution of 500 mM imidazole. Here, the enzyme is present in the elution and does not emit any fluorescence (Figure 5.26a). Contrarily, *Mac*MCM^{FL-GSC}-FAM cannot interact with the nickel column alone and hence is observed as a fluorescent band in the flow through (Figure 5.26b). Mixing equimolar concentrations of *Mac*MCM^{FL-GSC}-FAM and His$_{10}$-*Mac*MCM$^{\Delta WHD}$ in an ATP/Mg^{2+} buffer generates elution of fluorescently labelled *Mac*MCM^{FL-GSC}-FAM in both the flow through and the elution (Figure

5.26c). This strongly suggests that *Mac*MCM$^{FL\text{-}GSC}$-FAM is active and can interact with other *Mac*MCM constructs when attached to a bulky conjugate.

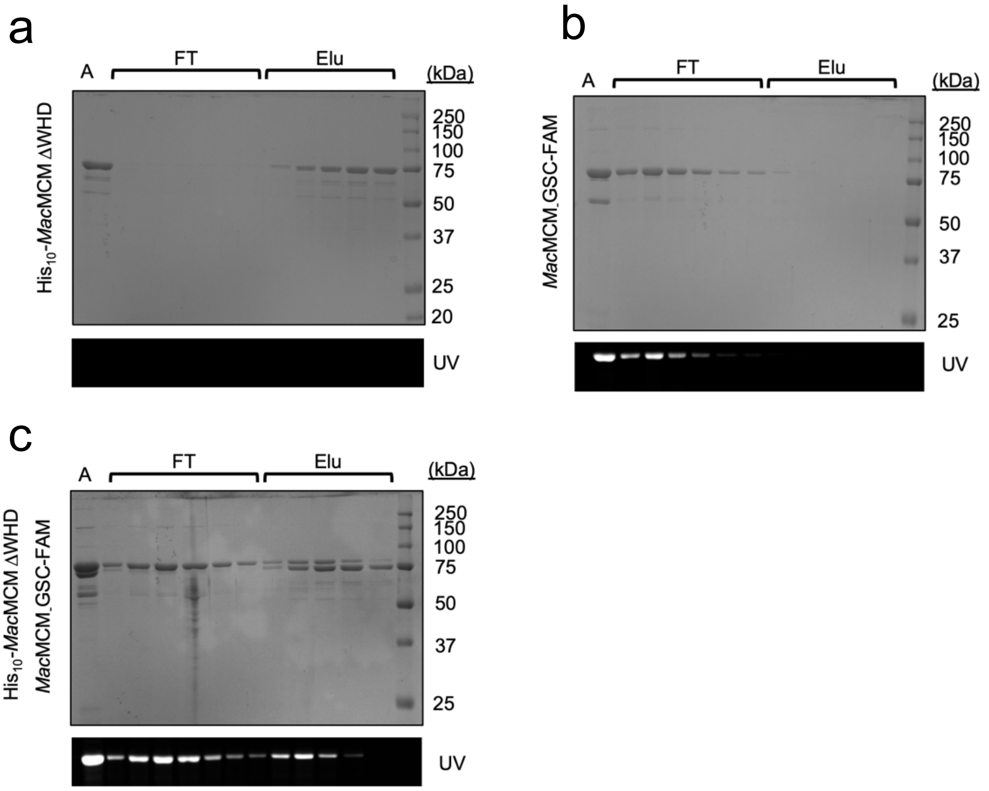


Figure 5.26: Interactions of His$_{10}$-*Mac*MCM$^{\Delta WHD}$ with labelled, *Mac*MCM^{FL}.
(a-c) 1 mg of protein was applied (A) onto a 1 mL HP-Ni-NTA column pre-equilibrated in 100 mM NaCl, 20 mM Tris-Cl pH 8.0, 5 % glycerol, 1 mM ATP, 10 mM $MgCl_2$. Where two proteins were present, the 1 mg comprised of an equimolar ratio of each enzyme. The loaded sample is noted to the left hand of the figure. At each stage, 1 mL fractions were collected. Bound proteins were then washed with 10 mL of the above buffer (FT). Bound proteins were then eluted (Elu), with a buffer containing 500 mM Imidazole. 5 µL each fraction was then analyzed by electrophoresis on a 12 % SDS-PAGE. Gels were imaged on a Typhoon Imager (GE Healthcare) to determine incorporation of fluorescein (UV) and then stained using a Coomassie-based approach to determine protein.

5.10.3 Improvement of noise and frequency of events: Forced hexamerization

One potential problem with using MCM as a motor within a nanopore flow cell comes from the oligomerization propensity of the enzyme. As discussed previously, MCM subunits are in equilibrium. In a nanopore sequencing experiment the concentration of the enzyme has to be equivalent to the number of adapter ends to ensure the integrity of the stall. However, the concentration of adapters is low (~100s pM). At this point, we have insufficient evidence to conclude whether MCM can efficiently assemble at these subunit concentrations. Therefore, one strategy would be to remove the equilibria by covalently forcing the enzyme into a hexamer around DNA (Figure 5.27a).

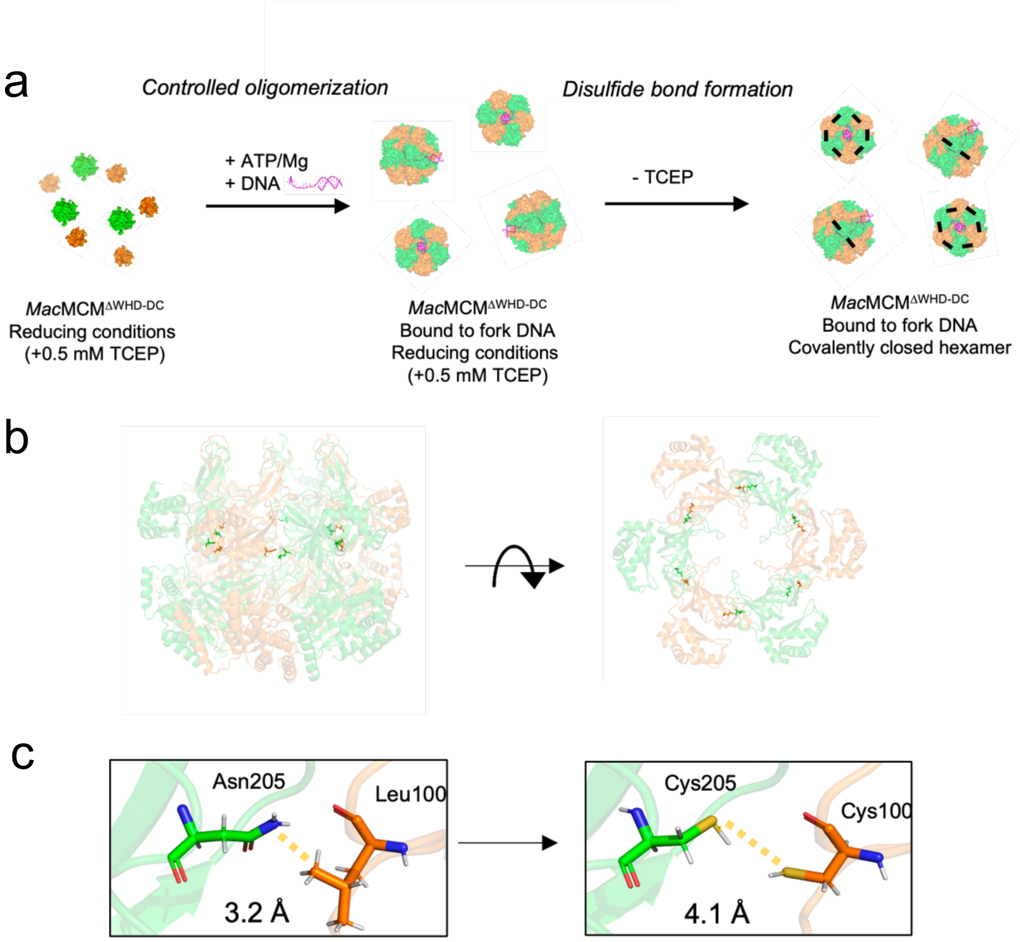


Figure 5.27: Design of a DNA-bound, covalently closed *Mac*MCM$^{\Delta WHD}$.
(a) Theoretical basis for *Mac*MCM$^{\Delta WHD-DC}$ cross-linking. Hexamerization of *Mac*MCM$^{\Delta WHD-DC}$ is induced through addition of ATP/Mg^{2+} and DNA. The concentration of reducing agent is slowly reduced through spin concentration-based dialysis. Two closely situated cysteine residues on opposing interfaces form a covalent disulfide bond. **(b)** The chosen location for engineered cysteine residues within the planar OB-fold ring of the N-terminal domain. Protein is shown in the ribbon format. **(c)** Cartoon representation of the residues selected for mutation to cysteine. Both leucine 100 and aspartate 205 are situated in neighbouring loop regions of the OB-fold.

To achieve this, a double-cysteine mutant was designed to generate a covalently linked hexamer. Based on our structural data of *Mac*MCM$^{\Delta WHD.E418Q}$, two closely situated residues were selected for mutation to cysteines: asparagine 205 and leucine 100. Both residues were present in two linker regions of the OB-fold that help to create the MCM subunit-subunit interface (Figure 5.27b). In both instances, the amino acid side chains project towards one another (Figure 5.27c).

Mutants were generated on *Mac*MCM$^{\Delta WHD}$ by mutagenesis and purified to near homogeneity as previously described (Figure 5.9). To ensure the reaction could be performed controllably, purification was performed in the presence of at least 0.5 mM reducing agent, where the final size exclusion chromatography buffer contained 0.5 mM TCEP. Protein was stored at a lower concentration of 3 mg/mL to lower intrinsic reaction propensity. Briefly, protein was extensively dialysed in a spin concentrator (~10,000-fold) in the crystallography buffer with the presence of 1 mM ATP/10 mM $MgCl_2$ and a short DNA oligonucleotide $(ACTG)_{16}$. After the final centrifugation, samples were placed in the fridge overnight at 4 °C. Subsequently, protein was then passed over a S6 10/300 size exclusion column equilibrated in the crystallography buffer.

Analysis of size exclusion profiles reveals that the enzyme can form hexamers in the absence of ligands (Figure 5.28a). The reaction mixture elutes with two peaks: the first peak exhibits at an elution volume comparable to that of the hexameric *SsoPfu*MCM, whilst the second peak is consistent with the monomeric *Mac*MCM$^{\Delta WHD}$ species. Fractions were then resolved by SDS-PAGE in the presence or absence of DTT. Without DTT, multiple bands are observed (Figure 5.28b). When DTT is added only single bands are present, suggesting that covalent products are mediated by cysteine residues (Figure 5.28c). Further inspection of reaction products suggests that at least 6 covalent species are present (Figure 5.28d). These bands correspond to a covalent hexamer and the intermediate reaction products. DNA is not observed to co-elute with DNA. It is likely that the assembly reactions require further optimization to improve the efficiency of covalent hexamerization and entrapment of DNA.

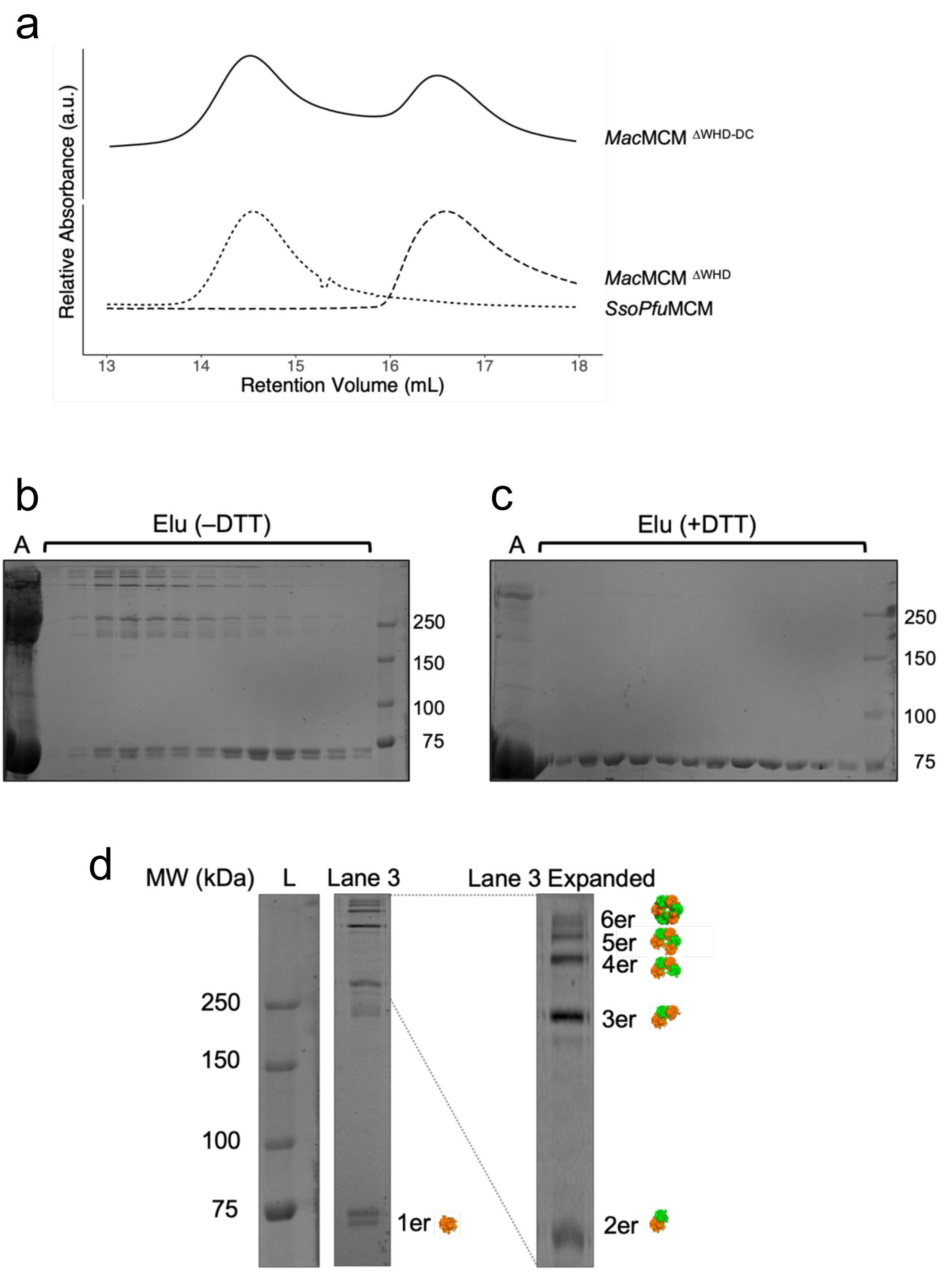


Figure 5.28: *Mac*MCM$^{\Delta WHD\text{-}DC}$ can controllably be assembled into a covalent hexamer.
(a) Complex formation was initiated by spin concentration dialysis. Protein was mixed with a solution containing 100 mM NaCl, 20 mM Tris-Cl pH 8.0, 5 % Glycerol, 1/10 mM ATP/Mg^{2+} and a 1:1 ratio of protein to (ACTG)$_{16}$. Protein was dialysed extensively, then incubated at 4 ºC overnight. Samples were then passed over a Superose 6 Increase 10/300 GL column in absence of ATP/Mg^{2+}. Absorbance of the eluate was monitored at 280 nm. **(b)** 0.4 mL fractions from part (a) were analysed by electrophoresis on a 12 % SDS-PAGE without DTT in the loading dye. Protein was visualized using Coomassie dye. **(c)** The gel in (b) was repeated, however, without DTT in the loading dye. **(d)** Magnified lanes from gel (b).

5.11 Chapter conclusions

This chapter outlines the preliminary work to assess whether MCM helicases can be incorporated into ONTs sequencing platform. ONTs own commercial motors are highly engineered to ensure sequencing occurs efficiently with a good signal to noise ratio. Currently, there is an interest to develop new motors that will reduce sequencing preparation steps and improve the error rate of sequencing reads. ONT sequencing demands motors that are active in a salty environment at room temperature. Previously we demonstrated activity of MCMs at room temperature. In this chapter, we identified the optimal buffer conditions for performing nanopore experiments. Crucially, glutamate should be used as the anion of choice for MCMs, whilst chloride is largely inhibitory. Adaptors are a central component to ONT sequencing workflow, and they must be tailored for each motor used. Each motor must be stalled in the correct orientation to perform sequencing. Here, 2x iSp18 moieties are sufficient for stalling an MCM, when the stoichiometry of protein: DNA is at 1: 1.

Initial ONT flow cell studies suggest that MCM may be unwinding DNA above the nanopore, however events are rare and noisy. We have begun to investigate mechanisms for reducing MCM-specific noise and increasing the capture rates of MCM-bound DNA. The resolved structure of *Mac*MCM has provided excellent information for engineering MCM. Thus far, we have demonstrated an ability to chemically modify MCMs and to generate covalently closed hexamers. Future research will need to evaluate the success of these engineering strategies in a flow cell.

Chapter 6 – Discussion

6.1 Thesis summary

MCM are an essential component of the replisome in both archaea and eukaryotes, catalysing strand separation ahead of the replication fork. Recent advances in remapping the archaeal tree of life suggest that archaea are not only related to eukaryotes, they are the direct ancestors [32,33]. Archaeal enzymes offer simplified models for understanding the core structure and function of shared homologous enzymes, such as MCM. It is acknowledged that many archaeal enzyme models like MCM, are not very representative of this newfound diversity and can be poor systems for studying the relationship between archaeal and eukaryotic enzymes [232]. Before this study, understanding of archaeal MCM largely stemmed from studies of MCM from two thermophilic organisms (*M. thermautotrophicus* and *S. solfataricus*), which are positioned in the centre of the archaeal tree.

ONTs MinION platform requires helicase enzymes that are active at room temperature. Based on historic enzyme studies, we speculated that traditional MCM models from thermophilies would not be active at room temperature [365]. The industrial collaboration with ONT has greatly facilitated the discovery of new archaeal MCM models. This thesis offers a fresh perspective on the relationship between homohexameric archaeal MCM and its heterohexameric eukaryotic descendent MCM2-7 and explores the use of MCM in biotechnological applications.

6.2 Identification of new MCM models

Adaptation of a cuvette-based real-time helicase assay into a high throughput format, has been instrumental for the characterization of new MCM models in this thesis [190,363]. It is now possible to routinely screen hundreds of different buffer, orthologue, and mutant combinations to rapidly compare and identify important characteristics of MCM. This thesis examined at least 150 unique reaction conditions using a 96-well plate format and there is scope in the future to increase throughput further into 392-well plate formats and beyond. Here, on a single reaction plate it was possible to increase the total number of MCMs ever biochemically characterized by a factor of two, whilst simultaneously comparing many of the traditional models. Notably, all the traditional thermophilic archaeal MCM models were straightforward to purify, however they were largely inactive at room temperature as expected. Conversely, many of the MCM from mesophilic archaea were more difficult to purify but more active at room temperature. Historically, MCMs have been examined using end-point band shift assays. The use of a real-time assay permitted for the first-time important observations about the population kinetics of DNA unwinding by archaeal MCM that have not previously been observed.

The sole MCM from the ectosymbiotic mesophilic organism, *Mancarchaeum acidiphilum* exhibited a distinctly unique kinetic event ('lag time') before maximum unwinding was achieved. Characterization of this kinetic event in *Mac*MCM biochemically and biophysically has profound implications for both evolutionary and biotechnological studies on MCM.

6.3 The link between *Mac*MCM lag time and MCM2-7 ring closure

Characterization of lag time in *Mac*MCM suggested immediate similarities with the '*slow kinetic step*' observed during closure of the MCM2-5 gate in eukaryotes [59]. Both kinetic events occur on a comparable time scale (minutes) that cannot be attributed to diffusion alone [59,60,157,164]. In both cases, the distinct kinetic step is dependent on oligomerization propensity. Gel filtration demonstrates *Mac*MCM self-associates poorly, whilst MCM2-5 subunits do not co-elute *in vitro* [58]. For *Mac*MCM, it was demonstrated that reduction of protein and ATP concentration extends lag time exponentially, consistent with ATP driven oligomerization (Figure 6.1a–b). Equivalently, well defined hexameric complexes such as *SsoPfu*MCM or the eukaryotic subcomplex MCM4,6,7 exhibit minimal lag-time [59]. This implies that the kinetics of hexamerization may be conserved from archaea to eukaryotes. In eukaryotes, the MCM2-5 gate is a structurally defined cleft through which dsDNA is threaded during replication initiation [163]. It is not well defined whether or how an equivalent gate exists in archaea, where all the subunits are identical. At a minimum, the formation of a gate would require some induction of asymmetry, which could occur spontaneously or by interaction with protein partners such as Cdc6/Orc [105].

ATP-dependent oligomerization has not previously been reported for an archaeal MCM. ATP-dependent oligomerization is observed not only for other hexameric helicases such as E1[254], SV40 L-Tag [366], and T7 gp41 [64], but also for hexameric AAA+ proteins, such as ClpB [367] and NSF [368]. The response of *Mac*MCM^{FL} lag time with respect to ATP concentration is notable (Figure 6.1b). When the concentration of ATP is below the estimated $K_{d.app}$ of *Mac*MCM ATP binding, lag time extends exponentially. One may speculate biological explanations for the relationship between ATP concentration and MCM regulation. Whilst the MCM from the similarly mesophilic *M. barkeri* was highly active, this enzyme lacks an obvious lag time. Notably, *M. acidiphilum* is an endosymbiotic species that lacks genes involved in key metabolic processes [18]. DNA replication is a metabolically expensive process for cells. Weak assembly under limiting ATP concentrations (or recruitment to origins) may allow MCMs to act as an autoinhibitory sensor that can prevent cell cycle progression.

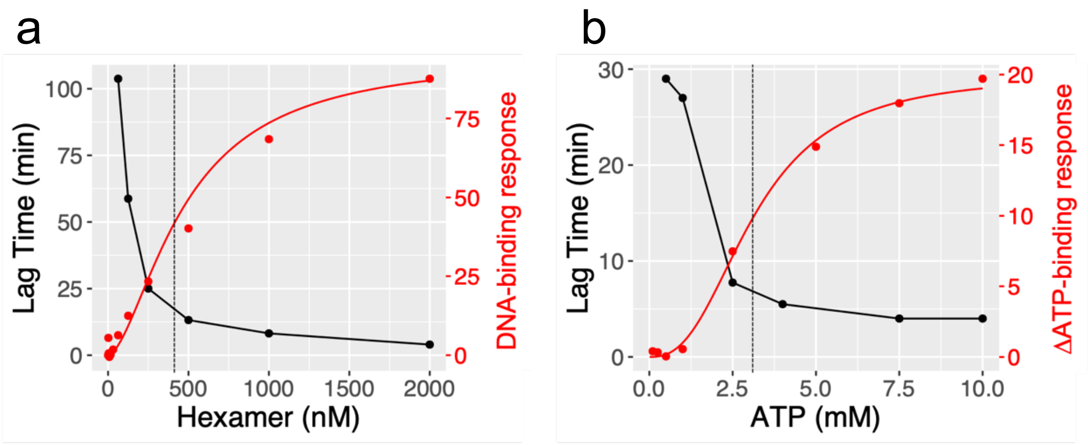


Figure 6.1: Direct comparison of factors influencing lag time.
(a) Lag time is directly compared against DNA-binding response measured at the same concentration of protein. The dotted line is the estimated affinity of MCM to DNA when no ligand is present. The red line represents the measured binding *Mac*MCM^{FL} to DNA by fluorescence anisotropy. **(b)** Lag time is directly compared against ATP-binding response measured at the same concentration of ATP. The red line represents the change in aggregation temperature in response to ATP measured by light scattering. The dotted line is the estimated affinity of MCM to ATP.

6.4 *Mac*MCM adaptations to MCM oligomerization

Based on the hexameric MCM structure solved in this thesis, two orthogonal *in silico* analyses suggest the *Mac*MCM interface is more eukaryotic in nature than MCM from thermophilic archaea. Protein flexibility is an important determinant of enzymatic activity, allowing efficient conformational change under the optimal growth conditions of an organism [369]. It is established that enzymes from thermophilic organisms typically contain a higher number of salt bridge interactions that may support stability at high temperatures [370]. Predictably, interfaces within the structure of the hexameric MCM from *S. solfataricus* contain over 2 and 4 times the respective number of salt bridges and hydrogen bonds compared to both *Mac*MCM and *Sce*MCM2-7. These interactions are assumed to stabilize *Sso*MCM at high temperature, however, at low temperatures we demonstrate *Sso*MCM is largely inactive. Previous studies conclude that salt bridges rigidify active sites making enzymes from thermophilic organisms less active at low temperature [181]. Removal of salt bridges previously in archaeal MCM has yielded clear changes in oligomerization propensity and activity [99,120].

Analysis of the degree and sequential distribution of buried residues within MCM subunits further implies that *Mac*MCM more closely resembles the core interface of MCM2-7 than *Sso*MCM. In the future, it will be beneficial to expand this analysis to include more MCMs. For example, enzymes from psychrophilic organisms (organisms that live in extreme cold), may be expected to contain even fewer inter subunit interactions [280]. We are currently limited by the lack of quality, near full-length hexameric MCM structures deposited in the PDB. Nevertheless, these results underpin the importance of choosing appropriate models that reflect the environmental restraints subjected to eukaryotic enzymes. This finding should be considered important beyond the field of MCM.

6.5 Analogous routes of MCM hexamer closure

Deciphering the steps that underpin lag time requires an understanding of the conformational change that MCM undergoes between monomeric and oligomeric states. Unfortunately, it was only possible to capture the closed hexameric state of *Mac*MCM, limiting the conclusions that may be drawn. It is however possible to use published structures to speculate about the conformational changes involved in ATP-dependent assembly of *Mac*MCM^{FL} onto DNA.

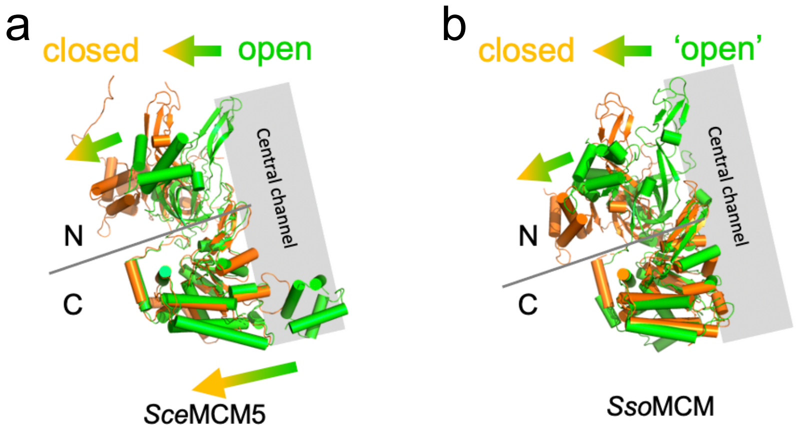


Figure 6.2: Global subunit conformational change involved in MCM oligomerization.
All protein are visualized in the ribbon format, where alpha helices are represented by cylinders. **(a)** *Sce*MCM5 was extracted from the closed (orange, PDB: 6EYC)[140] and open (green, PDB:5XF8)[158] MCM structures. **(b)** A subunit of *Sso*MCM was extracted from the hexameric (orange, PDB: 6MII)[100] and monomeric/ 'open' (green, PDB:3F9V)[123] structures. In both instances structures were structurally aligned in PyMol using the AAA+ domain as a reference. N and C represent the position of the N- and C-terminal domains respectively.

Recent cryo-EM studies suggest that *Sce*MCM5 undergoes large conformational changes between open and closed hexameric states [140,158,163]. The conformational changes are more prominent in *Sce*MCM5 than the neighbouring *Sce*MCM2 subunit, which lacks the WHD. In terms of sequence length, MCM5 is also the most comparable *Sce*MCM2-7 subunit with archaeal MCM. Using the core AAA+ domain as a reference for structure alignment, the N-terminal and winged helix domains of *Sce*MCM5 undergo significant rotation between open and closed states (Figure 6.2a). The N-terminal domain rotates about 25 Å (measured from the ZnF), whilst the WHD moves out of the central channel and is no longer identifiable in the electron density.

A similar analysis can be performed between two crystal structures of *Sso*MCM that capture the 'open'/monomeric and closed hexameric states [100,123]. The 'open' structure was elucidated with the WHD present in the crystallization construct, however the WHD was not identifiable within the electron density [123]. In this *Sso*MCM^{FL} structure, *Sso*MCM crystallized as a lattice of monomeric subunits. The structure was particularly low resolution at 4.4 Å, and yielded poor R-factors, of 0.41/0.48 (R_{work}/R_{free})[123]. The limiting data quality was probably due to the exceptionally high solvent content (75%) and large unit cell size. At this resolution, it is possible to identify the secondary structure within the electron density. When structurally aligned with a subunit from the closed hexameric structure of *Sso*MCM, the interdomain movement between N and C-terminal domains are strikingly like observations made for *Sce*MCM5 (Figure 3.2b), where the ZnF is observed to rotate 26 Å between monomeric and closed hexameric states.

Recent advances of *in silico* structure prediction techniques such as AlphaFold allow us to predict the structure of apo *Mac*MCM monomers with near experimental confidence [371]. A monomeric model of *Mac*MCM^{FL} was generated using AlphaFold[371]. Across the AlphaFold predicted model, the atomic positioning of residues within each domain are astonishingly accurate with respect to the experimentally derived data for *Mac*MCM$^{\Delta WHD.E418Q}$ (Figure 6.3a–b). The calculated all-atom RMSD are 0.7 Å and 0.9 Å for the N- and C-terminal domains respectively. The AlphaFold model was also able to correctly predict the tetrahedral co-ordination geometry by the 4 ZnF cysteine residues (Figure 6.3c). These highly accurate subdomain model predictions would have been beneficial during this thesis, which required extensive trial and error using MCM homology models to solve the phase problem by molecular replacement. Indeed, researchers elsewhere are already exploiting these predictions for effective molecular replacement [372]. Currently, it is not possible for AlphaFold to predict hexameric MCMs from sequence, although this will be expected in the future.

Whilst this apo *Mac*MCM model is perhaps somewhat speculative, the results are encouraging. The predicted rotation of the N and C-terminal domains between monomeric and hexameric states are on par with experimentally derived structural data for *Sce*MCM5 and *Sso*MCM at around 30 Å (Figure 6.3d). In the open form, the WHD is predicted to project away from the AAA+ fold into the putative central channel of the hexamer, like MCM5. From this apo *Mac*MCM^{FL} AlphaFold prediction alone, it is not possible to reconstitute a hexamer due to significant atomic clashes between neighbouring N-terminal and winged helix domains. This suggests that extensive remodelling is required, which is supported by our biochemical data.

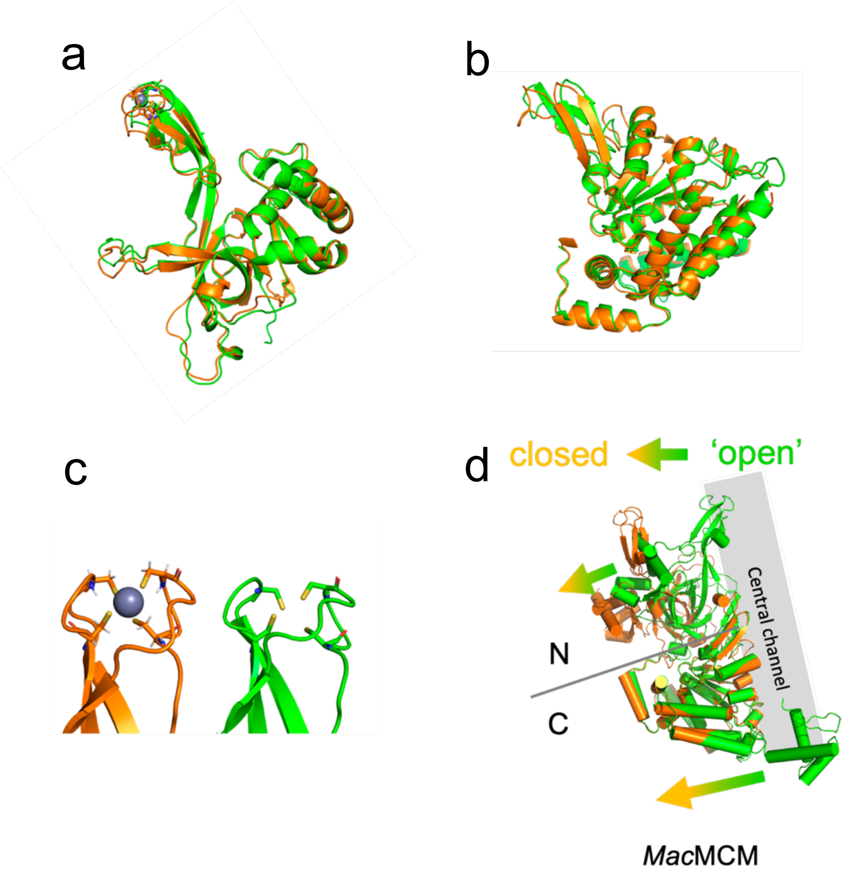


Figure 6.3: Predicted conformational changed involved in *Mac*MCM oligomerization.
A subunit of *Mac*MCM was extracted from the hexameric crystal structure derived in this thesis (orange). The structure of monomeric *Mac*MCM^{FL} was predicted using AlphaFold (green). **(a)** The N-terminal domains of the models were structurally aligned in PyMol and displayed in the ribbon format. **(b)** The C-terminal domains of the two models were structurally aligned in PyMol and displayed in the ribbon format. **(c)** The AlphaFold model correctly predicts the tetrahedral geometry of the cysteines in *Mac*MCM zinc fingers. The cysteine residues are represented by sticks. The grey sphere represents the zinc atom from the experimentally derived model. **(d)** The two models were aligned in PyMol using the AAA+ domain as a reference. N and C represent the position of the N- and C-terminal domains respectively. The models are displayed in the cartoon format, where alpha helices are represented by cylinders.

6.6 The winged helix domain: an autoinhibitory oligomerization regulator?

Ring closure of MCM2-7 around DNA is dependent on the repositioning of MCM5 WHD from the central channel of the hexamer [158,162,163]. This slow transition Is dependent on the hydrolysis of ATP, where inactive mutants or non-hydrolysable ATP analogues do not support efficient loading of MCM2-7 onto DNA [59]. Equivalently, *Mac*MCM^{FL} loading onto DNA was only supported when ATP hydrolysis was present. Non-hydrolysable analogues and inactive hydrolysis mutants were not able to efficiently load *Mac*MCM^{FL} onto DNA when measured by fluorescent anisotropy and SEC. ATP binding alone does not support DNA binding or hexamerization of *Mac*MCM^{FL}.

Using a *Mac*MCM$^{\Delta WHD}$ truncation mutant, it was possible to decouple the involvement of the WHD in assembly. Removal of the WHD in *Mac*MCM completely ablates the observation of lag time. It is also worth noting that the lag time event may still be present for *Mac*MCM$^{\Delta WHD}$, but not measurable under current time resolution of the assays used. Nevertheless, the data strongly suggests that the WHD has a regulatory purpose in *Mac*MCM lag time akin to MCM5 WHD. To the best of knowledge, the kinetics of MCM2-7 have not been determined in absence of MCM5 WHD, however it is speculated here that removal may significantly accelerate ring closure.

It is important to distinguish that the WHD alone does not itself prohibit hexamerization; removal of the WHD does not reconstitute a population of apo *Mac*MCM$^{\Delta WHD}$ hexamers. Equally, previously studied full length archaeal MCMs are generally reported to form hexamers in solution [82,98,99,101,104]. Instead, it is speculated that when the interactions of the hexamer interface are sufficiently weak (e.g., in enzymes from a mesophilic organism), the regulatory function of the WHD becomes measurable. Likewise, if the kinetics of interface formation are favourable where a high number of inter subunit interactions are present (e.g., *Sso*MCM), hexamers may form spontaneously at low temperature. Spontaneous hexamer formation would forcibly bypass inhibitory effects of the WHD that typically limit oligomerization. This accounts for the observed hexameric state and lack of lag time observed in previously studied full-length archaeal MCM [82,104].

6.7 How does ATP support hexamer formation?

Where the WHD is absent, ATP binding alone is sufficient to support the formation of *Mac*MCM$^{\Delta WHD}$ homohexamers. This suggests that ATP has a cooperative role in MCM hexamerization. Alluding to previous MCM models, it is possible to discern important conformational changes involved in ATP binding and hexamerization.

During ATP dependent MCM2-7 ring closure, an additional helix becomes structured into the electron density of the MCM5 AAA+ domain (Figure 6.4a)[139,158]. Equally, between monomeric and hexameric structures of *Sso*MCM, significant conformational change exists for the equivalent helix (Figure 6.4b)[100,123]. Here the *Sso*MCM helix undergoes a 90 ° rotation when nucleotide is present. The cooperative restructuring of this helix upon ATP binding has two potential routes and roles.

First, this helix provides residues that form a hydrophobic pocket on the *cis*- side of the active site. Functionally, this pocket may support binding of the hydrophobic adenosine

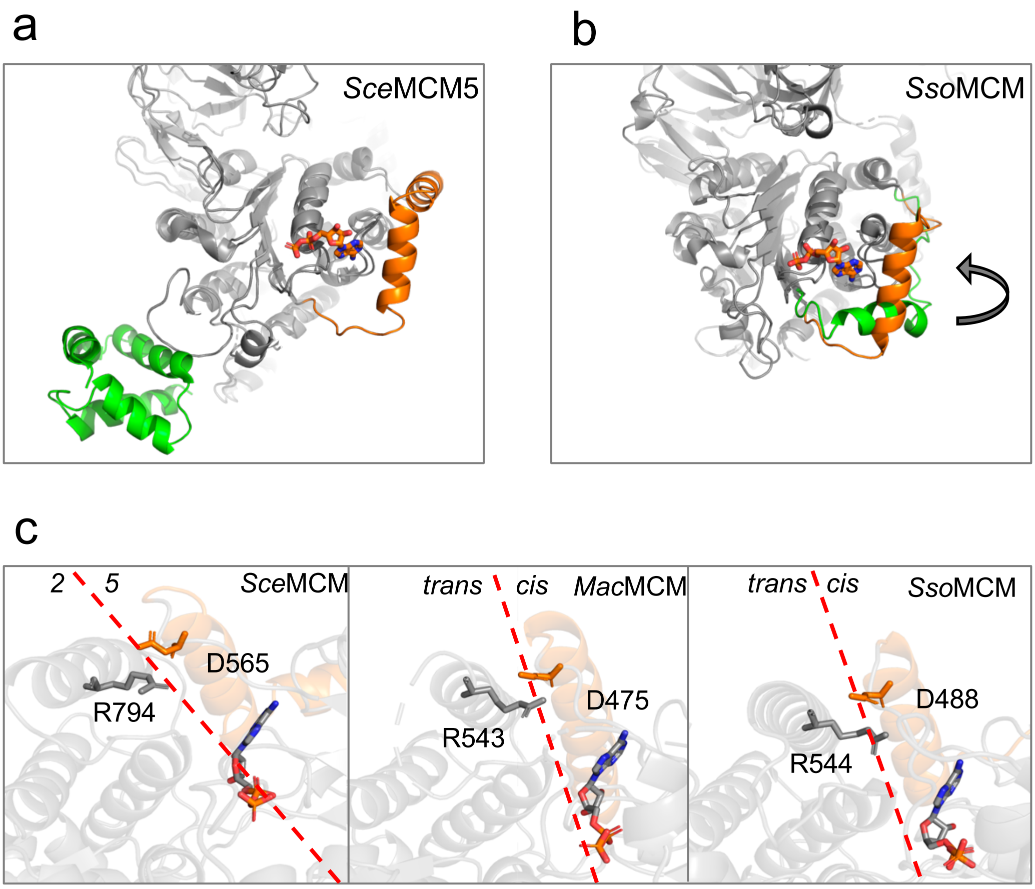


Figure 6.4: Conformational change in a helix supports ATP dependent oligomerization. **(a)** Conformational changes in the C-terminal domain of MCM5 (ribbon format) observed between open (PDB: 5XF8)[158] and closed (PDB: 6EYC)[140] MCM2-7. **(b)** Conformational changes in the C-terminal domain of *Sso*MCM (ribbon format) observed between monomeric (PDB:3F9V)[123] and hexameric (PDB:6MII)[100] structures. Between the two states, a helix rotates 90º to accommodate ATP. In both parts (a) and (b), colouring highlights the major differences in conformational state (open: green, closed: orange). Nucleotide is represented in sticks in orange. **(c)** ATP dependent restructuring of a structurally conserved helix (orange) promotes formation of an inter subunit salt bridge. The salt-bridge pair are represented as sticks.This is conserved between all 3 MCM protein structures (ribbon format).

moiety. ATP mutagenesis studies have largely focused on the charged phosphate co-ordinating motifs required for ATP hydrolysis [99]. However, ATP is an amphiphilic molecule, and the precise contributions of the hydrophobic interactions are almost completely unexplored.

Second, this *cis*-acting helix when formed interacts directly with a helical bundle of the neighbouring subunit (e.g., MCM2) to form a stable interface. Formation of this helix-helix interface between MCM2 and 5 concludes MCM2-7 loading onto DNA [163]. Removal of this helix in *Sso*MCM yields inactive mutants, which are monomeric in solution [123]. Subtle mutation of a single salt-bridge from this inter subunit helix pair (D488A) is sufficient to yield a monomeric phenotype for the apoenzyme. This *Sso*MCM D488A mutant also exhibited elevated helicase activity [99]. The helicase activity of *Sso*MCM D488A was previously assessed under end point band shift assays and so the population kinetics of this mutant remains unknown [99]. One may speculate that D488A would share similar kinetics to *Mac*MCM

The salt bridge forming aspartate residue is conserved from *Sso*MCM to *Mac*MCM and *Sce*MCM (Figure 6.4c). By probing the different roles of this helix further in *Mac*MCM^{FL}, it may be possible to generate monomeric mutants of *Mac*MCM with different, fixed conformational states. This would allow further structural investigation into the conformational changes involved in *Mac*MCM^{FL} hexamerization. Removal of key

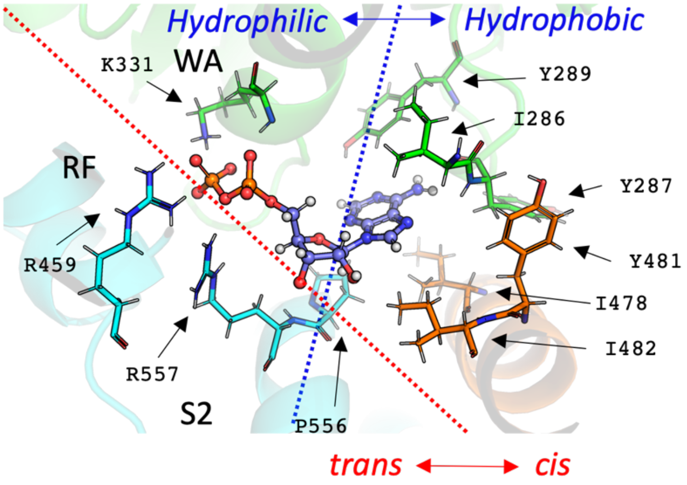


Figure 6.5: ATP binding to the active site of *Mac*MCM supports oligomerization.
Residues from neighbouring subunits co-ordinate binding of the phosphate groups of ATP. The adenosine moiety binds to a hydrophobic pocket in the *cis* side of the ATPase active site. This hydrophobic pocket is formed on addition of ATP by restructuring of an alpha helix (orange). This alpha helix is then able to form inter subunit salt bridges with the neighbouring subunit. Other *Cis*-acting residues are coloured in green, *trans*-acting residues are coloured in cyan. ADP is represented as a ball and stick model. WA: Walker A motif, RF: Arginine finger, S2: Sensor-2 motif.

hydrophobic residues in the *cis* side of the active site, may limit binding of ATP to MCM subunits and therefore prevent the formation of the hydrophobic pocket. Removal of the conserved salt bridges formed by the helix formation may allow binding of ATP to the subunit and remodelling of the helix without hexamerization.

Within the hexameric *Mac*MCM structure, the hydrophobic pocket forming helix is structured around ADP (Figure 6.5). *Cis* and *trans* residues co-ordinate the phosphate residues as expected. Co-operative binding of ATP by both neighbouring subunits may support the formation of the active site and support oligomerization.

6.8 How does the winged helix domain limit oligomerization?

Historical studies have implied that removal of the winged-helix domain improves ATP hydrolysis ~2-fold [83]. ATP hydrolysis is required for *Mac*MCM^{FL} binding onto DNA; thus, it is important to determine whether the WHD limits assembly by modulating ATP hydrolysis or by acting as a steric block to hexamerization and DNA binding. The data presented here would seem to support the latter hypothesis. Using a ^{31}P NMR assay, we were unable to demonstrate a clear increase in ATPase activity where the WHD is removed. However, it is important to note that this assay was performed under highly saturating concentrations of ATP (50 mM), that are over 16-times the estimated $K_{d.app}$ of MCM to ATP (3.14 mM). Other MCM studies suggest that saturation of the AAA+-fold may cause substrate inhibition [81]. It is possible to investigate ATP hydrolysis under more physiological conditions using alternative NMR-based techniques [373]

Gel filtration and order of addition helicase experiments imply that DNA and ATP hydrolysis are required together for *Mac*MCM^{FL} to cooperatively assemble into a DNA bound homohexamer. This suggests that DNA may have role in repositioning the WHD in *Mac*MCM. However, studies imply that the WHD of MCM does not itself bind to DNA tightly [83], instead the WHD primarily acts as a recruitment motif for ORC [105,162,163]. Therefore, MCM interactions with DNA that promote repositioning of the WHD must come from alternative structural motifs. It is demonstrated here and elsewhere that hexamerization and DNA binding are synergistic: hexamerization is required for DNA binding by the N-terminal OB-fold of MCM where co-elution with DNA does not occur with monomeric oligomers [119,128]; DNA stabilizes the formation of a MCM hexamer [119].

Comparisons of cooperative DNA binding can be drawn from DNA-bound OB-fold structures of alternative proteins such as SSB [374]. In SSB, a positively charged cleft is formed between two beta-hairpin turns that flank the antiparallel beta-strands [131]. Single-stranded DNA binds through this cleft via electrostatic interactions (Figure 6.6a). However, in MCM one of the beta-hairpin turns is absent and is instead the site of the structural ZnF insertion (Figure 6.6b). In part, the lack of a functional cleft may underpin why MCM monomers fail to associate with DNA. Instead, it is possible to form 6 functional DNA-binding clefts when an MCM is arranged as a hexamer (Figure 6.6c). In this instance one can consider each complete OB-fold motif to be formed of both *cis*- and *trans*-acting residues. Neutralisation of negatively charged DNA by basic residues from adjacent subunits may cooperatively support MCM hexamerization. Supporting this hypothesis, in the *Mac*MCM hexamer structure, phosphate co-ordination by both *cis*- and *trans*-acting basic residues is observed.

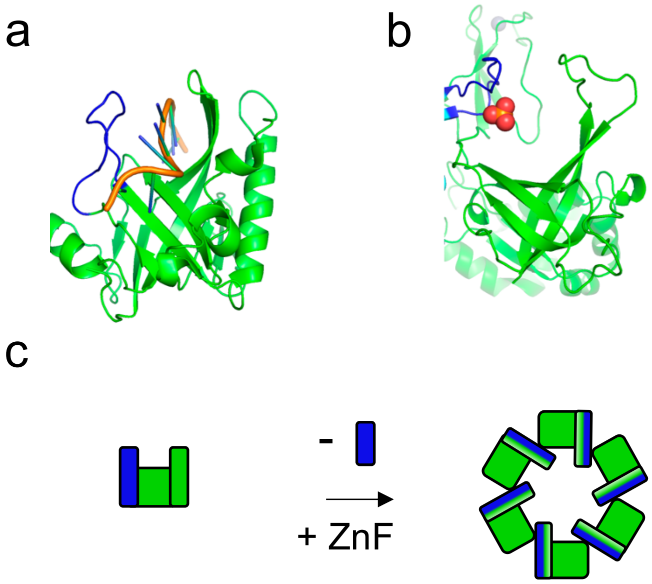
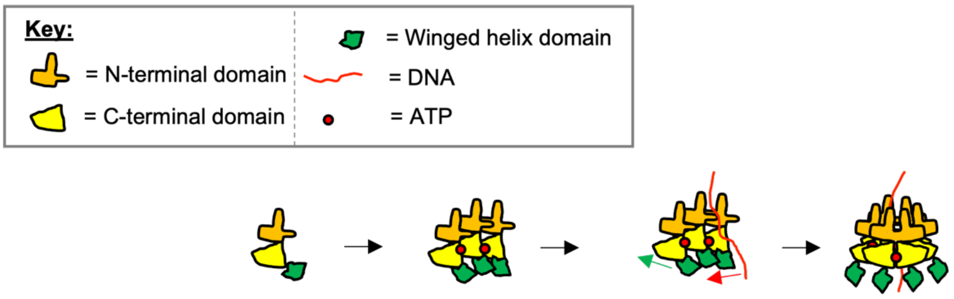


Figure 6.6: Cooperative DNA binding by adjacent subunits.
(a) DNA-binding (orange) in the OB-fold of SSB (PDB: 5ODL)[374]. Protein is shown in the ribbon format. **(b)** Phosphate-binding (orange spheres) to the OB-fold of *Mac*MCM. The blue hairpin represents the contribution of the neighbouring subunit to form a functional DNA-binding cleft. Protein is displayed in the ribbon format. **(c)** Schematic outlining how a single hairpin can form 6 functional clefts in an MCM hexamer.

The role of the C-terminal DNA-binding PS1β hairpins in *Mac*MCM hexamer formation were not evaluated in this study. Interestingly, mutation of PS1β residues on MCM5 is linked with Meier-Gorlin dwarfism [375]. This study concluded that pathologically MCM5 and MCM2 display decreased levels of chromatin association and therefore the PS1β hairpin may also be involved in initial recruitment and assembly.

Taken together, DNA interacting at distal sites along the MCM hexamer may provide means to support hexamerization and repositioning of the WHD by DNA. Indeed, cryo-EM studies suggest that as DNA is loaded into the MCM2-7 ring by ORC, the MCM5 WHD is displaced from the central channel by dsDNA [163]. In *Mac*MCM^{FL}, ATP binding may support higher order oligomerization, whilst ATP hydrolysis may allow translocation/conformational change along DNA that supports displacement of the WHD. A spiral staircase translocation mechanism[100], where DNA binds to multiple subunits around and through the ring would also support DNA-mediated hexamerization, where the DNA may act as an inverted belt. The inability for either ATP or DNA alone to limit the rate of *Mac*MCM^{FL} lag time supports synergistic roles in ligand mediated MCM assembly, which is consistent with eukaryotic MCM2-7.

Table 6.1: Co-operative ATP hydrolysis and DNA-binding mediate slow hexamerization of *Mac*MCM^{FL}.



Condition	No ligand	+ATP hydrolysis/ATP binding	+DNA +ATP **hydrolysis**	+ Time
State	Monomer	Efficient hexamerization inhibited by WHD	ATP hydrolysis and DNA binding drives efficient repositioning of WHD	Stable DNA bound hexamer formed on DNA
Supporting evidence (technique)	When no ligand is present *MacMCM* primarily exists as a monomer in solution (SEC-MALLS/AUC)	When ATP is added to *MacMCM* or *Mac*MCM$^{FL.E418Q}$, neither enzyme elute as a hexamer (SEC) When the WHD is removed, ATP binding alone is sufficient to promote formation of a , *Mac*MCM$^{\Delta WHD.E418Q}$ hexamer (SEC) Reduction of ATP concentration increases lag time exponentially (helicase assay)	*MacMCM* binds to DNA only when ATP hydrolysis is present (fluorescent anisotropy) DNA and ATP are both required together. Neither component supplied alone reduces lag time (order of addition helicase assay) Reduction of DNA binding increases the dependency of protein concentration on lag time (helicase assay)	A stable DNA-bound *MacMCM* hexamer is observed when the previous conditions are met (SEC)

6.9 Future directions for *Mac*MCM

This thesis has identified an archaeal MCM with properties and broad conformational changes analogous to the eukaryotic MCM2-7. Future studies will need to address at near atomic resolution the conformational changes occurring within the enzyme with a focus on loading onto DNA and unwinding. Now many of the core properties of this enzyme have been characterized, it will be beneficial to further explore the role of accessory proteins in assisting MCM function in both replication initiation and synthesis. Notably, the genome of *M. acidiphilum* encodes homologous proteins to both eukaryotic Cdc45 (RecJ), GINS and Orc/Cdc6 [18]. Furthermore, a greater focus may also be placed on the substrate choice. In this thesis, the DNA onto which MCMs are loaded are generally non-specific single stranded or forked DNA substrates. It would also be interesting to investigate the loading of *Mac*MCM onto specific *M. acidiphilum* double-stranded origin of replication sequences. Moreover, since the start of this PhD, numerous important archaeal phyla have been identified, such as Asgard [35]. These are widely regarded to be the archaeal lineage from which eukaryotes evolved. Therefore, examining the properties of a mesophilic Asgard archaeal MCM will be of great interest. This will be possible through the robust MCM characterization pipelines outlined in this thesis.

Archaeal MCM studies are still limited to inferring structure-function relationships by linking biochemistry and crystal structures. Crystal structures, whilst generating high resolution data, represent an average conformation for all molecules in a crystal lattice. Comparably, techniques such as cryo-EM permit observation of different conformational states of an enzyme. *Mac*MCM is uniquely suited for cryo-EM. First, *Mac*MCM is a large roughly symmetrical enzyme, which immediately lends itself to the benefits of cryo-EM analysis [94]. Second, the resolved crystal structure of the *Mac*MCM hexamer, will allow docking of subunits into calculated electron density maps [376]. Third, we have developed robust protocols for controlling the oligomeric state of MCM through: mutagenesis; incubation of ATP and DNA; use of chemically modified DNA substrates; chemical crosslinking. Fourth, the relatively slow conformational change of *Mac*MCM (minutes), may permit exploration of time-resolved conformation dynamics [157]. Many standard macromolecular conformational changes are significant orders of magnitude faster than *Mac*MCM, and require specialized stop-flow style setups to capture dynamics [157,377]. Alternatively, slowly hydrolysable ATP analogues, such as ATPγS may also achieve this [260]. Fifth, we can purify *Mac*MCM to exceptionally high purity due to its monomeric purification status. As the enzyme does not form a DNA-binding hexamer, co-purification of trapped

DNA that may lead to misinterpretation of results is less likely. The open state of *Mac*MCM also allows greater accessibility to DNase treatment. We observe 0 % DNA contamination in *Mac*MCM purifications, whilst partial contamination was observed in many of the other MCMs tested even after high salt wash treatment. Further, the lack of subunit-subunit interactions prevents co-purification of partially degraded contaminants that may be difficult to separate.

Cryo-EM can be well complemented by a range of other dynamic biophysical techniques that employ NMR and fluorescence [378]. To date, there exists minimal studies examining the solution dynamics of MCM via NMR spectroscopy [118]. Robust NMR toolboxes exist for exploring the dynamics of proteins up to 1 MDa in size [379]. There are numerous bottlenecks for studying proteins by NMR spectroscopy, pertaining mainly to the yield, stability and solubility of a given sample [380]. High sample solubility is important to ensure that proteins can be concentrated to a high enough degree that a strong NMR signal can be obtained. The solubility of *Mac*MCM has been recorded as high as 360 µM in certain buffers. High stability ensures that proteins do not degrade over the course of dataset collection. The stability of *Mac*MCM is impressive, with a baseline T_m of 62 °C, whilst samples have been left at room temperature for over a week and retained 100 % activity and structure (data not shown). Many bespoke NMR experiments require the use of expensive isotopic labels that are incorporated during cell growth, hence achieving a high yield is an important economic consideration [381,382]. *Mac*MCM routinely yields > 40 mg/ L of culture for a 3-step purification. NMR does not require the same degree of purity as crystallography, and a 2-step procedure may further improve yields to ~60-80 mg/ L [383]. It is noted protein yields from minimal ^{2}H labelled media will likely be lower than standard LB media[384].

Many [^{1}H ^{13}C]-methyl-labelling strategies exist for exploring the conformational dynamics of both local and global conformational change within large macromolecular complexes. This could be used explore the changes in nuclei environment in response to addition of ligands such as ATP and DNA. An ambitious NMR approach would deploy [^{1}H ^{13}C]-labelling of methyl groups on amino acids such as Ile, Leu and Val (ILV), in an otherwise perdeuterated protein. This study would explore global conformational changes of entire MCM complexes, as has been achieved previously in other enzyme systems [385]. This strategy does not require full NMR resonance assignment of an entire MCM. For a *Mac*MCM hexamer, each NMR spectrum may contain up to 6 x (I = 55, L = 51, V = 38) unique NMR peaks that would be difficult to identity. It may be possible to transfer assignments from a simpler apo

*Mac*MCM monomer to larger oligomers. Structure-guided mutagenesis of residues (ILV → X) can permit discrimination of functionally important peaks.

Alternatively, local dynamics may be explored by chemical introduction of a single NMR probe onto each *Mac*MCM subunit. As explored in chapter 5, we have proven an ability to site-specifically label engineered cysteine mutants, whilst maintaining the integrity of the cysteine rich ZnF. Many ^{13}C thiol reagents, such as S-methyl-[^{13}C] methanethiosulfonate (MMTS), are commercially available and can be easily incorporated onto reactive cysteine residues [386]. This labelling strategy permits observation of local conformational dynamics during complex formation without requiring deconvolution of a complex spectrum. Obvious targets for labelling include the N-terminal OB-fold, the AAA+ domain and the WHD. Ideally, each [^{13}C]-labelled hexamer will contain 6-probes, hence labelling will also inform of any conformational asymmetry.

Understanding the dynamics of MCM in greater detail will further establish the relationship between archaeal and eukaryotic MCM, whilst providing us with valuable information for engineering MCM into ONT sequencing applications.

6.10 Summary: MCM for use in ONT sequencing applications

Since the initial release of the MinION early access programme in 2013, ONT sequencing has become a cornerstone sequencing product for numerous fields. Many of the early issues associated with nanopore sequencing, such as low accuracy, have been solved through improvements to both the data analysis and biochemistry within the flow cell itself. The biochemical components within the flow cell such as the pore, motor and buffer have been extensively engineered so that sequencing occurs as efficiently as possible with minimal noise [323]. Assessing the performance of new motors in a Nanopore flow cell is not trivial and requires years of understanding and work.

Biological components of a nanopore sequencing experiment may contain intrinsic biases towards certain DNA sequences. It is underexplored, but the use of different motor enzymes may improve the accuracy of DNA sequencing [351]. Each motor is expected to interact with DNA using different mechanisms, and hence will entail a signature error profile. These may account for mechanistic differences between enzymes when translocating along DNA, such as alternative interactions (base-stacking versus electrostatic interactions) or rate differences between AT or CG tracts [75]. When multiple motors are combined the error may be accounted for thereby allowing better discrimination of the nucleotide sequence.

The first biotechnological aim was to assess whether an MCM can be used as the motor within the MinION flow cell. Many of the initial concerns related to whether an archaeal MCM could interact with and unwind DNA under the high salt required for ONT sequencing. It was demonstrated in chapter 5 that MCM retained high activity, however only when the anion was glutamate. It was also shown that an MCM-specific adapter could be developed by efficiently stalling the enzyme by chemical modification of the encircled strand. It is also noted that only *Mac*MCM$^{\Delta WHD}$ exhibited processivity of around 90 bp, at low nM concentrations of hexamer. Longer substrates may be examined in the future. Preliminary Nanopore experiments suggest that there may be events consistent with enzyme-controlled movement, however the events are rare, noisy, and inconclusive. In relation to this first aim, future work needs to address these issues.

6.11 Future directions for use of MCM as a motor

6.11.1 Pore capture

To improve the pore capture rate of MCM-bound DNA we must consider the factors affecting *Mac*MCM DNA binding. As demonstrated in chapter 3, *Mac*MCM assembly is dependent on protein concentration, ATP, and DNA. Most of our understanding about oligomerization comes from gel filtration at protein concentrations above 2 µM, however, in an ONT sequencing experiment this needs to be equimolar with the DNA (pM) to ensure the efficiency of the stall. It is unknown whether *Mac*MCM will form a hexamer at these concentrations. For example, a radiochromatography technique suggests that at at 9 nM, *Mth*MCM is predominantly a monomer in the absence of ATP or DNA [104]. It may be possible using our fluorescent labelling approaches to determine oligomeric state by sensitive single molecule techniques [387,388].

Structure guided mutagenesis of *Mac*MCM$^{\Delta WHD}$ has allowed design of a covalently cross-linked MCM variant. This has the advantage that hexamerization is decoupled from an equilibrium process that dissociates the enzyme on dilution. It has been proven that the cross-linked MCM is able to form a covalent hexamer in solution. The reaction process is yet to be fully optimized, and hence yield is currently insufficient to determine activity using the helicase assay. In theory, cross-linking *Mac*MCM around adapter DNA will prevent disassembly of the motor when captured by the pore.

Another question that needs to be addressed is whether an MCM tolerates force on the captured, MCM encircled strand. Electrophoretic DNA when translocating through a pore exerts a strong force on the helicase enzyme. Many hexameric helicases such as DnaB and T4 bacteriophage gp41 are intolerant to forces supplied to the encircled strand [389,390]. DnaB exhibits stimulation when force is applied on the excluded strand [389]. As DnaB is a 5′–3′ helicase, it encircles the lagging strand during DNA synthesis. Strand dependency on force may mimic the force experienced by DnaB when polymerase directly synthesizes the leading strand. Lower forces may be expected in nature on the lagging strand, which is synthesised towards the helicase. Indeed polymerase mediated force on the leading strand has previously been demonstrated to stimulate the unwinding activity for MCM [72,73]. However, the 5′–3′ motor gp41 was intolerant to forces applied to either strand suggesting that there is no comprehensive model that describes the force dependency of hexameric helicases [390].

6.11.2 MCM noise

Protein-dependent noise likely stems from the large relative size of MCM versus traditional monomeric sequencing motors. The size of MCM is comparable to that of the CsgG nanopore itself [346]. It is likely that upon pore capture MCMs do not sit stably on top of the pore, instead the enzyme likely moves around causing significant, stochastic disruption to the current flow. Reduction of enzyme motion around the pore could be achieved by improving the interaction between the pore and the motor. This could be through either targeted mutagenesis of MCM to generate a favourable interaction with the pore, or by physically coupling the enzyme to the pore via tethering [364]. An additional source of noise is the direction of DNA translocation through the pore. Studies suggest that 3′–5′ translocation is generally noisier than 5′–3′ translocation, hence ONTs DNA sequencing motor unwinds with 5′–3′ polarity [391]. It may therefore be useful to evaluate alternative ways of capturing MCM bound DNA. This could include capture of the lagging 5′ strand of DNA (Figure 6.7b), or assessment of alternative ‘inny’ or ‘outy’ sequencing approaches (Figure 6.7c). Nanopore sequencing has largely used the ‘inny’ approach, whereby DNA is pulled through the nanopore under force. Here the motor is acting as a brake to limit translocation [392]. The alternative ‘outy’ approach uses a motor to pull single stranded DNA back through the pore against the electrical field [392]. This approach has not yet been widely adopted for nanopore sequencing. Whilst it is possible to strip a double stranded section of DNA using a nanopore [393], MCM are able to mid-load onto and translocate along ssDNA.

The self-loading ability of MCMs suggests MCM could be compatible for an adapter-less 'outy' approach of nanopore sequencing.

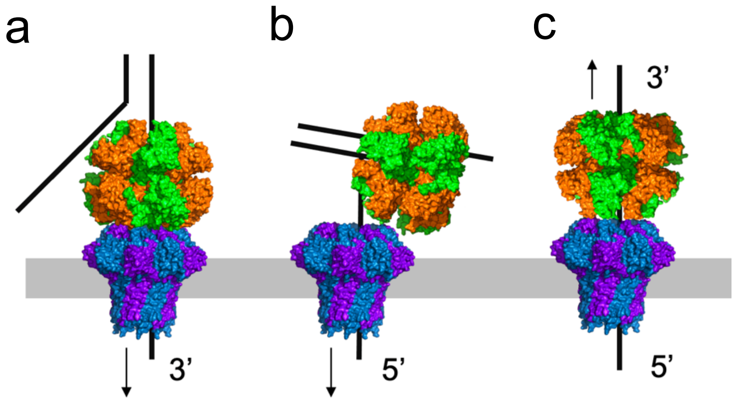


Figure 6.7: Possible orientations of MCM in ONT experiments.
All proteins are visualized in the surface format. **(a)** MCM encircles the strand of DNA which is captured by the pore. The MCM unwinds DNA in line with the electrical field. **(b)** The excluded strand is captured by the pore. This may lead to better signal to noise ratio through 5'-3' pore capture. **(c)** ssDNA is pulled back through the pore by MCM against the electrophoretic force.

6.11.3 Future development of an integrated motor-pore

Assessing whether an MCM is functional in an ONT flow cell in bulk solution is an entirely different question to whether a membrane embedded MCM is active. The structure of *Mac*MCM helicase allows speculation of how this goal may be achieved through either direct insertion of the whole MCM or by coupling MCM to a pore. Recently, the performance of E1 helicase was assessed when embedded within a membrane [394]. In this approach, E1 helicase appears to self-assemble as a hexamer within a lipid bilayer. This enzyme retained activity and using electrophysiology it was possible to determine the size of a translocating DNA strand [394]. However, at this stage no base-level discrimination has been proven. An alternative engineering method used chemical engineering to conjugate hydrophobic porphyrin moieties onto the surface of a thermophage portal protein [395]. Porphyrin subsequently allowed docking of the enzyme into the membrane. These approaches may be adapted for MCM, however the size of the MCM pore is possibly too wide for it to be used as the sole sequencing pore. Usually, a nanopore is only wide enough to allow translocation of ssDNA, however the central channel of MCMs is large enough to accommodate dsDNA. This may let unrestricted transport of any DNA through the pore. Furthermore, the large pore size may yield a poor signal to noise ratio, that prevents nucleobase discrimination.

An alternative approach may instead be derived whereby a hexameric pore forming sequence is engineered onto the C-terminus of MCM. In nature, some hexameric AAA+ ATPases, such as mitochondrial YME1L, natively form membrane tethered complexes [396]. This approach may include *de novo* design of membrane-embeddable pores or conjugation biologically existing pores onto an engineered MCM [397,398]. This approach would tailor the constriction point of the pore to prevent dsDNA translocation. Importantly, any successful approach will need to prove the activity of the enzyme, a good signal to noise ratio and an efficient return on the number of bases sequenced per minute. Achievement of this goal in the future will ultimately allow ONT to perform nanopore sequencing without any library preparation step. All the users would need to provide is DNA.

Chapter 7 – Supplementary Information

7.1 DNA and protein sequences of recombinant MCM

The core two enzymes (*Mac*MCM and *SsoPfu*MCM) are annotated with highlighted mutation sites. The His_{10} tag cleavage site is also noted (|).

>*Mac*MCM

```
ATGGGCAGCAGCCACCACCACCACCATCATCATCACCACCACAGCGGTGGCAGCGGTGGCGAAAACCTGTACTTCCAGGGTATGG
CGGAGCAGCAAACCAGCAGCCTGAAGCTGCTGTTCGACGAATTTCTGGAGAGCTACTATAGCGATGAAATCAAAGACATCATTAT
CAAGTTCCCGAACAAACGTAGCCTGCCGGTGAACATCAGCGATCTGGAGGAATTTGATCCGGACACCGCGACCAACCTGATTGCG
GATCCGGAAATTATCATTGACGCGGCGAACGAGAGCCTGATGGGCAAGCTGGCGGGTCTGAACTTCGACACCTACATTCCGCACG
TGCGTTTTTATAACCAAAGCATCAACACCCCGATGGTGCTGAACGTTGGTAGCGCGTATATTAACAAATTTGTTAGCATCGATGC
GCTGGTTGTGAAGCGTAGCGATATTCGTCCGAAAATCCGTGACGCGGTGTTCGTTTGCACCTTTTGCAACGCGAAGGTTAAAGCG
AACCTGGAAAAAGAGGAGATCCCGAAAGTGTGCCCGGAGTGCAAGAAACGTACCCTGAAGATTGTTCCGGAGGAAAGCAGCTTCT
TTAACAGCCAGAAAATCGCGGTTCAAGACCCGCTGGAACGTCTGAGCGGCAGCATTCCGACCTGGCAGCTGGAGGCGTGGCTGGA
CGATGACCTGGTGAACATGGCGATCCCGGGCGATCGTATCGAAATTAGCGGTGTTCTGAAGATTCGTCCGCGTAAAGATAGCCGT
GGCAAGGTTGACCCGAGCATCTACAGCATGTATCTGAACGTGACCAGCCTGGAAACCAAGCAGAAAGAGTTCGCGGATATCGACA
TTAGCGAAGACGAGGAACGTCAAATTAAGGAACTGAGCAAAGATCCGGAGATCTTTAACAAGGTGACCCAAAGCGTTGCGCCGAG
CATTTACGGCTATAACGAGATCAAACAGGCGGTTGCGCTGCAACTGTTTGGTGGCACCCCGGGTAAGAAACTGGTTGATGGTGGC
CAGATCCGTAGCGACATGCACATCCTGCTGATTGGCGACCCGGGTAGCGCGAAGACCCGTATTCTGCAAAGCGTGAGCCGTCTGG
TTCCGAAGGGCATCTACGTGAGCGGTAAAAGCGTTACCGGTGGCGGTCTGACCGCGGTGGCGGAACGTGATGACTTCAGCGAGGG
CGGTTGGACCCTGAAAGCGGGTGCGATGGTTCTGGGTAACGGCGGTATCGTGGCGATTGATGAGTTTGACAAAATCAGCGAGGAA
GACACCGCGGCGCTGCATGAAGCGCTGGAGAGCCAGACCATTAGCGTGGCGAAGGCGGGCATCATTGCGACCTTCAACGCGAAAG
CGAGCGTTCTGGCGGCGGCGAACCCGAAGTTCGGTCGTTTTGATCCGCACAAATACCCGGCGGAACAGTTTGACATCAGCCCGAC
CCTGCTGAGCCGTTTCGATCTGATCTTTCCGATTCGTGATATCATGGACACCGAGCTGGACAAGAGCATTGCGAACTATATCCTG
AACCAACACGAAGCGGCGGGTGCGGCGATTGCGGATGTTGAGAGCAGCGTTGCGGCGGAGGAACCGCCGATTGAACACAGCCTGC
TGAAGAAATACATCGCGTATGCGAAACGTTACGTGATGCCGCGTCTGAGCGAGGAAGCGAGCAACCGTATCAAGGAGTACTATGT
TGACCTGCGTCGTGCGGGCAGCATGAAAGGTGCGACCCCGATTACCCCGCGTCAGATTGAAGGTCTGATCCGTATGGCGGAGGCG
AGCGCGAAGAGCCAACTGCGTGATGTGGTTAGCGTGAAAGACGCGAACCTGGCGATCAGCCTGAGCGAATACATGCTGAAGACCC
TGGCGGTTGACACCGAGGGTCGTACCGATATTGACACCATCCTGACCGGTATGCCGCGTGAGAAGGTGGATAAAATTAACGTTAT
CCTGGACGCGGTGAAGAAACTGGAGGAAATGGATGGCAGCGCGAAAATTGACCGTATCTTCGAGGAAGTTGCGAAACAGGGTGTG
GACAAGAACACCGCGAACAAATATATTAACGAACTGGAGCAAAGCGGCGATATCTTTAGCCCGAAGATGGGTATCATTAAAGTGG
TTCGTCACGAGGAAGAGTAA
```

Amino acid sequence:

```
MGSSHHHHHHHHHHSGGSGGENLYFQ|GMAEQQTSSLKLLFDEFLESYYSDEIKDIIIKFPNKRSLPVNISDLEEFD
PDTATNLIADPEIIIDAANESLMGKLAGLNFDTYIPHVRFYNQSINTPMVLNVGSAYINKFVSIDALVVKRSDIRPK
IRDAVFVCTFCNAKVKANLEKEEIPKVCPECKKRTLKIVPEESSFFNSQKIAVQDPLERLSGSIPTWQLEAWLDDDL
VNMAIPGDRIEISGVLKIRPRKDSRGKVDPSIYSMYLNVTSLETKQKEFADIDISEDEERQIKELSKDPEIFNKVTQ
SVAPSIYGYNEIKQAVALQLFGGTPGKKLVDGGQIRSDMHILLIGDPGSAKTRILQSVSRLVPKGIYVSGKSVTGGG
LTAVAERDDFSEGGWTLKAGAMVLGNGGIVAIDEFDKISEEDTAALHEALESQTISVAKAGIIATFNAKASVLAAAN
PKFGRFDPHKYPAEQFDISPTLLSRFDLIFPIRDIMDTELDKSIANYILNQHEAAGAAIADVESSVAAEEPPIEHSL
LKKYIAYAKRYVMPRLSEEASNRIKEYYVDLRRAGSMKGATPITPRQIEGLIRMAEASAKSQLRDVVSVKDANLAIS
LSEYMLKTLAVDTEGRTDIDTILTGMPREKVDKINVILDAVKKLEEMDGSAKIDRIFEEVAKQGVDKNTANKYINEL
EQSGDIFSPKMGIIKVVRHEEE
```

>*SsoPfu*MCM

```
ATGGGTAGCAGCCACCACCACCACCATCATCATCACCACCACAGCGGTGGCAGCGGTGGCGAGAACCTGTACTTCCAGGGTAGCC
TGGAAATCCCGAGCAAGCAAATTGACTATCGTGATGTTTTCATCGAGTTTCTGACCACCTTTAAGGGCAACAACAACCAGAACAA
ATACATCGAGCGTATTAACGAACTGGTTGCGTATCGTAAGAAAAGCCTGATCATTGAGTTCAGCGACGTGCTGAGCTTTAACGAG
AACCTGGCGTACGAAATCATTAACAACACCAAGATCATTCTGCCGATCCTGGAAGGTGCGCTGTACGACCACATTCTGCAGCTGG
ATCCGACCTATCAACGTGACATCGAGAAAGTGCACGTTCGTATCGTGGGCATTCCGCGTGTTATCGAACTGCGTAAGATTCGTAG
CACCGACATCGGTAAACTGATCACCATTGATGGCATCCTGGTGAAGGTTACCCCGGTTAAAGAGCGTATCTACAAGGCGACCTAT
AAACACATTCACCCGGACTGCATGCAGGAGTTCGAATGGCCGGAAGATGAGGAAATGCCGGAAGTGCTGGAAATGCCGACCATCT
GCCCGAAGTGCGGTAAACCGGGCCAATTTCGTCTGATTCCGGAGAAGACCAAACTGATCGACTGGCAGAAGGCGGTTATTCAAGA
ACGTCCGGAGGAAGTGCCGAGCGGTCAGCTGCCGCGTCAACTGGAGATCATTCTGGAAGACGATCTGGTTGATAGCGCGCGTCCG
GGTGACCGTGTGAAAGTTACCGGCATCCTGGACATTAAGCAGGATAGCCCGGTGAAACGTGGCAGCCGTGCGGTTTTCGACATCT
ACATGAAAGTGAGCAGCATTGAAGTGAGCCAGAAAGTTCTGCAAGAGCTGGAAATCAGCCCGGAGGAAGAGCAAATCATTAAGGA
ACTGGCGAAGCGTAAAGACATCGTGGATGCGATTGTTGATAGCATCGCGCCGGCGATTTACGGTTATAAGGAAGTTAAGAAAGGC
ATTGCGCTGGCGCTGTTTGGTGGCGTGAGCCGTAAACTGCCGGATGGTACCCGTCTGCGTGGCGACATCCACGTGCTGCTGGTTG
GTGACCCGGGCGTTGCGAAGAGCCAGATTCTGCGTTATGTGGCGAACCTGGCGCCGCGTGCGATCTATACCAGCGGTAAAAGCAG
CAGCGCGGCGGGTCTGACCGCGGCGGCGGTGCGTGATGAGTTCACCGGTGGCTGGGTTCTGGAAGCGGGTGCGCTGGTGCTGGCG
GATGGTGGCTATGCGCTGATTGACGAGCTGGATAAGATGAGCGACCGTGATCGTAGCGTTATCCACGAGGCGCTGGAACAGCAAA
CCATCAGCATTAGCAAAGCGGGTATTACCGCGACCCTGAACGCGCGTACCACCGTTATTGCGGCGGCGAACCCGAAGCAGGGCCG
TTTCAACCGTATGAAAAACCCGTTTGAACAAATCGATCTGCCGCCGACCCTGCTGAGCCGTTTCGATCTGATCTTTGTTCTGATT
GACGAGCCGGACGATAAGATCGATAGCGAAGTTGCGCGTCACATTCTGCGTGTGCGTCGTGGTGAGAGCGAAGTGGTTGCGCCGA
AAATCCCGCACGAGATTCTGCGTAAGTACATCGCGTATGCGCGTAAAAACATCCACCCGGTGATTAGCGAAGAGGCGATGGAAGA
GATCGAAAAATACTATGTGCGTATGCGTAAGAGCGTTAAGAAAACCAAAGGTGAAGAGGAAGGCATTCCGCCGATCCCGATTACC
GCGCGTCAGCTGGAGGCGCTGATCCGTCTGAGCGAAGCGCACGCGCGTATGCGTCTGAGCCCGATTGTGACCCGTGAGGATGCGC
GTGAAGCGATTAAGCTGATGGAGTACACCCTGAAACAAATCGCGATGGATTAA
```

Amino acid sequence:

```
MGSSHHHHHHHHHHSGGSGGENLYFQ|GSLEIPSKQIDYRDVFIEFLTTFKGNNNQNKYIERINELVAYRKKSLIIE
FSDVLSFNENLAYEIINNTKIILPILEGALYDHILQLDPTYQRDIEKVHVRIVGIPRVIELRKIRSTDIGKLITIDG
ILVKVTPVKERIYKATYKHIHPDCMQEFEWPEDEEMPEVLEMPTICPKCGKPGQFRLIPEKTKLIDWQKAVIQERPE
EVPSGQLPRQLEIILEDDLVDSARPGDRVKVTGILDIKQDSPVKRGSRAVFDIYMKVSSIEVSQKVLQELEISPEEE
QIIKELAKRKDIVDAIVDSIAPAIYGYKEVKKGIALALFGGVSRKLPDGTRLRGDIHVLLVGDPGVAKSQILRYVAN
LAPRAIYTSGKSSSAAGLTAAAVRDEFTGGWVLEAGALVLADGGYALIDELDKMSDRDRSVIHEALEQQTISISKAG
ITATLNARTTVIAAANPKQGRFNRMKNPFEQIDLPPTLLSRFDLIFVLIDEPDDKIDSEVARHILRVRRGESEVVAP
KIPHEILRKYIAYARKNIHPVISEEAMEEIEKYYVRMRKSVKKTKGEEEGIPPIPITARQLEALIRLSEAHARMRLS
PIVTREDAREAIKLMEYTLKQIAMD
```

>*Afu*MCM

```
ATGGGTAGCAGCCACCACCACCACCATCATCATCACCACCACAGCGGTGGCAGCGGTGGCGAGAACCTGTACTTTCAGGGTATGG
GTATCAGCAGCCCGGCGCTGTGGACCGAATTCTTTGAGCGTTACTATCGTGAGGAAATTAACAAACTGGCGTACAAGCTGAAAAG
CGGTGGCGACGGTCGTAGCCTGTATGTGAACTTCGTTCGTGATCTGAGCATCTTTCAAGAAGGTAAACTGGGCGAGGAACTGATT
GAAAAACCGGATGAAGTGCTGGTTCATGCGGAGCGTGGCCTGGCGAACGCGACCAACATCTACGGTGTGAGCCTGGAAGGCTGCA
AACCGCGTTTCTATAGCCTGCCGACCGCGCGTAAGGTTCTGATCCGTAACCTGCGTGCGGAGCACATTGGTAAATTTATGGCGAT
CGAAGGCATTGTGCGTAAGGTTACCGAGGTGCGCCCGCGTATCGTGGAAGCGGCGTTCGCGTGCCTGAACTGCGGTAGCATTACC
ATGGTTCCGCAGGAAGACAGCCAGCTGCGTCAACCGTTCGAATGCAGCAAATGCAGCACCAAGAAAATGATCTTTCTGCCGGACA
GCAGCATTAGCGTTGATAGCCAGCGTGTGAAGATCCAAGAATATCCGGAGAACCTGCGTGGTGGCGAGCAGCCGCAAACCATCGA
CGTGATTCTGGAAGGTGATCTGGCGGGCAGCGTTAACCCGGGTGACCGTGTGATCATTAACGGCATCGTTCGTGCGAAACCGCGT
GGTCTGGGCCAGCGTAAGATGACCCACATGGATCTGTACATTGAGGGTAACAGCGTTGAGGTGCTGCAGCAAGAATATGAGGAAT
TTGAAATCACCGAGAAAGACCGTGAGCTGATTATGCAGCTGGCGGCGAGCGACGATATCTACGAAAAAATTGTTAAGAGCATCGC
GCCGAGCATTTATGGCCACGAGGATGTGAAGCTGGCGATCGCGCTGCAACTGTTCGGTGGCGTTCCGAAGAAACTGCCGGACGGT
ACCGAAATTCGTGGCGATATCCACATTCTGCTGGTTGGTGACCCGGGCGTGGCGAAAAGCCAGCTGCTGAAGTACGTGCACCGTA
TCGCGCCGCGTAGCGTTTATACCACCGGTAAAGGCACCACCACCGCGGGTCTGACCGCGACCGCGGTTCGTGACGAGGTGGATGG
TCGTTGGACCCTGGAAGCGGGTGCGCTGGTGCTGGCGGACAAGGGCATCGCGCTGGTTGACGAGATTGATAAAATGCGTAAGGAA
GATACCAGCGCGCTGCATGAAGCGCTGGAGCAGCAAACCATCAGCGTGGCGAAAGCGGGTATCAACGCGATTCTGAAAGCGCGTT
GCGCGCTGCTGGGTGCGGCGAACCCGAAATACGGCCGTTTCGAGAAGTTTACCCCGGTTCCGGAACAGATCGAGATGAGCCCGAC
CCTGCTGAGCCGTTTCGACCTGATTTTTGTGCTGAAGGACGAGCCGGATGAGGAAAAGGATAAACGTCTGGTTGAACACATCCTG
TACAGCCACCAGCTGGGTGAAATGACCGAGAAGGCGAAAAACGTGGCGGCGGAGTATGATGAGGAATTTATTCGTCAACGTAGCG
AGCGTATCGTTCCGGAAATTGACCCGGATCTGCTGCGTAAATACATCGCGTATGCGCGTAAGACCGTTTACCCGGTGCTGACCGA
CGAAGCGAAGGAGAAAATTAAGGAGTTCTATCTGAGCCTGCGTAGCCGTGTGAAAGAAAACAGCCCGGTTCCGATCACCGCGCGT
CAGCTGGAGAGCATTGTTCGTCTGGCGGAAGCGAGCGCGCGTGTGCGTCTGAGCGATCGTGTTGAACCGGAGGACGTTGATCGTG
TGATCGAGATTATGATGCGTAGCCTGCGTGAAATCGCGGTTGACCCGGAAACCGGCGAGATGGACATTGATCTGGCGTATAGCGG
CACCAGCAAGACCCAGCGTGATCGTATCATGATTCTGAAGAAAATCATTGAGCAACTGGAGGAAGAGCACGAACGTGGTGTGCCG
GAAGAGCTGATCCTGGAAGAGGCGGAAAAAGAGGGCATCGACCGTACCAAAGCGAAGGAGATTCTGAGCAAACTGAAGCTGCACG
GTGAAGTGTACACCCCGAAGCACGGCCACTATAAACTGGTTAGCAAGCTGTAA
```

Amino acid sequence:

```
MGSSHHHHHHHHHHSGGSGGENLYFQGMGISSPALWTEFFERYYREEINKLAYKLKSGGDGRSLYVNFVRDLSIFQE
GKLGEELIEKPDEVLVHAERGLANATNIYGVSLEGCKPRFYSLPTARKVLIRNLRAEHIGKFMAIEGIVRKVTEVRP
RIVEAAFACLNCGSITMVPQEDSQLRQPFECSKCSTKKMIFLPDSSISVDSQRVKIQEYPENLRGGEQPQTIDVILE
GDLAGSVNPGDRVIINGIVRAKPRGLGQRKMTHMDLYIEGNSVEVLQQEYEEFEITEKDRELIMQLAASDDIYEKIV
KSIAPSIYGHEDVKLAIALQLFGGVPKKLPDGTEIRGDIHILLVGDPGVAKSQLLKYVHRIAPRSVYTTGKGTTTAG
LTATAVRDEVDGRWTLEAGALVLADKGIALVDEIDKMRKEDTSALHEALEQQTISVAKAGINAILKARCALLGAANP
KYGRFEKFTPVPEQIEMSPTLLSRFDLIFVLKDEPDEEKDKRLVEHILYSHQLGEMTEKAKNVAAEYDEEFIRQRSE
RIVPEIDPDLLRKYIAYARKTVYPVLTDEAKEKIKEFYLSLRSRVKENSPVPITARQLESIVRLAEASARVRLSDRV
EPEDVDRVIEIMMRSLREIAVDPETGEMDIDLAYSGTSKTQRDRIMILKKIIEQLEEEHERGVPEELILEEAEKEGI
DRTKAKEILSKLKLHGEVYTPKHGHYKLVSKL
```

>*Ape*MCM

```
ATGGGTAGCAGCCACCACCACCACCATCATCATCACCACCACAGCGGTGGCAGCGGTGGCGAGAACCTGTACTTTCAGGGCATGG
CGGAGGAAATGCTGAGCGGCGAGGAAACCCTGGCGGTGGGTGAACGTTTCAAGACCTTTCTGGAAAACTTCCGTACCGAGGAAGG
TAAGCTGAAATACGTGGAAGCGATCCGTCGTATGATTAACTATGAGGAAACCAGCCTGGAAGTTGAGTTTAAGGACCTGTACCGT
TATGATCCGCTGCTGAGCGAAATCCTGCTGGAGAAGCCGCGTGAATTCCTGAAAGAAGCGAGCGAGGCGCTGAAAGAGATTGTGG
CGCAGGAAAGCCCGGAGTATGCGCAAGGTCGTGTTTTCACCCCGCGTTTTACCGGCCTGTTCGATACCGAACGTATCCGTGACAT
TGGCAGCGATCACGTGGGCAAGCTGGTTCAAATCAACGGTATTGTGACCCGTATGCACCCGCGTGCGACCCGTATGGTTCGTGCG
CGTTTTCGTCACGATCGTTGCGGTGCGGAGTTCTGGTGGCCGGCGAACGAAGACGAGGTGCTGGGTGAACGTATCGAGCGTCCGA
GCATTTGCCCGGTGTGCGGCGAAGGTGGCGGTAAATTTACCCTGGTTCGTGACAAAAGCCTGTACATCGATTGGCAGAAAATTAT
GGTGCAAGAACGTCCGGAGGATGTTCCGGGCGGTCAGATCCCGCGTAGCATTGAGGTTCACCTGAGCCGTGATCTGGTGGAAAAG
GTTCGTCCGGGTGACCGTGTGAAAATCGTTGGCGTGGTTGGTCTGCAGAGCTTCAGCAGCAGCAGCACCCTGTACAGCCTGTATA
TGGAGGCGAACAGCATCCTGCTGGAGGAAAAGATTCTGGAGGAAGTGAGCATCACCCGTGAGGATGAGGAAAAAATTCTGCAGCT
GAGCCGTGACCCGTGGATCAAGGAAAAAATCATTGCGAGCATCGCGCCGACCATTTATGGTCACTGGGATCTGAAAGAGGCGATC
GCGCTGCTGCTGTTTGGCGGTGTTCCGAAGCAACGTCCGGATGGTACCCGTACCCGTGGTGATATCCACGTGCTGTTCGTTGGTG
ACCCGGGTGTGGCGAAGAGCCAGCTGCTGCAAAGCACCGCGCAAGTTGCGCCGCGTGTGGTTTATACCACCGGTAAAGGTAGCAC
CGCGGCGGGTCTGACCGCGGCGGTGCTGCGTGATCCGCGTACCGGCGAGTATTTTCTGGAGGCGGGTGCGCTGGTGCTGGCGGAT
GGCGGTATCGCGGTTATTGACGAGTTCGATAAGATGAGCAAAGAAGACCGTGGTGTGATCCACGAAGCGATGGAGCAGCAAACCG
TTAGCATCGCGAAGGCGGGCATTAAAGCGACCCTGAGCGCGCGTGCGAGCCTGCTGGCGGCGGGCAACCCGAAGTTTGGTTACTA
TGACCCGAGCCGTAGCTTCGTGGACAACGTTGATCTGCCGGCGCCGATCATTAGCCGTTTTGACCTGATTTTCGTGGTTCGTGAT
GTTATCGAGCGTAGCCGTGACGAAATGCTGGCGAGCTACGTGCTGGAGACCCACACCAACGTTGAGCTGTTTAAGCCGGAAATTG
ACCCGGATCTGCTGCGTAAATATATCGCGTTCGCGCGTAAGCACGTGAAACCGCGTCTGACCCCGCAGGCGAAGAAACTGCTGAA
GGACTTTTACGTTGAGATGCGTAGCAGCGCGCTGCACCACAGCAGCCAGGAAGGTGCGAAACCGGTGCCGATTACCACCCGTCAA
CTGGAGGCGCTGATCCGTCTGACCGAAGCGCACGCGCGTATGAGCCTGAAACAGGAAGCGACCGAGGAAGATGCGATCGCGGCGA
TCCGTATTATGACCAGCGTTCTGCAAAGCATTGGCCTGGATCTGGAAACCGGCGAGATCGACATTGGTATCATTATGACCGGCGC
GAGCTTCCGTAGCCGTAAGATCATGAGCGAAGTGCTGGACCTGATCAAAAGCATTGTTGAGGAAGAGCGTGGCGGTCAGGGTTGC
```

GTGCGTGCGAGCGAGATTGTTCGTCGTCTGGGCGAAAAGAACATCCCGGAAGAGAAAGTGCGTGACGCGATTGATAAGCTGTACC
GTCAAGGTCTGATCATTGAGATCCGTACCGAATGCTATAAAGCGGTTTAA

Amino acid sequence:
MGSSHHHHHHHHHHSGGSGGENLYFQGMAEEMLSGEETLAVGERFKTFLENFRTEEGKLKYVEAIRRMINYEETSLE
VEFKDLYRYDPLLSEILLEKPREFLKEASEALKEIVAQESPEYAQGRVFTPRFTGLFDTERIRDIGSDHVGKLVQIN
GIVTRMHPRATRMVRARFRHDRCGAEFWWPANEDEVLGERIERPSICPVCGEGGGKFTLVRDKSLYIDWQKIMVQER
PEDVPGGQIPRSIEVHLSRDLVEKVRPGDRVKIVGVVGLQSFSSSSTLYSLYMEANSILLEEKILEEVSITREDEEK
ILQLSRDPWIKEKIIASIAPTIYGHWDLKEAIALLLFGGVPKQRPDGTRTRGDIHVLFVGDPGVAKSQLLQSTAQVA
PRVVYTTGKGSTAAGLTAAVLRDPRTGEYFLEAGALVLADGGIAVIDEFDKMSKEDRGVIHEAMEQQTVSIAKAGIK
ATLSARASLLAAGNPKFGYYDPSRSFVDNVDLPAPIISRFDLIFVVRDVIERSRDEMLASYVLETHTNVELFKPEID
PDLLRKYIAFARKHVKPRLTPQAKKLLKDFYVEMRSSALHHSSQEGAKPVPITTRQLEALIRLTEAHARMSLKQEAT
EEDAIAAIRIMTSVLQSIGLDLETGEIDIGIIMTGASFRSRKIMSEVLDLIKSIVEEERGGQGCVRASEIVRRLGEK
NIPEEKVRDAIDKLYRQGLIIEIRTECYKAV

>*Hvo*MCM
ATGGGTAGCAGCCACCACCACCACCATCATCATCACCACCACAGCGGTGGCAGCGGTGGCGAAAACCTGTACTTCCAAGGCATGG
CGCAAGCGCCGCAGAACCGTGACCTGACCGAGCGTTTCATCGAATTTTACCGTAACTACTATCGTGAGGAAATTGGTACCCTGGC
GCAGCAATATCCGAAGGAGAAACGTAGCCTGCACATCGATTACGACGATCTGTATCGTTTTGACAGCGAACTGGCGGACGATTAC
ATTACCAAACCGGGTCAATTCCAGGAGTATGCGGAGGAAGCGCTGCGTCTGTTTGATCTGCCGGCGGATGTGAAGCTGGGTCAGG
CGCACGTTCGTATGCGTAACCTGCCGGAAACCGTGGACATCCGTAACCTGCGTGTTAACGACGATCACATCGGTACCCTGATTAG
CGTGCAAGGCATCGTTCGTAAAGCGACCGACGTGCGTCCGAAGATTACCGAGGCGGCGTTCGAATGCCAGCGTTGCGGTACCATG
AGCTACATCCCGCAAGGCGATGGTGGCTTTCAGGAGCCGCACGAATGCCAAGGTTGCGAACGTCAGGGCCCGTTCCGTATTGACT
TTGATCAAAGCAACTTCGTGGACAGCCAAAAACTGCGTGTTCAGGAGAGCCCGGAAGGTCTGCGTGGTGGCGAGACCCCGCAGAG
CATCGACATTAACCTGAGCGACGATGTGACCGGTAAAGTTACCGCGGGCGATCACGTGACCGTGGTTGGTGTTCTGCACATCGAA
CAGCAAACCAGCGGCAACGAGAAGACCCCGGTGTTCGACTACTATATGGAAGGTATCAGCCTGACCATTGAGGATGAGGAATTTG
AGGACATGGAAATCAGCGACGAAGATGTGGCGGAGATTGTTGAACTGAGCAACGATCCGGCGATCTATGAGAAGATGGTGGAAAG
CGTTGCGCCGGCGATTTACGGTTATGAGCAAGAAAAAATCGCGATGATTCTGCAGCTGTTCAGCGGCGTTACCAAGCACCTGCCG
GATGGTAGCCGTATCCGTGGCGACCTGCACATGCTGCTGATTGGCGACCCGGGTACCGGCAAAAGCCAGATGCTGAGCTACATCC
GTCACATTGCGCCGCGTAGCGTGTATACCAGCGGCAAGGGTAGCAGCAGCGCGGGTCTGACCGCGGCGGCGGTTCGTGACGATTT
TGGTGATGGTCAGCAATGGACCCTGGAGGCGGGTGCGCTGGTGCTGGCGGATAAAGGCATCGCGGCGGTTGATGAACTGGACAAG
ATGCGTCCGGAGGACCGTAGCGCGATGCACGAGGGTCTGGAACAGCAACAGATCAGCGTGAGCAAAGCGGGCATTAACGCGACCC
TGAAGAGCCGTTGCAGCCTGCTGGGTGCGGCGAACCCGAAATACGGCCGTTTCGATCAATATGAGCCGATCGGCGAACAGATTGA
CCTGGAACCGGCGCTGATCAGCCGTTTCGATCTGATTTTTACCGTTACCGACGATCCGGACCCGGATGAAGACAGCAAACTGGCG
GACCACATCCTGAAGACCAACTACGCGGGCGAGCTGAACACCCAGCGTACCAACGTGGCGAACAGCGAGTTTACCGAACAACAGG
TGGATGCGGTTACCGACGAGGTTGCGCCGACCATCGATGCGGACCTGCTGCGTAAATACATTGCGTATGCGAAGCGTACCTGCTA
CCCGACCATGACCGATGAGGCGAAGGAAGTGATCCGTGATTTCTATGTTGACTTTCGTGCGCGTGGTGCGGATGAAGATGCGCCG
GTGCCGGTTACCGCGCGTAAACTGGAAGCGCTGGTTCGTCTGGGTGAGGCGAGCGCGCGTGTGCGTCTGAGCGATAAGGTTACCC
GTGAGGACGCGGAACGTGTGACCGGTATCGTTGAGAGCTGCCTGCGTGATATTGGTATGGACCCGGAGACCGGCGAATTCGATGC
GGACATCGTTGAAACCGGCCGTAGCAAAACCCAACGTGACCGTATCAAGAACCTGCTGGAGCTGATTCGTACCATGCAGGAAGAG
TACGAGGAAGGTGCGCCGCACGAGGAAGTGCTGGAGCGTGCGAACAGCGAACTGAACATGGATGAGAAAACCGTTAACGATCAGC
TGGACAAGCTGAAAATGAAGGGTGATATTTACGAGCCGCGTGGCGACGTTTATCGTGCGACCTAA

Amino acid sequence:
MGSSHHHHHHHHHHSGGSGGENLYFQGMAQAPQNRDLTERFIEFYRNYYREEIGTLAQQYPKEKRSLHIDYDDLYRF
DSELADDYITKPGQFQEYAEEALRLFDLPADVKLGQAHVRMRNLPETVDIRNLRVNDDHIGTLISVQGIVRKATDVR
PKITEAAFECQRCGTMSYIPQGDGGFQEPHECQGCERQGPFRIDFDQSNFVDSQKLRVQESPEGLRGGETPQSIDIN
LSDDVTGKVTAGDHVTVVGVLHIEQQTSGNEKTPVFDYYMEGISLTIEDEEFEDMEISDEDVAEIVELSNDPAIYEK
MVESVAPAIYGYEQEKIAMILQLFSGVTKHLPDGSRIRGDLHMLLIGDPGTGKSQMLSYIRHIAPRSVYTSGKGSSS
AGLTAAAVRDDFGDGQQWTLEAGALVLADKGIAAVDELDKMRPEDRSAMHEGLEQQQISVSKAGINATLKSRCSLLG
AANPKYGRFDQYEPIGEQIDLEPALISRFDLIFTVTDDPDPDEDSKLADHILKTNYAGELNTQRTNVANSEFTEQQV
DAVTDEVAPTIDADLLRKYIAYAKRTCYPTMTDEAKEVIRDFYVDFRARGADEDAPVPVTARKLEALVRLGEASARV
RLSDKVTREDAERVTGIVESCLRDIGMDPETGEFDADIVETGRSKTQRDRIKNLLELIRTMQEEYEEGAPHEEVLER
ANSELNMDEKTVNDQLDKLKMKGDIYEPRGDVYRAT

>*Kcr*MCM
ATGGGTAGCAGCCACCACCACCACCATCATCATCACCACCACAGCGGTGGCAGCGGTGGCGAAAACCTGTACTTCCAAGGCATGG
TTGAGAAAAGCCCGCTGATGAGCAGCGAGGAACTGGTGGAAAAGTACAAGAGCTTCATCCGTTACTATCGTGACGAAAACAACGA
GCCGATTTATCAGAAGGCGCTGGCGCAACTGATCGAGGAACAGCGTCGTAGCCTGAGCGTTAACTGGTACCACCTGTATAACTTC
AACCCGGACTTTCGTGAAATCGCGGAGGATATTGTTATGAACCCGAGCCTGCACATCAGCGCGGGTAGCAGCGCGATTAAAGAAC
TGGTGATGGAGCTGATGCCGATGACCGAGGAATTCCGTATCTACAGCGAGGGCGATTTCCACCTGCGTTTTTATAACGTTCCGAC
CAAGGCGAGCTTCCGTGACCTGACCAAATTTAGCATCGGTCGTCTGATCGAAATTGAGGGCATCATTACCCGTGTTAGCGACATT
TACGATAAGCTGGTGCGTGCGAGCTTCATCTGCACCAACTGCGGTCGTATCGAGGAAATTGATATCATTGGCGAAAAGCTGCGTG
TGCTGGAAAAATGCCCGGAGTGCGGTGCGCCGATGAAGCTGGACCACGAGATGAGCAAATTTATCCGTTGGCGTAGCGTTCGTAT
TCAGGAACGTCCGGAGGATCTGCCGCCGGGTATGATGCCGGAACACGTGGACGGCATCCTGACCGACGATATTGTTGACGATGTG
AAGCCGGGTGATCGTGTTCGTGTGACCGGCATCATTCGTATCAAACCGGCGCGTCGTGATGAAGGTCGTGAGGGTCTGATTTACA
AGCGTTATCTGGAAATCATTCACGTTGAGGTGCCGAACCGTGTTTACGAAAAGCTGGAGATTACCCCGGAGGATGAGGAAGAGAT
CCTGAAACTGAGCGAACGTGAGGACCTGGAAGAGCTGATCGTGAAGAGCATTGCGCCGAGCGTTTTCGGTTGGGCGGATGTGAAA
CGTGCGATTGCGTATGCGCTGTTTGGTGGCAGCACCAAGATCCTGGCGGATGGTAGCAAAGTTCGTGGCGAAATTAACGTTCTGC
TGGTTGGTGACCCGGGCGTGGCGAAGAGCCAACTGCTGAAATATACCGCGCAGCTGGCGCCGCGTGGTCTGTATACCACCGGTAA
AGGTAGCACCGCGGCGGGTCTGACCGCGGCGGTGGTTCGTGATAGCGCGACCGGTGGCTGGACCCTGGAAGCGGGTGCGCTGGTT
CTGGCGGACATGGGCGTGGCGTGCATCGACGAATTCGATAAAATGAGCGAGGACGATCGTCGTAGCATTCACGAAGCGATGGAGC
AGCAAACCATCAGCATTGCGAAGGCGGGTATCGTTGCGACCCTGAACGCGCGTACCACCATCATTGCGGCGGCGAACCCGAAGAA
AGGCAAATACGACGATTATGTTACCGTGGCGGAAAACATCAACCTGCCGCCGACCATTCTGAGCCGTTTCGACCTGGTGTTTATC
ATGAAGGATCGTCCGGGTGTTGAAAGCGACAGCATGGTGGCGGAGCACATCCTGATTACCCGTATGGGCCGTAACCCGGAGGCGA
AACCGCCGATTGACCCGAACCTGCTGAAGAAATACATCGCGTATGCGAAGCAAAACATCGACCCGATTCTGACCGATGAAGCGGC
GGAGCGTATCAAGAACTACTATGTTGATGTGCGTGGTCGTGGCATCAAAGAGAGCGAAGAGGGTATTGTTCAAGACCTGATCAGC

```
ATTACCCCGCGTCAGCTGGAAGCGCTGATCCGTCTGAGCGAGGCGCGTGCGCGTATGCACCTGCGTCGTGAAGTGACCGCGGAAG
ATGCGGAGATGGCGATCAACCTGATGGAGATTACCCTGAAGGGTGCGGCGTACGATATCGTTAGCGGCCACTTCGACATTACCGG
TTGGATGACCGGCATCAGCTTCCCGGAAGTGAAGCGTCGTGAGGTGGTTTTTCAGATCATTAAACAACTGGCGGAAGGTAGCGAG
GACGGCCTGGTTGACCGTGATGTGGTTGTGCGTATGGCGGCGGAACGTCTGAACCTGAAGGGTAAAGAATATGTGATCGAGGATA
TTCTGCGTAAACTGAACGAGGACGGCCTGATCATTTTTCCGCCGGGTGGCAAGATCCGTCTGATTTAA
```

Amino acid sequence:

```
MGSSHHHHHHHHHHSGGSGGENLYFQGMVEKSPLMSSEELVEKYKSFIRYYRDENNEPIYQKALAQLIEEQRRSLSV
NWYHLYNFNPDFREIAEDIVMNPSLHISAGSSAIKELVMELMPMTEEFRIYSEGDFHLRFYNVPTKASFRDLTKFSI
GRLIEIEGIITRVSDIYDKLVRASFICTNCGRIEEIDIIGEKLRVLEKCPECGAPMKLDHEMSKFIRWRSVRIQERP
EDLPPGMMPEHVDGILTDDIVDDVKPGDRVRVTGIIRIKPARRDEGREGLIYKRYLEIIHVEVPNRVYEKLEITPED
EEEILKLSEREDLEELIVKSIAPSVFGWADVKRAIAYALFGGSTKILADGSKVRGEINVLLVGDPGVAKSQLLKYTA
QLAPRGLYTTGKGSTAAGLTAAVVRDSATGGWTLEAGALVLADMGVACIDEFDKMSEDDRRSIHEAMEQQTISIAKA
GIVATLNARTTIIAAANPKKGKYDDYVTVAENINLPPTILSRFDLVFIMKDRPGVESDSMVAEHILITRMGRNPEAK
PPIDPNLLKKYIAYAKQNIDPILTDEAAERIKNYYVDVRGRGIKESEEGIVQDLISITPRQLEALIRLSEARARMHL
RREVTAEDAEMAINLMEITLKGAAYDIVSGHFDITGWMTGISFPEVKRREVVFQIIKQLAEGSEDGLVDRDVVVRMA
AERLNLKGKEYVIEDILRKLNEDGLIIFPPGGKIRLI
```

>*Mba*MCM

```
ATGGGCAGCAGCCACCACCACCACCATCATCATCACCACCACAGCGGTGGCAGCGGTGGCGAAAACCTGTACTTCCAGGGTATGA
CCGAAACCGAGAGCAAGTGGGATGAGAAGCTGAAACGTTTCTTTAAGGACTACTACTGGAACGAAATCCTGCAGCTGGCGAACGA
GTATCCGGACCAACGTAGCCTGAGCGTGGATTTCACCGACATTGAAAAGTTTGATCGTGAACTGAGCAAAGAGTTTCTGGACCAC
CCGGAGGAACTGATCAAGGCGGCGGAAGCGGCGCTGAAAGAGATTGATCTGCCGGTTGAGAAGAGCCTGGAGGAAGCGCACGTGC
GTGTTATCCGTATTCCGAACCGTATCCCGATTCGTGACCTGCGTAGCAAACACCTGAGCCGTTTCATCGCGATTGAGGGTATGAT
CCGTAAGGCGACCGAGGTGCGCCCGCGTATTACCAAAGCGGCGTTCGAATGCCTGCGTTGCGGCCACATCACCTTTGTTGATCAG
AACAGCTTCAAGTTTGAGGAACCGTTTGCGGGTTGCGAAGACGAGAACTGCGGTAAGAAAGGCCCGTTCAAAGTGCGTATCGAGG
ATAGCACCTTTATTGACGCGCAAAAGCTGCAGATCCAAGAGAGCCCGGAAAACCTGAAAGGTGGCAGCCAGCCGCAAAGCCTGGA
AGTTGATAGCGAGGACGATCTGACCGGTAACGTGACCCCGGGCGACCGTGTTATCATTAACGGTATCCTGAAAAGCCGTCAGCGT
ACCCTGAAGGATGGCAAAAGCACCTTCTACGACCTGGTGCTGGAAGCGAACAGCATTGAGCACCTGGACAAGGATTATGACGAAC
TGGAGATCACCGCGGAGGATGAGGAAGAGATTCTGGAACTGAGCCACGACCCGGAGATCTACAACAAAATCATTAGCAGCGTGGC
GCCGAGCATCTACGGTTATGAAGATATTAAAGAGGCGCTGGCGCTGCAGCTGTTCAGCGGCGTGGTTAAAAACCTGCCGGATGGT
AGCCGTATCCGTGGCGACATCCACATTATGCTGGTTGGTGACCCGGGCATTGCGAAGAGCCAACTGCTGCGTTATGTGGTTAAAC
TGAGCCCGCGTGGTGTGTTTACCAGCGGCCGTAGCGCGAGCGCGAGCGGTCTGACCGCGGCGGCGGTTAAGGACGATCTGAACGA
TGGCCGTTGGACCATTGAGGGTGGCGCGCTGGTGATGGCGGACATGGGCATTGCGGCGGTTGATGAAATGGACAAGATGAAAACC
GAGGATAAAAGCGCGCTGCACGAGGCGATGGAACAGCAAACCATCAGCATTGCGAAGGCGGGTATCATTGCGACCCTGAAAAGCC
GTTGCGCGCTGCTGGGTGCGGCGAACCCGAAGTATGGCCGTTTCGACCGTTATGAAAGCCTGGCGGAGCAGATCAGCATGCCGCC
AGCGCTGCTGAGCCGTTTCGATCTGATTTTTGTGCTGCTGGATACCCCGGACCACAACATGGACACCAAGATCGCGAACCACATT
CTGCAGAGCCACTACGCGGGCGAGCTGTTCGAACAACGTGAGCGTCTGCCGGGCAGCCACATCAAAGAAGACTTTGTGGAAGCGG
AGATGGAAGTGATCGAACCGGTTATTCAGCCGGAGCTGATGCGTAAGTACGTTGCGTATGCGCGTAAAAACGTGTTCCCGGTTAT
GGAAGAGGATGCGAAAGCGTACCTGATCAGCTTTTATACCGACCTGCGTAAGACCGGCGAGAGCAAAAACACCCCGGTGCCGGTT
ACCGCGCGTCAACTGGAAGCGCTGGTGCGTCTGAGCGAGGCGAGCGCGCGTGTGCGTCTGAGCAACACCGTTACCCTGGAAGATG
CGAAGCGTACCATCCGTATTGAGATGAACTGCCTGAAAAACGTGGGTGTTGACCCGGAAACCGGCGTGCTGGATGCGGACATCCT
GGCGAGCGGTACCAGCATGAGCCAACGTAACAAGATCAAGATCCTGCGTGATATCATTAAGAAAGTTAGCGAGAAGCACCCGGGT
GCGAAAGCGCCGCTGGAAGAGGTGTACGCGATCGCGGAGAACGAACACGGCATTGACCGTGTTCACGCGGAAGAGAACATCAAGA
AAATGAAGCAGCGTGGTGATCTGCTGAGCCCGGACCAAAACCACATTCGTCTGGTTTAA
```

Amino acid sequence:

```
MGSSHHHHHHHHHHSGGSGGENLYFQGMTETESKWDEKLKRFFKDYYWNEILQLANEYPDQRSLSVDFTDIEKFDRE
LSKEFLDHPEELIKAAEAALKEIDLPVEKSLEEAHVRVIRIPNRIPIRDLRSKHLSRFIAIEGMIRKATEVRPRITK
AAFECLRCGHITFVDQNSFKFEEPFAGCEDENCGKKGPFKVRIEDSTFIDAQKLQIQESPENLKGGSQPQSLEVDSE
DDLTGNVTPGDRVIINGILKSRQRTLKDGKSTFYDLVLEANSIEHLDKDYDELEITAEDEEEILELSHDPEIYNKII
SSVAPSIYGYEDIKEALALQLFSGVVKNLPDGSRIRGDIHIMLVGDPGIAKSQLLRYVVKLSPRGVFTSGRSASASG
LTAAAVKDDLNDGRWTIEGGALVMADMGIAAVDEMDKMKTEDKSALHEAMEQQTISIAKAGIIATLKSRCALLGAAN
PKYGRFDRYESLAEQISMPPALLSRFDLIFVLLDTPDHNMDTKIANHILQSHYAGELFEQRERLPGSHIKEDFVEAE
MEVIEPVIQPELMRKYVAYARKNVFPVMEEDAKAYLISFYTDLRKTGESKNTPVPVTARQLEALVRLSEASARVRLS
NTVTLEDAKRTIRIEMNCLKNVGVDPETGVLDADILASGTSMSQRNKIKILRDIIKKVSEKHPGAKAPLEEVYAIAE
NEHGIDRVHAEENIKKMKQRGDLLSPDQNHIRLV
```

>*Mha*MCM

```
ATGGGCAGCAGCCACCACCACCACCATCATCATCACCACCACAGCGGTGGCAGCGGTGGCGAGAACCTGTACTTCCAGGGTATGA
CCGAGGAAAAGTGGGAAAACAAATTCCGTGACTTTCTGAAGCGTTACTGCTGGCACGATATCCTGAAACTGGCGAACGAGTATCC
GGAACTGCGTAGCATTGAGGTTAACTTCACCGACCTGGAGCAATTTGATCGTGAACTGAGCGAGGAACTGCTGCAGACCCCGGAC
GAAGTTATCCCGAGCGCGGAGGAAGCGCTGAAGCAGATTGAGCTGCCGGTGGAAAAACAGCTGCACGACGCGCACATCCAATTCA
CCAGCATTCCGAACAAGGTGACCATCCGTGATCTGCGTAGCAACCACCTGCTGAAATTTATCGCGGTTGAGGGTATGATTCGTAA
GGCGACCGAAGTGCGTCCGAAAATTACCAACGCGGCGTTCTACTGCATGCGTTGCGAGACCGTTAACTATGTGCCGCAAAGCGGT
CCGAAGTTTGTTGAGCCGGGCGAATGCGAGGAAGAGAGCTGCGGCAAGCGTGGCCCGTTCAAACTGCTGATCGACAAGAGCAACT
TTATTGATGCGCAGAAGCTGCAGGTGCAAGAGAGCCCGGAAAGCCTGAAAGGTGGCAGCCAGCCGCAAAGCATCGACGTTGATGC
GGAAGACGAGCTGGCGGGTATTGTGAAGCCGGGCGATCGTGTGGTTGTGAACGGCATCCTGCGTAGCCACCAGCGTACCACCCGT
GAGGGTAAAAGCACCTTCTACGACCTGGTTCTGCACTGCAACAGCATCGAATATCTGGATCAAGAGTTTGACGAACTGGATATTA
GCCCGGAAGAGGAAGACGAGATCATTGAACTGAGCAACGATCCGCAGATCTACAACAAGATCATTAAAAGCATCGCGCCGAGCAT
TTACGGTTATGAGAACATTAAGGAAGCGCTGACCCTGCAACTGTTCAGCGGCGTGCCGAAAAGCCTGCCGGATGGTGGCCGTGTT
CGTGGTGATATCCACCTGCTGCTGGTTGGTGACCCGGGCATTGCGAAGAGCCAGCTGCTGCGTTATATGGTTAAACTGAGCCCGC
GTGGTGTGTTTGCGAGCGGCAAGAGCGCGAGCAGCAGCGGTCTGACCGCGGCGGCGGTTAAAGACGATCTGGGTGATGGCCGTTG
GACCCTGGAAGCGGGTGCGCTGGTGATGGCGGATATGGGTGTTGCGGCGGTGGACGAGATGGATAAGATGAGCCGTGAAGACAAA
AGCGCGCTGCACGAGGCGATGGAACAGCAAACCATCAGCGTTGCGAAGGCGGGTATTCTGGCGACCCTGAAAAGCCGTTGCGCGC
TGCTGGGTGCGGCGAACCCGAAGTATGGCCGTTTCGATCGTTATGAGGGTCTGGCGGAACAAATCAACATGCCGCCAGCGCTGAT
TAGCCGTTTCGACCTGATCTTTATTCTGCTGGACGTGCCGGATAGCAAGATGGATGCGAACATCGCGAACCACATTCTGAAAAGC
```

```
CACTACGCGGGCGAGCTGTATGAACAGTGGGACAAGCTGAGCACCAGCACCATCACCCAGGAAAAGGTTGCGAGCCACCAAAAAG
TGATCCTGCCGGAAATTGAGACCGAACTGCTGCGTAAATACGTTGCGTATGCGCGTCGTATGGTGTACCCGATCATGGAGGAAGA
GGCGCGTCAACACCTGGTTAACTTTTATCTGGACCTGCGTAAAATGGGCGAGAACAAGGATAGCCCGGTTCCGGTTACCGCGCGT
CAGCTGGAGGCGCTGGTTCGTCTGGCGGAAAGCAGCGCGCGTATCCGTCTGAGCAACACCGTGACCCTGGAAGACGCGAAGCGTA
CCACCAAAATTAGCCTGGCGTGCATGAAACAAGTTGGTGTGGACCCGGATACCGGTGCGCTGGACGTTGATGTGATCGCGAGCGG
CACCAGCAAAAGCCAGCGTGACAAGATCCACATTCTGCAAGATATCATTAAACACGTTAGCCAAAAGCATGCGGGTGGCAAAGCG
CCGCTGGATGAGGTGTACGAAGAGGCGAGCAGCGAAAACATCGACCGTGAGCACGCGGAAGATCTGATTCAGAAGATGAAACGTA
CCGGTGACCTGCTGGCGCCGGATAAGAAACACATCCGTCTGGTGTAA
```

Amino acid sequence:

```
MGSSHHHHHHHHHHSGGSGGENLYFQGMTEEKWENKFRDFLKRYCWHDILKLANEYPELRSIEVNFTDLEQFDRELS
EELLQTPDEVIPSAEEALKQIELPVEKQLHDAHIQFTSIPNKVTIRDLRSNHLLKFIAVEGMIRKATEVRPKITNAA
FYCMRCETVNYVPQSGPKFVEPGECEEESCGKRGPFKLLIDKSNFIDAQKLQVQESPESLKGGSQPQSIDVDAEDEL
AGIVKPGDRVVVNGILRSHQRTTREGKSTFYDLVLHCNSIEYLDQEFDELDISPEEEDEIIELSNDPQIYNKIIKSI
APSIYGYENIKEALTLQLFSGVPKSLPDGGRVRGDIHLLLVGDPGIAKSQLLRYMVKLSPRGVFASGKSASSSGLTA
AAVKDDLGDGRWTLEAGALVMADMGVAAVDEMDKMSREDKSALHEAMEQQTISVAKAGILATLKSRCALLGAANPKY
GRFDRYEGLAEQINMPPALISRFDLIFILLDVPDSKMDANIANHILKSHYAGELYEQWDKLSTSTITQEKVASHQKV
ILPEIETELLRKYVAYARRMVYPIMEEEARQHLVNFYLDLRKMGENKDSPVPVTARQLEALVRLAESSARIRLSNTV
TLEDAKRTTKISLACMKQVGVDPDTGALDVDVIASGTSKSQRDKIHILQDIIKHVSQKHAGGKAPLDEVYEEASSEN
IDREHAEDLIQKMKRTGDLLAPDKKHIRLV
```

>*Mka*MCM

```
ATGGGTAGCAGCCACCACCACCACCATCATCATCACCACCACAGCGGTGGCAGCGGTGGCGAGAACCTGTACTTCCAGGGCATGG
AGATGGAGCGTGAGTTTGAGGAAGCGCTGCGTAACAGCAAGACCTTCCTGCGTAGCATCGATCGTGTTATTACCGACTATCCGAA
ATGGCGTACCGTGGTTGTGGACCTGGAGGAATTCGACGAACCGGATATTGCGTTTGCTGTGAGCGACGATGTTGTGGAGGCGATG
AAGGTTGTGCAACGTGTGGCGATGGAACTGATCAAGAAAGAGCGTCCGGACGTTGATCGTGTTTGGGTGGAATTCCGTGGTAGCC
CGATTCGTCTGCGTGCGCGTGATATGAGCGTTGAATTTAAAGACCGTCTGGTTACCGTGGAGGGCATCGTTCGTCGTGTGGATAA
CGTTGCGGCGGAAGTGGTTCGTGTGGAAGCGGAGTGCCCGCAGTGCGGTAACCGTTTCGAAGTGCGTCGTCGTGAGTACCGTCCG
GACGTTCGTTGCCCGAACTGCGGCATGCGTTGCGAACCGGATGAGCTGTTTTACACCGACTATCAGCTGGTTGTGCTGCAAGAAG
CGCCGGAGCACGTTCGTGGTGGCGAACAACCGGCGACCGTTGAAGTGGAGTTTCGTTATGATCACATTAACCGTGTGCGTCCGGG
TGACCGTGTTCGTGTTACCGCGGTTCCGCGTGTGCGTCTGCCGAGCAGCAGCCCGCGTCCGGGTGATACCGGTGAAATTGTGCTG
GAAGCGCATGGTGTTGAGCGTAGCGACAGCCCGCTGCCGGAGGAAGATCTGCGTTTCACCCAGGACGAAGTTGAGCGTTTTGAGG
AACTGGCGGAAGGTGATCCGCTGGGCGAATTTGTGGAGGCGGTTGCGCCGCACATCCACGGTCACGAAGTGATTAAGAAAGCGGT
TAGCCTGCAGCTGTTTAGCTGCGTGGAGGAAGGTCAAATCCGTGAACGTGTTCACGTGCTGATTGTGGGTGACCCGGCGACCGCG
AAGAGCCAGATCCTGCAACACGTGATTGAGCACCTGGCGCCGCGTGGTGTTTACGTGAGCGCGCAGCACGTTACCGGTGCGGGTC
TGACCGCGGCGGCGGAACGTACCGAGGATGGTTGGACCCTGGAAGCGGGTGCGGTTGTGATGGCGGATGGTGGCGTGATCGCGAT
TGACGAGCTGGATAAAGCGAGCCGTGGTGATCTGAACGCGCTGCTGGAAGCGATGGAGAGCGGCAAGATCAGCGTTGCGAAAGCG
GGCATTACCACCACCCTGAACGCGCGTTGCGCGGTTCTGGCGGCGGCGAACCCGGAAGCGGGCCGTTGGCAGGGTGGCCACCCGA
TCGAGGAAATTAACCTGGATCCGGCGCTGCTGAGCCGTTTCGACGTTATCCTGTTTACCCGTGACGAACCGGATCCGGAGCAAGA
CAAACTGGTGGCGGAACGTATGATGGAGGCGTTCGATGGTGAATTTGACGAAATCGAGGGCAAGTATGAGCTGCTGCGTCGTTAC
GTGCTGTATGCGACCAAAGAATTCCCGAACGTTACCATTAGCGAGGATGCGCGTGAGGAACTGCGTGACTGGTTTGTTAGCGCGC
GTCAGGAAGCGGCGGACCGTATCGATGAGGGTGACCTGCGTACCGTTCCGGTGACCCGTCGTCAAATGGGCAGCGTTCTGCGTCT
GGCGCGTGCGAGCGCGCGTATGCGTCTGAGCGAAACCGTGGGTCGTGGCGATGTTAGCGTGGCGCTGAGCGTTGTGGAGGAATTC
ATGAAGGAAGTGATGCAGGAAGACGGTGTGCTGGACGCGGATGTTATCGAGACCGGCAAGCCGAAAAGCGTGCGTGAAGTTCGTG
AGTACGTTCTGAAGGTTGTGCGTAAACTGGCGAAGAAACACGAAGACGGTGTGCCGAAACGTGAGATCGTGAAGGCGGTTAAACA
CCGTGTGAGCCGTGAGCGTGTTGAGGAAATTCTGGACGATCTGGTGGAGGAAGGTAGCCTGCTGCAACCGCGTCCGGGCGTTTAT
CTGCCGATGTAA
```

Amino acid sequence:

```
MGSSHHHHHHHHHHSGGSGGENLYFQGMEMEREFEEALRNSKTFLRSIDRVITDYPKWRTVVVDLEEFDEPDIAFAV
SDDVVEAMKVVQRVAMELIKKERPDVDRVWVEFRGSPIRLRARDMSVEFKDRLVTVEGIVRRVDNVAAEVVRVEAEC
PQCGNRFEVRRREYRPDVRCPNCGMRCEPDELFYTDYQLVVLQEAPEHVRGGEQPATVEVEFRYDHINRVRPGDRVR
VTAVPRVRLPSSSPRPGDTGEIVLEAHGVERSDSPLPEEDLRFTQDEVERFEELAEGDPLGEFVEAVAPHIHGHEVI
KKAVSLQLFSCVEEGQIRERVHVLIVGDPATAKSQILQHVIEHLAPRGVYVSAQHVTGAGLTAAAERTEDGWTLEAG
AVVMADGGVIAIDELDKASRGDLNALLEAMESGKISVAKAGITTTLNARCAVLAAANPEAGRWQGGHPIEEINLDPA
LLSRFDVILFTRDEPDPEQDKLVAERMMEAFDGEFDEIEGKYELLRRYVLYATKEFPNVTISEDAREELRDWFVSAR
QEAADRIDEGDLRTVPVTRRQMGSVLRLARASARMRLSETVGRGDVSVALSVVEEFMKEVMQEDGVLDADVIETGKP
KSVREVREYVLKVVRKLAKKHEDGVPKREIVKAVKHRVSRERVEEILDDLVEEGSLLQPRPGVYLPM
```

>*Mth*MCM

```
ATGGGTAGCAGCCACCACCACCACCATCATCATCACCACCACAGCGGTGGCAGCGGTGGCGAGAACCTGTACTTCCAGGGCATGA
TGAAAACCGTGGACAAGAGCAAAACCCTGACCAAGTTCGAGGAATTCTTTAGCCTGCAAGACTATAAAGATCGTGTTTTTGAGGC
GATTGAAAAGTACCCGAACGTGCGTAGCATCGAGGTTGACTATCTGGATCTGGAAATGTTCGACCCGGATCTGGCGGATCTGCTG
ATTGAGAAACCGGACGATGTGATCCGTGCGGCGCAGCAAGCGATTCGTAACATCGACCGTCTGCGTAAGAACGTGGATCTGAACA
TCCGTTTTAGCGGTATTAGCAACGTTATCCCGCTGCGTGAACTGCGTAGCAAATTCATTGGCAAGTTTGTGGCGGTTGACGGCAT
CGTGCGTAAAACCGATGAGATTCGTCCGCGTATCGTGAAGGCGGTTTTCGAATGCCGTGGTTGCATGCGTCACCACGCGGTTACC
CAGAGCACCAACATGATCACCGAACCGAGCCTGTGCAGCGAATGCGGTGGCCGTAGCTTCCGTCTGCTGCAAGACGAGAGCGAAT
TTCTGGATACCCAGACCCTGAAACTGCAAGAGCCGCTGGAAAACTTGAGCGGTGGCGAGCAGCCGCGTCAAATTACCGTGGTTCT
GGAAGACGATCTGGTGGACACCCTGACCCCGGGTGATATCGTGCGTGTTACCGGCACCCTGCGTACCGTTCGTGACGAGCGTACC
AAACGTTTCAAGAACTTCATCTACGGTAACTACACCGAATTCCTGGAGCAGGAATTTGAGGAACTGCAAATTAGCGAGGAAGACG
AGGAAAAGATCAAAGAGCTGGCGGGCGATCCGAACATCTACGAAAAAATCATTCGTAGCACCGCGCCGAGCATTCACGGTTATCG
TGAGGTTAAGGAAGCGATCGCGCTGCAGCTGTTCGGTGGCACCGGCAAAGAGCTGGACGATAAGACCCGTCTGCGTGGTGACATC
CACATTCTGATCGTGGGCGATCCGGGTATTGGCAAAAGCCAAATGCTGAAATACGTTAGCAAGCTGGCGCCGCGTGGTATCTATA
CCAGCGGCAAGGGTACCAGCGGTGTGGGTCTGACCGCGGCGGCGGTTCGTGATGAGTTTGGTGGCTGGAGCCTGGAAGCGGGTGC
GCTGGTGCTGGGTGACAAGGGTAACGTGTGCGTTGACGAACTGGATAAGATGCGTGAGGAAGATCGTAGCGCGATTCACGAGGCG
CTGGAACAGCAAACCATTAGCATCGCGAAGGCGGGTATCATGGCGACCCTGAACAGCCGTTGCAGCGTTCTGGCGGCGGCGAACC
CGAAATTCGGCCGTTTTGACAGCTACAAGAGCATTGCGGAGCAGATCGATCTGCCGAGCACCATTCTGAGCCGTTTCGACCTGAT
```

```
CTTTGTGGTTGAGGACAAACCGGATGAGGAAAAGGACCGTGAACTGGCGCGTCACATCCTGAAAACCCACAAGGAAGACCACATG
CCGTTCGAGATTGATCCGGAACTGCTGCGTAAATACATCGCGTATGCGCGTAAGAACGTGCGTCCGGTTCTGACCGACGAGGCGA
TGCAGGTGCTGGAAGATTTTTACGTTAGCATGCGTGCGAGCGCGGCGGATGAGGATAGCCCGGTGCCGATCACCGCGCGTCAACT
GGAGGCGCTGGTTCGTCTGAGCGAAGCGAGCGCGAAGATTAAACTGAAGGAGCACGTGGAGGCGGAAGACGCGCGTAAAGCGATC
AAGCTGAGCCAGGCGTGCCTGAAACAAGTGGGTTATGATCCGGAAACCGGCAAAATTGACATCGATAAGGTTGAGGGTCGTACCC
CGAAAAGCGAACGTGACAAGTTCCGTCTGCTGCTGGAGCTGATCAAGGAGTACGAAGACGATTATGGTGGCCGTGCGCCGACCAA
CATTCTGATCACCGAAATGATGGACCGTTACAACGTGAGCGAGGAAAAAGTTGAGGAACTGATTCGTATCCTGAAAGATAAGGGT
GCGATTTTTGAACCGGCGCGTGGCTATCTGAAGATCGTTTAATAAAAGCTT
```

Amino acid sequence:

```
MGSSHHHHHHHHHHSGGSGGENLYFQGMMKTVDKSKTLTKFEEFFSLQDYKDRVFEAIEKYPNVRSIEVDYLDLEMF
DPDLADLLIEKPDDVIRAAQQAIRNIDRLRKNVDLNIRFSGISNVIPLRELRSKFIGKFVAVDGIVRKTDEIRPRIV
KAVFECRGCMRHHAVTQSTNMITEPSLCSECGGRSFRLLQDESEFLDTQTLKLQEPLENLSGGEQPRQITVVLEDDL
VDTLTPGDIVRVTGTLRTVRDERTKRFKNFIYGNYTEFLEQEFEELQISEEDEEKIKELAGDPNIYEKIIRSTAPSI
HGYREVKEAIALQLFGGTGKELDDKTRLRGDIHILIVGDPGIGKSQMLKYVSKLAPRGIYTSGKGTSGVGLTAAAVR
DEFGGWSLEAGALVLGDKGNVCVDELDKMREEDRSAIHEALEQQTISIAKAGIMATLNSRCSVLAAANPKFGRFDSY
KSIAEQIDLPSTILSRFDLIFVVEDKPDEEKDRELARHILKTHKEDHMPFEIDPELLRKYIAYARKNVRPVLTDEAM
QVLEDFYVSMRASAADEDSPVPITARQLEALVRLSEASAKIKLKEHVEAEDARKAIKLSQACLKQVGYDPETGKIDI
DKVEGRTPKSERDKFRLLLELIKEYEDDYGGRAPTNILITEMMDRYNVSEEKVEELIRILKDKGAIFEPARGYLKIV
```

>*Nac*MCM

```
ATGGGCAGCAGCCACCACCACCACCATCATCATCACCACCACAGCGGTGGCAGCGGTGGCGAGAACCTGTACTTCCAGGGTATGG
GCAGCGGTGGCTGCATTTATGTGAACCGTCAAAACGAAATCCTGATTTGGATCAAAGAGAACTACCCGGACACCGATTTCGCGAC
CAGCAAGATCTACGAGGAATATCAGGAAGAGGAAGAGTACGAGGGTAAAGAGAGCACCTTTTATAACCTGATTAACAAGCTGTGC
AGCAAAAACCTGCTGGAACGTCCGGAGCACGGCAAGTACCGTATCAACACCCGTGGCAAGCGTAAAGCGGAGTATCTGATTGGTG
GCGAGGAAATCGACGCGGAAAACGAGGATCTGGAGGAAATTCGTCTGCAGCTGCTGGACTTCCTGGATGAACGTAAAGAGGAGAT
TAAACAAAGCATCGCGGACAGCACCGCGTTCAAGCTGAAACTGAGCGAACTGGACAAATTTAACCCGGAGCTGATCGATTACTTC
GAAAACAACCCGGAGAAATTTCTGGACGCGGTGGATAAGGCGTTTAACACCGTGCTGGACGTTAGCAGCCGTATTGATTACACCA
TCGACTGCGATGTTGACTACTGGGAAATTCCGCTGTATAAGGCGCGTAGCAACGAGTATCGTAAGAAACTGATCACCGTGCAGGG
CACCGTTGAAAGCGCGAGCGATTTCCACCAAGAGCTGGTTAGCGCGGTTTTTGAATGCAGCCAGTGCGGCGAGCGTTACGAAAAA
GAGCAAGACAGCGCGAAGCTGAAAAGCCCGTATAAGTGCGAATGCGGCAGCAAGAAATTCGGTGAGGTGTCTCGTAACCTGATTA
ACCTGGTTAGCTTTCGTATCAGCAACGAAAAAGGCCACAACGAGTACATCAAGGCGGACTTCCGTAGCAGCAGCATTACCAGCGA
TATCCAGGAAGCGTTTAAACCGGGTCAAGAGCTGCGTGTGACCGGTGTTGCGCATGGTCAGCCGATTGGCAAGGATAGCAGCAAA
GTGGAACCGATCCTGAAGGTGGTTAGCTTCAAACCGCAGGACAGCCAAAAAGTGCTGGAGGATTATAAGAAAGAGGAAATCAGCC
TGCTGATGGGCAAGGTTGACACCCTGGAAAACCCGTTCCAGAGCTTTGCGACCAGCATTGCGCCGAGCATCGTTGAACAAGAGTT
TGCGAAGAAAGTGGTTGCGGCGAGCCTGATTGGTGGCAAGGCGACCGGTGGCCAGGGTGGCAGCGGTCGTATCCACAGCCTGCTG
CTGAGCAACCCGGGTAGCGGCAAAAGCGACATCCAAAACTTCGTGAAGGAAACCTTTAGCAACGTTGAGCTGGCGGATGGTAGCA
ACGCGACCGGTCCGGCGCTGACCGCGACCGTGGAGCAGGAAGAGGGTAACTACCGTCTGCGTGCGGGCAAGCTGGTGTATGCGGA
CGAAGGTGTTCTGTGCCTGGACGAGTTCGATAAGATGAACAAAAACGATAGCAGCCGTCTGAACACCGCGATGACCAGCAAGACC
TTTCCGATTGACAAAGCGAGCATCAACGCGGAGCTGCCGGGTAACGCGACCGTGGTTGCGACCGGTAACTTCGAAGACTACGTGG
ACGATATGGAGTTTGTTAAAGACAGCATTCCGGATCACGCGGAAAGCCTGCTGGACCGTTTCCACCTGATCTATGCGATGCGTAG
CCCGGATAACCAGGAAAAAGTTCAAGACGCGATTTTCGATAGCTTTAACAACAGCAGCCAGGAGAAGGCGAGCACCGATTTTGAC
GAGGAAGAGCTGGTGCTGTACCGTGAACTGGCGGCGAGCAAGGACCCGGAGCTGACCGATGAAAGCGTTGAGCTGCTGAAGAAAT
GGCTGCGTGGCCAAAAGGAGATCGCGGACAGCAAGGGTAACAGCAGCTTCAAGACCGATAGCAACCGTCACCTGATGGCGCTGGG
TCTGCTGACCACCATGTTTGCGAAAAGCCGTCTGAGCAGCAAGACCAACGAAGAGGACGCGGAACGTGCGGTGAAGCTGTTCATG
CGTTGCCGTAACAGCCTGGGCCTGTATGATGGTGACACCGATCAGCGTAGCCAAAAGGTTAAAGCGTAA
```

Amino acid sequence:

```
MGSSHHHHHHHHHHSGGSGGENLYFQGMGSGGCIYVNRQNEILIWIKENYPDTDFATSKIYEEYQEEEEYEGKESTF
YNLINKLCSKNLLERPEHGKYRINTRGKRKAEYLIGGEEIDAENEDLEEIRLQLLDFLDERKEEIKQSIADSTAFKL
KLSELDKFNPELIDYFENNPEKFLDAVDKAFNTVLDVSSRIDYTIDCDVDYWEIPLYKARSNEYRKKLITVQGTVES
ASDFHQELVSAVFECSQCGERYEKEQDSAKLKSPYKCECGSKKFGEVSRNLINLVSFRISNEKGHNEYIKADFRSSS
ITSDIQEAFKPGQELRVTGVAHGQPIGKDSSKVEPILKVVSFKPQDSQKVLEDYKKEEISLLMGKVDTLENPFQSFA
TSIAPSIVEQEFAKKVVAASLIGGKATGGQGGSGRIHSLLLSNPGSGKSDIQNFVKETFSNVELADGSNATGPALTA
TVEQEEGNYRLRAGKLVYADEGVLCLDEFDKMNKNDSSRLNTAMTSKTFPIDKASINAELPGNATVVATGNFEDYVD
DMEFVKDSIPDHAESLLDRFHLIYAMRSPDNQEKVQDAIFDSFNNSSQEKASTDFDEEELVLYRELAASKDPELTDE
SVELLKKWLRGQKEIADSKGNSSFKTDSNRHLMALGLLTTMFAKSRLSSKTNEEDAERAVKLFMRCRNSLGLYDGDT
DQRSQKVKA
```

>*Neq*MCM

```
ATGGGCAGCAGCCACCACCACCACCATCATCATCACCACCACAGCGGTGGCAGCGGTGGCGAGAACCTGTACTTCCAGGGTATGG
AGGAACTGGAAAAGCTGGAACTGGAGAAATACCTGCTGAAGAAAATGAACGAGGCGAACGACATTCTGTATATCAGCCTGAAGGA
ACTGGAGGAACTGGGCCTGATTAACCTGATCGACGAGATTCTGGATAACCCGGAGAAAGCGATCGAACAGATTAAGACCATCGTG
AAAGAAATTCAAGAGGAATATGCGCTGACCAAGGTTGACTACTATATCGCGTTCACCGATGTTCAGTACTATCGTAACGTGAAAA
TTCGTGAGCTGCGTAGCCACCACCTGAACAAGCTGGTTGCGATCGAAGGCATCATTAAACAAAGCAGCATGGTTAAGCCGGTGCT
GAAACGTGCGGTTTTTCGTCACAGCTGCGGTTACGAAGTGGAGAAGGAAATTAAAAGCATCAGCGACAAGATCAGCAAACCGAAG
AAATGCCCGAAGTGCAACAAAAGCGGTGACTGGGAGATTGTGGAGGAAGAGTATATCGATATTCAGCGTCTGGTTCTGGAAGAAC
TGCCGGAGAACCTGACCGGTGGCGCGCAACCGGAACGTGTTACCGCGATTCTGAAGGACAAACTGGTGGAACCGAAGATCAACGA
TAAAACCGTTCCGGGTGCGCGTGTTCGTATTGTGGGCATCCCGCGTACCGCGAAGCTGACCGAGAAAGGCGCGATCTACGATATC
CTGATTGAAGTTAACAACATTGAGTTCCTGGAAAAGAACATCACCGACATCGTTATTACCAACAAGGATCTGGTGGAGATTAAAG
AAATCGCGAACAGCAACAACCCGCTGGACCTGCTGGTTGAGAACTTCGCGCCGAGCATCTTCGGTTACGATTACATTAAGAAAGC
GATCCTGCTGCAGATGGTTGGTGGCGTGAAGAAAATTCGTCGTGACGGTACCAAAGTGCGTGGCCACATCCACATTCTGCTGGTT
GGTGATCCGGGCACCGCGAAGAGCACCCTGCTGAAGTATGCGGCGGAAGTGGCGCCGCGTGGTCGTTATGTGAGCGGTACCAGCG
CGACCGCGGTTGGTCTGGTGGCGGTGGTTGTGCGTGACGAACTGCTGAAAGTGTGGAGCATCGATGCGGGTCCGATGGTTCTGGC
GAACGGTGGCCTGCTGGCGCTGGACGAGATTGAAAAGCTGGGTAAAAACGAGCTGATGATCCTGCACGAGGCGATGGAACAAGGT
AGCGTGACCATTAGCAAGGCGGGCATCCACGTTACCCTGAAAACCGAGACCAGCGTGCTGGCGGCGGCGAACCCGAAATTTGGCC
```

```
GTTGGGACGATAACCTGAGCCTGGTTGAACAGATCGCGATTCCGCCGACCATCCTGAACCGTTTCGACCTGATTTTTCTGATCCG
TGATAAGCCGGGTAAAGACTACGATGAGCAACTGGCGGAACGTGTGCTGGAGAGCTATGTTGAAGACGTGGATCTGGCGATCCCG
GTGGACCTGCTGCGTAAGTACATTCTGTATGTTCGTAAGAACATCAAACCGCGTCTGAGCAACGAGGCGATTGCGCGTATCAAAG
ATTTCTTTGTTAGCCTGCGTGAGAAGAGCCAGGAACTGAAAGCGGTGCCGATTAGCACCCGTCAACTGGAGAGCATCGTTCGTCT
GGCGGAAGCGAGCGCGCGTATTCGTTTCAGCGATATCGTGGAGAAGGAAGACGCGGATCTGGCGATCGAGCTGACCAAACGTTTT
CTGGAAGAGGCGGGTGTTGACCCGGAAAGCAAGGTGATCGATATTACCATCCTGGAGAGCGGCAAGCCGCGTAGCAAAATTGAAA
AGCAGAAACTGCTGCTGCAGCTGATCAAGCAACTGGACAGCGGCGAGGGCGTGAGCGAGAAAGAACTGATCGAGAAGGCGAAAGA
ATACGGTCTGATGGATAGCGAGATTGAACAACTGCTGTACTATCTGAAAACCAGCGGTGCGGTGTTTGAAATTAAGCCGGGCATC
CTGAAAGCGGTTTAA
```

Amino acid sequence:

```
MGSSHHHHHHHHHHSGGSGGENLYFQGMEELEKLELEKYLLKKMNEANDILYISLKELEELGLINLIDEILDNPEKA
IEQIKTIVKEIQEEYALTKVDYYIAFTDVQYYRNVKIRELRSHHLNKLVAIEGIIKQSSMVKPVLKRAVFRHSCGYE
VEKEIKSISDKISKPKKCPKCNKSGDWEIVEEEYIDIQRLVLEELPENLTGGAQPERVTAILKDKLVEPKINDKTVP
GARVRIVGIPRTAKLTEKGAIYDILIEVNNIEFLEKNITDIVITNKDLVEIKEIANSNNPLDLLVENFAPSIFGYDY
IKKAILLQMVGGVKKIRRDGTKVRGHIHILLVGDPGTAKSTLLKYAAEVAPRGRYVSGTSATAVGLVAVVVRDELLK
VWSIDAGPMVLANGGLLALDEIEKLGKNELMILHEAMEQGSVTISKAGIHVTLKTETSVLAAANPKFGRWDDNLSLV
EQIAIPPTILNRFDLIFLIRDKPGKDYDEQLAERVLESYVEDVDLAIPVDLLRKYILYVRKNIKPRLSNEAIARIKD
FFVSLREKSQELKAVPISTRQLESIVRLAEASARIRFSDIVEKEDADLAIELTKRFLEEAGVDPESKVIDITILESG
KPRSKIEKQKLLLQLIKQLDSGEGVSEKELIEKAKEYGLMDSEIEQLLYYLKTSGAVFEIKPGILKAV
```

>*Nma*MCM

```
ATGGGCAGCAGCCACCACCACCACCATCATCATCACCACCACAGCGGTGGCAGCGGTGGCGAAAACCTGTACTTCCAGGGTATGA
GCAGCGCGCAAACCAGCACCTTTACCGACAGCGCGCTGAGCGATAAGGTGAAAGAATTCCTGACCCGTTTTAAAGATGCGAACGG
CGAGTACAAGTATGTTCAGGAAATCGACGAGATGATGCCGAAGAACAGCAAATACATCATTGTGGACTATAACGATCTGATTGTT
GAACCGGAGATCATTAGCATCTTCAGCGAAAACCCGGACCGTATTTTCGATGCGTTTAGCCGTGCGATCAAAGAGGCGCTGCAGA
CCCGTTTTCCGGACTACGCGGAAAAGATTAAAGATGAAGTGCGTGTTCGTCTGGTGAACTATCCGAGCGAACGTAGCCTGCGTCA
AATCAACGCGGAGACCATCGGTACCATTACCAGCGTGAGCGGCATGGTGGTTCGTGCGAGCGAAGTTAAGCCGCTGGCGAAAGAG
CTGATCTTCGTGTGCCCGGACGAACACCAGACCAAGGTTATCCAAATTAAGGGTATGGATGTGAAGGTTCCGGTGGTTTGCGACA
ACCCGAACTGCAAGCAGCGTGACTTTGATCTGAAACCGGAGGCGAGCAAGTTCATCGACTTTCAGATCATGCGTCTGCAAGAACT
GCCGGAGGATCTGCCGCCGGGTCAGCTGCCGCACTACATTGACGTGACCGTTCGTCAAGACCTGGTTGATAACGCGCGTCCGGGT
GATCGTATCGTGCTGACCGGCGTGGTTCGTGTTGAACAAGAGAGCGTGACCGGTGTTACCCGTGGTCACAGCGGCCTGTATCGTC
TGCGTATTGAGGGTAACAACATCGAGTTTCTGGGTGGCCGTGGCAGCAAGACCAGCCGTAAAATTGAACGTGAGGAAATCAGCCC
GGAGGAAGAGAAGATGATCAAAGCGCTGGCGGCGAGCCCGGATGTGTACCAGCGTCTGATTGATAGCTTTGCGCCGCACATCCAG
GGTCAAAGCCTGATCAAAGAAGCGATTCTGCTGCTGATCGTTGGCAGCAACCAACGTCCGCTGGGTGACGGCAGCAAGATTCGTG
GTGATATCAACGTGTTCCTGGTTGGTGACCCGGGCACCGCGAAAAGCGAGATGCTGAAGTTTTGCAGCCGTATCGCGCCGCGTGG
TCTGTATACCAGCGGTCGTGGCAGCACCGCGGCGGGTCTGACCGCGGCGGTGGTTCGTGATAAAACCGGTATTATGATGCTGGAA
GCGGGTGCGGTGGTTCTGGGTGACCAGGGCCTGGTTAGCATCGACGAATTCGATAAGATGAAACCGGAGGATCGTAGCGCGCTGC
ACGAAGTGATGGAGCAGCAAAGCGCGAGCATTGCGAAGGGTGGCATCGTTGCGACCCTGAACGCGCGTACCAGCATTCTGGCGGC
GGCGAACCCGATGTACGGCAAATATGACCCGTTCAAGAACATTACCGAAAACGTGAACCTGCCGATCCCGCTGCTGACCCGTTTC
GACCTGATTTTTGTGGTTCGTGATATCCCGACCAAAGAACGTGACGAGCAGATCGCGCGTCACATCATTGAGCTGCACACCCCGC
AAGGCACCGATAAGAAAAGCGTGGTTGACGTTGATCTGCTGACCAAATACCTGAGCTATGCGAAGCGTGGTACCCCGGATCTGAC
CAAAGAAGCGGAGCAGAAGATTCTGGACTACTATCTGGAAATGCGTAACGTGGAGAGCGAAGAGATGATCACCGTTACCCCGCGT
CAACTGGAAGGTATCATTCGTCTGAGCACCGCGCGTGCGCGTCTGCTGATGAAGGACAAAGTGGAAGAGGAAGATGCGGAGCGTG
CGATCTTCCTGATTCAGAGCATGCTGCAAGACGCGGGTGTGGATGTTAACACCGGCAAAGTGGACCTGGGTGTTCTGCAGGGCAA
ACCGCGTAGCGAAGTTAGCAAGATGCAACTGTTTATGGATATTCTGAAAGGTCTGGAGGGCGACAACAAGATCCCGGTGGAGGAA
AAGGCGTTCGTTAAAGAACTGGAGAAGAGCGAGAAATTTACCGAGGAAGAGGCGCGTAACTACATTCGTCGTATGCTGCGTGAAG
CGAGCATCTACGAGAGCAAGCCGGGTCACTATAACCGTGTTTAA
```

Amino acid sequence:

```
MGSSHHHHHHHHHHSGGSGGENLYFQGMSSAQTSTFTDSALSDKVKEFLTRFKDANGEYKYVQEIDEMMPKNSKYII
VDYNDLIVEPEIISIFSENPDRIFDAFSRAIKEALQTRFPDYAEKIKDEVRVRLVNYPSERSLRQINAETIGTITSV
SGMVVRASEVKPLAKELIFVCPDEHQTKVIQIKGMDVKVPVVCDNPNCKQRDFDLKPEASKFIDFQIMRLQELPEDL
PPGQLPHYIDVTVRQDLVDNARPGDRIVLTGVVRVEQESVTGVTRGHSGLYRLRIEGNNIEFLGGRGSKTSRKIERE
EISPEEEKMIKALAASPDVYQRLIDSFAPHIQGQSLIKEAILLLIVGSNQRPLGDGSKIRGDINVFLVGDPGTAKSE
MLKFCSRIAPRGLYTSGRGSTAAGLTAAVVRDKTGIMMLEAGAVVLGDQGLVSIDEFDKMKPEDRSALHEVMEQQSA
SIAKGGIVATLNARTSILAAANPMYGKYDPFKNITENVNLPIPLLTRFDLIFVVRDIPTKERDEQIARHIIELHTPQ
GTDKKSVVDVDLLTKYLSYAKRGTPDLTKEAEQKILDYYLEMRNVESEEMITVTPRQLEGIIRLSTARARLLMKDKV
EEEDAERAIFLIQSMLQDAGVDVNTGKVDLGVLQGKPRSEVSKMQLFMDILKGLEGDNKIPVEEKAFVKELEKSEKF
TEEEARNYIRRMLREASIYESKPGHYNRV
```

>*Pfu*MCM

```
ATGGGTAGCAGCCACCACCACCACCATCATCATCACCACCACAGCGGTGGCAGCGGTGGCGAAAACCTGTACTTTCAGGGCATGG
ATCGTGAGGAAATGATCGAGCGTTTCGCGAACTTTCTGCGTGAGTACACCGACGAAGATGGTAACCCGGTGTATCGTGGCAAAAT
CACCGACCTGCTGACCATTACCCCGAAGCGTAGCGTTGCGATCGATTGGATGCACCTGAACAGCTTCGACAGCGAACTGGCGCAC
GAGGTGATCGAAAACCCGGAGGAAGGTATTAGCGCGGCGGAGGATGCGATCCAGATTGTGCTGCGTGAGGACTTTCAACGTGAAG
ATGTTGGTAAAATTCACGCGCGTTTCTACAACCTGCCGGAGACCCTGATGGTGAAAGACATCGGCGCGGAACACATCAACAAGCT
GATTCAAGTTGAGGGTATCGTGACCCGTGTTGGCGAAATTAAACCGTTTGTGAGCGTTGCGGTGTTCGTTTGCAAGGATTGCGGT
CACGAGATGATCGTGCCGCAGAAACCGTATGAGAGCCTGGAAAAGGTTAAGAAATGCGAGCAATGCGGCAGCAAAAACATTGAAC
TGGACGTGAACAAGAGCAGCTTCGTTAACTTTCAGAGCTTCCGTATCCAAGATCGTCCGGAAACCCTGAAAGGTGGCGAGATGCC
GCGTTTTATCGACGGTATTCTGCTGGACGATATCGTGGACGTTGCGCTGCCGGGTGATCGTGTGATCGTTACCGGCATTCTGCGT
GTGGTTCTGGAGAAGCGTGAAAAAACCCCGATCTTCCGTAAAATTCTGGAAGTGAACCACATCGAACCGGTTAGCAAGGAGATCC
AGGAGCTGGAAATTAGCCCGGAGGAAGAGCAAATCATTAAGGAACTGGCGAAGCGTAAAGACATCGTGGATGCGATTGTTGATAG
CATCGCGCCGGCGATTTACGGTTATAAAGAGGTGAAGAAAGGCATTGCGCTGGCGCTGTTTGGTGGCGTTAGCCGTAAGCTGCCG
GATGGTACCCGTCTGCGTGGCGATATCCATGTGCTGCTGGTTGGTGACCCGGGCGTGGCGAAAAGCCAGATTCTGCGTTACGTTG
CGAACCTGGCGCCGCGTGCGATCTATACCAGCGGCAAGAGCAGCAGCGCGGCGGGTCTGGCGGCGGCGGTTCGTGATGAGTTCAC
CGGTGGCTGGGTGCTGGAAGCGGGTGCGCTGGTTCTGGCGGATGGTGGCTACGCGCTGATTGACGAACTGGATAAAATGAGCGAC
```

```
CGTGATCGTAGCGTGATCCACGAGGCGCTGGAACAGCAAACCATCAGCATTAGCAAGGCGGGTATTACCGCGACCCTGAACGCGC
GTACCACCGTTATCGCGGCGGCGAACCCGAAACAGGGCCGTTTTAACCGTATGAAGAACCCGTTCGAGCAAATCGACCTGCCGCC
GACCCTGCTGAGCCGTTTTGACCTGATCTTCGTGCTGATTGATGAGCCGGACGATAAAATCGACAGCGAAGTGGCGCGTCACATT
CTGCGTGTTCGTCGTGGTGAGAGCGAAGTGGTTGCGCCGAAGATCCCGCACGAAATTCTGCGTAAATACATCGCGTATGCGCGTA
AGAACATCCACCCGGTTATTAGCGAAGAGGCGATGGAAGAGATCGAGAAGTACTATGTGCGTATGCGTAAAAGCGTTAAGAAAAC
CAAGGGTGAAGAGGAAGGCATTCCGCCGATCCCGATTACCGCGCGTCAGCTGGAAGCGCTGATTCGTCTGAGCGAGGCGCATGCG
CGTATGCGTCTGAGCCCGATTGTGACCCGTGAGGATGCGCGTGAAGCGATTAAACTGATGGAGTACACCCTGAAGCAGATCGCGA
TGGACGAAACCGGTCAAATCGATGTTACCATTCTGGAGCTGGGCCAGAGCGCGCGTAAGCTGAGCAAAATCGAAAAGATTCTGGA
TATCATTGAGAAACTGCAAAAGACCAGCGAACGTGGTGCGCACGTGAACGACATTCTGGAGGAAGCGAAGAAAGCGGGCATCGAG
AAACAGGAAGCGCGTGAGATTCTGGAAAAACTGCTGGAGAAGGGTCAAATCTATATGCCGGAAAGCGGCTACTATAAGACCGTTT
AA
```

Amino acid sequence:

```
MGSSHHHHHHHHHHSGGSGGENLYFQGMDREEMIERFANFLREYTDEDGNPVYRGKITDLLTITPKRSVAIDWMHLN
SFDSELAHEVIENPEEGISAAEDAIQIVLREDFQREDVGKIHARFYNLPETLMVKDIGAEHINKLIQVEGIVTRVGE
IKPFVSVAVFVCKDCGHEMIVPQKPYESLEKVKKCEQCGSKNIELDVNKSSFVNFQSFRIQDRPETLKGGEMPRFID
GILLDDIVDVALPGDRVIVTGILRVVLEKREKTPIFRKILEVNHIEPVSKEIQELEISPEEEQIIKELAKRKDIVDA
IVDSIAPAIYGYKEVKKGIALALFGGVSRKLPDGTRLRGDIHVLLVGDPGVAKSQILRYVANLAPRAIYTSGKSSSA
AGLAAAVRDEFTGGWVLEAGALVLADGGYALIDELDKMSDRDRSVIHEALEQQTISISKAGITATLNARTTVIAAAN
PKQGRFNRMKNPFEQIDLPPTLLSRFDLIFVLIDEPDDKIDSEVARHILRVRRGESEVVAPKIPHEILRKYIAYARK
NIHPVISEEAMEEIEKYYVRMRKSVKKTKGEEEGIPPIPITARQLEALIRLSEAHARMRLSPIVTREDAREAIKLME
YTLKQIAMDETGQIDVTILELGQSARKLSKIEKILDIIEKLQKTSERGAHVNDILEEAKKAGIEKQEAREILEKLLE
KGQIYMPESGYYKTV
```

>*PfuSso*MCM

```
ATGGGTAGCAGCCACCACCACCACCATCATCATCACCACCACAGCGGTGGCAGCGGTGGCGAGAACCTGTACTTTCAGGGCATGG
ACCGTGAGGAAATGATCGAACGTTTCGCGAACTTTCTGCGTGAGTACACCGACGAAGATGGTAACCCGGTTTATCGTGGCAAGAT
TACCGATCTGCTGACCATCACCCCGAAACGTAGCGTGGCGATCGACTGGATGCACCTGAACAGCTTCGATAGCGAGCTGGCGCAC
GAGGTTATTGAAAACCCGGAGGAAGGTATCAGCGCGGCGGAAGACGCGATCCAGATTGTTCTGCGTGAGGACTTTCAACGTGAAG
ATGTGGGCAAGATTCACGCGCGTTTCTACAACCTGCCGGAGACCCTGATGGTTAAGGATATCGGCGCGGAACACATTAACAAACT
GATCCAAGTGGAGGGTATTGTGACCCGTGTTGGCGAAATCAAGCCGTTTGTGAGCGTTGCGGTGTTCGTTTGCAAAGACTGCGGT
CACGAGATGATCGTTCCGCAGAAGCCGTATGAGAGCCTGGAAAAAGTGAAGAAATGCGAGCAATGCGGCAGCAAGAACATTGAAC
TGGATGTGAACAAAAGCAGCTTCGTTAACTTTCAGAGCTTCCGTATCCAAGACCGTCCGGAGACCCTGAAAGGTGGCGAAATGCC
GCGTTTTATCGATGGCATTCTGCTGGACGATATCGTGGACGTTGCGCTGCCGGGTGATCGTGTGATTGTTACCGGCATCCTGCGT
GTGGTTCTGGAGAAGCGTGAAAAAACCCCGATTTTCCGTAAGATCCTGGAAGTGAACCACATTGAACCGGTTAGCAAAGAGATCG
ACGAAGTTATCATTAGCGAGGAAGATGAGAAGAAAATTAAGGACCTGGCGAAAGATCCGTGGATCCGTGACCGTATCATTAGCAG
CATTGCGCCGAGCATCTACGGTCACTGGGAGCTGAAGGAAGCGCTGGCGCTGGCGCTGTTTGGTGGCGTGCCGAAAGTTCTGGAA
GACACCCGTATTCGTGGCGATATCCACATTCTGATCATTGGTGATCCGGGCACCGCGAAGAGCCAGATGCTGCAATTCATCAGCC
GTGTGGCGCCGCGTGCGGTTTATACCACCGGTAAAGGTAGCACCGCGGCGGGTCTGACCGCGGCGGTGGTTCGTGAGAAAGGTAC
CGGCGAGTACTATCTGGAAGCGGGTGCGCTGGTGCTGGCGGATGGTGGCATCGCGGTTATTGACGAGATCGATAAAATGCGTGAC
GAAGATCGTGTGGCGATCCACGAGGCGATGGAACAGCAAACCGTGAGCATTGCGAAGGCGGGTATCGTTGCGAAACTGGCGCGTG
CGGCGGTGATTGCGGCGGGTAACCCGAAGTTTGGCCGTTACATCAGCGAGCGTCCGGTTAGCGACAACATTAACCTGCCGCCGAC
CATCCTGAGCCGTTTTGATCTGATCTTCATTCTGAAAGACCAGCCGGGCGAGCAAGATCGTGAACTGGCGAACTACATCCTGGAC
GTGCACAGCGGCAAGAGCACCAAAAACATCATTGACATTGATACCCTGCGTAAGTACATCGCGTATGCGCGTAAGTATGTTACCC
CGAAAATTACCAGCGAAGCGAAAAACCTGATCACCGACTTCTTTGTGGAGATGCGTAAGAAAAGCAGCGAAACCCCGGATAGCCC
GATCCTGATTACCCCGCGTCAGCTGGAGGCGCTGATCCGTATTAGCGAAGCGTATGCGAAGATGGCGCTGAAAGCGGAAGTGACC
CGTGAGGACGCGGAACGTGCGATCAACATTATGCGTCTGTTCCTGGAAAGCGTGGGTGTTGATTAA
```

Amino acid sequence:

```
MGSSHHHHHHHHHHSGGSGGENLYFQGMDREEMIERFANFLREYTDEDGNPVYRGKITDLLTITPKRSVAIDWMHLN
SFDSELAHEVIENPEEGISAAEDAIQIVLREDFQREDVGKIHARFYNLPETLMVKDIGAEHINKLIQVEGIVTRVGE
IKPFVSVAVFVCKDCGHEMIVPQKPYESLEKVKKCEQCGSKNIELDVNKSSFVNFQSFRIQDRPETLKGGEMPRFID
GILLDDIVDVALPGDRVIVTGILRVVLEKREKTPIFRKILEVNHIEPVSKEIDEVIISEEDEKKIKDLAKDPWIRDR
IISSIAPSIYGHWELKEALALALFGGVPKVLEDTRIRGDIHILIIGDPGTAKSQMLQFISRVAPRAVYTTGKGSTAA
GLTAAVVREKGTGEYYLEAGALVLADGGIAVIDEIDKMRDEDRVAIHEAMEQQTVSIAKAGIVAKLARAAVIAAGNP
KFGRYISERPVSDNINLPPTILSRFDLIFILKDQPGEQDRELANYILDVHSGKSTKNIIDIDTLRKYIAYARKYVTP
KITSEAKNLITDFFVEMRKKSSETPDSPILITPRQLEALIRISEAYAKMALKAEVTREDAERAINIMRLFLESVGVD
```

> *Sso*MCM

```
ATGGGTAGCAGCCACCACCACCACCATCATCATCACCACCACAGCGGTGGCAGCGGTGGCGAGAACCTGTACTTCCAGGGTATGG
AAATTCCGAGCAAGCAAATCGACTATCGTGATGTTTTCATCGAATTTCTGACCACCTTTAAGGGCAACAACAACCAGAACAAATA
CATCGAGCGTATTAACGAACTGGTGGCGTATCGTAAGAAAAGCCTGATCATTGAATTCAGCGATGTTCTGAGCTTTAACGAGAAC
CTGGCGTACGAAATCATTAACAACACCAAGATCATTCTGCCGATTCTGGAGGGTGCGCTGTACGACCACATCCTGCAGCTGGATC
CGACCTATCAACGTGACATCGAAAAAGTGCACGTTCGTATCGTTGGCATTCCGCGTGTGATTGAGCTGCGTAAGATCCGTAGCAC
CGATATTGGTAAACTGATCACCATTGACGGCATCCTGGTGAAGGTTACCCCGGTGAAAGAACGTATTTACAAGGCGACCTATAAA
CACATCCACCCGGATTGCATGCAGGAGTTCGAATGGCCGGAGGACGAGGAAATGCCGGAAGTGCTGGAAATGCCGACCATCTGCC
CGAAGTGCGGTAAACCGGGCCAATTTCGTCTGATCCCGGAAAAGACCAAACTGATTGACTGGCAGAAGGCGGTGATCCAAGAGCG
TCCGGAGGAAGTTCCGAGCGGTCAGCTGCCGCGTCAACTGGAGATCATTCTGGAAGACGATCTGGTGGATAGCGCGCGTCCGGGT
GACCGTGTGAAAGTTACCGGCATCCTGGACATTAAGCAGGATAGCCCGGTTAAACGTGGCAGCCGTGCGGTGTTCGATATTTACA
TGAAAGTTAGCAGCATCGAAGTGAGCCAAAAAGTTCTGGATGAAGTGATCATTAGCGAGGAAGACGAGAAGAAAATTAAGGACCT
GGCGAAAGATCCGTGGATCCGTGACCGTATCATTAGCAGCATTGCGCCGAGCATCTATGGTCACTGGGAGCTGAAGGAAGCGCTG
GCGCTGGCGCTGTTTGGTGGCGTGCCGAAGGTTCTGGAAGATACCCGTATTCGTGGCGACATCCACATTCTGATCATTGGTGATC
CGGGCACCGCGAAGAGCCAGATGCTGCAATTTATCAGCCGTGTTGCGCCGCGTGCGGTGTACACCACCGGTAAAGGTAGCACCGC
GGCGGGTCTGACCGCGGCGGTGGTTCGTGAGAAAGGTACCGGCGAGTACTATCTGGAAGCGGGTGCGCTGGTGCTGGCGGATGGT
GGCATCGCGGTTATTGACGAGATCGATAAAATGCGTGACGAAGATCGTGTGGCGATCCACGAGGCGATGGAACAGCAAACCGTGA
GCATCGCGAAGGCGGGTATTGTTGCGAAACTGAACGCGCGTGCGGCGGTGATTGCGGCGGGTAACCCGAAGTTCGGCCGTTATAT
CAGCGAGCGTCCGGTTAGCGATAACATCAACCTGCCGCCGACCATTCTGAGCCGTTTCGACCTGATCTTTATTCTGAAAGATCAG
```

CCGGGCGAGCAAGACCGTGAACTGGCGAACTACATCCTGGACGTTCACAGCGGCAAGAGCACCAAAAACATCATTGACATTGATA
CCCTGCGTAAGTACATCGCGTATGCGCGTAAGTATGTGACCCCGAAAATTACCAGCGAAGCGAAAAACCTGATCACCGATTTCTT
TGTTGAGATGCGTAAGAAAAGCAGCGAAACCCCGGACAGCCCGATCCTGATTACCCCGCGTCAGCTGGAGGCGCTGATCCGTATT
AGCGAAGCGTACGCGAAGATGGCGCTGAAAGCGGAAGTGACCCGTGAGGATGCGGAACGTGCGATCAACATTATGCGTCTGTTTC
TGGAGAGCGTGGGTGTTGACATGGAAAGCGGCAAGATCGACATTGATACCATCATGACCGGCAAGCCGAAAAGCGCGCGTGAGAA
GATGATGAAGATCATCGAAATCATCGATAGCCTGGCTGTGAGCAGCGAATGCGCGAAGGTTAAAGACATCCTGAAAGAGGCGCAG
CAAGTGGGTATCGAGAAGAGCAACATTGAAAAACTGCTGACCGACATGCGTAAGAGCGGCATCATTTACGAGGCGAAACCGGAAT
GCTATAAGAAAGTTTAA

Amino acid sequence:

MGSSHHHHHHHHHHSGGSGGENLYFQGMEIPSKQIDYRDVFIEFLTTFKGNNNQNKYIERINELVAYRKKSLIIEFS
DVLSFNENLAYEIINNTKIILPILEGALYDHILQLDPTYQRDIEKVHVRIVGIPRVIELRKIRSTDIGKLITIDGIL
VKVTPVKERIYKATYKHIHPDCMQEFEWPEDEEMPEVLEMPTICPKCGKPGQFRLIPEKTKLIDWQKAVIQERPEEV
PSGQLPRQLEIILEDDLVDSARPGDRVKVTGILDIKQDSPVKRGSRAVFDIYMKVSSIEVSQKVLDEVIISEEDEKK
IKDLAKDPWIRDRIISSIAPSIYGHWELKEALALALFGGVPKVLEDTRIRGDIHILIIGDPGTAKSQMLQFISRVAP
RAVYTTGKGSTAAGLTAAVVREKGTGEYYLEAGALVLADGGIAVIDEIDKMRDEDRVAIHEAMEQQTVSIAKAGIVA
KLNARAAVIAAGNPKFGRYISERPVSDNINLPPTILSRFDLIFILKDQPGEQDRELANYILDVHSGKSTKNIIDIDT
LRKYIAYARKYVTPKITSEAKNLITDFFVEMRKKSSETPDSPILITPRQLEALIRISEAYAKMALKAEVTREDAERA
INIMRLFLESVGVDMESGKIDIDTIMTGKPKSAREKMMKIIEIIDSLAVSSECAKVKDILKEAQQVGIEKSNIEKLL
TDMRKSGIIYEAKPECYKKV

7.2 Primers

Table 7.1: Primers used for mutagenesis.
Mutation primers noted with asterix (*) were designed for multi-site mutagenesis (Agilent), and only require 1 sequencing primer per site.

Construct	Mutation	Primer direction	Sequence (5'–3')
*Mac*MCM.^{FL}	ΔWHD	Forward	TGAAGACCCTGGCGGTTGACTAGTAGGGTCGTACCGATATTGAC
		Reverse	GTCAATATCGGTACGACCCTACTAGTCAACCGCCAGGGTCTTCA
*Mac*MCM.^{FL}	K119A	Forward	TCGATGCGCTGGTTGTGGCGCGTAGCGATATTCGTC
		Reverse	GACGAATATCGCTACGCGCCACAACCAGCGCATCGA
*Mac*MCM.^{FL}	R124A	Forward	CTGGTTGTGAAGCGTAGCGATATTGCTCCGAAAATCCGT
		Reverse	ACGGATTTTCGGAGCAATATCGCTACGCTTCACAACCAG
*Mac*MCM.^{FL}	K176A	Forward	CAGCTTCTTTAACAGCCAGGCAATCGCGGTTCAAGACCCG
		Reverse	CGGGTCTTGAACCGCGATTGCCTGGCTGTTAAAGAAGCTG
*Mac*MCM.^{FL}	R224A	Forward	GTTCTGAAGATTCGTCCGGCTAAAGATAGCCGTGGCAA
		Reverse	TTGCCACGGCTATCTTTAGCCGGACGAATCTTCAGAAC
*Mac*MCM.^{FL}	EEE-GSC	Forward	AGATGGGTATCATTAAAGTGGTTCGTCACGGGTCATGCTAATAAAAGCTTGGATCCGGCTGCTAACAA
		Reverse	TTGTTAGCAGCCGGATCCAAGCTTTTATTAGCATGACCCGTGACGAACCACTTTAATGATACCCATCT
*Mac*MCM.$^{\Delta WHD}$	E418Q	Forward	GCGGTATCGTGGCGATTGATCAGTTTGACAAAATCAG
		Reverse	CTGATTTTGTCAAACTGATCAATCGCCACGATACCGC
*Mac*MCM$^{.\Delta WHD}$	L100C	Forward*	TATACGCGCTACCAACGTTGCACACCATCGGGGTGTTGATG
*Mac*MCM$^{.\Delta WHD}$	N205C	Forward*	CCCGGGATCGCCATGCACACCAGGTCATCGTC
*SsoPfu*MCM	K130A	Forward	CATTGATGGCATCCTGGTGGCGGTTACCCCGGTTAAAGAG
		Reverse	CTCTTTAACCGGGGTAACCGCCACCAGGATGCCATCAATG

7.3 Optimal growth temperatures of selected archaea

Note: *Nanohaloarchaeum SG9* has never successfully been grown in culture. Therefore, the ideal growth temperature range is estimated by proxy from the habitat the organism was identified (*Nanohaloarchaeum SG9* was identified in the Atacama Desert).

Table 7.2: The optimal growth temperatures of selected archaeal species.

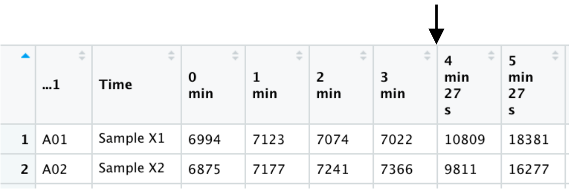
	Temperature (ºC)				
MCM	**Min**	**Optimum**	**Max**	**Phyla**	**Reference**
***Afu*MCM**	60	83	85	Euryarchaeota	399
***Ape*MCM**	70	90	97	TACK	400
***Hvo*MCM**	30	42	55	Euryarchaeota	401
***Kcr*MCM**	55	85	90	TACK	402
***Mac*MCM**	10	37	45	DPANN	403
***Mba*MCM**	30	35	45	Euryarchaeota	247
***Mha*MCM**	30	40	55	Euryarchaeota	404
***Mka*MCM**	84	98	110	Euryarchaeota	241
***Mth*MCM**	40	65	75	Euryarchaeota	405
***Nac*MCM**	16	33	50	DPANN	-
***Neq*MCM**	75	80	95	DPANN	406
***Nma*MCM**	15	28	32	TACK	407
***Pfu*MCM**	70	95	103	Euryarchaeota	408
***Sso*MCM**	55	75	90	TACK	409

7.4 Helicase script

This script outlines a general analysis of the fluorescent helicase data, where raw data is converted into standardized values. The experiment performed here was a single repeat of a protein concentration versus helicase activity. To adapt this script to other experiments, only the lines highlighted in yellow need to be adjusted. These lines ensure that the correct control wells are used for adjusting the signal. This code may easily be adjusted into a single function. In this experiment, the buffer conditions were the same, hence only single controls were required.

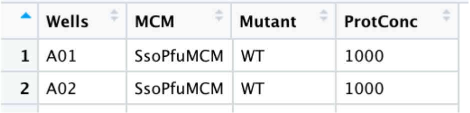
Data required:

1) .xlsx file from Clariostar plate reader (BMG Labtech). Will be loaded into R, in the format below. ATP is added at the arrow. Data can be reproducibly adjusted to this point.



	...1	Time	0 min	1 min	2 min	3 min	4 min 27 s	5 min 27 s
1	A01	Sample X1	6994	7123	7074	7022	10809	18381
2	A02	Sample X2	6875	7177	7241	7366	9811	16277

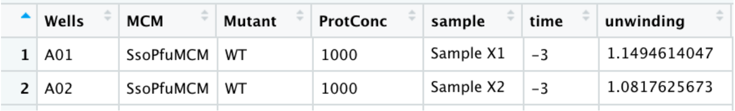
2) .csv Metadata file. This is made by the user in a format compatible for use with the plater package. It carries the information about what is in each experimental well. It will be loaded into R as:



	Wells	MCM	Mutant	ProtConc
1	A01	SsoPfuMCM	WT	1000
2	A02	SsoPfuMCM	WT	1000

Key output:

A long format dataset of unwinding values, expressed as a percentage of maximum fluorescent. Time is adjusted so ATP is added at t=0.



	Wells	MCM	Mutant	ProtConc	sample	time	unwinding
1	A01	SsoPfuMCM	WT	1000	Sample X1	-3	1.1494614047
2	A02	SsoPfuMCM	WT	1000	Sample X2	-3	1.0817625673

7.4.1 Standardizing fluorescent helicase data to the no helicase and maximum fluorescence controls.

```
library(readxl)     # for reading excel files
library(tidyr)      # for manipulating data from wide to long format
library(dplyr)      # group based analysis (i.e. means or max())
library(ggplot2)    # for plotting the graphs, requires long format data
library(lubridate)  # for translating data in the %min/ %sec format to
                    # seconds
library(plater)     # essential for plate based analysis (loading in meta-
data)

# Set your working directory (i.e. where your files are located)
setwd("/../directory")
df_wide <- read_excel("../fhelicase_20210512_ProtConcSsoPfuMacP1.xlsx",
                                                           skip = 13)



# Create a metadata table of wells and their identities
plate_layout <- read_plate(file = "fhelicase_20210513_ProtConc1.csv",
                                              well_ids_column = "Wells")

# Rename columns to MCM
names(df_wide)[1] <- "Wells"
names(df_wide)[2] <- "sample"

# Merge raw data and metadata
df_wide <- na.omit(merge(x = plate_layout, y = df_wide, by = "Wells",
                                                        all.y = TRUE))

# Convert data from wide to long format
final   <- rev(names(df_wide))[1]
df_long <- gather(df_wide, time, unwinding, `0 min`: final)

# Extract all MF + NH rows
MF <- subset(df_long, df_long$MCM=="MF")
NH <- subset(df_long, df_long$MCM=="NH")

# Remove from df_long
df_long <- subset(df_long, df_long$Well!="MF")
df_long <- subset(df_long, df_long$Well!="NH")

# Calculate averages (or maximums) if controls by groups
MF_avg <- MF %>%
   group_by(MCM) %>% # Add in extra groups here if required, e.g. Salt
   summarize(avg=mean(unwinding))

NH_avg <- NH %>%
   group_by(MCM) %>% # Add in extra groups here if required, e.g. Salt
   summarize(avg=mean(unwinding))

# Subtract appropriate NH values in unwinding column
rel.unwind <- c(0) # Creates a list with just the value 0
```

```r
# Initiates a loop that goes through each row in df_long
# in each loop, the correct controls are identified.
# the no helicase control is subtracted from the raw data.
# Data are then divided by the maximum fluorescence to get
# the proportion substrate unwound.
for (i in 1:nrow(df_long)) {

 # Extract the row
  unwind.row <- df_long[i,]

 # Change the highlighted section if different controls are used.
 # E.g, subset(MF_avg, MF_avg$ConcSalt==unwind.row$ConcSalt, select = avg)
 # ^ this would select controls where different salts + concentrations are
used.
 # Extract the values from MF_avg + NH according to these variables.
  baselineMF <- subset(MF_avg,
                         select = avg)

  baselineNH <- subset(NH_avg,
                         select = avg)

  # Extract raw unwinding value
  unwind.row <- unwind.row[1,ncol(unwind.row)]

  # Subtract appropriate NH
  unwind.row <- unwind.row - baselineNH$avg

  # Divide by appropriate MF
  unwind.row <- unwind.row / baselineMF$avg

  # Merge to list
  rel.unwind <- c(rel.unwind, unwind.row)

}
# Remove first zero that was required to create list
rel.unwind <- rel.unwind[-1]

# Merge to data frame
df_long <- cbind(df_long, rel.unwind)

# Convert time in % min /% sec format to just seconds.
# Data is recorded every 60 seconds, but when the ATP is added,
# 59/60 times, time is no longer divisible by 60.
# (i.e. 1 min 0s, becomes 1 min 23s. See Data required 1)
# This disparity can be used to identify where ATP was added
df_long$time <- as.numeric(as.period(df_long$time, unit = "sec"))

# Organising time so time=0
# Get a list of the unique time values
time_zero <- unique(df_long$time)

t1 <- c(0)
for (i in 1:length(time_zero)) {

  # Set a variable
  x <- time_zero[i]
  if (x%%60 != 0) { # Where is time no longer divisible by 60?
    t1 <- c(t1, x)
  }
}
t1 <- t1[-1]

# Get the minimum value (t=1)
t1 <- min(t1)

# Get t=0
t0 <- time_zero[which(time_zero==t1)-1]
df_long$time <- df_long$time - t0 # Adjusts all values to this point
df_long$time <- df_long$time/60 # Convert to minutes

# Convert unwinding to a %.
# Usually write file after this line. E.g. write.csv(df_long, name.csv)
df_long$unwinding <- df_long$rel.unwind * 100

# tidy up environment
rm(list=setdiff(ls(), "df_long"))
```

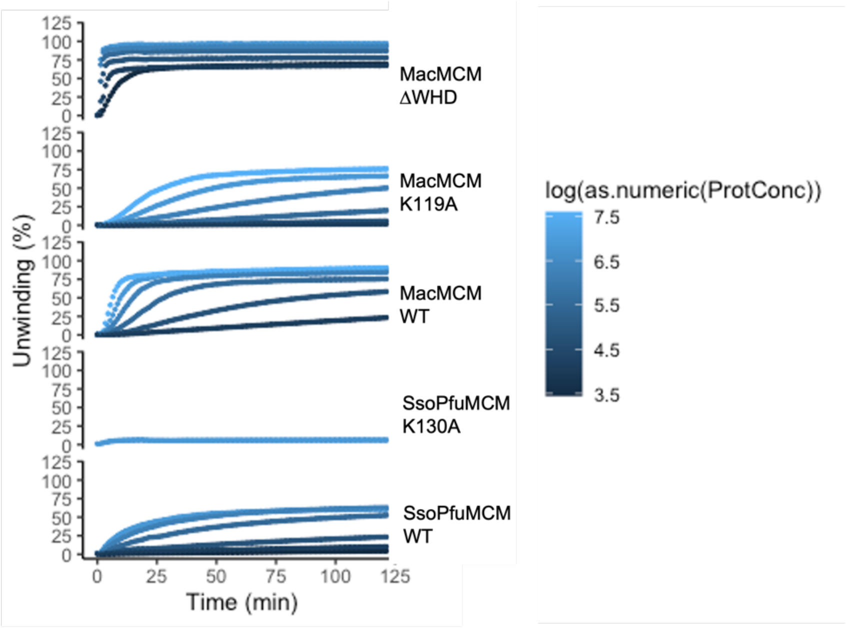
7.4.2 Downstream analysis

Calculation of summary statistics:

This output data can then be subjected to routine downstream analyses. E.g., calculation of means, sd, se:

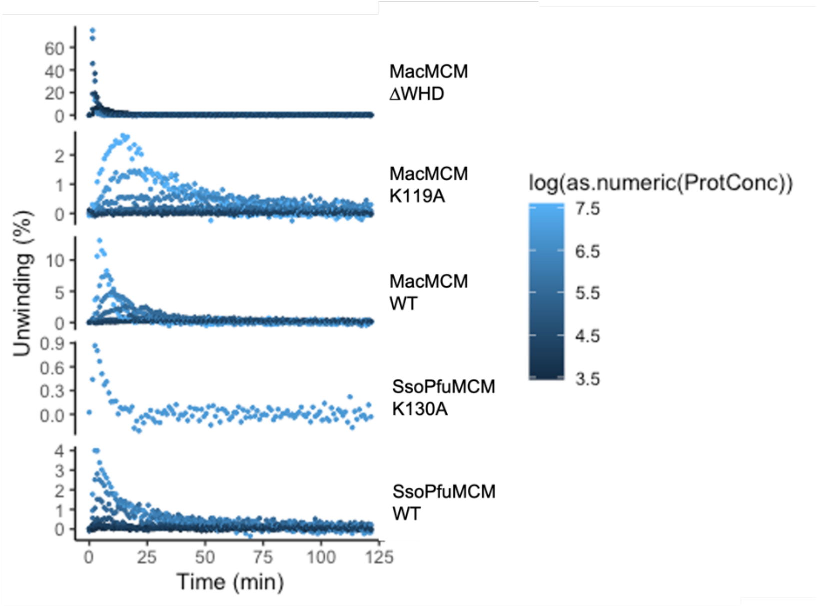
T=30 measurements can be easily extracted from this data (net unwinding).

```
# calculate mean, n, sd and se
df_long<-df_long %>%
  group_by(time, MCM, Mutant, ProtConc) %>% # adjust this if more conditions used
  summarize(avg=mean(unwinding), n=n(), sd=sd(unwinding), se=sd/sqrt(n))

# remove controls
df_long <- subset(df_long, df_long$MCM!='MF'&df_long$MCM!='NH')

# merging
df_long$expt <- paste(df_long$MCM, df_long$Mutant) # change if more conditions
```



Calculation of lag time:

```
laggy <- df_long %>%
    group_by(MCM, Mutant, ProtConc, expt) %>% #
    arrange(time) %>%
    mutate(diff = avg - lag(avg, default = first(avg), n=1)) # first deriv
ative, where x2-x1=1
```

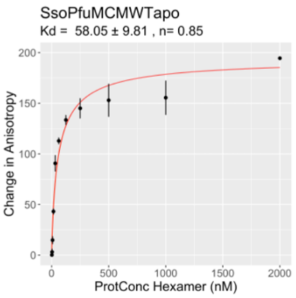


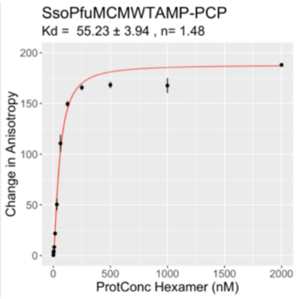
Lag time values may then be easily extracted:

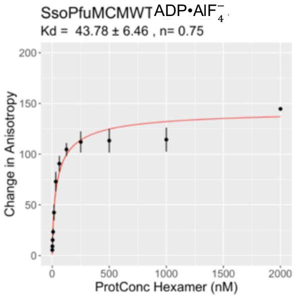
```
laggy_max <- laggy %>%
  group_by(MCM, Mutant, ProtConc, expt)%>%
  filter(diff == max(diff))
```

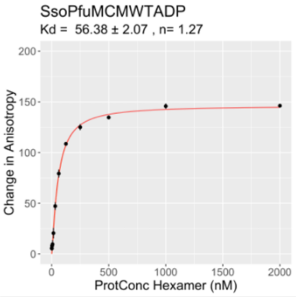
7.5 Anisotropy fits

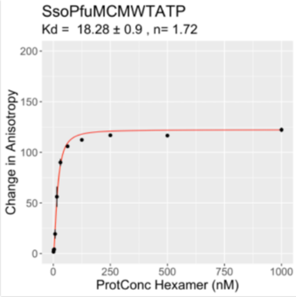
The equilibrium dissociation constant (K_d) was calculated from anisotropy experiments (section 3.8.2), from a Langmuir binding isotherm fit with Hill coefficient, *n* (see section 2.5.3).

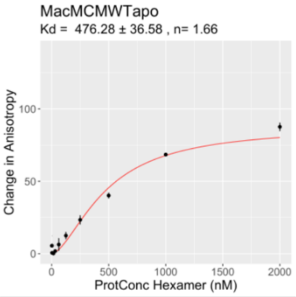


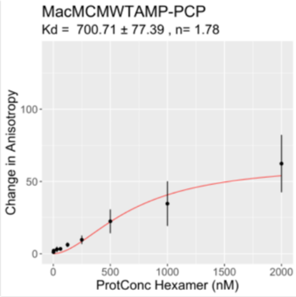


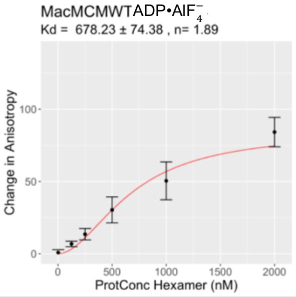


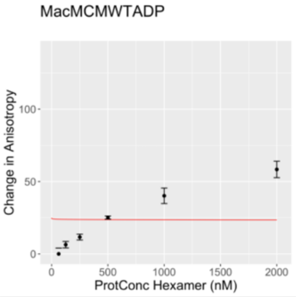


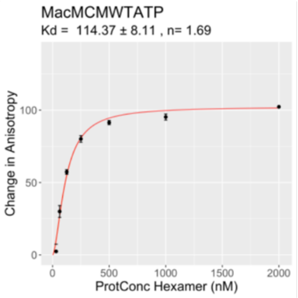




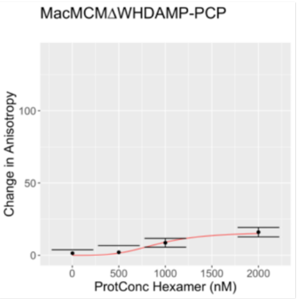


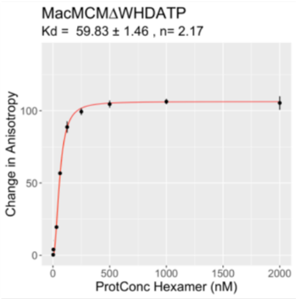


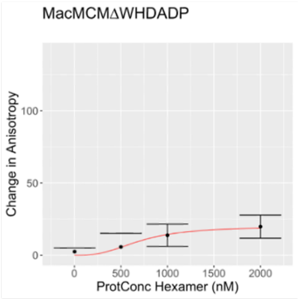


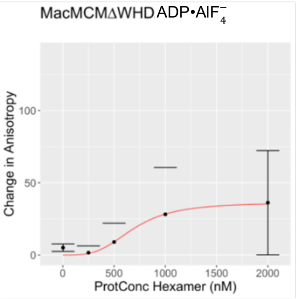


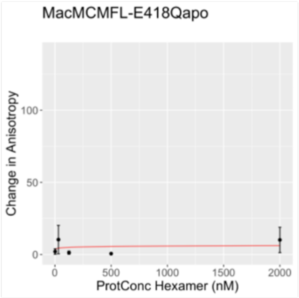
Continued…

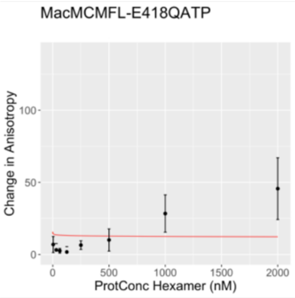

MacMCMΔWHDAMP-PCP
Change in Anisotropy
0
50
100
0
500
1000
1500
2000
ProtConc Hexamer (nM)

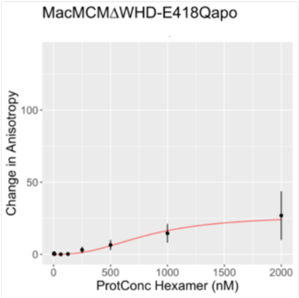

MacMCMΔWHDATP
Kd = 59.83 ± 1.46 , n= 2.17
Change in Anisotropy
0
50
100
0
500
1000
1500
2000
ProtConc Hexamer (nM)

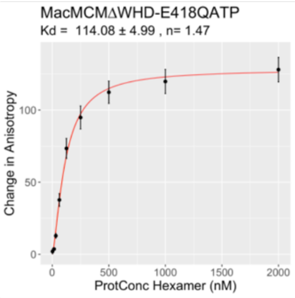

MacMCMΔWHDADP
Change in Anisotropy
0
50
100
0
500
1000
1500
2000
ProtConc Hexamer (nM)


MacMCMΔWHD ADP•AlF$_4^-$
Change in Anisotropy
0
50
100
0
500
1000
1500
2000
ProtConc Hexamer (nM)


MacMCMFL-E418Qapo
Change in Anisotropy
0
50
100
0
500
1000
1500
2000
ProtConc Hexamer (nM)

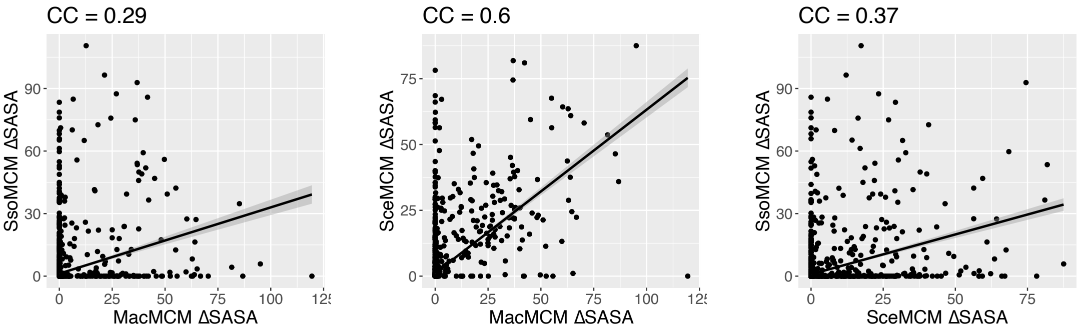

MacMCMFL-E418QATP
Change in Anisotropy
0
50
100
0
500
1000
1500
2000
ProtConc Hexamer (nM)


MacMCMΔWHD-E418Qapo
Change in Anisotropy
0
50
100
0
500
1000
1500
2000
ProtConc Hexamer (nM)


MacMCMΔWHD-E418QATP
Kd = 114.08 ± 4.99 , n= 1.47
Change in Anisotropy
0
50
100
0
500
1000
1500
2000
ProtConc Hexamer (nM)

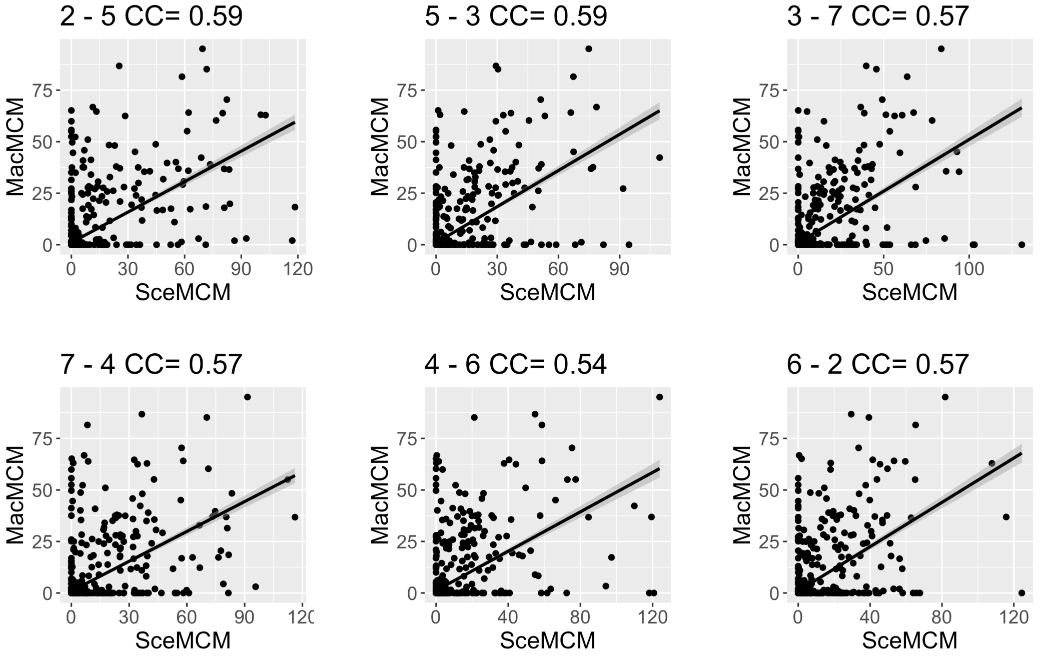
7.6 Core interface correlation plots

Correlation plots comparing *Mac*MCM, *Sso*MCM (PDB: 6MII)[100] and *Sce*MCM (PDB 6EYC) [140] interfaces (see section 4.4.3.2).



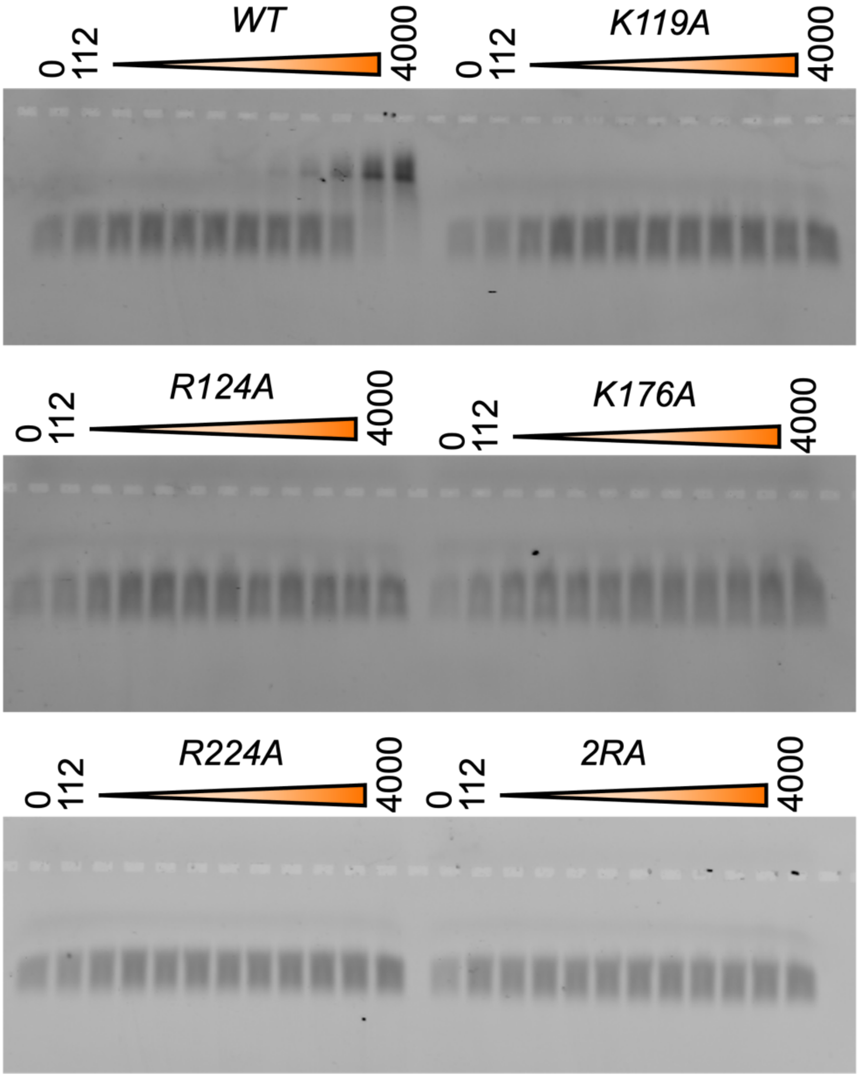
7.7 Correlation of *Mac*MCM and *Sce*MCM interfaces

Correlation plots comparing *Mac*MCM interfaces with each interface of *Sce*MCM (PDB: 6EYC)[140]. See section 4.4.3.2.



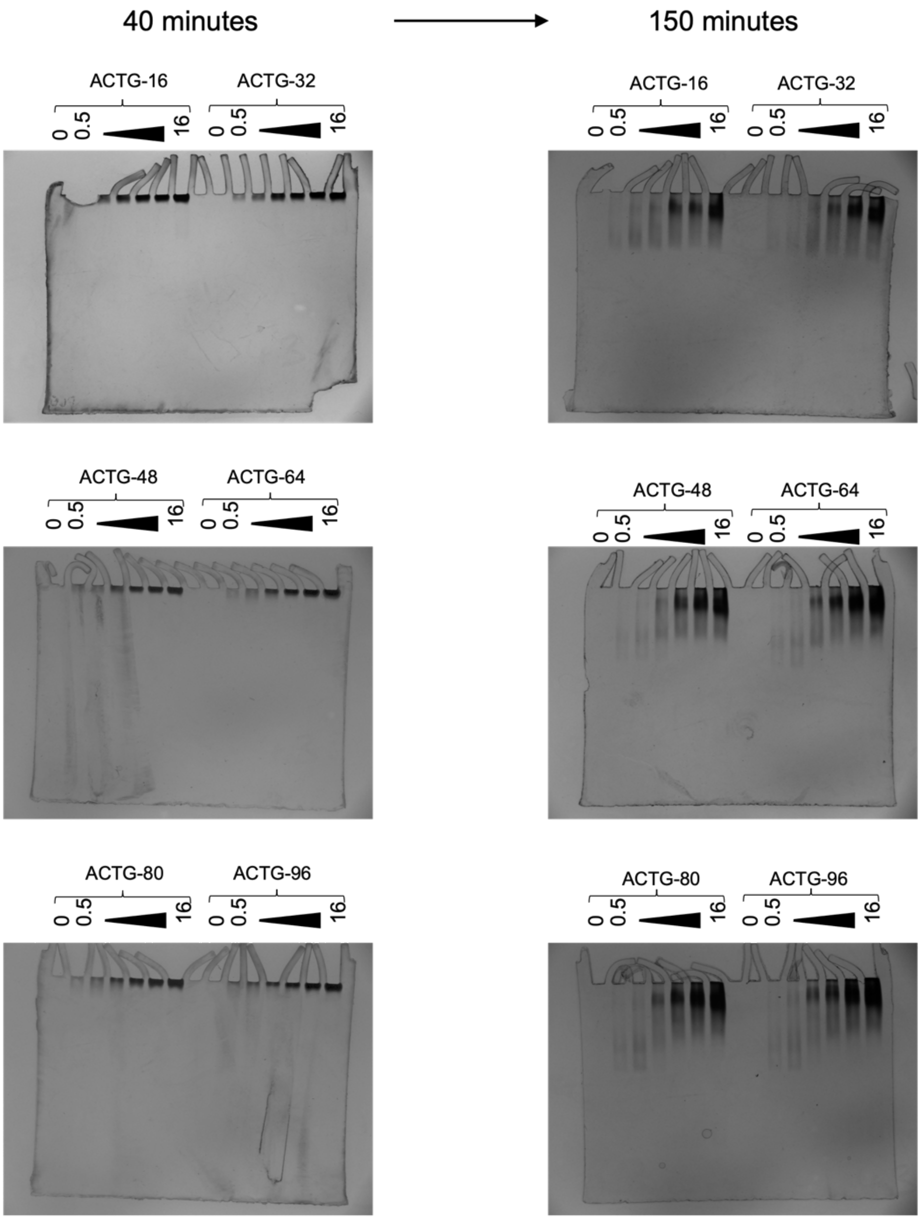
7.8 DNA binding mutants EMSA

EMSA of *Mac*MCM^{FL} DNA-binding mutants with a forked DNA substrate. Experiment as outlined in section 4.5.3.



7.9 Native PAGE – Coomassie stained gels

Gels in section 5.8.3, were stained with Coomassie dye after SYBR-GOLD staining, thus demonstrating the migration of protein through the gel.



Chapter 8 - References

1. Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5088–5090 (1977).

2. Stanier, R. Y. & Van Niel, C. B. The concept of a bacterium. *Arch. Mikrobiol.* **42**, 17–35 (1962).

3. Woese, C. R., Kandler, O. & Wheelis, M. L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 4576–4579 (1990).

4. Baker, B. J. *et al.* Diversity, ecology and evolution of Archaea. *Nat. Microbiol.* **5**, 887–900 (2020).

5. Imachi, H. *et al.* Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature* **577**, 519–525 (2020).

6. Sun, Y., Liu, Y., Pan, J., Wang, F. & Li, M. Perspectives on cultivation strategies of Archaea. *Microb. Ecol.* **79**, 770–784 (2020).

7. Bult, C. J. *et al.* Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273**, 1058–1073 (1996).

8. Hua, Z.-S. *et al.* Ecological roles of dominant and rare prokaryotes in acid mine drainage revealed by metagenomics and metatranscriptomics. *ISME J.* **9**, 1280–1294 (2015).

9. Iverson, V. *et al.* Untangling genomes from metagenomes: revealing an uncultured class of marine euryarchaeota. *Science* **335**, 587–590 (2012).

10. Collins, R. E., Rocap, G. & Deming, J. W. Persistence of bacterial and archaeal communities in sea ice through an Arctic winter. *Environ. Microbiol.* **12**, 1828–1841 (2010).

11. Guy, L. & Ettema, T. J. G. The archaeal “TACK” superphylum and the origin of eukaryotes. *Trends Microbiol.* **19**, 580–587 (2011).

12. Bitan-Banin, G., Ortenberg, R. & Mevarech, M. Development of a gene knockout system for the halophilic archaeon *Haloferax volcanii* by use of the pyrE gene. *J. Bacteriol.* **185**, 772–778 (2003).

13. Wagner, M. *et al.* Versatile Genetic Tool Box for the Crenarchaeote *Sulfolobus acidocaldarius*. *Front. Microbiol.* **3**, 214 (2012).

14. Castelle, C. J. *et al.* Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Curr. Biol.* **25**, 690–701 (2015).

15. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).

16. Zaremba-Niedzwiedzka, K. *et al.* Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).

17. Huber, H. *et al.* A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature* **417**, 63–67 (2002).

18. Golyshina, O. V. *et al.* “ARMAN” archaea depend on association with euryarchaeal host in culture and in situ. *Nat. Commun.* **8**, 60 (2017).

19. Castelle, C. J. *et al.* Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nat. Rev. Microbiol.* **16**, 629–645 (2018).

20. Beam, J. P. *et al.* Ancestral absence of electron transport chains in patescibacteria and DPANN. *Front. Microbiol.* **11**, 1848 (2020).

21. Dombrowski, N. *et al.* Undinarchaeota illuminate DPANN phylogeny and the impact of gene transfer on archaeal evolution. *Nat. Commun.* **11**, 3939 (2020).

22. Jahn, U. *et al. Nanoarchaeum equitan*s and *Ignicoccus hospitalis*: new insights into a unique, intimate association of two archaea. *J. Bacteriol.* **190**, 1743–1750 (2008).

23. Zengler, K. & Zaramela, L. S. The social network of microorganisms - how auxotrophies shape complex communities. *Nat. Rev. Microbiol.* **16**, 383–390 (2018).

24. Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179 (2015).

25. Seitz, K. W., Lazar, C. S., Hinrichs, K.-U., Teske, A. P. & Baker, B. J. Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. *ISME J.* **10**, 1696–1705 (2016).

26. Klinger, C. M., Spang, A., Dacks, J. B. & Ettema, T. J. G. Tracing the archaeal origins of eukaryotic membrane-trafficking system building blocks. *Mol. Biol. Evol.* **33**, 1528–1541 (2016).

27. Akıl, C. & Robinson, R. C. Genomes of Asgard archaea encode profilins that regulate actin. *Nature* **562**, 439–443 (2018).

28. Yutin, N. & Koonin, E. V. Archaeal origin of tubulin. *Biol. Direct* **7**, 10 (2012).

29. Neveu, E., Khalifeh, D., Salamin, N. & Fasshauer, D. Prototypic SNARE proteins are encoded in the genomes of Heimdallarchaeota, potentially bridging the gap between the prokaryotes and eukaryotes. *Curr. Biol.* **30**, 2468-2480.e5 (2020).

30. Yutin, N., Wolf, M. Y., Wolf, Y. I. & Koonin, E. V. The origins of phagocytosis and eukaryogenesis. *Biol. Direct* **4**, 9 (2009).

31. Da Cunha, V., Gaia, M., Nasir, A. & Forterre, P. Asgard archaea do not close the debate about the universal tree of life topology. *PLoS genetics* vol. 14 e1007215 (2018).

32. Da Cunha, V., Gaia, M., Gadelle, D., Nasir, A. & Forterre, P. Lokiarchaea are close relatives of euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. *PLoS Genet.* **13**, e1006810 (2017).

33. Zhou, Z., Liu, Y., Li, M. & Gu, J.-D. Two or three domains: a new view of tree of life in the genomics era. *Appl. Microbiol. Biotechnol.* **102**, 3049–3058 (2018).

34. Lake, J. A., Henderson, E., Oakes, M. & Clark, M. W. Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **81**, 3786–3790 (1984).

35. Williams, T. A., Cox, C. J., Foster, P. G., Szöllősi, G. J. & Embley, T. M. Phylogenomics provides robust support for a two-domains tree of life. *Nat. Ecol. Evol.* **4**, 138–147 (2020).

36. Jain, S., Caforio, A. & Driessen, A. J. M. Biosynthesis of archaeal membrane ether lipids. *Front. Microbiol.* **5**, 641 (2014).

37. Knopp, M., Stockhorst, S., van der Giezen, M., Garg, S. G. & Gould, S. B. The asgard archaeal-unique contribution to protein families of the eukaryotic common ancestor was 0.3. *Genome Biol. Evol.* **13**, (2021).

38. Booth, A. & Doolittle, W. F. Eukaryogenesis, how special really? *Proc. Natl. Acad. Sci. U. S. A.* **112**, 10278–10285 (2015).

39. O'Donnell, M. E. & Li, H. The ring-shaped hexameric helicases that function at DNA replication forks. *Nat. Struct. Mol. Biol.* **25**, 122–130 (2018).

40. Ishino, Y. & Ishino, S. Rapid progress of DNA replication studies in Archaea, the third domain of life. *Sci. China Life Sci.* **55**, 386–403 (2012).

41. LeBowitz, J. H. & McMacken, R. The *Escherichia coli* dnaB replication protein is a DNA helicase. *J. Biol. Chem.* **261**, 4738–4748 (1986).

42. Gefter, M. L., Hirota, Y., Kornberg, T., Wechsler, J. A. & Barnoux, C. Analysis of DNA polymerases II and 3 in mutants of *Escherichia coli* thermosensitive for DNA synthesis. *Proc. Natl. Acad. Sci. U. S. A.* **68**, 3150–3153 (1971).

43. Barry, E. R. & Bell, S. D. DNA replication in the archaea. *Microbiol. Mol. Biol. Rev.* **70**, 876–887 (2006).

44. O'Donnell, M., Langston, L. & Stillman, B. Principles and concepts of DNA replication in bacteria, archaea, and eukarya. *Cold Spring Harb. Perspect. Biol.* **5**, (2013).

45. Bell, S. P. & Kaguni, J. M. Helicase loading at chromosomal origins of replication. *Cold Spring Harb. Perspect. Biol.* **5**, (2013).

46. Vashee, S., Simancek, P., Challberg, M. D. & Kelly, T. J. Assembly of the human origin recognition complex. *J. Biol. Chem.* **276**, 26666–26673 (2001).

47. Toledo, L. I. *et al.* ATR prohibits replication catastrophe by preventing global exhaustion of RPA. *Cell* **155**, 1088–1103 (2013).

48. Devbhandari, S. & Remus, D. Rad53 limits CMG helicase uncoupling from DNA synthesis at replication forks. *Nat. Struct. Mol. Biol.* **27**, 461–471 (2020).

49. Moreno, S. P. & Gambus, A. Mechanisms of eukaryotic replisome disassembly. *Biochem. Soc. Trans.* **48**, 823–836 (2020).

50. Sedlackova, H. *et al.* Equilibrium between nascent and parental MCM proteins protects replicating genomes. *Nature* **587**, 297–302 (2020).

51. Petropoulos, M., Champeris Tsaniras, S., Taraviras, S. & Lygerou, Z. Replication licensing aberrations, replication stress, and genomic instability. *Trends Biochem. Sci.* **44**, 752–764 (2019).

52. Maine, G. T., Sinha, P. & Tye, B. K. Mutants of *S. cerevisiae* defective in the maintenance of minichromosomes. *Genetics* **106**, 365–385 (1984).

53. Chong, J. P., Thömmes, P. & Blow, J. J. The role of MCM/P1 proteins in the licensing of DNA replication. *Trends Biochem. Sci.* **21**, 102–106 (1996).

54. Crevel, G., Ivetic, A., Ohno, K., Yamaguchi, M. & Cotterill, S. Nearest neighbour analysis of MCM protein complexes in *Drosophila melanogaster*. *Nucleic Acids Res.* **29**, 4834–4842 (2001).

55. Ishimi, Y. A DNA helicase activity is associated with an MCM4, -6, and -7 protein complex. *J. Biol. Chem.* **272**, 24508–24513 (1997).

56. Kanter, D. M., Bruck, I. & Kaplan, D. L. Mcm subunits can assemble into two different active unwinding complexes. *J. Biol. Chem.* **283**, 31172–31182 (2008).

57. Sato, M. *et al.* Electron microscopic observation and single-stranded DNA binding activity of the Mcm4,6,7 complex. *J. Mol. Biol.* **300**, 421–431 (2000).

58. Davey, M. J., Indiani, C. & O'Donnell, M. Reconstitution of the Mcm2-7p heterohexamer, subunit arrangement, and ATP site architecture. *J. Biol. Chem.* **278**, 4491–4499 (2003).

59. Bochman, M. L. & Schwacha, A. Differences in the single-stranded DNA binding activities of MCM2-7 and MCM467: MCM2 and MCM5 define a slow ATP-dependent step. *J. Biol. Chem.* **282**, 33795–33804 (2007).

60. Coster, G., Frigola, J., Beuron, F., Morris, E. P. & Diffley, J. F. X. Origin licensing requires ATP binding and hydrolysis by the MCM replicative helicase. *Mol. Cell* **55**, 666–677 (2014).

61. Xu, Y. *et al.* Archaeal orthologs of Cdc45 and GINS form a stable complex that stimulates the helicase activity of MCM. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 13390–13395 (2016).

62. Bochman, M. L. & Schwacha, A. The Mcm2-7 complex has *in vitro* helicase activity. *Mol. Cell* **31**, 287–293 (2008).

63. Hesketh, E. L. *et al.* DNA induces conformational changes in a recombinant human minichromosome maintenance complex. *J. Biol. Chem.* **290**, 7973–7979 (2015).

64. Jose, D., Weitzel, S. E., Jing, D. & von Hippel, P. H. Assembly and subunit stoichiometry of the functional helicase-primase (primosome) complex of bacteriophage T4. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 13596–13601 (2012).

65. Pacek, M., Tutter, A. V., Kubota, Y., Takisawa, H. & Walter, J. C. Localization of MCM2-7, Cdc45, and GINS to the site of DNA unwinding during eukaryotic DNA replication. *Mol. Cell* **21**, 581–587 (2006).

66. Takayama, Y. *et al.* GINS, a novel multiprotein complex required for chromosomal DNA replication in budding yeast. *Genes Dev.* **17**, 1153–1165 (2003).

67. Moir, D., Stewart, S. E., Osmond, B. C. & Botstein, D. Cold-sensitive cell-division-cycle mutants of yeast: isolation, properties, and pseudoreversion studies. *Genetics* **100**, 547–563 (1982).

68. Moyer, S. E., Lewis, P. W. & Botchan, M. R. Isolation of the Cdc45/Mcm2–7/GINS (CMG) complex, a candidate for the eukaryotic DNA replication fork helicase. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 10236–10241 (2006).

69. Conti, C. *et al.* Replication fork velocities at adjacent replication origins are coordinately modified during DNA replication in human cells. *Mol. Biol. Cell* **18**, 3059–3067 (2007).

70. Burnham, D. R., Kose, H. B., Hoyle, R. B. & Yardimci, H. The mechanism of DNA unwinding by the eukaryotic replicative helicase. *Nat. Commun.* **10**, 2159 (2019).

71. Kose, H. B., Xie, S., Cameron, G., Strycharska, M. S. & Yardimci, H. Duplex DNA engagement and RPA oppositely regulate the DNA-unwinding rate of CMG helicase. *Nat. Commun.* **11**, 3713 (2020).

72. Lewis, J. S. *et al.* Single-molecule visualization of *Saccharomyces cerevisiae* leading-strand synthesis reveals dynamic interaction between MTC and the replisome. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 10630–10635 (2017).

73. Wasserman, M. R., Schauer, G. D., O’Donnell, M. E. & Liu, S. Replication fork activation is enabled by a single-stranded DNA gate in CMG helicase. *Cell* **178**, 600–611 (2019).

74. Stano, N. M. *et al.* DNA synthesis provides the driving force to accelerate DNA unwinding by a helicase. *Nature* **435**, 370–373 (2005).

75. Manosas, M., Xi, X. G., Bensimon, D. & Croquette, V. Active and passive mechanisms of helicases. *Nucleic Acids Res.* **38**, 5518–5526 (2010).

76. Barberis, M., Spiesser, T. W. & Klipp, E. Replication origins and timing of temporal replication in budding yeast: how to solve the conundrum? *Curr. Genomics* **11**, 199–211 (2010).

77. Sakai, H. D. & Kurosawa, N. *Saccharolobus caldissimu*s 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 solfataricu*s comb. nov. and *Sulfolobus shibatae* as *Saccharolobus shibatae* comb. nov. *Int. J. Syst. Evol. Microbiol.* **68**, 1271–1278 (2018).

78. Kelman, Z., Lee, J. K. & Hurwitz, J. The single minichromosome maintenance protein of *Methanobacterium thermoautotrophicum* DeltaH contains DNA helicase activity. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 14783–14788 (1999).

79. Chong, J. P., Hayashi, M. K., Simon, M. N., Xu, R. M. & Stillman, B. A double-hexamer archaeal minichromosome maintenance protein is an ATP-dependent DNA helicase. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 1530–1535 (2000).

80. Atanassova, N. & Grainge, I. Biochemical characterization of the minichromosome maintenance (MCM) protein of the crenarchaeote *Aeropyrum pernix* and its

interactions with the origin recognition complex (ORC) proteins. *Biochemistry* **47**, 13362–13370 (2008).

81. Grainge, I., Scaife, S. & Wigley, D. B. Biochemical analysis of components of the pre-replication complex of *Archaeoglobus fulgidus*. *Nucleic Acids Res.* **31**, 4888–4898 (2003).

82. Carpentieri, F., De Felice, M., De Falco, M., Rossi, M. & Pisani, F. M. Physical and functional interaction between the mini-chromosome maintenance-like DNA helicase and the single-stranded DNA binding protein from the crenarchaeon *Sulfolobus solfataricus*. *J. Biol. Chem.* **277**, 12118–12127 (2002).

83. Barry, E. R., McGeoch, A. T., Kelman, Z. & Bell, S. D. Archaeal MCM has separable processivity, substrate choice and helicase domains. *Nucleic Acids Res.* **35**, 988–998 (2007).

84. Yoshimochi, T., Fujikane, R., Kawanami, M., Matsunaga, F. & Ishino, Y. The GINS complex from *Pyrococcus furiosus* stimulates the MCM helicase activity. *J. Biol. Chem.* **283**, 1601–1609 (2008).

85. Shin, J.-H., Jiang, Y., Grabowski, B., Hurwitz, J. & Kelman, Z. Substrate requirements for duplex DNA translocation by the eukaryal and archaeal minichromosome maintenance helicases. *J. Biol. Chem.* **278**, 49053–49062 (2003).

86. Shin, J.-H., Santangelo, T. J., Xie, Y., Reeve, J. N. & Kelman, Z. Archaeal minichromosome maintenance (MCM) helicase can unwind DNA bound by archaeal histones and transcription factors. *J. Biol. Chem.* **282**, 4908–4915 (2007).

87. Shin, J.-H. & Kelman, Z. The replicative helicases of bacteria, archaea, and eukarya can unwind RNA-DNA hybrid substrates. *J. Biol. Chem.* **281**, 26914–26921 (2006).

88. Graham, B. W., Schauer, G. D., Leuba, S. H. & Trakselis, M. A. Steric exclusion and wrapping of the excluded DNA strand occurs along discrete external binding paths during MCM helicase unwinding. *Nucleic Acids Res.* **39**, 6585–6595 (2011).

89. Xia, Y. *et al.* The helicase activity of hyperthermophilic archaeal MCM is enhanced at high temperatures by lysine methylation. *Front. Microbiol.* **6**, 1247 (2015).

90. Jenkinson, E. R. & Chong, J. P. J. Minichromosome maintenance helicase activity is controlled by N- and C-terminal motifs and requires the ATPase domain helix-2 insert. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 7613–7618 (2006).

91. Schermerhorn, K. M., Tanner, N., Kelman, Z. & Gardner, A. F. High-temperature single-molecule kinetic analysis of thermophilic archaeal MCM helicases. *Nucleic Acids Res.* **44**, 8764–8771 (2016).

92. Lang, S. & Huang, L. The *Sulfolobus solfataricus* GINS complex stimulates DNA binding and processive DNA unwinding by minichromosome maintenance helicase. *J. Bacteriol.* **197**, 3409–3420 (2015).

93. Marinsek, N. *et al.* GINS, a central nexus in the archaeal DNA replication fork. *EMBO Rep.* **7**, 539–545 (2006).

94. Oyama, T. *et al.* Architectures of archaeal GINS complexes, essential DNA replication initiation factors. *BMC Biol.* **9**, 28 (2011).

95. Wu, Z., Liu, J., Yang, H. & Xiang, H. DNA replication origins in archaea. *Front. Microbiol.* **5**, 179 (2014).

96. Jae-Ho Shin, Rachel Mauro, Melamud, E. & Rajesh Kasiviswanathan. Cloning and Partial Characterization of the *Methanococcoides burtonii* minichromosome maintenance (MCM) Helicase. *Bios* **77**, 37–41 (2006).

97. Goswami, K., Arora, J. & Saha, S. Characterization of the MCM homohexamer from the thermoacidophilic euryarchaeon *Picrophilus torridus*. *Sci. Rep.* **5**, 9057 (2015).

98. Pucci, B., De Felice, M., Rossi, M., Onesti, S. & Pisani, F. M. Amino acids of the *Sulfolobus solfataricus* mini-chromosome maintenance-like DNA helicase involved in DNA binding/remodeling. *J. Biol. Chem.* **279**, 49222–49228 (2004).

99. Moreau, M. J., McGeoch, A. T., Lowe, A. R., Itzhaki, L. S. & Bell, S. D. ATPase site architecture and helicase mechanism of an archaeal MCM. *Mol. Cell* **28**, 304–314 (2007).

100. Meagher, M., Epling, L. B. & Enemark, E. J. DNA translocation mechanism of the MCM complex and implications for replication initiation. *Nat. Commun.* **10**, 3117 (2019).

101. Graham, B. W. *et al.* Control of hexamerization, assembly, and excluded strand specificity for the *Sulfolobus solfataricus* MCM helicase. *Biochemistry* **57**, 5672–5682 (2018).

102. Haugland, G. T., Shin, J.-H., Birkeland, N.-K. & Kelman, Z. Stimulation of MCM helicase activity by a Cdc6 protein in the archaeon *Thermoplasma acidophilum*. *Nucleic Acids Res.* **34**, 6337–6344 (2006).

103. Chen, Y.-J. *et al.* Structural polymorphism of *Methanothermobacter thermautotrophicus* MCM. *J. Mol. Biol.* **346**, 389–394 (2005).

104. Shin, J.-H., Heo, G.-Y. & Kelman, Z. The *Methanothermobacter thermautotrophicus* MCM helicase is active as a hexameric ring. *J. Biol. Chem.* **284**, 540–546 (2009).

105. Samson, R. Y., Abeyrathne, P. D. & Bell, S. D. Mechanism of archaeal MCM helicase recruitment to DNA replication origins. *Mol. Cell* **61**, 287–296 (2016).

106. Cannone, G., Visentin, S., Palud, A., Henneke, G. & Spagnolo, L. Structure of an octameric form of the minichromosome maintenance protein from the archaeon Pyrococcus abyssi. *Sci. Rep.* **7**, 42019 (2017).

107. Slaymaker, I. M. *et al.* Mini-chromosome maintenance complexes form a filament to remodel DNA structure and topology. *Nucleic Acids Res.* **41**, 3446–3456 (2013).

108. Jenkinson, E. R. *et al.* Mutations in subdomain B of the minichromosome maintenance (MCM) helicase affect DNA binding and modulate conformational transitions. *J. Biol. Chem.* **284**, 5654–5661 (2009).

109. De Felice, M. *et al.* A CDC6-like factor from the archaea *Sulfolobus solfataricus* promotes binding of the mini-chromosome maintenance complex to DNA. *J. Biol. Chem.* **279**, 43008–43012 (2004).

110. Rothenberg, E., Trakselis, M. A., Bell, S. D. & Ha, T. MCM forked substrate specificity involves dynamic interaction with the 5′-tail. *J. Biol. Chem.* **282**, 34229–34234 (2007).

111. Miller, J. M. & Enemark, E. J. Archaeal MCM proteins as an analog for the eukaryotic Mcm2-7 helicase to reveal essential features of structure and function. *Archaea* **2015**, 305497 (2015).

112. Fletcher, R. J. *et al.* The structure and function of MCM from archaeal *M. thermoautotrophicum*. *Nat. Struct. Biol.* **10**, 160–167 (2003).

113. Poplawski, A., Grabowski, B., Long, S. E. & Kelman, Z. The zinc finger domain of the archaeal minichromosome maintenance protein is required for helicase activity. *J. Biol. Chem.* **276**, 49371–49377 (2001).

114. Krishna, S. S., Majumdar, I. & Grishin, N. V. Structural classification of zinc fingers: survey and summary. *Nucleic Acids Res.* **31**, 532–550 (2003).

115. Liu, W., Pucci, B., Rossi, M., Pisani, F. M. & Ladenstein, R. Structural analysis of the *Sulfolobus solfataricus* MCM protein N-terminal domain. *Nucleic Acids Res.* **36**, 3235–3243 (2008).

116. Kasiviswanathan, R., Shin, J.-H., Melamud, E. & Kelman, Z. Biochemical characterization of the *Methanothermobacter thermautotrophicus* minichromosome maintenance (MCM) helicase N-terminal domains. *J. Biol. Chem.* **279**, 28358–28366 (2004).

117. Koonin, E. V. A common set of conserved motifs in a vast variety of putative nucleic acid-dependent ATPases including MCM proteins involved in the initiation of eukaryotic DNA replication. *Nucleic Acids Res.* **21**, 2541–2547 (1993).

118. Wiedemann, C. *et al.* Structure and regulatory role of the C-terminal winged helix domain of the archaeal minichromosome maintenance complex. *Nucleic Acids Res.* **43**, 2958–2967 (2015).

119. Froelich, C. A., Nourse, A. & Enemark, E. J. MCM ring hexamerization is a prerequisite for DNA-binding. *Nucleic Acids Res.* **43**, 9553–9563 (2015).

120. Fletcher, R. J. *et al.* Double hexamer disruption and biochemical activities of *Methanobacterium thermoautotrophicum* MCM. *J. Biol. Chem.* **280**, 42405–42410 (2005).

121. Shima, N., Buske, T. R. & Schimenti, J. C. Genetic screen for chromosome instability in mice: Mcm4 and breast cancer. *Cell Cycle* **6**, 1135–1140 (2007).

122. Shima, N. *et al.* A viable allele of Mcm4 causes chromosome instability and mammary adenocarcinomas in mice. *Nat. Genet.* **39**, 93–98 (2007).

123. Brewster, A. S. *et al.* Crystal structure of a near-full-length archaeal MCM: functional insights for an AAA+ hexameric helicase. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 20191–20196 (2008).

124. Miller, J. M., Arachea, B. T., Epling, L. B. & Enemark, E. J. Analysis of the crystal structure of an active MCM hexamer. *Elife* **3**, e03433 (2014).

125. Perera, H. M. & Trakselis, M. A. Amidst multiple binding orientations on fork DNA, Saccharolobus MCM helicase proceeds N-first for unwinding. *Elife* **8**, (2019).

126. Graham, B. W. *et al.* DNA interactions probed by Hydrogen-Deuterium Exchange (HDX) Fourier transform ion cyclotron resonance mass spectrometry confirm external binding sites on the Minichromosomal Maintenance (MCM) helicase. *J. Biol. Chem.* **291**, 12467–12480 (2016).

127. Mohammed Khalid, A. A., Parisse, P., Medagli, B., Onesti, S. & Casalis, L. Atomic Force Microscopy Investigation of the Interactions between the MCM Helicase and DNA. *Materials* **14**, (2021).

128. Froelich, C. A., Kang, S., Epling, L. B., Bell, S. P. & Enemark, E. J. A conserved MCM single-stranded DNA binding element is essential for replication initiation. *Elife* **3**, e01993 (2014).

129. Arias-Palomo, E., Puri, N., O'Shea Murray, V. L., Yan, Q. & Berger, J. M. Physical basis for the loading of a bacterial replicative helicase onto DNA. *Mol. Cell* **74**, 173-184.e4 (2019).

130. Enemark, E. J. & Joshua-Tor, L. Mechanism of DNA translocation in a replicative hexameric helicase. *Nature* **442**, 270–275 (2006).

131. Raghunathan, S., Kozlov, A. G., Lohman, T. M. & Waksman, G. Structure of the DNA binding domain of *E. coli* SSB bound to ssDNA. *Nat. Struct. Biol.* **7**, 648–652 (2000).

132. McGeoch, A. T., Trakselis, M. A., Laskey, R. A. & Bell, S. D. Organization of the archaeal MCM complex on DNA and implications for the helicase mechanism. *Nat. Struct. Mol. Biol.* **12**, 756–762 (2005).

133. Sakakibara, N. *et al.* Coupling of DNA binding and helicase activity is mediated by a conserved loop in the MCM protein. *Nucleic Acids Res.* **36**, 1309–1320 (2008).

134. Ogura, T. & Wilkinson, A. J. AAA+ superfamily ATPases: common structure--diverse function. *Genes Cells* **6**, 575–597 (2001).

135. Zhang, X. & Wigley, D. B. The "glutamate switch" provides a link between ATPase activity and ligand binding in AAA+ proteins. *Nat. Struct. Mol. Biol.* **15**, 1223–1227 (2008).

136. Prieß, M., Göddeke, H., Groenhof, G. & Schäfer, L. V. Molecular mechanism of ATP hydrolysis in an ABC transporter. *ACS Cent. Sci.* **4**, 1334–1343 (2018).

137. Barry, E. R., Lovett, J. E., Costa, A., Lea, S. M. & Bell, S. D. Intersubunit allosteric communication mediated by a conserved loop in the MCM helicase. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 1051–1056 (2009).

138. Crampton, D. J., Mukherjee, S. & Richardson, C. C. DNA-induced switch from independent to sequential dTTP hydrolysis in the bacteriophage T7 DNA helicase. *Mol. Cell* **21**, 165–174 (2006).

139. Li, N. *et al.* Structure of the eukaryotic MCM complex at 3.8 Å. *Nature* **524**, 186–191 (2015).

140. Croll, T. I. ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallogr. D Struct. Biol.* **74**, 519–530 (2018).

141. Sheu, Y.-J. & Stillman, B. Cdc7-Dbf4 phosphorylates MCM proteins via a docking site-mediated mechanism to promote S phase progression. *Mol. Cell* **24**, 101–113 (2006).

142. Bruck, I., Kanter, D. M. & Kaplan, D. L. Enabling association of the GINS protein tetramer with the mini chromosome maintenance (Mcm)2-7 protein complex by phosphorylated Sld2 protein and single-stranded origin DNA. *J. Biol. Chem.* **286**, 36414–36426 (2011).

143. Christopher F. J. Hardy, Dryga, O., Seematter, S., Paula M. B. Pahl & Sclafani, R. A. mcm5/cdc46-bob1 bypasses the requirement for the S phase activator Cdc7p. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 3151–3155 (1997).

144. Fletcher, R. J. & Chen, X. S. Biochemical activities of the BOB1 mutant in *Methanobacterium thermoautotrophicum* MCM. *Biochemistry* **45**, 462–467 (2006).

145. Fei, L. & Xu, H. Role of MCM2-7 protein phosphorylation in human cancer cells. *Cell Biosci.* **8**, 43 (2018).

146. Nitani, N., Yadani, C., Yabuuchi, H., Masukata, H. & Nakagawa, T. Mcm4 C-terminal domain of MCM helicase prevents excessive formation of single-stranded DNA at stalled replication forks. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 12973–12978 (2008).

147. Kimura, H., Ohtomo, T., Yamaguchi, M., Ishii, A. & Sugimoto, K. Mouse MCM proteins: complex formation and transportation to the nucleus. *Genes Cells* **1**, 977–993 (1996).

148. Leon, R. P., Tecklenburg, M. & Sclafani, R. A. Functional conservation of beta-hairpin DNA binding domains in the Mcm protein of *Methanobacterium thermoautotrophicum* and the Mcm5 protein of *Saccharomyces cerevisiae*. *Genetics* **179**, 1757–1768 (2008).

149. Ramey, C. J. & Sclafani, R. A. Functional conservation of the pre-sensor one beta-finger hairpin (PS1-hp) structures in mini-chromosome maintenance proteins of *Saccharomyces cerevisiae* and archaea. *G3* **4**, 1319–1326 (2014).

150. Hanson, P. I. & Whiteheart, S. W. AAA+ proteins: have engine, will work. *Nat. Rev. Mol. Cell Biol.* **6**, 519–529 (2005).

151. Bochman M.L., Bell S.P. & Schwacha A. Subunit organization of Mcm2-7 and the unequal role of active sites in ATP hydrolysis and viability. *Mol. Cell. Biol.* **28**, 5865–5873 (2008).

152. Sterner, J. M., Dew-Knight, S., Musahl, C., Kornbluth, S. & Horowitz, J. M. Negative regulation of DNA replication by the retinoblastoma protein is mediated by its association with MCM7. *Mol. Cell. Biol.* **18**, 2748–2757 (1998).

153. Ilves, I., Petojevic, T., Pesavento, J. J. & Botchan, M. R. Activation of the MCM2-7 helicase by association with Cdc45 and GINS proteins. *Mol. Cell* **37**, 247–258 (2010).

154. Kang, S., Warner, M. D. & Bell, S. P. Multiple functions for Mcm2-7 ATPase motifs during replication initiation. *Mol. Cell* **55**, 655–665 (2014).

155. Bell, S. P. & Stillman, B. ATP-dependent recognition of eukaryotic origins of DNA replication by a multiprotein complex. *Nature* **357**, 128–134 (1992).

156. Remus, D. *et al.* Concerted loading of Mcm2-7 double hexamers around DNA during DNA replication origin licensing. *Cell* **139**, 719–730 (2009).

157. Miller, T. C. R., Locke, J., Greiwe, J. F., Diffley, J. F. X. & Costa, A. Mechanism of head-to-head MCM double-hexamer formation revealed by cryo-EM. *Nature* **575**, 704–710 (2019).

158. Zhai, Y. *et al.* Open-ringed structure of the Cdt1-Mcm2-7 complex as a precursor of the MCM double hexamer. *Nat. Struct. Mol. Biol.* **24**, 300–308 (2017).

159. Samel, S. A. *et al.* A unique DNA entry gate serves for regulated loading of the eukaryotic replicative helicase MCM2-7 onto DNA. *Genes Dev.* **28**, 1653–1666 (2014).

160. Frigola, J., Remus, D., Mehanna, A. & Diffley, J. F. X. ATPase-dependent quality control of DNA replication origin licensing. *Nature* **495**, 339–343 (2013).

161. Fernández-Cid, A. *et al.* An ORC/Cdc6/MCM2-7 complex is formed in a multistep reaction to serve as a platform for MCM double-hexamer assembly. *Mol. Cell* **50**, 577–588 (2013).

162. Yuan, Z. *et al.* Structural basis of Mcm2-7 replicative helicase loading by ORC-Cdc6 and Cdt1. *Nat. Struct. Mol. Biol.* **24**, 316–324 (2017).

163. Yuan, Z. *et al.* Structural mechanism of helicase loading onto replication origin DNA by ORC-Cdc6. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 17747–17756 (2020).

164. Sun, J. *et al.* Cryo-EM structure of a helicase loading intermediate containing ORC-Cdc6-Cdt1-MCM2-7 bound to DNA. *Nat. Struct. Mol. Biol.* **20**, 944–951 (2013).

165. Guerrero-Puigdevall, M., Fernandez-Fuentes, N. & Frigola, J. Stabilisation of half MCM ring by Cdt1 during DNA insertion. *Nat. Commun.* **12**, 1746 (2021).

166. Abid Ali, F. *et al.* Cryo-EM structure of a licensed DNA replication origin. *Nat. Commun.* **8**, 1–10 (2017).

167. Lei, M. *et al.* Mcm2 is a target of regulation by Cdc7–Dbf4 during the initiation of DNA synthesis. *Genes Dev.* **11**, 3365–3374 (1997).

168. Eickhoff, P. *et al.* Molecular basis for ATP-hydrolysis-driven DNA translocation by the CMG helicase of the eukaryotic replisome. *Cell Rep.* **28**, 2673-2688.e8 (2019).

169. Rzechorzek, N. J., Hardwick, S. W., Jatikusumo, V. A., Chirgadze, D. Y. & Pellegrini, L. CryoEM structures of human CMG-ATPγS-DNA and CMG-AND-1 complexes. *Nucleic Acids Res.* **48**, 6980–6995 (2020).

170. Yuan, Z. *et al.* DNA unwinding mechanism of a eukaryotic replicative CMG helicase. *Nat. Commun.* **11**, 688 (2020).

171. Baretić, D. *et al.* Cryo-EM structure of the fork protection complex bound to CMG at a replication fork. *Mol. Cell* **78**, 926-940.e13 (2020).

172. Costa, A. *et al.* The structural basis for MCM2-7 helicase activation by GINS and Cdc45. *Nat. Struct. Mol. Biol.* **18**, 471–477 (2011).

173. Yuan, Z. *et al.* Structure of the eukaryotic replicative CMG helicase suggests a pumpjack motion for translocation. *Nat. Struct. Mol. Biol.* **23**, 217–224 (2016).

174. Sun, J. *et al.* The architecture of a eukaryotic replisome. *Nat. Struct. Mol. Biol.* **22**, 976–982 (2015).

175. Langston, L. & O'Donnell, M. Action of CMG with strand-specific DNA blocks supports an internal unwinding mode for the eukaryotic replicative helicase. *Elife* **6**, (2017).

176. Walters, A. D. & Chong, J. P. J. *Methanococcus maripaludis*: an archaeon with multiple functional MCM proteins? *Biochem. Soc. Trans.* **37**, 1–6 (2009).

177. Walters, A. D. & Chong, J. P. J. Non-essential MCM-related proteins mediate a response to DNA damage in the archaeon *Methanococcus maripaludis*. *Microbiology* **163**, 745–753 (2017).

178. Samson, R. Y. *et al.* Specificity and function of archaeal DNA replication initiator proteins. *Cell Rep.* **3**, 485–496 (2013).

179. Arias-Palomo, E., O’Shea, V. L., Hood, I. V. & Berger, J. M. The bacterial DnaC helicase loader is a DnaB ring breaker. *Cell* **153**, 438–448 (2013).

180. Shin, J.-H., Heo, G. Y. & Kelman, Z. The *Methanothermobacter thermautotrophicus* Cdc6-2 protein, the putative helicase loader, dissociates the minichromosome maintenance helicase. *J. Bacteriol.* **190**, 4091–4094 (2008).

181. Lam, S. Y., Yeung, R. C. Y., Yu, T.-H., Sze, K.-H. & Wong, K.-B. A rigidifying salt-bridge favors the activity of thermophilic enzyme at high temperatures at the expense of low-temperature activity. *PLoS Biol.* **9**, e1001027 (2011).

182. Studier, F. W. & Moffatt, B. A. Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *J. Mol. Biol.* **189**, 113–130 (1986).

183. Hellman, L. M. & Fried, M. G. Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nat. Protoc.* **2**, 1849–1861 (2007).

184. Fried, M. G. Measurement of protein-DNA interaction parameters by electrophoresis mobility shift assay. *Electrophoresis* **10**, 366–376 (1989).

185. Swinehart, D. F. The Beer-Lambert Law. *J. Chem. Educ.* **39**, 333 (1962).

186. Carter, S. G. & Karl, D. W. Inorganic phosphate assay with malachite green: an improvement and evaluation. *J. Biochem. Biophys. Methods* **7**, 7–13 (1982).

187. Drummen, G. P. C. Fluorescent probes and fluorescence (microscopy) techniques--illuminating biological and biomedical research. *Molecules* **17**, 14067–14090 (2012).

188. Marras, S. A. E., Kramer, F. R. & Tyagi, S. Efficiencies of fluorescence resonance energy transfer and contact-mediated quenching in oligonucleotide probes. *Nucleic Acids Res.* **30**, e122 (2002).

189. Cooper, A. Spectroscopy. in *Biophysical Chemistry* 21–69 (2004).

190. Bjornson, K. P., Amaratunga, M., Moore, K. J. & Lohman, T. M. Single-turnover kinetics of helicase-catalyzed DNA unwinding monitored continuously by fluorescence energy transfer. *Biochemistry* **33**, 14306–14316 (1994).

191. Pollard, T. D. A guide to simple and informative binding assays. *Mol. Biol. Cell* **21**, 4061–4067 (2010).

192. Heyduk, T. & Lee, J. C. Application of fluorescence energy transfer and polarization to monitor Escherichia coli cAMP receptor protein and lac promoter interaction. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 1744–1748 (1990).

193. minpack.lm: R Iiterface to the Levenberg-Marquardt nonlinear least-squares algorithm found in MINPACK, plus support for bounds. *Comprehensive R Archive Network (CRAN)* https://cran.r-project.org/web/packages/minpack.lm/index.html.

194. Loving, G. S., Sainlos, M. & Imperiali, B. Monitoring protein interactions and dynamics with solvatochromic fluorophores. *Trends Biotechnol.* **28**, 73–83 (2010).

195. Blech, M. Analysis of formulation-dependent colloidal and conformational stability of monoclonal antibodies. (2015).

196. Ó'Fágáin, C., Cummins, P. M. & O'Connor, B. F. Gel-filtration chromatography. *Methods Mol. Biol.* **1485**, 15–25 (2017).

197. Wyatt, P. J. Light scattering and the absolute characterization of macromolecules. *Anal. Chim. Acta* **272**, 1–40 (1993).

198. Dark, W. A. UV and dRI detectors in liquid chromatography: The workhorse detectors. *J. Chromatogr. Sci.* **24**, 495–498 (1986).

199. Zimm, B. H. The scattering of light and the radial distribution function of high polymer solutions. *J. Chem. Phys.* **16**, 1093–1099 (1948).

200. Cole, J. L., Lary, J. W., P Moody, T. & Laue, T. M. Analytical ultracentrifugation: sedimentation velocity and sedimentation equilibrium. *Methods Cell Biol.* **84**, 143–179 (2008).

201. Cooper, A. Hydrodynamics. in *Biophysical Chemistry* 82–98 (2004).

202. Lamm, O. Zur Bestimmung von Konzentrationsgradienten mittels gekrümmter Lichtstrahlen. *Zeitschrift für Physikalische Chemie* vol. 138A 313–331 (1928).

203. Schuck, P. Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and lamm equation modeling. *Biophys. J.* **78**, 1606–1619 (2000).

204. Harding, S. E., Rowe, A. J. & Horton, J. C. *Analytical ultracentrifugation in biochemistry and polymer science*. (Royal Society of Chemistry, 1992).

205. Louche, A., Salcedo, S. P. & Bigot, S. Protein-protein interactions: Pull-down assays. *Methods Mol. Biol.* **1615**, 247–255 (2017).

206. Blümich, B. Introduction to compact NMR: A review of methods. *Trends Analyt. Chem.* **83**, 2–11 (2016).

207. Wlodawer, A., Minor, W., Dauter, Z. & Jaskolski, M. Protein crystallography for aspiring crystallographers or how to avoid pitfalls and traps in macromolecular structure determination. *FEBS J.* **280**, 5705–5736 (2013).

208. Taylor, G. The phase problem. *Acta Crystallogr. D Biol. Crystallogr.* **59**, 1881–1890 (2003).

209. Bellizzi, J. J., Widom, J., Kemp, C. W. & Clardy, J. Producing selenomethionine-labeled proteins with a baculovirus expression vector system. *Structure* **7**, R263-7 (1999).

210. Rose, J. P., Wang, B.-C. & Weiss, M. S. Native SAD is maturing. *IUCrJ* **2**, 431–440 (2015).

211. Garman, E. & Murray, J. W. Heavy-atom derivatization. *Acta Crystallogr. D Biol. Crystallogr.* **59**, 1903–1913 (2003).

212. Powell, H. R. X-ray data processing. *Biosci. Rep.* **37**, (2017).

213. Diederichs, K. & Andrew Karplus, P. Improved R -factors for diffraction data analysis in macromolecular crystallography. *Nat. Struct. Biol.* **4**, 269–275 (1997).

214. Karplus, P. A. & Diederichs, K. Linking crystallographic model and data quality. *Science* **336**, 1030–1033 (2012).

215. Rossmann, M. G., Blow, D. M. & IUCr. The detection of sub-units within the crystallographic asymmetric unit. *Acta Crystallogr.* **15**, 24–31 (1962).

216. Read, R. J. Pushing the boundaries of molecular replacement with maximum likelihood. *Acta Crystallogr. D Biol. Crystallogr.* **57**, 1373–1382 (2001).

217. Chothia, C. & Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826 (1986).

218. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).

219. Terwilliger, T. C. *et al.* Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallogr. D Biol. Crystallogr.* **64**, 61–69 (2008).

220. Cowtan, K. D. & Zhang, K. Y. Density modification for macromolecular phase improvement. *Prog. Biophys. Mol. Biol.* **72**, 245–270 (1999).

221. Lehmann, M. S. & Pebay-Peyroula, E. Location of the sulfur atoms from the phased anomalous map using native protein data can be very helpful in tracing the peptide chain. *Acta Crystallogr. B* **48 (Pt 1)**, 115–116 (1992).

222. Read, R. J. *et al.* A new generation of crystallographic validation tools for the protein data bank. *Structure* **19**, 1395–1412 (2011).

223. Holcomb, J. *et al.* Protein crystallization: Eluding the bottleneck of X-ray crystallography. *AIMS Biophys* **4**, 557–575 (2017).

224. Gorrec, F. The MORPHEUS protein crystallization screen. *J. Appl. Crystallogr.* **42**, 1035–1042 (2009).

225. Winter, G. xia2 : an expert system for macromolecular crystallography data reduction. (2010) doi:10.1107/S0021889809045701.

226. Terwilliger, T. SOLVE and RESOLVE: automated structure solution, density modification and model building. *J. Synchrotron Radiat.* **11**, 49–52 (2004).

227. Cowtan, K. The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr. D Biol. Crystallogr.* **62**, 1002–1011 (2006).

228. Afonine, P. V. *et al.* Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D Biol. Crystallogr.* **68**, 352–367 (2012).

229. Liebschner, D. *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr. D Struct. Biol.* **75**, 861–877 (2019).

230. Cavallo, L., Kleinjung, J. & Fraternali, F. POPS: A fast algorithm for solvent accessible surface areas at atomic and residue level. *Nucleic Acids Res.* **31**, 3364–3366 (2003).

231. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

232. Kelman, L. M., O'Dell, W. B. & Kelman, Z. Unwinding 20 years of the archaeal minichromosome maintenance helicase. *J. Bacteriol.* **202**, (2020).

233. Cline, J., Braman, J. C. & Hogrefe, H. H. PCR fidelity of *Pfu* DNA polymerase and other thermostable DNA polymerases. *Nucleic Acids Res.* **24**, 3546–3551 (1996).

234. Sakakibara, N. *et al.* Cloning, purification, and partial characterization of the *Halobacterium* sp. NRC-1 minichromosome Maintenance (MCM) helicase. *Open Microbiol. J.* **2**, 13–17 (2008).

235. Gómez-Llorente, Y., Fletcher, R. J., Chen, X. S., Carazo, J. M. & San Martín, C. Polymorphism and double hexamer structure in the archaeal minichromosome maintenance (MCM) helicase from *Methanobacterium thermoautotrophicum*. *J. Biol. Chem.* **280**, 40909–40915 (2005).

236. Akanuma, S. *et al.* Establishment of mesophilic-like catalytic properties in a thermophilic enzyme without affecting its thermal stability. *Sci. Rep.* **9**, 9346 (2019).

237. Bishara, A. *et al.* High-quality genome sequences of uncultured microbes by assembly of read clouds. *Nat. Biotechnol.* (2018) doi:10.1038/nbt.4266.

238. Meagher, M. & Enemark, E. J. Structure of a double hexamer of the *Pyrococcus furiosus* minichromosome maintenance protein N-terminal domain. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **72**, 545–551 (2016).

239. Meagher, M., Spence, M. N. & Enemark, E. J. Structure of a dimer of the *Sulfolobus solfataricus* MCM N-terminal domain reveals a potential role in MCM ring opening. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **77**, 177–186 (2021).

240. Grainge, I. *et al.* Biochemical analysis of a DNA replication origin in the archaeon *Aeropyrum pernix*. *J. Mol. Biol.* **363**, 355–369 (2006).

241. Kurr, M. *et al. Methanopyrus kandleri*, gen. and sp. nov. represents a novel group of hyperthermophilic methanogens, growing at 110°C. *Arch. Microbiol.* **156**, 239–247 (1991).

242. Elkins, J. G. *et al.* A korarchaeal genome reveals insights into the evolution of the Archaea. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 8102–8107 (2008).

243. Walker, C. B. *et al. Nitrosopumilus maritimus* genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 8818–8823 (2010).

244. Pohlschroder, M. & Schulze, S. *Haloferax volcanii. Trends Microbiol.* **27**, 86–87 (2019).

245. Guan, Y. *et al.* Comparative genomics of the genus *Methanohalophilus*, including a newly isolated strain from Kebrit deep in the Red Sea. *Front. Microbiol.* **10**, 839 (2019).

246. Crits-Christoph, A. *et al.* Functional interactions of archaea, bacteria and viruses in a hypersaline endolithic community. *Environ. Microbiol.* **18**, 2064–2077 (2016).

247. Lambie, S. C. *et al.* The complete genome sequence of the rumen methanogen *Methanosarcina barkeri* CM1. *Stand. Genomic Sci.* **10**, 57 (2015).

248. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

249. Bae, B. *et al.* Insights into the architecture of the replicative helicase from the structure of an archaeal MCM homolog. *Structure* **17**, 211–222 (2009).

250. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

251. Pucci, B. *et al.* Modular organization of the *Sulfolobus solfataricus* mini-chromosome maintenance protein. *J. Biol. Chem.* **282**, 12574–12582 (2007).

252. Morales, E. S., Parcerisa, I. L. & Ceccarelli, E. A. A novel method for removing contaminant Hsp70 molecular chaperones from recombinant proteins. *Protein Sci.* **28**, 800–807 (2019).

253. Glasel, J. A. Validity of nucleic acid purities monitored by 260nm/280nm absorbance ratios. *Biotechniques* **18**, 62–63 (1995).

254. Whelan, F. *et al.* A flexible brace maintains the assembly of a hexameric replicative helicase during DNA unwinding. *Nucleic Acids Res.* **40**, 2271–2283 (2012).

255. Laue, T. M. & Stafford, W. F., 3rd. Modern applications of analytical ultracentrifugation. *Annu. Rev. Biophys. Biomol. Struct.* **28**, 75–100 (1999).

256. Lebowitz, J., Lewis, M. S. & Schuck, P. Modern analytical ultracentrifugation in protein science: a tutorial review. *Protein Sci.* **11**, 2067–2079 (2002).

257. Laue, T. M., Shah, B., Ridgeway, T. M. & Pelletier, S. L. Computer-aided interpretation of sedimentation data for proteins. (1992).

258. Dong, F., Gogol, E. P. & von Hippel, P. H. The phage T4-coded DNA replication helicase (gp41) forms a hexamer upon activation by nucleoside triphosphate. *J. Biol. Chem.* **270**, 7462–7473 (1995).

259. Wiegand, T. *et al.* The conformational changes coupling ATP hydrolysis and translocation in a bacterial DnaB helicase. *Nat. Commun.* **10**, 31 (2019).

260. Jean, N. L., Rutherford, T. J. & Löwe, J. FtsK in motion reveals its mechanism for double-stranded DNA translocation. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 14202–14208 (2020).

261. Sim, A. Y. L., Lipfert, J., Herschlag, D. & Doniach, S. Salt dependence of the radius of gyration and flexibility of single-stranded DNA in solution probed by small-angle x-ray scattering. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **86**, 021901 (2012).

262. Fu, Y., Slaymaker, I. M., Wang, J., Wang, G. & Chen, X. S. The 1.8-Å crystal structure of the N-terminal domain of an archaeal MCM as a right-handed filament. *J. Mol. Biol.* **426**, 1512–1523 (2014).

263. Dessau, M. A. & Modis, Y. Protein crystallization for X-ray crystallography. *J. Vis. Exp.* (2011) doi:10.3791/2285.

264. Heras, B. *et al.* Dehydration converts DsbG crystal diffraction from low to high resolution. *Structure* **11**, 139–145 (2003).

265. Lobley, C. M. C. *et al.* A generic protocol for protein crystal dehydration using the HC1b humidity controller. *Acta Crystallogr D Struct Biol* **72**, 629–640 (2016).

266. Tong, Y., Dong, A., Xu, X. & Wernimont, A. Salvage or recovery of failed targets by in situ proteolysis. *Methods Mol. Biol.* **1140**, 179–188 (2014).

267. Vuillard, L., Rabilloud, T., Leberman, R., Berthet-Colominas, C. & Cusack, S. A new additive for protein crystallization. *FEBS Lett.* **353**, 294–296 (1994).

268. Hassell, A. M. *et al.* Crystallization of protein-ligand complexes. *Acta Crystallogr. D Biol. Crystallogr.* **63**, 72–79 (2007).

269. Hoeppner, A., Schmitt, L. & Smits, S. H. J. Proteins and their ligands: Their importance and how to crystallize them. in *Advanced Topics on Crystal Growth* (ed. Ferreira, S. O.) (IntechOpen, 2013).

270. Vedadi, M. *et al.* Chemical screening methods to identify ligands that promote protein stability, protein crystallization, and structure determination. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 15835–15840 (2006).

271. Müller, I. Guidelines for the successful generation of protein-ligand complex crystals. *Acta Crystallogr D Struct Biol* **73**, 79–92 (2017).

272. Patel, A. *et al.* ATP as a biological hydrotrope. *Science* **356**, 753–756 (2017).

273. Hersch, G. L., Burton, R. E., Bolon, D. N., Baker, T. A. & Sauer, R. T. Asymmetric interactions of ATP with the AAA+ ClpX6 unfoldase: allosteric control of a protein machine. *Cell* **121**, 1017–1027 (2005).

274. Kim, Y.-C., Snoberger, A., Schupp, J. & Smith, D. M. ATP binding to neighbouring subunits and intersubunit allosteric coupling underlie proteasomal ATPase function. *Nat. Commun.* **6**, 8520 (2015).

275. Gorrec, F. The MORPHEUS II protein crystallization screen. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **71**, 831–837 (2015).

276. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).

277. Mevarech, M., Frolow, F. & Gloss, L. M. Halophilic enzymes: proteins with a grain of salt. *Biophys. Chem.* **86**, 155–164 (2000).

278. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).

279. Grant, B. J., Rodrigues, A. P. C., ElSawy, K. M., McCammon, J. A. & Caves, L. S. D. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics* **22**, 2695–2696 (2006).

280. Reed, C. J., Lewis, H., Trejo, E., Winston, V. & Evilia, C. Protein adaptations in archaeal extremophiles. *Archaea* **2013**, 373275 (2013).

281. Chan, C.-H., Yu, T.-H. & Wong, K.-B. Stabilizing salt-bridge enhances protein thermostability by reducing the heat capacity change of unfolding. *PLoS One* **6**, e21624 (2011).

282. Nair, P. Sequencing ancient DNA. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 2401 (2014).

283. Brown, E., Dessai, U., McGarry, S. & Gerner-Smidt, P. Use of whole-genome sequencing for food safety and public health in the United States. *Foodborne Pathog. Dis.* **16**, 441–450 (2019).

284. Yang, Y., Xie, B. & Yan, J. Application of next-generation sequencing technology in forensic science. *Genomics Proteomics Bioinformatics* **12**, 190–197 (2014).

285. Ekblom, R. & Galindo, J. Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* **107**, 1–15 (2011).

286. Esplin, E. D., Oei, L. & Snyder, M. P. Personalized sequencing and the future of medicine: discovery, diagnosis and defeat of disease. *Pharmacogenomics* **15**, 1771–1790 (2014).

287. Kraemer, M. U. G. *et al.* Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence. *Science* (2021) doi:10.1126/science.abj0113.

288. Fiedler, K., Lazzaro, S., Lutz, J., Rauch, S. & Heidenreich, R. mRNA Cancer Vaccines. *Recent Results Cancer Res.* **209**, 61–85 (2016).

289. Dolgin, E. Unlocking the potential of vaccines built on messenger RNA. *Nature* **574**, S10–S12 (2019).

290. Doroschak, K. *et al.* Rapid and robust assembly and decoding of molecular tags with DNA-based nanopore signatures. *Nat. Commun.* **11**, 5454 (2020).

291. Lin, K. N., Volkel, K., Tuck, J. M. & Keung, A. J. Dynamic and scalable DNA-based information storage. *Nat. Commun.* **11**, 2981 (2020).

292. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 560–564 (1977).

293. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463–5467 (1977).

294. Smith, L. M., Fung, S., Hunkapiller, M. W., Hunkapiller, T. J. & Hood, L. E. The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. *Nucleic Acids Res.* **13**, 2399–2412 (1985).

295. Ansorge, W., Sproat, B., Stegemann, J., Schwager, C. & Zenke, M. Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. *Nucleic Acids Res.* **15**, 4593–4602 (1987).

296. Prober, J. M. *et al.* A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238**, 336–341 (1987).

297. Kambara, H., Nishikawa, T., Katayama, Y. & Yamaguchi, T. Optimization of parameters in a DNA sequenator using fluorescence detection. *Biotechnology* **6**, 816–821 (1988).

298. Swerdlow, H. & Gesteland, R. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Res.* **18**, 1415–1419 (1990).

299. Luckey, J. A. *et al.* High speed DNA sequencing by capillary electrophoresis. *Nucleic Acids Res.* **18**, 4417–4421 (1990).

300. Crossley, B. M. *et al.* Guidelines for Sanger sequencing and molecular assay monitoring. *J. Vet. Diagn. Invest.* **32**, 767–775 (2020).

301. Staden, R. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res.* **6**, 2601–2610 (1979).

302. Mullis, K. *et al.* Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb. Symp. Quant. Biol.* **51 Pt 1**, 263–273 (1986).

303. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135–1145 (2008).

304. Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M. & Nyrén, P. Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.* **242**, 84–89 (1996).

305. Leamon, J. H. *et al.* A massively parallel PicoTiterPlate based platform for discrete picoliter-scale polymerase chain reactions. *Electrophoresis* **24**, 3769–3777 (2003).

306. Nyrén, P. Enzymatic method for continuous monitoring of DNA polymerase activity. *Anal. Biochem.* **167**, 235–238 (1987).

307. Mashayekhi, F. & Ronaghi, M. Analysis of read length limiting factors in Pyrosequencing chemistry. *Anal. Biochem.* **363**, 275–287 (2007).

308. Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L. & Welch, D. M. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* **8**, R143 (2007).

309. Fedurco, M., Romieu, A., Williams, S., Lawrence, I. & Turcatti, G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res.* **34**, e22 (2006).

310. Seo, T. S. *et al.* Four-color DNA sequencing by synthesis on a chip using photocleavable fluorescent nucleotides. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 5926–5931 (2005).

311. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).

312. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).

313. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).

314. Chen, Y.-C., Liu, T., Yu, C.-H., Chiang, T.-Y. & Hwang, C.-C. Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PLoS One* **8**, e62856 (2013).

315. Clarke, L. A., Rebelo, C. S., Gonçalves, J., Boavida, M. G. & Jordan, P. PCR amplification introduces errors into mononucleotide and dinucleotide repeat sequences. *Mol. Pathol.* **54**, 351–353 (2001).

316. Pfeiffer, F. *et al.* Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci. Rep.* **8**, 1–14 (2018).

317. Frommer, M. *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 1827–1831 (1992).

318. Levene, M. J. *et al.* Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299**, 682–686 (2003).

319. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).

320. Rhoads, A. & Au, K. F. PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics* **13**, 278–289 (2015).

321. Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **21**, 30 (2020).

322. Weirather, J. L. *et al.* Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Res.* **6**, 100 (2017).

323. Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nat. Biotechnol.* **34**, 518–524 (2016).

324. Coulter, W. H. Means for counting particles suspended in a fluid. *US Patent* (1953).

325. Cao, C. *et al.* Single-molecule sensing of peptides and nucleic acids by engineered aerolysin nanopores. *Nat. Commun.* **10**, 4918 (2019).

326. Storm, A. J., Chen, J. H., Ling, X. S., Zandbergen, H. W. & Dekker, C. Fabrication of solid-state nanopores with single-nanometre precision. *Nat. Mater.* **2**, 537–540 (2003).

327. Schneider, G. F. *et al.* DNA Translocation through Graphene Nanopores. *Nano Lett.* **10**, 3163–3167 (2010).

328. Venta, K. *et al.* Differentiation of short, single-stranded DNA homopolymers in solid-state nanopores. *ACS Nano* **7**, 4629–4636 (2013).

329. Kim, M. J., Wanunu, M., Bell, D. C. & Meller, A. Rapid fabrication of uniformly sized nanopores and nanopore arrays for parallel DNA analysis. *Adv. Mater.* **18**, 3149–3153 (2006).

330. Venkatesan, B. M. & Bashir, R. Nanopore sensors for nucleic acid analysis. *Nat. Nanotechnol.* **6**, 615–624 (2011).

331. Plesa, C., van Loo, N. & Dekker, C. DNA nanopore translocation in glutamate solutions. *Nanoscale* **7**, 13605–13609 (2015).

332. Yeh, L.-H., Zhang, M., Joo, S. W. & Qian, S. Slowing down DNA translocation through a nanopore by lowering fluid temperature. *Electrophoresis* **33**, 3458–3465 (2012).

333. Larkin, J. *et al.* Slow DNA transport through nanopores in hafnium oxide membranes. *ACS Nano* **7**, 10121–10128 (2013).

334. Banerjee, S. *et al.* Slowing DNA transport using graphene-DNA interactions. *Adv. Funct. Mater.* **25**, 936–946 (2015).

335. Wang, C., Sensale, S., Pan, Z., Senapati, S. & Chang, H.-C. Slowing down DNA translocation through solid-state nanopores by edge-field leakage. *Nat. Commun.* **12**, 140 (2021).

336. Song, L. *et al.* Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore. *Science* **274**, 1859–1866 (1996).

337. Kasianowicz, J. J., Brandin, E., Branton, D. & Deamer, D. W. Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 13770–13773 (1996).

338. Cockroft, S. L., Chu, J., Amorin, M. & Ghadiri, M. R. A single-molecule nanopore device detects DNA polymerase activity with single-nucleotide resolution. *J. Am. Chem. Soc.* **130**, 818–820 (2008).

339. Lieberman, K. R. *et al.* Processive replication of single DNA molecules in a nanopore catalyzed by phi29 DNA polymerase. *J. Am. Chem. Soc.* **132**, 17961–17972 (2010).

340. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228–232 (2016).

341. Heron AJ, Alves DA, Clarke J, Crawford ML, Garalde DR, Hall G, Turner DJ, White J. Enzyme stalling method. *World Patent* (2014).

342. Rang, F. J., Kloosterman, W. P. & de Ridder, J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.* **19**, 90 (2018).

343. Craig, J. M. *et al.* Determining the effects of DNA sequence on Hel308 helicase translocation along single-stranded DNA using nanopore tweezers. *Nucleic Acids Res.* **47**, 2506–2513 (2019).

344. Caldwell, C. C. & Spies, M. Helicase SPRNTing through the nanopore. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 11809–11811 (2017).

345. de Lannoy, C., de Ridder, D. & Risse, J. The long reads ahead: de novo genome assembly using the MinION. *F1000Res.* **6**, 1083 (2017).

346. Goyal, P. *et al.* Structural and mechanistic insights into the bacterial amyloid secretion channel CsgG. *Nature* **516**, 250–253 (2014).

347. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* **20**, 129 (2019).

348. Payne, A., Holmes, N., Rakyan, V. & Loose, M. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics* **35**, 2193–2198 (2019).

349. Prall, T. M. *et al.* Consistent ultra-long DNA sequencing with automated slow pipetting. *BMC Genomics* **22**, 182 (2021).

350. Gilpatrick, T. *et al.* Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nat. Biotechnol.* **38**, 433–438 (2020).

351. Brown, C. Nobody Expects the Strandish Exposition. https://nanoporetech.com/resource-centre/video/lc21/nobody-expects-the-strandish-exposition (2021).

352. Loose, M., Malla, S. & Stout, M. Real-time selective sequencing using nanopore technology. *Nat. Methods* **13**, 751–754 (2016).

353. Kovaka, S., Fan, Y., Ni, B., Timp, W. & Schatz, M. C. Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. *Nat. Biotechnol.* **39**, 431–441 (2021).

354. Van der Verren, S. E. *et al.* A dual-constriction biological nanopore resolves homonucleotide sequences with high fidelity. *Nat. Biotechnol.* **38**, 1415–1420 (2020).

355. Maglia, G., Restrepo, M. R., Mikhailova, E. & Bayley, H. Enhanced translocation of single DNA molecules through alpha-hemolysin nanopores by manipulation of internal charge. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 19720–19725 (2008).

356. You, Z. & Masai, H. DNA binding and helicase actions of mouse MCM4/6/7 helicase. *Nucleic Acids Res.* **33**, 3033–3047 (2005).

357. Nova, I. C. *et al.* Investigating asymmetric salt profiles for nanopore DNA sequencing with biological porin MspA. *PLoS One* **12**, e0181599 (2017).

358. Hofmeister, F. *Zur Lehre von der Wirkung der Salze: Zweite Mittheilung*. (1888).

359. Göpfrich, K., Kulkarni, C. V., Pambos, O. J. & Keyser, U. F. Lipid nanobilayers to host biological nanopores for DNA translocations. *Langmuir* **29**, 355–364 (2013).

360. Stellwagen, E. & Stellwagen, N. C. Electrophoretic Mobility of DNA in Solutions of High Ionic Strength. *Biophys. J.* **118**, 2783–2789 (2020).

361. Haugland, G. T., Rollor, C. R., Birkeland, N.-K. & Kelman, Z. Biochemical characterization of the minichromosome maintenance protein from the archaeon *Thermoplasma acidophilum*. *Extremophiles* **13**, 81–88 (2009).

362. Cortez, D. Replication-Coupled DNA Repair. *Mol. Cell* **74**, 866–876 (2019).

363. Özeş, A. R., Feoktistova, K., Avanzino, B. C., Baldwin, E. P. & Fraser, C. S. Real-time fluorescence assays to monitor duplex unwinding and ATPase activities of helicases. *Nat. Protoc.* **9**, 1645–1661 (2014).

364. Stranges, P. B. *et al.* Design and characterization of a nanopore-coupled polymerase for single-molecule DNA sequencing by synthesis on an electrode array. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E6749–E6756 (2016).

365. Vieille, C. & Zeikus, G. J. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiol. Mol. Biol. Rev.* **65**, 1–43 (2001).

366. Onwubiko, N. O. *et al.* SV40 T antigen interactions with ssDNA and replication protein A: a regulatory role of T antigen monomers in lagging strand DNA replication. *Nucleic Acids Res.* **48**, 3657–3677 (2020).

367. Mogk, A. *et al.* Roles of individual domains and conserved motifs of the AAA+ chaperone ClpB in oligomerization, ATP hydrolysis, and chaperone activity. *J. Biol. Chem.* **278**, 17615–17624 (2003).

368. Yu, R. C., Hanson, P. I., Jahn, R. & Brünger, A. T. Structure of the ATP-dependent oligomerization domain of N-ethylmaleimide sensitive factor complexed with ATP. *Nat. Struct. Biol.* **5**, 803–811 (1998).

369. Mukherjee, J. & Gupta, M. N. Increasing importance of protein flexibility in designing biocatalytic processes. *Biotechnology Reports* **6**, 119–123 (2015).

370. Kumar, S., Tsai, C. J. & Nussinov, R. Factors enhancing protein thermostability. *Protein Eng.* **13**, 179–191 (2000).

371. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

372. Flower, T. G. & Hurley, J. H. Crystallographic molecular replacement using an in silico-generated search model of SARS-CoV-2 ORF8. *Protein Sci.* **30**, 728–734 (2021).

373. Mas, G. *et al.* Structural investigation of a chaperonin in action reveals how nucleotide binding regulates the functional cycle. *Sci. Adv.* **4**, eaau4196 (2018).

374. Cernooka, E., Rumnieks, J., Tars, K. & Kazaks, A. Structural basis for DNA recognition of a single-stranded DNA-binding protein from Enterobacter phage Enc34. *Sci. Rep.* **7**, 1–10 (2017).

375. Vetro, A. *et al.* MCM5: a new actor in the link between DNA replication and Meier-Gorlin syndrome. *Eur. J. Hum. Genet.* **25**, 646–650 (2017).

376. Allen, G. S. & Stokes, D. L. Modeling, docking, and fitting of atomic structures to 3D maps from cryo-electron microscopy. *Methods Mol. Biol.* **955**, 229–241 (2013).

377. Dandey, V. P. *et al.* Time-resolved cryo-EM using Spotiton. *Nat. Methods* **17**, 897–900 (2020).

378. Gauto, D. F. *et al.* Integrated NMR and cryo-EM atomic-resolution structure determination of a half-megadalton enzyme complex. *Nat. Commun.* **10**, 2697 (2019).

379. Kerfah, R., Plevin, M. J., Sounier, R., Gans, P. & Boisbouvier, J. Methyl-specific isotopic labeling: a molecular tool box for solution NMR studies of large proteins. *Curr. Opin. Struct. Biol.* **32**, 113–122 (2015).

380. Pellecchia, M. *et al.* Perspectives on NMR in drug discovery: a technique comes of age. *Nat. Rev. Drug Discov.* **7**, 738–745 (2008).

381. Boisbouvier, J. & Kay, L. E. Advanced isotopic labeling for the NMR investigation of challenging proteins and nucleic acids. *J. Biomol. NMR* **71**, 115–117 (2018).

382. Hoffmann, B., Löhr, F., Laguerre, A., Bernhard, F. & Dötsch, V. Protein labeling strategies for liquid-state NMR spectroscopy using cell-free synthesis. *Prog. Nucl. Magn. Reson. Spectrosc.* **105**, 1–22 (2018).

383. Acton, T. B. *et al.* Preparation of protein samples for NMR structure, function, and small-molecule screening studies. *Methods Enzymol.* **493**, 21–60 (2011).

384. Azatian, S. B., Kaur, N. & Latham, M. P. Increasing the buffering capacity of minimal media leads to higher protein yield. *J. Biomol. NMR* **73**, 11–17 (2019).

385. Vahidi, S. *et al.* An allosteric switch regulates *Mycobacterium tuberculosis* ClpP1P2 protease function as established by cryo-EM and methyl-TROSY NMR. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 5895–5906 (2020).

386. Galiakhmetov, A. R., Kovrigina, E. A., Xia, C., Kim, J.-J. P. & Kovrigin, E. L. Application of methyl-TROSY to a large paramagnetic membrane protein without perdeuteration: 13C-MMTS-labeled NADPH-cytochrome P450 oxidoreductase. *J. Biomol. NMR* **70**, 21–31 (2018).

387. Sierecki, E. *et al.* Nanomolar oligomerization and selective co-aggregation of α-synuclein pathogenic mutants revealed by single-molecule fluorescence. *Sci. Rep.* **6**, 37630 (2016).

388. Wang, Q., Serban, A. J., Wachter, R. M. & Moerner, W. E. Single-molecule diffusometry reveals the nucleotide-dependent oligomerization pathways of *Nicotiana tabacum* Rubisco activase. *J. Chem. Phys.* **148**, 123319 (2018).

389. Ribeck, N., Kaplan, D. L., Bruck, I. & Saleh, O. A. DnaB helicase activity is modulated by DNA geometry and force. *Biophys. J.* **99**, 2170–2179 (2010).

390. Ribeck, N. & Saleh, O. A. DNA unwinding by ring-shaped T4 helicase gp41 is hindered by tension on the occluded strand. *PLoS One* **8**, e79237 (2013).

391. Muzard, J., Martinho, M., Mathé, J., Bockelmann, U. & Viasnoff, V. DNA translocation and unzipping through a nanopore: some geometrical effects. *Biophys. J.* **98**, 2170–2178 (2010).

392. Brown, C. Single molecule "strand" sequencing using protein nanopores and scalable electronic devices. (2012).

393. Sauer-Budge, A. F., Nyamwanda, J. A., Lubensky, D. K. & Branton, D. Unzipping kinetics of double-stranded DNA in a nanopore. *Phys. Rev. Lett.* **90**, 238101 (2003).

394. Sun, K. *et al.* Active DNA unwinding and transport by a membrane-adapted helicase nanopore. *Nat. Commun.* **10**, 5083 (2019).

395. Cressiot, B. *et al.* Porphyrin-assisted docking of a thermophage portal protein into lipid bilayers: nanopore engineering and characterization. *ACS Nano* **11**, 11931–11945 (2017).

396. Shi, H., Rampello, A. J. & Glynn, S. E. Engineered AAA+ proteases reveal principles of proteolysis at the mitochondrial inner membrane. *Nat. Commun.* **7**, 13301 (2016).

397. Zaccai, N. R. *et al.* A de novo peptide hexamer with a mutable channel. *Nat. Chem. Biol.* **7**, 935–941 (2011).

398. Xu, C. *et al.* Computational design of transmembrane pores. *Nature* **585**, 129–134 (2020).

399. Beeder, J., Nilsen, R. K., Rosnes, J. T., Torsvik, T. & Lien, T. *Archaeoglobus fulgidus* isolated from hot North Sea oil field waters. *Appl. Environ. Microbiol.* **60**, 1227–1231 (1994).

400. Sako, Y. *et al. Aeropyrum pernix* gen. nov., sp. nov., a novel aerobic hyperthermophilic archaeon growing at temperatures up to 100 degrees C. *Int. J. Syst. Bacteriol.* **46**, 1070–1077 (1996).

401. Robinson, J. L. *et al.* Growth kinetics of extremely halophilic archaea (family halobacteriaceae) as revealed by arrhenius plots. *J. Bacteriol.* **187**, 923–929 (2005).

402. Miller-Coleman, R. L. *et al.* Korarchaeota diversity, biogeography, and abundance in Yellowstone and Great Basin hot springs and ecological niche modeling based on machine learning. *PLoS One* **7**, e35964 (2012).

403. Golyshina, O. V. *et al.* The novel extremely acidophilic, cell-wall-deficient archaeon *Cuniculiplasma divulgatum* gen. nov., sp. nov. represents a new family, Cuniculiplasmataceae fam. nov., of the order Thermoplasmatales. *Int. J. Syst. Evol. Microbiol.* **66**, 332–340 (2016).

404. Boone, D. R. *et al.* Isolation and Characterization of *Methanohalophilus portucalensis* sp. nov. and DNA Reassociation Study of the Genus Methanohalophilus. *Int. J. Syst. Bacteriol.* **43**, 430–437 (1993).

405. Liu, C. *et al.* Comparative proteomic analysis of *Methanothermobacter thermautotrophicus* reveals methane formation from H2 and CO2 under different temperature conditions. *Microbiologyopen* **8**, e00715 (2019).

406. Paper, W. *et al. Ignicoccus hospitalis* sp. nov., the host of "*Nanoarchaeum equitans*." *Int. J. Syst. Evol. Microbiol.* **57**, 803–808 (2007).

407. Qin, W. *et al. Nitrosopumilus maritimus* gen. nov., sp. nov., *Nitrosopumilus cobalaminigenes* sp. nov., *Nitrosopumilus oxyclinae* sp. nov., and *Nitrosopumilus ureiphilus* sp. nov., four marine ammonia-oxidizing archaea of the phylum Thaumarchaeota. *Int. J. Syst. Evol. Microbiol.* **67**, 5067–5079 (2017).

408. Weinberg, M. V., Schut, G. J., Brehm, S., Datta, S. & Adams, M. W. W. Cold shock of a hyperthermophilic archaeon: *Pyrococcus furiosus* exhibits multiple responses to a suboptimal growth temperature with a key role for membrane-bound glycoproteins. *J. Bacteriol.* **187**, 336–348 (2005).

409. Zaparty, M. *et al.* "Hot standards" for the thermoacidophilic archaeon *Sulfolobus solfataricus*. *Extremophiles* **14**, 119–142 (2010).